NLP, AKBC, and The Future of Science

Tom Hope

Assistant Professor, The Hebrew University of Jerusalem
Research Scientist, Allen Institute for AI (Semantic Scholar)

CAROLVS
II
SOCIETATIS
REGALIS
AVTHOR
ET
PATRONVS
SOCIETATIS PRAESES
ARTIVM INSTAVRATOR

THE HISTORY OF THE Royal-Society OF LONDON,

For the Improving of

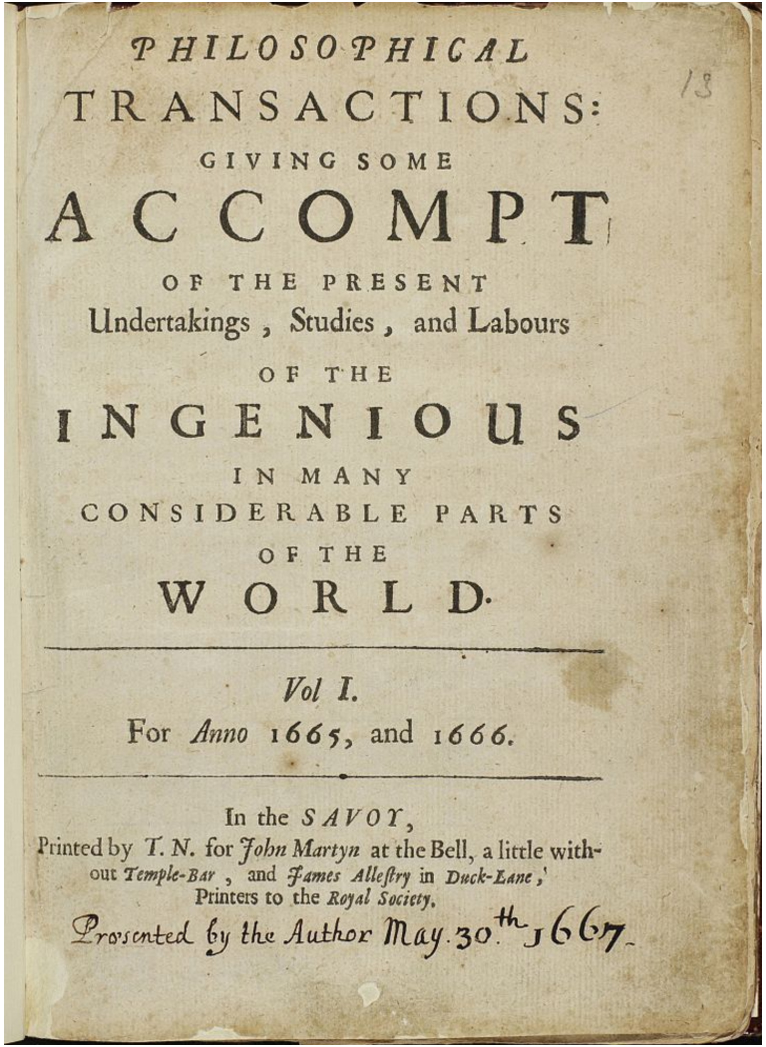
NATURAL KNOWLEDGE.

BY

THO. SPRAT.

LONDON,
Printed by *T. R.* for *J. Martyn* at the *Bell* without
Temple-bar, and *J. Alleſtry* at the *Roſe* and *Crown* in
Duck-lane, Printers to the *Royal Society*.
MDCLXVII.

1665: **first scientific journal** was published!



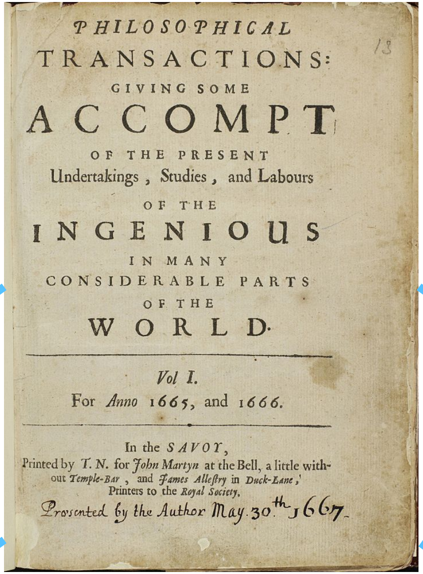
13

PHILOSOPHICAL
TRANSACTIONS:
GIVING SOME
ACCOMPT
OF THE PRESENT
Undertakings, Studies, and Labours
OF THE
INGENIOUS
IN MANY
CONSIDERABLE PARTS
OF THE
WORLD.

Vol I.
For *Anno* 1665, and 1666.

In the *SAVOY*,
Printed by *T. N.* for *John Martyn* at the Bell, a little without *Temple-Bar*, and *James Alleſtry* in *Duck-Lane*,
Printers to the *Royal Society*.

Presented by the Author May. 30th 1667.



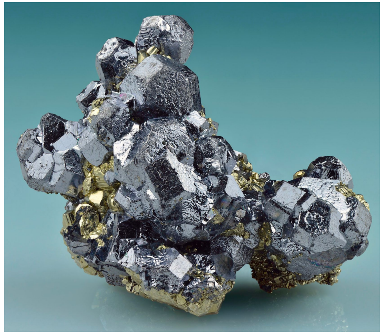
13

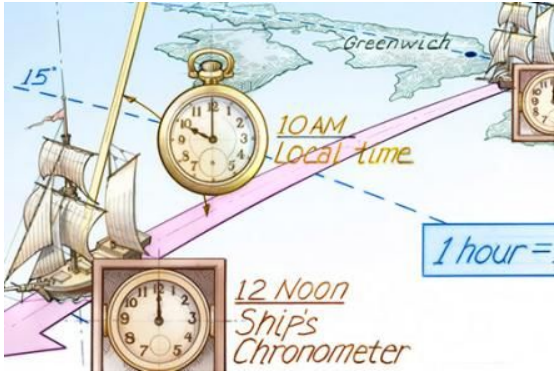
PHILOSOPHICAL
TRANSACTIONS:
GIVING SOME
ACCOMPT
OF THE PRESENT
Undertakings, Studies, and Labours
OF THE
INGENIOUS
IN MANY
CONSIDERABLE PARTS
OF THE
WORLD.

Vol I.
For *Anno* 1665, and 1666.

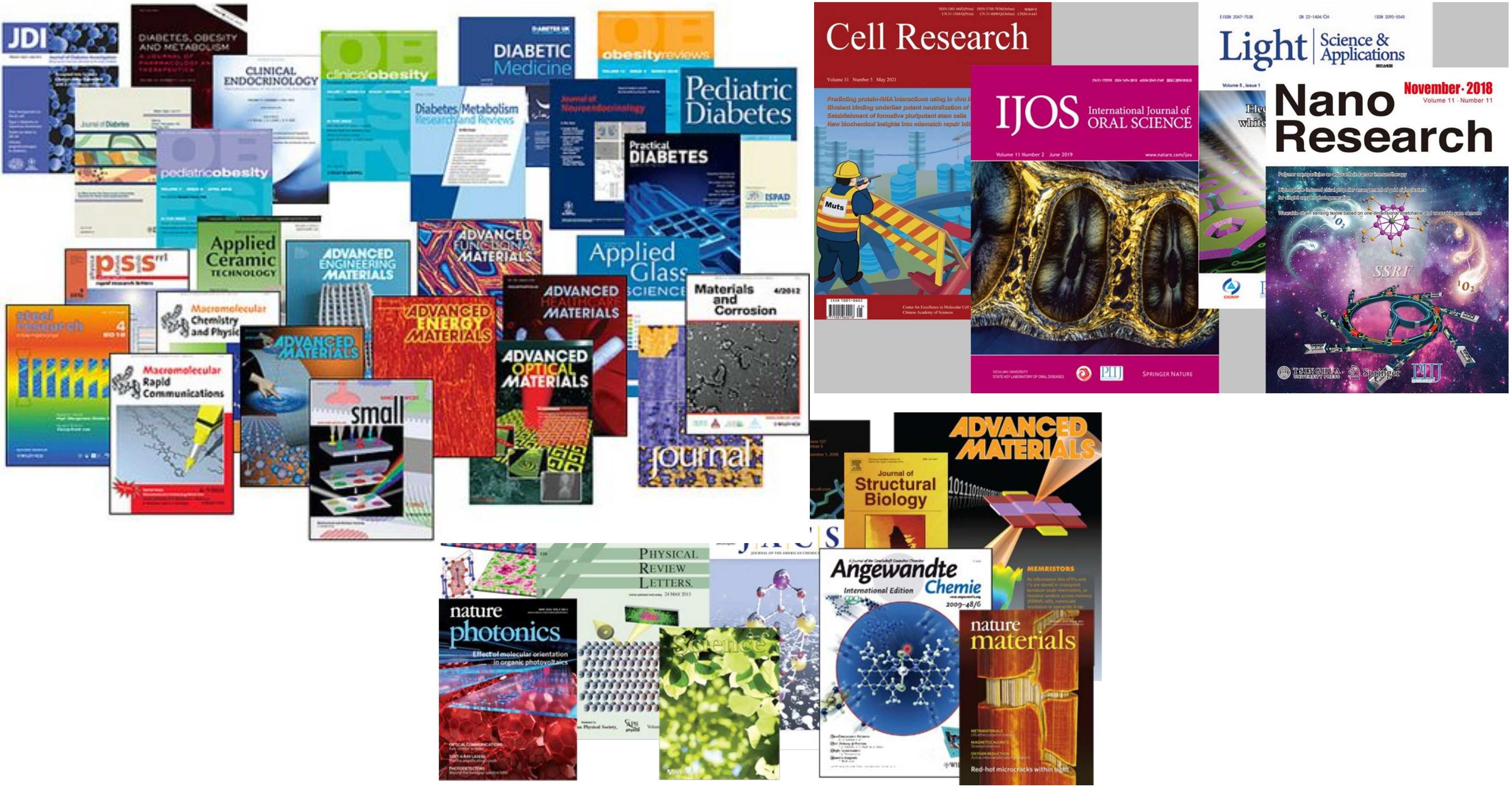
In the *SAVOY*,
Printed by *T. N.* for *John Martyn* at the Bell, a little without *Temple-Bar*, and *James Alleſtry* in *Duck-Lane*,
Printers to the *Royal Society*.

Presented by the Author May.30th 1667.



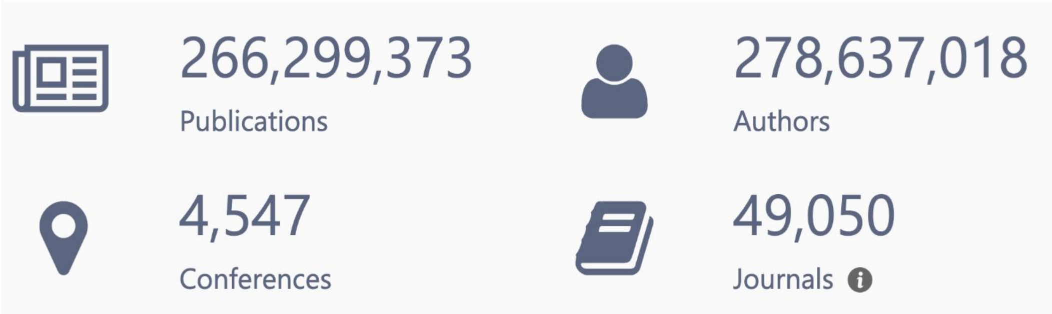


Scientific Journals, 2022…


JDI
DIABETES, OBESITY AND METABOLISM
CLINICAL ENDOCRINOLOGY
clinicalobesity
DIABETIC Medicine
obesityreviews
Journal of Diabetes
pediatricobesity
Diabetes/Metabolism Research and Reviews
Journal of Neuroendocrinology
Pediatric Diabetes
Practical DIABETES
ISPAD
pss rrl
Applied Ceramic TECHNOLOGY
ADVANCED ENGINEERING MATERIALS
ADVANCED FUNCTIONAL MATERIALS
Applied Glass SCIENCE
steel research
Macromolecular Chemistry and Physics
ADVANCED HEALTHCARE MATERIALS
Materials and Corrosion
4/2012
ADVANCED ENERGY MATERIALS
ADVANCED MATERIALS
Macromolecular Rapid Communications
small
ADVANCED OPTICAL MATERIALS
journal
Cell Research
IJOS International Journal of ORAL SCIENCE
SPRINGER NATURE
Light Science & Applications
Nano Research
November · 2018
Volume 11 · Number 11
SSRF
TSINGHUA UNIVERSITY PRESS
Springer
Journal of Structural Biology
ADVANCED MATERIALS
MEMRISTORS
PHYSICAL REVIEW LETTERS
JACS
Angewandte Chemie International Edition
2009–48/6
nature photonics
Effect of molecular orientation in organic photovoltaics
Science
nature materials
Red-hot microcracks within sight

Google Scholar

Looking for some relevant science



266,299,373
Publications

278,637,018
Authors

4,547
Conferences

49,050
Journals

Source:
Microsoft Academic

Example: Biomedical Literature

>1 million biomedical papers *every year*

Explosion of Scientific Information

Literature

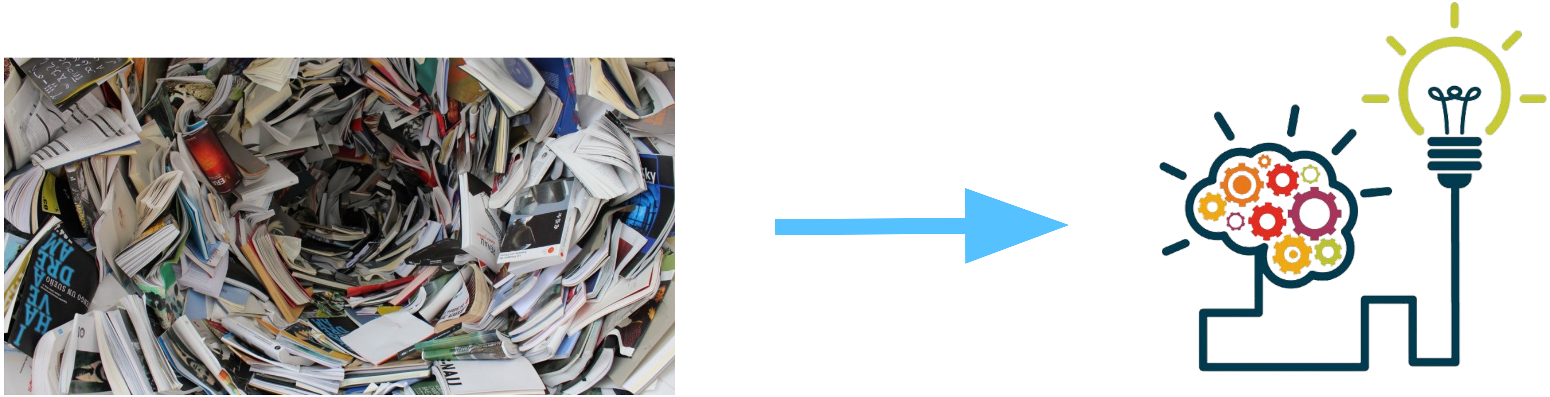
Scientific knowledge bases

Resources, code libraries

Online discussions

…

Opportunity:
Augment & Scale Scientific Discovery



AI systems that harness humanity's
collective scientific knowledge

Research Tasks

Research Tasks

Attention to areas of interest

(How do we keep track?)

Problem identification & prioritization
(How do we select what to work on?)

Attention to areas of interest
(How do we keep track?)

Research Tasks

Problem identification & prioritization
(How do we select what to work on?)

Forming directions
(How do we generate solutions?)

Research Tasks

Attention to areas of interest
(How do we keep track?)

Problem identification & prioritization
(How do we select what to work on?)

Forming directions
(How do we generate solutions?)

Attention to areas of interest
(How do we keep track?)

Experimentation, analysis

Research Tasks

Problem identification & prioritization
(How do we select what to work on?)

Forming directions
(How do we generate solutions?)

Attention to areas of interest
(How do we keep track?)

Experimentation, analysis

Research Tasks

Learning, understanding

Problem identification & prioritization
(How do we select what to work on?)

Forming directions
(How do we generate solutions?)

Attention to areas of interest
(How do we keep track?)

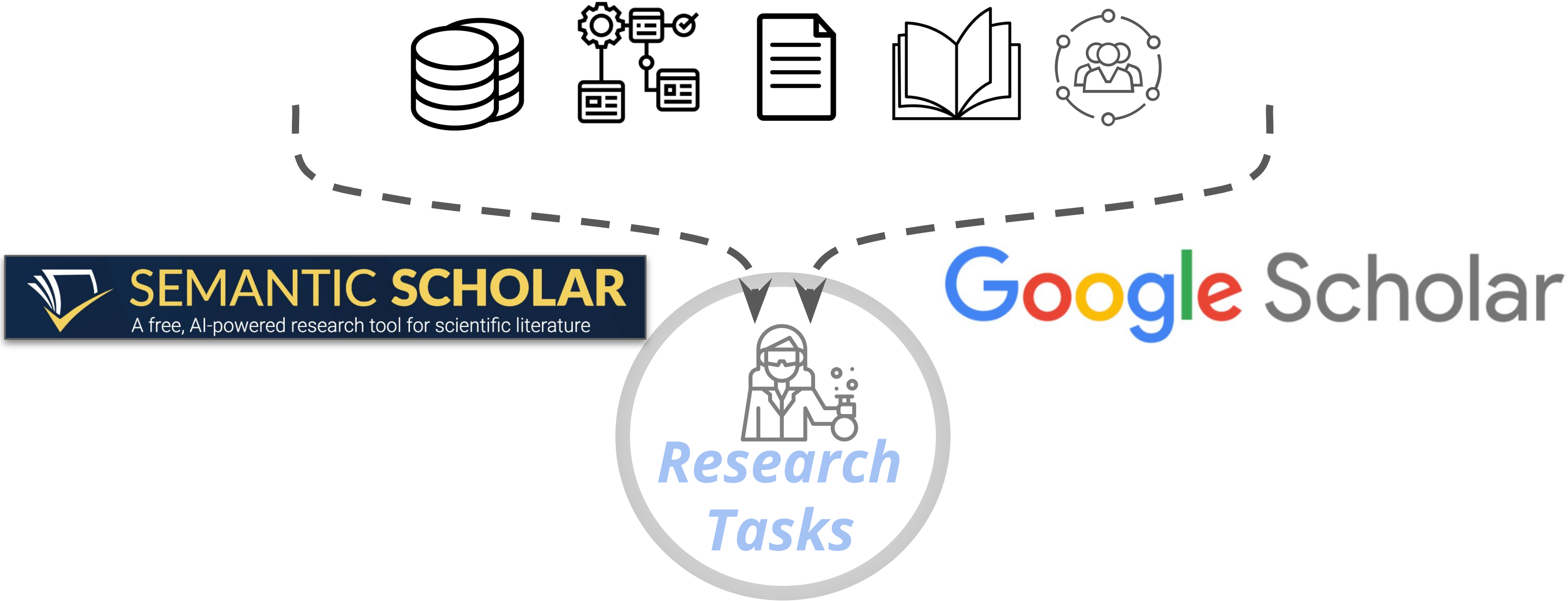
Experimentation, analysis

Research Tasks

Learning, understanding

Communication, collaboration

Scientific Knowledge



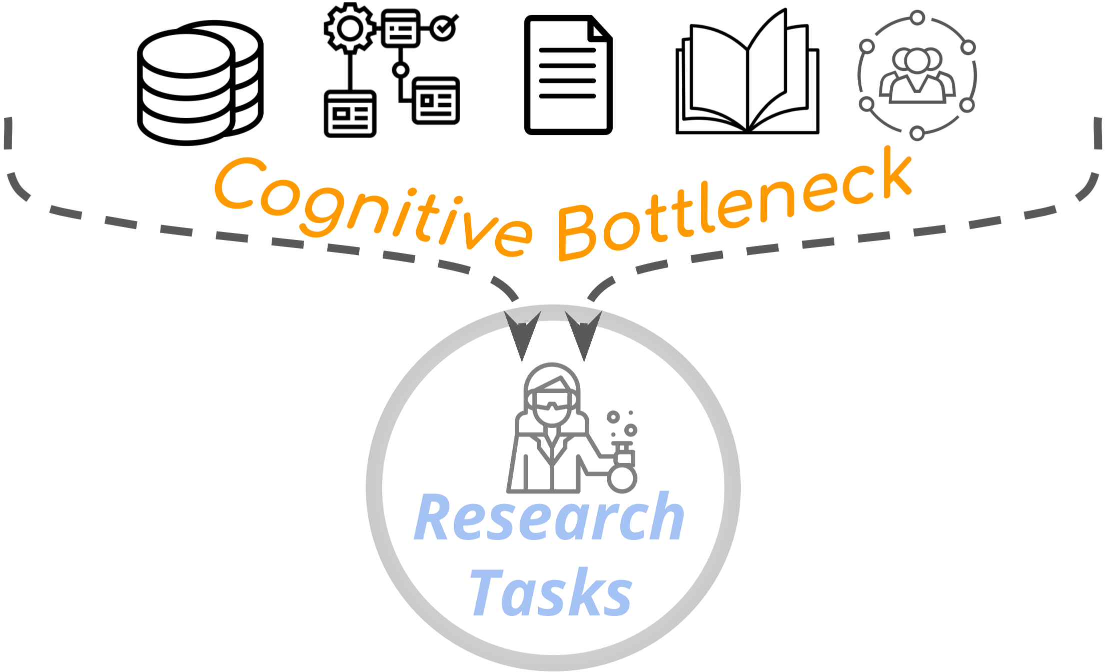
SEMANTIC SCHOLAR

A free, AI-powered research tool for scientific literature

Google Scholar

Research Tasks

Scientific Knowledge



Cognitive Bottleneck

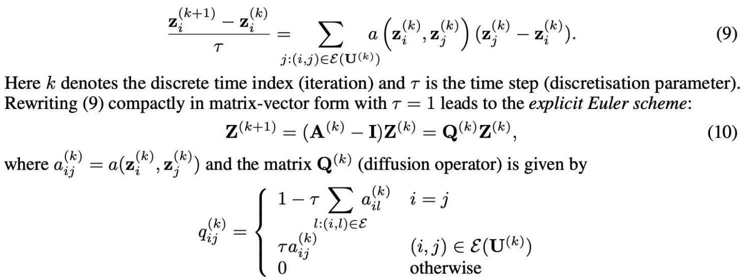
Research Tasks

Scientific Knowledge: Challenges

Large-scale, diverse

Rapidly evolving

Deeply technical



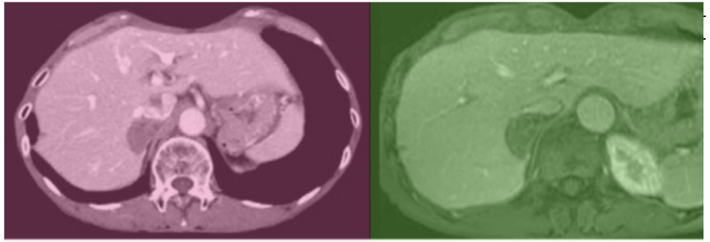
$$\frac{\mathbf{z}_i^{(k+1)} - \mathbf{z}_i^{(k)}}{\tau} = \sum_{j:(i,j)\in\mathcal{E}(\mathbf{U}^{(k)})} a\left(\mathbf{z}_i^{(k)}, \mathbf{z}_j^{(k)}\right)(\mathbf{z}_j^{(k)} - \mathbf{z}_i^{(k)}). \tag{9}$$

Here k denotes the discrete time index (iteration) and τ is the time step (discretisation parameter). Rewriting (9) compactly in matrix-vector form with $\tau = 1$ leads to the *explicit Euler scheme*:

$$\mathbf{Z}^{(k+1)} = (\mathbf{A}^{(k)} - \mathbf{I})\mathbf{Z}^{(k)} = \mathbf{Q}^{(k)}\mathbf{Z}^{(k)}, \tag{10}$$

where $a_{ij}^{(k)} = a(\mathbf{z}_i^{(k)}, \mathbf{z}_j^{(k)})$ and the matrix $\mathbf{Q}^{(k)}$ (diffusion operator) is given by

$$q_{ij}^{(k)} = \begin{cases} 1 - \tau \sum_{l:(i,l)\in\mathcal{E}} a_{il}^{(k)} & i = j \\ \tau a_{ij}^{(k)} & (i,j) \in \mathcal{E}(\mathbf{U}^{(k)}) \\ 0 & \text{otherwise} \end{cases}$$



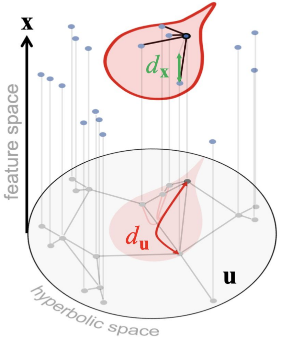
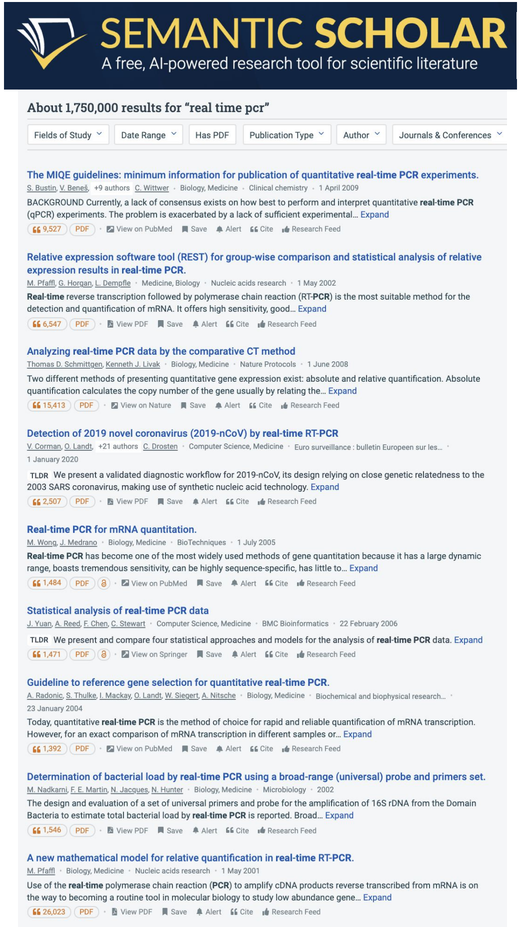


Figure 2: Graph Beltrami flow with

Limited Modes of Interaction



SEMANTIC SCHOLAR

A free, AI-powered research tool for scientific literature

About 1,750,000 results for "real time pcr"

Fields of Study ⌄ | Date Range ⌄ | Has PDF | Publication Type ⌄ | Author ⌄ | Journals & Conferences ⌄

The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments.
S. Bustin, V. Beneš, +9 authors C. Wittwer · Biology, Medicine · Clinical chemistry · 1 April 2009
BACKGROUND Currently, a lack of consensus exists on how best to perform and interpret quantitative **real-time PCR** (qPCR) experiments. The problem is exacerbated by a lack of sufficient experimental… Expand
9,527 PDF · View on PubMed Save Alert Cite Research Feed

Relative expression software tool (REST) for group-wise comparison and statistical analysis of relative expression results in real-time PCR.
M. Pfaffl, G. Horgan, L. Dempfle · Medicine, Biology · Nucleic acids research · 1 May 2002
Real-time reverse transcription followed by polymerase chain reaction (RT-**PCR**) is the most suitable method for the detection and quantification of mRNA. It offers high sensitivity, good… Expand
6,547 PDF · View PDF Save Alert Cite Research Feed

Analyzing real-time PCR data by the comparative CT method
Thomas D. Schmittgen, Kenneth J. Livak · Biology, Medicine · Nature Protocols · 1 June 2008
Two different methods of presenting quantitative gene expression exist: absolute and relative quantification. Absolute quantification calculates the copy number of the gene usually by relating the… Expand
15,413 PDF · View on Nature Save Alert Cite Research Feed

Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR
V. Corman, O. Landt, +21 authors C. Drosten · Computer Science, Medicine · Euro surveillance : bulletin Europeen sur les… · 1 January 2020
TLDR We present a validated diagnostic workflow for 2019-nCoV, its design relying on close genetic relatedness to the 2003 SARS coronavirus, making use of synthetic nucleic acid technology. Expand
2,507 PDF · View PDF Save Alert Cite Research Feed

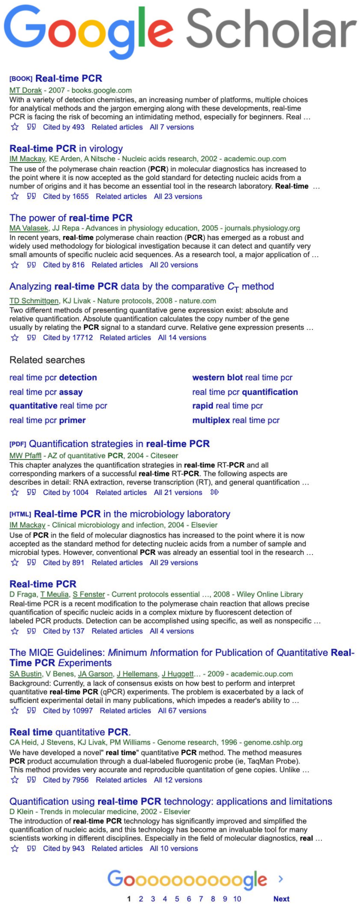
Real-time PCR for mRNA quantitation.
M. Wong, J. Medrano · Biology, Medicine · BioTechniques · 1 July 2005
Real-time PCR has become one of the most widely used methods of gene quantitation because it has a large dynamic range, boasts tremendous sensitivity, can be highly sequence-specific, has little to… Expand
1,484 PDF · View on PubMed Save Alert Cite Research Feed

Statistical analysis of real-time PCR data
J. Yuan, A. Reed, F. Chen, C. Stewart · Computer Science, Medicine · BMC Bioinformatics · 22 February 2006
TLDR We present and compare four statistical approaches and models for the analysis of **real-time PCR** data. Expand
1,471 PDF · View on Springer Save Alert Cite Research Feed

Guideline to reference gene selection for quantitative real-time PCR.
A. Radonic, S. Thulke, I. Mackay, O. Landt, W. Siegert, A. Nitsche · Biology, Medicine · Biochemical and biophysical research… · 23 January 2004
Today, quantitative **real-time PCR** is the method of choice for rapid and reliable quantification of mRNA transcription. However, for an exact comparison of mRNA transcription in different samples or… Expand
1,392 PDF · View on PubMed Save Alert Cite Research Feed

Determination of bacterial load by real-time PCR using a broad-range (universal) probe and primers set.
M. Nadkarni, F. E. Martin, N. Jacques, N. Hunter · Biology, Medicine · Microbiology · 2002
The design and evaluation of a set of universal primers and probe for the amplification of 16S rDNA from the Domain Bacteria to estimate total bacterial load by **real-time PCR** is reported. Broad… Expand
1,546 PDF · View PDF Save Alert Cite Research Feed

A new mathematical model for relative quantification in real-time RT-PCR.
M. Pfaffl · Biology, Medicine · Nucleic acids research · 1 May 2001
Use of the **real-time** polymerase chain reaction (**PCR**) to amplify cDNA products reverse transcribed from mRNA is on the way to becoming a routine tool in molecular biology to study low abundance gene… Expand
26,023 PDF · View PDF Save Alert Cite Research Feed



Google Scholar

[BOOK] **Real-time PCR**
MT Dorak - 2007 - books.google.com
With a variety of detection chemistries, an increasing number of platforms, multiple choices for analytical methods and the jargon emerging along with these developments, real-time PCR is facing the risk of becoming an intimidating method, especially for beginners. Real …
☆ 99 Cited by 493 Related articles All 7 versions

Real-time PCR in virology
IM Mackay, KE Arden, A Nitsche - Nucleic acids research, 2002 - academic.oup.com
The use of the polymerase chain reaction (**PCR**) in molecular diagnostics has increased to the point where it is now accepted as the gold standard for detecting nucleic acids from a number of origins and it has become an essential tool in the research laboratory. **Real-time** …
☆ 99 Cited by 1655 Related articles All 23 versions

