
Citation Detective:

A Public Dataset to Improve and Quantify Wikipedia Citation Quality at Scale

Ai-Jou Chou, National Chiao Tung University

Guilherme Gonçalves, Google

Sam Walton, Wikimedia Foundation

Miriam Redi, Wikimedia Foundation



<https://github.com/AikoChou/citationdetective>

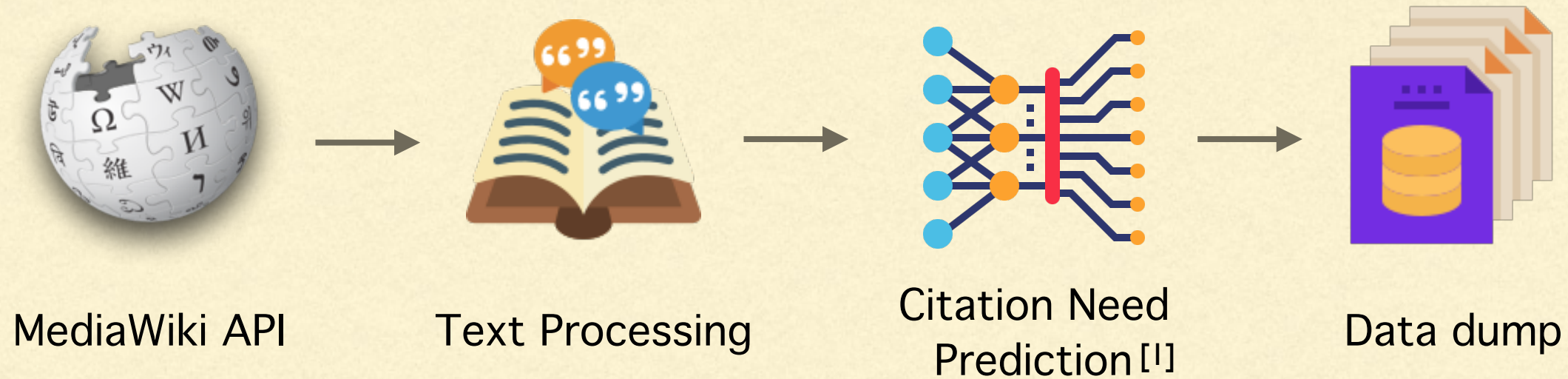


THE WEB
CONFERENCE

April 20-24
TAIPEI
2020

Citation Detective

- An end-to-end system periodically release public, usable dataset exposing sentences classified as missing citations.



Use Case: Integration into Citation Hunt

Citation Hunt

The Wikipedia snippet below is not backed by a [reliable source](#). Can you find one?

Click **I got this!** to go to Wikipedia and fix the snippet, or **Next!** to see another one. Good luck!

In page [Ekman spiral](#):

The classic Ekman spiral has been observed under sea ice,^[1] but observations remain rare in open-ocean conditions. This is due both to the fact that the turbulent mixing in the surface layer of the ocean has a strong diurnal cycle and to the fact that surface waves can destabilize the Ekman spiral. Ekman spirals are also found in the atmosphere. Surface winds in the Northern Hemisphere tend to blow to the left of winds aloft.



I got this!

Next!

[1] Miriam Redi, Besnik Fetahu, Jonathan Morgan, Dario Taraborelli. 2019. Citation Needed: A Taxonomy and Algorithmic Assessment of Wikipedia's Verifiability.

[2] Citation Hunt (https://meta.wikimedia.org/wiki/Citation_Hunt)

Quantify Citation Quality

- Citation Quality: the proportion of “well sourced” sentences in an article.

