HarriGT: Linking news articles to scientific literature

James Ravenscroft[1,3,4], Amanda Clare[2] and Maria Liakata[1,3]

[1]Centre for Scientific Computing, University of Warwick, CV4 7AL, United Kingdom
[2]Department of Computer Science, Aberystwyth University, SY23 3DB, United Kingdom
[3]Alan Turing Institute, 96 Euston Rd, London, NW1 2DBB, United Kingdom
[4]Filament AI, CargoWorks, 1-2 Hatfields, London, SE1 9PG, United Kingdom

Abstract

Being able to reliably link scientific works to the newspaper articles that discuss them could provide a breakthrough in the way we rationalise and measure the impact of science on our society. Linking these articles is challenging because the language used in the two domains is very different, and the gathering of online resources to align the two is a substantial information retrieval endeavour. We present HarriGT, a semi-automated tool for building corpora of news articles linked to the scientific papers that they discuss. Our aim is to facilitate future development of information-retrieval tools for newspaper/scientific work citation linking. HarriGT retrieves newspaper articles from an archive containing 17 years of UK web content. It also integrates with 3 large external citation networks, leveraging named entity extraction, and document classification to surface relevant examples of scientific literature to the user. We also provide a tuned candidate ranking algorithm to highlight potential links between scientific papers and newspaper articles to the user, in order of likelihood. HarriGT is provided as an open source tool (http://harrigt.xyz).

1 Introduction

For scientists, understanding the ways in which their work is being reported by journalists and the subsequent societal impact of these reports remains an overwhelming task. Research funding councils have also become increasingly interested in the impact that the research that they fund produces. These motivating factors, combined with suggestions that traditional citation-based metrics such as JIF (Garfield, 2006) and h-index (Hirsch, 2005) are not as transparent as once thought (Cronin, 1984; Bornmann and Daniel, 2008) have catalyzed the development of metrics to measure scientific impact in society, policy and the economy (recently termed "comprehensive scientific impact" (Ravenscroft et al., 2017)). Evaluation programmes such as the UK's Research Excellence Framework (REF) Impact Case Study (REF 2014, 2012) and the United States' STAR Metrics programme (Lane and Bertuzzi, 2010) set the current state of the art in comprehensive scientific impact metrics. However, both processes involve significant manual labour and introduce human subjectivity into their evaluation processes. Ravenscroft et al. (2017) recently showed that there is negligible correlation between citation-based metrics and REF scores and called for the development of an objective, automated metric for measuring comprehensive impact. As part of the US-funded FUSE project, McKeown et al. (2016) developed a method for measuring the use of technical terms over time in scientific works as a proxy for scientific impact. McKeown's work, whilst primarily focusing on scientific literature, represents a significant step towards deeper understanding of scientific impact beyond citations.

Our assumption is that the perception of researchers' work as reflected in the mainstream media is an important means of measuring comprehensive impact, useful both to researchers themselves as well as funding bodies. However, one of the main barriers to building an automated solution to assessing such comprehensive impact is a lack of training data. In this paper, we present and discuss our tool, HarriGT, which facilitates ground truth collection for a corpus of news articles linked to the scientific works that they discuss. In this way we aim to lay the groundwork for future stud-

Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics-System Demonstrations, pages 19–24
Melbourne, Australia, July 15 - 20, 2018.

ies that help scientists understand societal perception and impact of their work through the media.

2 Background

Citation extraction from news articles reporting on scientific topics remains a challenging and relatively unexplored task. There are no conventions, formal or informal, for citing a scientific work in a news article. Scientific journalists often omit key information about who funded or even carried out a given study from their reports making identification of the work very difficult (Bubela et al., 2009). Journalists also frequently quote academics who were not directly involved in a scientific work in their stories, further confusing attempts to automate citation extraction (Conrad, 1999). Louis and Nenkova (2013) found that the quality of scientific reporting varies greatly even between journalists within the same publishing venue.

On the other hand, parsing and understanding citations between scientific works is a domain that has seen a lot of attention from academia in recent years. Citations in scientific papers are relatively well structured and formulaic. As such, pattern-based extraction mechanisms have been found to yield good citation extraction results (Councill et al., 2008). Disambiguation of the scientific work and authors to which a citation refers can be a much more challenging task. This especially applies in cases where authors have ambiguous names (e.g. J. Smith). One approach is to assign scientific works and authors unique identifiers such that there is no ambiguity in cited works (DOI and ORCID respectively) (Paskin, 2015; Butler, 2012). A more pragmatic approach is needed to disambiguate publications and authors for which no DOI or ORCID ID have been assigned. Huang and Ertekin (2006) present a method for disambiguation of authors using a learned distance metric that takes into account author's known names, affiliations and venues that they typically publish at. Similar approaches have led to the creation of citation networks that store relationships between huge volumes of scientific works. Networks such as CiteSeerX (Wu et al., 2014), Microsoft Academic Knowledge Graph and Scopus provide external access via APIs for research and application development purposes.

