GRAND CHALLENGES | OCTOBER 1–4, 2017 | WASHINGTON, DC, USA

Mitigating Information Overload with Influence Search

Accelerating literature-based discovery across domains using a conceptual influence graph

GUS HAHN-POWELL, MARCO A. VALENZUELA-ESCÁRCEGA, ZECHY WONG, MIHAI SURDEANU

Computational Language Understanding Lab at the University of Arizona, Tucson, AZ, USA

Motivation: Undiscovered Public Knowledge

“Knowledge can be public, yet undiscovered, if independently created fragments are logically related but never retrieved, brought together, and interpreted.” —SWANSON,1986

GROWTH OF LITERATURE

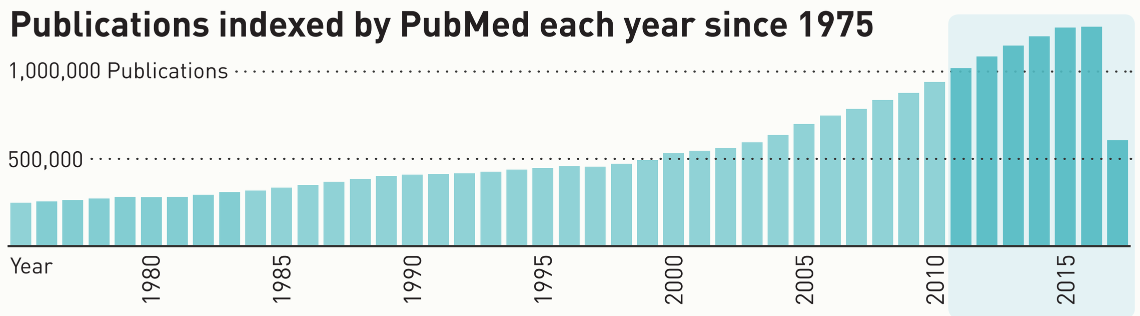

Figure 1. Annual number of publications indexed by PubMed.[6] Since 2011, there have been over 1 million new publications each year.

TASK

Definition:

Semantic search of literature along influence relations:

X causes Y causes Z

Beyond keyword search:

— What are causes of Z?

— How are X and Z causally connected?

Data:

— All PubMed abstracts (>26M)

— +100K full-text Open Access publications relevant to children's health

— *Soon:* Paywall publications, including Cochrane's systematic reviews

APPROACH

Step 1. Machine reading

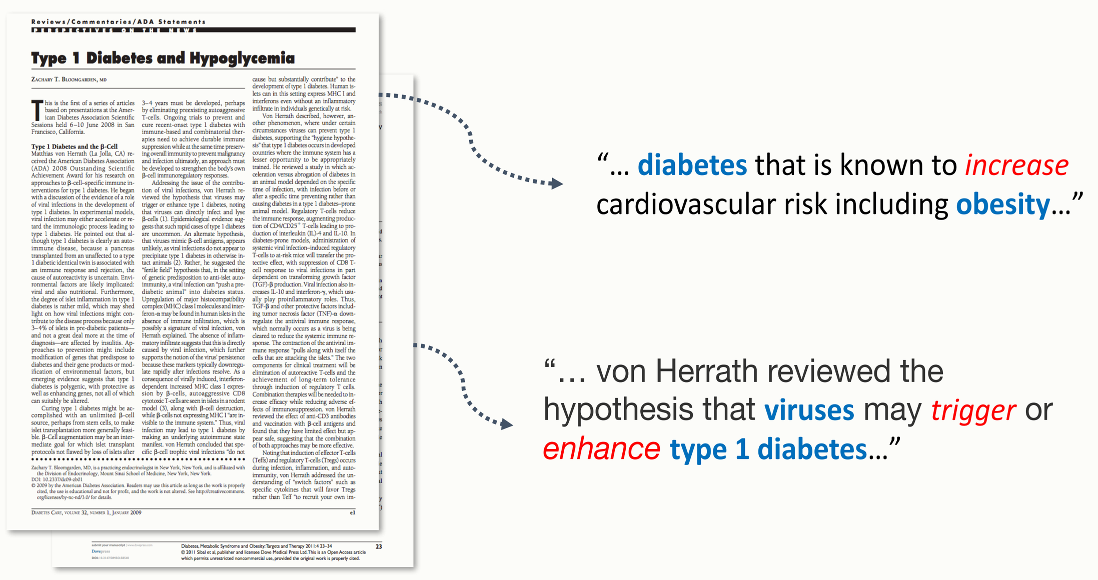

Figure 2. Influence relation fragments are extracted from scientific publications.

