Validation and Analysis of Hypothesis Generation Systems

Justin Sybrandt

Talk Outline

Warning: is actually two talks

- Overview + Background
- Large-Scale Validation of Hypothesis Generation Systems via Candidate Ranking
- Are Abstracts Enough for Hypothesis Generation?

Overview

Problem Overview

- Medical research is expensive and risky
- Text mining can identify fruitful research directions before expensive experiments

WSJ

Pfizer Ends Hunt for Drugs to Treat Alzheimer’s and Parkinson's

Hypothesis Generation

- NIH provides 27 million abstracts
- 2-4 thousand added daily
- Lack of communication leads to undiscovered connections
- Hypothesis generation finds *implicitly* published relationships

Moliere

Automatic Biomedical Hypothesis Generation System

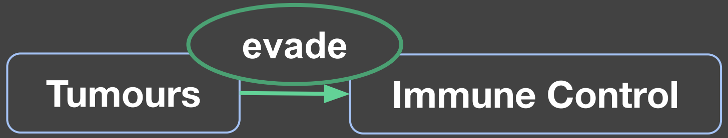
- Presented at KDD’17
- Validated against small number of historical examples
- Relied on expert input to interpret results
- Original Pipeline
    - Data Collection
    - Network Construction
    - Relevant Abstract Identification
    - Topic Modeling

Data Collection

Abstracts & n-grams

Tumours evade immune control by creating hostile microenvironments that perturb T cell metabolism and effector function.

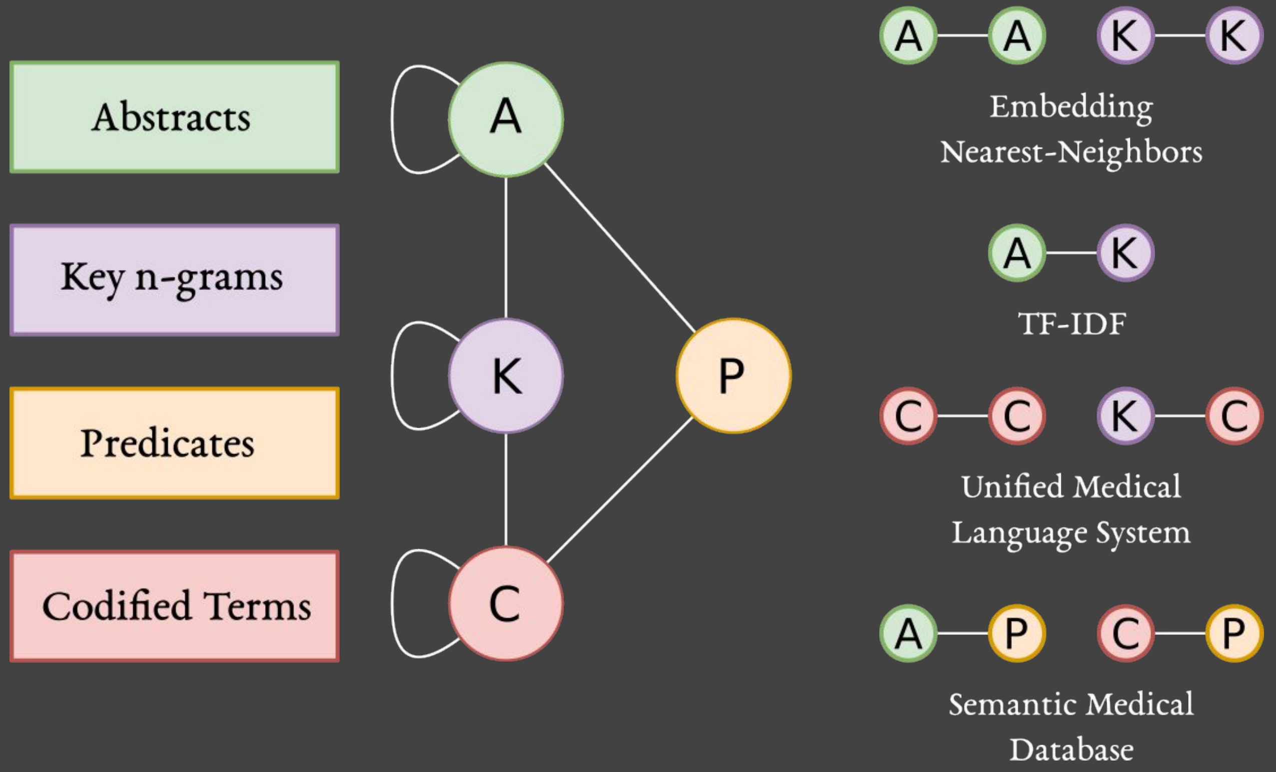
Predicates



Codified Terms

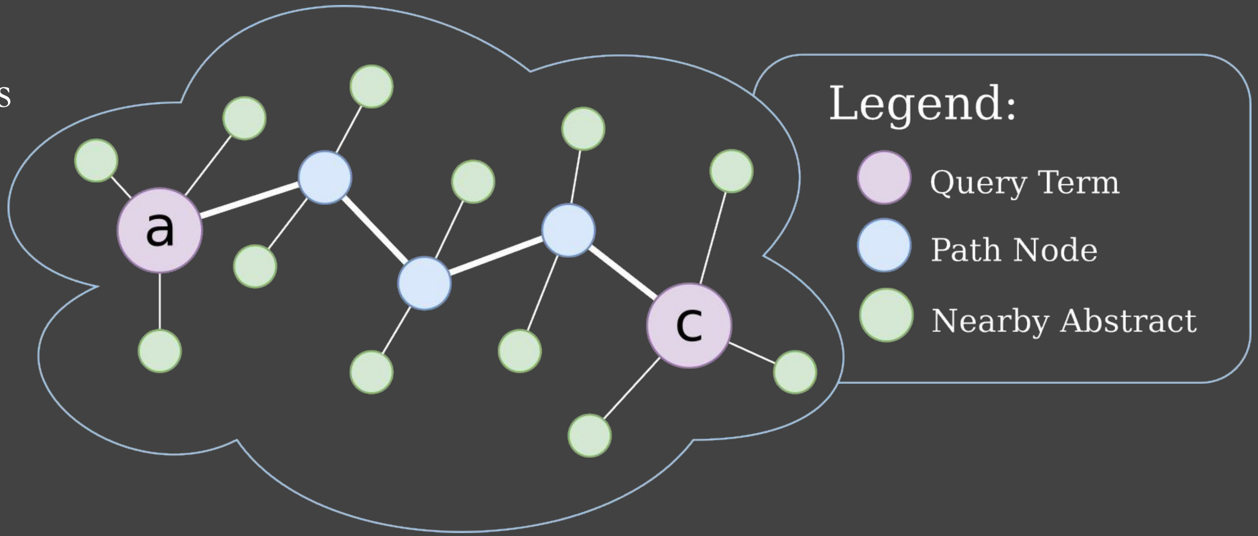
Neoplasms
Tumor - Tumour
Oncological abnormality