Beyond academia, references to scientific work are common across a number of domains. The popular encyclopedia website, Wikipedia, relies upon outbound citation to establish its veracity concerning matters of science (Nielsen, 2007). Whilst DOI links to articles are often used, in many cases, only the title, publication name and author names are provided leading to a structured extraction and disambiguation problem similar to that outlined above. (Nielsen, 2007; Kousha and Thelwall, 2017; Nielsen et al., 2017).

Since academia as a whole has begun to adapt to online publishing, academics have become individually accustomed to sharing work through digital channels and social media. This has led to the development of systems such as Altmetric.com (Adie and Roe, 2013), that monitor social media posts as well as some mainstream media outlets for mentions of scientific works via DOI links. By their own admission, altmetric toolmakers still struggle to identify all mentions of scientific works, focusing only on articles with a DOI or some other unique identifier (Liu and Adie, 2013).

Extraction and disambiguation of references to scientific works in news articles is the task that has motivated the development of HarriGT. We seek to facilitate construction of a human-curated corpus of newspaper articles that have been explicitly linked to scientific works. Such corpora could be used to build machine learning models that are able to connect news articles to scientific works automatically. Using HarriGT we have already started the creation of such a corpus. At time of writing the corpus consists of 304 newspaper articles linked to one or more scientific paper. The corpus is growing incrementally and can be downloaded via the tool.

3 System Overview

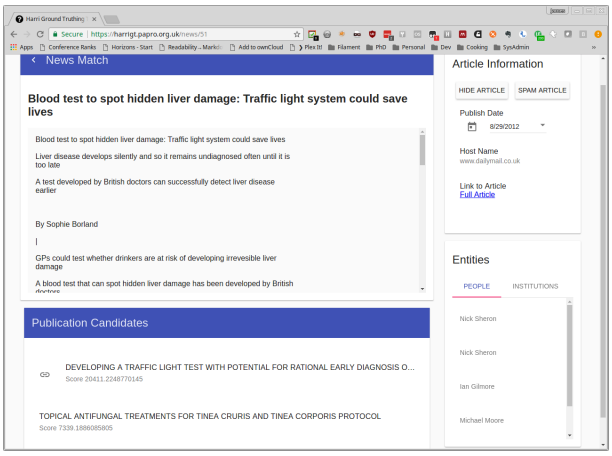

Figure 1: HarriGT Web UI shows a news article and related candidate scientific papers

HarriGT provides a system that brings together historical news articles from web archives stored in the widely used open source WARC format. The system automatically ingests and parses newspaper articles and searches citation graphs for relevant candidate papers that the user is able to **link** or **hide** or mark as **spam**. A diagram explaining this process is available on the HarriGT website. In this paper, we worked with a UK national web archive (JISC and the Internet Archive, 2013) and candidate scientific papers connected to citation graphs from Microsoft, Scopus and Springer. The news article and candidate scientific papers are presented in a web interface, enabling a user to quickly decide whether each candidate is linked to the news article or not. This section discusses the components involved in this process in detail and outlines some of the challenges we faced in the system creation.

3.1 News Corpus Retrieval

In order to build a comprehensive corpus of news articles, we worked with the JISC Web Archive, a comprehensive scrape of the *.uk* top-level domain between 1996 and 2013. Content is stored in Web Archive (WARC) compressed format and an index file containing metadata about every URL that was scraped and a pointer to the related content within the WARC structure was made available. The JISC Web Archive is approximately 62 Terabytes in size, so identifying and filtering relevant content became a primary concern[1].

We initially decided to restrict our investigation to news articles between 2011 and late 2013 which coincided with REF 2014. We compiled a list of web addresses for local and national UK news outlets via a Wikipedia article[2] in order to reduce the number of hostnames that our tool should inspect down to 205. The archive index files also provided metadata about the type of each WARC entry and whether the original scrape was successful or not (e.g. whether the URL was invalid). This brought down the total number of WARC entries to be examined to approximately 11.5 million. Requests to the BLOB store hosting the web archive were optimised through a script that identified batches of URLS archived in the same BLOB.