Step 2. Assembly

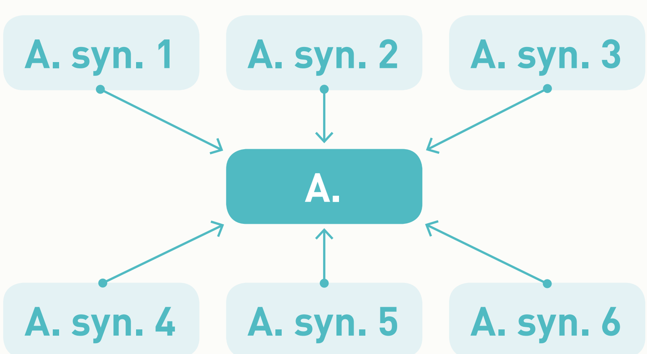

Figure 3. Synonymous entities are linked before stitching together edges. For example, "obesity" and "incidence of obesity" are linked together in this step.

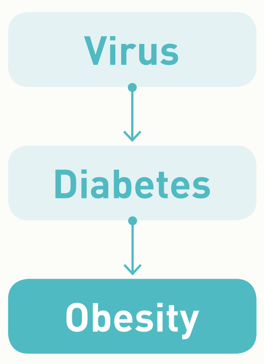

Figure 4. The puzzle pieces are assembled to form a snapshot of a model.

DISCLOSURE

The authors declare a financial interest in lum.ai, which licenses the intellectual property involved in this research. This interest has been properly disclosed to the University of Arizona Institutional Review Committee and is managed in accordance with its conflict of interest policies.

APPROACH (CONT.)

Step 3. Search

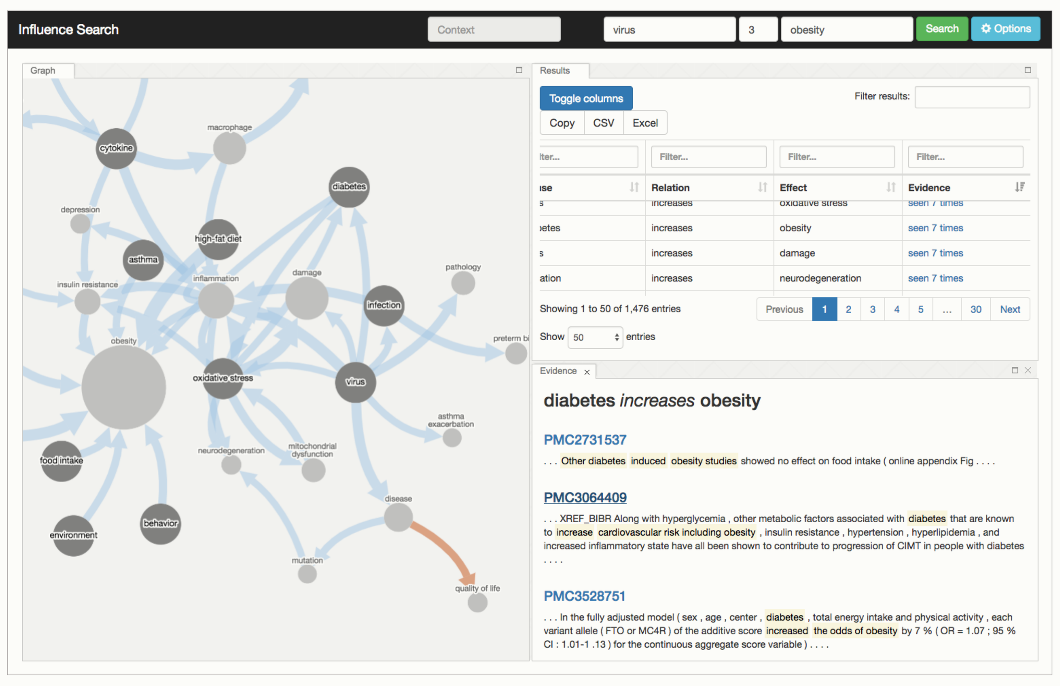

Figure 5. Example of results page for the query “how do viruses indirectly cause obesity?”

