News2PubMed: A Browser Extension for Linking Health News to Medical Literature

Jun Wang
Independent Researcher
Syracuse, New York, USA
junwang4@gmail.com

Bei Yu
Syracuse University
Syracuse, New York, USA
byu@syr.edu

ABSTRACT

This demo system presents a browser extension that allows the reader of a health news article to quickly retrieve related medical/health research papers. This system can help news editors and readers fact-check health news for incorrect or exaggerated claims, such as making causal claims from correlational findings or inference of animal studies to humans. Linking health news to the original research papers is not a trivial task, as links are largely missing in science news reports. To link health news to medical literature, our system includes a new named-entity recognition function to extract journal names, and a new Elasticsearch-based search engine to incorporate rich metadata into the search strategy. This paper also introduces a new dataset for evaluating the performance of the proposed search system.

CCS CONCEPTS

• **Information systems** → **Web applications**; **Specialized information retrieval**; *Digital libraries and archives*; *Information extraction*; • **Computing methodologies** → **Information extraction**; *Natural language processing*; • **Applied computing** → *Health informatics*.

KEYWORDS

health news; literature linking; elastic search engine; named-entity recognition; sentence classification; browser extension

ACM Reference Format:
Jun Wang and Bei Yu. 2021. News2PubMed: A Browser Extension for Linking Health News to Medical Literature. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21), July 11–15, 2021, Virtual Event, Canada.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3404835.3462788

1 INTRODUCTION

Health news is an important science communication channel to introduce the latest health research findings to the general public. However, errors and exaggerated claims, such as causal claims from correlational findings or inference of animal studies to humans, have been found quite often in health news, as scientists and media watchdogs like HealthNewsReview have warned [3, 10–12]. This misinformation in health news may cause unintended consequences like creating false hopes for patients, and undermine public trust in science as a result.

A natural solution to this problem is fact-checking health news against the original research papers, or at least presenting the original paper together with the news to readers for convenient comparison. The first step of this solution is to link health news to the original papers. However, this is not a trivial task since currently, such links are largely missing in science news reports, as including links is not a standard practice for traditional news outlets [5]. Moreover, existing links are often outdated or point to incorrect sources. For example, based on our calculation, more than 25% of the links included in Reuters' health news articles are outdated, with many of them pointing to the scientific journals' homepage, which often shows the latest publications instead of the specific research paper cited in the news article.

The problem of missing links or incorrect links calls for an automated solution to matching news articles with their corresponding research papers [5, 8]. However, linking news to literature is still a relatively unexplored research area. To our best knowledge, [8] is the only published study that has tackled this exact problem. They developed a system named *HarriGT* to link news articles to research papers. In addition, *altmetric.com*, arguably the most widely used altmetrics tool, briefly described their proprietary techniques for picking up mentions of research works in news [5]. Both studies modeled the linking problem as an information retrieval task with a two-stage approach: first, use named-entity recognition (NER) or text mining techniques to extract metadata such as author and affiliation names from news articles; second, use the metadata to query research literature databases to find corresponding research papers. The performance of *altmetric.com* is not disclosed. HarriGT team reported a top-1 accuracy of 0.59, indicating a need for improvement toward a deployable tool [8].

To date, the existing approaches have utilized a limited number of extracted metadata and off-the-shelf search engines. The advancement of natural language processing and information retrieval offers new ideas to solve the linking problem. In this study, we developed a demo system that can link health news articles to the original research papers with high accuracy. Compared to previous systems, this demo system includes a new NER function to extract metadata such as journal names, and a new elastic search-based engine to incorporate rich metadata and content information into the search strategy.

This demo system has been implemented as a browser extension: when a user visits a health news article online, she can click an injected search button to retrieve corresponding research papers.

SIGIR '21, July 11–15, 2021, Virtual Event, Canada.

ACM ISBN 978-1-4503-8037-9/21/07…$15.00
https://doi.org/10.1145/3404835.3462788

This system can be used by news editors, writers, or the general public to fact check health news for incorrect or exaggerated claims.