[1]The JISC Web Archive is accessible for research purposes at data.webarchive.org.uk/opendata/ukwa.ds.2/

[2]https://en.wikipedia.org/wiki/List_of_newspapers_in_the_United_Kingdom

3.2 News Article Pre-Processing & Filtering

The contents of the archives were typically HTML and thus we needed to extract the title and body of each news story. HTML layouts can vary significantly between sites but news articles follow a typical layout and thus extraction of content fields can be carried out using rules and patterns rather than a machine learning approach. For our purposes we found that the open source library newspaper[3] was highly effective and gave us access to an article's title, authors, publication date and other metadata.

During the process we realised that some news articles had been duplicated in the archive. This can occur when a web crawler retrieves a URL that has been generated erroneously by the scraper script or the website being scraped. This can lead to multiple links to the same content. Examples include incorrectly appending search keywords, pagination information and other parameters into URLs that do not require these parameters.

To get around this problem, we introduced a hashing system, taking the SHA256 hash of the title body text from each article and only accepting new content if its hash is not already known.

We found that using the science section of the newspapers to filter suitable articles led to exclusion of relevant material. A second approach was to only accept articles that pass two high-level keyword filters. The first, simpler check is to see whether or not an article contains one or more keywords: *science, scientist, professor, doctor, academic, journal, research, publish, report*. We deliberately chose these keywords as a simplistic filter to reduce the amount of current affairs/celebrity gossip news that was initially accepted into our system. For the second of our filters, we ran a Named Entity Recognition (NER) algorithm[4] that provided multi-word expression identification and classification for names, locations and geo-political entities. From the results of the NER execution, we only accepted articles with at least one organisation containing *University, College* or *Institute*.

The final step in the pre-processing pipeline is identification of each article's publication date. Publication date is one of the most salient features in our paper candidate scoring algorithm discussed below. Recent digital news articles give their date

[3]http://newspaper.readthedocs.io/en/latest/

[4]SpaCy 2.0 https://spacy.io/

Model Type	Accuracy	F1-Score
SVM	0.94	0.94
Naive Bayes	0.82	0.86

Table 1: Micro-averaged Results from Spam Models. Spam Articles: 2085, Ham Articles: 840

of publication in their HTML metadata. However, for many of the old articles in the web archive, this information was not present. For articles with no known publication date, we first attempted to retrieve the same URL from the live internet where much of the original content is still available but with updated layouts and metadata. If the content can't be found, we used a set of regular expressions (found within the newspaper library mentioned above) to try and find the date in the article HTML. Failing all else, we simply asked the user to try and identify the publication date manually within the user interface.

The retrieval and pre-processing steps are rather time consuming, taking a modern workstation (Intel i7 Quad Core @ 3.5Ghz, 16GB RAM) approximately 24 hours to process 20k news articles. We therefore ingest content into HarriGT in batches using a small Apache Hadoop cluster.

3.3 'Spam' Filtering

Our keyword filter during pre-processing removes a large number of general interest articles that do not discuss scientific work. There are still a number of articles that pass this initial screening that are off topic. We address this issue by including a machine learned "spam" model into HarriGT. Within the user interface, news articles can be marked as **spam** if they contain little relevant scientific content. The model is re-trained using new examples from the **spam** and **link** categories as the user continues to tag articles.

We trained two machine learning models to address the problem, a Naive Bayes classifier and a Support Vector Machine. We used Grid Search to identify the best training hyper-parameters for feature extraction and the models. The optimal feature hyper-parameters were found to be unigram and bigram bag-of-words features with TF-IDF weighting, maximum document frequency of 75% and a maximum vocabulary size of 10,000. We found that an SVM with a linear kernel and $C = 1$ produced the best results and used this model in the live system. Table 3.3 shows our model results after 4 iterations of training and use.

Given the size of the corpus, the hardware environment that the model was required to support and the positive results from the SVM mode, we decided not to explore deep learning approaches to spam filtering.

3.4 Citation Graph Integration

In order to provide candidate scientific works for each newspaper article, we required integration with rich sources of metadata for as many scientific disciplines as possible. We decided to integrate HarriGT with the Microsoft Academic Knowledge[5], Scopus[6] and Springer[7] APIs. These APIs all provide broad, up to date coverage of known academic works. Each API had a different search endpoint with differing query languages and syntax that had to be catered for.

Each of the APIs returns metadata such as title, names and affiliations of authors, name of publishing venue and date of publication. In most cases each API returned a DOI so that each work could be uniquely identified and hyperlinked via the HarriGT interface. This allowed us to deduplicate items returned by more than one API.

