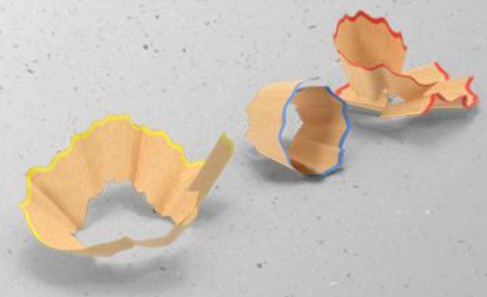
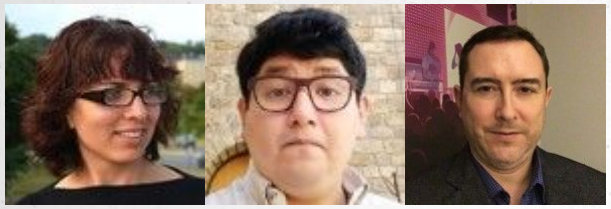
JSALT 2023
Evaluation and Conference Systems

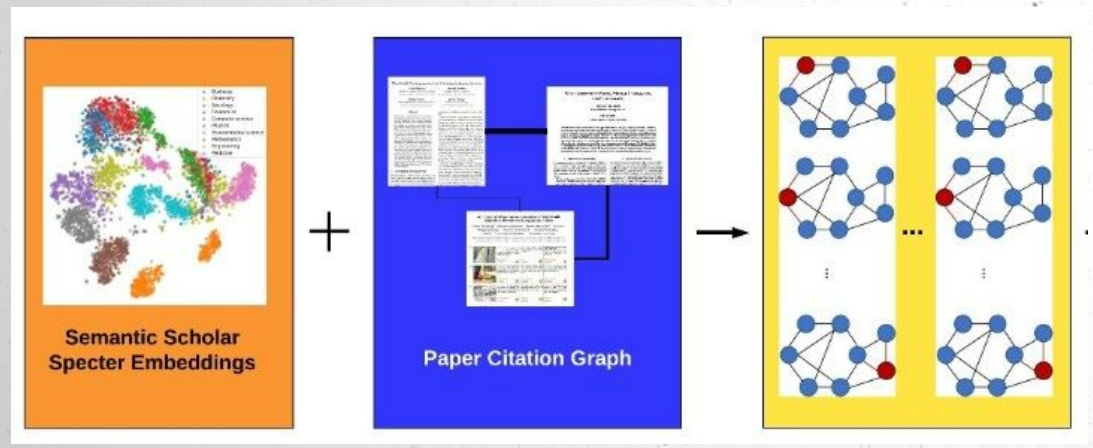
Shabnam Tafreshi, Rodolfo Zevallos and John E. Ortega
May 14, 2023
Northeastern University



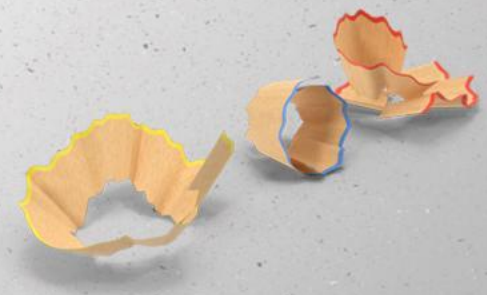
Overview

What we will cover today

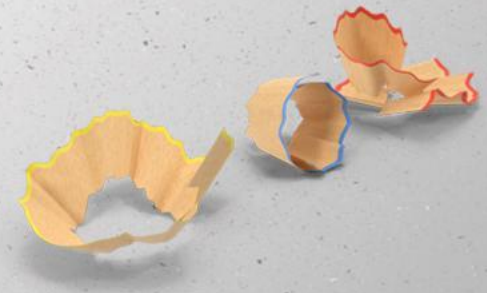




- Introduction - Who We Are
- Evaluation
- Conference Systems
- Delivery Proposal
- Winding Down

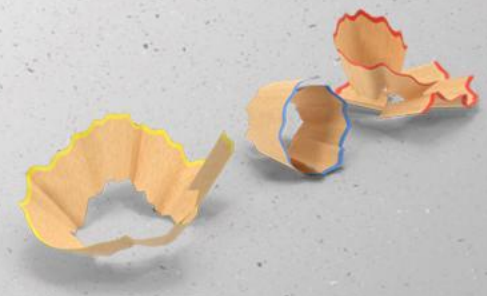


Who are we?