The power of **real-time PCR**
MA Valasek, JJ Repa - Advances in physiology education, 2005 - journals.physiology.org
In recent years, **real-time** polymerase chain reaction (**PCR**) has emerged as a robust and widely used methodology for biological investigation because it can detect and quantify very small amounts of specific nucleic acid sequences. As a research tool, a major application of …
☆ 99 Cited by 816 Related articles All 20 versions

Analyzing **real-time PCR** data by the comparative C_T method
TD Schmittgen, KJ Livak - Nature protocols, 2008 - nature.com
Two different methods of presenting quantitative gene expression exist: absolute and relative quantification. Absolute quantification calculates the copy number of the gene usually by relating the **PCR** signal to a standard curve. Relative gene expression presents …
☆ 99 Cited by 17712 Related articles All 14 versions

Related searches

real time pcr **detection**	**western blot** real time pcr
real time pcr **assay**	real time pcr **quantification**
quantitative real time pcr	**rapid** real time pcr
real time pcr **primer**	**multiplex** real time pcr

[PDF] Quantification strategies in **real-time PCR**
MW Pfaffl - AZ of quantitative **PCR**, 2004 - Citeseer
This chapter analyzes the quantification strategies in **real-time** RT-**PCR** and all corresponding markers of a successful **real-time** RT-**PCR**. The following aspects are describes in detail: RNA extraction, reverse transcription (RT), and general quantification …
☆ 99 Cited by 1004 Related articles All 21 versions »

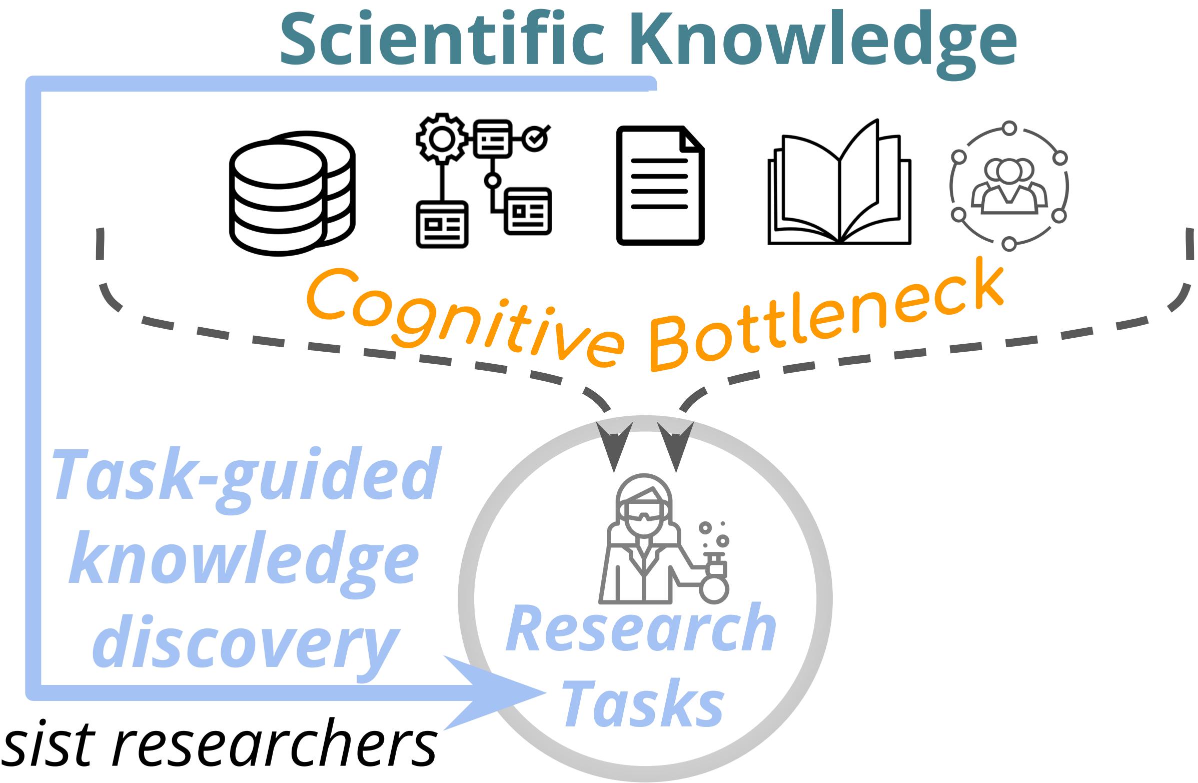
[HTML] **Real-time PCR** in the microbiology laboratory
IM Mackay - Clinical microbiology and infection, 2004 - Elsevier
Use of **PCR** in the field of molecular diagnostics has increased to the point where it is now accepted as the standard method for detecting nucleic acids from a number of sample and microbial types. However, conventional **PCR** was already an essential tool in the research …
☆ 99 Cited by 891 Related articles All 29 versions

Real-time PCR
D Fraga, T Meulia, S Fenster - Current protocols essential …, 2008 - Wiley Online Library
Real-time PCR is a recent modification to the polymerase chain reaction that allows precise quantification of specific nucleic acids in a complex mixture by fluorescent detection of labeled PCR products. Detection can be accomplished using specific, as well as nonspecific …
☆ 99 Cited by 137 Related articles All 4 versions

The MIQE Guidelines: *M*inimum *I*nformation for Publication of Quantitative **Real-Time PCR** *E*xperiments
SA Bustin, V Benes, JA Garson, J Hellemans, J Huggett… - 2009 - academic.oup.com
Background: Currently, a lack of consensus exists on how best to perform and interpret quantitative **real-time PCR** (qPCR) experiments. The problem is exacerbated by a lack of sufficient experimental detail in many publications, which impedes a reader's ability to …
☆ 99 Cited by 10997 Related articles All 67 versions

Real time quantitative **PCR**.
CA Heid, J Stevens, KJ Livak, PM Williams - Genome research, 1996 - genome.cshlp.org
We have developed a novel" **real time**" quantitative **PCR** method. The method measures **PCR** product accumulation through a dual-labeled fluorogenic probe (ie, TaqMan Probe). This method provides very accurate and reproducible quantitation of gene copies. Unlike …
☆ 99 Cited by 7956 Related articles All 12 versions

Quantification using **real-time PCR** technology: applications and limitations
D Klein - Trends in molecular medicine, 2002 - Elsevier
The introduction of **real-time PCR** technology has significantly improved and simplified the quantification of nucleic acids, and this technology has become an invaluable tool for many scientists working in different disciplines. Especially in the field of molecular diagnostics, **real** …
☆ 99 Cited by 943 Related articles All 10 versions



Scientific Knowledge

Cognitive Bottleneck

Task-guided knowledge discovery

Research Tasks

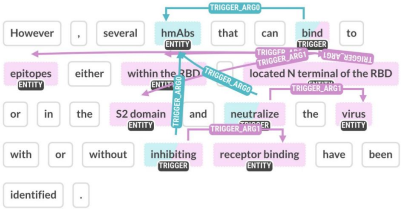
Assist researchers ***throughout their tasks***

NLP for Science

NLP for Science

NAACL 2021, 2022

AAAI-22



However , several hmAbs that can bind to epitopes either within the RBD , located N terminal of the RBD or in the S2 domain and neutralize the virus with or without inhibiting receptor binding have been identified .

Extraction

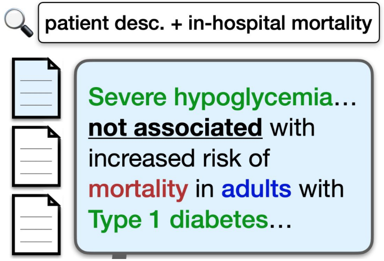
Organizing the world's scientific knowledge

NLP for Science

Extraction

NAACL 2022 CHI22 New Orleans, LA KDD

Literature Retrieval



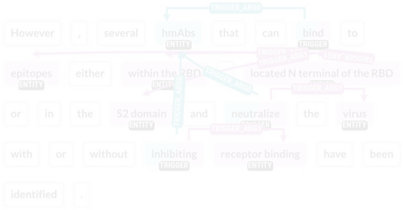
patient desc. + in-hospital mortality

Severe hypoglycemia… **not associated** with increased risk of mortality in adults with Type 1 diabetes…

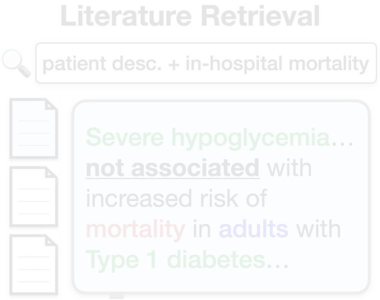
Retrieval

Finding information to boost research, decision-making

NLP for Science



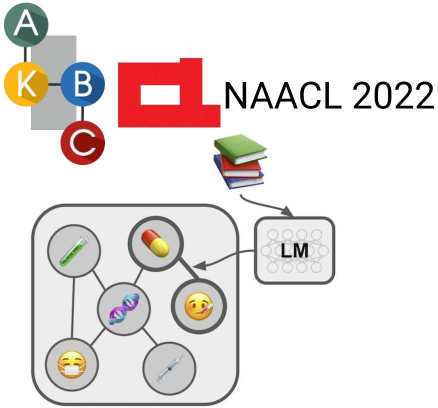
Extraction


Literature Retrieval

patient desc. + in-hospital mortality

Severe hypoglycemia… **not associated** with increased risk of mortality in adults with Type 1 diabetes…

Retrieval



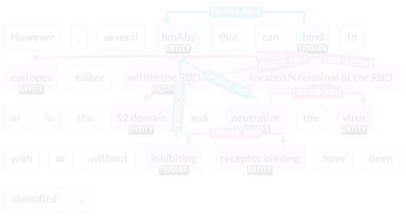
Inference

Making predictions, generating hypotheses

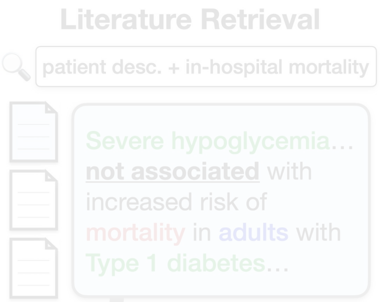
NLP for Science

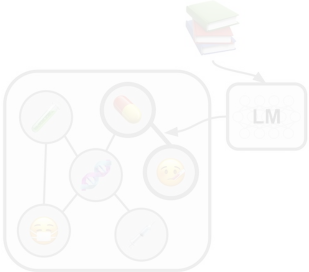
CHI22
New Orleans, LA

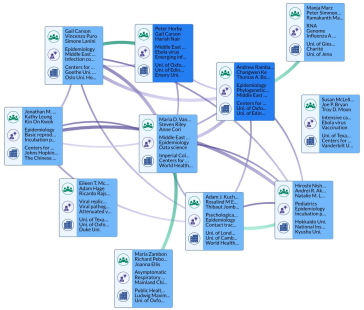
EMNLP 2020, 2022



However , several hmAbs that can bind to
ENTITY TRIGGER
epitopes either within the RBD located N terminal of the RBD
ENTITY ENTITY
or in the S2 domain and neutralize the virus
ENTITY ENTITY
with or without inhibiting receptor binding have been
TRIGGER ENTITY
identified .







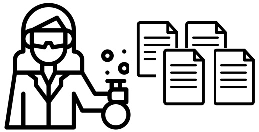
Extraction

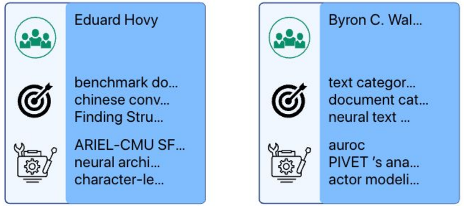
Retrieval

Inference

Interaction

NLP-powered exploratory interfaces for science





Eduard Hovy

benchmark do...
chinese conv...
Finding Stru...

ARIEL-CMU SF...
neural archi...
character-le...

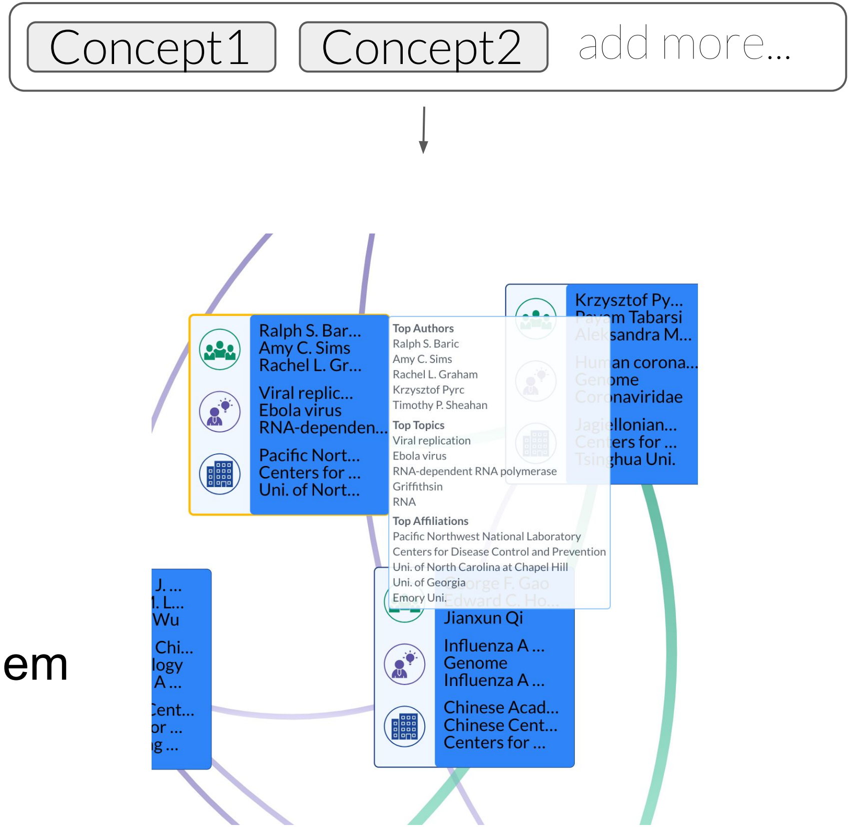
Byron C. Wal...

text categor...
document cat...
neural text ...

auroc
PIVET 's ana...
actor modeli...

Input: Researcher's papers - - - -▶ **Retrieved Knowledge:** Researchers who inspire **novel directions**

Input:
Items of interest
(*concepts, keywords...)*



Top Authors
Ralph S. Baric
Amy C. Sims
Rachel L. Graham
Krzysztof Pyrc
Timothy P. Sheahan

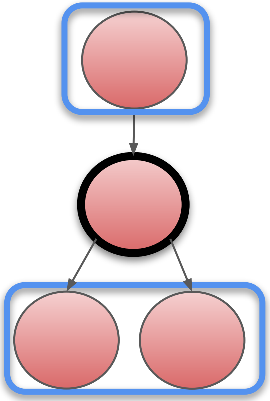
Top Topics
Viral replication
Ebola virus
RNA-dependent RNA polymerase
Griffithsin
RNA

Top Affiliations
Pacific Northwest National Laboratory
Centers for Disease Control and Prevention
Uni. of North Carolina at Chapel Hill
Uni. of Georgia
Emory Uni.

Retrieved Knowledge:
Groups of researchers
and the **links** between them

Input: *Problem description*

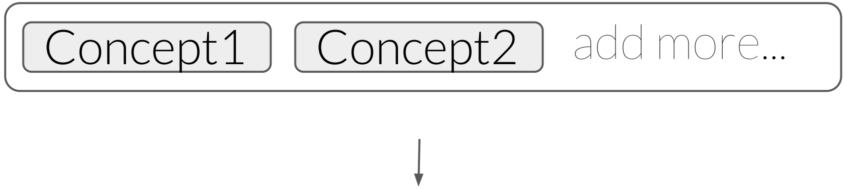
Retrieved Knowledge: Related problems for **forming new perspectives**



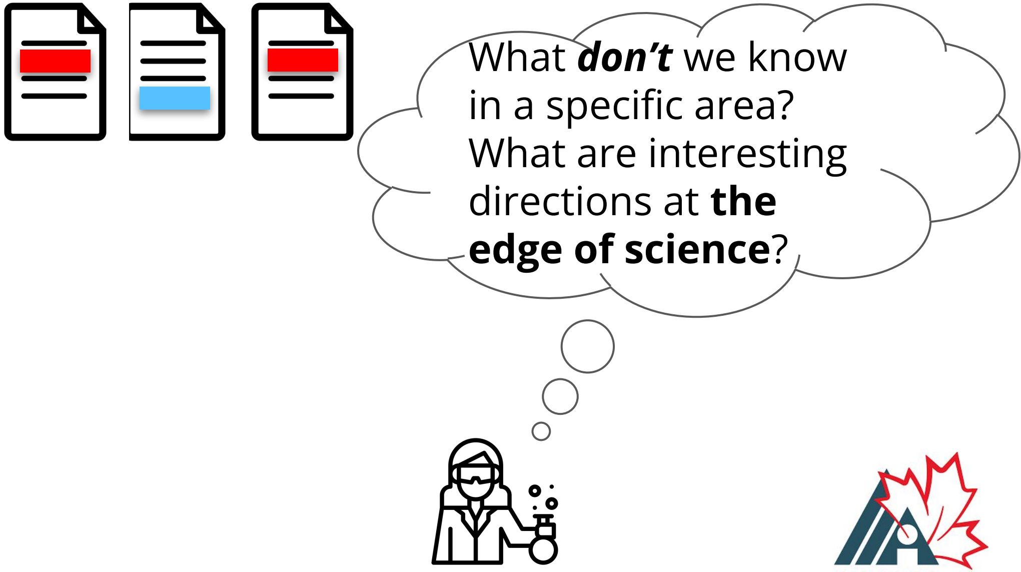
Best Research Paper

KDD CHI22 New Orleans, LA

Input:
Items of interest
(*concepts, keywords...*)

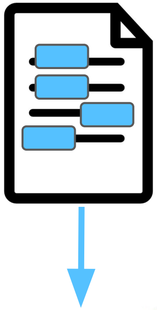


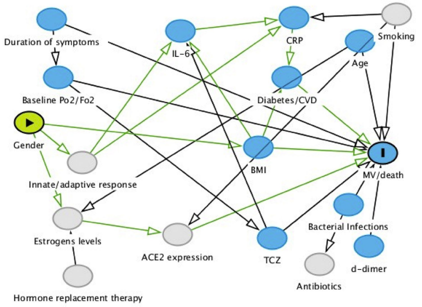
Retrieved Knowledge:
Specific mentions of
challenges, gaps in knowledge,
directions and hypotheses



What ***don't*** we know in a specific area? What are interesting directions at **the edge of science**?

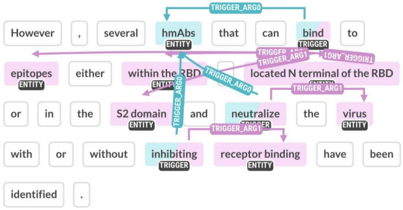
Extracting KBs + search interfaces for discovering **causal mechanisms** and **drug combination treatments**





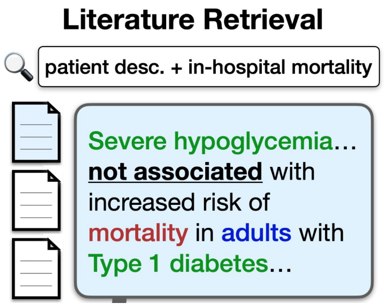
NAACL 2021, 2022

In Today's Talk:



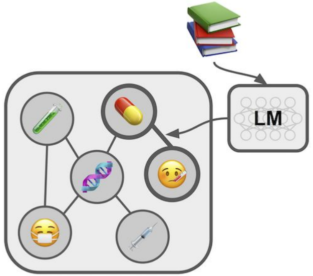
However , several hmAbs that can bind to
ENTITY
TRIGGER
TRIGGER_ARG0
TRIGGER_ARG1
epitopes either within the RBD located N terminal of the RBD
or in the S2 domain and neutralize the virus
with or without inhibiting receptor binding have been
identified .

Extraction

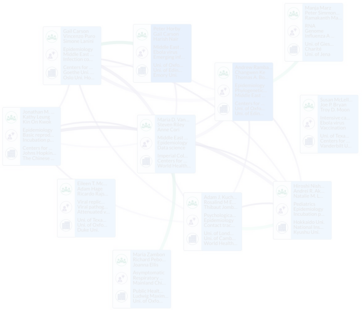


Literature Retrieval
patient desc. + in-hospital mortality
Severe hypoglycemia… not associated with increased risk of mortality in adults with Type 1 diabetes…

Retrieval



Inference



Interaction

In Today’s Talk:

1. Hierarchical Cross-Document Coreference
2. Document Similarity & Retrieval
3. Literature-Augmented Predictions

SciCo: Hierarchical Cross-Document Coreference for Scientific Concepts

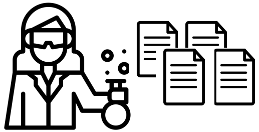
Outstanding Paper Award

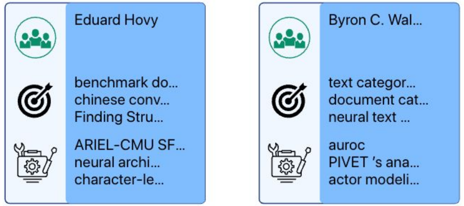
AUTOMATED KNOWLEDGE BASE 2021 | irvine CONSTRUCTION

Arie Cattan, Sophie Johnson, Daniel Weld, Ido Dagan, Iz Beltagy, Doug Downey & **Tom Hope**

AI2

Motivation: Author Matching





Eduard Hovy

benchmark do...
chinese conv...
Finding Stru...

ARIEL-CMU SF...
neural archi...
character-le...

Byron C. Wal...

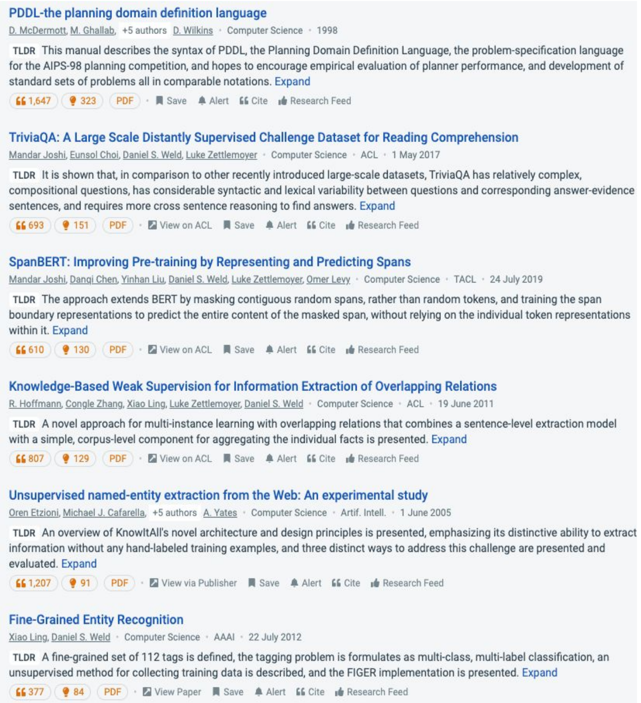
text categor...
document cat...
neural text ...

auroc
PIVET 's ana...
actor modeli...

Input: Researcher's papers ----▶ **Retrieved Knowledge:** Researchers who inspire **novel directions**

Authors who work on related tasks or use similar methods

Dan Weld



PDDL-the planning domain definition language
D. McDermott, M. Ghallab, +5 authors D. Wilkins · Computer Science · 1998
TLDR This manual describes the syntax of PDDL, the Planning Domain Definition Language, the problem-specification language for the AIPS-98 planning competition, and hopes to encourage empirical evaluation of planner performance, and development of standard sets of problems all in comparable notations. Expand
1,647 323 PDF · Save Alert Cite Research Feed

TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension
Mandar Joshi, Eunsol Choi, Daniel S. Weld, Luke Zettlemoyer · Computer Science · ACL · 1 May 2017
TLDR It is shown that, in comparison to other recently introduced large-scale datasets, TriviaQA has relatively complex, compositional questions, has considerable syntactic and lexical variability between questions and corresponding answer-evidence sentences, and requires more cross sentence reasoning to find answers. Expand
693 151 PDF · View on ACL Save Alert Cite Research Feed

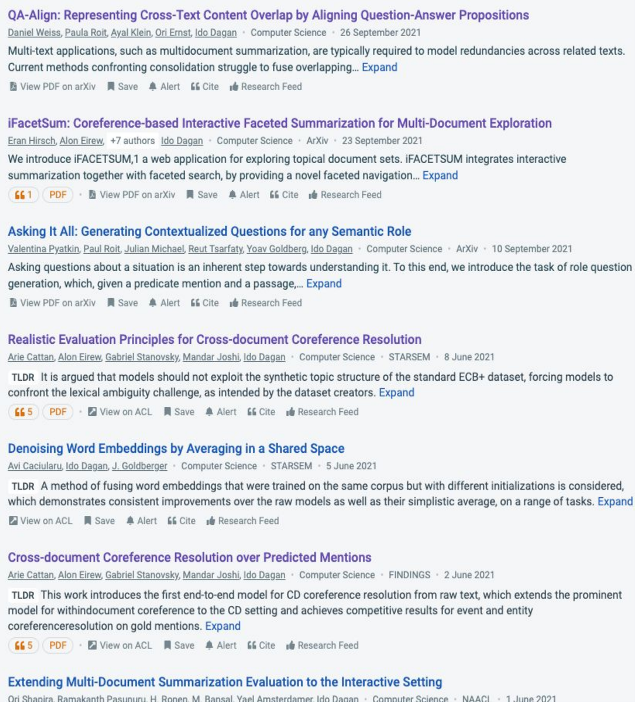
SpanBERT: Improving Pre-training by Representing and Predicting Spans
Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, Omer Levy · Computer Science · TACL · 24 July 2019
TLDR The approach extends BERT by masking contiguous random spans, rather than random tokens, and training the span boundary representations to predict the entire content of the masked span, without relying on the individual token representations within it. Expand
610 130 PDF · View on ACL Save Alert Cite Research Feed

Knowledge-Based Weak Supervision for Information Extraction of Overlapping Relations
R. Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, Daniel S. Weld · Computer Science · ACL · 19 June 2011
TLDR A novel approach for multi-instance learning with overlapping relations that combines a sentence-level extraction model with a simple, corpus-level component for aggregating the individual facts is presented. Expand
807 129 PDF · View on ACL Save Alert Cite Research Feed

Unsupervised named-entity extraction from the Web: An experimental study
Oren Etzioni, Michael J. Cafarella, +5 authors A. Yates · Computer Science · Artif. Intell. · 1 June 2005
TLDR An overview of KnowItAll's novel architecture and design principles is presented, emphasizing its distinctive ability to extract information without any hand-labeled training examples, and three distinct ways to address this challenge are presented and evaluated. Expand
1,207 91 PDF · View via Publisher Save Alert Cite Research Feed

Fine-Grained Entity Recognition
Xiao Ling, Daniel S. Weld · Computer Science · AAAI · 22 July 2012
TLDR A fine-grained set of 112 tags is defined, the tagging problem is formulates as multi-class, multi-label classification, an unsupervised method for collecting training data is described, and the FIGER implementation is presented. Expand
377 84 PDF · View Paper Save Alert Cite Research Feed

Ido Dagan



QA-Align: Representing Cross-Text Content Overlap by Aligning Question-Answer Propositions
Daniel Weiss, Paula Roit, Ayal Klein, Ori Ernst, Ido Dagan · Computer Science · 26 September 2021
Multi-text applications, such as multidocument summarization, are typically required to model redundancies across related texts. Current methods confronting consolidation struggle to fuse overlapping… Expand
View PDF on arXiv Save Alert Cite Research Feed

iFacetSum: Coreference-based Interactive Faceted Summarization for Multi-Document Exploration
Eran Hirsch, Alon Eirew, +7 authors Ido Dagan · Computer Science · ArXiv · 23 September 2021
We introduce iFACETSUM,1 a web application for exploring topical document sets. iFACETSUM integrates interactive summarization together with faceted search, by providing a novel faceted navigation… Expand
1 PDF · View PDF on arXiv Save Alert Cite Research Feed

Asking It All: Generating Contextualized Questions for any Semantic Role
Valentina Pyatkin, Paul Roit, Julian Michael, Reut Tsarfaty, Yoav Goldberg, Ido Dagan · Computer Science · ArXiv · 10 September 2021
Asking questions about a situation is an inherent step towards understanding it. To this end, we introduce the task of role question generation, which, given a predicate mention and a passage,… Expand
View PDF on arXiv Save Alert Cite Research Feed

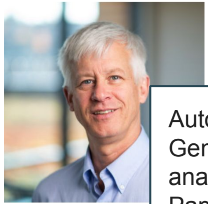
Realistic Evaluation Principles for Cross-document Coreference Resolution
Arie Cattan, Alon Eirew, Gabriel Stanovsky, Mandar Joshi, Ido Dagan · Computer Science · STARSEM · 8 June 2021
TLDR It is argued that models should not exploit the synthetic topic structure of the standard ECB+ dataset, forcing models to confront the lexical ambiguity challenge, as intended by the dataset creators. Expand
5 PDF · View on ACL Save Alert Cite Research Feed

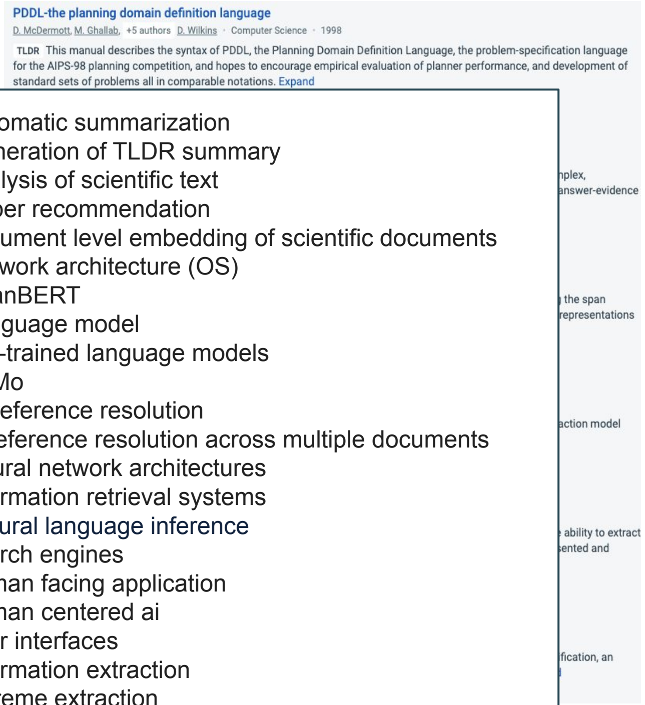
Denoising Word Embeddings by Averaging in a Shared Space
Avi Caciularu, Ido Dagan, J. Goldberger · Computer Science · STARSEM · 5 June 2021
TLDR A method of fusing word embeddings that were trained on the same corpus but with different initializations is considered, which demonstrates consistent improvements over the raw models as well as their simplistic average, on a range of tasks. Expand
View on ACL Save Alert Cite Research Feed

Cross-document Coreference Resolution over Predicted Mentions
Arie Cattan, Alon Eirew, Gabriel Stanovsky, Mandar Joshi, Ido Dagan · Computer Science · FINDINGS · 2 June 2021
TLDR This work introduces the first end-to-end model for CD coreference resolution from raw text, which extends the prominent model for withindocument coreference to the CD setting and achieves competitive results for event and entity coreferenceresolution on gold mentions. Expand
5 PDF · View on ACL Save Alert Cite Research Feed

Extending Multi-Document Summarization Evaluation to the Interactive Setting
Ori Shapira, Ramakanth Pasunuru, H. Ronen, M. Bansal, Yael Amsterdamer, Ido Dagan · Computer Science · NAACL · 1 June 2021

Authors who work on related tasks or use similar methods





PDDL-the planning domain definition language

D. McDermott, M. Ghallab, +5 authors D. Wilkins · Computer Science · 1998

TLDR This manual describes the syntax of PDDL, the Planning Domain Definition Language, the problem-specification language for the AIPS-98 planning competition, and hopes to encourage empirical evaluation of planner performance, and development of standard sets of problems all in comparable notations. Expand

nplex, answer-evidence

the span representations

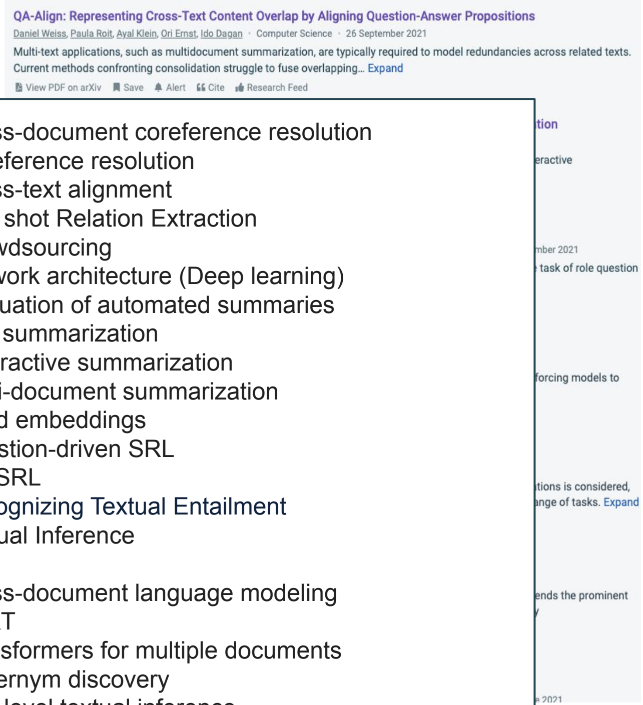
action model

ability to extract ented and

fication, an

Dan Wel

Automatic summarization
Generation of TLDR summary
analysis of scientific text
Paper recommendation
document level embedding of scientific documents
Network architecture (OS)
SpanBERT
Language model
Pre-trained language models
ELMo
Coreference resolution
coreference resolution across multiple documents
Neural network architectures
information retrieval systems
Natural language inference
search engines
human facing application
human centered ai
user interfaces
information extraction
extreme extraction
self training event extraction system'
….



