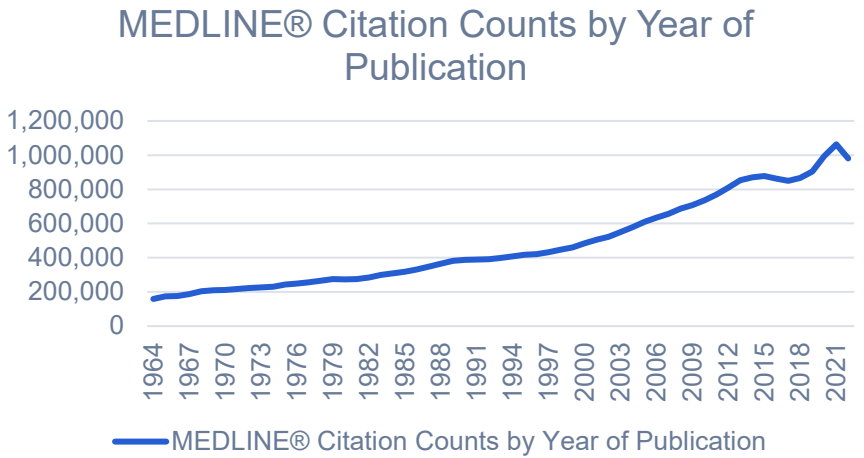
SciMON 🧪 : Scientific Inspiration Machines Optimized for Novelty

Qingyun Wang[1], Doug Downey[2], Heng Ji[1], Tom Hope[2,3]
[1]University of Illinois at Urbana-Champaign
[2]Allen Institute for Artificial Intelligence (AI2)
[3]The Hebrew University of Jerusalem
qingyun4@illinois.edu

Motivation: Augmenting Human Innovation

- Millions of scientific papers are published every year
  - More than 1M papers are added to PubMed every year, bringing the total number of papers to over 36M

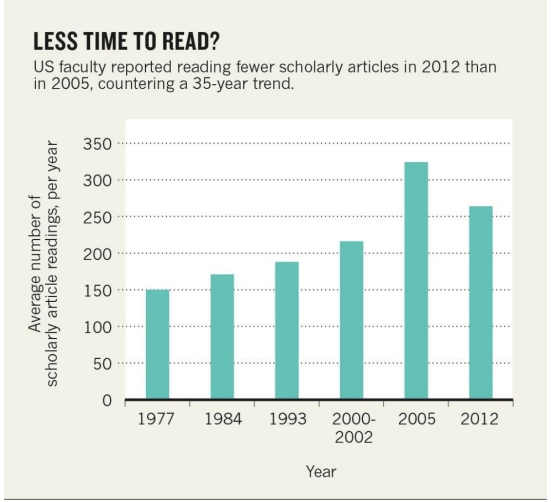


https://www.nlm.nih.gov/bsd/medline_cit_counts_yr_pub.html

Landhuis, E. (2016). Scientific literature: Information overload. Nature, 535(7612), 457-458

Motivation: Augmenting Human Innovation

- Human's reading ability keeps almost the same across years
  - US scientists estimated that they read, on average, only about 300 papers per year



Landhuis, E. (2016). Scientific literature: Information overload. Nature, 535(7612), 457-458

Why do we want AI-Assisted Hypothesis Generation?

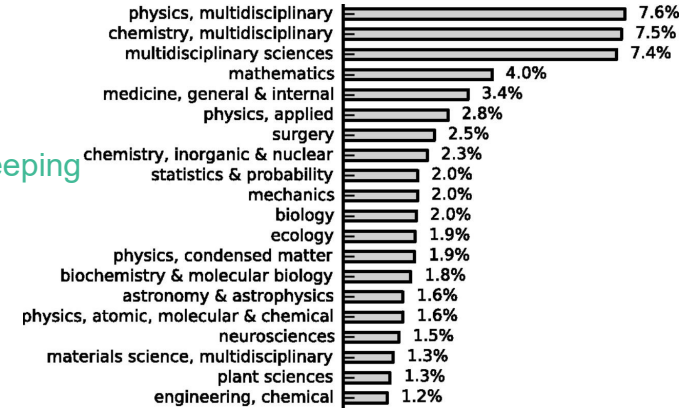
- “Sleeping beauties” in science: Discoveries that lay dormant and largely unnoticed for long periods of time before suddenly attracting great attention
    - Examples include a now famous 1935 paper by Einstein, Podolsky, and Rosen on quantum mechanics; a 1936 paper by Wenzel on waterproofing materials; and a 1958 paper by Rosenblatt on artificial neural networks

Clauset, A., Larremore, D. B., & Sinatra, R. (2017). Data-driven predictions in the science of science. Science, 355(6324), 477-480.
Ke, Q., Ferrara, E., Radicchi, F., & Flammini, A. (2015). Defining and identifying sleeping beauties in science. Proceedings of the National Academy of Sciences, 112(24), 7426-7431.
Foster, J. G., Rzhetsky, A., & Evans, J. A. (2015). Tradition and innovation in scientists’ research strategies. American sociological review, 80(5), 875-908.