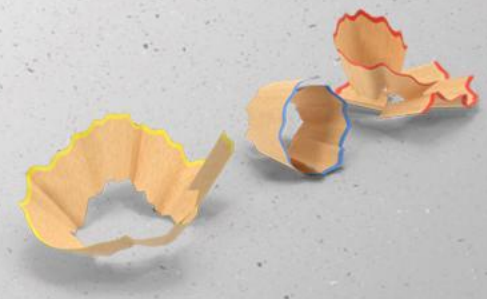
I believe that bias in models and linguistics is clearly inevitable. It is our responsibility to find those errors...

-- Shabnam Tafreshi



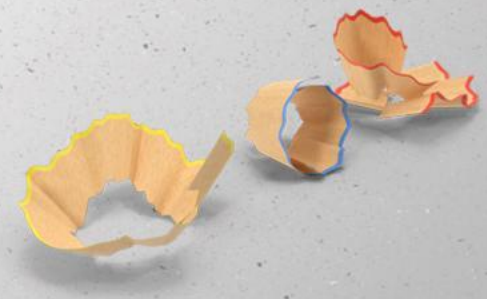
Solving the low-resource language problem for Quechua and other indigenous languages has not been solved. Languages are going extinct and we need to help.

-- Rodolfo Zevallos

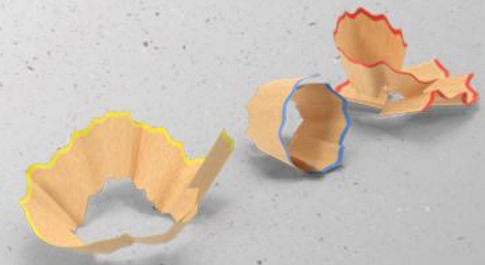


Natural language processing is a means to an end. Artificial intelligence is the act of applying mathematical reason to automate human tasks. The low-resource task has a lot of opportunity for everyone to solve!

-- John E. Ortega



Evaluation



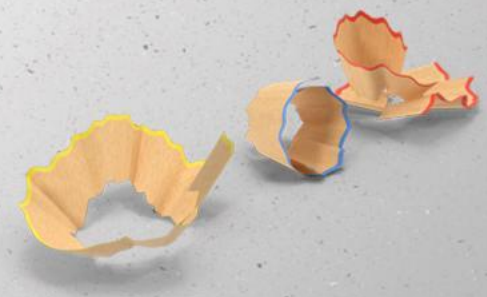
OGB-LSC:

A Large-Scale Challenge for Machine Learning on Graphs

Datasets:

- MAG240M - Node-Level Prediction
- WikiKG90M - Link-Level Prediction (corruption)
- PCQM4M - Graph-Level Prediction (molecular energy)

No human-in-the-loop methods!

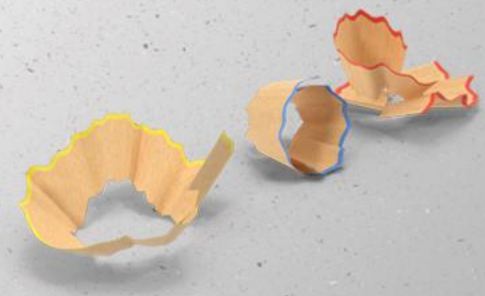


SciRepEval:

A Multi-Format Benchmark for Scientific Document Representations

- Scientific document representations.
- 25 tasks across classification, regression, ranking and search.
- Embeddings and more
- MAP evaluation
- Kendall evaluation
- F1 Score
- Peer-review matching based on previous literature
    - 0-3 relevance rating scale
    - Paper embeddings
    - Cosine similarity

No new data captured from the latest systems!



Expertise Modeling for Matching Papers with Reviewers

- Language-model approach
- Author in existing set compared to non-existing set
- Topic-based
- Author-Persona-Topic (APT)
- Evaluation with topics, LDA, sampling, and more.
- Close to what we want to do.
- 2006 Evaluation from NIPS papers

Does not use/create recent data or deep-learning (HF) models!