DOMAIN-INDEPEDENT MACHINE READING

Reading

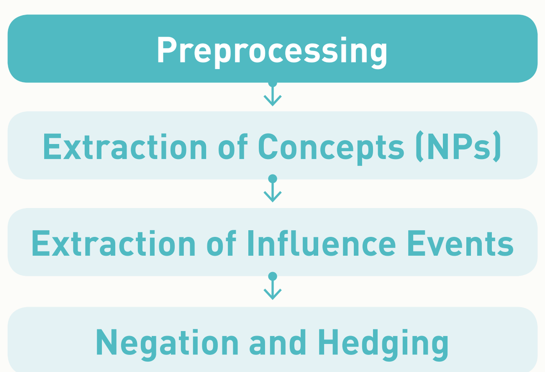

Figure 6. Entities (concepts) are expanded noun phrases (NPs) in the style of open IE.[1] Influence events are extracted using a grammar developed using our own information extraction framework.[2,5,4]

Assembly: Entity linking via grouping on shared content lemmas

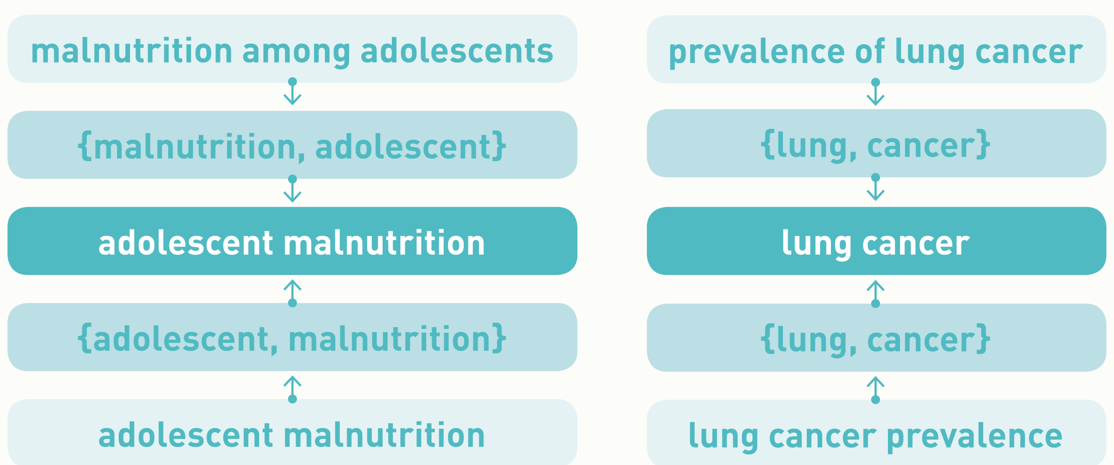

Figure 7. Concept mentions are simplified and linked by retaining only the lemma forms of the essential tokens (nouns, adjectives, & verbs).