Why do we want AI-Assisted Hypothesis Generation?

- "Sleeping beauties" in science: Discoveries that lay dormant and largely unnoticed for long periods of time before suddenly attracting great attention
    - A systematic analysis of nearly 25 million publications in the natural and social sciences over the past 100 years found that sleeping beauties occur in all fields of study

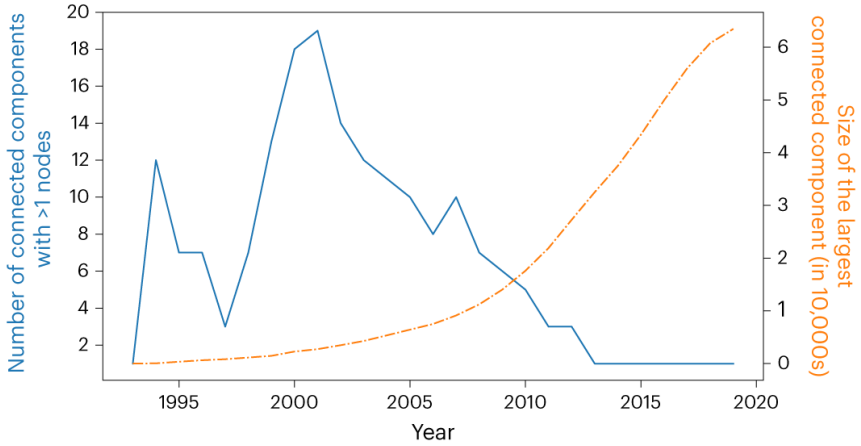
Top 20 disciplines producing Sleeping Beauties in science



Clauset, A., Larremore, D. B., & Sinatra, R. (2017). Data-driven predictions in the science of science. Science, 355(6324), 477-480.
Ke, Q., Ferrara, E., Radicchi, F., & Flammini, A. (2015). Defining and identifying sleeping beauties in science. Proceedings of the National Academy of Sciences, 112(24), 7426-7431.
Foster, J. G., Rzhetsky, A., & Evans, J. A. (2015). Tradition and innovation in scientists' research strategies. American sociological review, 80(5), 875-908.

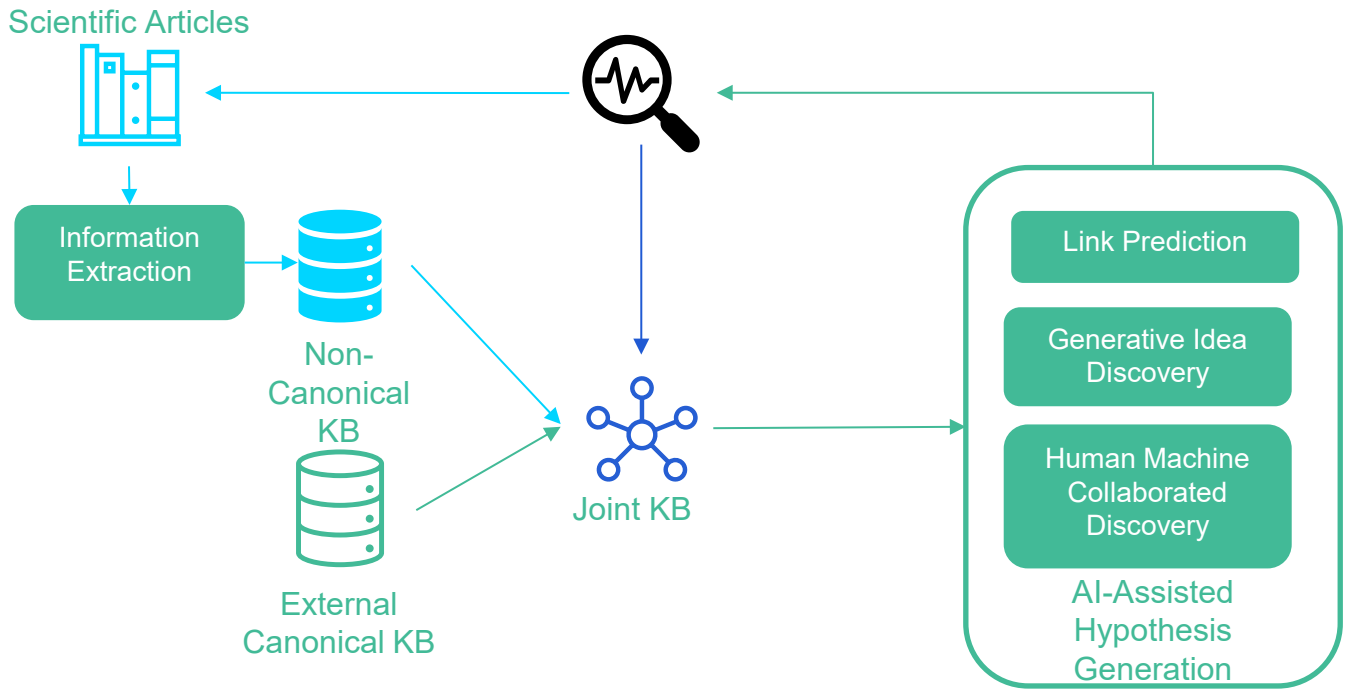
Why do we want AI-Assisted Hypothesis Generation?