Articles typically talk about the institution that a scientific work was carried out at and independently the name of the author e.g. "Cambridge Researchers have found that... Dr Smith who led the study said..." making automatic extraction of reference information very difficult. Therefore, we use NER to identify all names and institutions in the article and run citation graph queries for each permutation. For example: "A study run by Oxford and Cambridge universities found that... Dr Jones who led the study said..." would yield two queries: (Jones, Oxford), (Jones, Cambridge). Searches are bounded by the article's publication date plus-or-minus 90 days.

3.5 Candidate Scoring Implementation

The retrieval mechanism described above tends to overgenerate links between news articles and scientific publications, resulting in 0.19 precision. Therefore it is important to have a mechanism for ranking these further, to avoid spurious links and only show the user the most prominent ones for further verification. To address this we propose a simple but effective mechanism based on the Levenshtein Ratio. Each news article is associated

[5] https://labs.cognitive.microsoft.com/en-us/project-academic-knowledge

[6] https://dev.elsevier.com/index.html

[7] https://dev.springer.com/

with a set of C candidate scientific works c_i where $i \in [0, C]$ are found using the retrieval method discussed above. News articles contain two sets of entity mentions of interest: A set of N peoples' names n_j and a set of O organization names o_j. We also record the number of times each entity is mentioned M_j. For each candidate scientific work c_i, we identify a set of A_i authors' names a_i^k and their respective academic affiliations u_i^k. We also note the publication date of each news article D and the publication date of each candidate scientific work P_i.

For a given news article, we score each candidate scientific work c_i by summing over the square of Levenshtein Ratio ($L_r(x, y)$) of each pair of mentions of names and authors:

$$S_i^{per} = \sum_{j=0}^{N} M_j \sum_{k=0}^{A_i} L_r(n_j, a_k^i)^2$$

A similar calculation is carried out for organisation mentions and affiliations.

$$S_i^{org} = \sum_{j=0}^{O} M_j \sum_{k=0}^{A_i} L_r(o_j, u_k^i)^2$$

The Levenshtein Ratio is a simple, effective measure that has been used for assessing NE similarity (Moreau et al., 2008). We also calculate Δ_D, the number of days between the publication date of the news article, D and the scientific work P_i. In cases where the candidate article has multiple publication dates (for example, online publication versus print publication), Δ_D is calculated for all publication dates and the smallest value is retained.

$$\Delta_D = \min_n(\sqrt{(D - P_i^n)^2})$$

Finally, we calculate an overall score S_i for each article by normalizing S_i^{per} and S_i^{org} by their respective numbers of distinct entity mentions and then dividing by Δ_D like so:

$$S_i = (\frac{S_i^{per}}{N} + \frac{S_i^{org}}{O}) \times \frac{1}{\Delta_D}$$

Candidates are ranked according to their S_i score in descending order so that the highest scoring candidates are presented to the user first.

3.6 Candidate Scoring Evaluation

To evaluate our candidate scoring technique, we use it to retrieve the N-best candidates for news articles with known links to one or more scientific papers. For each of the news articles in our ground truth collection, we retrieved all candidate scientific works from the citation graphs as described in section 3.4 above. We then use the scoring algorithm from section 3.5 above to rank the candidates then check to see whether actual linked papers appear in the top 1,3 and 5 results (Top-K accuracy).

	Top-1	**Top-3**	**Top-5**
Accuracy	0.59	0.83	0.90

Table 2: Top-K Accuracy for Scoring Algorithm

We identified a small number of reasons for suboptimal ranking. Newspaper articles occasionally focus around candidate works published months earlier. In some cases, incorrect publication dates are being reported by the scientific paper APIs. In both cases, our system strongly penalizes candidates in terms of Δ_D. HarriGT's ranking algorithm also weakly penalizes candidates that have multiple authors in cases where only one author (often the lead) is mentioned in the newspaper text. This effect is amplified when work by the same lead author with fewer or no co-authors is also found since these candidates are preferred and filtered to the top of the list.

HarriGT's recall is not bounded by the candidate ranking algorithm but by the queries and results from our integration with Scopus, Microsoft and Springer APIs. HarriGT allows the user to **hide** news articles that are scientific but for which no relevant candidates are recommended. This action is distinct from marking an item as **spam**, which indicates that it has no scientific value and should be excluded from the corpus.