2 RELATED WORKS

HarriGT is the only published system on linking news to science literature [8]. The authors developed a corpus of about 300 news articles and applied general-purpose NER techniques to recognizing personal names and institutions in the news articles. Then named entities-based queries were sent to literature search engines (i.e. Microsoft Academic, Scopus, and Springer) to find research papers published within ±90 days. The retrieved papers were further re-ranked to ensure the most relevant papers were at the top of the search result. The system achieved a top-1 accuracy of 0.59 on a manually-curated data set.

A company named `altmetric.com`, arguably the most widely used site that monitors over 2000 mainstream news outlets as well as other types of media, developed a text mining system to extract journal names and author names from news articles and then used the extracted information to query `CrossRef` with a time window of ±45 days from the date of news release [1, 6]. (CrossRef is a major DOI registration agency that has registered more than 124 million works in May 2021, a nearly 10-percent increase from 2020.) Neither algorithm details nor performance have been disclosed to the public, although their brief report admitted room for improvement [5].

The above approaches have a number of limitations. First, although general-purpose NER techniques can recognize some metadata items such as authors and affiliations as named entities (*persons* and *organizations*), specialized NER models are needed to identify more metadata items that are not targets of common NER tools, such as journal names.

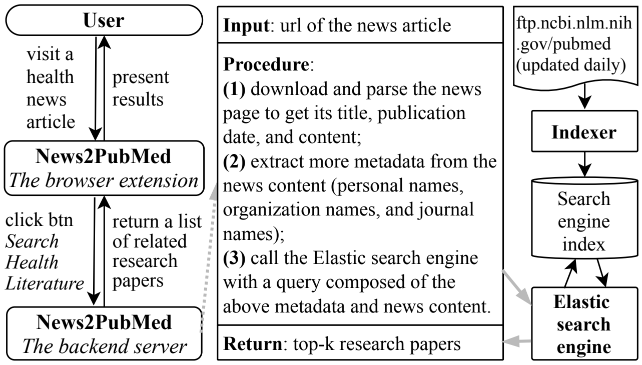
Second, the current search strategy uses a limited number of metadata items, such as author and affiliation in [8], while the usefulness of rich metadata and content information has not been investigated. To some extent, locating the research paper cited by a news article can be considered a navigational search task [4]. Prior user studies on navigational search have shown that compared to exploratory search, navigational search might benefit from longer queries [2]. Hence, it is worth exploring whether expanding queries with more metadata and content data would help link news articles to research papers.

Third, current literature search engines such as CrossRef lack flexibility in matching news articles to research papers. For example, given a news article, one would expect that the publication date of its corresponding paper should be as close as possible to the the news release date. However, current search engines only allow one to specify a date range (e.g. *news_release_date±90* days) which treats papers within the range equally and makes papers outside impossible to retrieve. Another example is that none of current engines supports full-text content search.

These limitations call for more research on developing better search strategies for linking news articles to research papers.

3 SYSTEM OVERVIEW

Fig. 1 illustrates the overview of the system. When a user reads a news article online, she can find a button *Search Health Literature*, which is injected by our browser extension, underneath the headline of the news article. If the user clicks the button, the url of the news article will be sent to our backend server, which will then download and parse the article to get its headline, publication date, and content. The server will also extract other metadata from the news content such as personal names, organizations, and journal names, using NER techniques. These metadata will be assembled to construct a query to a homemade elastic search engine. Then, the top-k research papers from the search result will be returned and presented to the user.


User
visit a health news article
present results
News2PubMed
The browser extension
click btn Search Health Literature
return a list of related research papers
News2PubMed
The backend server
ftp.ncbi.nlm.nih.gov/pubmed (updated daily)
Indexer
Search engine index
Elastic search engine

Input: url of the news article

Procedure:
(1) download and parse the news page to get its title, publication date, and content;
(2) extract more metadata from the news content (personal names, organization names, and journal names);
(3) call the Elastic search engine with a query composed of the above metadata and news content.

Return: top-k research papers

Figure 1: Overview of the system.