- Most papers build on existing knowledge to formulate new innovations
  - Foster et al. (2015) shows that more than 60% of 6.4 million papers in biomedicine and chemistry published between 1934 and 2008 report findings that build on existing knowledge and provide additional innovations and improvements



Clauset, A., Larremore, D. B., & Sinatra, R. (2017). Data-driven predictions in the science of science. Science, 355(6324), 477-480.
Ke, Q., Ferrara, E., Radicchi, F., & Flammini, A. (2015). Defining and identifying sleeping beauties in science. Proceedings of the National Academy of Sciences, 112(24), 7426-7431.
Foster, J. G., Rzhetsky, A., & Evans, J. A. (2015). Tradition and innovation in scientists' research strategies. American sociological review, 80(5), 875-908.

Types of AI-Assisted Hypothesis Generation

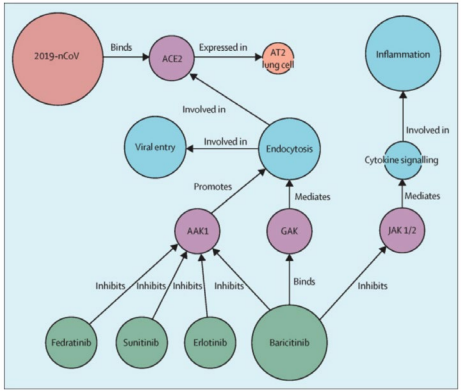


Swanson, D. R. (1986). Undiscovered public knowledge. The Library Quarterly, 56(2), 103-118.

Background: Scientific Knowledge Discovery

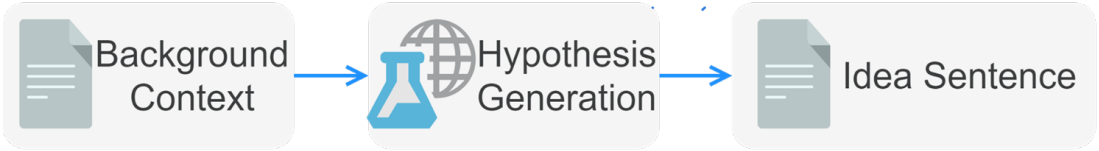
- Literature-based Discovery
  - Predict missing links in KG (e.g., drug→disease)
  - Can lead to important discoveries

  But
  - Limited to curated entities and relations
  - Limited to certain domains
  - Cannot model nuanced contexts
    - (e.g., target application settings, requirements and constraints, motivations and challenges)



Contextualized Literature-based Discovery

- Input
  - Current problems, motivations, experimental settings and constraints
  - A seed term that should be a focus point of the generated idea
- Output
  - A generated novel hypothesis as a natural language sentence