We evaluate the recall of our tool by considering items marked as **link** to be retrieved and deemed relevant and items marked as **hide** to be retrieved but for which no relevant items could be found. Thus defining recall as:

$$recall = \frac{|\{linked\}|}{|\{linked\} \cup \{hidden\}|}$$

At the time of writing, the recall of the system is 0.57. This figure may be lower than the actual figure, since papers are occasionally classified as 'hidden' by annotators if several strong candidates are presented and they are unsure which paper to link to. We expect that this figure will get stronger with more use.

4 Conclusion & Future Work

We have presented HarriGT, the first tool for rapidly establishing links between scientific works and the newspaper articles that discuss them. We have shown that using a combination of NLP techniques and proposing a simple but effective candidate ranking algorithm, it is possible to construct a linked corpus of scientific articles and news articles for future analysis of the impact of scientific articles in news media. The tool could also have other uses such as the discovery of primary sources for scientific news. Future work will explore the role of time and other content in this task. Our open source tool has been constructed with use of the JISC corpus in mind, but could be used with other sources of news also. HarriGT produces useful ranking and good recall and is ready for use with a large corpus. HarriGT is available to try out at `http://www.harrigt.xyz` and we welcome feedback from volunteer users.

Acknowledgments

We thank the EPSRC (grant EP/L016400/1) for funding us through the University of Warwick's CDT in Urban Science, the Alan Turing Institute and British Library for providing resources.

References

E Adie and W Roe. 2013. Altmetric: Enriching scholarly content with article-level discussion and metrics. *Learned Publishing* 26(1):11–17.

L Bornmann and H Daniel. 2008. What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation* 64(1):45–80.

T Bubela et al. 2009. Science communication reconsidered 27(6):514–518.

D Butler. 2012. Scientists: your number is up. *Nature* 485(7400):564–564.

P Conrad. 1999. Uses of expertise: sources, quotes, and voice in the reporting of genetics in the news. *Public Understanding of Science* 8:285–302.

I G Councill et al. 2008. ParsCit: An open-source CRF Reference String Parsing Package. *LREC '08: Proceedings of the 6th International Conference on Language Resources and Evaluation* 2008(3):661–667.

B Cronin. 1984. *The citation process: The role and significance of citations in scientific communication*. T. Graham London.

E Garfield. 2006. History and Meaning of the Journal Impact Factor. *JAMA* 295(1):90–93.

J E Hirsch. 2005. An index to quantify an individual's scientific research output. *Proc Natl Acad Sci USA* 102(46):16569–16572.

J Huang and S Ertekin. 2006. Fast Author Name Disambiguation in CiteSeer .

JISC and the Internet Archive. 2013. JISC UK Web Domain Dataset (1996-2013).

K Kousha and M Thelwall. 2017. Are wikipedia citations important evidence of the impact of scholarly articles and books? *Journal of the Association for Information Science and Technology* 68(3):762–779.

J Lane and S Bertuzzi. 2010. The STAR METRICS project: current and future uses for S&E workforce data. In *Science of Science Measurement Workshop, held Washington DC*.

J Liu and E Adie. 2013. Five challenges in altmetrics: A toolmaker's perspective. *Bulletin of the American Society for Information Science and Technology* 39:31–34.

A Louis and A Nenkova. 2013. A corpus of science journalism for analyzing writing quality. *Dialogue and Discourse* 4(2):87–117.

K McKeown et al. 2016. Predicting the Impact of Scientific Concepts Using Full-Text Features. *Journal of the Association of for Information Science and Technology* 67(11):2684–2696.

E Moreau et al. 2008. Robust Similarity Measures for Named Entities Matching pages 593–600.

F Nielsen et al. 2017. Scholia and scientometrics with Wikidata. In *Joint Proceedings of the 1st International Workshop on Scientometrics and 1st International Workshop on Enabling Decentralised Scholarly Communication*.

F Årup Nielsen. 2007. Scientific citations in Wikipedia. *First Monday* 12(8).

N Paskin. 2015. The digital object identifier: From ad hoc to national to international. In *The Critical Component: Standards in the Information Exchange Environment*, ALCTS (American Library Association Publishing).

J Ravenscroft et al. 2017. Measuring scientific impact beyond academia: An assessment of existing impact metrics and proposed improvements. *PLOS ONE* 12(3):e0173152.

REF 2014. 2012. Assessment framework and guidance on submissions. http://www.ref.ac.uk/2014/pubs/2011-02/.

J Wu et al. 2014. CiteSeerX : AI in a Digital Library Search Engine. *Proceedings of the Twenty-Sixth Annual Conference on Innovative Applications of Artificial Intelligence* 36(3):2930–2937.