4 SEARCH ENGINE DEVELOPMENT AND EVALUATION

Existing literature search engines such as CrossRef are designed for processing queries that consist of research paper metadata, such as authors, titles, and journal names. However, news articles are a different genre which may not have these metadata readily available to formulate a query. For example, news articles rarely mention the title of a research paper, probably to avoid jargons or technical terms. The author and journal name of a research paper may be mentioned, but they need to be first extracted from the news content to be used in a query.

In this section, we describe our approach that can better accommodate the news genre: (1) developing a new Elasticsearch-based search engine that allows to construct powerful queries from enriched metadata and news content; (2) extracting metadata from news content, with a focus on creating a Reuters news-derived dataset to train a specialized journal name extraction model; and (3) evaluating the search engine's effectiveness.

4.1 Developing a new literature search engine

To accommodate the news genre, we built a new literature search engine based on Elasticsearch (https://www.elastic.io/), the most popular enterprise search engine that is based on the open-source Lucene library. Elasticsearch provides two advantages for our task. First, it supports a type of decay function that can decay relevance score depending on how far a value is from a given origin. This is a very useful feature for finding research papers whose publication dates should be as close as possible to the news release date. None of existing literature search engines has such function: they only support the common date range query like *from_date* and *until_date*. Second, to the best of our knowledge, none of existing engines supports the Elasticsearch-like relevance scoring function that can not only sum up the individual scores calculated from each query component, but also weight them to reflect the importance of certain metadata (e.g., a personal name may get more weight when it occurs multiple times in a news article).

To build our homemade literature search engine, we also need to develop a literature corpus to populate the search engine index (see Fig. 1). The metadata in our corpus include journal title, author names, affiliations, paper title, paper abstract, and publication date (which includes `year`, `month`, and `day`). When there are multiple publication dates (e.g. print date and online publication date), the earliest date is used because a news article is usually released after the research paper is accepted or put online. In addition, since many journals have alternative names, for each journal we augmented its name with alternative names, through the use of NLM medical journal name database. [1]

We considered four large corpora, including CrossRef (120m items), SemanticScholar (190m), Microsoft Academic (250m), and PubMed (30m). All of them provide snapshots for their latest collections, and two of them (SemanticScholar and PubMed) are free to download. After examining the four corpora, we decided to use PubMed since it contains the most complete metadata items. In comparison, only 10% of CrossRef's collection contain the affiliation information. SemanticScholar provides neither affiliation nor detailed publication date except year. Microsoft Academic does not have detailed publication date information. Note that although PubMed is already the largest biomedical literature database, it still does not cover all health/medical literature. For example, in a study of health news exaggeration [12], a small number of research papers reported in UK news agencies are not included in PubMed. In the future, we would like to extend the domain from health/medical to other science fields, by combining data from CrossRef and Microsoft Academic or SemanticScholar to create a more comprehensive literature search engine.

4.2 Named-entity metadata extraction

As seen in Fig. 1, the query to the backend literature search engine is constructed with two types of metadata: the original metadata (news title, publication date, and news content) and the extracted named entities (author names, affiliations, and journal names). Here we describe how we extract these named-entity metadata.

For author name extraction, we adopt a simple approach by extracting personal names through the use of Stanford's recent neural network-based NLP package, called *Stanza* [7]. Though sometimes the personal names mentioned in a news story refer to other people than the authors of the reported study, we found that the high recall ensured the Elastic search engine to be robust to queries with mixed true and false author names. So we decided to take this simple approach.

For author affiliation extraction, we used Stanza-extracted organizations to represent author affiliations. Similar to the author name identification task, although some extracted organizations are not research affiliations, the Elastic search engine is robust to queries with mixed true and false affiliation names.

Journal name identification. Since journal name identification is not included in standard named-entity recognition models, we developed a new, two-step method to extract journal names from news articles: first, develop a *journal sentence filter* model to remove sentences that did not mention any journal; second, develop a specialized *journal name NER* model to extract journal names from the sentences that mention journals.

