SciEv: Finding Scientific Evidence Papers for Scientific News

Dr. Jian Wu

Assistant Professor

Computer Science Department

Old Dominion University, Norfolk, Virginia

OLD DOMINION
UNIVERSITY

Motivation

- Scientific news is an important medium to disseminate scientific knowledge to the general public.
- However, scientific news, such as posts on social media, does not consistently cite or faithfully present facts in the source, which is usually scientific papers.
- **Research Question**: how to automatically find evidence from scientific papers given scientific news?
- **Research Impacts**: verifying scientific facts and democratizing scientific knowledge to any news readers from **scientists** to **government officers** to the **general public citizens**.

Examples

true news

People With Schizophrenia Have a Strange, Telling Response to The Epstein Barr Virus

HEALTH 11 January 2019 By CARLY CASSELLA

(KatarzynaBialasiewicz/iStock)

Schizophrenia is a severe mental disorder that has been baffling scientists for decades. It's one of the most enigmatic afflictions out there, and no matter how much research is carried out, the mysteries continue to pile on.

fake news

Don't look now, but Arctic sea ice mass has grown almost 40% since 2012

Sunday, October 01, 2017 by: Tracey Watson
Tags: Arctic, Arctic sea ice extent, climate change, global warming, ICE, ice flows, real science, sea ice

This article may contain statements that reflect the opinion of the author

127K VIEWS

supporting evidence paper

JOURNAL ARTICLE

Schizophrenia is Associated With an Aberrant Immune Response to Epstein–Barr Virus FREE

Faith Dickerson, Lorraine Jones-Brando, Glen Ford, Giulio Genovese, Cassie Stallings, Andrea Origoni, Colm O'Dushlaine, Emily Katsafanas, Kevin Sweeney, Sunil Khushalani ...
Show more
Author Notes

Schizophrenia Bulletin, Volume 45, Issue 5, September 2019, Pages 1112–1119,
https://doi.org/10.1093/schbul/sby164
Published: 20 November 2018

refuting evidence paper

Earth's Future

Research Article | Open Access

Global warming releases microplastic legacy frozen in Arctic Sea ice

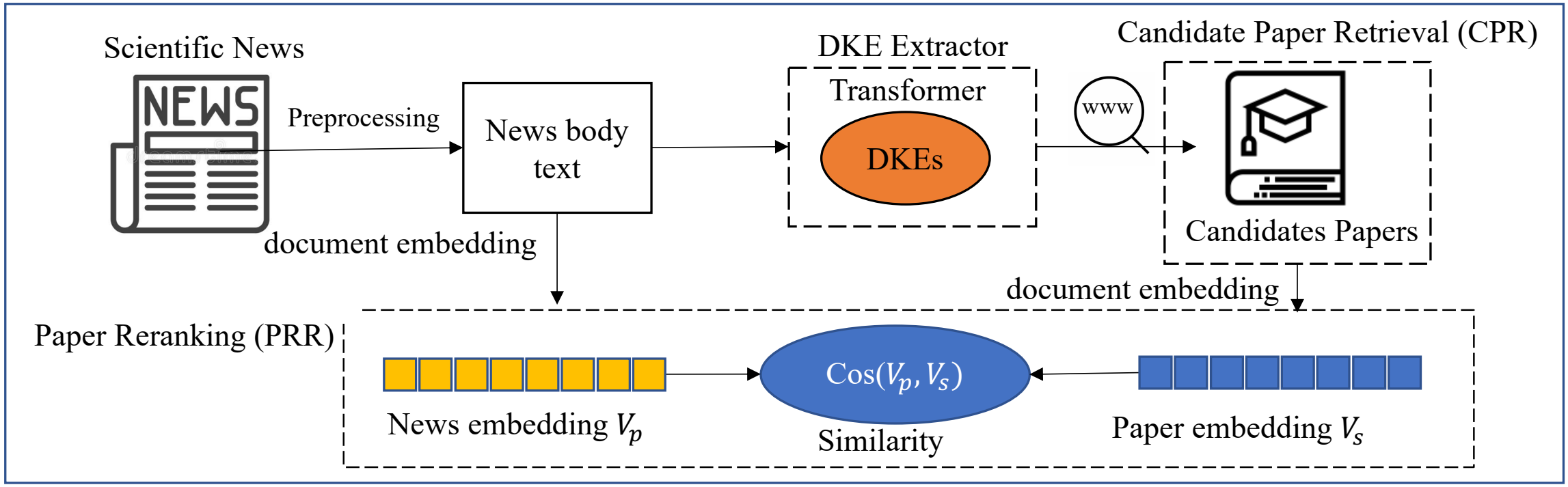
Rachel W. Obbard, Saeed Sadri, Ying Qi Wong, Alexandra A. Khitun, Ian Baker, Richard C. Thompson