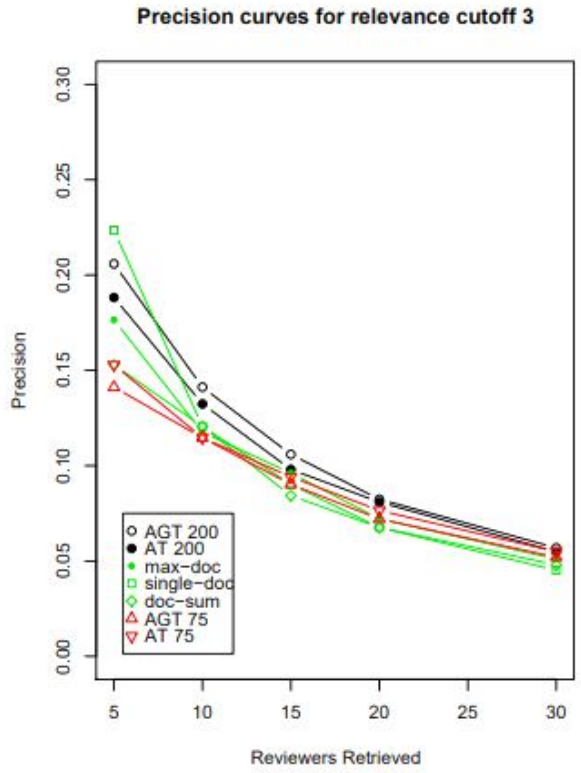
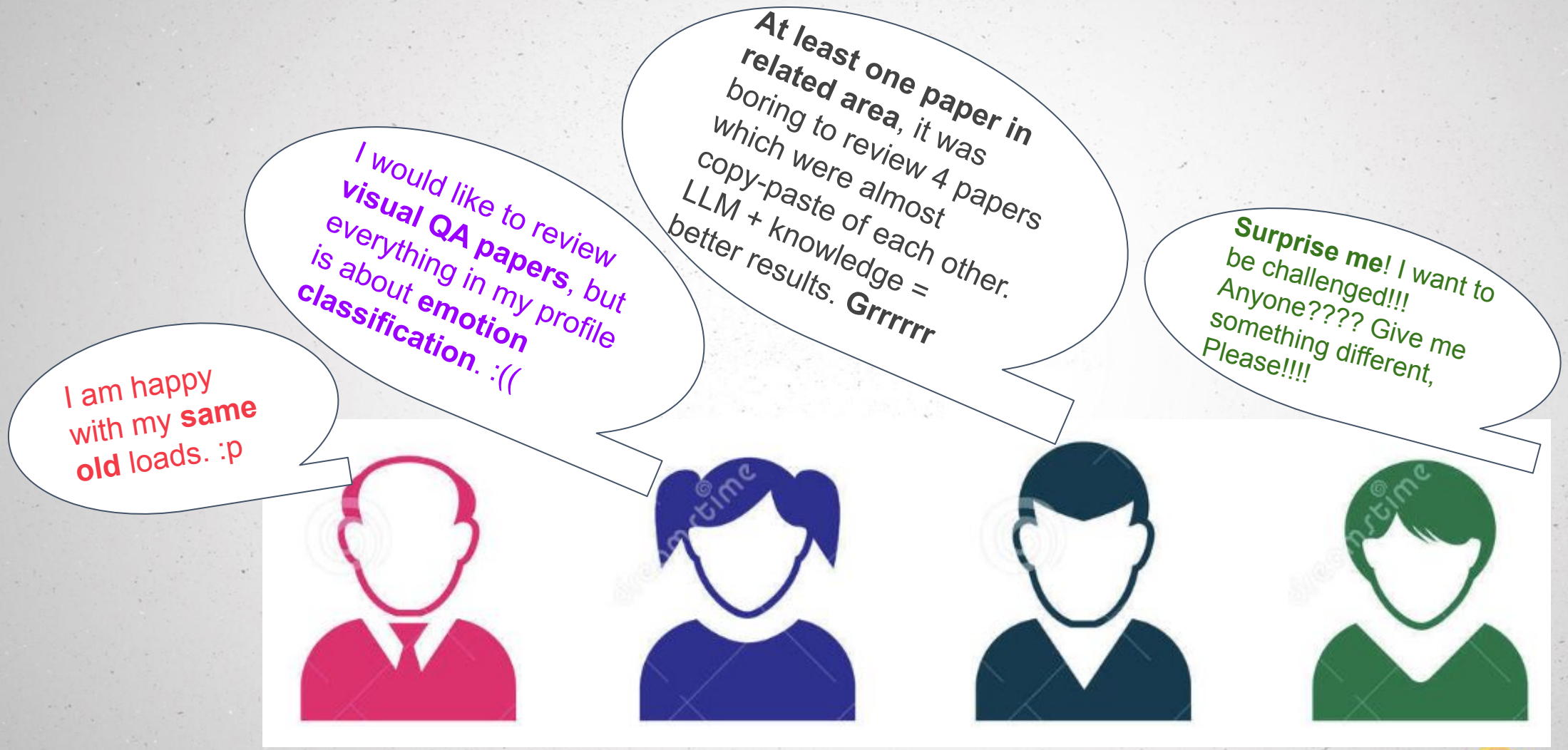