To facilitate the development of the filter and NER models, we created a dataset using *Reuters* health news articles. We chose Reuters as the news source for developing our dataset because it has a large accessible archive of health news, including 8,600 news articles which contained one or more bitly urls linking to related research publications. Using the bitly urls, we can identify the doi of the journal papers through fetching and parsing the linked web pages. During the procedure of doi identification, however, we found that not all of those bitly urls are reliable: more than a quarter of the links were outdated or redirected to the entrance page of a journal website, which often features the latest journal issue instead of the specific research paper cited by the news article. These invalid links were excluded from the data set. In addition, for simplicity, in this demo paper, we focused on news articles that discuss only one study, namely, those with only one bitly url. In the end, we obtained 5,533 valid pairs of news articles and research papers.

To develop the *journal sentence filter*, we first built a training data set that includes a set of sentences that mention journal names (which will be referred to as *journal sentences*) and another set of sentences that do not (*non-journal sentences*). For each pair of linked news article and research paper in the Reuters corpus, since we know the name of the journal that publishes the research paper, we can locate the specific journal sentence that mentions the journal name. From the 5,533 pairs, we were able to locate 4,876 journal sentences. For the remaining 657 (12%) news articles, the original journal names were not found since the sentences used an alternative name of a journal, such as "PNAS" for "Proceedings of the National Academy of Sciences of the United States of America". Regardless of mentioning original journal names or alternative names, since journal sentences share common linguistic patterns when referring to a journal (e.g. many journal sentences contain such words as *journal*, *published*, *report*, or *write*), our journal sentence prediction model can generalize well to both cases.

We then randomly sampled from all news articles another set of 4,876 sentences as non-journal sentences. This balanced dataset was used to train a journal sentence classification model. We compared the performance of two text classification methods: Linear SVM (count vectorizer with 1,2,3-grams, which is better than tf-idf vectorizer) and BERT (case-based). To have a fair evaluation, we adopted a group-based training-test set split method: the journal sentences with the same journal name should appear either in the training or in the test set. For LinearSVM, the classification accuracy is 98.6%, and for BERT, it is 99.7%. Hence we adopted the BERT method in our system.

We also used the BERT-based pretrained model to fine-tune a journal NER model. On the 4,876 journal sentences, via 5-fold cross-validation, our fine-tuned journal NER model can obtain an accuracy of 99.9% and F1 of 0.989 (both precision and recall are 0.989) at the level of full named entity [9]. In contrast, if running the *Stanza* NER on those journal sentences and using the recognized

[1] https://www.nlm.nih.gov/databases/download/catalog.html

ORG entities as the journal names, we can only get an accuracy of 94.2% and F1 of 0.378 [9]. This demonstrates that it is necessary to develop a specialized journal NER module.

4.3 Search engine evaluation

Our dataset for evaluating the new elastic search engine is the above described 5,533 valid pairs of news articles and research papers.

Definition of top-k accuracy. For a news article, if its associated paper is in the top-k list of the search result, we count it as a *top-k hit*. Suppose a corpus has n news articles, if there are m top-k hits, we say the top-k accuracy for the corpus is $\frac{m}{n}$.

We chose to use top-k accuracy instead of other measurements such as nDCG (normalized discounted cumulative gain) because in our evaluation setting, as mentioned above, our dataset only includes those news articles that discuss one research work.

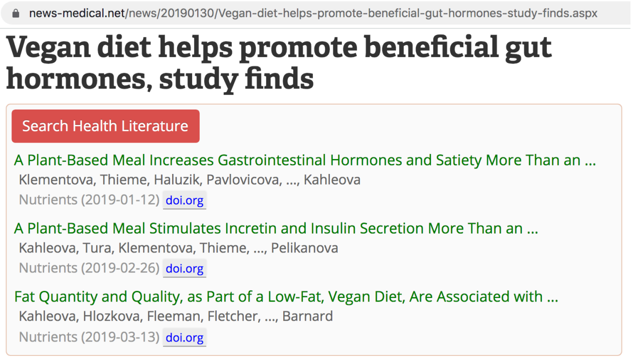
Table 1: Performance of running search experiments with various metadata settings.

