Determining the Credibility of Science Communication

Isabelle Augenstein
Dpt. of Computer Science
University of Copenhagen
augenstein@di.ku.dk

Abstract

Most work on scholarly document processing assumes that the information processed is trustworthy and factually correct. However, this is not always the case. There are two core challenges, which should be addressed: 1) ensuring that scientific publications are credible – e.g. that claims are not made without supporting evidence, and that all relevant supporting evidence is provided; and 2) that scientific findings are not misrepresented, distorted or outright misreported when communicated by journalists or the general public. I will present some first steps towards addressing these problems and outline remaining challenges.

1 The Life Cycle of Scientific Research

Scientific research is highly diverse not just when it comes to the topic of study, but also how studies are conducted, how the resulting research is described and when and where it is published. However, what different fields still have in common is a certain life cycle, starting with planning a study and ending with promoting the research post-publication, in the hopes of the article finding readership and having an impact.

Scholary document processing aims to support researchers throughout this life cycle of scientific research, by offering various tools to automate otherwise manual processes. Most research within scholarly document processing has focused on supporting information discovery for finding related work. Most prominently, research has focused on methods to condense scientific documents, using entity extraction and linking, keyphrase or relation extraction (Augenstein et al., 2017; Augenstein and Søgaard, 2017; Wright et al., 2019; Gábor et al., 2018; Ammar et al., 2018) or automatic summarisation (Collins et al., 2017; Yasunaga et al., 2019).

Once papers are written and submitted for peer review, it is pertinent to evaluate them fairly and objectively. This process is far from straightforward, as, among others, reviewers have certain biases, including against truly novel research (Rogers and Augenstein, 2020; Bhattacharya and Packalen, 2020). Research has thus focused on automatically generating peer reviews from paper content (Wang et al., 2020), as well as on studying how well review scores can be predicted from review texts (Kang et al., 2018; Plank and van Dalen, 2019).

Finally, post-publication, the impact of scientific work can be tracked, using citations and citation counts as a proxy for this. It is again worth noting that there are significant biases in this – e.g. author information is among the, if not the most salient feature for predicting citation counts (Yan et al., 2011; Holm et al., 2020). Looking further into what papers are cited and why, Mohammad (2020b,a) find that there are significant topical as well as gender biases when it comes to who is cited and by whom.

2 Credibility and Veracity of Science Communication

While all of the work referenced above is important in supporting researchers, it neglects one crucial aspect, namely that it assumes the resulting scientific documents and broader communication about them are credible and supported by the underlying evidence. Though it is the task of peer reviewers to spot issues regarding credibility, and the task of journalists to check their sources when they report on scientific studies, distortions, exaggerations and outright misrepresentations can still happen.

The ongoing COVID-19 pandemic has highlighted the disastrous and direct consequences misreporting of scientific findings can have on our everyday lives, yet, there is still relatively little work on detecting issues in the credibility of scientific writing. This especially holds for detecting smaller nuances of untrustworthy scientific writing, whereas there is comparatively more work on de-

Proceedings of the Second Workshop on Scholarly Document Processing, pages 1–6
June 10, 2021. ©2021 Association for Computational Linguistics

Biology
Wood Frogs (Rana sylvatica) are a charismatic species of frog common in much of North America. They breed in explosive choruses over a few nights in late winter to early spring. *The incidence in Wood Frogs was associated with a die-off of frogs during the breeding chorus in the Sylamore District of the Ozark National Forest in Arkansas ~~(Trauth et al., 2000)~~.*

Computer Science
Land use or cover change is a direct reflection of human activity, such as land use, urban expansion, and architectural planning, on the earth's surface caused by urbanization ~~[1]~~. Remote sensing images are important data sources that can efficiently detect land changes. *Meanwhile, remote sensing image-based change detection is the change identification of surficial objects or geographic phenomena through the remote observation of two or more different phases ~~[2]~~.*

Table 1: Excerpts from training samples in CITEWORTH (Wright and Augenstein, 2021) from the Biology and Computer Science fields. Green sentences are cite-worthy sentences, from which citation markers are removed during dataset construction.

tecting outright scientific misinformation (Vijjali et al., 2020; Lima et al., 2021).

Here, we highlight two important and so far understudied tasks to address issues with such smaller nuances of untrustworthy scientific writing, which can come into play at different stages of the life cycle of scientific research. The first one is *cite-worthiness detection*, which is about detecting whether or not a sentence ought to contain a citation to prior work. This task could help to ensure that claims are not made without supporting evidence, i.e. support researchers in writing more trustworthy scientific publications.

