Improving Wikipedia Citations

A semi-automated system for fixing, improving, and repairing citations across Wikipedia

Cat Ball — catball@uw.edu
2023-09-07

Quick Intro

Cat Ball — catball@uw.edu

- UW Computational Linguistics Master’s Student
- Day job as an SRE at a tech company
- May start applying to PhD programs next year

(Why? and How?)

Why?

how is this work useful?

- Add more citation information for readers & improve readability
    - If sources are easily accessible, more people may read them
- Metadata analysis
    - Help understand what sources influence which articles
    - Gather metadata about scholarly papers and resources
- Citation error corrections
- Improve machine readability
    - Including allowing other, existing citation analysis tools to run

→ An aside

“Why not fully automate it?”

- Failing to fix a citation?
  - Acceptable
  - Page is the same as before
- Accidentally making a citation inaccurate?
  - *Not acceptable*
  - Possible to generate harmful misinformation

How?

use the computer, mostly

- Parse all citations from all articles
- Parse sources from citations
- Validate citation data
- Generate a new citation with fixes
- Web UI for humans to review & submit suggested changes

- EZ PZ right...?

How?

use the computer, mostly

- Parse all citations from all articles
- Parse sources from citations
- Validate citation data
- GENERATE A NEW CITATION WITH FIXES
- Web UI for humans to review & submit suggested changes

- ~~EZ PZ right...?~~

Service Overview

- **Frontend**
    - Web UI displays current citation & suggested citation
    - Allows user to approve suggested change, make revisions, or mark the current citation as good
- **Backend**
    - From database, get a pair of (current citation, suggested citation) for user requests
    - Check that the current citation in the database is still what lives on the live page
    - Send revisions to Wikipedia based on user reviews
- **Periodic jobs**
    - Get new Wikipedia data dumps
    - Update database with changes
    - Generate suggested fixes for citations (i.e. the hard part)

Service Overview

- **Frontend**
  - Web UI displays current citation & suggested citation
  - Allows user to approve suggested change, make revisions, or mark the current citation as good
- **Backend**
  - From datab t citation, suggested citation) for user requests
  - Check that e database is still what lives on the live page
  - Send revisions to Wikipedia based on user reviews
- **Periodic jobs**
  - Get new Wikipedia data dumps
  - Update database with changes
  - Generate suggested fixes for citations (i.e. the hard part)

Let’s focus on this

Populate the database

at a high level

- Download most recent monthly Wikipedia data dump
    - Much faster to process a batch of data from our disk than to call the Wikipedia API over the network for every article
- Check which Wikipedia articles have changed since last database update
- Parse citations from new and modified articles
- Classify which citations can be improved
- Generate improved citation, put it in the database for human review

Citations with templates

<ref> {{cite journal | ... }} </ref>

- Template keys tell us what each bit of text describes
- e.g. title, publisher, date, DOI, etc.
- Validating data is easier when you know what text is describing!

```
It is the most abundant form of [[ordinary matter]] in the
[[universe]], being mostly associated with [[star]]s, including the
[[Sun]].
<ref name="Itptma2013a">
{{cite book
|last1=Chu
|first1=P.K.
|last2=Lu
|first2=XinPel
|date=2013
|title=Low Temperature Plasma Technology: Methods and Applications
|page=3
|publisher=CRC Press
|isbn=978-1-4665-0990-0}}
</ref>
```

Citations without templates

just some text with between <ref> tags

- <ref> tag may contain full citation or abbreviated citation
- Citation may be in any format
- Difficult for machines to determine which words describe what data
- Sometimes <ref> tags may not even contain a citation

This is how short citations look in the edit box:

```
The Sun is pretty big,<ref>Miller 2005, p. 23.</ref> but the Moon is not so big.<ref>Brown 2006, p. 46.</ref> The Sun is also quite hot.<ref>Miller 2005, p. 34.</ref>

== Notes ==
{{reflist}}

== References ==
* Brown, Rebecca (2006). "Size of the Moon", ''Scientific American'', 51 (78).
* Miller, Edward (2005). ''The Sun''. Academic Press.
```

This is how they look in the article:

The Sun is pretty big,[1] but the Moon is not so big.[2] The Sun is also quite hot.[3]

Notes

1. ^ Miller 2005, p. 23.
2. ^ Brown 2006, p. 46.
3. ^ Miller 2005, p. 34.

References

- Brown, Rebecca (2006). "Size of the Moon", *Scientific American*, 51 (78).
- Miller, Edward (2005). *The Sun*. Academic Press.