Network Construction



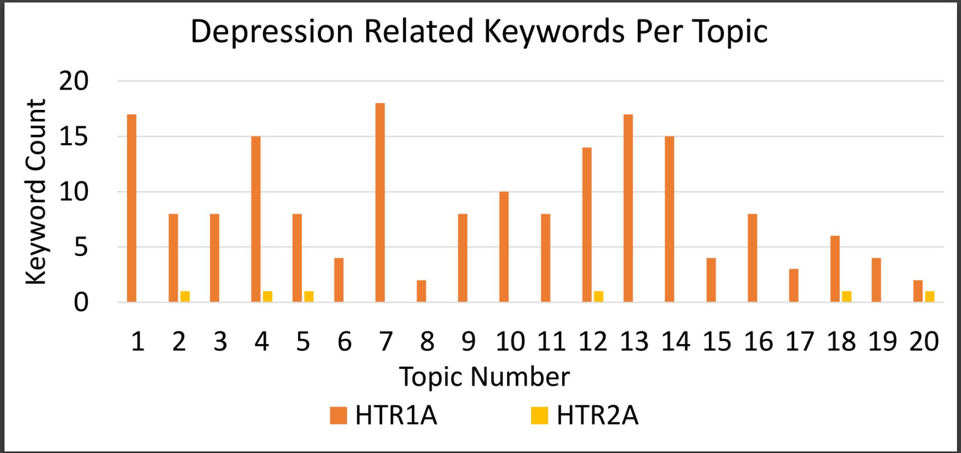
Relevant Abstract Identification

- Select two query nodes
- Find shortest path
- Locate nearby abstracts
- Collect sub-corpus



Extract Information

- Apply LDA topic modeling
- Explore patterns in fuzzy clusters
- Original limitations:
    - Expert analysis
    - No numerical results
    - Lots of data, time consuming



Large-Scale Validation of Hypothesis Generation Systems via Candidate Ranking

Justin Sybrandt[1], Michael Shtutman[2], Ilya Safro[1]

[1] Clemson U. - School of Computing
[2] U. of S. Carolina - Drug Discovery and Biomedical Sciences

Validation

Does it work?

- Challenges
  - Lack of datasets
  - Problematic false positive / negative
- We propose a scalable approach
- Verification through lab studies

Existing Validation

- Existing validation methods [1]
  - Replicate Swanson's experiments
  - Statistical evaluation
  - Incorporate expert opinion
  - Publish in medicine
- Complications
  - Human in the loop
  - Consumes expert time
  - Small validation sets

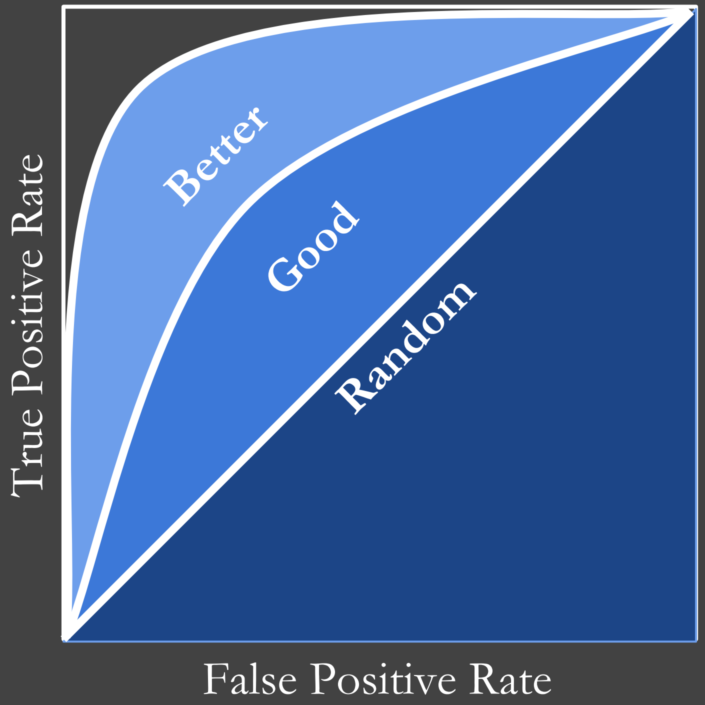
[1] M. Yetisgen-Yildiz and W. Pratt, "Evaluation of literature-based discovery systems," in Literature-based discovery. Springer, 2008, pp. 101–113.