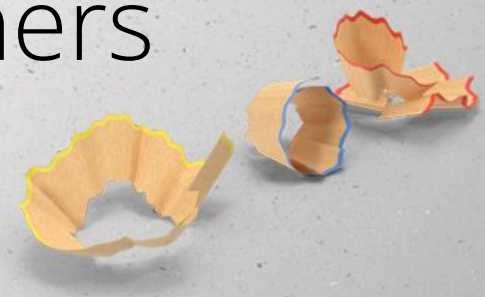
Figure 3: The precision of each model as more documents are retrieved for relevance cutoff 3. The same general patterns are present at this level of relevance as in the lower-cutoff evaluation. The topic models with 200 topics are the best overall, while the single-document author language model has the highest precision in the first five reviewers retrieved.

Reviewers in the Loop



Annotation Design

- Reviewers' Feedback - create "**Reviewer Desire Taxonomy**"
- Reviewers' Motivation - behavioural study of reviewers given the current load (ARR, ACL, Worskshors, etc.) and in general
- Quality of Reviews - self-ranking or others ranking the reviews



Annotation Design

Reviewers' Feedback

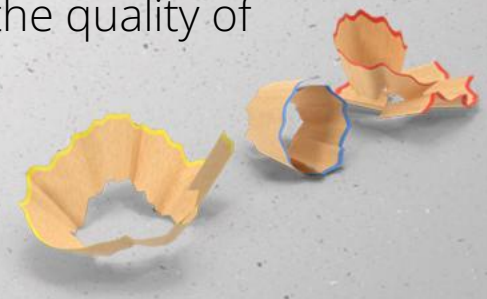
- How satisfy you are with your paper load?
- Will you review again?
- Do you want to receive different papers from your area?
- Do you want to be surprised?
- Do you want one/two papers to be from sister area of your expertise?

Reviewers' Motivation

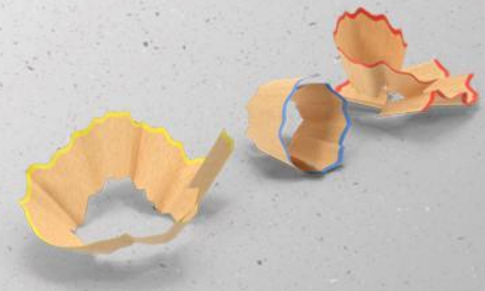
- How excited you were during review time?
- Do you think there are so many papers in your court?
- When did you start reading your papers?