The second task is *exaggeration detection*, which is to determine whether a statement describing the findings of a scientific study exaggerates them, e.g. by claiming that two variables are strongly correlated when in reality they only co-occur. We argue that this task could be useful to verify if popular science reporting faithfully describes scientific research, or also to determine whether citation sentences (sentences which contain a citation; also called *citances*) faithfully describe the research documented in the cited papers.

2.1 Cite-Worthiness Detection

The CITEWORTH Dataset To study cite-worthiness detection, we first introduce a new rigorously curated dataset, CITEWORTH (Wright and Augenstein, 2021), for cite-worthiness detection from scientific articles. It is created from S2ORC, the Semantic Scholar Open Research Corpus (Lo et al., 2020). CITEWORTH consists of 1.2M sentences, balanced across 10 diverse scientific fields. While others have studied this task for few and/or narrow domains (Sugiyama et al., 2010; Färber et al., 2018), and have also studied very related tasks, such as claim check-worthiness detection (Wright and Augenstein, 2020a) or citation recommendation (Jürgens et al., 2018), this is the largest and most diverse dataset for this task to date.

An excerpt of our introduced dataset, CITEWORTH can be found in Table 1. The dataset curation process involves: 1) data filtering, to identify credible papers with relevant metadata such as venue information; 2) citation span identification and masking, of which we only keep papers with citation spans at the end of sentences to avoid rendering sentences ungrammatical; 3) discarding paragraphs without citations, or where not all sentences have citation spans in accordance with our heuristics; 4) evenly sampling paragraphs, such that the resulting dataset is equally balanced for the domains of Biology, Medicine, Engineering, Chemistry, Psychology, Computer Science, Materials Science, Economics, Mathematics, and Physics.

Given this dataset, we then study: how cite-worthy sentences can be detected automatically; to what degree there are domain shifts between how different fields use citations; and if cite-worthiness data can be used to perform transfer learning to downstream scientific text tasks.

Methods for Cite-Worthiness Detection We find that the best performance can be achieved by a Longformer-based model (Beltagy et al., 2020), which encodes entire paragraphs in papers and jointly predicts cite-worthiness labels for each of the sentences contained in the paragraph. Additional gains in recall can be achieved by using positive unlabelled learning, as documented in Wright and Augenstein (2020a) for the related task

Exaggerated Claims
Press Release: Players of the game rock paper scissors subconsciously copy each other's hand shapes, significantly increasing the chance of the game ending in a draw, according to new research.
Abstract: Specifically, the execution of either a rock or scissors gesture by the blind player was predictive of an imitative response by the sighted player.

Exaggerated Advice
Press Release: Parents should dilute fruit juice with water or opt for unsweetened juices, and only allow these drinks during meals.
Abstract: Manufacturers must stop adding unnecessary sugars and calories to their FJJDS.

Table 2: Examples of exaggerated claims and exaggerated advice given in press releases about scientific papers.

of claim check-worthiness detection. Our best-performing model outperforms baselines such as a carefully fine-tuned SciBERT (Beltagy et al., 2019) by over 5 points in F1.

Domain Differences To study domain effects, we perform a cross-evaluation, where we hold out one domain for testing and evaluate model performance on that, and compare this against an in-domain evaluation setting, where all domains observed at test time are also observed at training time. We find that there is a high variance in the maximum performance for each field (σ = 3.32), and between different fields on the same test data, despite large pretrained Transformer models being relatively invariant across domains (Wright and Augenstein, 2020b). This suggests stark differences in how different fields employ citations.

Downstream Applicability We evaluate our models on downstream scientific document processing tasks from Beltagy et al. (2019), which can be grouped into: named entity recognition tasks; relation extraction tasks; and text classification tasks. Specifically, we use our best-performing model, pre-trained for cite-worthiness detection and masked language modelling, and fine-tune them for 10 different downstream tasks. We find that improvements over the state of the art can be achieved for two citation intent classification tasks.

2.2 Exaggeration Detection

We frame exaggeration detection in the context of popular science communication. Specifically, we ask the question: how can one automatically detect if popular science articles overstate the claims made in scientific articles?

Prior work has shown that exaggeration of findings of scientific articles is highly prevalent (Sumner et al., 2014; Bratton et al., 2019; Woloshin et al., 2009; Woloshin and Schwartz, 2002). Exaggeration can mean a sensationalised take-away of the applicability of the work in terms, i.e. giving advice for which there is no scientific basis. Moreover, the strength of the main causal claims and conclusions of a paper can be exaggerated. Table 2 shows examples of those two types of claims from the datasets curated by Sumner et al. (2014) and Bratton et al. (2019), which we use in our work.

Prior work (Yu et al., 2019, 2020; Li et al., 2017) uses datasets based on PubMed abstracts and paired press releases from EurekAlert.[1] Their core limitations of is that they are limited to only observational studies from PubMed, which have structured abstracts, which strongly simplifies the task of identifying the main claims of a paper. This also holds for the test settings they consider, meaning that the proposed models have a limited applicability.