First published: 20 May 2014 | https://doi.org/10.1002/2014EF000240 | Citations: 569

Research Challenges

1. Different from news articles, research papers are written for domain scientists and are written in a different style, what do we use to link these two types of media?
   - **Domain knowledge entity (DKE) (**Wu et al. 2020 JCDL**)**
2. There are hundreds of millions of research papers published, how do we narrow down the search space to find the closely relevant papers given a news article?
   - **A two-step retrieval system**
   - Step 1: shallow features with high recall; Step 2: deep features with high precision
3. What is the best way to represent the text when matching the content of a news article against research papers?
   - **TF-IDF vs. Language Models**

SciEv System Architecture



Hoque et al. (2019 K-CAP)
Hoque et al. (2022 arXiv:2205.00126)

Datasets

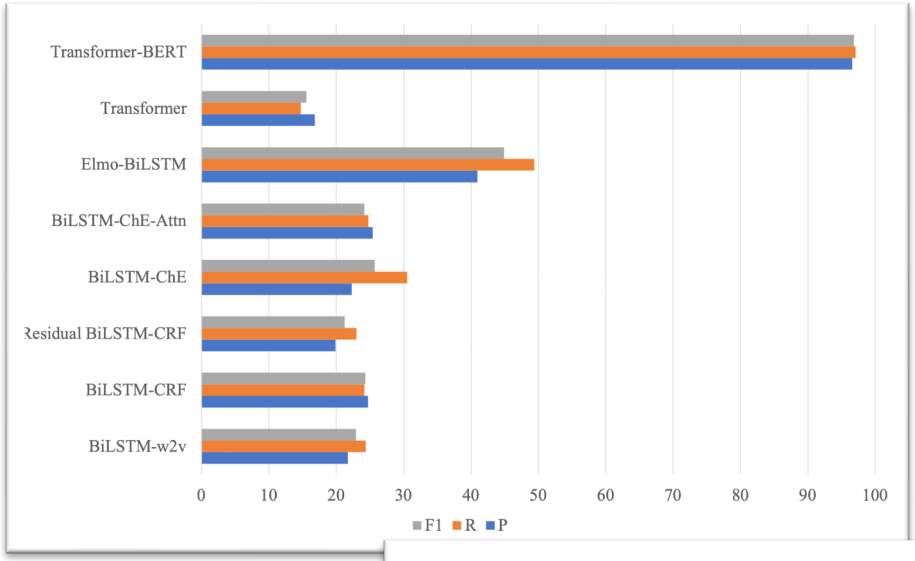
- Domain knowledge entity (DKE) extraction:
  - SemEval 2017 Task 10 dataset (Augenstein et al. 2017)
  - OA-STM dataset (Brack et al. 2020)
- 2-stage retrieval model (Candidate paper retrieval + reranking):
  - An in-house dataset consisting of 100 manually curated (news,paper) pairs from ScienceAlert

Example data:

News (full text)	Paper (metadata + abstract)
Lyme bacteria survive 28-day course of antibiotics months after infection (2017)	Embers et al. Cell (2017): Variable manifestations, diverse seroreactivity and post-treatment persistence in non-human primates exposed to *Borrelia burgdorferi* by tick feeding
There's Another Link Between Our Gut And The Brain That Could Help Avoid Dementia (2018)	Faraco et al. Nature (2018): Dietary salt promotes neurovascular and cognitive dysfunction through a gut-initiated TH17 response
Mind-Melting Study Says Our Universe Is an Expanding Bubble in Another Dimension (2019)	Souvik Banerjee et al. Physical Review Letters (2019): Emergent de Sitter Cosmology from Decaying Anti–de Sitter Space

Hoque et al. (2022 arXiv:2205.00126)

Evaluation



The transformer model was almost perfect for domain knowledge entity extraction.

- DKEs are more effective than keyphrases and named entities to link scientific news to scientific papers
- TFIDF is more effective than language models to represent text when matching scientific news against scientific papers
- The 2-stage system identifies relevant scientific papers within a reasonable time.

Beat the next best baseline by 11-26% for P@K.