	Experiment	Top-1	Top-3	Top-5	Top-10	Time*
1	TiDr	0.075	0.104	0.132	0.162	.025s
2	TiDd	0.166	0.262	0.301	0.358	.074s
3	TiDdCo	0.881	0.923	0.942	0.955	.64s
4	TiDdCoAu	0.908	0.953	0.957	0.976	.67s
5	TiDdCoAuAf	0.925	0.955	0.966	0.977	.82s
6	TiDdCoAuAfJo	0.957	0.976	0.985	0.989	.86s

Ti: Title / Dr: Date range (±45 days) / Dd: Date decay
Co: Content / Au: Author / Af: Affiliation / Jo: Journal

Table 1 shows how the use of enriched metadata query and the date decay function affects the retrieval performance. Row 1 (TiDr) shows that when using news title with a publication date range of 45 days in queries, the search engine can only achieve a top-1 accuracy of 0.075. Row 2 (TiDd) shows that replacing the above date range with a date decay function[2] in a query can double the performance. Row 3 (TiDdCo) shows that adding the first 300 tokens of news content into a query can make a huge boost to the retrieval performance. This boost is made possible by adding to the search engine database index a text field that consists of all relevant information, including paper title, author names and affiliations, journal name, and paper abstract. Rows 4-6 (TiDdCoAu, TiDdCoAuAf, and TiDdCoAuAfJo) show that adding author names, affiliations, and journal names to a query can further improve the top-k accuracy. In summary, the new search engine is able to harness enriched metadata to achieve more accurate linking between news articles and research papers.

Table 1 also displays the time needed to run each experiment in terms of a single search task. The experiment of using news content as part of a query (Row 3) takes 9 times as much time as the one without the use of news content (Row 2), suggesting there is a tradeoff between linking accuracy and response time.



news-medical.net/news/20190130/Vegan-diet-helps-promote-beneficial-gut-hormones-study-finds.aspx

Vegan diet helps promote beneficial gut hormones, study finds

Search Health Literature

A Plant-Based Meal Increases Gastrointestinal Hormones and Satiety More Than an ...
Klementova, Thieme, Haluzik, Pavlovicova, ..., Kahleova
Nutrients (2019-01-12) doi.org

A Plant-Based Meal Stimulates Incretin and Insulin Secretion More Than an ...
Kahleova, Tura, Klementova, Thieme, ..., Pelikanova
Nutrients (2019-02-26) doi.org

Fat Quantity and Quality, as Part of a Low-Fat, Vegan Diet, Are Associated with ...
Kahleova, Hlozkova, Fleeman, Fletcher, ..., Barnard
Nutrients (2019-03-13) doi.org

Figure 2: Screenshot of a user clicking on an injected button *Search Health Literature* on a health news page.

5 BROWSER EXTENSION DEMO

When a user visits an article on a news website that is supported by our browser extension, she will see an injected button *Search Health Literature*, underneath the headline of the news article (see Fig. 2). If the user clicks on the search button, the top 3 journal papers in the search result will be presented inside a box, in which the user can click the title of the paper to view its abstract or click "`doi.org`" to access the paper on its official website. If none of the match scores of the top 3 papers passes some threshold value, the user will be informed "sorry, the system could not find any related health literature". Multiple reasons could contribute to a failed search: (1) a recent news article may report a study that has not been published or registered in the PubMed metadata collection; (2) the news article reports something other than research, e.g. government policy; (3) the extracted metadata do not match with any records in the search engine index due to mistakes that occurred during the procedure of metadata extraction.

To facilitate users to check out our demo system, we provide a page at https://junwang4.github.io/demo-news2pubmed. There a user can find information about how to install the browser extension. In addition, the user can play with a list of news items that are randomly sampled from three news sources: `reuters.com`, `news-medical.net`, and `eurekalert.org`. For readers who are not familiar with the latter two, `news-medical` is a leading open-access medical and life science hub with about 300,000 health news items, and `eurekalert` is a major platform for universities, journal publishers, medical centers, and other organizations to post research news releases. Note that for `news-medical` and `eurekalert`, many of their news articles, especially the ones published earlier, do not include any journal reference; and for `reuters`, all news articles shown here are from our above evaluation dataset, and thus they have a reference link at the end, though more than a quarter of the links are outdated. In the above web page, alongside the headline of a news article, the publication date of the article is also displayed so that a user can compare news articles published during different time periods and thus appreciate how the browser extension could be used to help one locate related literature.

