Automatically Labeling Low Quality Content on Wikipedia by Leveraging Patterns in Editing Behavior

Sumit Asthana, Sabrina Tobar Thommel, Dr. Aaron Halfaker, Dr. Nikola Banovic

WMF research showcase, March 2022
Wikipedia’s Goal as an encyclopedia

Wikipedia aims to be an encyclopedia of high quality with well written, grammatically correct and neutral articles.
English Wikipedia article quality statistics

Number of articles by quality

- ~0% best quality (FA)
- ~1% second best quality (GA)

Quality increases
Too many articles and too few editors has made article improvement a challenging task
Article improvement – add, assess, improve
We need automation

'Gala' is an early-ripening firm-fleshed edible potato, for which good eating and keeping qualities are claimed. It was developed by the based plant breeder Norika. In 2010, it was chosen as Thuringia's potato of the year in a public contest to which nearly 1000 consumers contributed, and in which 'Laura' was second place.
Automating article improvements using ML

Data labeling
Crowdsourcing or Featured articles
- Positive examples
- Negative examples
- Training & Testing Data

Classifier Training & Testing
- Training Data
- Testing Data
- Past research datasets
- Sequence-based RNN model
- Validation results

Classifier Deployment
- Sentences from current Wikipedia articles
- Sequence-based RNN model
- Sentences that need improvement
Automating article improvements using ML
Automating article improvements using ML
Automating article improvements using ML
Challenges to automating article improvement using ML
A new data quality labeling approach

A labeling approach that leverages implicit editing behavior to automatically label the quality of Wikipedia sentences
Automated labeling of sentence quality
Automated labeling of sentence quality
was born on July 27th, 1948 into the suburbs of... [0-9]+

...the compound is harmful {{citation-needed}}... {{[a-z]+}}
“It was during 1996 season that Ricketson was first picked for City Origin due to his outstanding defence.”
• Many Wikipedia edits are performed with a specific intention
  • e.g. adding citation, removing bias, clarifying content.

• Sentences modified in Wikipedia edits can be used as positive examples of sentence improvement.
Semantic categories demonstrated

• Citations: adding a citation

• Point-of-view: rewrite using encyclopedic, neutral tone; remove bias; apply due weight.

• Clarification: specify or explain an existing fact or meaning by example or discussion without adding new information.
Lateral flow tests are devices used for detecting the presence or absence of a target analyte in a sample, without the need for specialized equipment. They are widely used for home testing, laboratory use, and for medical diagnostics. A home pregnancy test is a well-known application of lateral flow tests.
Semantic category rules on Wikipedia edits

Lateral flow test: Difference between revisions

From Wikipedia, the free encyclopedia

-- Previous edit

Revision as of 02:55, 9 June 2013 (edit)
Hipal (talk | contribs)

m (Reverted edits by 131.111.184.97 (talk) to last version by 128.243.253.112)

Intent: Citation

Line 1:

{{Context|date=October 2009}}

"Lateral flow tests""""<ref>
[[Immunochromatographic]] Assays are simple devices intended to detect the presence (or absence) of a target analyte in sample (matrix) without the need for specialized and costly equipment, though many lab based applications exist that are supported by reading equipment. Typically, these tests are used for medical diagnostics either for home testing, [[point of care]] testing, or laboratory use. A widely spread and well known application is the home pregnancy test.

Revision as of 14:47, 13 August 2013 (edit (undo))
Yetisen (talk | contribs)

m (A reference was added.)

Line 1:

{{Context|date=October 2009}}

"Lateral flow tests""""<ref>
[[Immunochromatographic]] Assays are simple devices intended to detect the presence (or absence) of a target analyte in sample (matrix) without the need for specialized and costly equipment, though many lab based applications exist that are supported by reading equipment. <ref name="Yetisen2013">{{cite journal | author = Yetisen A. K. | year = 2013 | title = Paper-based microfluidic point-of-care diagnostic devices | url = http://pubs.rsc.org/en/content/articlelanding/2013/lf/c3lc50169b | journal = Lab on a Chip | volume= 13 | issue= 12 | pages= 2210–2251 | doi=10.1039/C3LC50169H}}</ref> Typically, these tests are used for medical diagnostics either for home testing, [[point of care]] testing, or laboratory use. A widely spread and well known application is the home pregnancy test.
### Semantic category rules – citations