Parsing citations from articles

considerations

- Articles (*usually*) have (*mostly*) valid syntax
    - Of that syntax, we're only looking for <ref> tags and {{citation templates}}
- <ref> tags aren't nested
    - but {{templates}} *can* be nested
- There are several current citation styles
    - and some deprecated citation styles
- We don't care much about text outside the citations

Parsing citations from articles

options

- **Use the MediaWiki parser**
  - Huge, complex PHP codebase
  - Useful to reference, but not for just extracting citations
- **Use a 3rd-party citation parser**
  - Existing options limited in functionality, or no longer work
  - Possibly useful to reference, or as modules for sub-tasks
- **Roll my own parser**
  - I can only blame myself for my bugs 

Parsing Strategies

get all the structured citation data out

- Recursive descent parser
  - Relatively simple to implement, well-understood
  - Make a small grammar of the symbols used in templates and refs
  - Allows us to classify items inside citation templates while we parse
- Near-term prototype parser
  - Get contents of <ref> tags with Python's html.parser.HTMLParser module
  - Use a sad regex on each set of ref tag contents to identify any templates
  - Some string splits and hope to get keys and values from templates

Yet more parsing

citations that aren't templated

```
The Sun is pretty big,<ref>Miller 2005, p. 23.</ref> but the Moon is not so big.<ref>Brown 2006, p. 46.</ref> The Sun is also quite hot.<ref>Miller 2005, p. 34.</ref>

== Notes ==
{{reflist}}

== References ==
* Brown, Rebecca (2006). "Size of the Moon", ''Scientific American'', 51 (78).
* Miller, Edward (2005). ''The Sun''. Academic Press.
```

- Citation data that wasn't in a template is now just a blob of text
- The <ref> tag may or may not contain the whole citation
    - the <ref> tag may or may not contain *any* citation
- So what are our options?

Parsing untemplated citations

some options

- Cherry-pick easily-recognizable identifiers
    - ISBNs, DOIs, etc
    - maybe URLs (if they point to a known domain where we can easily fetch metadata)
    - Get metadata from authoritative sources; validate by checking if parts of authoritative metadata are present in the citation (e.g., author, title, etc).
- Match citations using known style formats
    - Crossref classifier

Note on Crossref style classifier

Existing work

- Crossref model for citation style classification
    - Training data is reference text from papers across a variety of style formats, represented as a CBOW
    - Author evaluates accuracy with naive Bayes, logistic regression, linear SVCs, and random forests
- Relation to this work
    - If citations are nicely formatted in a known style, we can extract identifiers based on this
    - However, more often the problem may be citations not using a particular style
- Could their published model be fine-tuned with more “sloppy” styles we sample?
    - Still limited to a fairly narrow set of existing citation styles
    - Extracting data could still be an issue if the citation deviates too far from the standard

Parsing untemplated citations

work in progress

- Current plan exploratory
- Evaluate several methods
    - What proportion of untemplated citations can be benefited by each method?
    - What is the accuracy of each method?
- Continue applying learnings from CLMS as I progress

Where were we?

generating citations

- Now that we’re confident what source this citation is referring to:
    - Validate consistency of citation data
    - Combine metadata from authoritative source with local data (e.g. page numbers, quote blocks, etc)
    - Generate templated citation from metadata
    - Compare generated and original citation

Validating identifiers

- Are all datums in a citation consistent with each other?
    - i.e. do different identifiers in the citation reference the same source?
    - May need fuzzy matching for similar-but-different
- Is there missing data that is relevant to add?
- Is the formatting readable? Does this citation use a citation template?
    - *Should* it use a citation template? (usually, yes)
    - Which citation template?

Things NOT being validated

things that need thorough human understanding

Is this source reliable?

- Lists of trusted / untrusted sources *could* be constructed, *but*...
- Some sources may be reliable for some topics, but not others
- Context is critical!

Does the citation support the claim?

- Can’t check paywalled sources
- Requires semantic understanding of both the claim and the source
- *Context is critical!!*

Authoritative citation metadata sources

existing datasets

- Notable resources:
    - WorldCat (thank you for the API access!)
    - Initiative for Open Citations (I4OC)
    - Open Citation Corpus
    - Refcat

Authoritative citation metadata sources

considerations

- Local corpora faster to use
- Unsure of completeness and quality between sources; needs evaluation
- Possible to coalesce data from multiple sources for a more complete record
    - Resource intensive; better suited for periodic batch jobs

From here

- Generate citation in a useful format
- Put it in the database
- Let users evaluate suggested revisions from the web UI

Experiment progression

iterative experiments across issue classes

- We’re correcting potentially many types of citation issues
- Initially generate fixes for a single subset of citation issues
- Analyze accept / reject / revise feedback from human evaluators
- Continue improving and iterating through more types citation fixes, repeating evaluation for each type of fix

Questions? Comments? Feedback?

ask me anything!