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
Citation Detective

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

MediaWiki API  →  Text Processing  →  Citation Need Prediction[1]  →  Data dump

**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."

[2] Citation Hunt (https://meta.wikimedia.org/wiki/Citation_Hunt)
Citation Quality: the proportion of “well sourced” sentences in an article.

Quantify Citation Quality

Breakdown of Citation Quality by Topic

- Biography
- Sports
- Biology
- Medicine & Health