

NL IAS AUTOMATIC TEXT SUMMARIZATION

By Anna Yesypenko, Computer Science Major at Cornell University
Advised by Lauren M. Stuart and Drs. Julia M. Taylor and Victor Raskin

Problem Statement

- Text summarization is a problem supporting information assurance
- To develop a new algorithm that provides useful, accurate summaries, I plan to ontologically compare the first paragraph of an American news report, which can be viewed as a summary written by journalists, to a computer-generated summary of rest of the article.

Ways to Approach Automatic Summarization

PREVIOUS ATTEMPTS – EXTRACTIVE

Use a subset of words, phrases, or sentences from the original text

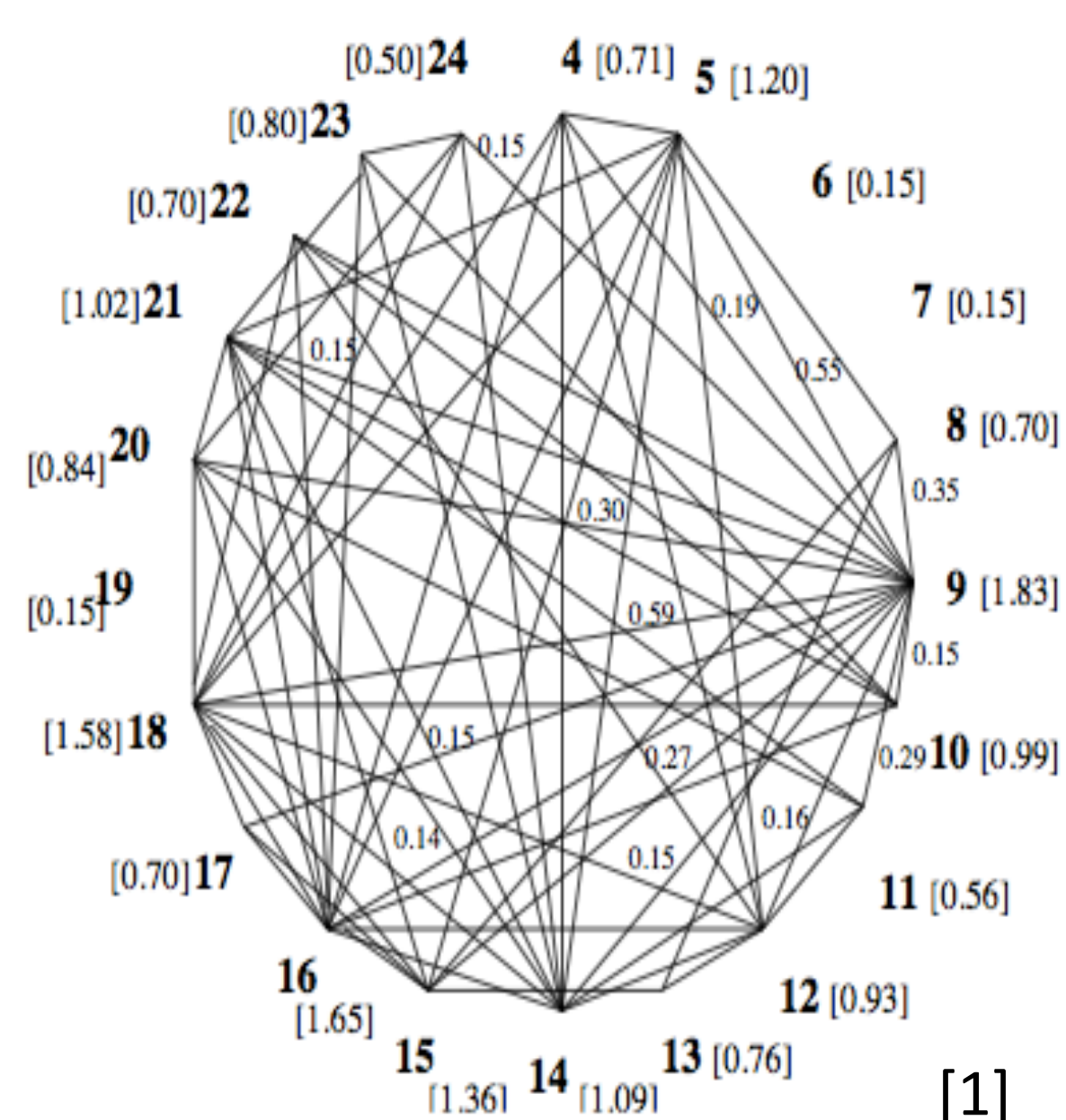
Identify keywords in the text (can be multiple documents) [2]

FLAWS

- Difficult to remove overlapping sentences
 - “Alice ate lunch with Greg at Gino’s.”
 - “Alice ordered a steak, and Greg ordered a hamburger at Gino’s.”
 - “Alice and Greg ordered entrees at Gino’s.”
- Inaccurate frequency analysis
 - “Allan Smith robbed a 7-11 in Tallahassee, Florida.”
 - “The man entered the convenience store late at night.”