QA-Align: Representing Cross-Text Content Overlap by Aligning Question-Answer Propositions

Daniel Weiss, Paula Roit, Ayal Klein, Ori Ernst, Ido Dagan · Computer Science · 26 September 2021

Multi-text applications, such as multidocument summarization, are typically required to model redundancies across related texts. Current methods confronting consolidation struggle to fuse overlapping… Expand

tion

eractive

mber 2021

task of role question

forcing models to

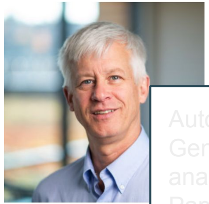
tions is considered, nge of tasks. Expand

ends the prominent

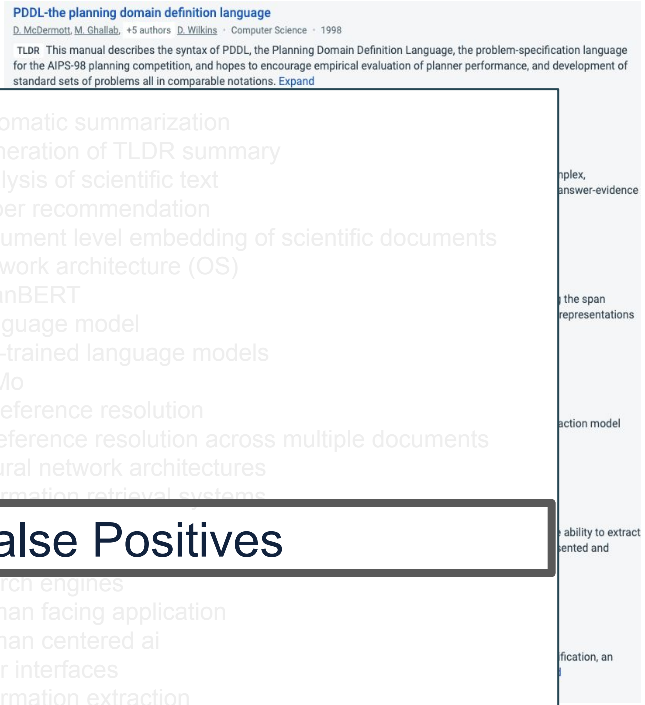
2021

Ido Dag

Cross-document coreference resolution
Coreference resolution
Cross-text alignment
Few shot Relation Extraction
Crowdsourcing
Network architecture (Deep learning)
Evaluation of automated summaries
Text summarization
Abstractive summarization
Multi-document summarization
Word embeddings
Question-driven SRL
QA-SRL
Recognizing Textual Entailment
Textual Inference
NLP
Cross-document language modeling
BERT
Transformers for multiple documents
Hypernym discovery
Low-level textual inference
….



Dan Wel



PDDL-the planning domain definition language

D. McDermott, M. Ghallab, +5 authors D. Wilkins · Computer Science · 1998

TLDR This manual describes the syntax of PDDL, the Planning Domain Definition Language, the problem-specification language for the AIPS-98 planning competition, and hopes to encourage empirical evaluation of planner performance, and development of standard sets of problems all in comparable notations. Expand

nplex,
answer-evidence

the span
representations

action model

ability to extract
sented and

fication, an

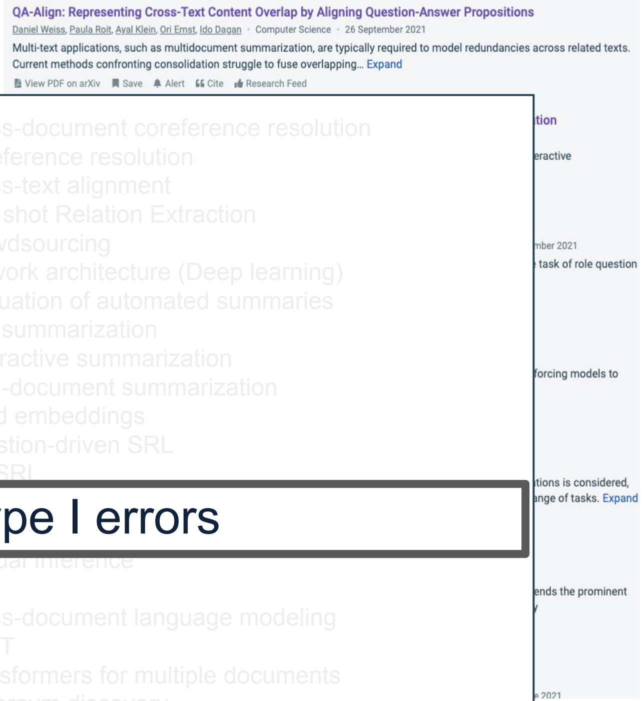
Automatic summarization
Generation of TLDR summary
analysis of scientific text
Paper recommendation
document level embedding of scientific documents
Network architecture (OS)
SpanBERT
Language model
Pre-trained language models
ELMo
Coreference resolution
coreference resolution across multiple documents
Neural network architectures
information retrieval systems

False Positives

search engines
human facing application
human centered ai
user interfaces
information extraction
extreme extraction
self training event extraction system'
….

Diversity

Ido Dag



QA-Align: Representing Cross-Text Content Overlap by Aligning Question-Answer Propositions

Daniel Weiss, Paula Roit, Ayal Klein, Ori Ernst, Ido Dagan · Computer Science · 26 September 2021

Multi-text applications, such as multidocument summarization, are typically required to model redundancies across related texts. Current methods confronting consolidation struggle to fuse overlapping… Expand

View PDF on arXiv Save Alert Cite Research Feed

tion

eractive

mber 2021
task of role question

forcing models to

tions is considered,
nge of tasks. Expand

ends the prominent

2021

Cross-document coreference resolution
Coreference resolution
Cross-text alignment
Few shot Relation Extraction
Crowdsourcing
Network architecture (Deep learning)
Evaluation of automated summaries
Text summarization
Abstractive summarization
Multi-document summarization
Word embeddings
Question-driven SRL
QA-SRL

Type I errors

Textual inference
NLP
Cross-document language modeling
BERT
Transformers for multiple documents
Hypernym discovery
Low-level textual inference
….

Dan Wel

PDDL-the planning domain definition language

D. McDermott, M. Ghallab, +5 authors D. Wilkins · Computer Science · 1998

TLDR This manual describes the syntax of PDDL, the Planning Domain Definition Language, the problem-specification language for the AIPS-98 planning competition, and hopes to encourage empirical evaluation of planner performance, and development of standard sets of problems all in comparable notations. Expand

Automatic summarization
Generation of TLDR summary
analysis of scientific text
Paper recommendation
document level embedding of scientific documents

Network architecture (systems)

SpanBERT
Language model
Pre-trained language models
ELMo
Coreference resolution
coreference resolution across multiple documents
Neural network architectures
information retrieval systems
Natural language inference
search engines
human facing application
human centered ai
user interfaces
information extraction
extreme extraction
self training event extraction system'
….

Ambiguity

Ido Dag

QA-Align: Representing Cross-Text Content Overlap by Aligning Question-Answer Propositions

Daniel Weiss, Paula Roit, Ayal Klein, Ori Ernst, Ido Dagan · Computer Science · 26 September 2021

Multi-text applications, such as multidocument summarization, are typically required to model redundancies across related texts. Current methods confronting consolidation struggle to fuse overlapping… Expand

View PDF on arXiv Save Alert Cite Research Feed

Cross-document coreference resolution
Coreference resolution
Cross-text alignment
Few shot Relation Extraction
Crowdsourcing

Network architecture (deep learning)

Evaluation of automated summaries
Text summarization
Abstractive summarization
Multi-document summarization
Word embeddings
Question-driven SRL
QA-SRL
Recognizing Textual Entailment
Textual Inference
NLP
Cross-document language modeling
BERT
Transformers for multiple documents
Hypernym discovery
Low-level textual inference
….

Dan Wel

PDDL-the planning domain definition language

D. McDermott, M. Ghallab, +5 authors D. Wilkins · Computer Science · 1998

TLDR This manual describes the syntax of PDDL, the Planning Domain Definition Language, the problem-specification language for the AIPS-98 planning competition, and hopes to encourage empirical evaluation of planner performance, and development of standard sets of problems all in comparable notations. Expand

Automatic summarization
Generation of TLDR summary
analysis of scientific text
Paper recommendation
document level embedding of scientific documents
Network architecture (OS)
SpanBERT
Language model
ELMo
Coreference resolution
coreference resolution across multiple documents
Neural network architectures
information retrieval systems
Natural language inference
search engines
human facing application
human centered ai
user interfaces
information extraction

Pre-trained language models

extreme extraction
self training event extraction system'
….

plex, answer-evidence

the span representations

action model

ability to extract ented and

fication, an

Hierarchy

Ido Dag

QA-Align: Representing Cross-Text Content Overlap by Aligning Question-Answer Propositions

Daniel Weiss, Paula Roit, Ayal Klein, Ori Ernst, Ido Dagan · Computer Science · 26 September 2021

Multi-text applications, such as multidocument summarization, are typically required to model redundancies across related texts. Current methods confronting consolidation struggle to fuse overlapping… Expand

View PDF on arXiv · Save · Alert · Cite · Research Feed

Cross-document coreference resolution
Coreference resolution
Cross-text alignment
Few shot Relation Extraction
Crowdsourcing
Network architecture (Deep learning)
Evaluation of automated summaries
Text summarization
Abstractive summarization
Multi-document summarization
Word embeddings
Question-driven SRL
QA-SRL
Recognizing Textual Entailment
Textual Inference
NLP

Cross-document language modeling

BERT
Transformers for multiple documents
Hypernym discovery
Low-level textual inference
….

tion

eractive

mber 2021

task of role question

forcing models to

tions is considered, nge of tasks. Expand

ends the prominent

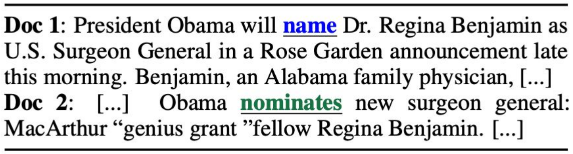
e 2021

Limitations of Previous Work

Cross-Document Coreference Resolution (CDCR)

- No abstract technical concepts
  - No work in science!
- No cross-document **hierarchy**

ECB+[1]



Doc 1: President Obama will **name** Dr. Regina Benjamin as U.S. Surgeon General in a Rose Garden announcement late this morning. Benjamin, an Alabama family physician, [...]
Doc 2: [...] Obama **nominates** new surgeon general: MacArthur "genius grant "fellow Regina Benjamin. [...]

[1] Cybulska, Agata and P. Vossen. "Using a sledgehammer to crack a nut? Lexical diversity and event coreference resolution." LREC (2014).

Goal: *Address cross-document ambiguity, diversity and hierarchy together*

New task: *cross-document ambiguity, diversity, hierarchy*

Input:

Concept mentions in scientific papers

… **self-driving cars** have made it increasingly urgent …

… navigation for **autonomous vehicles** in real-life traffic …

… transformer models in **computer vision** …

… we use **categorical image generation** …

… the problem of **generating images** …

New task: *cross-document ambiguity, diversity, hierarchy*

Input:
Concept mentions in scientific papers

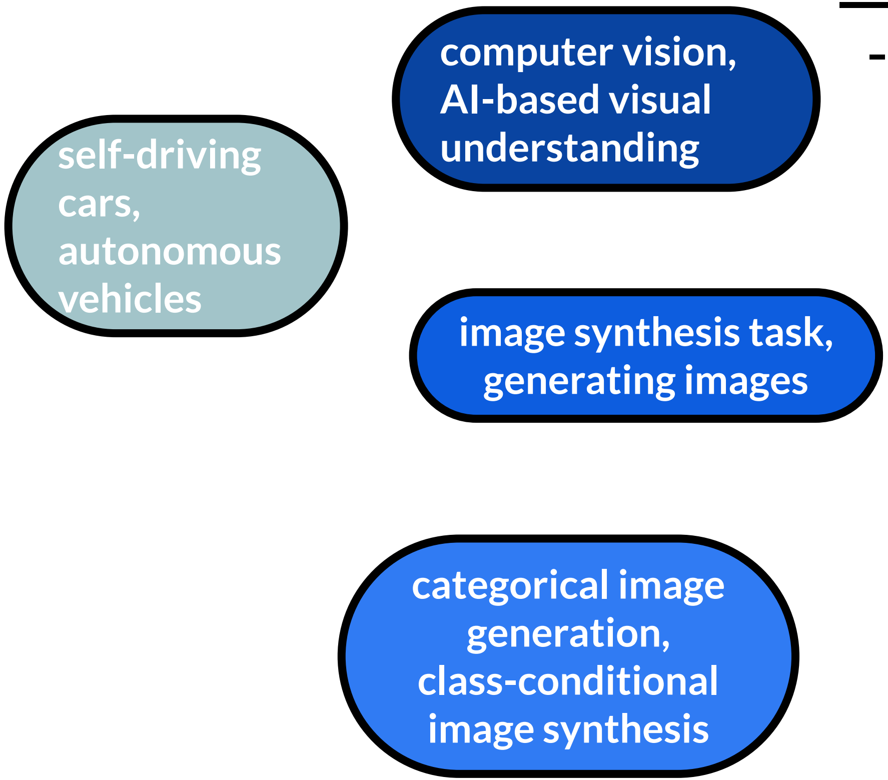
... **self-driving cars** have made it increasingly urgent ...

... navigation for **autonomous vehicles** in real-life traffic ...

... transformer models in **computer vision** ...

... we use **categorical image generation** ...

... the problem of **generating images** ...



self-driving cars, autonomous vehicles

computer vision, AI-based visual understanding

image synthesis task, generating images

categorical image generation, class-conditional image synthesis

Output:
- Coreference Clusters

New task: *cross-document ambiguity, diversity, hierarchy*

Input:
Concept mentions in scientific papers

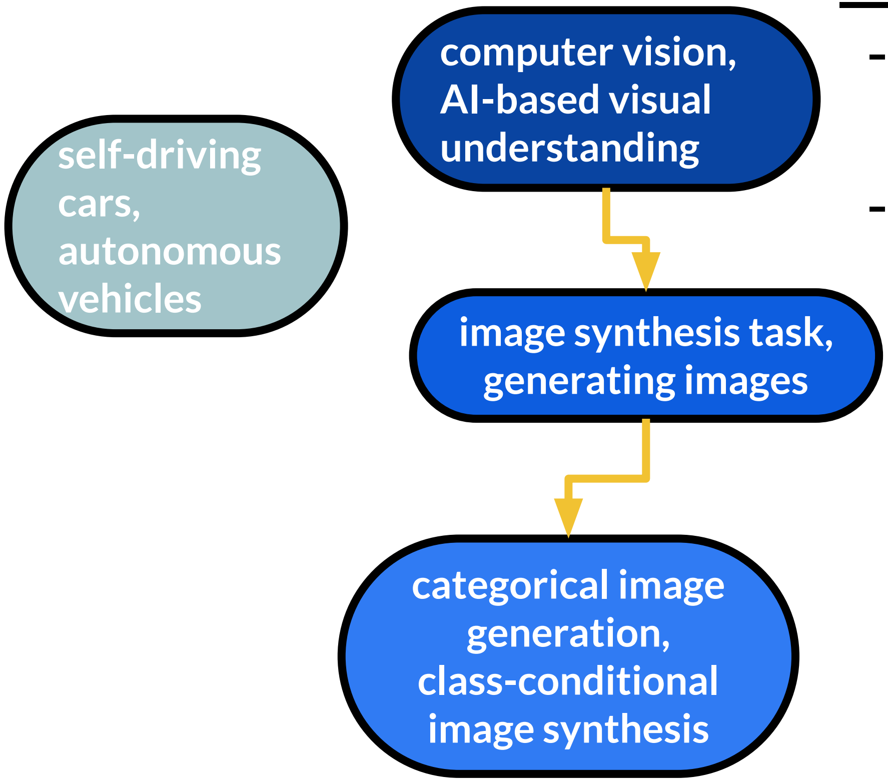
... **self-driving cars** have made it increasingly urgent ...

... navigation for **autonomous vehicles** in real-life traffic ...

... transformer models in **computer vision** ...

... we use **categorical image generation** ...

... the problem of **generating images** ...


self-driving cars, autonomous vehicles
computer vision, AI-based visual understanding
image synthesis task, generating images
categorical image generation, class-conditional image synthesis

Output:

- Coreference Clusters
- Referential Hierarchy

New task: *cross-document ambiguity, diversity, hierarchy*

Input:
Concept mentions in scientific papers

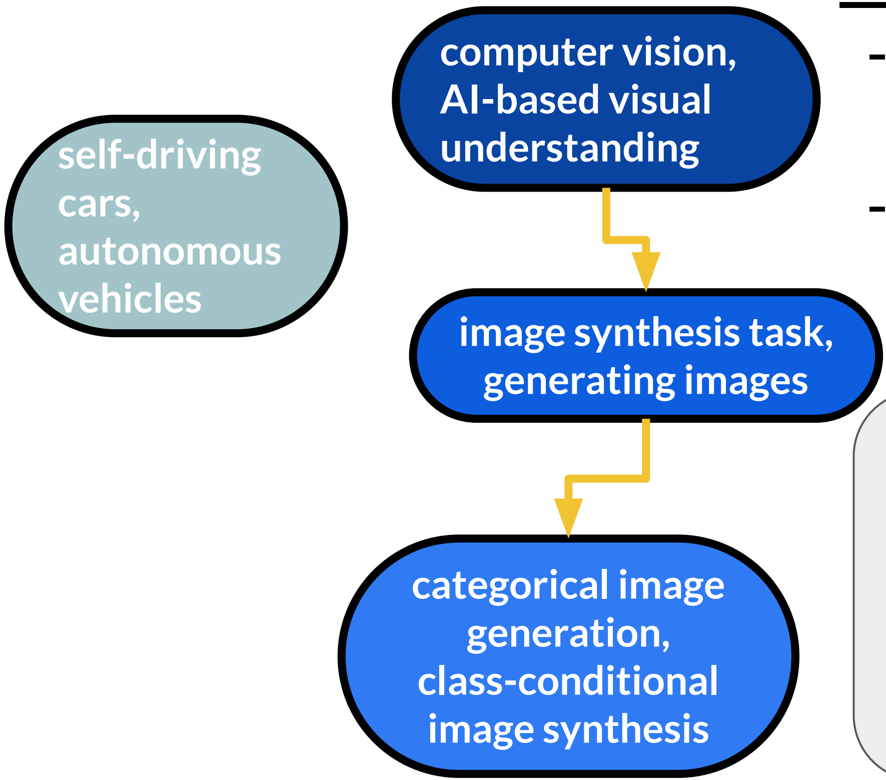
... **self-driving cars** have made it increasingly urgent ...

... navigation for **autonomous vehicles** in real-life traffic ...

... transformer models in **computer vision** ...

... we use **categorical image generation** ...

... the problem of **generating images** ...

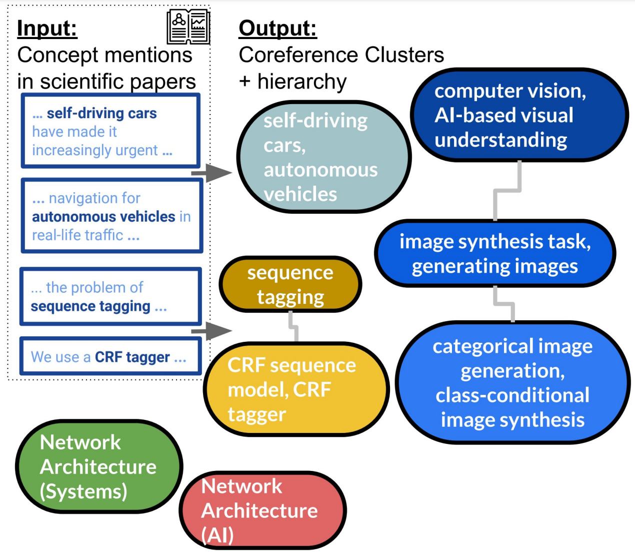


self-driving cars, autonomous vehicles
computer vision, AI-based visual understanding
image synthesis task, generating images
categorical image generation, class-conditional image synthesis

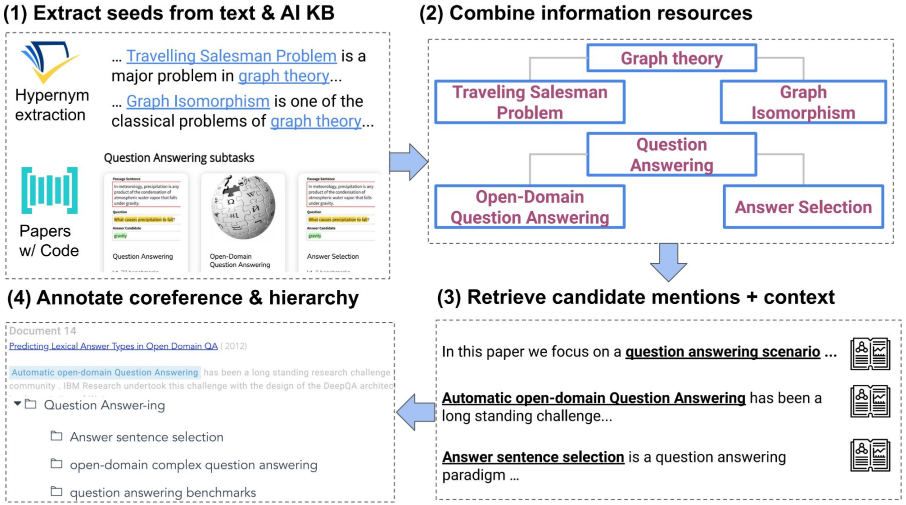
Output:
- Coreference Clusters
- Referential Hierarchy

Mentioning a concept → implicit reference to parent concept

Hierarchical Cross-Document Coreference Resolution (H-CDCR)


Input:
Concept mentions in scientific papers
... **self-driving cars** have made it increasingly urgent ...
... navigation for **autonomous vehicles** in real-life traffic ...
... the problem of **sequence tagging** ...
We use a **CRF tagger** ...
Output:
Coreference Clusters + hierarchy
self-driving cars, autonomous vehicles
computer vision, AI-based visual understanding
image synthesis task, generating images
categorical image generation, class-conditional image synthesis
sequence tagging
CRF sequence model, CRF tagger
Network Architecture (Systems)
Network Architecture (AI)

SciCo: A new large-scale dataset annotated by domain experts



(1) Extract seeds from text & AI KB

Hypernym extraction

... Travelling Salesman Problem is a major problem in graph theory...

... Graph Isomorphism is one of the classical problems of graph theory...

Papers w/ Code

Question Answering subtasks

Passage Sentence
In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity.
Question
What causes precipitation to fall?
Answer Candidate
gravity

Question Answering

Open-Domain Question Answering

Passage Sentence
In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity.
Question
What causes precipitation to fall?
Answer Candidate
gravity

Answer Selection

(2) Combine information resources

Graph theory

Traveling Salesman Problem

Graph Isomorphism

Question Answering

Open-Domain Question Answering

Answer Selection

(3) Retrieve candidate mentions + context

In this paper we focus on a **question answering scenario** ...

Automatic open-domain Question Answering has been a long standing challenge...

Answer sentence selection is a question answering paradigm ...

(4) Annotate coreference & hierarchy

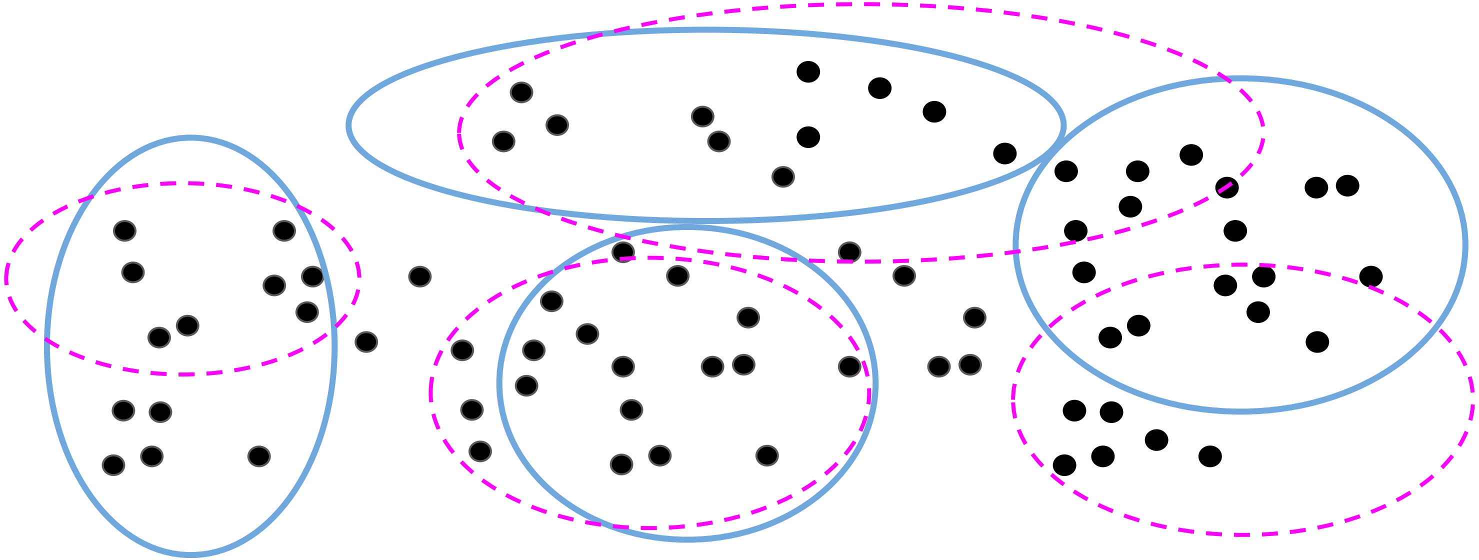
Document 14

Predicting Lexical Answer Types in Open Domain QA (2012)

Automatic open-domain Question Answering has been a long standing research challenge community . IBM Research undertook this challenge with the design of the DeepQA architec

- Question Answer-ing
  - Answer sentence selection
  - open-domain complex question answering
  - question answering benchmarks

Novel evaluation metrics for *hierarchical* cross-document coreference resolution



Custom Baseline Models

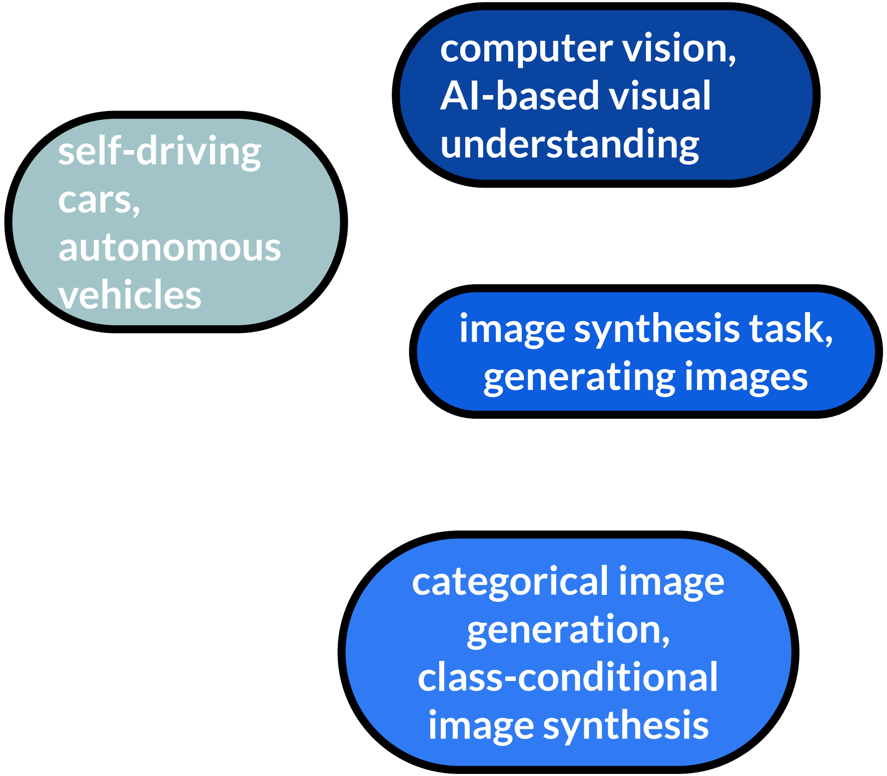
Baseline I: Two-Step Model

Two steps:

1. **Concept clusters:** Apply existing SOTA CDCR model
2. **Hierarchy**: Find relations between predicted clusters

Baseline I: Two-Step Model

✅ We have clusters (using SOTA trained on SciCo)



computer vision, AI-based visual understanding

self-driving cars, autonomous vehicles

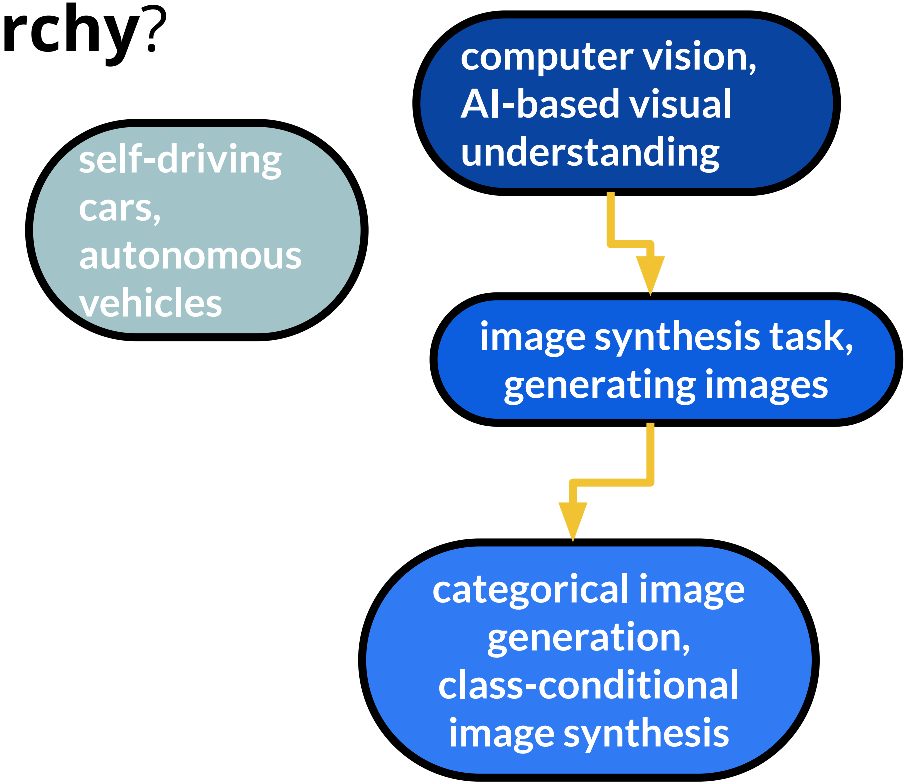
image synthesis task, generating images

categorical image generation, class-conditional image synthesis

Baseline I: Two-Step Model

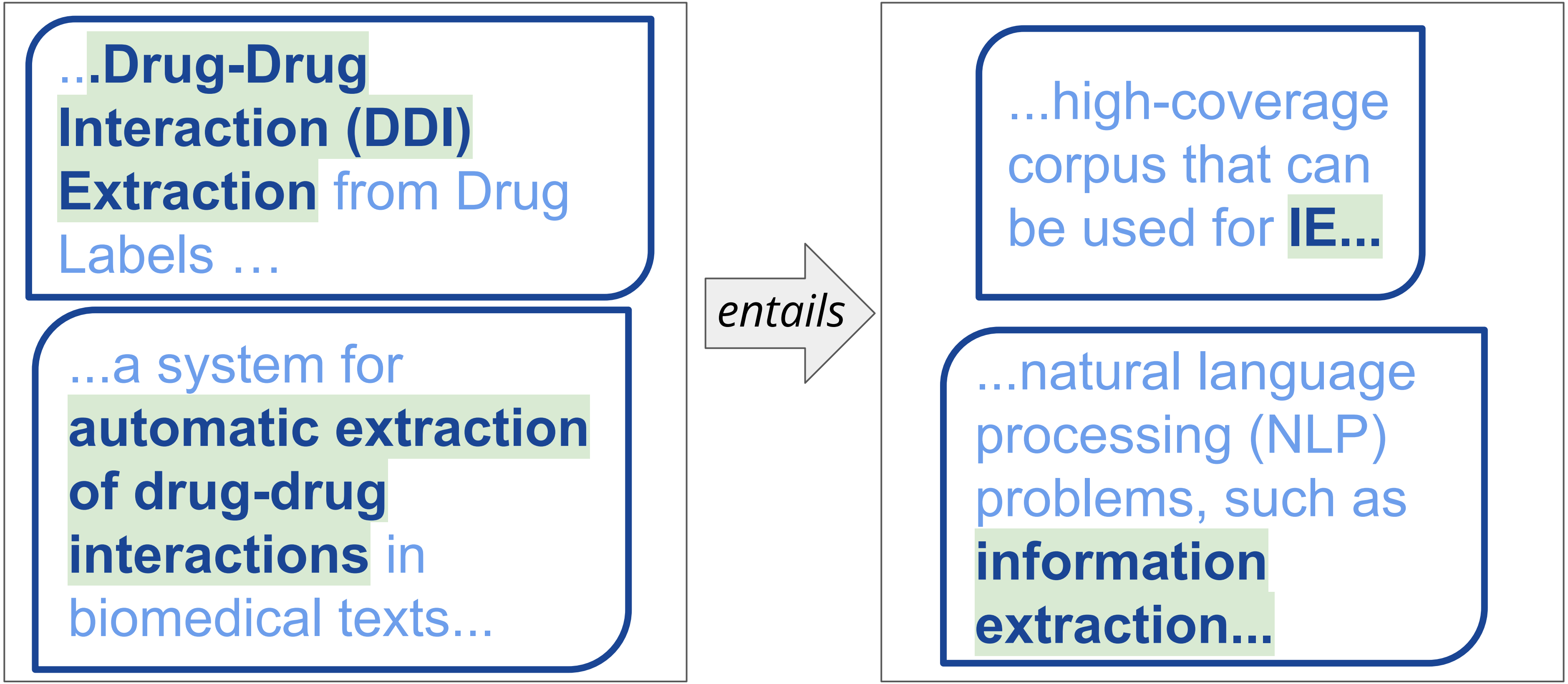
✅ We have clusters (using SOTA trained on SciCo)

❓ How do we **infer the hierarchy**?


self-driving cars, autonomous vehicles
computer vision, AI-based visual understanding
image synthesis task, generating images
categorical image generation, class-conditional image synthesis

Intuition: Referential hierarchy as ‘**multi-document textual entailment**’

Intuition: Referential hierarchy as ‘**multi-document textual entailment**’



...**Drug-Drug Interaction (DDI) Extraction** from Drug Labels ...