Drug Discovery and Candidate Selection

- Drug companies must prioritize investments
- Thousands of targets narrow to handful of candidates
- Drug discovery is a ranking problem

Validation through Candidate Ranking

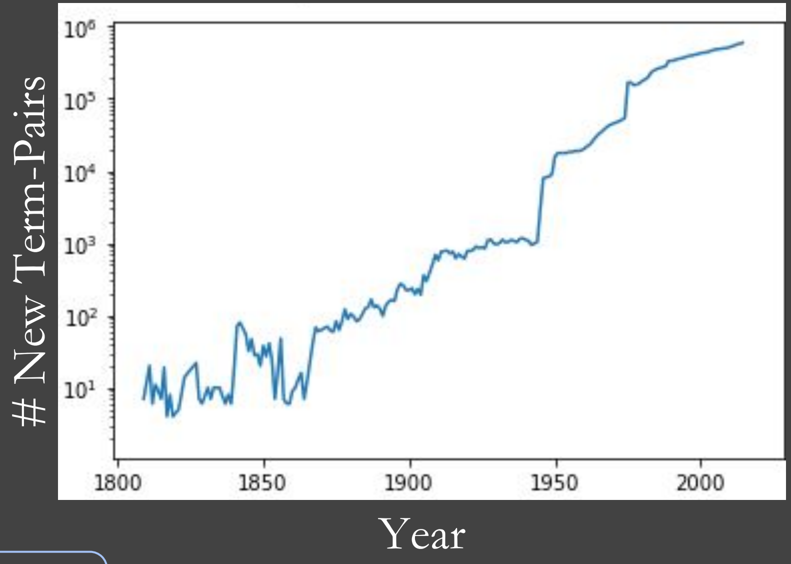
- New validation approach inspired by drug discovery
- Rank recent hypotheses by plausibility
- Requires
  - Positive & negative samples
  - Ranking criteria
- Produces area under ROC curve



Collecting Recent Hypotheses

- Assume abstracts are a reasonable summary
- Identify original term-pairs from each year
- Select cut year for validation (2010)
- Record pairs newer than cut year
- Published Set

Prevalence of New Hypotheses in Medical Literature



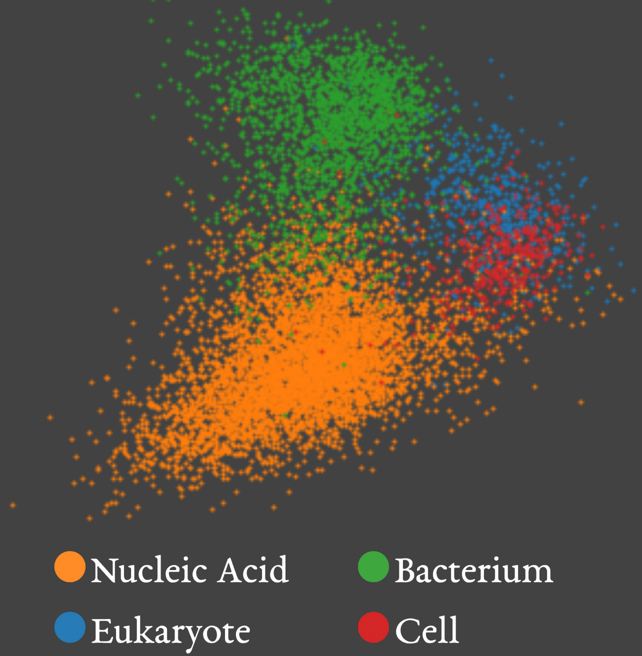
Collecting Negative Samples

- Select term subset present at cut year
- Randomly pair terms
- Record sampled pairs that do not occur in literature
- Generate samples equal to published set
- Noise Set

Creating a Ranking Criteria

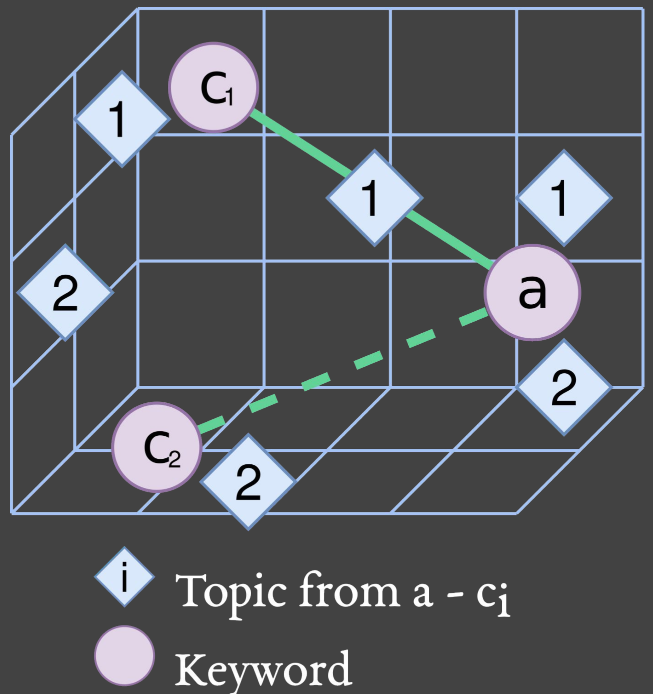
- Extract numeric features from topic model results
- Learn correlation between features and plausibility
- Generate a collection of measurements
  - Embedding based
  - Topic network based

Clusters of Words in Embedding Space



Embedding Measurements

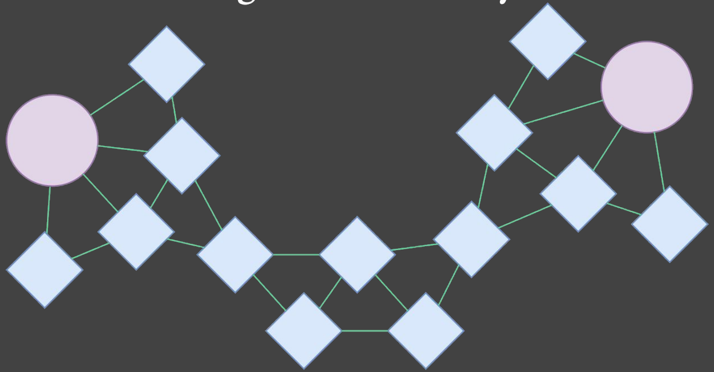
- Connected terms should...
  - Be similarly embedded
  - Share nearby topics
- Topic embeddings from centroids
- Measure L_2 distances and cosine similarity