Quality of Reviews

- Creating taxonomy
- Self-ranking or others rankings the quality of reviews



Conference Systems

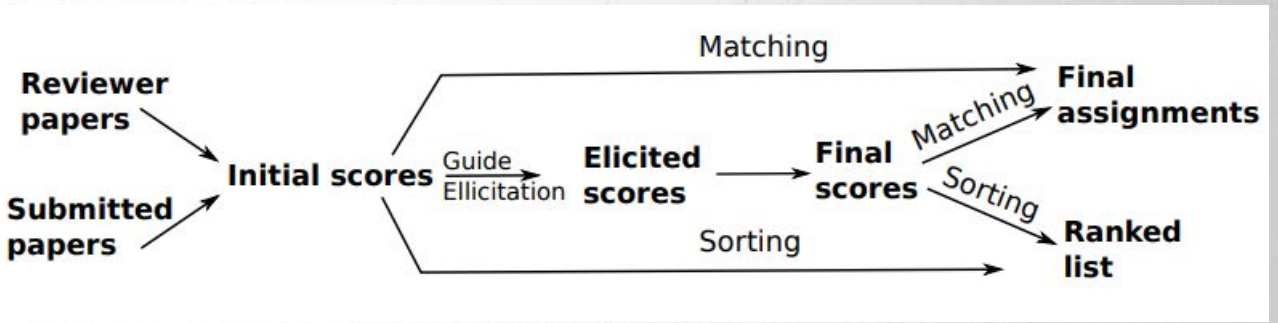


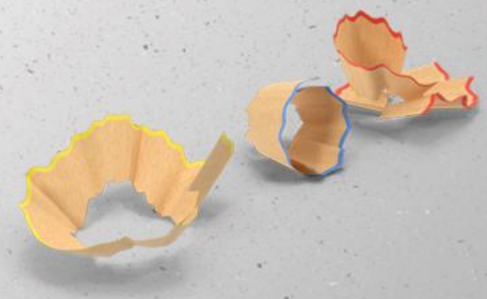
What are Conference Systems?

Tools that use machine learning and natural language processing to analyze and understand academic articles.

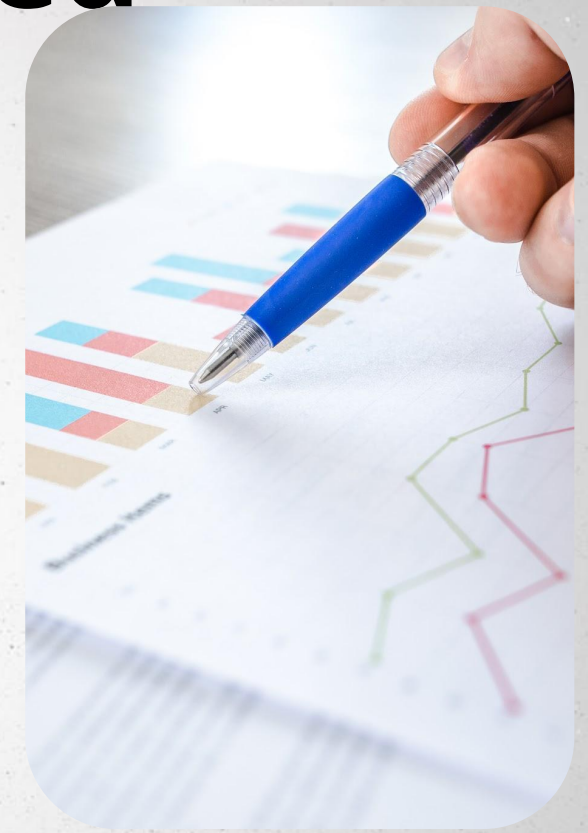
Provide recommendations for reviewers based on relevant articles in the reviewer's topic area.

Useful for researchers who want to stay up-to-date and reviewers looking for relevant articles to evaluate.

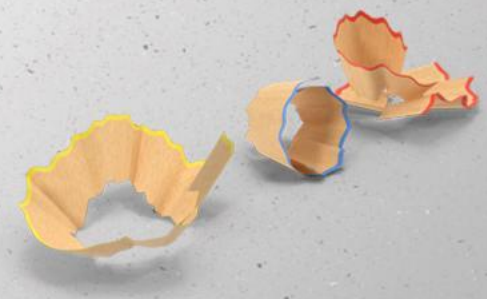




Techniques Currently Used



- Use techniques like feature extraction, pattern identification, and content classification.
- Take into account features such as the content of the article, references cited, metadata (category, tl&dr, conference, etc.) and user preferences.
- New techniques being explored, such as co-citation network analysis and semantic similarity identification between articles.



