state of
Machine learning at Wikipedia

Santhosh Thottingal
Principal Software Engineer, Wikimedia Foundation
Projects using Machine Learning at Wikipedia

- Use cases
- Guiding principles
- Product design
- Challenges
- Impact
Machine Translation
Easier translation of articles between languages.

Reusing work done by another community (notability, verifiability,...) lowers the risk of deletion.

It also expands the number of people who can contribute, as it requires a different set of skills compared to writing completely new content.
Language gap

English
6M

German
2M

Indonesian
657K

Telugu
84K
Plants in space are plants grown in outer space typically in a weightless but pressurized controlled environment in specific space gardens. In the context of human spaceflight, they can be consumed
Human curation of Machine translation

Oxygen is the chemical element with the symbol \( \text{O} \) and atomic number 8. It is a member of the chalcogen group in the periodic table, a highly reactive nonmetal, and an oxidizing agent that readily forms oxides with most elements as well as with other compounds. Oxygen is Earth's most abundant element, and after hydrogen and helium, it is the third-most abundant element in the universe. At standard temperature and pressure, two atoms of the element bind to form dioxygen, a colorless and odorless diatomic gas with the formula \( \text{O}_2 \). Dioxygen gas currently constitutes 20.95% of the Earth's atmosphere, though this has changed considerably over long periods of time. Oxygen makes up almost half of the Earth's crust in the form of oxides.\(^1\)

All plants, animals, and fungi need oxygen for cellular respiration, which extracts energy by the reaction of oxygen with molecules derived from food and produces carbon dioxide as a waste product. In
Machine Translation misuse prevention

Wikipedia Translate page

Your translation cannot be published because it contains too much unmodified text. View issues

Plants in space

English

view page

4 categories

Bahasa Indonesia

No categories

Issues

Your translation contains a total of 85% of unmodified text.

Automatic translation is provided only as a starting point. Make sure that the content is accurate and reads naturally in your language.

Your translation cannot be published without further editing.

Learn more
Content Translation

Impact

1.6 Million+

Articles published by translating
Combined, this would be a top 10 wikipedia

4%

Low deletion rate
Compared with 13% deletion rate of articles created without translation
Machine translation services

- Apertium
- Google
- Yandex
- LingoCloud
- Elia
- MinT
MinT: Supporting underserved languages with open machine translation

13 June 2023 by Pau Giner

Our vision is a world in which every single human being can freely share in the sum of all knowledge. Machine translation has the potential to help us achieve that vision by enabling more people to contribute content to Wikipedia in their native or preferred languages.

Content Translation, the tool used by Wikipedia editors to translate over one and a half million articles, uses machine translation as a starting point when it is available, making sure to keep humans in the loop by encouraging them to improve the initial translation and controlling how much it is edited. In this case, providing automation while keeping the humans in control helps Wikipedia editors to become more productive while producing quality content. However, not all languages have good quality machine translation available for editors to benefit from.

We are launching MinT in order to expand the current machine translation support. MinT ("Machine in Translation") is a new translation service by the Wikimedia Foundation
MinT

A self hosted Neural Machine Translation service by Wikipedia

Serves multiple MT models and provides a single API interface

- **NLLB**
  Generic model by Meta

- **NLLB-Wikipedia**
  Wikipedia Optimized models

- **OpusMT**
  For low resource languages

- **SoftCatala**
  For English-Catalan

- **IndicTrans2**
  for 22 indic languages and english
MinT

A self hosted Neural Machine Translation service by Wikipedia

Serves multiple MT models and provides a single API interface

198 Languages

35924 Language pairs
02 Knowledge Integrity

AI article & edit quality assessment, vandalism patrol/prevention
Artificial Intelligence Aims to Make Wikipedia Friendlier and Better

The nonprofit behind Wikipedia is turning to machine learning to combat a long-standing decline in the number of editors.

By Tom Simonite  December 1, 2015

Software trained to know the difference between an honest mistake and intentional vandalism is being rolled out in an effort to make editing Wikipedia less psychologically bruising. It was developed by the Wikimedia Foundation, the nonprofit organization that supports Wikipedia.

AARON HALFAKER*, Microsoft, USA
R. STUART GEIGER**, University of California, San Diego, USA

Algorithmic systems—from rule-based bots to machine learning classifiers—have a long history of supporting the essential work of content moderation and other curation work in peer production projects. From countervandalism to task routing, basic machine prediction has allowed open knowledge projects like Wikipedia to scale to the largest encyclopedia in the world, while maintaining quality and consistency. However, conversations about how quality control should work and what role algorithms should play have generally been led by the expert engineers who have the skills and resources to develop and modify these complex algorithmic systems. In this paper, we describe ORES: an algorithmic scoring service that supports real-time scoring of wiki edits using multiple independent classifiers trained on different datasets. ORES decouples several activities that have typically all been performed by engineers: choosing or curating training data, building models to serve predictions, auditing predictions, and developing interfaces or automated agents that act on those predictions. This meta-algorithmic system was designed to open up socio-technical conversations about algorithms in Wikipedia to a broader set of participants. In this paper, we discuss the theoretical mechanisms of social change ORES enables and detail case studies in participatory machine learning around ORES from the 5 years since its deployment.