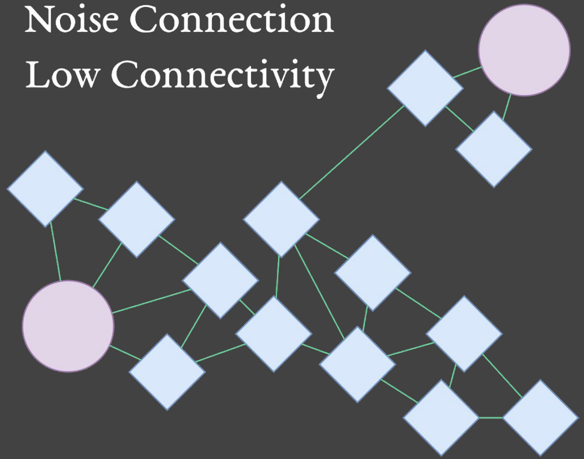
Topic Network Measurements

- Place terms and topics in network
- Edges formed by nearest-neighbors in embedding
- Add edges until path between terms appears
- Observed different network properties

Published Connection
High Connectivity



Noise Connection
Low Connectivity

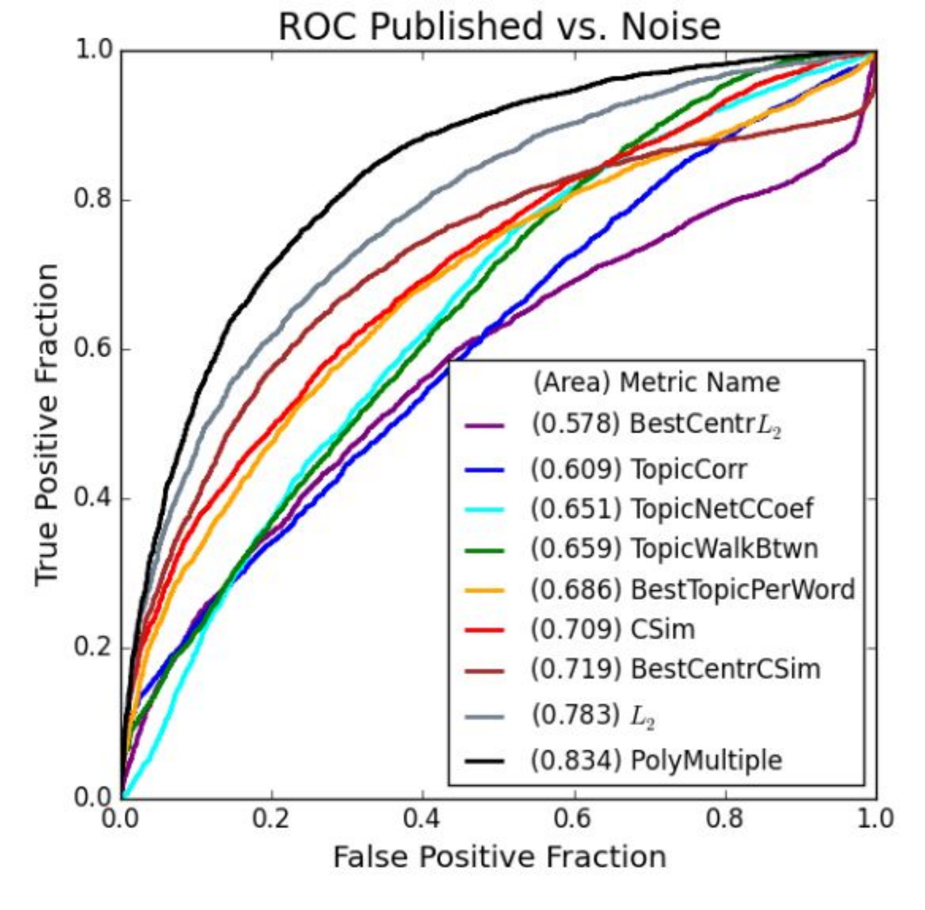


Polynomial Combination

- Each previous metric is heuristically backed
- Polynomial combination provides
  - Interpretable results
  - Improved performance
  - Easy fitting