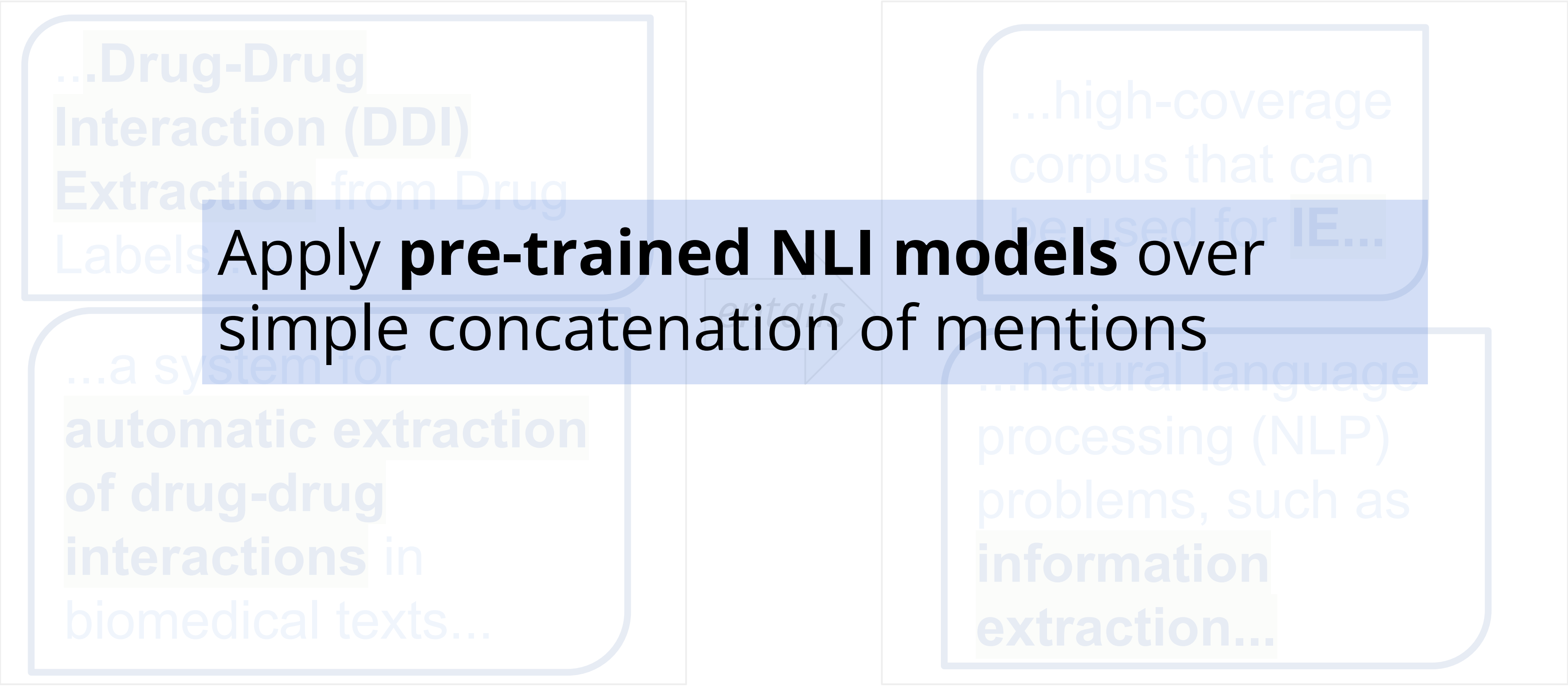
...a system for **automatic extraction of drug-drug interactions** in biomedical texts...

entails

...high-coverage corpus that can be used for **IE...**

...natural language processing (NLP) problems, such as **information extraction...**

Intuition: Referential hierarchy as ‘**multi-document entailment**’



...**Drug-Drug Interaction (DDI) Extraction** from Drug Labels...

...a system for **automatic extraction of drug-drug interactions** in biomedical texts...

entails

...high-coverage corpus that can be used for **IE**...

...natural language processing (NLP) problems, such as **information extraction**...

Apply **pre-trained NLI models** over simple concatenation of mentions

Unified Model

Unified model with multiclass formulation
for pairs of mentions m1,m2 with classes {$\asymp$, ←, →, None}

$$L = -\frac{1}{N} \sum_{\substack{m_1, m_2 \in \mathcal{M} \\ m_1 \neq m_2}} y \cdot log(f(m_1, m_2))$$

Unified Model

ƒ(m1,m2): **Cross-encode** mentions m1,m2 with entity markers

[CLS]…an experiment in **<m> definition extraction </m>** from legal texts ... [SEP] … natural language processing problems, such as **<m> information extraction </m>,** summarization and dialogue.... [SEP]

Unified Model

f(m1,m2): **Cross-encode** mention pairs m1,m2 with entity markers

[CLS] …an experiment in **<m> definition extraction </m>** from legal texts … [SEP] … natural language processing problems, such as **<m> information extraction </m>,** summarization and dialogue…. [SEP]

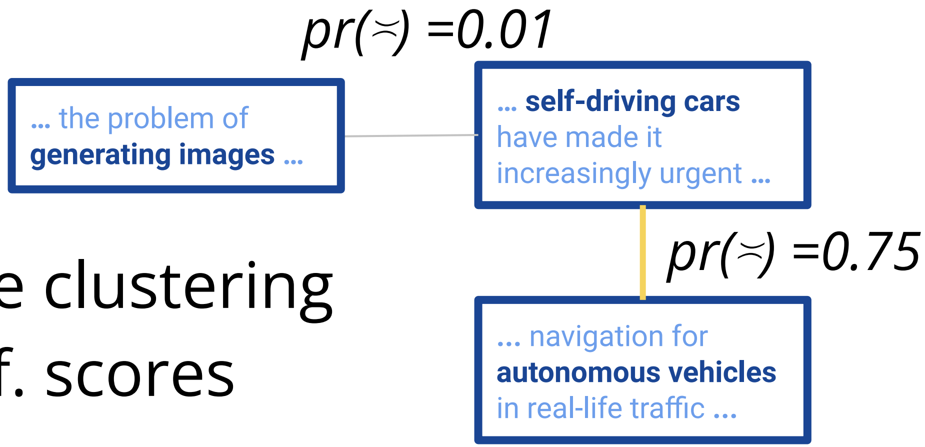
Leverage **wide context**:

- LongFormer[1]: Transformers for ***long-sequences***
- CDLM[2]: Transformers for ***multi-document*** tasks

[1] Beltagy, Iz, Matthew E. Peters and Arman Cohan. “Longformer: The Long-Document Transformer.” (2020)
[2] Caciularu, Avi, Arman Cohan, Iz Beltagy, Matthew E. Peters, Arie Cattan and Ido Dagan. “CDLM: Cross-Document Language Modeling.” (EMNLP FINDINGS 2021).

Inference:

1. **Clusters**: Agglomerative clustering over mention-pair coref. scores



$pr(\asymp) = 0.01$

… the problem of **generating images** …

… **self-driving cars** have made it increasingly urgent …

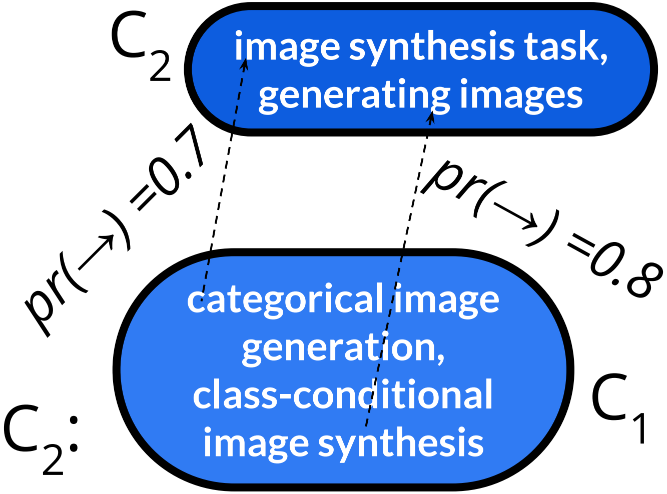
$pr(\asymp) = 0.75$

… navigation for **autonomous vehicles** in real-life traffic …

Inference:

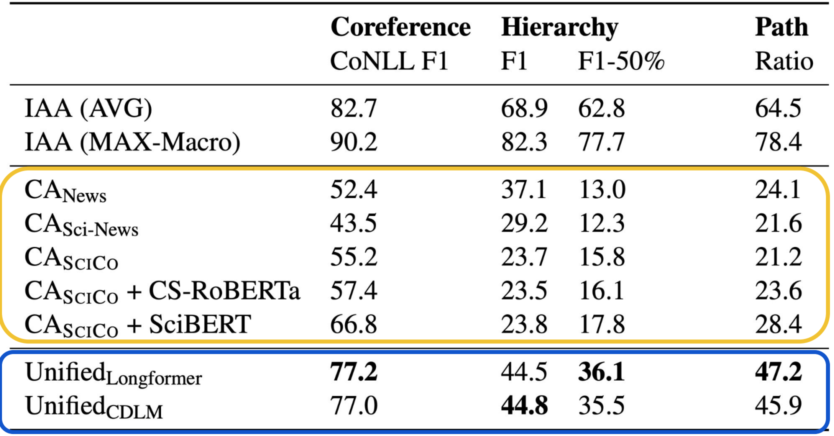
1. **Clusters**: Agglomerative clustering over mention-pair coref. scores
2. **Hierarchy**:
   a. Score (prob.) that C_1 is a child of C_2:

$$s(\mathcal{C}_1,\mathcal{C}_2)=\frac{1}{|\mathcal{C}_1|\cdot|\mathcal{C}_2|}\sum_{m_1\in\mathcal{C}_1}\sum_{m_2\in\mathcal{C}_2} f_{\text{is-child}}(m_1,m_2)$$


C_2
image synthesis task, generating images
pr(→) =0.7
pr(→) =0.8
categorical image generation, class-conditional image synthesis
C_1

   b. Greedy construction of hierarchy (to avoid cycles)

Results

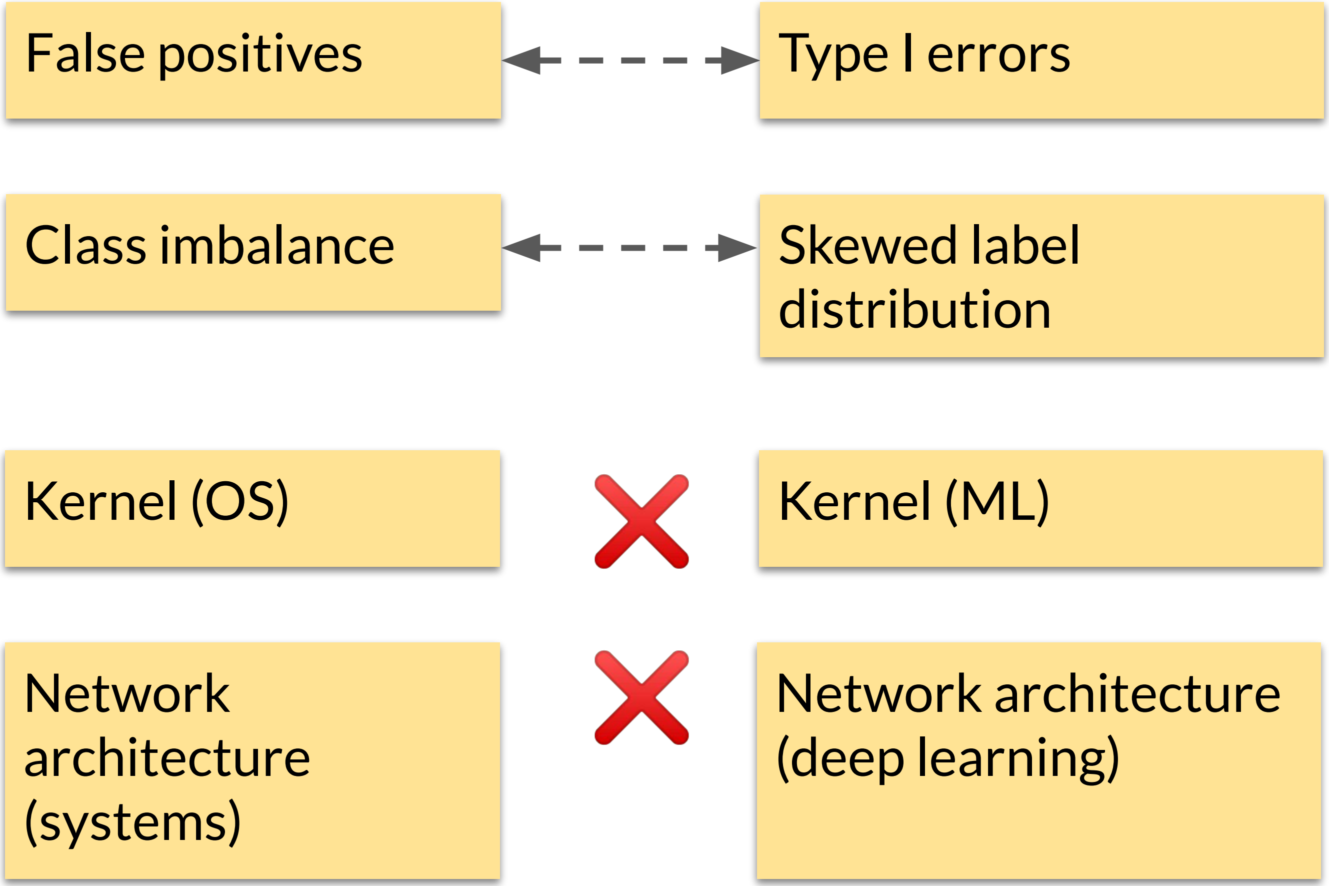


	Coreference	**Hierarchy**		**Path**
	CoNLL F1	F1	F1-50%	Ratio
IAA (AVG)	82.7	68.9	62.8	64.5
IAA (MAX-Macro)	90.2	82.3	77.7	78.4
CA$_{\text{News}}$	52.4	37.1	13.0	24.1
CA$_{\text{Sci-News}}$	43.5	29.2	12.3	21.6
CA$_{\text{SciCo}}$	55.2	23.7	15.8	21.2
CA$_{\text{SciCo}}$ + CS-RoBERTa	57.4	23.5	16.1	23.6
CA$_{\text{SciCo}}$ + SciBERT	66.8	23.8	17.8	28.4
Unified$_{\text{Longformer}}$	**77.2**	44.5	**36.1**	**47.2**
Unified$_{\text{CDLM}}$	77.0	**44.8**	35.5	45.9

Two-step models

Multi-class cross-encoder

Results



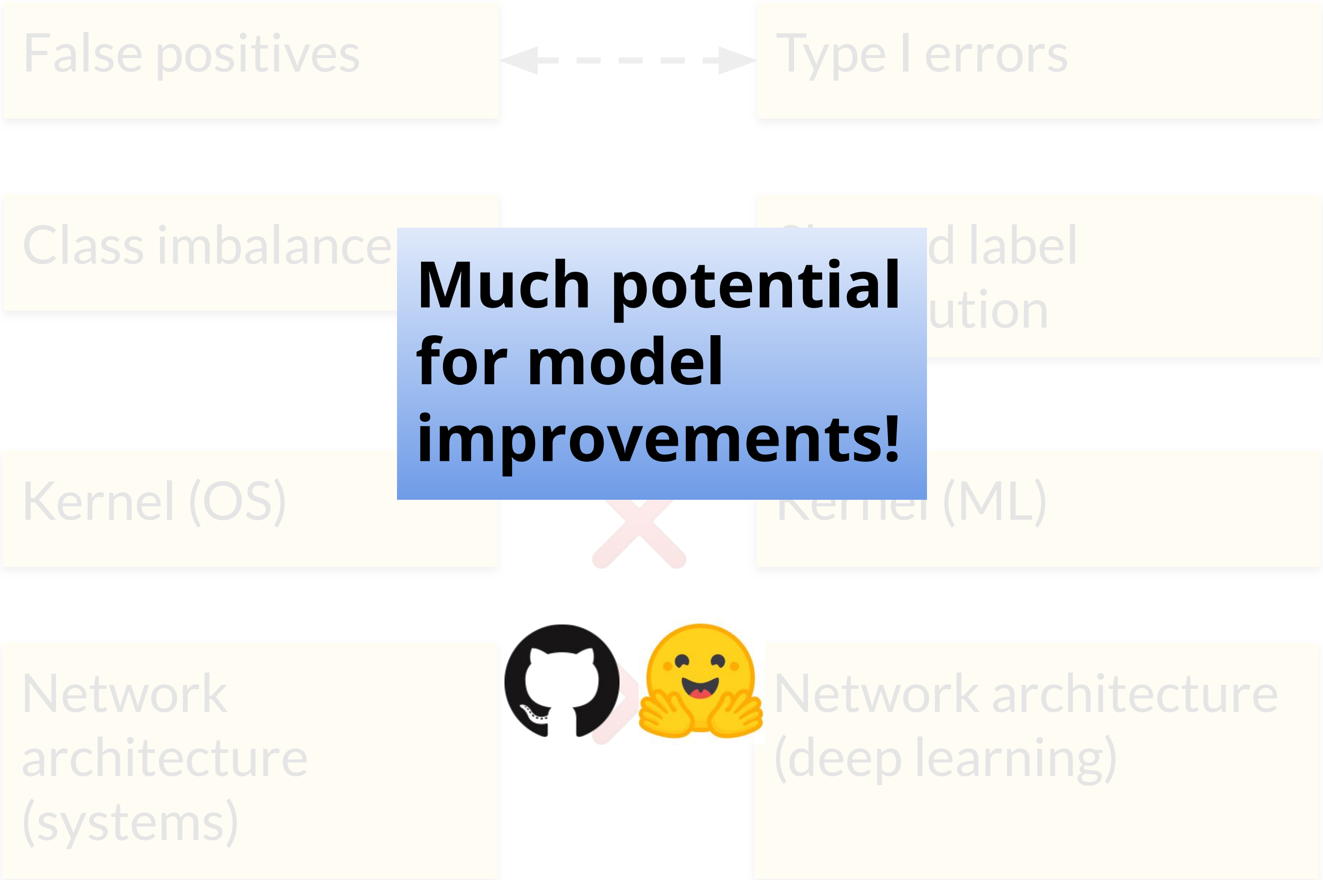
False positives ↔ Type I errors

Class imbalance ↔ Skewed label distribution

Kernel (OS) ❌ Kernel (ML)

Network architecture (systems) ❌ Network architecture (deep learning)

Results



False positives

Type I errors

Class imbalance

d label
ution

Kernel (OS)

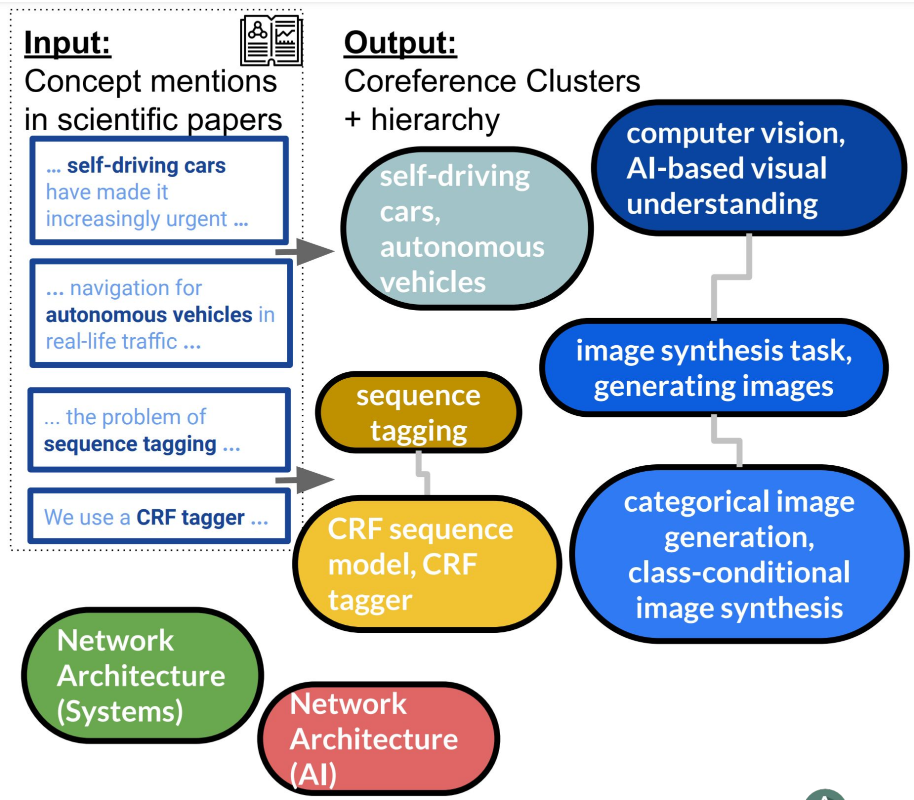
Kernel (ML)

Network architecture (systems)

Network architecture (deep learning)

Much potential for model improvements!

SciCo: Hierarchical Cross-Document Coreference



Input:
Concept mentions in scientific papers

... **self-driving cars** have made it increasingly urgent ...

... navigation for **autonomous vehicles** in real-life traffic ...

... the problem of **sequence tagging** ...

We use a **CRF tagger** ...

Output:
Coreference Clusters + hierarchy

self-driving cars, autonomous vehicles

computer vision, AI-based visual understanding

image synthesis task, generating images

categorical image generation, class-conditional image synthesis

sequence tagging

CRF sequence model, CRF tagger

Network Architecture (Systems)

Network Architecture (AI)

Outstanding Paper Award

AUTOMATED KNOWLEDGE BASE 2021 | irvine CONSTRUCTION

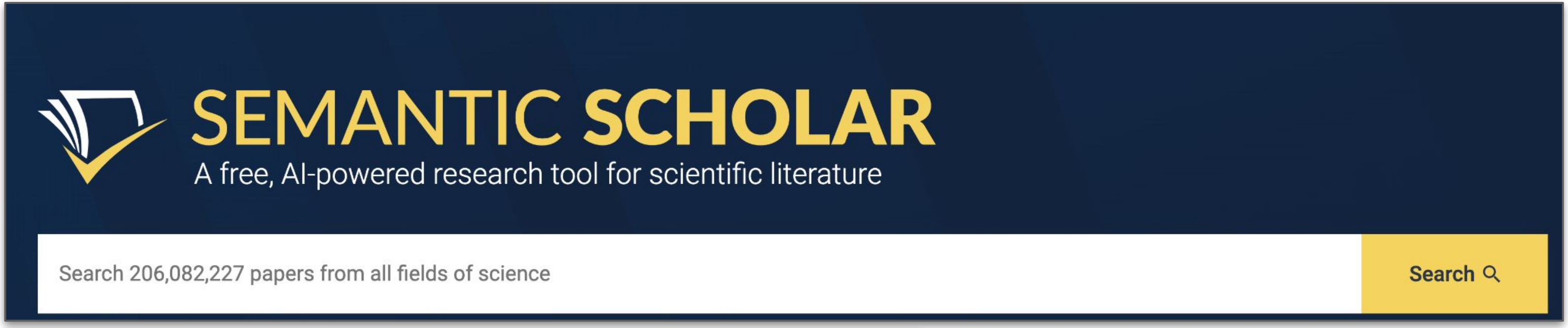
1. SciCo: Hierarchical Cross-Document Coreference
2. **Document Similarity & Retrieval**
3. Literature-Augmented Prediction

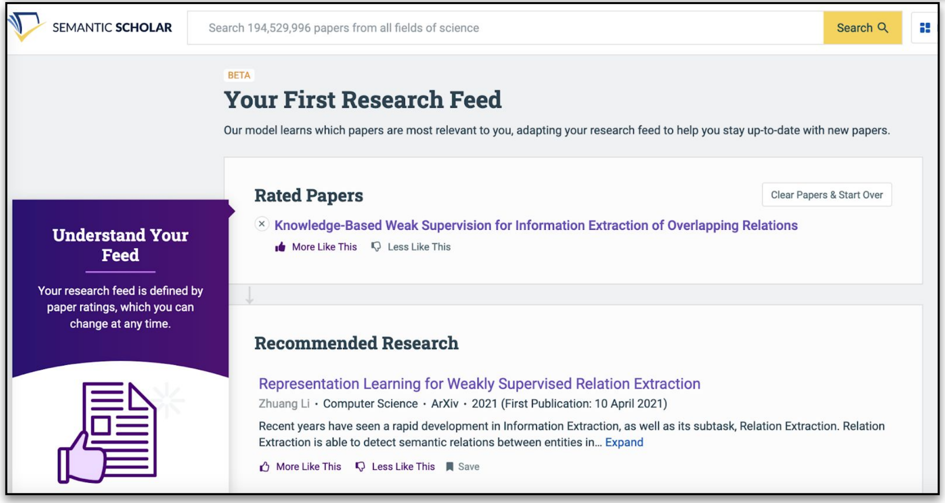
Multi-Vector Models with Textual Guidance for Fine-Grained Scientific Document Similarity

Sheshera Mysore, Arman Cohan & **Tom Hope**

NAACL 2022

Document Similarity for Science


SEMANTIC SCHOLAR
A free, AI-powered research tool for scientific literature
Search 206,082,227 papers from all fields of science
Search



Your First Research Feed

Our model learns which papers are most relevant to you, adapting your research feed to help you stay up-to-date with new papers.

Rated Papers

Clear Papers & Start Over

Knowledge-Based Weak Supervision for Information Extraction of Overlapping Relations

More Like This Less Like This

Understand Your Feed

Your research feed is defined by paper ratings, which you can change at any time.