CCS Concepts: • Networks → Online social networks; • Computing methodologies → Supervised learning by classification; • Applied computing → Sociology; • Software and its engineering → Software design techniques; • Computer systems organization → Cloud computing;

Additional Key Words and Phrases: Wikipedia; Reflection; Machine learning; Transparency; Fairness; Algorithms; Governance

AGM Reference Format:
Objective Revision Evaluation Service (ORES)

Edit quality: unknown good needs review damaging
Prediction Threshold preferences

Prediction Threshold

- May have problems (flags most problem edits but includes many false positives)
- Likely have problems (medium probability)
- Very likely have problems (flags few false positives but finds a smaller % of problem edits)
- Likely have problems (medium probability)

Change the "threshold" setting to make the options below broader or more selective.

- Highlight likely problem edits with colors and an "r" for "needs review"
- Show only likely problem edits (and hide probably good edits)
Revert Risk is now a service hosted in Lift Wing system.
<table>
<thead>
<tr>
<th>Technology</th>
<th>Revert Risk</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Language Agnostic</td>
</tr>
<tr>
<td>Characteristics</td>
<td>Can run in all Wikipedia Language Editions</td>
</tr>
<tr>
<td></td>
<td>Mainly Based on Meta-Data</td>
</tr>
<tr>
<td>Training Data</td>
<td>Implicit Annotations (past reverts)</td>
</tr>
<tr>
<td>Pros</td>
<td>Fast</td>
</tr>
<tr>
<td></td>
<td>Light on resources usage</td>
</tr>
<tr>
<td></td>
<td>Covers all languages</td>
</tr>
<tr>
<td></td>
<td>Lower accuracy on IP Edits.</td>
</tr>
<tr>
<td></td>
<td>Basic NLP power.</td>
</tr>
</tbody>
</table>
03 Structured Tasks

“add a link” and “add an image” to help new editors get started with easy tasks

WIKIMEDIA FOUNDATION
Add a link

Newcomer task

New editors review machine suggestions for making words in one Wikipedia article link to other Wikipedia articles.
“Add a link” is available via the Suggested edits feed on Homepage

Onboarding 1: Explains value and impact of this small contribution

Onboarding 2: “Human in the loop” reviews machine suggestions
Evaluating machine suggestions of specific text to make into links...

...as an easy and fast way of contributing

Encouragement to do more post-edit
Add a link

Algorithm
devolved by the WMF Research team automatically generates link recommendations for Wikipedia articles.

The model's performance is evaluated based on precision and recall. Based on manual feedback from editors, hard-coded rules are implemented to avoid unwanted linking (e.g. links to dates).

Machine-learning model

The model predicts the probability of a link in the article (anchor-text + target-page).

- Identify unlinked text that could potentially contain a link
- Generate candidate links by looking up existing links with this text
- Score candidates and pick the most likely as the target-page

Training

The model is trained with existing sentences of millions of positive (what is linked) and negative examples (what is not linked).
Add a link

**Impact**

**Activation**
Increase in probability that a newcomer makes their first edit

**Retention**
Increase in probability that a new editor is retained

**Productivity**
Increase in the number of edits newcomers make during their first couple of weeks

**Reverts**
Decrease in revert rates compared to baseline newcomer edits (although this comparison is imperfect)
Moderation burden
Burden on patrollers: More edits = more work for patrollers.

Wider language support
Language characteristic and complexity affects parsing the sentences. ML models perform relatively poor on low resource languages

Data scarcity
Data scarcity for small wikis cause less performant ML models
Optical Character Recognition

Document digitization
<table>
<thead>
<tr>
<th>System</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tesseract</td>
<td>Self hosted Open source OCR engine</td>
</tr>
<tr>
<td>Transkribus</td>
<td>Externally hosted OCR system. Used for digitizing historical and handwritten documents relevant for Wikisource.</td>
</tr>
<tr>
<td>Google Cloud Vision OCR</td>
<td>External service</td>
</tr>
</tbody>
</table>
05 Lift Wing

Machine learning hosting platform
Scalability
Microservices can be independently scaled based on demand, allowing for more efficient resource utilization and improved performance.

Flexibility
Microservices architecture enables the use of different languages, and frameworks for each model service, providing greater flexibility in development.

Faster Deployment
Smaller codebases and independent deployment of microservices enable faster and more frequent releases, accelerating release to production.

Fault Isolation:
Failure in one microservice is less likely to impact the entire system, improving overall system resilience and uptime.
Lift Wing

- API Gateway
- Model Cards
- Community Model Governance
- KServe
- k8s

Production environment
More machine learning use cases

**Topic Classification system**
Language agnostic link-based article topic classification - Label a given wikipedia article in any language to a topic

**Language identification**
Given a content snippet, this model can detect the language of the snippet. Supports ~200 languages.