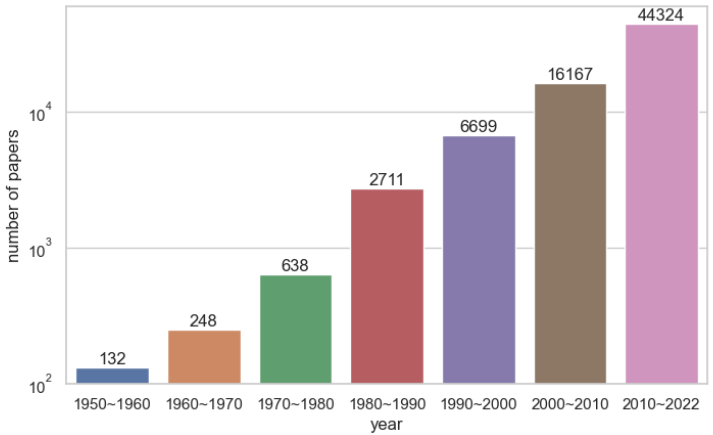
Seed Term: **knowledge acquisition**
Background: ... This requires plms to **integrate the information from all the sources** in a lifelong manner. Although this goal could be achieved by exhaustive pre-training on all the existing data, such a process is known to be **computationally expensive.**

Specifically, ELLE consists of (1) **function preserved model expansion**, which flexibly expands an existing PLM's width and depth to improve the efficiency of **knowledge acquisition** ...

Qin, Y., Zhang, J., Lin, Y., Liu, Z., Li, P., Sun, M., & Zhou, J. (2022). ELLE: Efficient lifelong pre-training for emerging data. ACL Findings 2022.

Dataset Construction

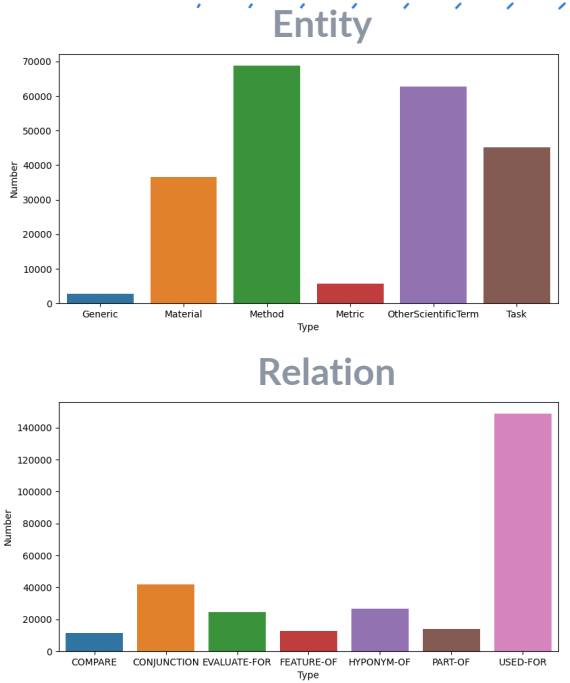
- Construct a corpus from 67,408 ACL Anthology papers from 1952 to 2022 with 5,946 papers from 2021, and 2,588 papers from 2022



Qin, Y., Zhang, J., Lin, Y., Liu, Z., Li, P., Sun, M., & Zhou, J. (2022). ELLE: Efficient lifelong pre-training for emerging data. ACL Findings 2022.

Dataset Construction

- Given a paper in the previous dataset, we perform the following steps to build a knowledge graph:
  - Named Entity Recognition (PLMarker)
  - Relation Extraction (PLMarker)
  - Coreference (SciCo)
  - Abbreviation Extraction (ScispaCy)



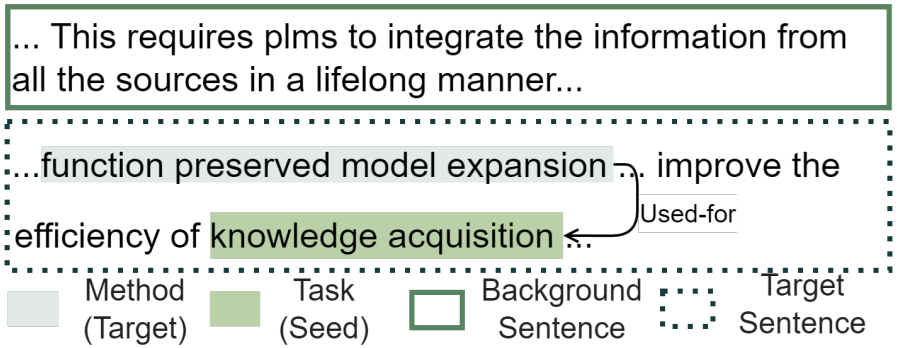
Ye, D., Lin, Y., Li, P., & Sun, M. (2022, May). Packed Levitated Marker for Entity and Relation Extraction. In Proceedings of ACL 2022.