Extract sentences that are topic sentence or that contain a high concentration of keywords

Get rid of overlapping sentences



GRAPHICAL REPRESENTATION OF NEWS ARTICLE

- Vertices represent sentences
- Edge between two vertices means that the two sentences contain the same words
- Algorithm uses high-degree vertices in summary.

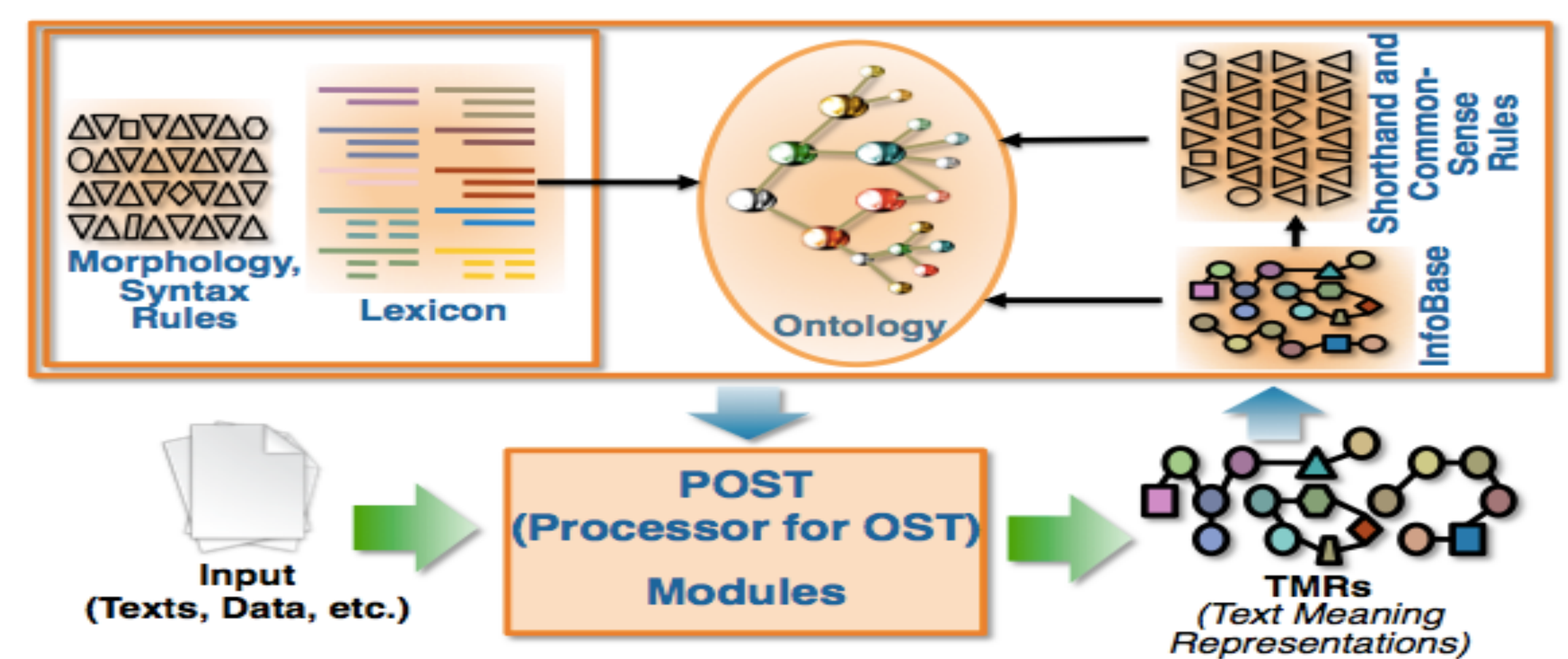
CITATIONS

- [1] Mihalcea, R. (2004). Graph-based ranking algorithm for sentence extraction, applied to text summarization. *Proceedings of the ACL 2004 on Interactive Poster and Demonstration Sessions*, (4), 20–es. doi:10.3115/1219044.1219064s
- [2] Radev, D. R., Jing, H., Styś, M., & Tam, D. (2004). Centroid-based summarization of multiple documents. *Information Processing & Management*, 40(6), 919–938. doi:10.1016/j.ipm.2003.10.006
- [3] Steenhusen, J. (2014, Jun 27) “Exclusive: U.S. government scientists retrace events leading to anthrax breach.” *Reuters* <http://www.reuters.com/article/2014/06/27/us-usa-anthrax-tests-idUSKBN0F210N20140627>