SEARCH UI: OVERVIEW

Supported Queries

— Direct & indirect causes and effects

— Direct & indirect paths linking a cause to an effect

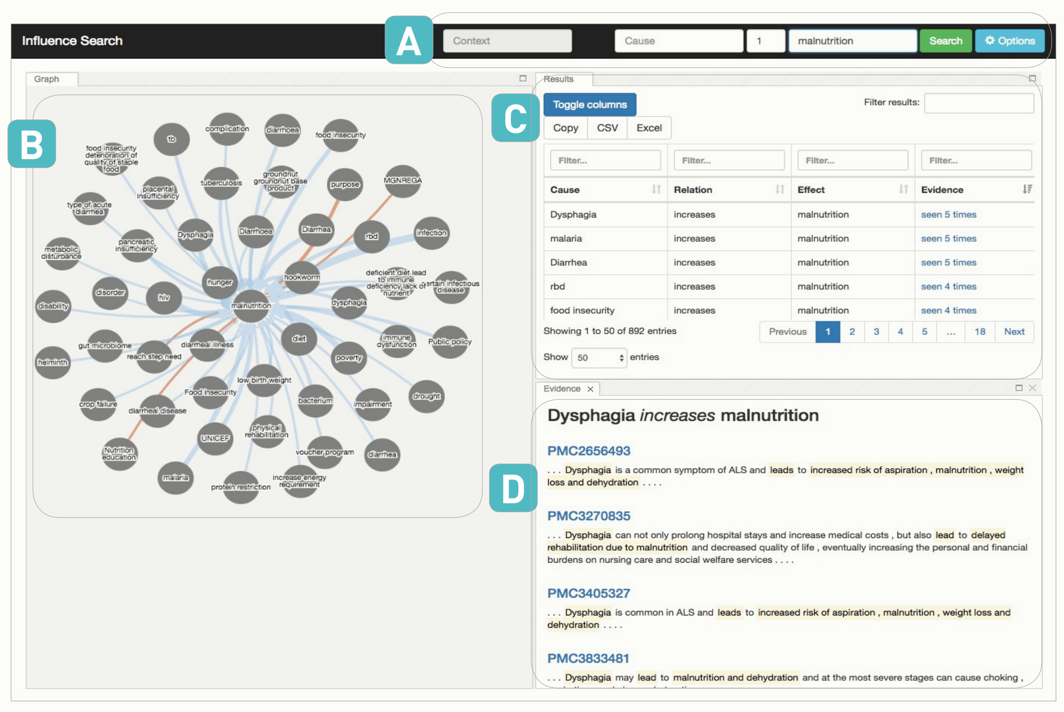

Figure 8. An annotated overview of the various components that comprise the search UI.

(A) Search boxes used to specify the desired cause and/or effect, and the maximum number of hops connecting them. (B) Network graph view of the results, indicating influence relations between concepts. Edge thickness corresponds to the evidence count. Orange edges indicate inhibition. Blue edges indicate promotion. (C) Table of results. Each row corresponds to an edge in the graph. Results can sorted by relevance score or evidence count, and can be further filtered by applying searches to individual columns or globally. Data can be exported to other table formats. (D) Textual evidence corresponding to a selected edge with cause, effect, and trigger highlighted. Paper IDs link back to the full text of the paper when available.

SEARCH UI: CUSTOMIZATION

Customizable views

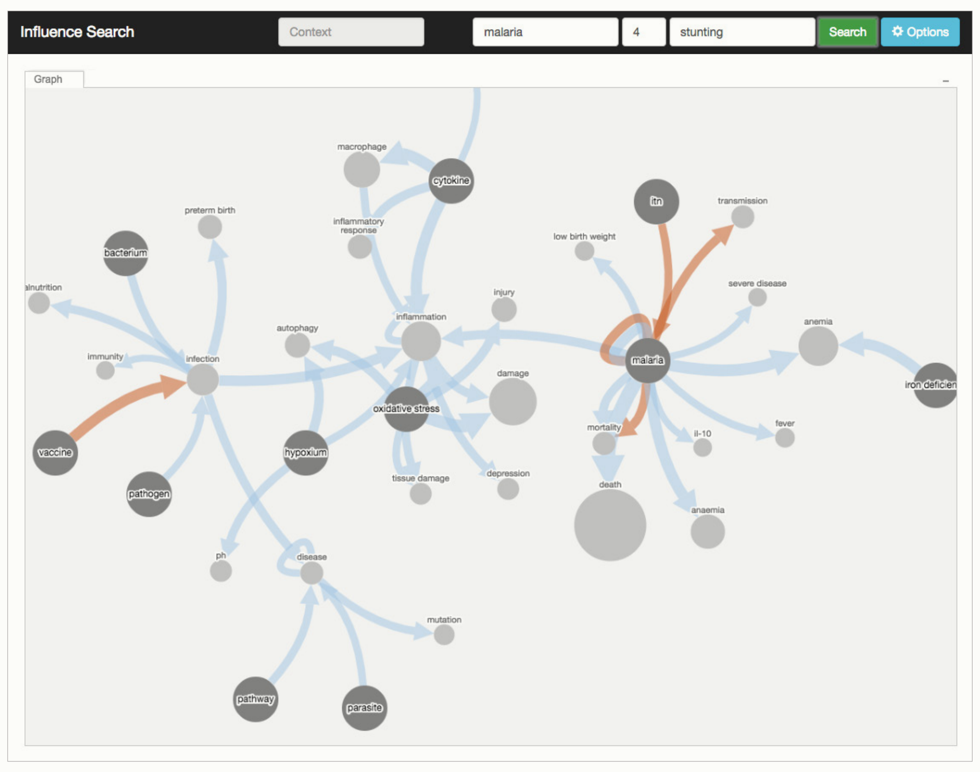

Figure 9. Results of a query for indirect effects of malaria on stunting. In this layout, the table and evidence views have been hidden away as minimized tabs to provide more space for the network graph view.