By contrast, we study how to best identify exaggerated claims in popular science communication in the wild, without highly curated data with annotations about core claims. This represents a more realistic experimental setup, which is more suited to supporting downstream use cases such as flagging exaggerated popular news articles as well as exaggerated summaries of scientific papers as referenced in other scientific papers.

Our method is a semi-supervised approach, which first identifies sentences containing claims in both scientific articles and popular science communication within the medical domain, then identifies the main conclusion of both articles, and lastly predicts to what degree popular science articles exaggerate those findings. We further analyse to what degree exaggeration of findings is correlated

[1] https://www.eurekalert.org/

with the perceived media bias of popular science communication outlets.

3 Conclusion

This paper discusses research avenues for automatically determining the credibility of science communication, both in terms of scientific papers and popular science communication. These avenues are put in the context of scholarly data processing more broadly, and how different tasks can be used to assist the life cycle of scientific research. While existing research has focused on developing models for assisting with information discovery, peer review and citation tracking, comparatively little work has been done on identifying non-credible claims and assisting authors in making sure their research is backed up by sufficient evidence where needed. The suggestion is therefore to focus on two tasks: cite-worthiness detection, to identify sentences requiring citations; and exaggeration detection, to identify cases in which scientific findings have been overstated. A core problem for both tasks is the lack of appropriate training data, which we address by introducing a new dataset, and a semi-supervised learning method, respectively. We hope our research will inspire future work on developing tools to assist authors and journalists in ensuring that research is described in a credible and evidence-based way.

Acknowledgements

The research documented in this paper has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 801199.

Thank you to Dustin Wright for the fruitful discussions and feedback on this extended abstract.

References

Waleed Ammar, Dirk Groeneveld, Chandra Bhagavatula, Iz Beltagy, Miles Crawford, Doug Downey, Jason Dunkelberger, Ahmed Elgohary, Sergey Feldman, Vu Ha, Rodney Kinney, Sebastian Kohlmeier, Kyle Lo, Tyler Murray, Hsu-Han Ooi, Matthew Peters, Joanna Power, Sam Skjonsberg, Lucy Wang, Chris Wilhelm, Zheng Yuan, Madeleine van Zuylen, and Oren Etzioni. 2018. Construction of the Literature Graph in Semantic Scholar. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, pages 84–91, New Orleans - Louisiana. Association for Computational Linguistics.

Isabelle Augenstein, Mrinal Das, Sebastian Riedel, Lakshmi Vikraman, and Andrew McCallum. 2017. SemEval 2017 task 10: ScienceIE - extracting keyphrases and relations from scientific publications. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 546–555, Vancouver, Canada. Association for Computational Linguistics.

Isabelle Augenstein and Anders Søgaard. 2017. Multi-task learning of keyphrase boundary classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 341–346, Vancouver, Canada. Association for Computational Linguistics.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3606–3611.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The Long-Document Transformer. *CoRR*, abs/2004.05150.

Jay Bhattacharya and Mikko Packalen. 2020. Stagnation and scientific incentives. Technical report, National Bureau of Economic Research.

Luke Bratton, Rachel C Adams, Aimée Challenger, Jacky Boivin, Lewis Bott, Christopher D Chambers, and Petroc Sumner. 2019. The Association Between Exaggeration in Health-Related Science News and Academic Press Releases: A Replication Study. *Welcome open research*, 4.

Ed Collins, Isabelle Augenstein, and Sebastian Riedel. 2017. A supervised approach to extractive summarisation of scientific papers. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 195–205, Vancouver, Canada. Association for Computational Linguistics.

Michael Färber, Alexander Thiemann, and Adam Jatowt. 2018. To Cite, or Not to Cite? Detecting Citation Contexts in Text. In *European Conference on Information Retrieval*, pages 598–603. Springer.

Kata Gábor, Davide Buscaldi, Anne-Kathrin Schumann, Behrang QasemiZadeh, Haifa Zargayouna, and Thierry Charnois. 2018. Semeval-2018 task 7: Semantic relation extraction and classification in scientific papers. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 679–688.

Andreas Nugaard Holm, Barbara Plank, Dustin Wright, and Isabelle Augenstein. 2020. Longitudinal citation prediction using temporal graph neural networks. *arXiv preprint arXiv:2012.05742*.

David Jürgens, Srijan Kumar, Raine Hoover, Dan McFarland, and Dan Jurafsky. 2018. Measuring the Evolution of a Scientific Field through Citation Frames. *Transactions of the Association for Computational Linguistics*, 6:391–406.