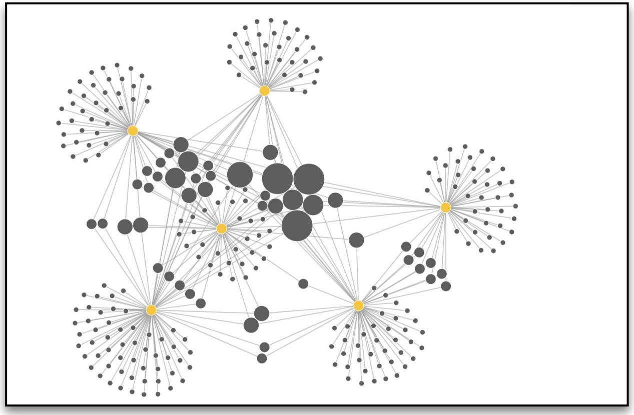
Recommended Research

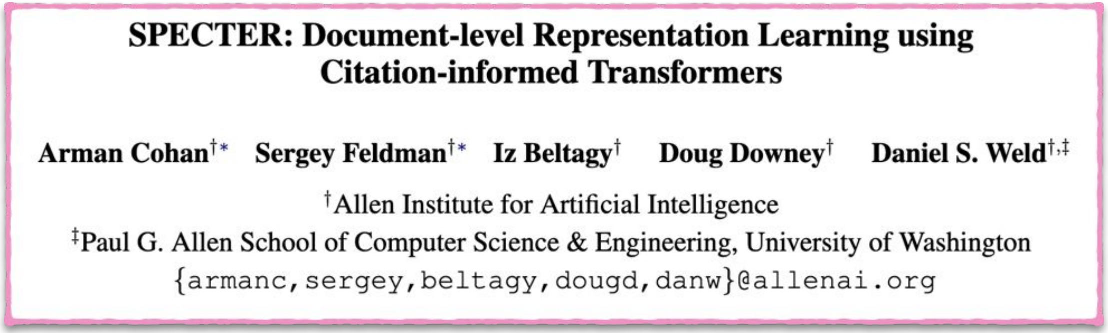
Representation Learning for Weakly Supervised Relation Extraction

Zhuang Li • Computer Science • ArXiv • 2021 (First Publication: 10 April 2021)

Recent years have seen a rapid development in Information Extraction, as well as its subtask, Relation Extraction. Relation Extraction is able to detect semantic relations between entities in... Expand

More Like This Less Like This Save





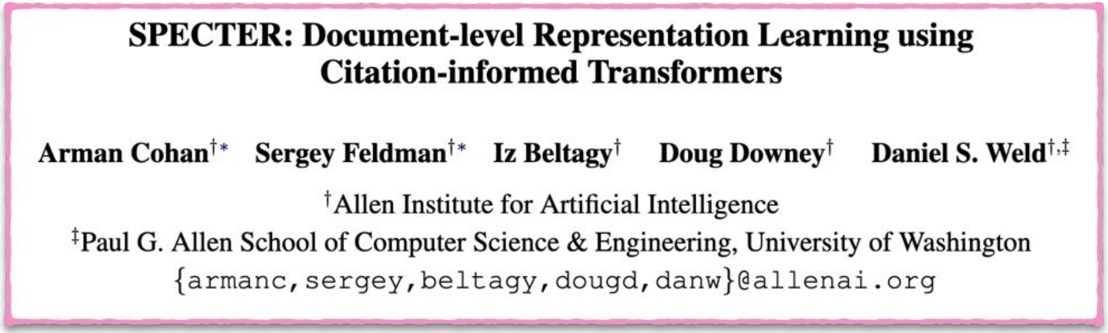
SPECTER: Document-level Representation Learning using Citation-informed Transformers

Arman Cohan[†*] **Sergey Feldman**[†*] **Iz Beltagy**[†] **Doug Downey**[†] **Daniel S. Weld**[†,‡]

[†]Allen Institute for Artificial Intelligence

[‡]Paul G. Allen School of Computer Science & Engineering, University of Washington

{armanc,sergey,beltagy,dougd,danw}@allenai.org



SPECTER: Document-level Representation Learning using Citation-informed Transformers

Arman Cohan[†*] **Sergey Feldman**[†*] **Iz Beltagy**[†] **Doug Downey**[†] **Daniel S. Weld**[†,‡]

[†]Allen Institute for Artificial Intelligence

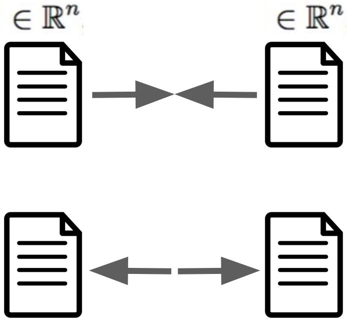
[‡]Paul G. Allen School of Computer Science & Engineering, University of Washington

{armanc,sergey,beltagy,dougd,danw}@allenai.org

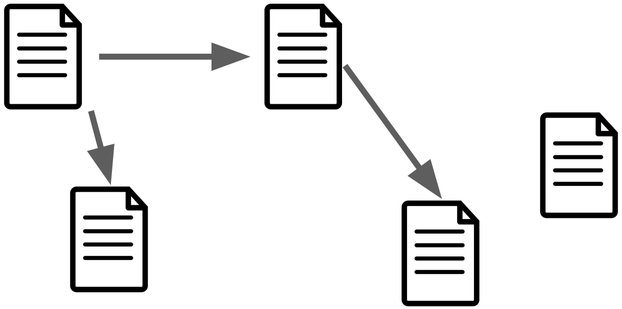
Contrastive learning: Learn embeddings of papers that:

Pull together papers that are related/similar

Pull apart papers that are unrelated/dissimilar

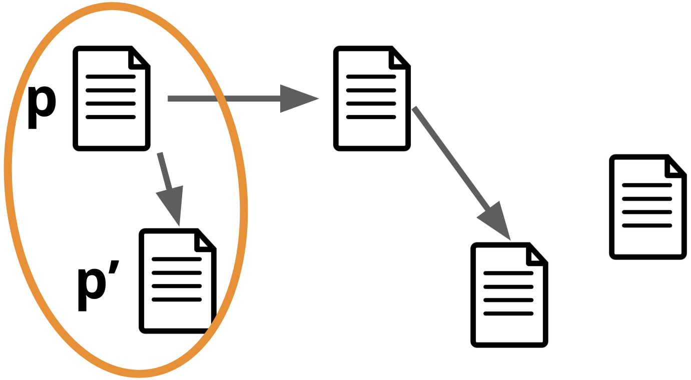


Citation network



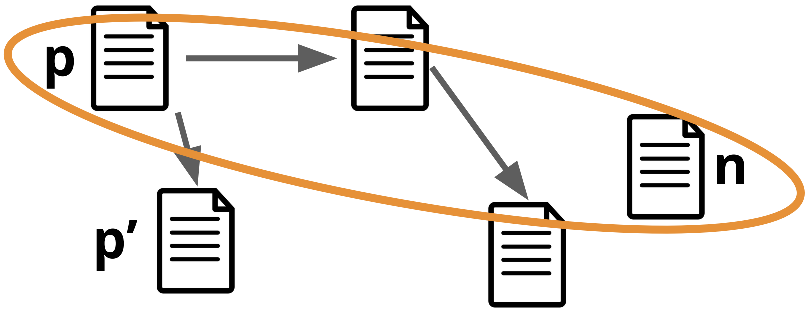
Related papers:
Paper p cites paper p'

Citation network



Unrelated papers:
No citation link between paper p and paper n

Citation network



$$f(p, p') < f(p, n)$$

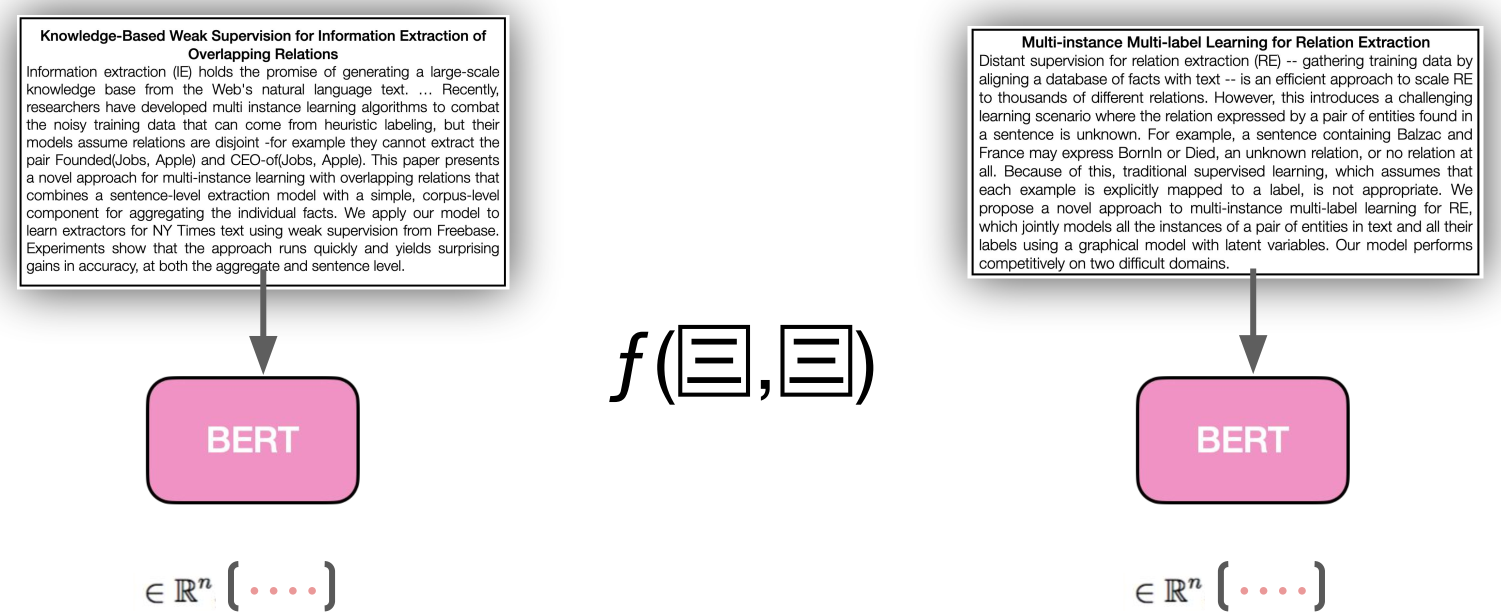
Distance between paper **p**, and a ***related*** paper **p'**

Distance between paper **p**, and a ***unrelated*** paper **n**

Contrastive training loss

$$\boxed{\mathcal{L}_f(p, p', n) = \texttt{max}[f(p, p') - f(p, n) + m, 0]}$$

f(🗎,🗎) = ❓



Knowledge-Based Weak Supervision for Information Extraction of Overlapping Relations

Information extraction (IE) holds the promise of generating a large-scale knowledge base from the Web's natural language text. … Recently, researchers have developed multi instance learning algorithms to combat the noisy training data that can come from heuristic labeling, but their models assume relations are disjoint -for example they cannot extract the pair Founded(Jobs, Apple) and CEO-of(Jobs, Apple). This paper presents a novel approach for multi-instance learning with overlapping relations that combines a sentence-level extraction model with a simple, corpus-level component for aggregating the individual facts. We apply our model to learn extractors for NY Times text using weak supervision from Freebase. Experiments show that the approach runs quickly and yields surprising gains in accuracy, at both the aggregate and sentence level.

Multi-instance Multi-label Learning for Relation Extraction

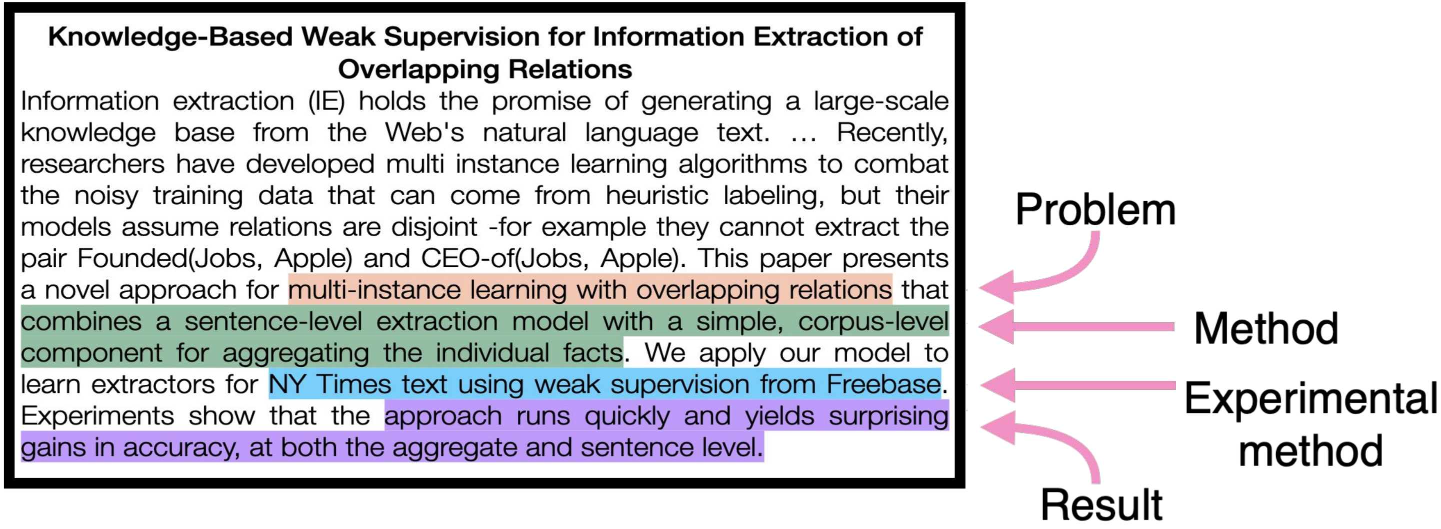
Distant supervision for relation extraction (RE) -- gathering training data by aligning a database of facts with text -- is an efficient approach to scale RE to thousands of different relations. However, this introduces a challenging learning scenario where the relation expressed by a pair of entities found in a sentence is unknown. For example, a sentence containing Balzac and France may express BornIn or Died, an unknown relation, or no relation at all. Because of this, traditional supervised learning, which assumes that each example is explicitly mapped to a label, is not appropriate. We propose a novel approach to multi-instance multi-label learning for RE, which jointly models all the instances of a pair of entities in text and all their labels using a graphical model with latent variables. Our model performs competitively on two difficult domains.

BERT
BERT
$f($☰$,$☰$)$
$\in \mathbb{R}^n$ [• • • •]
$\in \mathbb{R}^n$ [• • • •]

SPECTER: Distance based on one ***overall*** document vector

(e.g., Euclidean distance between p and p’ vector embeddings)

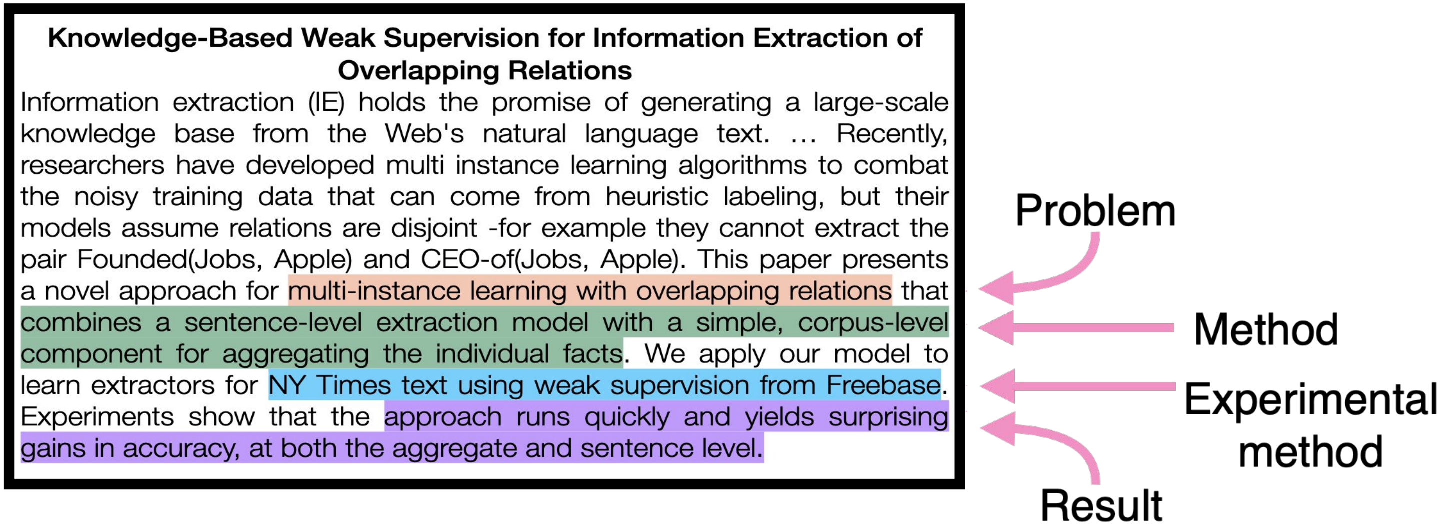
Scientific Documents are Multi-Faceted



Knowledge-Based Weak Supervision for Information Extraction of Overlapping Relations

Information extraction (IE) holds the promise of generating a large-scale knowledge base from the Web's natural language text. … Recently, researchers have developed multi instance learning algorithms to combat the noisy training data that can come from heuristic labeling, but their models assume relations are disjoint -for example they cannot extract the pair Founded(Jobs, Apple) and CEO-of(Jobs, Apple). This paper presents a novel approach for multi-instance learning with overlapping relations that combines a sentence-level extraction model with a simple, corpus-level component for aggregating the individual facts. We apply our model to learn extractors for NY Times text using weak supervision from Freebase. Experiments show that the approach runs quickly and yields surprising gains in accuracy, at both the aggregate and sentence level.

Scientific Documents are Multi-Faceted



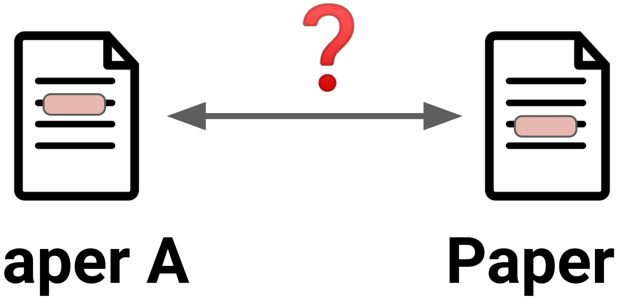
Knowledge-Based Weak Supervision for Information Extraction of Overlapping Relations

Information extraction (IE) holds the promise of generating a large-scale knowledge base from the Web's natural language text. … Recently, researchers have developed multi instance learning algorithms to combat the noisy training data that can come from heuristic labeling, but their models assume relations are disjoint -for example they cannot extract the pair Founded(Jobs, Apple) and CEO-of(Jobs, Apple). This paper presents a novel approach for multi-instance learning with overlapping relations that combines a sentence-level extraction model with a simple, corpus-level component for aggregating the individual facts. We apply our model to learn extractors for NY Times text using weak supervision from Freebase. Experiments show that the approach runs quickly and yields surprising gains in accuracy, at both the aggregate and sentence level.

Can ***aspect***-level modeling lead to better ***document***-level similarity?

How do we identify **fine-grained aspect similarity?**

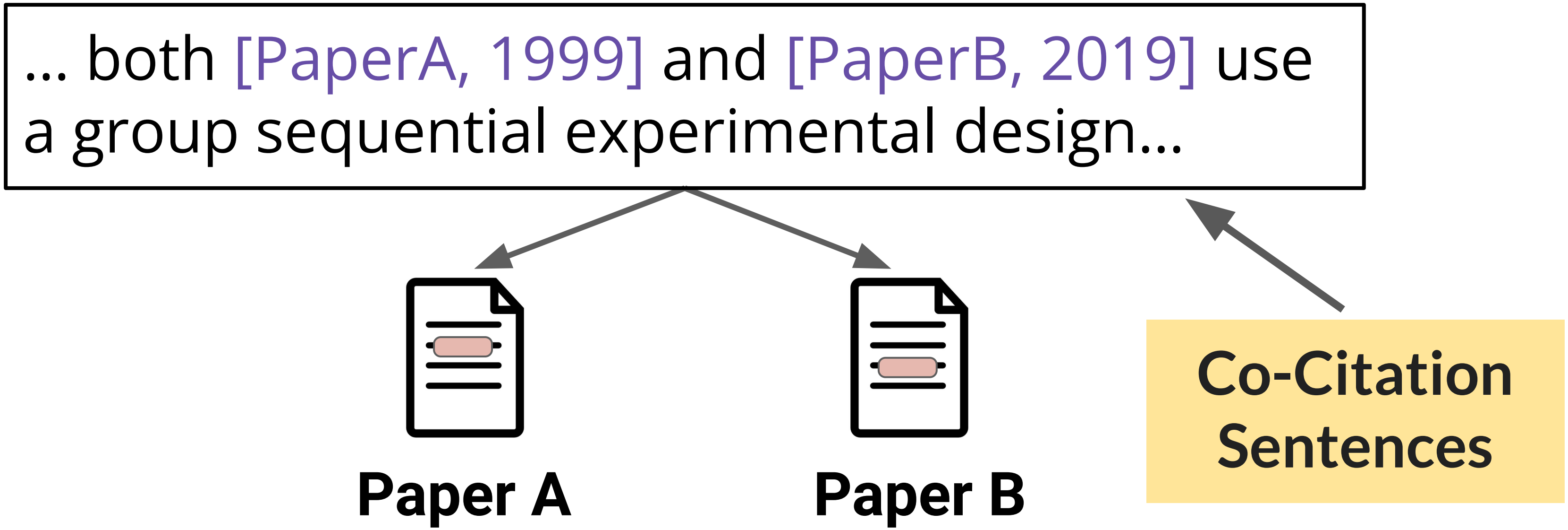
(Descriptions of methodologies, experiments, findings...)



How do we identify **fine-grained aspect similarity?**
(Descriptions of methodologies, experiments, findings…)

Paper A **Paper B**

No Gold Labels

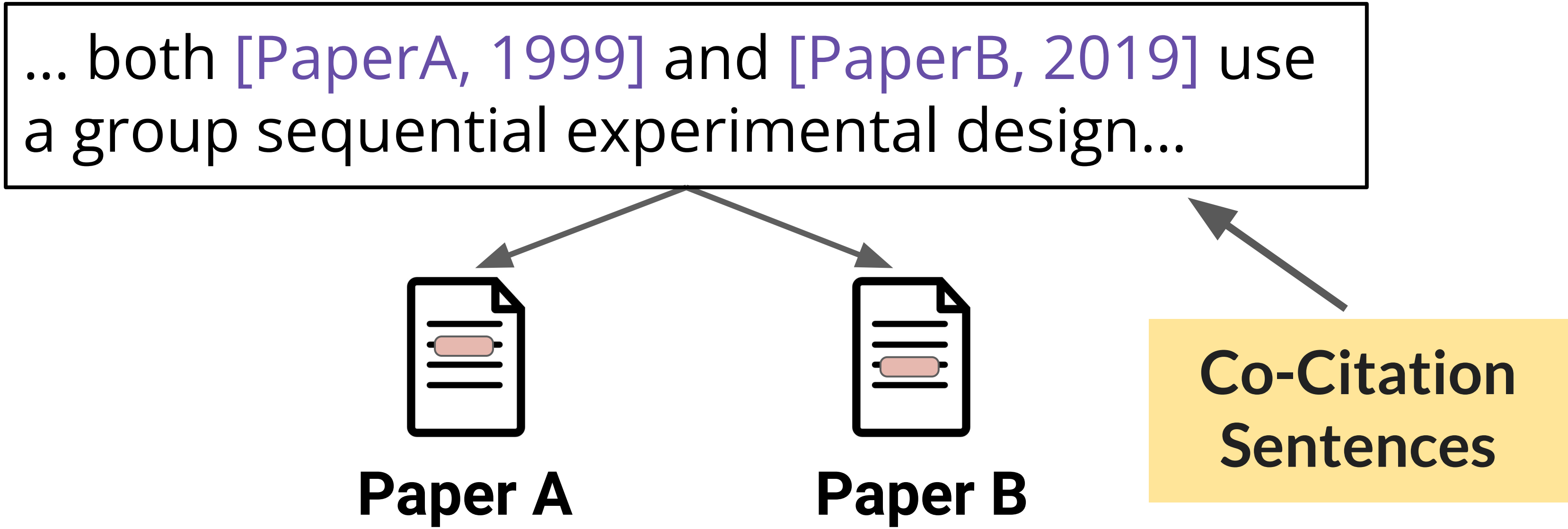


… both [PaperA, 1999] and [PaperB, 2019] use a group sequential experimental design…

Paper A

Paper B

Co-Citation Sentences



... both [PaperA, 1999] and [PaperB, 2019] use a group sequential experimental design...

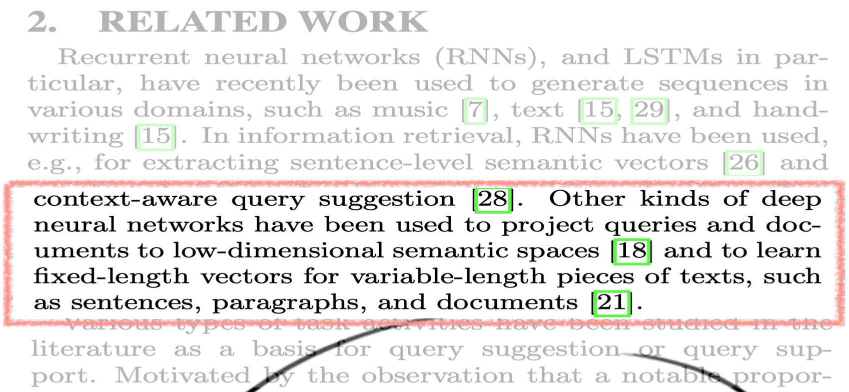
Paper A

Paper B

Co-Citation Sentences

Co-citation sentences provide explanations regarding *how* two papers are related

Textual supervision: co-citation contexts

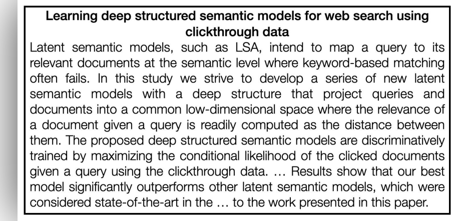


2. RELATED WORK

Recurrent neural networks (RNNs), and LSTMs in particular, have recently been used to generate sequences in various domains, such as music [7], text [15, 29], and handwriting [15]. In information retrieval, RNNs have been used, e.g., for extracting sentence-level semantic vectors [26] and context-aware query suggestion [28]. Other kinds of deep neural networks have been used to project queries and documents to low-dimensional semantic spaces [18] and to learn fixed-length vectors for variable-length pieces of texts, such as sentences, paragraphs, and documents [21].

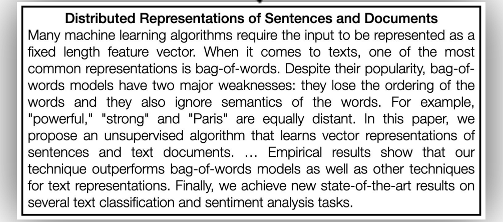
Various types of task activities have been studied in the literature as a basis for query suggestion or query support. Motivated by the observation that a notable propor-

Co-Citation Context



Learning deep structured semantic models for web search using clickthrough data

Latent semantic models, such as LSA, intend to map a query to its relevant documents at the semantic level where keyword-based matching often fails. In this study we strive to develop a series of new latent semantic models with a deep structure that project queries and documents into a common low-dimensional space where the relevance of a document given a query is readily computed as the distance between them. The proposed deep structured semantic models are discriminatively trained by maximizing the conditional likelihood of the clicked documents given a query using the clickthrough data. … Results show that our best model significantly outperforms other latent semantic models, which were considered state-of-the-art in the … to the work presented in this paper.

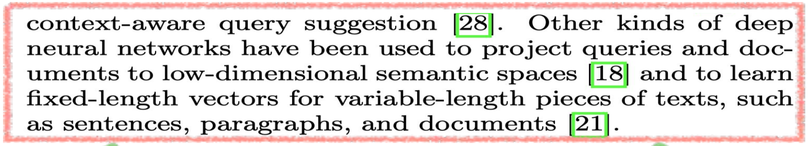


Distributed Representations of Sentences and Documents

Many machine learning algorithms require the input to be represented as a fixed length feature vector. When it comes to texts, one of the most common representations is bag-of-words. Despite their popularity, bag-of-words models have two major weaknesses: they lose the ordering of the words and they also ignore semantics of the words. For example, "powerful," "strong" and "Paris" are equally distant. In this paper, we propose an unsupervised algorithm that learns vector representations of sentences and text documents. … Empirical results show that our technique outperforms bag-of-words models as well as other techniques for text representations. Finally, we achieve new state-of-the-art results on several text classification and sentiment analysis tasks.

Co-Cited Abstracts

Textual supervision: co-citation contexts

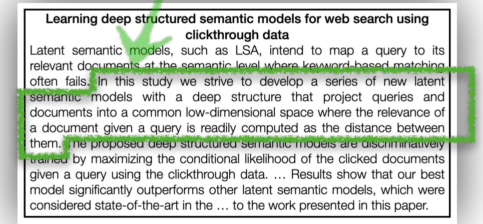


context-aware query suggestion [28]. Other kinds of deep neural networks have been used to project queries and documents to low-dimensional semantic spaces [18] and to learn fixed-length vectors for variable-length pieces of texts, such as sentences, paragraphs, and documents [21].

Co-Citation Context

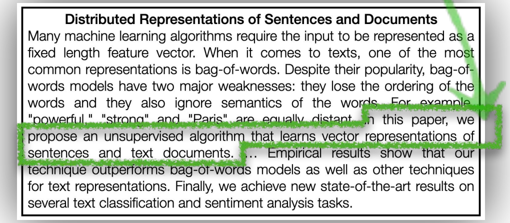
I. **Sentence alignment:**

Encode each sentence with BERT.
Find pair of sentences maximally similar
to the co-citation context (“aligned”).



Learning deep structured semantic models for web search using clickthrough data

Latent semantic models, such as LSA, intend to map a query to its relevant documents at the semantic level where keyword-based matching often fails. In this study we strive to develop a series of new latent semantic models with a deep structure that project queries and documents into a common low-dimensional space where the relevance of a document given a query is readily computed as the distance between them. The proposed deep structured semantic models are discriminatively trained by maximizing the conditional likelihood of the clicked documents given a query using the clickthrough data. … Results show that our best model significantly outperforms other latent semantic models, which were considered state-of-the-art in the … to the work presented in this paper.

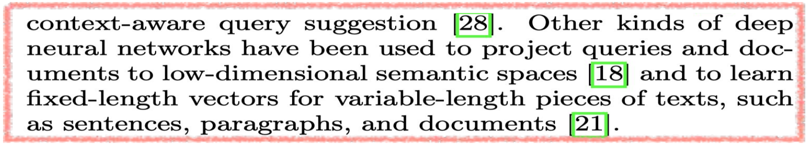


Distributed Representations of Sentences and Documents

Many machine learning algorithms require the input to be represented as a fixed length feature vector. When it comes to texts, one of the most common representations is bag-of-words. Despite their popularity, bag-of-words models have two major weaknesses: they lose the ordering of the words and they also ignore semantics of the words. For example "powerful," "strong" and "Paris" are equally distant. In this paper, we propose an unsupervised algorithm that learns vector representations of sentences and text documents. … Empirical results show that our technique outperforms bag-of-words models as well as other techniques for text representations. Finally, we achieve new state-of-the-art results on several text classification and sentiment analysis tasks.

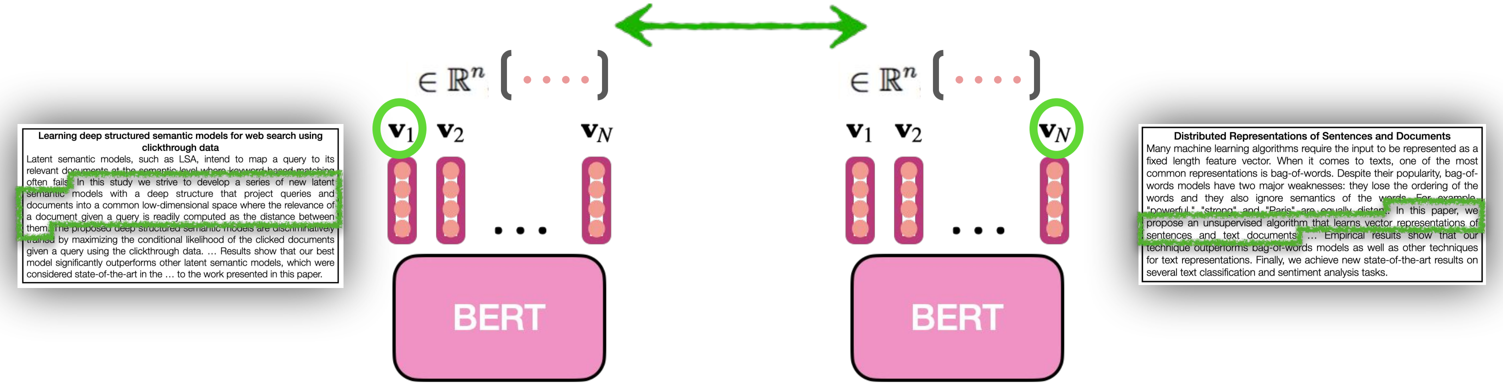
Co-Cited Abstracts

Textual supervision: co-citation contexts



context-aware query suggestion [28]. Other kinds of deep neural networks have been used to project queries and documents to low-dimensional semantic spaces [18] and to learn fixed-length vectors for variable-length pieces of texts, such as sentences, paragraphs, and documents [21].