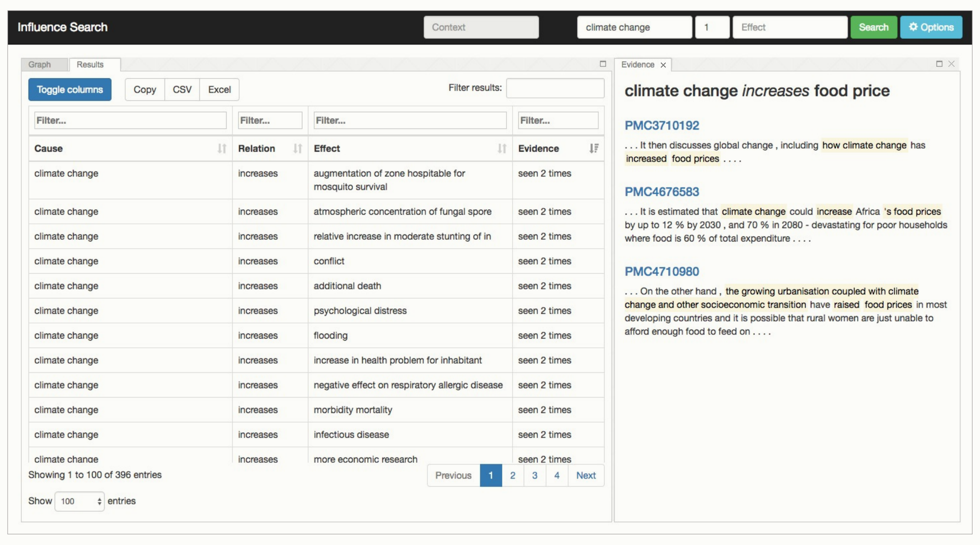

Figure 10. Results of a query for direct effects of climate change. In this configuration, the table view is shown beside the evidence panel.

Model creation workspace

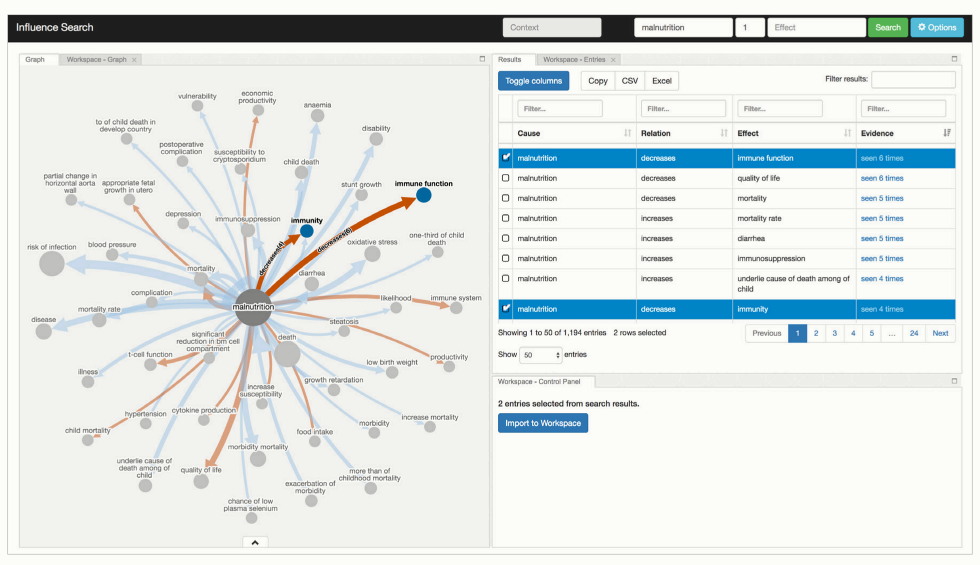

Figure 11. The user chooses which results to import into the persistent model creation workspace.