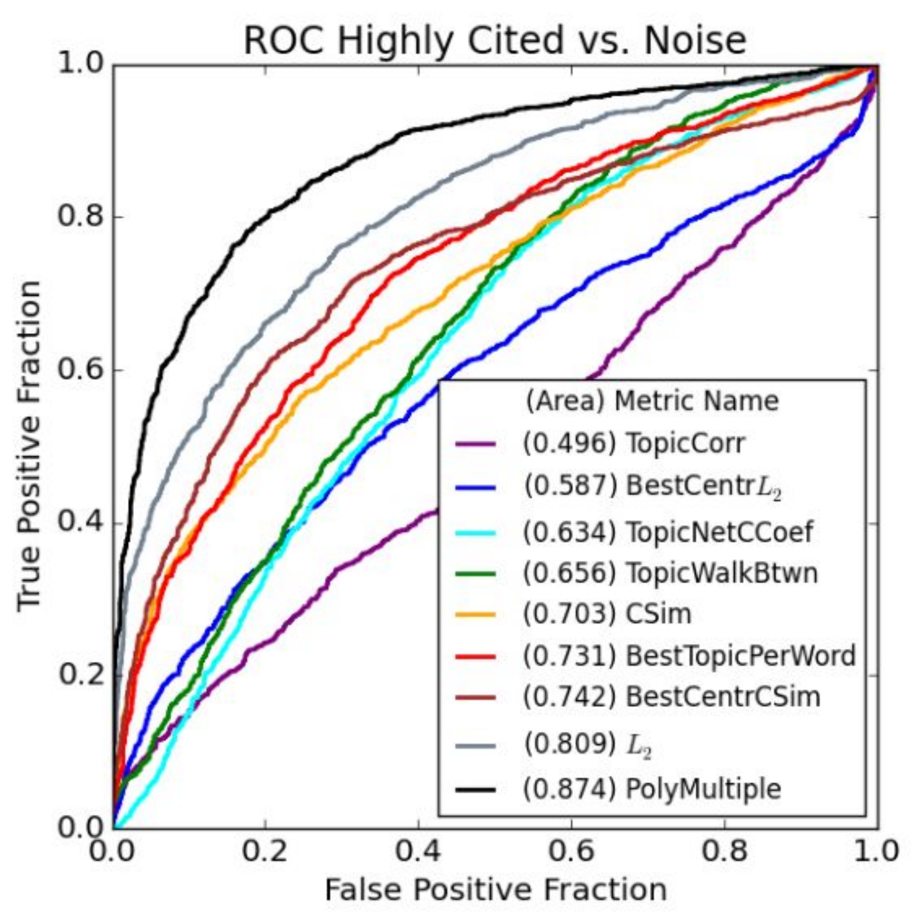
Results

- Represents 8,638 queries
- Cut year 2010
- Polynomial is top performer
- L_2 shows strength of embedding
- Topic information adds signal



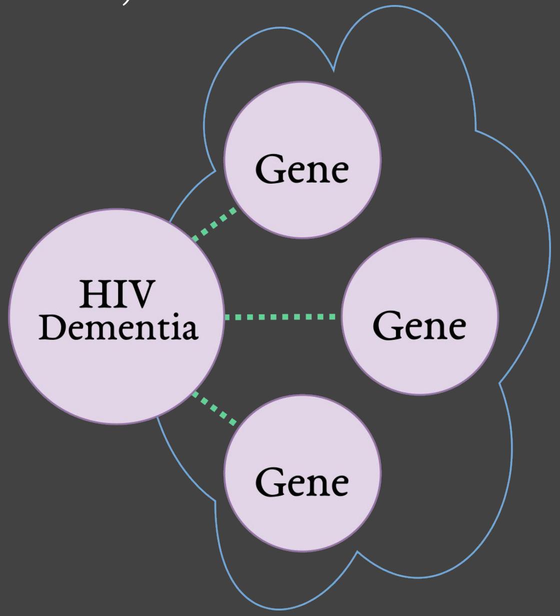
Results Highly Cited

- Represents 2,896 queries
- Subset to papers with 100 citations
- Performance improved
- Similar order of metric performance



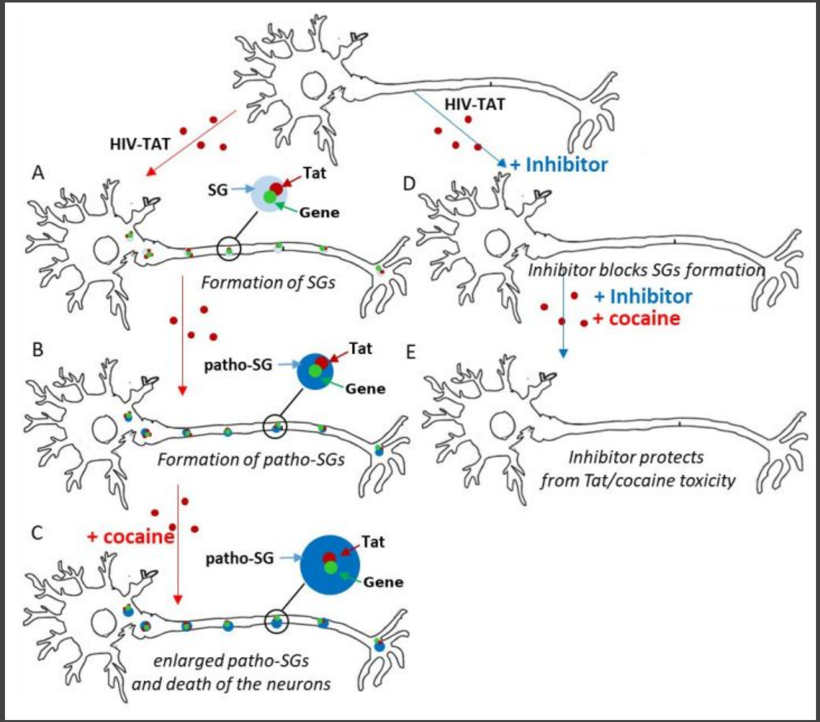
Verification in Lab Experiments

- Want to show that ranking method extends beyond validation experiment
- Focus on HIV-associated Neurodegenerative Disease (HAND)
  - ~30% of HIV patients over 60 have dementia
  - ~7% is typical rate
- Ran over 30k queries



New HAND-Gene Connection

- DDX3 identified in top 10% of genes
- Previously studied in relation to cancer
- Unexpected in this context
- Support from wet lab experiments
    - Rapidly age HIV+ neurons with cocaine
    - Cells with DDX3 inhibited survive
    - Cells with DDX3 active die



Summary : Validation

- Introduces a new validation method based on candidate ranking
    - Does not rely on expert input
    - Scales to large validation sets
- Proposed ranking metrics
    - Embedding based
    - Topic network based
- Validated our system, Moliere
    - Published vs. Noise
    - Highly Cited vs. Noise
- Applied ranking to real-world application
    - HIV associated dementia

See more online at:

sybrandt.com/2018/validation


Moliere
Validation
? ? ?

Are Abstracts Enough for Hypothesis Generation?

Justin Sybrandt, Angelo Carrabba, Alexander Herzog, Ilya Safro

Clemson U. - School of Computing

Motivation

- We now have a method to evaluate overall system performance
- Interesting questions:
  - What effect does corpus size and document length have on results?
  - How sensitive is a hypothesis generation system to input qualities?
  - How many papers does a hypothesis generation system need?
  - Are abstracts enough?