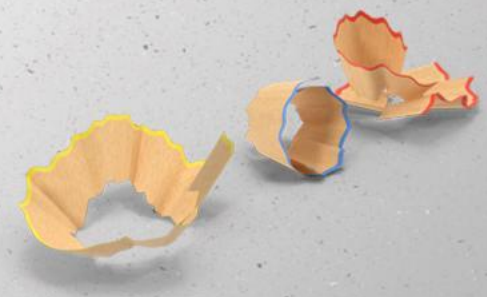
Core Features Used by Conference Systems

Article metadata

References cited

NLP techniques

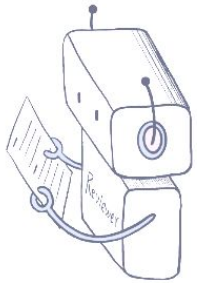
User Features



Some Conference Systems

SEMANTIC **SCHOLAR**

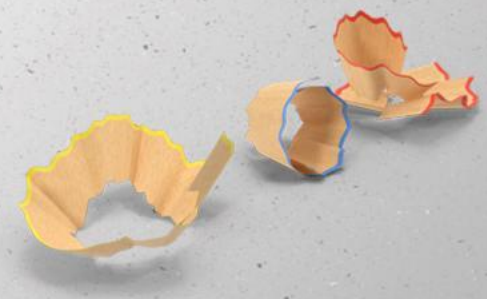
A free, AI-powered research tool for scientific literature