**Section alignment**
Identify missing section between two existing article pairs in any languages. Used in Section translation feature of Content Translation
<table>
<thead>
<tr>
<th>Service</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Machine Translation</td>
<td>Google, Yandex, Elia machine translation services in Content Translation</td>
</tr>
<tr>
<td>Text to Speech</td>
<td>The Phonos extension to read IPA use Google TTS API*</td>
</tr>
<tr>
<td>Machine Vision</td>
<td>Machine Vision use Google's Cloud Vision API to identify potential depicts statements for images in Commons.</td>
</tr>
<tr>
<td>Image to Text (OCR)</td>
<td>Wikisource use Google's OCR API, and Transkribus</td>
</tr>
<tr>
<td>Content moderation</td>
<td>Community Tech's CopyPatrol make use of Turnitin's API for detecting plagiarism between passages added to Wikipedia and external documents</td>
</tr>
<tr>
<td>Named Entity Recognition</td>
<td>Architecture team used Rosette to identify Wikidata items from text</td>
</tr>
</tbody>
</table>
Model Cards

to make open source, transparent, human-centered machine learning

on-wiki model cards for every model hosted on WMF servers

- Use case, users
- Training data
- Ethical considerations
- Owners
- License
- Model architecture
How can we predict what general topic an article is in, and do so consistently across many languages? Answering this question would be useful for various analyses of Wikipedia dynamics. However, it is difficult to group a very diverse range of Wikipedia articles into coherent, consistent topics manually across all Wikipedia projects.

This model is a new, language-agnostic approach to predicting which topic an article might be relevant to. It uses the wikilinks in a given Wikipedia article to predict which (0 to many) of a set of 64 topics are relevant to a given article. For example, Mount Everest might reasonably be associated with South Asia, East Asia, Sports, Earth, and the Environment.

The training data for this model was over 30 million Wikipedia articles spanning all languages on Wikipedia. Each article was represented as the list of wikidata items associated with its outlinks. This data originated from the editing activities of Wikipedia and Wikidata editors, and was collected in an automated fashion.

This model is deployed on LiftWing. Right now, it can be publicly accessed through a beta testing site. This model may be useful for high-level analyses of Wikipedia dynamics (pageviews, article quality, edit trends), filtering articles, and cross-language analytics. It should not be used for projects outside of Wikipedia, namespaces outside of 0, disambiguations, or redirects.
Machine learning at Wikipedia

Thank You
Santhosh Thottingal
La piedra de Rosetta es un fragmento de una antigua estela egipcia de granodiorita inscrita con un decreto publicado en Menfis en el año 196 a. C. en nombre del faraón Ptolomeo V. El decreto aparece en tres escrituras distintas: el texto superior en jeroglíficos egipcios, la parte intermedia en escritura demótica y la inferior en griego antiguo. Gracias a que presenta esencialmente el mismo contenido en las tres inscripciones, con diferencias menores entre ellas, esta piedra facilitó la clave para el entendimiento moderno de los jeroglíficos egipcios.

Originalmente dispuesta dentro de un templo, la estela fue probablemente trasladada durante la época paleocristiana o la Edad Media y finalmente usada como material de construcción en una fuente cerca de la localidad de Rashid (Rosetta), en el delta del Nilo. Allí fue hallada en 1799 por el soldado Pierre-François Bouchard durante la campaña francesa en Egipto. Las tropas británicas derrotaron a las francesas en Egipto en 1801 y la piedra original acabó en posesión inglesa bajo la Capitulación de Alejandría. Transportada a Londres, fue expuesta al público desde 1820 en el Museo Británico, donde es la pieza más visitada.

Debido a que fue el primer texto plurilingüe antiguo descubierto en tiempos modernos, la Piedra de Rosetta despertó el interés público por su potencial para descifrar la hasta entonces intangible escritura jeroglífica egipcia, y en consecuencia sus copias litográficas y de yeso comenzaron a circular entre los museos y los eruditos europeos. La primera traducción completa del texto en griego antiguo apareció en 1803, pero no fue hasta 1822 cuando Jean-François Champollion anunció en París el descifrado de los textos jeroglíficos egipcios, mucho antes de que los lingüistas fueran capaces de leer con seguridad otras inscripciones y textos del antiguo Egipto. Los principales avances de la traducción se debieron a la combinación de la ciencia clásica, el arqueología, la filología, la historia antigua, la arqueología e incluso la química, que permitieron descifrar los caracteres jeroglíficos antiguos.

Más tarde se descubrieron dos copias fragmentarias del mismo decreto, y en la actualidad se conocen varias inscripciones egipcias bilingües y trilingües, incluidos los decretos Ptolomeicos, como el Decreto de Canopus del 238 a. C. y el Decreto de Menfis de Ptolomeo IV, del 218 a. C. Por ello, aunque la Piedra de Rosetta no es única, fue un hito crucial para el desciframiento de los jeroglíficos egipcios.