Cattan, A., Johnson, S., Weld, D. S., Dagan, I., Beltagy, I., Downey, D., & Hope, T. (2021). SciCo: Hierarchical Cross-Document Coreference for Scientific Concepts. In 3rd AKBC.

Neumann, M., King, D., Beltagy, I., & Ammar, W. (2019). ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. In Proceedings of the 18th BioNLP Workshop and Shared Task.

Dataset Construction

- Perform scientific sentence classification to classify sentences from the abstract into five categories including *Background*, *Method*, *Objective*, *Other*, and *Result*
  - Select sentences with labels of *Background* and *Other* as background context
- Focus on *used-for* relations, which usually include tasks and methods



Split	Forward	Backward	Total
Train	55,884	58,426	114,310
Valid	7,938	8,257	16,195
Test	2,623	2,686	5,309

Arman Cohan, Iz Beltagy, Daniel King, Bhavana Dalvi, and Dan Weld. 2019. Pretrained language models for sequential sentence classification. EMNLP 2019
Qin, Y., Zhang, J., Lin, Y., Liu, Z., Li, P., Sun, M., & Zhou, J. (2022). ELLE: Efficient lifelong pre-training for emerging data. ACL Findings 2022.

Quality of IE Preprocessing

- Keep high-confidence outputs from IE models to reduce errors
- Perform manual quality evaluation for each preprocessing stage
  - Overall pass rate after all steps are applied is 79.7%

Stage	PL-Maker Entities	PL-Maker Used-for Relations	SciCo Coreference	Scispacy Abbreviation Detection	Sentence Classification
Precision	91.3%	65.4%	97.2%	100%	100%

Gold Test Subset Annotation

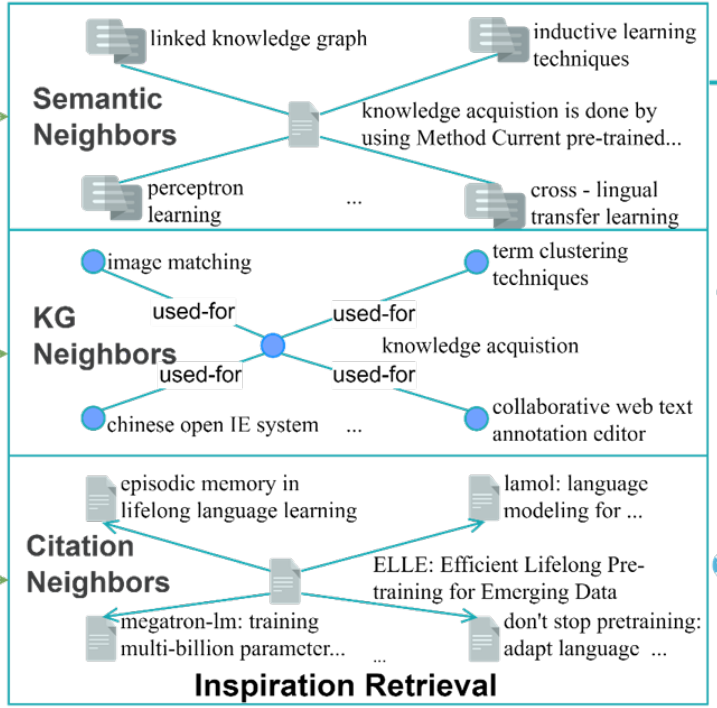
- Exclude instances with trivial overlap between ground truth and background
- Remove cases with irrelevant background
- Retain only instances where the target relation (from which the seed term is taken) is salient to the target sentence

input	context	entity	output	relation	rel_sent	Is the output trivally overlap with the context	IE is of sufficient quality (not generic, correct)	context contains relevant information for target relation (Conservative filter - only flag cases where context is highly irrelevant)	Relation is a part of the main idea propose d by the paper
extractive text summarizatio n is done by using Metric	transformer - based language models usually treat texts as linear sequences . however , most texts also have an inherent hierarchical structure , i.e. , parts of a text can be identified using their position in this hierarchy . in addition , section titles usually indicate the common topic of their respective sentences .	extractive text summarization	sota rouges	used for	We propose a novel approach to formulate , extract , encode and inject hierarchical structure information explicitly into an extractive summarization model based on a pre - trained , encoder - only Transformer language model (HiStruct+ model) , which improves SOTA ROUGEs for extractive summarization on PubMed and arXiv substantially .				