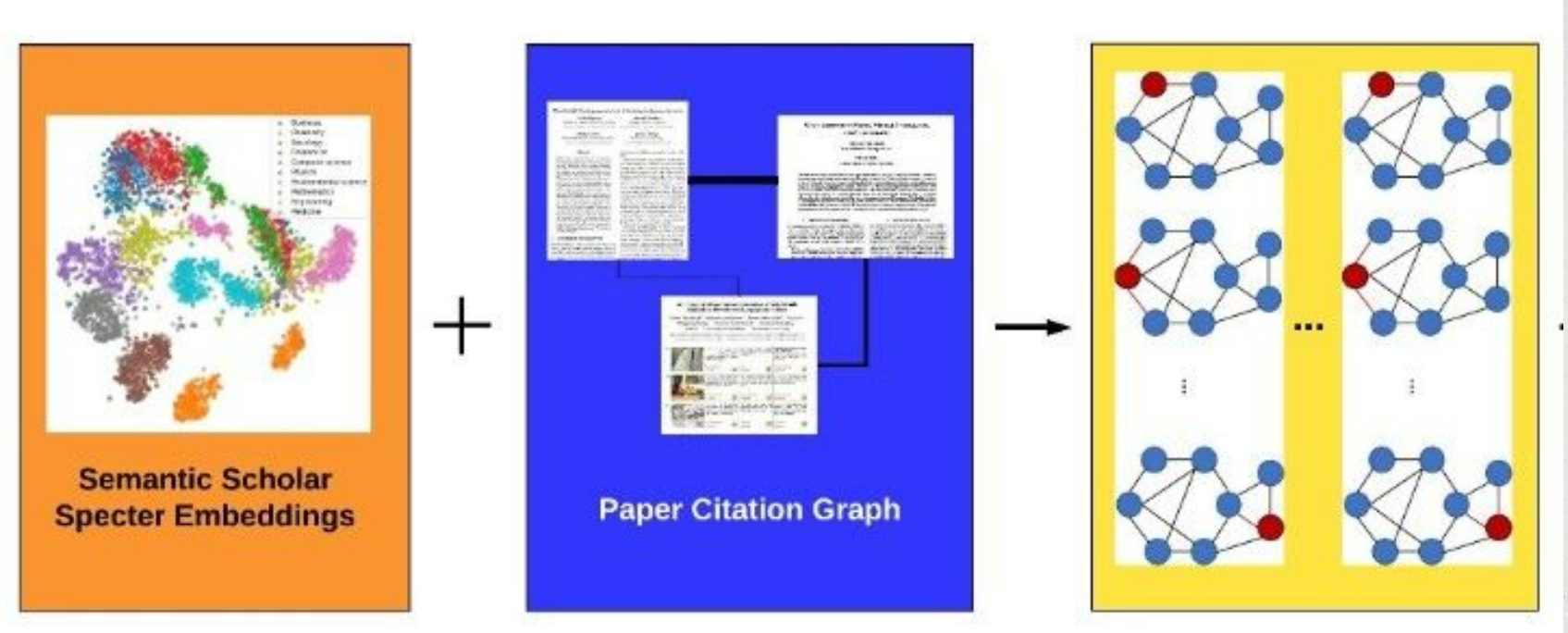
Review Advisor

Clarivate™ | ScholarOne Manuscripts Support

Peerwith

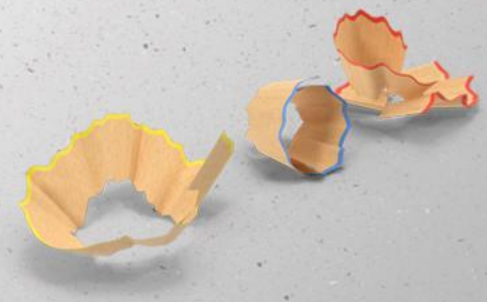


Ideas for Improving Academic Search

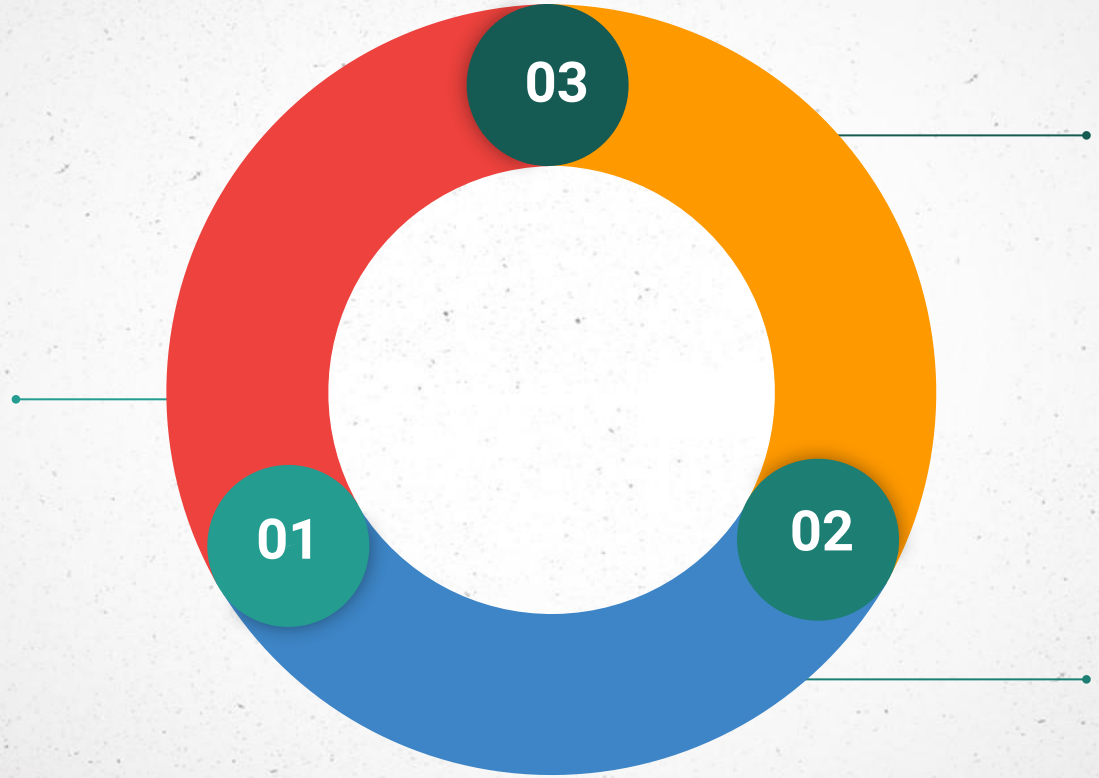


Text **Context** **Graph**

- Data-centric
- More features
- Power Graph



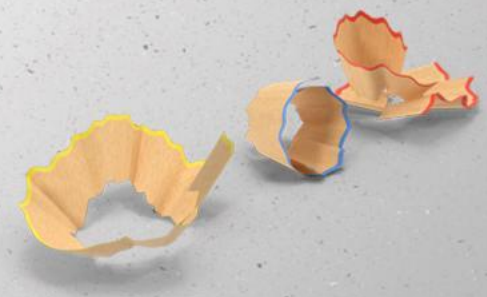
Ideas for Improving Academic Search



Idea A: Reviewer Prediction

Idea C: Relevant Paper Identification

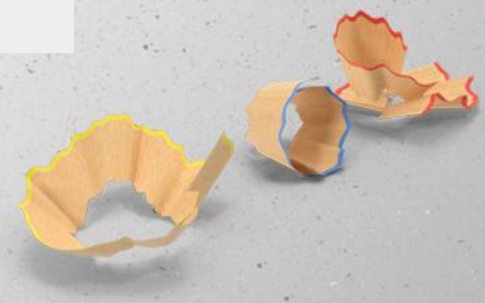
Idea B: Query Relevance



Idea A: Reviewer Prediction
Idea C: Relevant Paper Identification

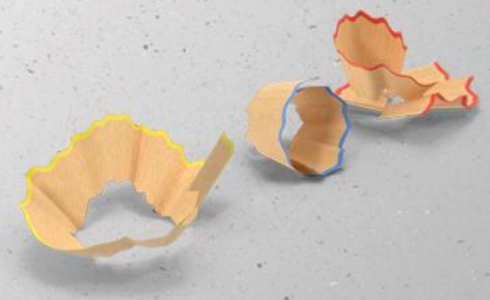
01	Qualifications	• Students are unqualified. • Reviewers should have published (h-index 5). • Reviewers should know the background literature. • (Reviewers may have read but not been cited)
02	Sympathetic	• Reviewers should have published in the area.