II. Learn contextualized **sentence embeddings** that **minimize distance** between the aligned sentences (w/ contrastive loss)



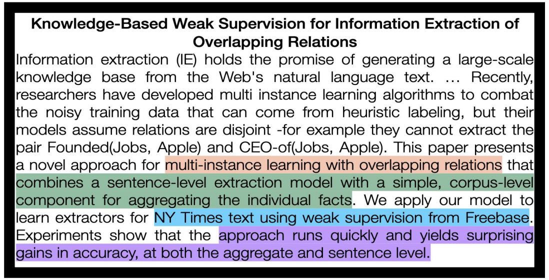
Learning deep structured semantic models for web search using clickthrough data

Latent semantic models, such as LSA, intend to map a query to its relevant documents at the semantic level where keyword-based matching often fails. In this study we strive to develop a series of new latent semantic models with a deep structure that project queries and documents into a common low-dimensional space where the relevance of a document given a query is readily computed as the distance between them. The proposed deep structured semantic models are discriminatively trained by maximizing the conditional likelihood of the clicked documents given a query using the clickthrough data. … Results show that our best model significantly outperforms other latent semantic models, which were considered state-of-the-art in the … to the work presented in this paper.

Distributed Representations of Sentences and Documents

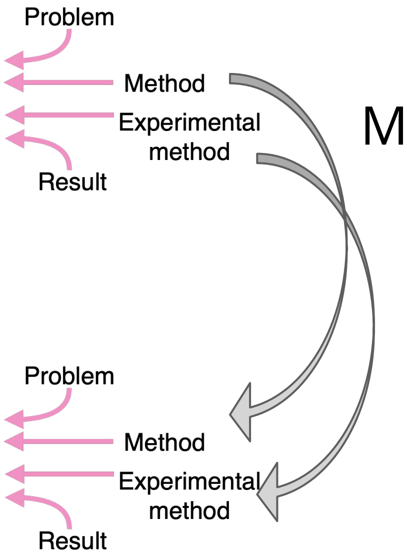
Many machine learning algorithms require the input to be represented as a fixed length feature vector. When it comes to texts, one of the most common representations is bag-of-words. Despite their popularity, bag-of-words models have two major weaknesses: they lose the ordering of the words and they also ignore semantics of the words. For example, "powerful," "strong" and "Paris" are equally distant. In this paper, we propose an unsupervised algorithm that learns vector representations of sentences and text documents … Empirical results show that our technique outperforms bag-of-words models as well as other techniques for text representations. Finally, we achieve new state-of-the-art results on several text classification and sentiment analysis tasks.

Multiple matches

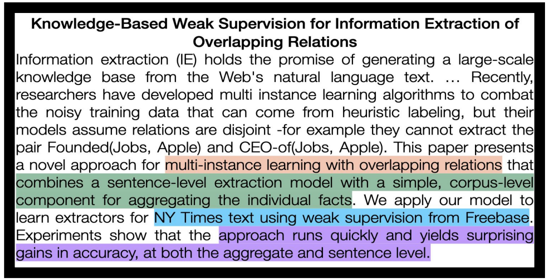


Knowledge-Based Weak Supervision for Information Extraction of Overlapping Relations

Information extraction (IE) holds the promise of generating a large-scale knowledge base from the Web's natural language text. … Recently, researchers have developed multi instance learning algorithms to combat the noisy training data that can come from heuristic labeling, but their models assume relations are disjoint -for example they cannot extract the pair Founded(Jobs, Apple) and CEO-of(Jobs, Apple). This paper presents a novel approach for multi-instance learning with overlapping relations that combines a sentence-level extraction model with a simple, corpus-level component for aggregating the individual facts. We apply our model to learn extractors for NY Times text using weak supervision from Freebase. Experiments show that the approach runs quickly and yields surprising gains in accuracy, at both the aggregate and sentence level.



Multiple aspect alignments



Knowledge-Based Weak Supervision for Information Extraction of Overlapping Relations

Information extraction (IE) holds the promise of generating a large-scale knowledge base from the Web's natural language text. … Recently, researchers have developed multi instance learning algorithms to combat the noisy training data that can come from heuristic labeling, but their models assume relations are disjoint -for example they cannot extract the pair Founded(Jobs, Apple) and CEO-of(Jobs, Apple). This paper presents a novel approach for multi-instance learning with overlapping relations that combines a sentence-level extraction model with a simple, corpus-level component for aggregating the individual facts. We apply our model to learn extractors for NY Times text using weak supervision from Freebase. Experiments show that the approach runs quickly and yields surprising gains in accuracy, at both the aggregate and sentence level.

$$f(p, p')$$

Distance between two **papers** p, p’

$$f(p, p') = \sum_{(i,i')\in\mathcal{A}_p\times\mathcal{A}_{p'}}$$

Distance between two **papers** p, p’

Sum over all pairs of aspects across p, p’

$$f(p, p') = \sum_{(i,i') \in \mathcal{A}_p \times \mathcal{A}_{p'}} w_{i,i'} \cdot d_{i,i'}$$

Distance between two **papers** p, p’

Distance between two **aspects** i, i’

Importance **weight** for
aspect pair i, i’

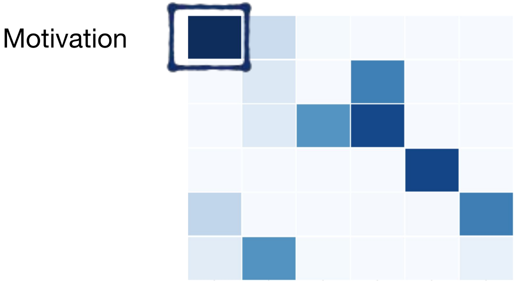
$$f(p, p') = \sum_{(i,i')\in\mathcal{A}_p\times\mathcal{A}_{p'}} w_{i,i'} \cdot d_{i,i'}$$

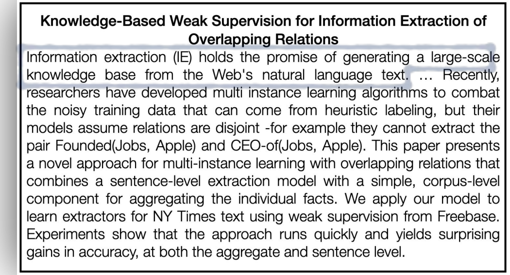
Distance between two
papers p, p’

Distance between two
aspects i, i’

$$w_{i,i'}$$

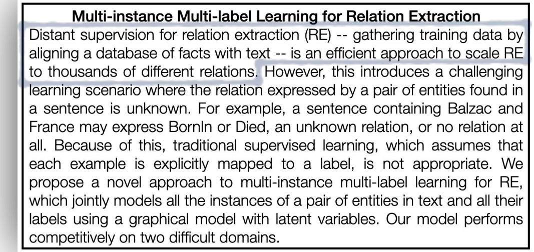
Aspect-level alignment weight matrix





Knowledge-Based Weak Supervision for Information Extraction of Overlapping Relations

Information extraction (IE) holds the promise of generating a large-scale knowledge base from the Web's natural language text. … Recently, researchers have developed multi instance learning algorithms to combat the noisy training data that can come from heuristic labeling, but their models assume relations are disjoint -for example they cannot extract the pair Founded(Jobs, Apple) and CEO-of(Jobs, Apple). This paper presents a novel approach for multi-instance learning with overlapping relations that combines a sentence-level extraction model with a simple, corpus-level component for aggregating the individual facts. We apply our model to learn extractors for NY Times text using weak supervision from Freebase. Experiments show that the approach runs quickly and yields surprising gains in accuracy, at both the aggregate and sentence level.

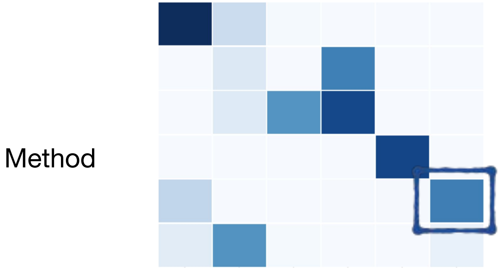


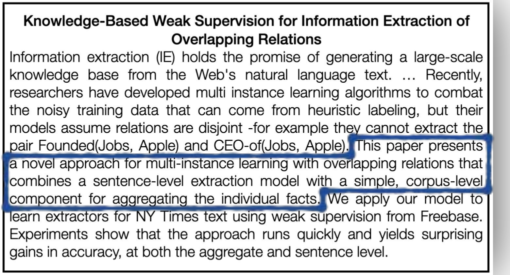
Multi-instance Multi-label Learning for Relation Extraction

Distant supervision for relation extraction (RE) -- gathering training data by aligning a database of facts with text -- is an efficient approach to scale RE to thousands of different relations. However, this introduces a challenging learning scenario where the relation expressed by a pair of entities found in a sentence is unknown. For example, a sentence containing Balzac and France may express BornIn or Died, an unknown relation, or no relation at all. Because of this, traditional supervised learning, which assumes that each example is explicitly mapped to a label, is not appropriate. We propose a novel approach to multi-instance multi-label learning for RE, which jointly models all the instances of a pair of entities in text and all their labels using a graphical model with latent variables. Our model performs competitively on two difficult domains.

$$w_{i,i'}$$

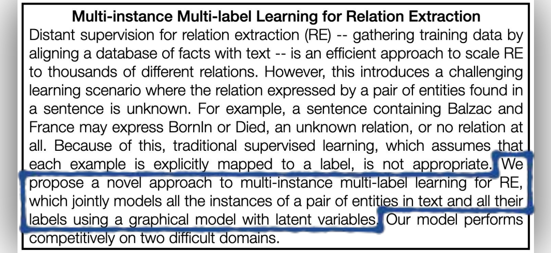
Aspect-level alignment weight matrix





Knowledge-Based Weak Supervision for Information Extraction of Overlapping Relations

Information extraction (IE) holds the promise of generating a large-scale knowledge base from the Web's natural language text. … Recently, researchers have developed multi instance learning algorithms to combat the noisy training data that can come from heuristic labeling, but their models assume relations are disjoint -for example they cannot extract the pair Founded(Jobs, Apple) and CEO-of(Jobs, Apple). This paper presents a novel approach for multi-instance learning with overlapping relations that combines a sentence-level extraction model with a simple, corpus-level component for aggregating the individual facts. We apply our model to learn extractors for NY Times text using weak supervision from Freebase. Experiments show that the approach runs quickly and yields surprising gains in accuracy, at both the aggregate and sentence level.

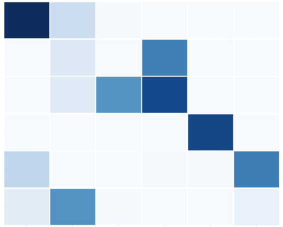


Multi-instance Multi-label Learning for Relation Extraction

Distant supervision for relation extraction (RE) -- gathering training data by aligning a database of facts with text -- is an efficient approach to scale RE to thousands of different relations. However, this introduces a challenging learning scenario where the relation expressed by a pair of entities found in a sentence is unknown. For example, a sentence containing Balzac and France may express BornIn or Died, an unknown relation, or no relation at all. Because of this, traditional supervised learning, which assumes that each example is explicitly mapped to a label, is not appropriate. We propose a novel approach to multi-instance multi-label learning for RE, which jointly models all the instances of a pair of entities in text and all their labels using a graphical model with latent variables. Our model performs competitively on two difficult domains.

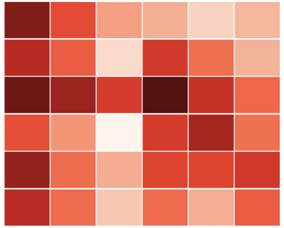
$$f(p, p') =$$

$w_{i,i'}$



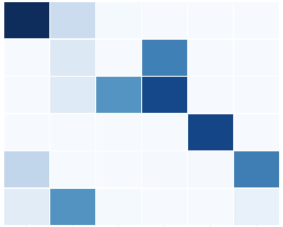
$$*$$

$d_{i,i'}$



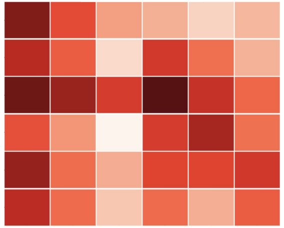
$$f(p, p') =$$

$w_{i,i'}$



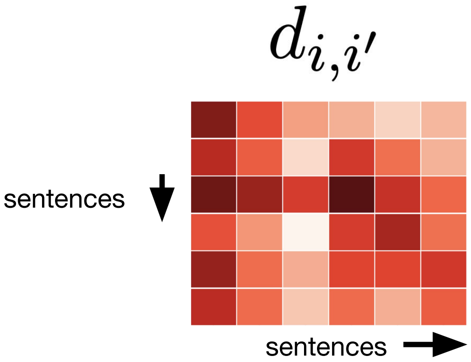
$*$

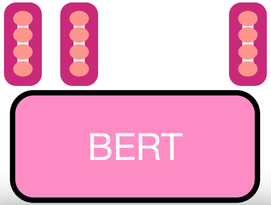
$d_{i,i'}$

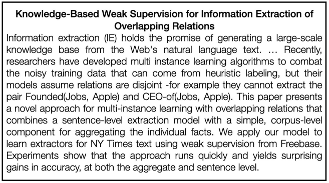


Learn aspect weights and distances such that
the document-level **contrastive loss** is minimized

$$\boxed{\mathcal{L}_f(p, p', n) = \texttt{max}[f(p, p') - f(p, n) + m, 0]}$$

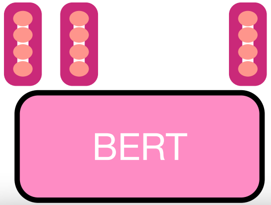


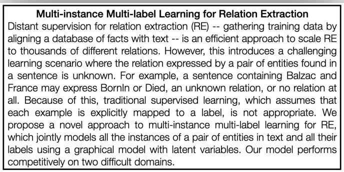




Knowledge-Based Weak Supervision for Information Extraction of Overlapping Relations

Information extraction (IE) holds the promise of generating a large-scale knowledge base from the Web's natural language text. … Recently, researchers have developed multi instance learning algorithms to combat the noisy training data that can come from heuristic labeling, but their models assume relations are disjoint -for example they cannot extract the pair Founded(Jobs, Apple) and CEO-of(Jobs, Apple). This paper presents a novel approach for multi-instance learning with overlapping relations that combines a sentence-level extraction model with a simple, corpus-level component for aggregating the individual facts. We apply our model to learn extractors for NY Times text using weak supervision from Freebase. Experiments show that the approach runs quickly and yields surprising gains in accuracy, at both the aggregate and sentence level.



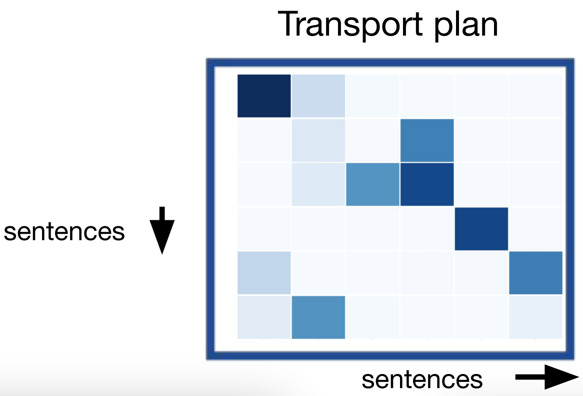


Multi-instance Multi-label Learning for Relation Extraction

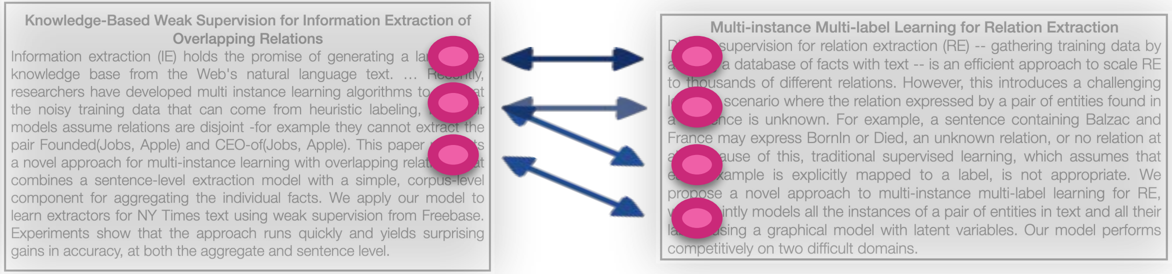
Distant supervision for relation extraction (RE) -- gathering training data by aligning a database of facts with text -- is an efficient approach to scale RE to thousands of different relations. However, this introduces a challenging learning scenario where the relation expressed by a pair of entities found in a sentence is unknown. For example, a sentence containing Balzac and France may express BornIn or Died, an unknown relation, or no relation at all. Because of this, traditional supervised learning, which assumes that each example is explicitly mapped to a label, is not appropriate. We propose a novel approach to multi-instance multi-label learning for RE, which jointly models all the instances of a pair of entities in text and all their labels using a graphical model with latent variables. Our model performs competitively on two difficult domains.

Optimal Transport:

Soft, sparse alignment between sets of aspects



$$\sum_{i \in p} \sum_{j \in p'} \hat{\mathbf{P}}[i,j] = 1$$

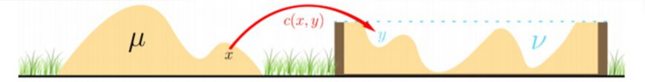


Knowledge-Based Weak Supervision for Information Extraction of Overlapping Relations

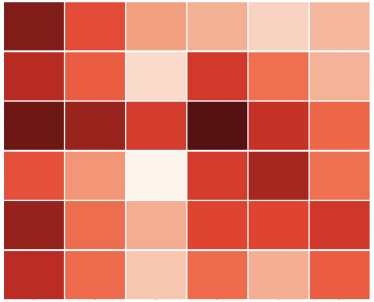
Information extraction (IE) holds the promise of generating a lar e knowledge base from the Web's natural language text. … Recently, researchers have developed multi instance learning algorithms to at the noisy training data that can come from heuristic labeling, ir models assume relations are disjoint -for example they cannot extract the pair Founded(Jobs, Apple) and CEO-of(Jobs, Apple). This paper ts a novel approach for multi-instance learning with overlapping relat at combines a sentence-level extraction model with a simple, corpus-level component for aggregating the individual facts. We apply our model to learn extractors for NY Times text using weak supervision from Freebase. Experiments show that the approach runs quickly and yields surprising gains in accuracy, at both the aggregate and sentence level.

Multi-instance Multi-label Learning for Relation Extraction

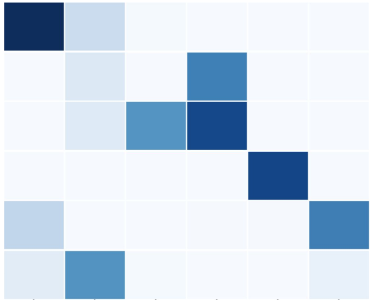
D supervision for relation extraction (RE) -- gathering training data by a a database of facts with text -- is an efficient approach to scale RE to thousands of different relations. However, this introduces a challenging l scenario where the relation expressed by a pair of entities found in a nce is unknown. For example, a sentence containing Balzac and France may express BornIn or Died, an unknown relation, or no relation at a ause of this, traditional supervised learning, which assumes that e xample is explicitly mapped to a label, is not appropriate. We propose a novel approach to multi-instance multi-label learning for RE, v intly models all the instances of a pair of entities in text and all their la sing a graphical model with latent variables. Our model performs competitively on two difficult domains.



Pairwise distances



Transport plan



$$\boxed{\hat{\mathbf{P}} = \underset{\mathbf{P}\in\mathcal{S}}{\texttt{argmin}}\langle \mathbf{D},\mathbf{P}\rangle}$$

Linear optimization problem, compatible with autodiff, GPUs

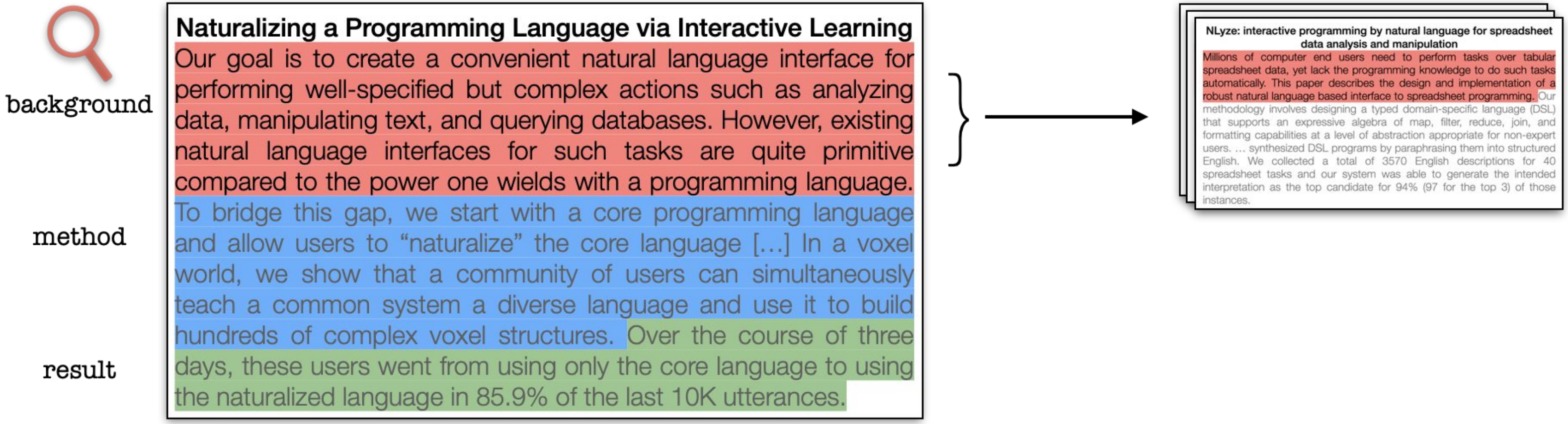
Evaluation

- Document-level similarity
    - Two biomedical paper datasets

Evaluation

- Document-level similarity
  - Two biomedical paper datasets

- Aspect-level similarity
  - Recent CS paper dataset



Naturalizing a Programming Language via Interactive Learning

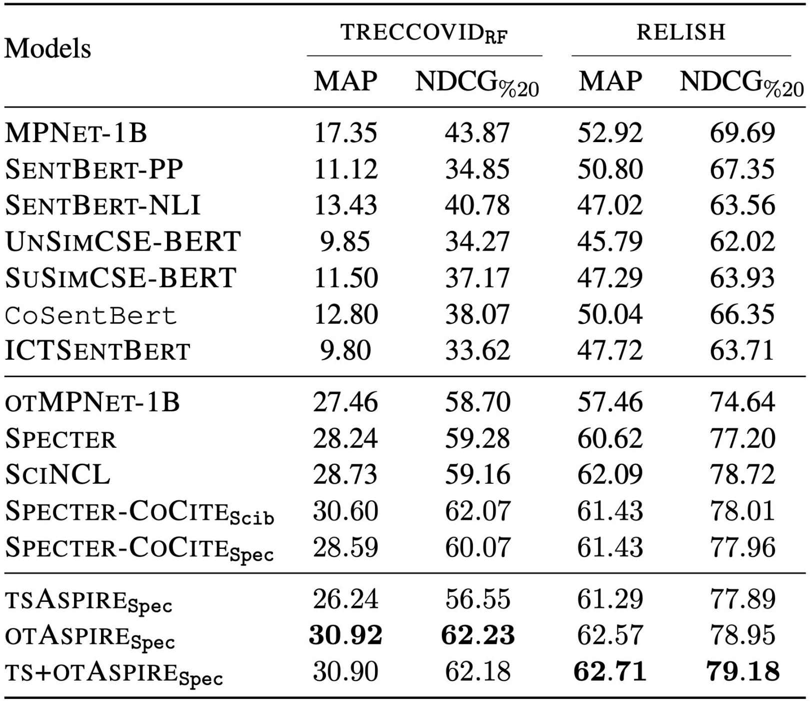
Our goal is to create a convenient natural language interface for performing well-specified but complex actions such as analyzing data, manipulating text, and querying databases. However, existing natural language interfaces for such tasks are quite primitive compared to the power one wields with a programming language. To bridge this gap, we start with a core programming language and allow users to "naturalize" the core language [...] In a voxel world, we show that a community of users can simultaneously teach a common system a diverse language and use it to build hundreds of complex voxel structures. Over the course of three days, these users went from using only the core language to using the naturalized language in 85.9% of the last 10K utterances.

NLyze: interactive programming by natural language for spreadsheet data analysis and manipulation

Millions of computer end users need to perform tasks over tabular spreadsheet data, yet lack the programming knowledge to do such tasks automatically. This paper describes the design and implementation of a robust natural language based interface to spreadsheet programming. Our methodology involves designing a typed domain-specific language (DSL) that supports an expressive algebra of map, filter, reduce, join, and formatting capabilities at a level of abstraction appropriate for non-expert users. … synthesized DSL programs by paraphrasing them into structured English. We collected a total of 3570 English descriptions for 40 spreadsheet tasks and our system was able to generate the intended interpretation as the top candidate for 94% (97 for the top 3) of those instances.

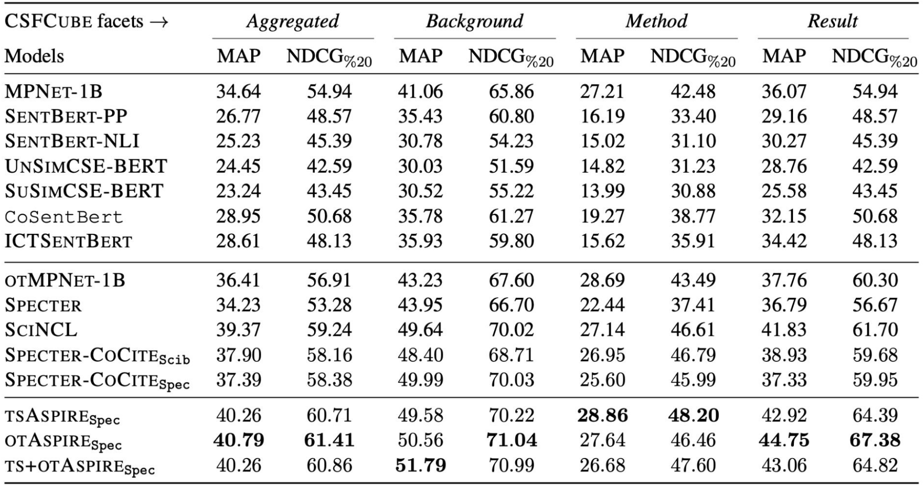
Evaluation

Document level



Models	TRECCOVID$_{\text{RF}}$		RELISH	
	MAP	NDCG$_{\%20}$	MAP	NDCG$_{\%20}$
MPNET-1B	17.35	43.87	52.92	69.69
SENTBERT-PP	11.12	34.85	50.80	67.35
SENTBERT-NLI	13.43	40.78	47.02	63.56
UNSIMCSE-BERT	9.85	34.27	45.79	62.02
SUSIMCSE-BERT	11.50	37.17	47.29	63.93
CoSentBert	12.80	38.07	50.04	66.35
ICTSENTBERT	9.80	33.62	47.72	63.71
OTMPNET-1B	27.46	58.70	57.46	74.64
SPECTER	28.24	59.28	60.62	77.20
SCINCL	28.73	59.16	62.09	78.72
SPECTER-COCITE$_{\text{Scib}}$	30.60	62.07	61.43	78.01
SPECTER-COCITE$_{\text{Spec}}$	28.59	60.07	61.43	77.96
TSASPIRE$_{\text{Spec}}$	26.24	56.55	61.29	77.89
OTASPIRE$_{\text{Spec}}$	**30.92**	**62.23**	62.57	78.95
TS+OTASPIRE$_{\text{Spec}}$	30.90	62.18	**62.71**	**79.18**

Aspect level



CSFCUBE facets →	*Aggregated*		*Background*		*Method*		*Result*	
Models	MAP	NDCG$_{\%20}$	MAP	NDCG$_{\%20}$	MAP	NDCG$_{\%20}$	MAP	NDCG$_{\%20}$
MPNET-1B	34.64	54.94	41.06	65.86	27.21	42.48	36.07	54.94
SENTBERT-PP	26.77	48.57	35.43	60.80	16.19	33.40	29.16	48.57
SENTBERT-NLI	25.23	45.39	30.78	54.23	15.02	31.10	30.27	45.39
UNSIMCSE-BERT	24.45	42.59	30.03	51.59	14.82	31.23	28.76	42.59
SUSIMCSE-BERT	23.24	43.45	30.52	55.22	13.99	30.88	25.58	43.45
CoSentBert	28.95	50.68	35.78	61.27	19.27	38.77	32.15	50.68
ICTSENTBERT	28.61	48.13	35.93	59.80	15.62	35.91	34.42	48.13
OTMPNET-1B	36.41	56.91	43.23	67.60	28.69	43.49	37.76	60.30
SPECTER	34.23	53.28	43.95	66.70	22.44	37.41	36.79	56.67
SCINCL	39.37	59.24	49.64	70.02	27.14	46.61	41.83	61.70
SPECTER-COCITE$_{\text{Scib}}$	37.90	58.16	48.40	68.71	26.95	46.79	38.93	59.68
SPECTER-COCITE$_{\text{Spec}}$	37.39	58.38	49.99	70.03	25.60	45.99	37.33	59.95
TSASPIRE$_{\text{Spec}}$	40.26	60.71	49.58	70.22	**28.86**	**48.20**	42.92	64.39
OTASPIRE$_{\text{Spec}}$	**40.79**	**61.41**	50.56	**71.04**	27.64	46.46	**44.75**	**67.38**
TS+OTASPIRE$_{\text{Spec}}$	40.26	60.86	**51.79**	70.99	26.68	47.60	43.06	64.82

Substantial gains in MAP, NDCG

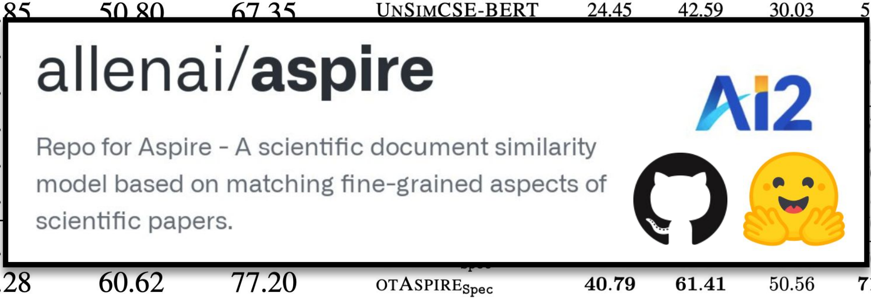
Evaluation

Document level

Models	TRECCOVID$_{RF}$		RELISH	
	MAP	NDCG$_{\%20}$	MAP	NDCG$_{\%20}$
MPNet-1B	17.35	43.87	52.92	69.69
SentBert-PP	11.12	34.85	50.80	67.35
SentBert-NLI	13.43	40.	[illegible]	[illegible]
UnSimCSE-BERT	9.85	34.	[illegible]	[illegible]
SuSimCSE-BERT	11.50	37.	[illegible]	[illegible]
CoSentBert	12.80	38.	[illegible]	[illegible]
ICTSentBert	9.80	33.	[illegible]	[illegible]
otMPNet-1B	27.46	58.	[illegible]	[illegible]
SPECTER	28.24	59.28	60.62	77.20
SciNCL	28.73	59.16	62.09	78.72
SPECTER-CoCite$_{Scib}$	30.60	62.07	61.43	78.01
SPECTER-CoCite$_{Spec}$	28.59	60.07	61.43	77.96
tsAspire$_{Spec}$	26.24	56.55	61.29	77.89
otAspire$_{Spec}$	**30.92**	**62.23**	62.57	78.95
ts+otAspire$_{Spec}$	30.90	62.18	**62.71**	**79.18**

Aspect level

CSFCube facets →	*Aggregated*		*Background*		*Method*		*Result*	
Models	MAP	NDCG$_{\%20}$	MAP	NDCG$_{\%20}$	MAP	NDCG$_{\%20}$	MAP	NDCG$_{\%20}$
MPNet-1B	34.64	54.94	41.06	65.86	27.21	42.48	36.07	54.94
SentBert-PP	26.77	48.57	35.43	60.80	16.19	33.40	29.16	48.57
SentBert-NLI	25.23	45.39	30.78	54.23	15.02	31.10	30.27	45.39
UnSimCSE-BERT	24.45	42.59	30.03	51.59	14.82	31.23	28.76	42.59
[illegible]	[illegible]	[illegible]	[illegible]	5.22	13.99	30.88	25.58	43.45
[illegible]	[illegible]	[illegible]	[illegible]	1.27	19.27	38.77	32.15	50.68
[illegible]	[illegible]	[illegible]	[illegible]	9.80	15.62	35.91	34.42	48.13
[illegible]	[illegible]	[illegible]	[illegible]	7.60	28.69	43.49	37.76	60.30
[illegible]	[illegible]	[illegible]	[illegible]	6.70	22.44	37.41	36.79	56.67
[illegible]	[illegible]	[illegible]	[illegible]	0.02	27.14	46.61	41.83	61.70
[illegible]	[illegible]	[illegible]	[illegible]	8.71	26.95	46.79	38.93	59.68
[illegible]	[illegible]	[illegible]	[illegible]	0.03	25.60	45.99	37.33	59.95
[illegible]	[illegible]	[illegible]	[illegible]	0.22	**28.86**	**48.20**	42.92	64.39
otAspire$_{Spec}$	**40.79**	**61.41**	50.56	**71.04**	27.64	46.46	**44.75**	**67.38**
ts+otAspire$_{Spec}$	40.26	60.86	**51.79**	70.99	26.68	47.60	43.06	64.82

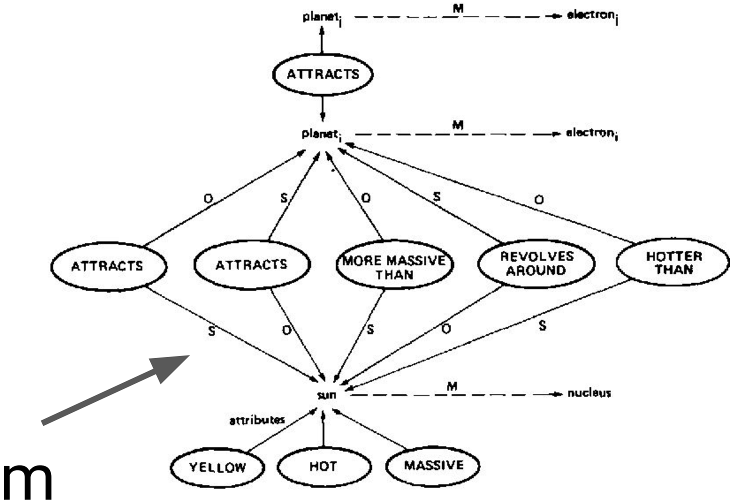

allenai/**aspire**

Repo for Aspire - A scientific document similarity model based on matching fine-grained aspects of scientific papers.