SciMON Overview

Problem/Motivation:
... streaming data of various sources may continuously grow ... requires plms to *integrate the information from all the sources* in a lifelong manner... pre-training on all existing data, such a process is expensive.
Seed Term:
knowledge acquisition

Input: Background Context



Initial Idea

...continual learning for knowledge acquisition... This approach is more efficient than exhaustive pre-training on all existing data...

Retrieved Similar Ideas ↓ Retrieval from Literature Papers

1. Continual learning (CL) aims to enable information systems to learn from a continuous data stream across time...

1st Novelty Iteration ↓ Novelty Threshold Check

...a method that combines continual learning with a dynamic knowledge distillation approach for efficient knowledge acquisition ...

Retrieved Similar Ideas ↓ Retrieval from Literature Papers

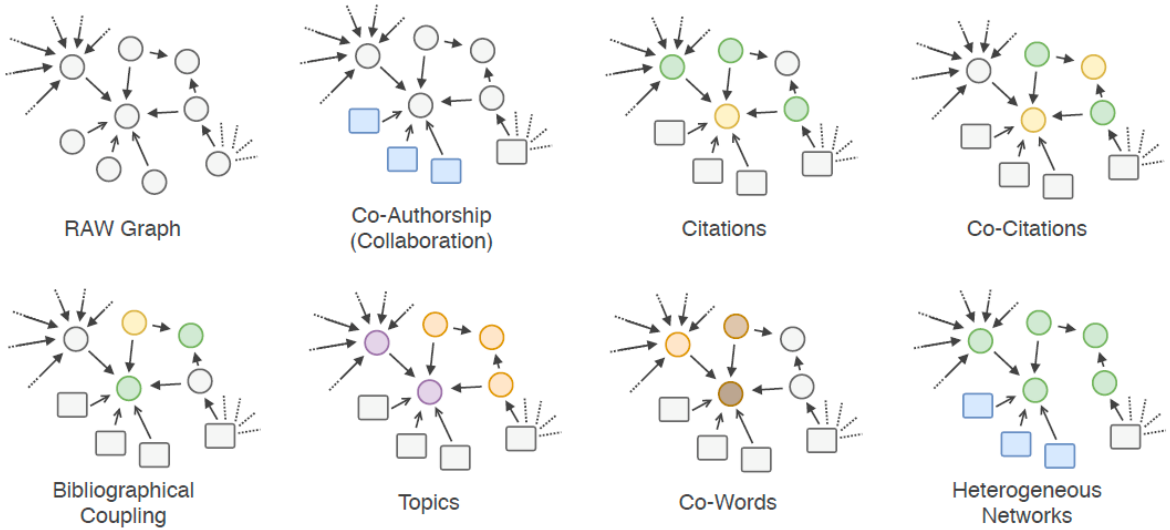
1. Different from previous knowledge distillation methods ... student model learns from teacher model for incremental knowledge extraction ...

Final Idea ↓ Novelty Threshold Check

... a method that leverages memory-augmented neural networks for knowledge acquisition in a lifelong learning scenario...