Challenges with Full Text

- Larger documents
  - ~15.6x more words
- Expensive to acquire
  - Abstracts are free
- Harder to parse
  - Figures, tables, references
  - Often must parse PDFs

Input Data from Other Systems

- Titles Only
  - ARROWSMITH - 1986
- Titles and Abstracts (+ external sources)
  - Moliere - 2017
  - Disease-Connect - 2015
  - BrainSCANr - 2010
  - ...
- Full Text
  - Watson for Drug Discovery - 2014

Input Data from Other Systems

- Titles Only
  - ARROWSMITH - 1986
- Titles and Abstracts (+ external sources)
  - Moliere - 2017
  - Disease-Connect - 2015
  - BrainSCANr - 2010
  - ...
- Full Text
  - Watson for Drug Discovery - 2014

- Proprietary system
- Designed after recommender systems [2]
- Most inference on term-document matrix [3]

[2] Spangler, Scott. *Accelerating Discovery: Mining Unstructured Information for Hypothesis Generation*. Chapman and Hall/CRC, 2015.

[3] He, Qi, Ming Ji, and W. Scott Spangler. "Mining strong relevance between heterogeneous entities from their co-occurrences." U.S. Patent Application No. 14/279,617.

Methodology

- Create datasets of variable corpus size and document size
  - Free abstracts from PubMed
  - Free full texts from PubMed Central
- Construct multiple “instances” of Moliere
  - Rebuild embedding, network, and queries
- Use previously discussed validation and ranking
  - Cut year 2015

Considered Corpora

- From PubMed
  - Entire dataset
  - Randomly sampled 1 / 2
  - Randomly sampled 1 / 4
  - Randomly sampled 1 / 8
  - Randomly sampled 1 / 16
- From PubMed Central*
  - Full Texts
  - Abstracts

* We restrict PMC to only papers released in plain text that contain abstracts.

Input Dataset Comparisons

	All of PubMed	PMC Full Text
# Documents (Millions)	**24**	**1**
Median Words Per Document	**71**	**1,594**
Unique Words (Millions)	**2.4**	**6.5**
Total Words (Billions)	**1.85**	**1.86**

Input Dataset Comparisons

	PMC Abstracts	PMC Full Text
# Documents (Millions)	**1**	**1**
Median Words Per Document	**102**	**1,594**
Unique Words (Millions)	**0.67**	**6.5**
Total Words (Billions)	**0.1**	**1.86**

Input Dataset Comparisons

	PMC Abstracts	1 / 16 PubMed
# Documents (Millions)	**1**	**1.5**
Median Words Per Document	**102**	**71**
Unique Words (Millions)	**0.67**	**0.35**
Total Words (Billions)	**0.1**	**0.1**

PMC vs. PubMed Quality Comparison

- PubMed contains some questionable "abstracts"
  - Translated
  - Incomplete records
  - Scanned from older documents
- PubMed Central
  - Much more recent
  - Authors submit their own full-text papers
  - Conform better to modern publication standards

Int J Trauma Nurs. 1999 Jan-Mar;5(1):38.

Just do it!

Feury KJ[1].

PMID: 10085830

J Fam Pract. 1999 Mar;48(3):230.

Ugly stepchildren?

Young R.

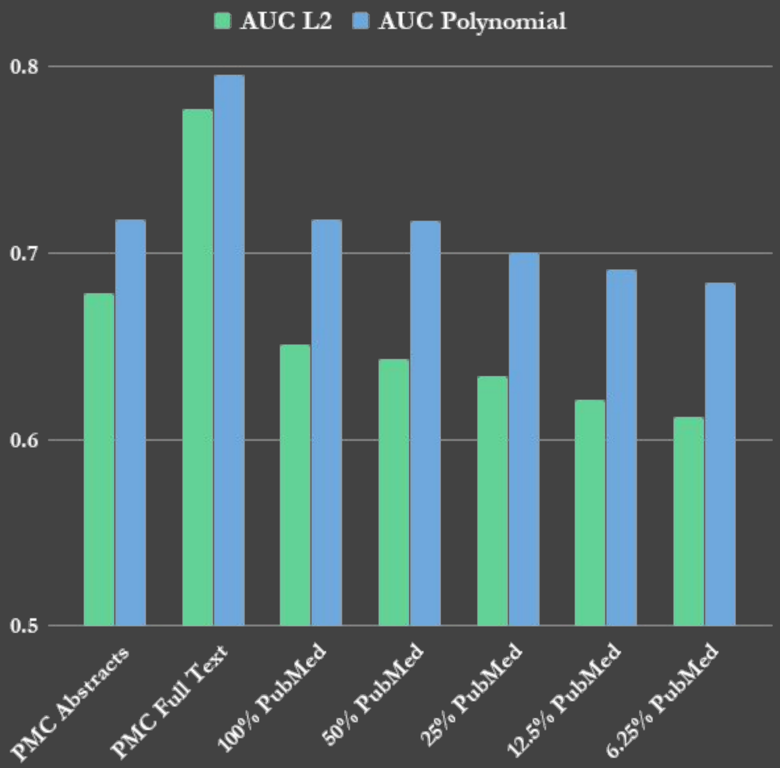
PMID: 10086771

Experiments

- Collected 2,000 validation pairs
  - Cut year 2015
  - Term pairs shared across all corpora
- Trained entire Moliere system per corpus
  - Embedding
  - Phrase Mining
  - Network Construction
  - Queries
  - Training Polynomial