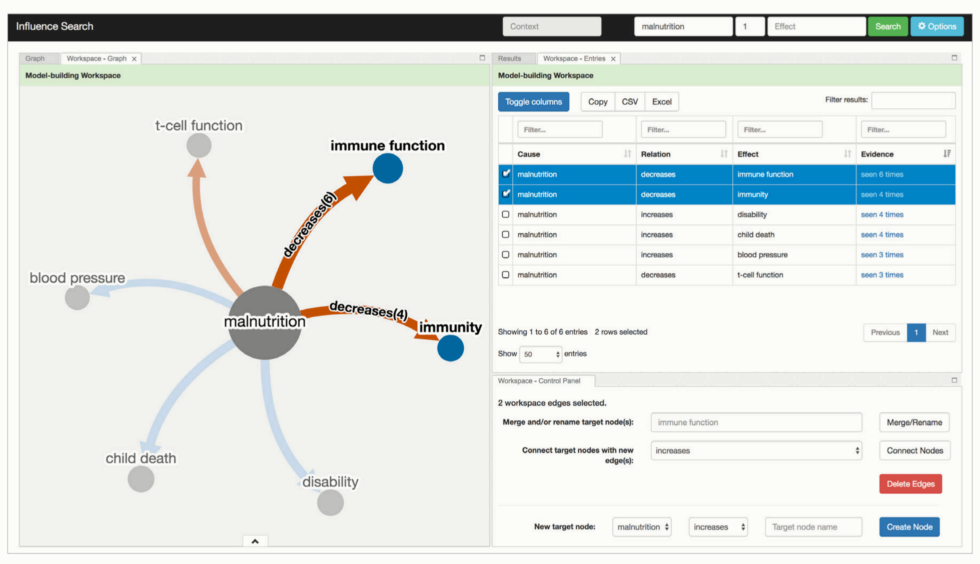

Figure 12. The user constructs a model for the task at hand in a dedicated workspace. This figure shows the process of merging two result nodes into a single one in the new model.

REFERENCES

1. Banko, M., Cafarella, M. J., Soderland, S., Broadhead, M., and Etzioni, O. (2007). Open information extraction from the web. In *IJCAI*, volume 7, pages 2670–2676.
2. Hahn-Powell, G., Bell, D., Valenzuela-Escárcega, M. A., and Surdeanu, M. (2016). This before that: Causal precedence in the biomedical domain. In *Proceedings of the 15th Workshop on Biomedical Natural Language Processing.*
3. Swanson, D. R. (1986). Undiscovered public knowledge. *The Library Quarterly*, 56(2):103–118.
4. Valenzuela-Escárcega, A. M., Hahn-Powell, G., Surdeanu, M., and Hicks, T. (2015). A domain-independent rule-based frame-work for event extraction. In *Proceedings of ACL-IJCNLP 2015 System Demonstrations*, pages 127–132. Association for Computational Linguistics and The Asian Federation of Natural Language Processing.
5. Valenzuela-Escarcega, M. A., Hahn-Powell, G., and Surdeanu, M. (2016). Odin’s runes: A rule language for information extraction. In *Proceedings of the Language Resources and Evaluation Conference.*
6. Wheeler, D. L., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Chetvernin, V., Church, D. M., DiCuccio, M., Edgar, R., Federhen, S., et al. (2007). Database resources of the national center for biotechnology information. *Nucleic acids research*, 36(suppl_1):D13–D21.

BILL & MELINDA GATES *foundation*

HBGD*ki*

Explore the prototype at http://clupublic.cs.arizona.edu:5001

Computational Language Understanding

THE UNIVERSITY OF ARIZONA

ACKNOWLEDGMENTS

The authors would like to thank Lyn Powell for her assistance with the children's health use case, and Zechy Wong for his contributions to improving the UI.

This work was funded in part by the Defense Advanced Research Projects Agency Big Mechanism program under ARO contract W911NF-14-1-0395. This work was funded in part by the Bill and Melinda Gates Foundation HBGD*ki* Initiative.