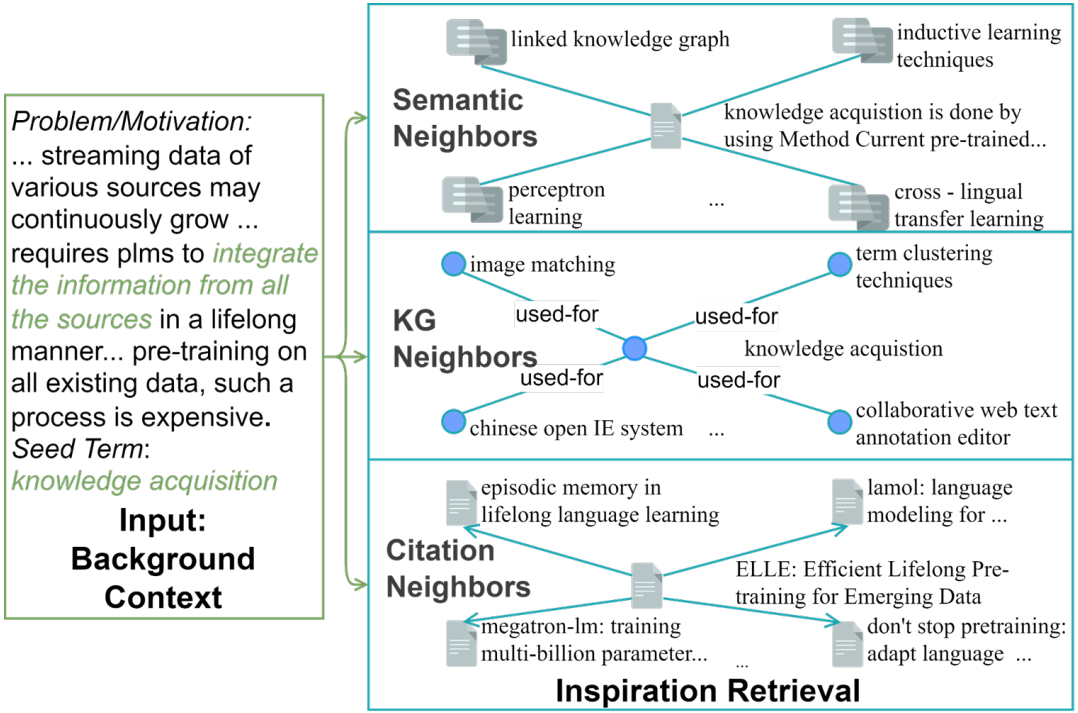
Inspiration Retrieval

- Can we leverage external knowledge graphs, such as citation information, to boost idea generation?



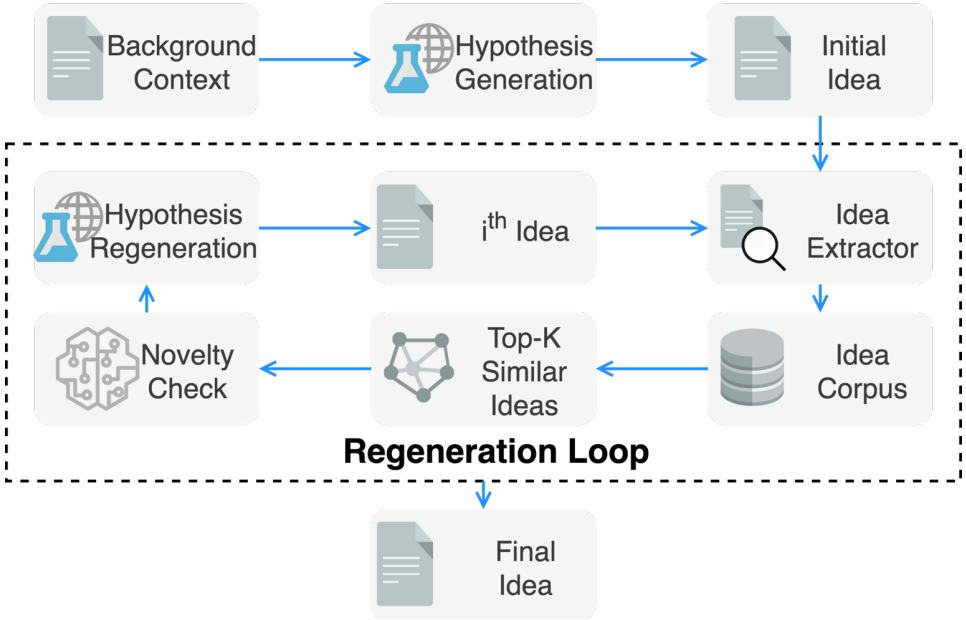
https://towardsdatascience.com/scholarly-network-analysis-22cd352c0f86

Inspiration Retrieval



- Semantic Neighbors
  - Ideas proposed for related problems in the training set
- KG Neighbors
  - Neighbors of the seed term in background KG
- Citation Neighbors
  - Cited paper title of given background context

Iterative Novelty Boosting



- We boost novelty iteratively by
  1. retrieving related work from literature reference examples
  2. measuring degree of novelty
  3. instructing the model to update idea to be more novel with respect to reference examples, conditioning on background context

Human Evaluation

- Comparing Outputs across Model Variants
  - Evaluate the generated hypothesis by considering each output's relevance to the context, novelty, clarity, and whether the idea is reasonable
  - GPT4FS and GPT4FS+KG **outperform** other models by a wide margin
  - GPT4 outputs tended to be **longer**, which may partially explain higher human preference

Human Evaluation

- Comparisons to Real Papers
  - The results are ranked according to the level of technical detail and innovation in comparison to each other and ground truth
  - 48% GPT4FS+KG shows **higher** technical detail
  - 45% GPT4FS+KG is **more novel**
  - Original ground truth ideas have **significantly higher** technical level and novelty in 85% of comparisons