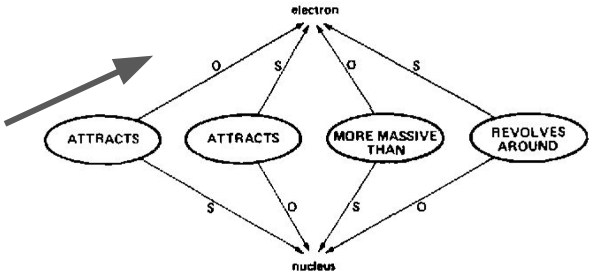
Substantial gains in MAP, NDCG

Analogical retrieval for finding cross-domain inspirations

Solar System



Atom

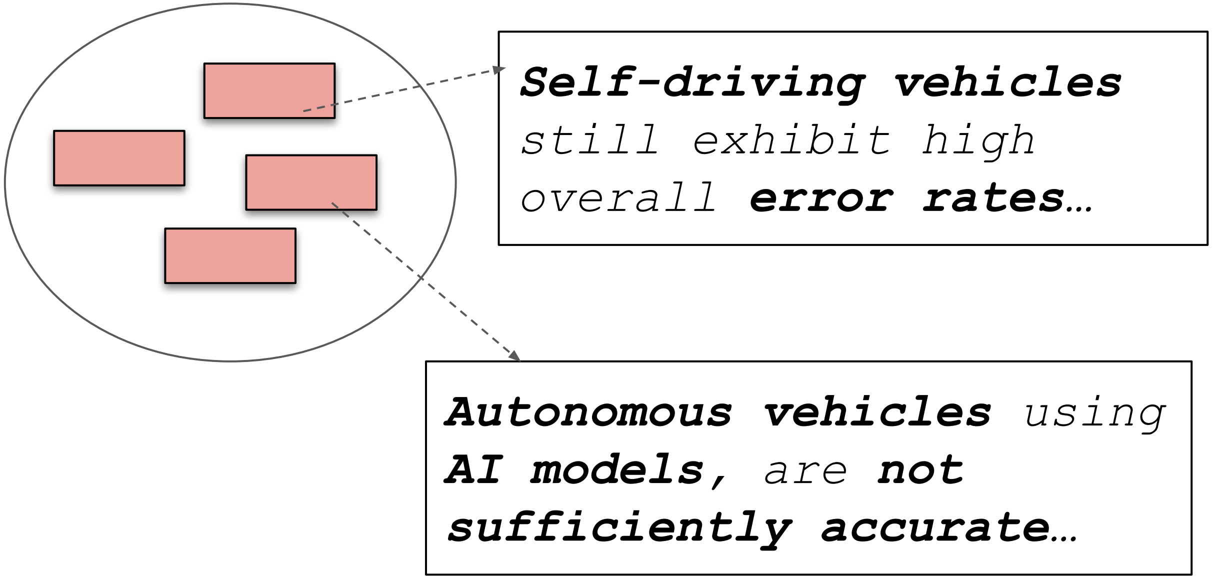


Best Research Paper

KDD

CHI22 New Orleans, LA

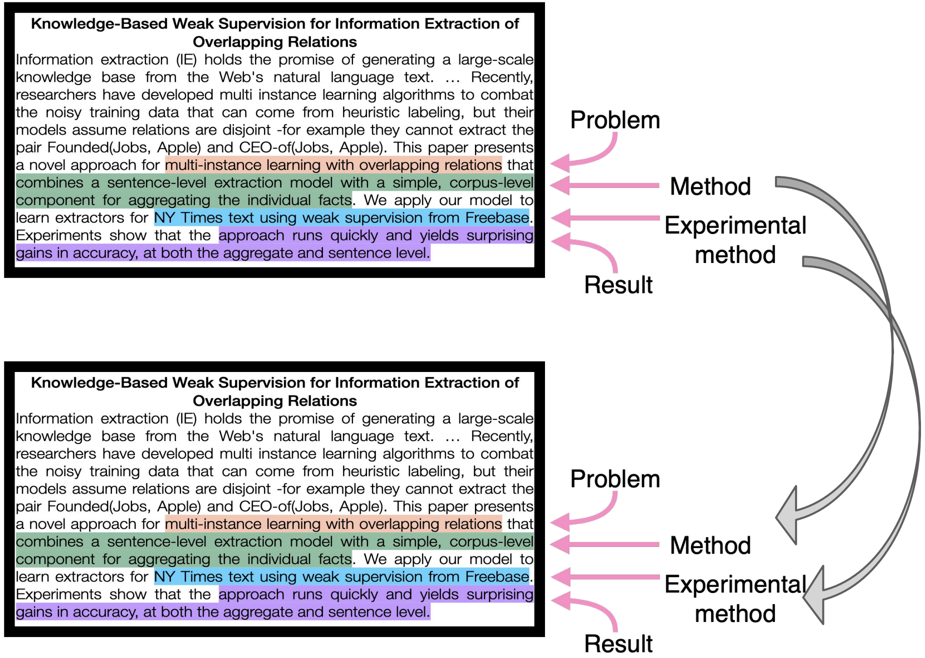
Cross-document coreference
+
Document Level Similarity?



Self-driving vehicles *still exhibit high overall* ***error rates…***

Autonomous vehicles *using* ***AI models,*** *are* ***not sufficiently accurate…***

Document Similarity & Retrieval with Aspect Alignments



Knowledge-Based Weak Supervision for Information Extraction of Overlapping Relations

Information extraction (IE) holds the promise of generating a large-scale knowledge base from the Web's natural language text. … Recently, researchers have developed multi instance learning algorithms to combat the noisy training data that can come from heuristic labeling, but their models assume relations are disjoint -for example they cannot extract the pair Founded(Jobs, Apple) and CEO-of(Jobs, Apple). This paper presents a novel approach for multi-instance learning with overlapping relations that combines a sentence-level extraction model with a simple, corpus-level component for aggregating the individual facts. We apply our model to learn extractors for NY Times text using weak supervision from Freebase. Experiments show that the approach runs quickly and yields surprising gains in accuracy, at both the aggregate and sentence level.

Problem

Method

Experimental method

Result

Knowledge-Based Weak Supervision for Information Extraction of Overlapping Relations

Information extraction (IE) holds the promise of generating a large-scale knowledge base from the Web's natural language text. … Recently, researchers have developed multi instance learning algorithms to combat the noisy training data that can come from heuristic labeling, but their models assume relations are disjoint -for example they cannot extract the pair Founded(Jobs, Apple) and CEO-of(Jobs, Apple). This paper presents a novel approach for multi-instance learning with overlapping relations that combines a sentence-level extraction model with a simple, corpus-level component for aggregating the individual facts. We apply our model to learn extractors for NY Times text using weak supervision from Freebase. Experiments show that the approach runs quickly and yields surprising gains in accuracy, at both the aggregate and sentence level.

Problem

Method

Experimental method

Result

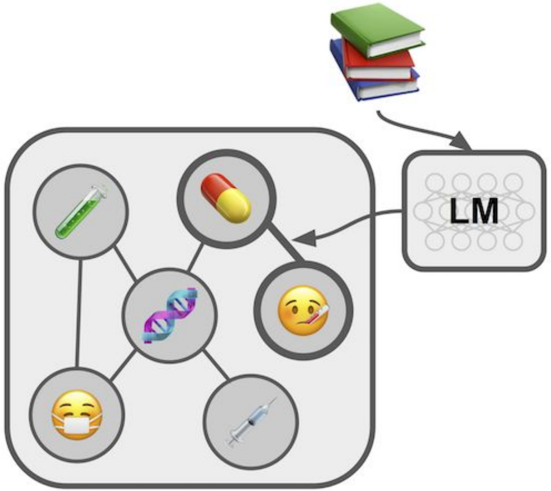
Best Research Paper KDD CHI22 New Orleans, LA NAACL 2022

From Retrieving & Extracting **Existing Knowledge**…

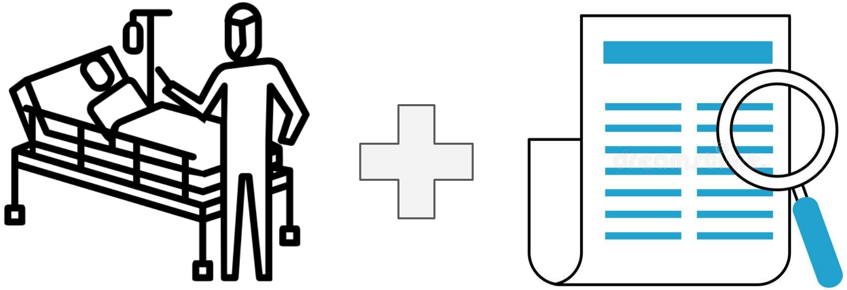
…To Predicting **New Knowledge**

1. SciCo: Hierarchical Cross-Document Coreference
2. Document Similarity & Retrieval
3. **Literature-Augmented Prediction**

Can **neural language models trained on biomedical corpora** (e.g., PubMed) be leveraged for predicting new links in biomedical knowledge graphs?

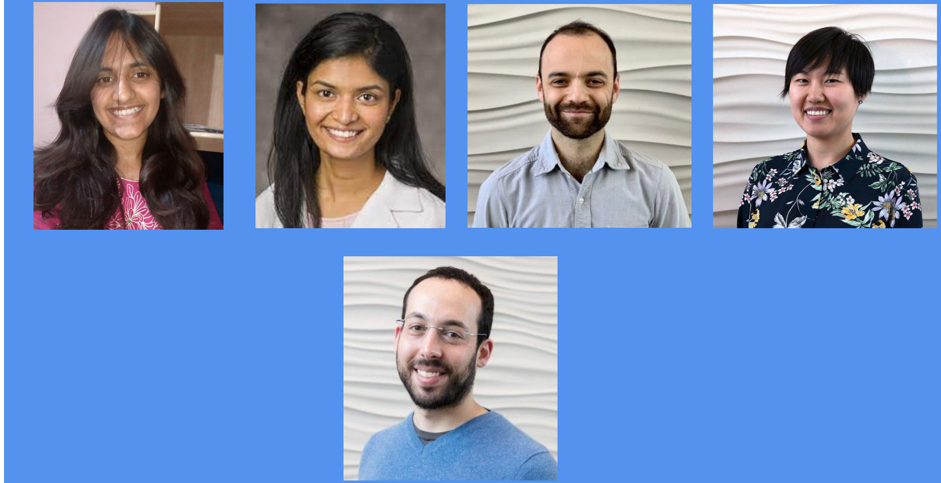


Can we enhance prediction of **clinical outcomes in hospital patients** by retrieving **patient-specific medical literature?**



Literature-Augmented Clinical Outcome Prediction

NAACL 2022



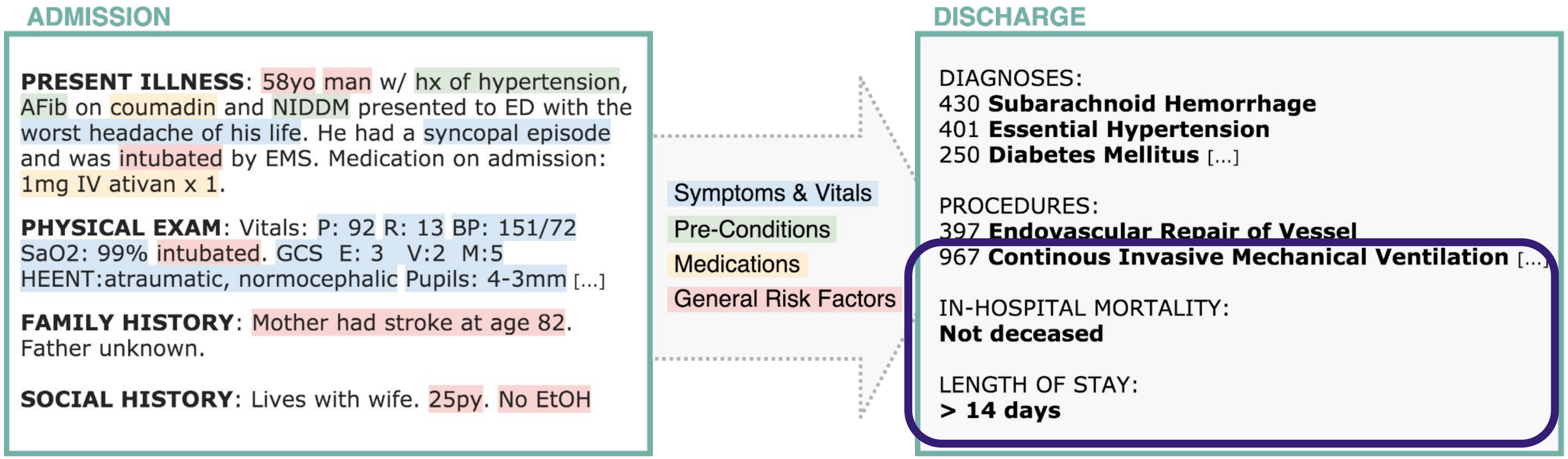
Aakanksha Naik, Sravanthi Parasa, Sergey Feldman, Lucy Lu Wang & **Tom Hope**

AI2

Predict clinical outcomes of ICU patients:

In-hospital mortality,
Prolonged mechanical ventilation,
Length of hospital stay,
...

Predict clinical outcomes of ICU patients:



ADMISSION

PRESENT ILLNESS: 58yo man w/ hx of hypertension, AFib on coumadin and NIDDM presented to ED with the worst headache of his life. He had a syncopal episode and was intubated by EMS. Medication on admission: 1mg IV ativan x 1.

PHYSICAL EXAM: Vitals: P: 92 R: 13 BP: 151/72 SaO2: 99% intubated. GCS E: 3 V:2 M:5 HEENT:atraumatic, normocephalic Pupils: 4-3mm [...]

FAMILY HISTORY: Mother had stroke at age 82. Father unknown.

SOCIAL HISTORY: Lives with wife. 25py. No EtOH

Symptoms & Vitals

Pre-Conditions

Medications

General Risk Factors

DISCHARGE

DIAGNOSES:
430 **Subarachnoid Hemorrhage**
401 **Essential Hypertension**
250 **Diabetes Mellitus** [...]

PROCEDURES:
397 **Endovascular Repair of Vessel**
967 **Continous Invasive Mechanical Ventilation** [...]

IN-HOSPITAL MORTALITY:
Not deceased

LENGTH OF STAY:
> 14 days

MIMIC

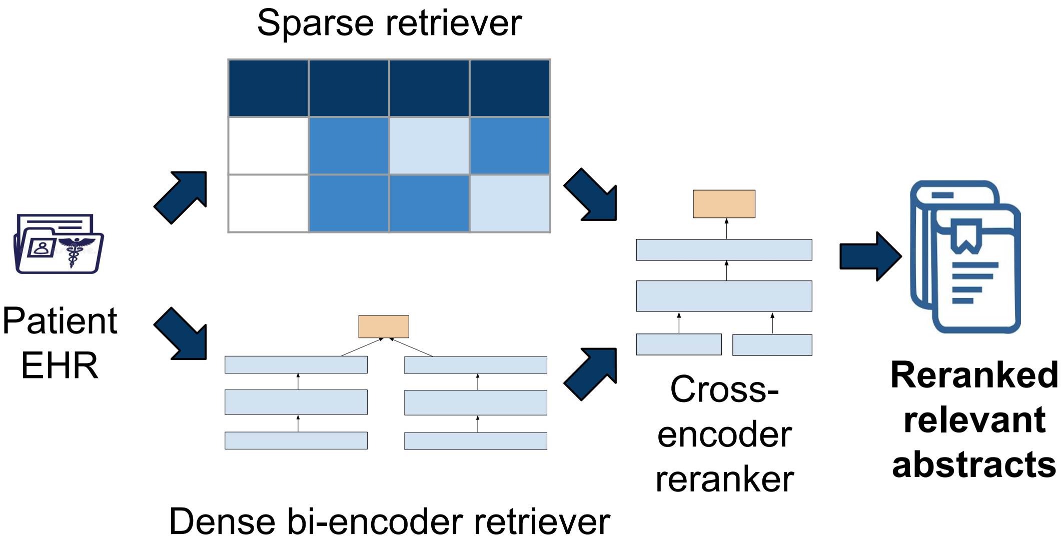
https://mimic.mit.edu/

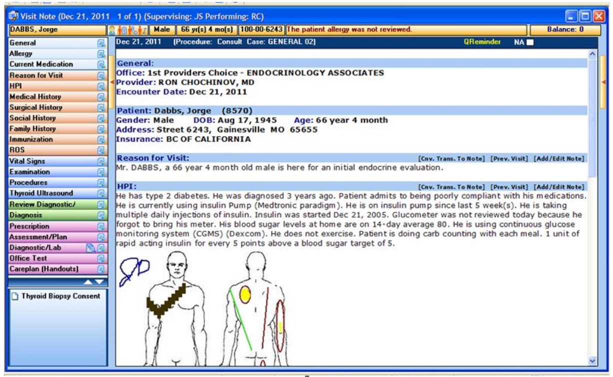
State-of-art models predict outcomes
from internal data (e.g., patient notes)

State-of-art models predict outcomes
from internal data (e.g., patient notes)

Can we improve performance by **adding patient-specific evidence from the literature?**

1. Retrieve patient-relevant literature





Visit Note (Dec 21, 2011 1 of 1) (Supervising: JS Performing: RC)

DABBS, Jorge | Male | 66 yr(s) 4 mo(s) | 100-00-6243 | The patient allergy was not reviewed. | Balance: 0

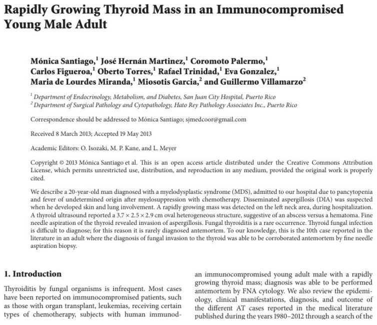
Dec 21, 2011 (Procedure: Consult Case: GENERAL 02) QReminder NA □

General:
Office: 1st Providers Choice - ENDOCRINOLOGY ASSOCIATES
Provider: RON CHOCHINOV, MD
Encounter Date: Dec 21, 2011

Patient: Dabbs, Jorge (8570)
Gender: Male DOB: Aug 17, 1945 Age: 66 year 4 month
Address: Street 6243, Gainesville MO 65655
Insurance: BC OF CALIFORNIA

Reason for Visit: [Cnv. Trans. To Note] [Prev. Visit] [Add/Edit Note]
Mr. DABBS, a 66 year 4 month old male is here for an initial endocrine evaluation.

HPI: [Cnv. Trans. To Note] [Prev. Visit] [Add/Edit Note]
He has type 2 diabetes. He was diagnosed 3 years ago. Patient admits to being poorly compliant with his medications. He is currently using insulin Pump (Medtronic paradigm). He is on insulin pump since last 5 week(s). He is taking multiple daily injections of insulin. Insulin was started Dec 21, 2005. Glucometer was not reviewed today because he forgot to bring his meter. His blood sugar levels at home are on 14-day average 80. He is using continuous glucose monitoring system (CGMS) (Dexcom). He does not exercise. Patient is doing carb counting with each meal. 1 unit of rapid acting insulin for every 5 points above a blood sugar target of 5.



Rapidly Growing Thyroid Mass in an Immunocompromised Young Male Adult

Mónica Santiago,[1] José Hernán Martinez,[1] Coromoto Palermo,[1] Carlos Figueroa,[1] Oberto Torres,[1] Rafael Trinidad,[1] Eva Gonzalez,[1] Maria de Lourdes Miranda,[1] Miosotis Garcia,[2] and Guillermo Villamarzo[2]

[1] *Department of Endocrinology, Metabolism, and Diabetes, San Juan City Hospital, Puerto Rico*
[2] *Department of Surgical Pathology and Cytopathology, Hato Rey Pathology Associates Inc., Puerto Rico*

Correspondence should be addressed to Mónica Santiago; sjmedcoor@gmail.com

Received 8 March 2013; Accepted 19 May 2013

Academic Editors: O. Isozaki, M. P. Kane, and L. Meyer

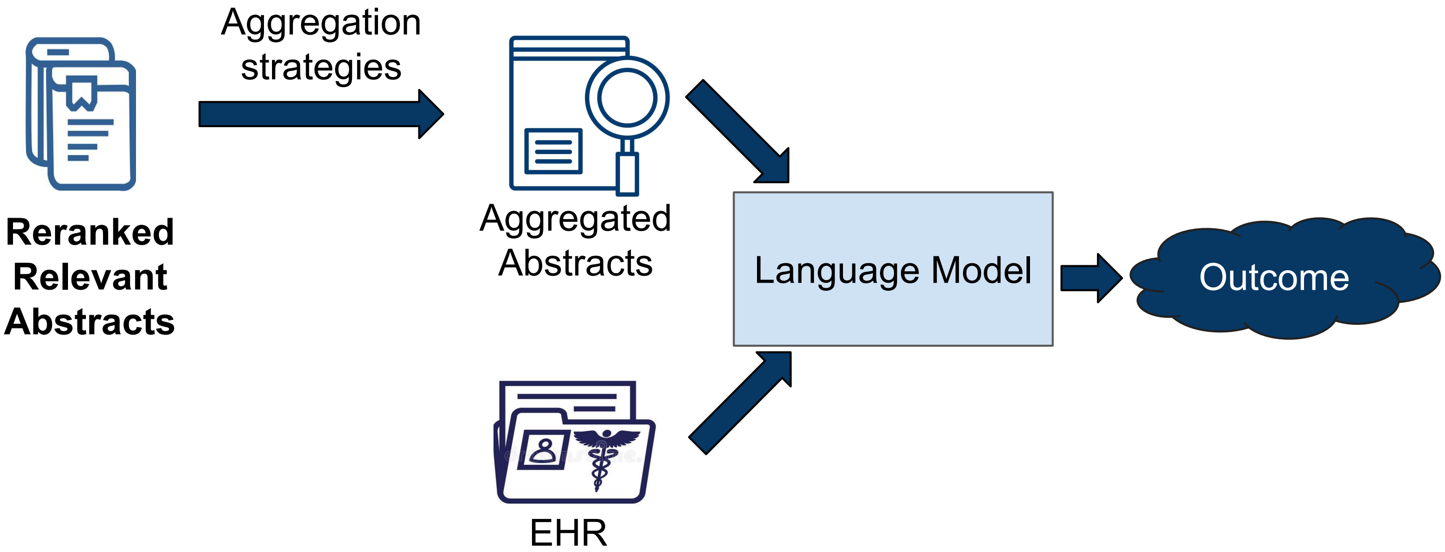
We describe a 20-year-old man diagnosed with a myelodysplastic syndrome (MDS), admitted to our hospital due to pancytopenia and fever of undetermined origin after myelosuppression with chemotherapy. Disseminated aspergillosis (DIA) was suspected when he developed skin and lung involvement. A rapidly growing mass was detected on the left neck area, during hospitalization. A thyroid ultrasound reported a 3.7 × 2.5 × 2.9 cm oval heterogeneous structure, suggestive of an abscess versus a hematoma. Fine needle aspiration of the thyroid revealed invasion of aspergillosis. Fungal thyroiditis is a rare occurrence. Thyroid fungal infection is difficult to diagnose; for this reason it is rarely diagnosed antemortem. To our knowledge, this is the 10th case reported in the literature in an adult where the diagnosis of fungal invasion to the thyroid was able to be corroborated antemortem by fine needle aspiration biopsy.

1. Introduction

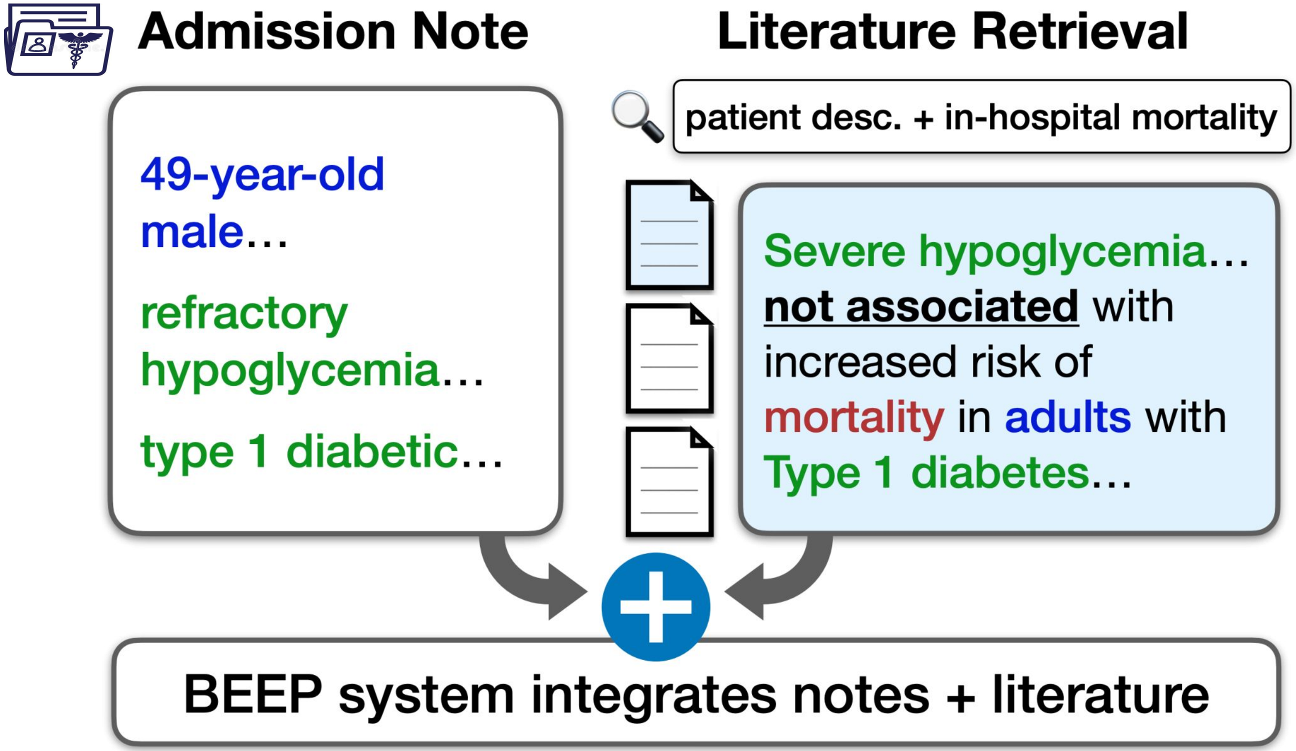
Thyroiditis by fungal organisms is infrequent. Most cases have been reported on immunocompromised patients, such as those with organ transplant, leukemias, receiving certain types of chemotherapy, subjects with human immunod-

an immunocompromised young adult male with a rapidly growing thyroid mass; diagnosis was able to be performed antemortem by FNA cytology. We also review the epidemiology, clinical manifestations, diagnosis, and outcome of the different AT cases reported in the medical literature published during the years 1980–2012 through a search of the

2. Literature-enhanced outcome prediction



BEEP: Biomedical Evidence-Enhanced Predictions



Admission Note

49-year-old male…

refractory hypoglycemia…

type 1 diabetic…

Literature Retrieval

patient desc. + in-hospital mortality

Severe hypoglycemia… **not associated** with increased risk of mortality in adults with Type 1 diabetes…

BEEP system integrates notes + literature

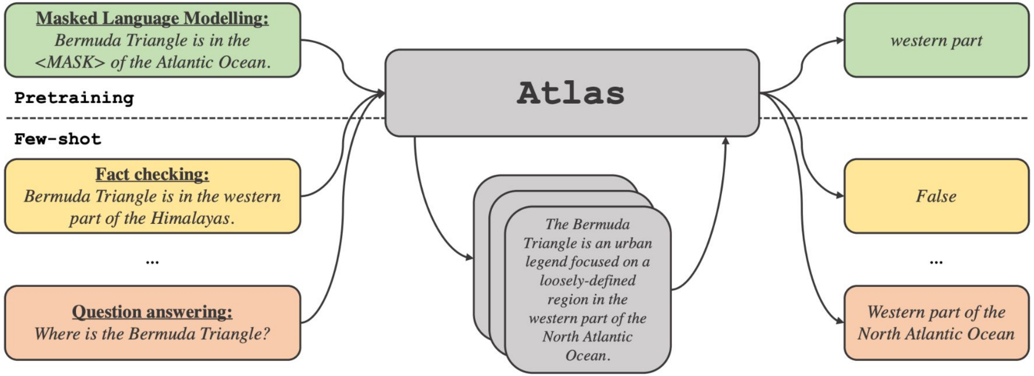
More accurate prediction: **Patient survives!**

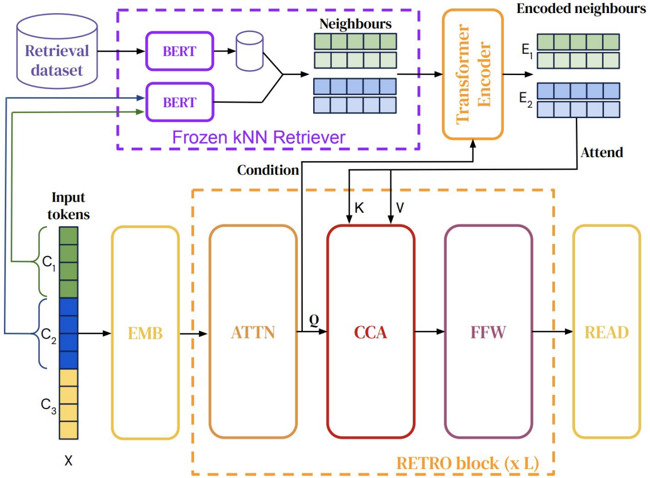
Adding literature boosts outcome prediction:

Up to **5 point increase** in overall F1/AUC scores

Over **25% increase** in precision@Top-K scores

Retrieval-augmented language models for clinical outcome prediction?

