Possibly obsoleted by FAIR project

Anyone who has done desk research carefully knows that many citations don't support the claim they're cited for - usually in a subtle way, but sometimes a total nonsequitur. Here's a fun list of 13 features we need to protect ourselves.

This seems to be a side effect of academia scaling so much in recent decades - it's not that scientists are more dishonest than other groups, it's that they don't have time to carefully read everything in their sub-sub-field (... while maintaining their current arms-race publication tempo).

Take some claim *P* which is below the threshold of obviousness that warrants a citation.

It seems relatively easy, given current tech, to answer: (1) "Does the cited article say *P*?" This question is closely related to document summarisation - not a solved task, but the state of the art is workable.

It is *very* hard to answer (2) "Is the cited article strong evidence for *P*?", mostly because of the lack of a ground-truth dataset.

Motivation

- Individual academic papers are often quite honest and good
- But when papers are cited, nuance/honesty is lost
    - manipulative interpretation of studies
    - exaggeration of evidence

- research is often thin on key questions, one might write a book about 1-3 papers e.g. commodification, superforecasters
- popularisations creates myths

Research question

A paper P makes a claim C, citing source paper S in support of C. We want to ask a few questions about this situation. In increasing order of difficulty:

1. Does S make claim C?
2. Is S strong evidence for C?
3. Is S trustworthy?
4. Is C true?

Let’s start by nailing down (1).

Does the cited article say this?

This is closely related to document summarisation. It’s also easy to **perturb** the labels and create known-false examples; simply edit key numbers in S to be some random digit (“45%” instead of “15%”).

Train a grandiosity detector. Larger claims are harder to justify for any S. So train a system that can detect overheated bullshit C; this is a penalty term on our eventual credence in P and S. (Either P is misciting S or S is overheated bullshit too.)

Other less promising angles

What does the *literature* say about T?

    Consider a cited article T and citing articles c_i

Sentiment / evaluative.

    (Doesn’t Scite do this? Yeah, but crappily)

Semantic Scholar API already gives you the snippets for citations!
Where are the labels?

Classify: neutral “background” citation or positive “load-bearing” citation or negative “foil” cite?

Take into account small n = |c_i| when judging

Is S trustworthy? Some candidate features and training setups

- Sentence level features
    - NLP (easy): Does the cited article say this?
    - NLP (super hard): Is the cited article strong evidence?
- Article level features
    - Deterministic: Retraction lookup (not that helpful)
    - Deterministic: OpenReview scores (not that helpful)
    - NLP: OpenReview / PubPeer comments
- Journal level features
    - *Trendiness as negative predictor.* See also the notion of a “tabloid” journal.
    - CiteScore rather than Clarivate

How do GopherCite do it?

TA note

Very exciting if this could work even somewhat reliably!

I am ignorant of ML unlike you, but it feels very ambitious what you're trying to do. For all but the most trivial cases (e.g. reporting a numeric estimate that is already well described by a sentence in the cited paper). For example at GiveWell a flaw I would find very often is someone cites an estimate from a paper to give evidence for a claim about the average effect of X, but the cited paper actually examined X' which is importantly different (this comes in degrees of course).

Even if the ML part could be done, would you need dozens of highly rigorous researchers to slave away manually creating the training data?

Or what about mathematical mistakes, like (somewhat contrived example) someone citing an exponentiated logistic regression coefficient as if it was not exponentiated?

And can you explain more how you expect to adapt document summarisation models / approaches?