[2] $Score(paper) = e^{\lambda \cdot max(0, |paper\_publication\_date - news\_release\_date| - 7)}$, where $\lambda = \log(0.5)/180$. Interpretation: 7 indicates that if the date difference is within 7 days, the match score will be 1 (perfect match); and 180 indicates if the date difference is 180 days, the score will be 0.5.

ACKNOWLEDGMENTS

This research is supported by the US National Science Foundation under grant 1952353 and the Microsoft Investigator Fellowship program. We thank the reviewers for helpful comments, and Albert Wang for help with proofreading.

REFERENCES

[1] Altmetric. 2017. Mainstream Media Outlets. https://web.archive.org/web/20170711041232/https://www.altmetric.com/about-our-data/our-sources/news/ [accessed 04-20-2021; web archived 07-11-2017].

[2] Kumaripaba Athukorala, Dorota Glowacka, Giulio Jacucci, Antti Oulasvirta, and Jilles Vreeken. 2016. Is exploratory search different? A comparison of information search behavior for exploratory and lookup tasks. *Journal of the Association for Information Science and Technology* 67, 11 (2016), 2635–2651. https://doi.org/10.1002/asi.23617

[3] Alan Cassels and Joel Lexchin. 2008. How well do Canadian media outlets convey medical treatment information?: Initial findings from a year and a half of media monitoring by Media Doctor Canada. *Open Medicine* 2, 2 (2008), e45–e48. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3090174/

[4] Bernard J Jansen, Danielle L Booth, and Amanda Spink. 2008. Determining the informational, navigational, and transactional intent of Web queries. *Information Processing & Management* 44, 3 (2008), 1251–1266. https://doi.org/10.1016/j.ipm.2007.07.015

[5] Jean Liu and Euan Adie. 2013. Five challenges in altmetrics: A toolmaker's perspective. *Bulletin of the American Society for Information Science and Technology* 39, 4 (2013), 31–34. https://doi.org/10.1002/bult.2013.1720390410

[6] Ansel MacLaughlin, John Wihbey, and David A Smith. 2018. Predicting news coverage of scientific articles. In *Twelfth International AAAI Conference on Web and Social Media*. https://ojs.aaai.org/index.php/ICWSM/article/view/14999

[7] Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Association for Computational Linguistics (ACL) System Demonstrations*. https://www.aclweb.org/anthology/2020.acl-demos.14/

[8] James Ravenscroft, Amanda Clare, and Maria Liakata. 2018. HarriGT: Linking news articles to scientific literature. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics-System Demonstrations* (2018), 19–24. https://www.aclweb.org/anthology/P18-4004/

[9] Erik Tjong Kim Sang. 2018. CoNLL-2000 Shared Task Evaluation. https://www.clips.uantwerpen.be/conll2000/chunking/output.html.

[10] Gary Schwitzer. 2008. How Do US Journalists Cover Treatments, Tests, Products, and Procedures? An Evaluation of 500 Stories. *PLOS Medicine* 5, 5 (2008), e95. https://doi.org/10.1371/journal.pmed.0050095

[11] David E Smith, Amanda J Wilson, and David A Henry. 2005. Monitoring the quality of medical news reporting: Early experience with media doctor. *Medical Journal of Australia* 183, 4 (2005), 190–193. https://doi.org/10.5694/j.1326-5377.2005.tb06992.x

[12] Petroc Sumner, Solveiga Vivian-Griffiths, Jacky Boivin, Andy Williams, Christos A Venetis, Aimee Davies, Jack Ogden, Leanne Whelan, Bethan Hughes, Bethan Dalton, Fred Boy Boy, and Christopher D Chambers. 2014. The association between exaggeration in health related science news and academic press releases: retrospective observational study. *BMJ* 349 (2014), g7015. https://doi.org/10.1136/bmj.g7015