Results overall

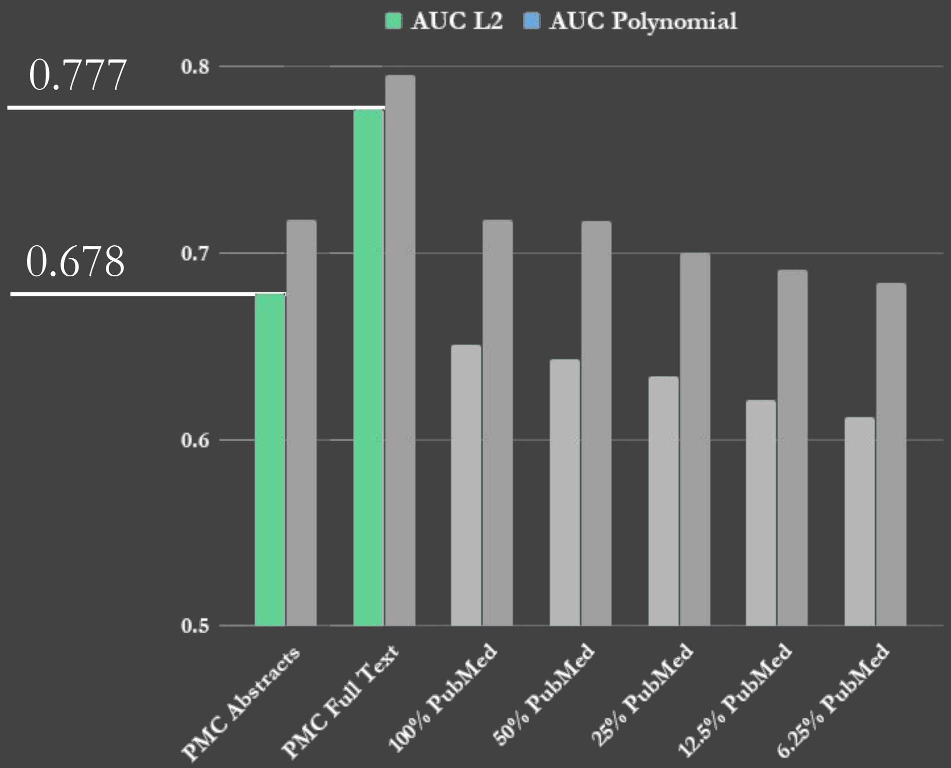
- We present full results in paper
- Focus here on L_2 and Polynomial
  - L_2 evaluates embedding quality
  - Polynomial evaluates max performance



* Lower performance than previously discussed. This work embeds text in R^{100} while the previous embeds in R^{500}.