Dongyeop Kang, Waleed Ammar, Bhavana Dalvi, Madeleine van Zuylen, Sebastian Kohlmeier, Eduard Hovy, and Roy Schwartz. 2018. A dataset of peer reviews (PeerRead): Collection, insights and NLP applications. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1647–1661, New Orleans, Louisiana. Association for Computational Linguistics.

Yingya Li, Jieke Zhang, and Bei Yu. 2017. An nlp analysis of exaggerated claims in science news. In *Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism*, pages 106–111.

Lucas Chaves Lima, Dustin Brandon Wright, Isabelle Augenstein, and Maria Maistro. 2021. University of copenhagen participation in trec health misinformation track 2020. *arXiv preprint arXiv:2103.02462*.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel S Weld. 2020. S2ORC: The Semantic Scholar Open Research Corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983.

Saif M. Mohammad. 2020a. Examining citations of natural language processing literature. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5199–5209, Online. Association for Computational Linguistics.

Saif M. Mohammad. 2020b. Gender gap in natural language processing research: Disparities in authorship and citations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7860–7870, Online. Association for Computational Linguistics.

Barbara Plank and Reinard van Dalen. 2019. CiteTracked: A Longitudinal Dataset of Peer Reviews and Citations. In *BIRNDL@ SIGIR*, pages 116–122.

Anna Rogers and Isabelle Augenstein. 2020. What can we do to improve peer review in NLP? In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1256–1262, Online. Association for Computational Linguistics.

Kazunari Sugiyama, Tarun Kumar, Min-Yen Kan, and Ramesh C Tripathi. 2010. Identifying Citing Sentences in Research Papers Using Supervised Learning. In *2010 International Conference on Information Retrieval & Knowledge Management (CAMP)*, pages 67–72. IEEE.

Petroc Sumner, Solveiga Vivian-Griffiths, Jacky Boivin, Andy Williams, Christos A Venetis, Aimée Davies, Jack Ogden, Leanne Whelan, Bethan Hughes, Bethan Dalton, et al. 2014. The Association Between Exaggeration in Health Related Science News and Academic Press Releases: Retrospective Observational Study. *BMJ*, 349.

Rutvik Vijjali, Prathyush Potluri, Siddharth Kumar, and Sundeep Teki. 2020. Two stage transformer model for COVID-19 fake news detection and fact checking. In *Proceedings of the 3rd NLP4IF Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 1–10, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

Qingyun Wang, Qi Zeng, Lifu Huang, Kevin Knight, Heng Ji, and Nazneen Fatema Rajani. 2020. ReviewRobot: Explainable paper review generation based on knowledge synthesis. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 384–397, Dublin, Ireland. Association for Computational Linguistics.

Steven Woloshin and Lisa M Schwartz. 2002. Press Releases: Translating Research Into News. *Jama*, 287(21):2856–2858.

Steven Woloshin, Lisa M Schwartz, Samuel L Casella, Abigail T Kennedy, and Robin J Larson. 2009. Press Releases by Academic Medical Centers: Not So Academic? *Annals of Internal Medicine*, 150(9):613–618.

Dustin Wright and Isabelle Augenstein. 2020a. Claim Check-Worthiness Detection as Positive Unlabelled Learning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 476–488, Online. Association for Computational Linguistics.

Dustin Wright and Isabelle Augenstein. 2020b. Transformer Based Multi-Source Domain Adaptation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7963–7974, Online. Association for Computational Linguistics.

Dustin Wright and Isabelle Augenstein. 2021. CiteWorth: Cite-Worthiness Detection for Improved Scientific Document Understanding. In *Findings of the Association for Computational Linguistics: ACL 2021*, Online. Association for Computational Linguistics.

Dustin Wright, Yannis Katsis, Raghav Mehta, and Chun-Nan Hsu. 2019. Normco: Deep disease normalization for biomedical knowledge base construction. In *AKBC*.

Rui Yan, Jie Tang, Xiaobing Liu, Dongdong Shan, and Xiaoming Li. 2011. Citation count prediction: learning to estimate future citations for literature. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, pages 1247–1252.

Michihiro Yasunaga, Jungo Kasai, Rui Zhang, Alexander R Fabbri, Irene Li, Dan Friedman, and Dragomir R Radev. 2019. Scisummnet: A large annotated corpus and content-impact models for scientific paper summarization with citation networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7386–7393.

Bei Yu, Yingya Li, and Jun Wang. 2019. Detecting Causal Language Use in Science Findings. In *EMNLP*, pages 4656–4666.

Bei Yu, Jun Wang, Lu Guo, and Yingya Li. 2020. Measuring Correlation-to-Causation Exaggeration in Press Releases. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4860–4872.