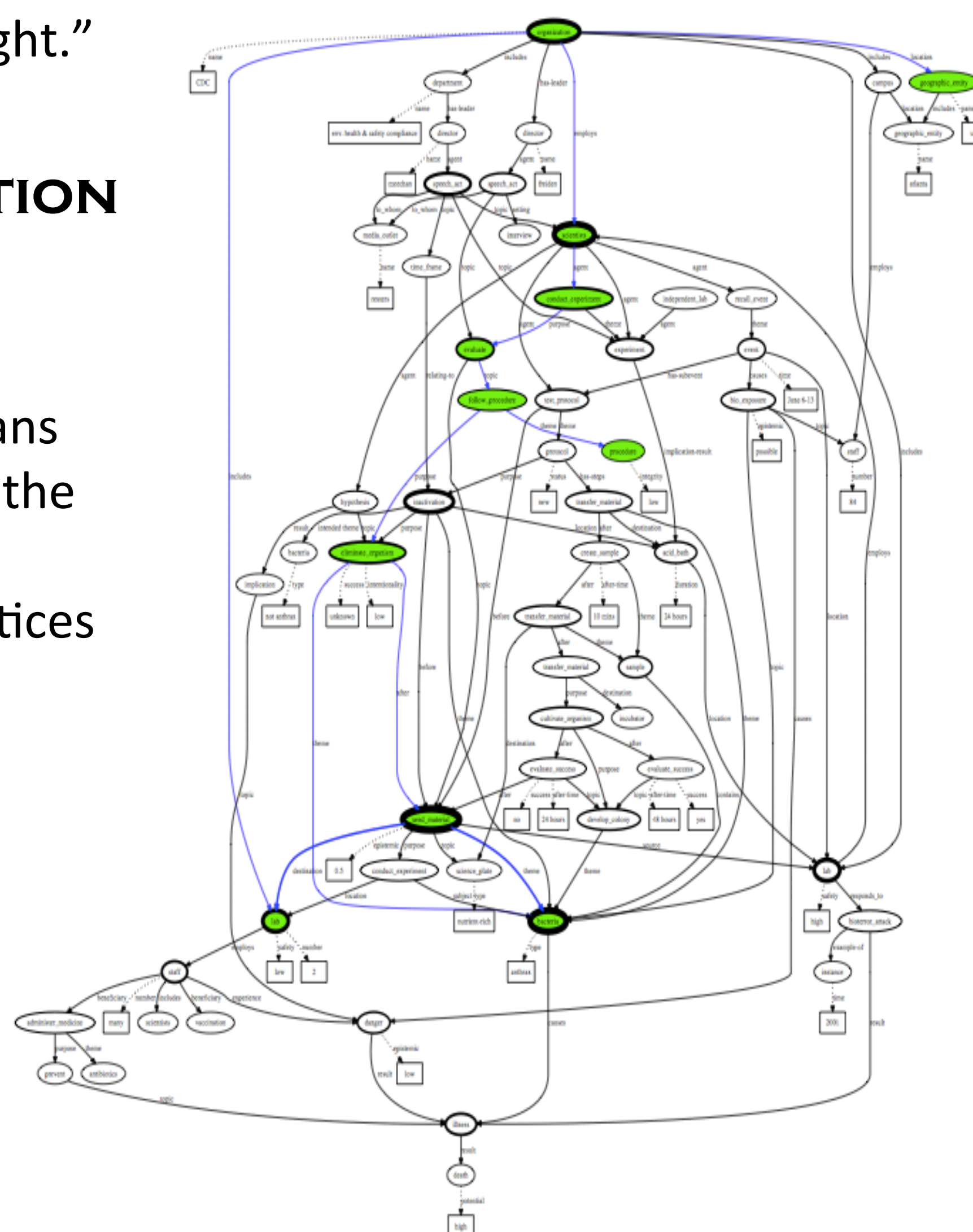
MY SOLUTION – ABSTRACTIVE

Build an internal text meaning representation (TMR) then construct a summary closer to what a human may generate

ONTOLOGICAL SEMANTICS TECHNOLOGY



CONDENSED TMR FROM ANTHRAX ARTICLE [3]



BASICS OF ALGORITHM

- Weigh edges by the frequency of their use in text
- Weigh vertices by the summation of their edge weights
- For each vertex:
 - Calculate the cost of each path leading to it, where each edge of the path has cost 1/(edge weight)
 - Remove edges corresponding to inefficient paths
- Remove remaining lowly-weighted vertices with out-degree 0