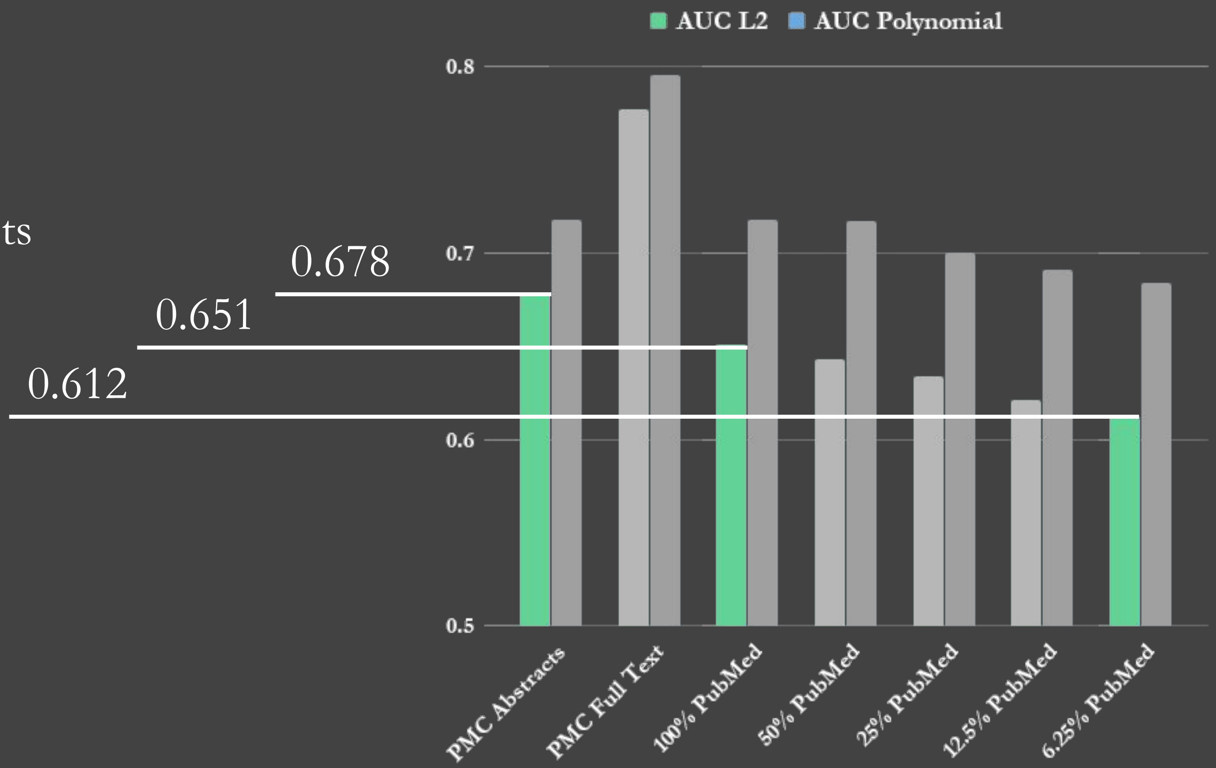
Findings

- Embedding
    - Full Text > Abstracts



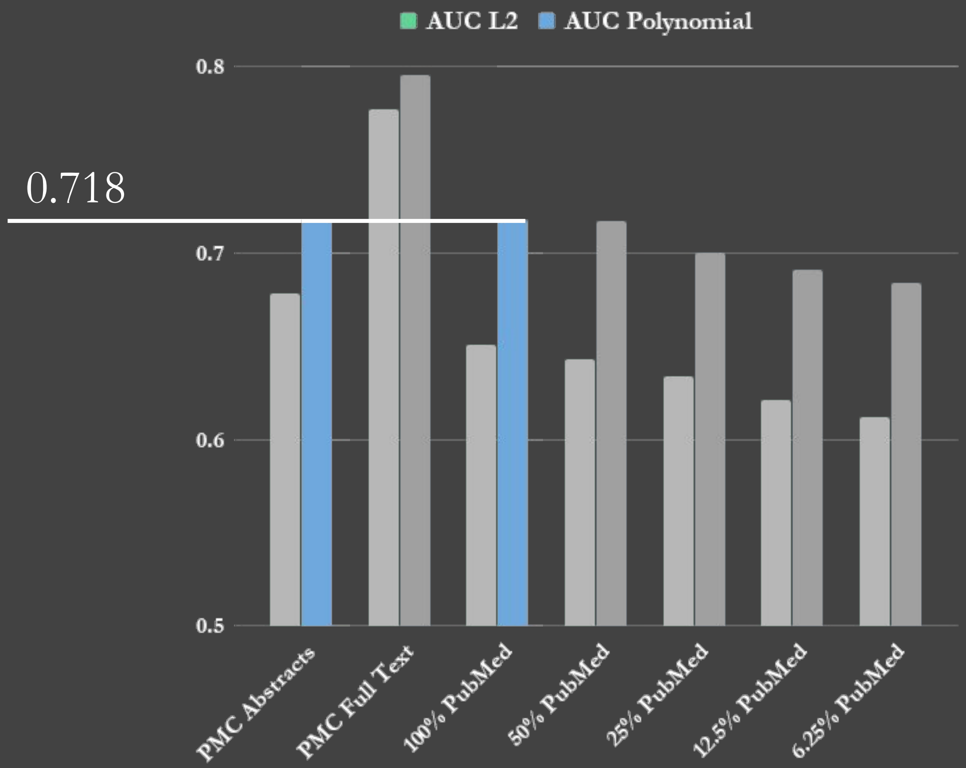
Findings

- Embedding
  - Full Text > Abstracts
  - Clean >> Many



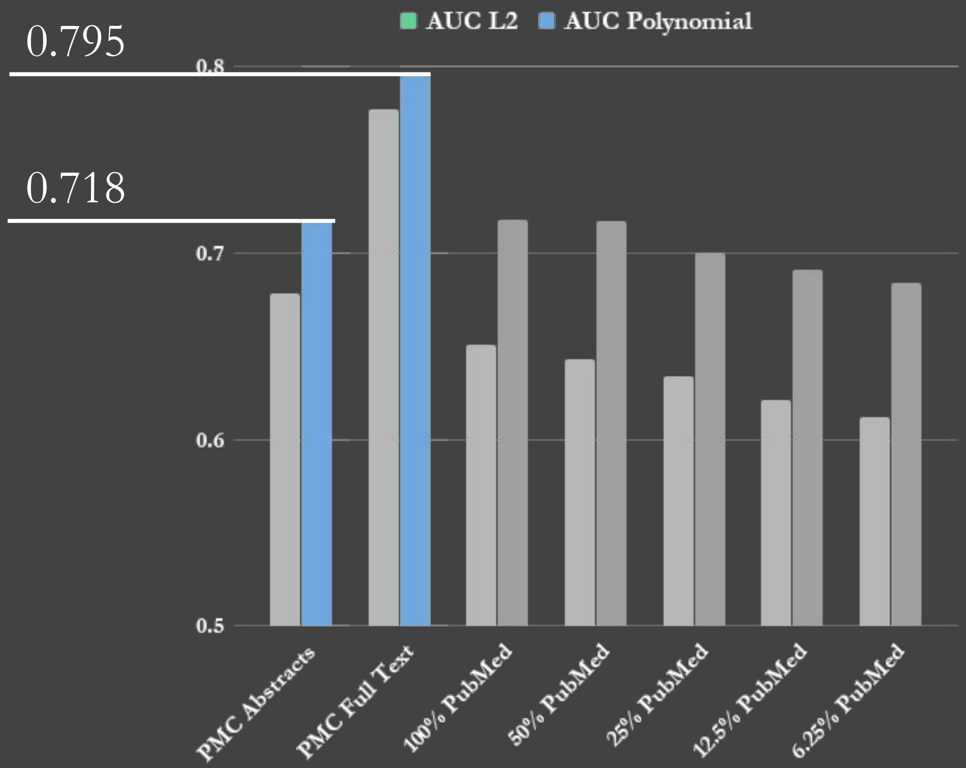
Findings

- Embedding
  - Full Text > Abstracts
  - Clean >> Many
- Max Performance
  - Clean = Many



Findings

- Embedding
    - Full Text > Abstracts
    - Clean >> Many
- Max Performance
    - Clean = Many
    - Full Text > Abstracts



The Downside to Full Text

- Increased single query runtime from 100s to ~4,500s
  - May be reasonable for specific searches
  - Does not scale to large candidate experiments
  - Most runtime during LDA topic models
- Full text topics are less interpretable

Answers

- What effect does corpus size and document length have on results?
  - Increasing either helps
  - Document length has more effect than corpus size
  - Documents that are too long negatively affect topic interpretability
- Effect
  - Removing very short documents likely to boost overall performance
  - Using automatically generated summaries may balance performance

Answers

- How sensitive is a hypothesis generation system to input qualities?
  - Significantly sensitive to short noisy documents
  - Performance predicated upon embedding
- Effect
  - Pre-trained word embeddings may boost performance

Answers

- How many papers does a hypothesis generation system need?
  - ~1 million perform well
  - Quality > Quantity
- Effect
  - Lower barrier to entry for cross-domain applications

Summary : Are Abstracts Enough?

- Explored multiple input corpora
  - PubMed vs. PubMed Central
- Found that longer documents increase performance
  - PMC abstracts are longer than Medline
  - Longer documents > larger quantity
- Significant runtime tradeoff
  - 45x runtime for 10% improvement
- Answer depends on the application

See more online at:

sybrandt.com/2018/abstracts