Human Evaluation: Iterative Novelty Boosting

- Compare the novelty-enhanced results against the previous generated ideas
- Examine the new terms added after filtering stopwords and generic words
- Ideas after novelty iterations are **longer** than initial ideas
- For ideas not considered more novel after applying our method, we do not observe a drop in novelty: the method either increases or maintains novelty

Type	GPT4FS	+SN	+CT	+KG
1st Novelty Δ (%)	+54.4	**+55.6**	+47.8	+46.7
2nd Novelty Δ (%)	-	+57.8	-	-
1st new terms Δ	**+23.1**	+22.8	+22.1	+21.9
2nd new terms Δ	-	+21.5	-	-

Qualitative Analysis

Input	*seed term*: **speech unit boundaries ;** *context* (abridged): ... generate partial sentence translation given a streaming speech input. existing approaches ... break the acoustic units in speech, as boundaries between acoustic units in speech are not even...
Initial Idea	A pause prediction model to identify **speech unit boundaries** ...
Iteration 1	A method that leverages acoustic and linguistic features to predict **speech unit boundaries** dynamically, ensuring smooth transitions ... differs from the existing research as it combines both acoustic properties and linguistic context ... adapting to variations in speaker characteristics, speaking styles, and languages.
Iteration 2	A novel method called Adaptive **Speech Unit Boundary** Detection (ASUBD) ... a combination of attention mechanisms to focus on relevant acoustic and linguistic features and reinforcement learning to guide the system to make optimal predictions of unit boundaries based on previous decisions...
Ground Truth	... an efficient monotonic segmentation module ... accumulate acoustic information incrementally and detect proper **speech unit boundaries**.

Dong, Q., Zhu, Y., Wang, M., & Li, L. (2021). Learning when to translate for streaming speech. ACL 2022.

Case Study: Biomedical

- Collect a dataset from PubMed papers from 1988 to 2024 and construct a biomedical dataset, including 4,767 papers before 2023/02, 642 papers from 2023/02 to 2023/08, and 299 papers after 2023/08
  - PubTator 3 for named entity recognition, relation extraction, entity coreference and linking, and entity normalization

Type	Meditron	+SN	+CT	+KG
Helpful (%)	35	80	60	50
vs. GT (%)	30	45	50	35

https://www.ncbi.nlm.nih.gov/research/pubtator3/

Chen, Z., Cano, A. H., Romanou, A., Bonnet, A., Matoba, K., Salvi, F., ... & Bosselut, A. (2023). Meditron-70b: Scaling medical pretraining for large language models. arXiv preprint arXiv:2311.16079.

Case Study: Biomedical

Input	*seed term*: **ARO10;** *context* (abridged): Huangjiu is known for its unique aroma, primarily attributed to its high concentration of b-phenylethanol (ranging from 40 to 130 mg/L).
Meditron	We then selected six key genes involved in the phenylalanine metabolism pathway and predicted that their product flux may affect the synthesis of b-phenylethanol.
+ CT	We found that the key amino acid residue that controls the activity of **Aro10p** was not conserved in wine yeast strains, which may explain the lower b-phenylethanol production in wine fermentation compared with that in Chinese huangjiu.
+ KG	Both target genes, SSA1 and **ARO10**, were deleted using the CRISPR-Cas9 genome editing system.
+ SN	Herein, we report that the key barrier for b-phenylethanol production in Huangjiu is **ARO10**, the only bi-functional amino acid decarboxylase in Saccharomyces cerevisiae.

Chen, Z., Cano, A. H., Romanou, A., Bonnet, A., Matoba, K., Salvi, F., ... & Bosselut, A. (2023). Meditron-70b: Scaling medical pretraining for large language models. arXiv preprint arXiv:2311.16079.

Conclusion

- Propose the first contextualized literature discovery task in which model input background contexts (e.g., problems, experimental settings, goals), and output *natural language ideas* grounded in literature
- Propose a novel approach that uses retrieval of *inspirations* from past scientific papers, and explicitly optimizes for novelty by iteratively comparing to prior papers and updating idea suggestions until sufficient novelty is achieved
- Design extensive evaluation experiments using human annotators with domain expertise to assess relevance, utility, novelty, and technical depth

Code and Data are public at:
https://github.com/EagleW/Scientific-Inspiration-Machines-Optimized-for-Novelty

Thank you!

Code and Data are public at:
https://github.com/EagleW/Scientific-Inspiration-Machines-Optimized-for-Novelty