03	Conflicts	• Avoid reviewers that co-author together
04	Diversity	• 3 reviewers per paper (one should be for experience, one for new blood, and one for outreach)

Is this realistic? We will have enough motivated reviewers?

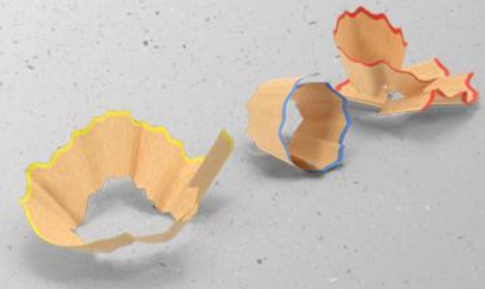


Idea B: Query Relevance

Features	Google Scholar	Idea
Citations	✓	✓
Keywords	✓	✓
History	✓	✓
Profile	✓	✓
Collaborations	✓	✓
Paper level	✓	✓
Date		✓
Abstract		✓
Country of origin		✓
Research method		✓
Topic		✓



Delivery Proposal



Winding Down

What should we use for a system and evaluation?

Human-in-the-loop

Reviewers Desire Taxonomy

New Conference System

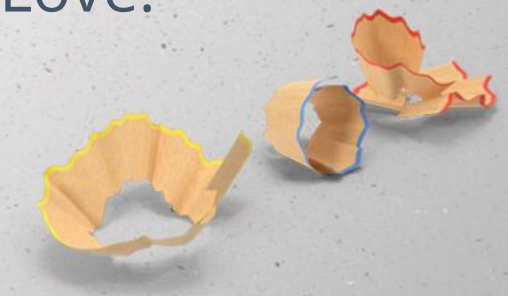
Include Preference

Evaluation Benchmarks from Previous Work

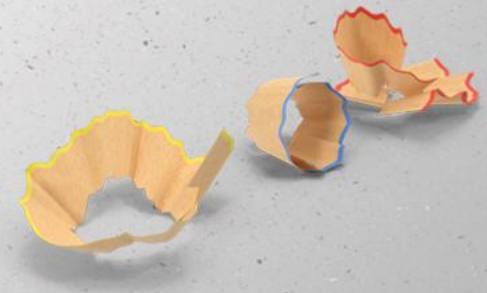
Creation of a New Eval Metric

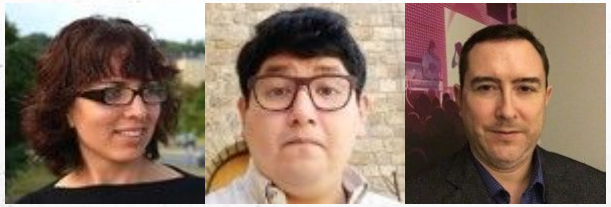
Website creation

Kiss and be happy! Love! Peace!



Winding Down





THANKS!