System setting name	Query type	DKE model	Text representation	P@K K=1/5/10/20/50					MRR	Average NDCG	T_{PRR} (sec)	T_{all} (sec)
Baseline1[1]	–	–	–	14%	–	–	–	38%	–	–	–	–
Baseline2[2]	KP[2]	TextRank	TFIDF	18%	23%	23%	28%	29%	0.19	.22	**0.8**	21.18
Baseline3[3]	NE[3]	CoreNLP	TFIDF	38%	44%	45%	46%	47%	0.40	0.44	**0.8**	**13.90**
Baseline4	DKE[4]	HESDK	TFIDF	39%	45%	49%	55%	60%	0.43	0.48	**0.8**	116.33
BERT-TFIDF	DKE	BERT	TFIDF	**50%**	**71%**	**74%**	80%	86%	**0.59**	**0.70**	**0.8**	66.70
BERT-d2vec	DKE	BERT	d2vec	20%	37%	41%	55%	69%	0.28	0.40	4.04	142.17
BERT-Doc2Vec	DKE	BERT	Doc2Vec	36%	51%	54%	64%	71%	0.43	0.52	365.59	459.71
BERT-WDoc2Vec	DKE	BERT	Weighted Doc2Vec	35%	55%	60%	68%	84%	0.43	0.55	350.90	473.14
BERT-SciBERT	DKE	BERT	SciBERT	19%	30%	36%	43%	67%	0.25	0.36	732.74	875.88
BERT-SBERT	DKE	BERT	SBERT	47%	69%	74%	82%	90%	0.57	0.69	10.12	119.43
BERT-SPECTER	DKE	BERT	SPECTER	47%	69%	73%	**84%**	**91%**	0.57	0.69	4.22	110.31

[1] Quoted from Harrison et al. [2019]. The specific corpus used is not available, making it impossible to make a fair comparison.
[2] Keyphrases extracted using TextRank Mihalcea and Tarau [2004].
[3] Named entities extracted using Stanford CoreNLP Manning et al. [2014].
[4] HESDK Wu et al. [2017].

Hoque et al. (2022, arXiv: 2205.00126)

Future Research

- Identify stances (including mixed stances) of evidence papers given scientific news
- Identify fine-granular evidence statements from evidence papers
- Automatically generate natural language justifications that summarize multiple evidence papers
- Investigate how to leverage humans with general knowledge instead of domain experts to identify trustfulness of and debunk fake scientific news

What Can We Contribute to REASON

- We can contribute to TA1 on **Identify Additional Evidence** from scientific papers.
- The LAMP-SYS Lab at ODU is specialized at natural language processing, natural language understanding, and information retrieval on **Scholarly Big Data**.
- Dr. Wu co-directs the **CiteSeerX** project, and he is familiar with how to build accessible, usable, scalable, and sustainable systems (Wu et al. 2021 BigData).
- We showcased a system that achieved promising results on identifying evidence papers given scientific news articles.

References

- Hoque et al. (2022 arXiv:2205.00126): Md Reshad Ul Hoque, Jiang Li, Jian Wu. **SciEv: Finding Scientific Evidence Papers for Scientific News**. In: arXiv:2205.00126, 2022.
- Wu et al. (2020 JCDL): Jian Wu, Md Reshad Ul Hoque, Gunnar W. Reiske, Michele C. Weigle, Brenda T. Bradshaw, Holly D. Gaff, Jiang Li, and Chiman Kwan. **A Comparative Study of Sequential Tagging Methods for Domain Knowledge Entity Recognition in Biomedical Papers**. In: Proceedings of the 2020 Joint Conferences on Digital Libraries (JCDL 2020), August 1--5, 2020, Virtual Event, China
- Brack et al. (2020): Brack, A., D'Souza, J., Hoppe, A., Auer, S., Ewerth, R. (2020). **Domain-Independent Extraction of Scientific Concepts from Research Articles**. In: , et al. Advances in Information Retrieval. ECIR 2020. Lecture Notes in Computer Science, vol 12035. Springer, Cham.
- Hoque et al. (2019 K-CAP): Md Reshad Ul Hoque, Jian Wu, Jiang Li, Chiman Kwan, Agnese Chiatti, and Dash Bradley. **Searching for Evidence of Scientific News in Scholarly Big Data**. In:Proceedings of the 10th International Conference on Knowledge Capture (K-CAP 2019), November 19-21, 2019, Marina del Rey, CA, USA.
- Augenstein et al. (2017): 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.