<table>
<thead>
<tr>
<th>Rule</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inserted contains “&lt;ref&gt;”</td>
<td>Presence of citation tag in added content</td>
</tr>
<tr>
<td>NOT is_citation_inserted_or_deleted</td>
<td>No &lt;ref&gt; tag inserted or removed</td>
</tr>
<tr>
<td>NOT is_template_inserted_or_deleted</td>
<td>No templates inserted or removed</td>
</tr>
<tr>
<td>NOT is_wikilink_inserted_or_deleted</td>
<td>No wikilinks inserted or removed</td>
</tr>
<tr>
<td>NOT is_infobox_inserted_or_deleted</td>
<td>No infoboxes inserted or removed</td>
</tr>
<tr>
<td>NOT is_multiline_inserted_or_deleted</td>
<td>Multiple lines not changed</td>
</tr>
</tbody>
</table>
Lateral flow test: Difference between revisions

From Wikipedia, the free encyclopedia

Revision as of 02:55, 9 June 2013 (edit)
Hipal (talk | contribs)
(m (Reverted edits by 131.111.184.97 (talk) to last version by 128.243.253.112))
← Previous edit

Revision as of 14:47, 13 August 2013 (edit) (undo)
Yetisen (talk | contribs)
(m (A reference was added.))
Next edit →

Intent: Citation

Line 1:

"needs citations": 1


Assays are simple devices intended to detect the presence (or absence) of a target analyte in sample (matrix) without the need for specialized and costly equipment, though many lab based applications exist that are supported by reading equipment. Typically, these tests are used for medical diagnostics either for home testing, [[point of care]] testing, or laboratory use. A widely spread and well known application is the home pregnancy test.
Evaluation of semantic edit intention labels

![Bar chart showing precision and recall for different categories](image)
Evaluation of semantic edit intention labels

High precision
Evaluation of semantic edit intention labels

- High precision
- Low recall
Evaluation of semantic edit intention labels

Inter-annotator agreement: 0.62, 0.02, 0.17

High precision
Low recall
Considerable disagreements
Evaluation of sentence quality labels

- Low quality sentences from edits
- High quality sentences from Featured Articles
- Machine Learning model

Examples:

- Needs Citations: 1946: ford sues the allies for damages done to his factories in dresden during the infamous bombing and wins compensation
- Has point-of-view: sharky and george was a very popular humorous show about two fish private detective.
Implications of learning from Wikipedia editing behaviors
Implications of automated labeling of sentence quality

Automated data labeling:
- Randomly sampled non-reverted Wikipedia edits
- Regular expressions
- Semantic-intention labeled edits
- Sentences from labeled edits (positive examples)
- Sentences from non-labeled edits (negative examples)
- Training & Testing Data

Classifier Training & Testing:
- Training Data
- Testing Data
- Past research datasets
- Sequence-based RNN model
- Validation results

Classifier Deployment:
- Sentences from current Wikipedia articles
- Sequence-based RNN model
- Sentences that need improvement

Fast
Flexible
Language independent
Article improvement – add, assess, improve
Article improvement – add, assess, improve

… → Article → Add content → Review → Improve content → …

Editor 1

Editor 2
Assisting new editors on common practices

• Recommending best practice changes to new editors.

• Can prevent follow-up edits and save other editor’s valuable time!
Assisting new editors on common practices

Create a system to encode best practices into editing experiences

In the forest that is Phabricator, this ticket is very much a seedling on the forest floor.
Read: this task is a gathering place/work in progress.

This parent task is intended to help gather and advance the thinking around how the visual editor might be enhanced to help people learn the social conventions (read: policies and guidelines) and exercise the judgement necessary to become productive and constructive Wikipedia editors.

Background

Visual editor’s growing popularity among people new to editing Wikipedia [1] suggests it has been reasonably successful at helping people learn the technical skills [ii] necessary to edit Wikipedia.

Trouble is, the edits these new people make often break/defy Wikipedia policies and guidelines.

This task is about exploring how the visual editor could be augmented/enhanced to help people learn these policies and guidelines and exercise the judgement necessary to become productive and constructive Wikipedia editors.
Inherent Ambiguity

• Some semantic categories can have multiple viewpoints.

• E.g., Neutral point-of-view is a highly debatable policy

• Models need to incorporate such ambiguity.
How can AI and humans work together in a task? Whose voice matters?

• Models trained on implicit editing behaviors capture descriptive behaviors (how editors interpret and operationalize Wikipedia norms and practices).

• Thus, these models can be used to understand the application of a policy in context (e.g., looking at the prediction of new sentences to understand the policy better)
“He had a very jovial personality”

Prediction: “biased” (91%)
What do we need for better AI?

Better data

Better AI

More expressive UI

Neutral point-of-view?

Clarification?
Thank You!
Sumit Asthana
asumit@umich.edu
http://sumitasthana.xyz