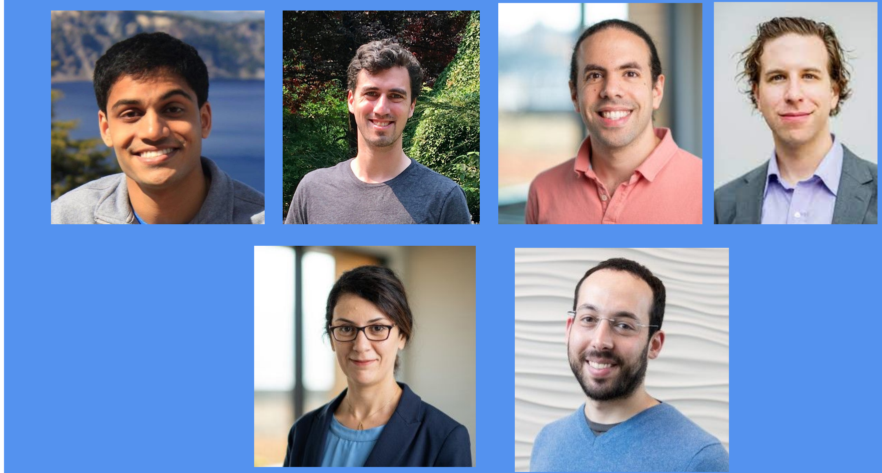

Masked Language Modelling:
Bermuda Triangle is in the <MASK> of the Atlantic Ocean.
Pretraining
Few-shot
Fact checking:
Bermuda Triangle is in the western part of the Himalayas.
...
Question answering:
Where is the Bermuda Triangle?
Atlas
The Bermuda Triangle is an urban legend focused on a loosely-defined region in the western part of the North Atlantic Ocean.
western part
False
...
Western part of the North Atlantic Ocean



Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, Edouard Grave, August 2022

Scientific Language Models for Biomedical Knowledge Base Completion

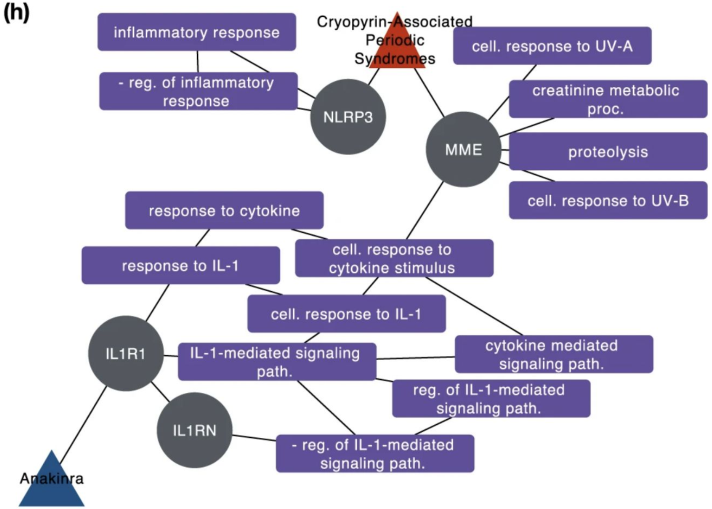
AUTOMATED KNOWLEDGE BASE 2021 | irvine CONSTRUCTION

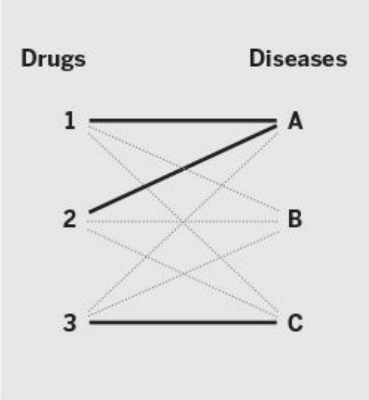


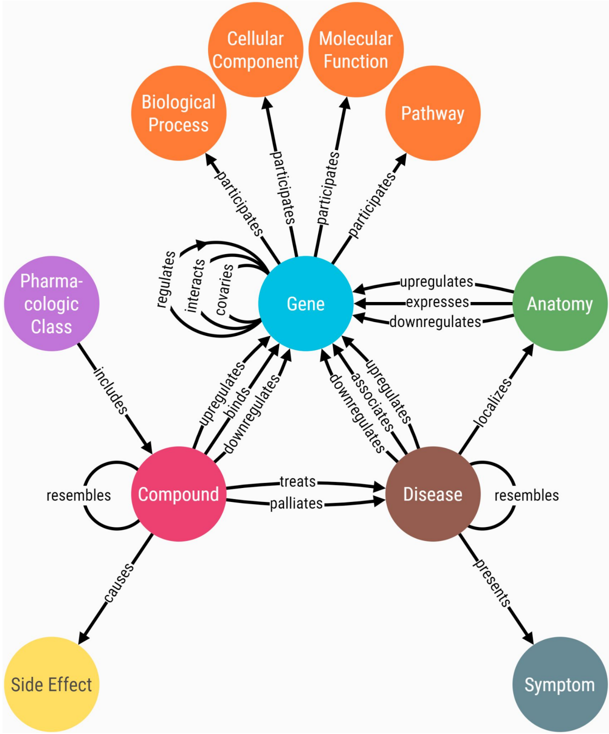
Rahul Nadkarni, David Wadden, Iz Beltagy, Noah A. Smith, Hannaneh Hajishirzi and **Tom Hope**

AI2

KG Completion for Drug Discovery

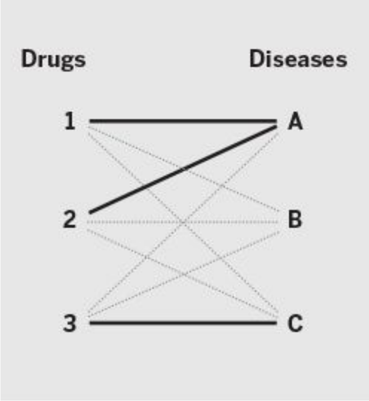

(h)
inflammatory response
- reg. of inflammatory response
Cryopyrin-Associated Periodic Syndromes
NLRP3
cell. response to UV-A
creatinine metabolic proc.
MME
proteolysis
cell. response to UV-B
response to cytokine
response to IL-1
cell. response to cytokine stimulus
cell. response to IL-1
IL1R1
IL-1-mediated signaling path.
cytokine mediated signaling path.
reg. of IL-1-mediated signaling path.
IL1RN
- reg. of IL-1-mediated signaling path.
Anakinra



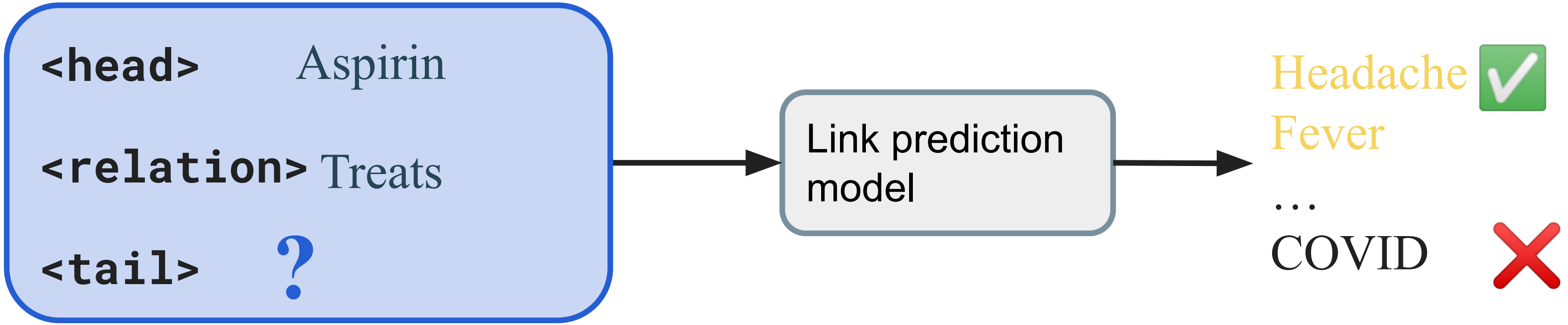


KG Completion for Drug Discovery

RepoDB: Drug-disease pairs intended for **drug repositioning** research



Knowledge Graph Link Prediction



<head> Aspirin

<relation> Treats

<tail> ?

Link prediction model

Headache ✅

Fever

…

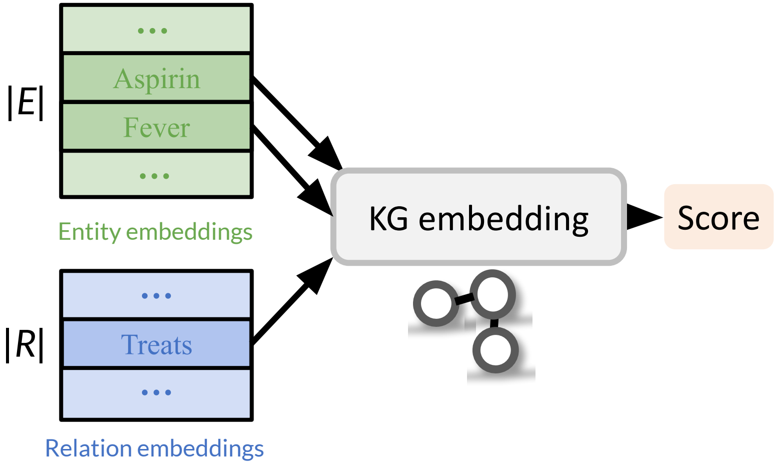
COVID ❌

Input: Structured query

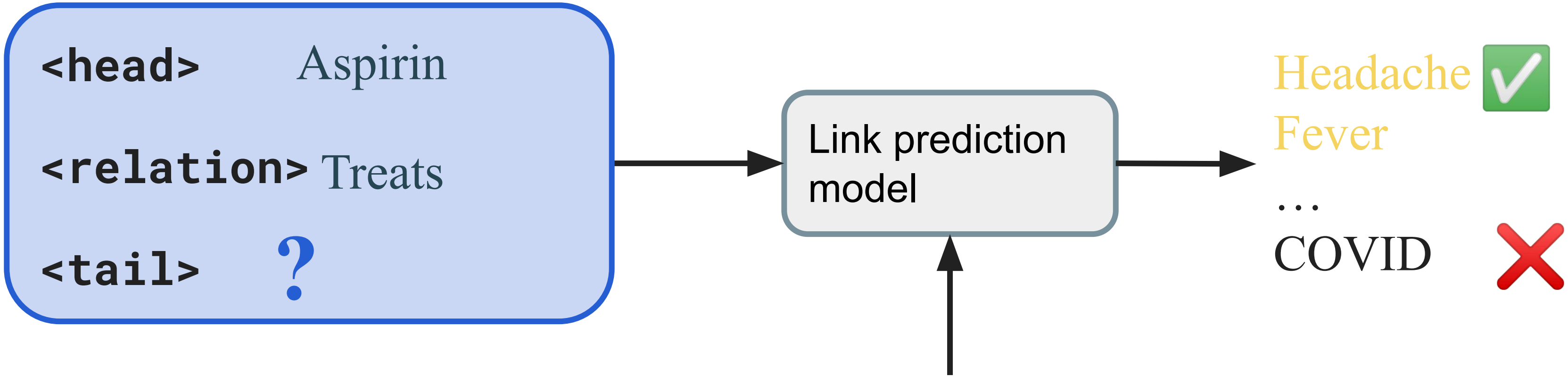
Output: Ranked candidates

KG Embedding (KGE)

Learn embeddings for entities and relations



Rich Textual Information...



<head> Aspirin

<relation> Treats

<tail> ?

Link prediction model

Headache ✅

Fever

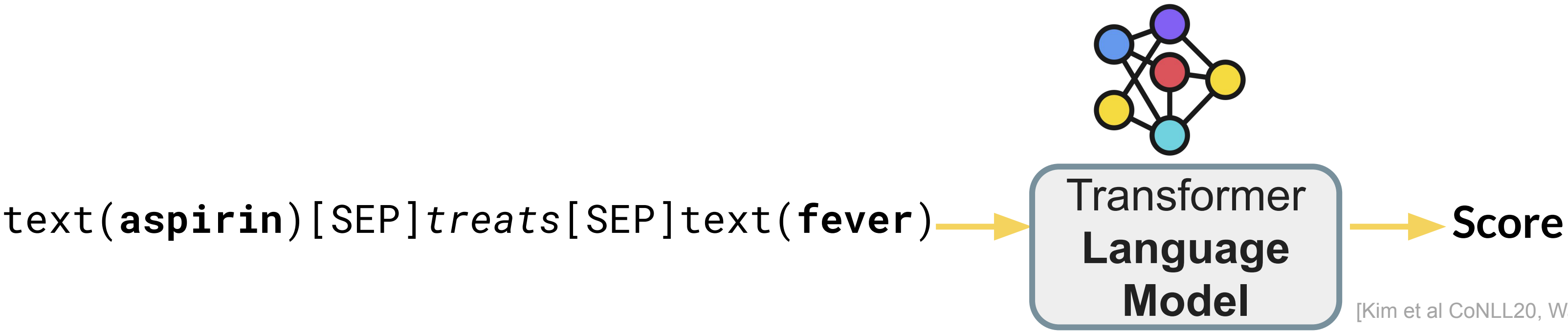
...

COVID ❌

DRUGBANK

```
Acetylsalicylic acid
Also known as Aspirin,
acetylsalicylic acid (ASA) has
both anti-inflammatory and
antipyretic...
```

Approach with Language Models



text(**aspirin**)[SEP]*treats*[SEP]text(**fever**)

[Kim et al CoNLL20, Wang et al WWW21]

Graph and Literature Language Models: Complementary Strengths

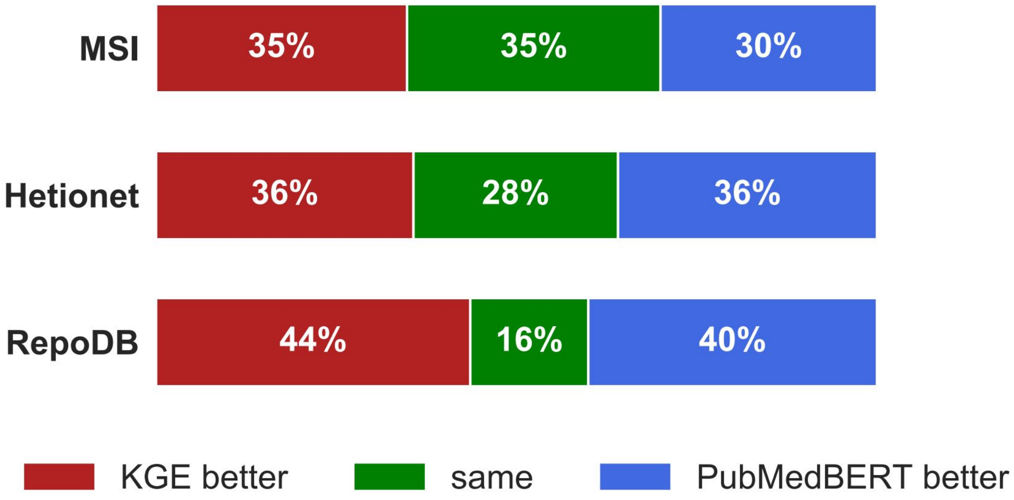
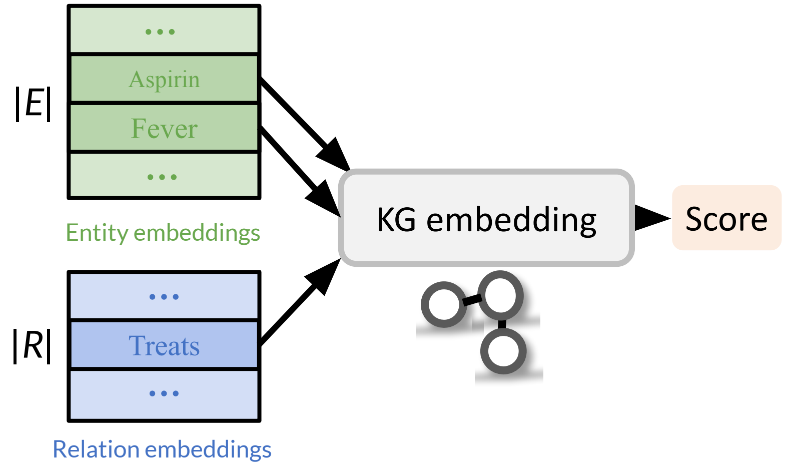
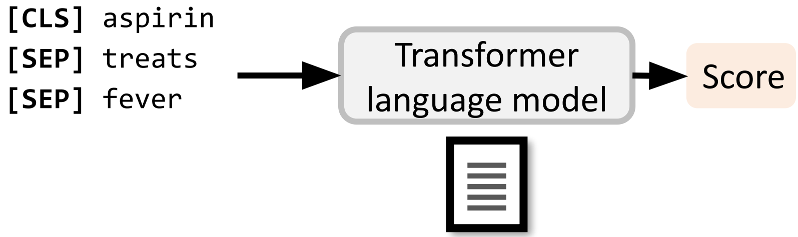


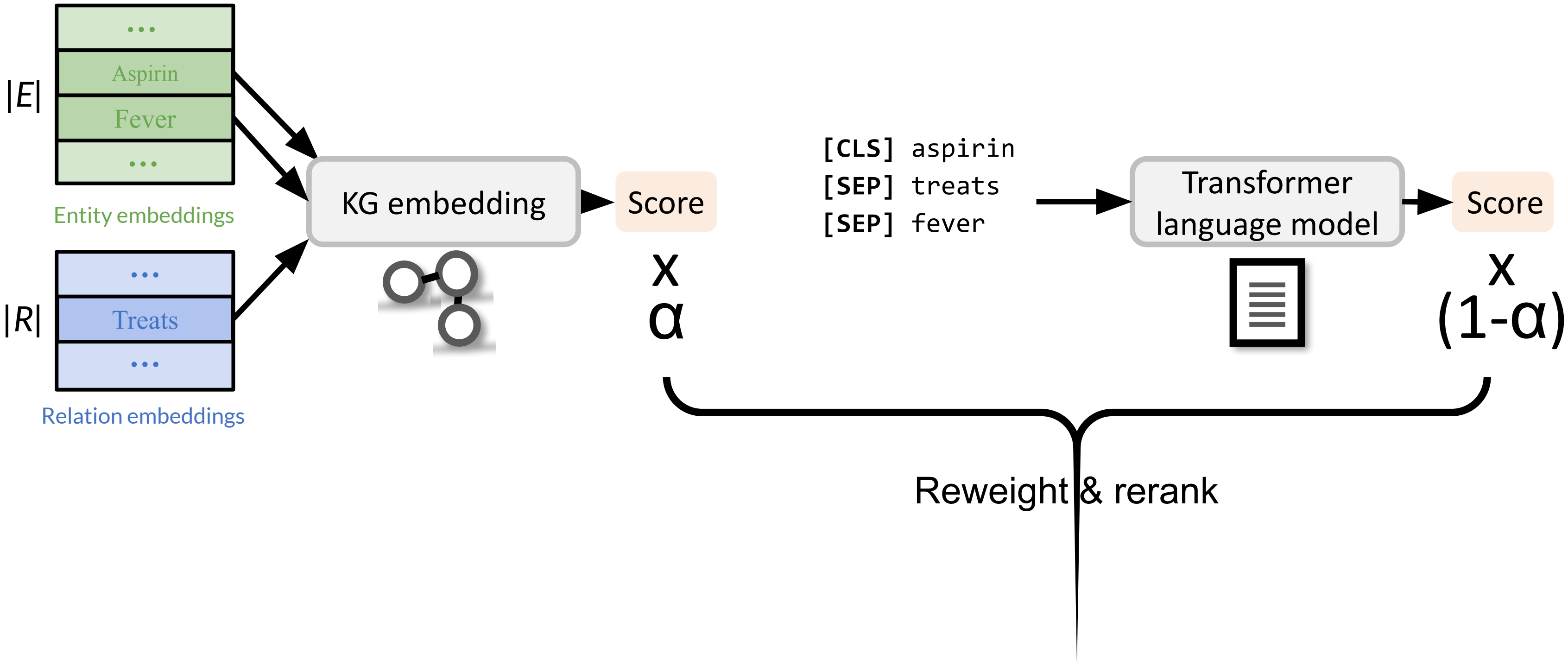
Figure 2: Fraction of test set examples where each model performs better.

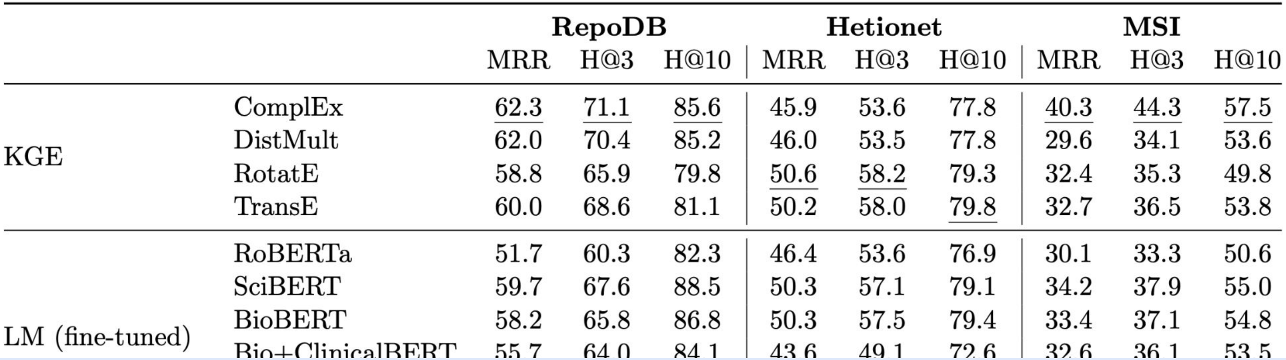
Cross-Modal Link Prediction





Cross-Modal Link Prediction





		RepoDB			Hetionet			MSI		
		MRR	H@3	H@10	MRR	H@3	H@10	MRR	H@3	H@10
KGE	ComplEx	62.3	71.1	85.6	45.9	53.6	77.8	40.3	44.3	57.5
	DistMult	62.0	70.4	85.2	46.0	53.5	77.8	29.6	34.1	53.6
	RotatE	58.8	65.9	79.8	50.6	58.2	79.3	32.4	35.3	49.8
	TransE	60.0	68.6	81.1	50.2	58.0	79.8	32.7	36.5	53.8
LM (fine-tuned)	RoBERTa	51.7	60.3	82.3	46.4	53.6	76.9	30.1	33.3	50.6
	SciBERT	59.7	67.6	88.5	50.3	57.1	79.1	34.2	37.9	55.0
	BioBERT	58.2	65.8	86.8	50.3	57.5	79.4	33.4	37.1	54.8
	Bio+ClinicalBERT	55.7	64.0	84.1	43.6	49.1	72.6	32.6	36.1	53.5

Integration of text and graph modalities provides **substantial relative improvements of 13–36% in** mean reciprocal rank (MRR).

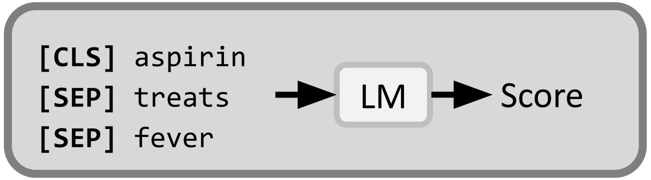
Multiple LM-based models further boosts results.

Cross-Modal Link Prediction: Challenges

	Accuracy (MRR)	**Efficiency** (inference sec, GPU)
KGE	0.33	2 x 10^1
→ Rerank w/ cross-encoder LM	0.38 (+0.05 MRR)	**1 x 10^6 11 days**

Wikipedia
The Free Encyclopedia

17K entities
206k edges



```
[CLS] aspirin
[SEP] treats
[SEP] fever
```

Cross-encoder LM
(**Slow**: combinatorial explosion)

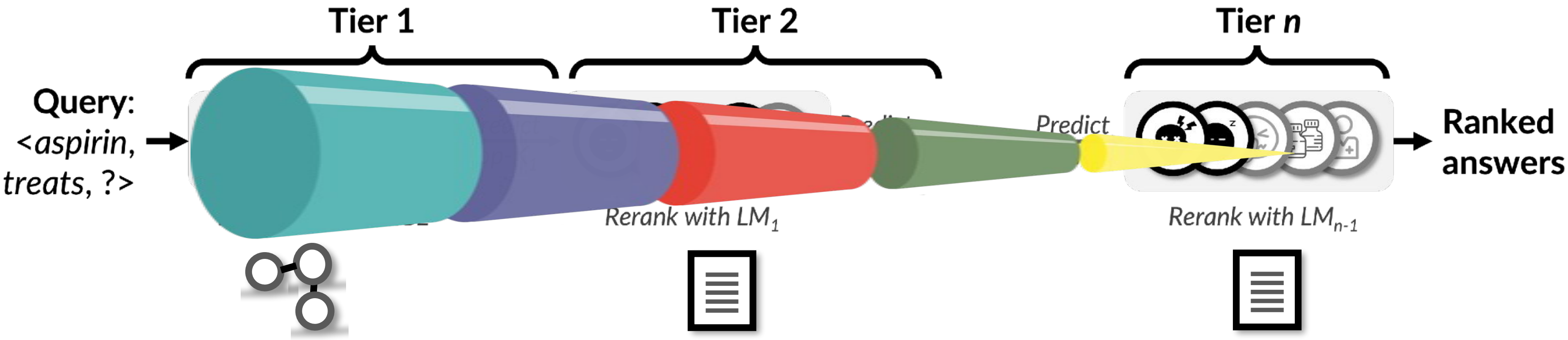
CascadER: Cross-Modal Cascading for Knowledge Graph Link Prediction

Tara Safavi, Doug Downey
and **Tom Hope**

AUTOMATED
KNOWLEDGE
BASE 2022 | London
CONSTRUCTION

AI2

Our Solution: CascadER



Models pass progressively smaller sets from one tier to the next, more expensive tiers repoDB

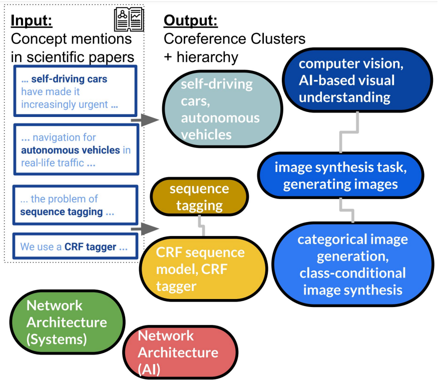
Drug Repositioning Database

WIKIPEDIA

The Free Encyclopedia

CascadER achieves SOTA accuracy while **improving efficiency by 1+ orders of magnitude** over most competitive ensemble baseline.

1. SciCo: Hierarchical Cross-Document Coreference
2. Document Similarity & Retrieval
3. Literature-Augmented Prediction



Input:
Concept mentions in scientific papers

... **self-driving cars** have made it increasingly urgent ...

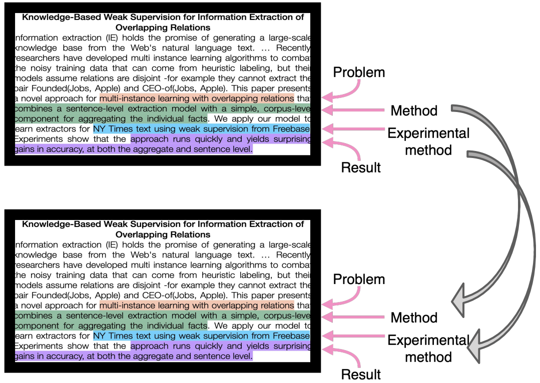
... navigation for **autonomous vehicles** in real-life traffic ...

... the problem of **sequence tagging** ...

We use a **CRF tagger** ...

Output:
Coreference Clusters + hierarchy

self-driving cars, autonomous vehicles
computer vision, AI-based visual understanding
image synthesis task, generating images
categorical image generation, class-conditional image synthesis
sequence tagging
CRF sequence model, CRF tagger
Network Architecture (Systems)
Network Architecture (AI)

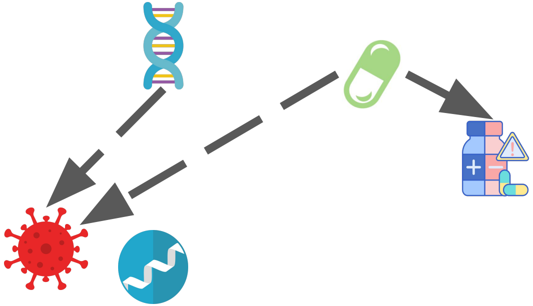


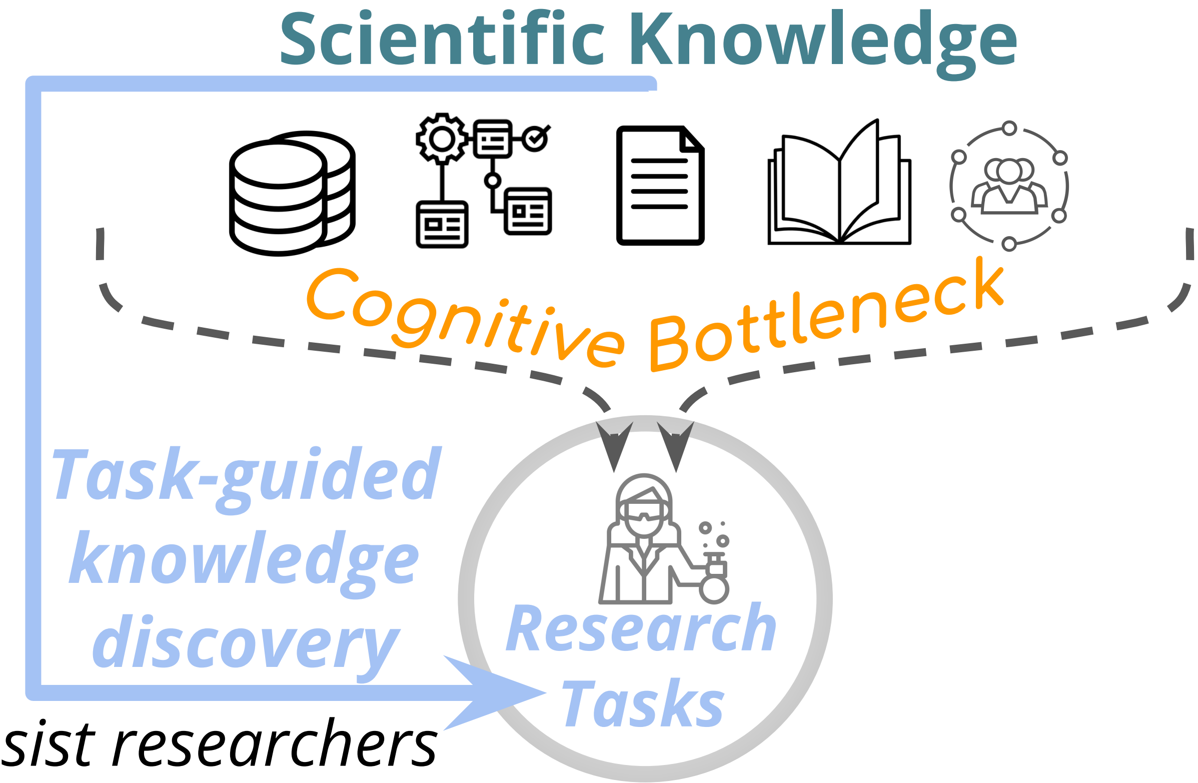
Knowledge-Based Weak Supervision for Information Extraction of Overlapping Relations

Information extraction (IE) holds the promise of generating a large-scale knowledge base from the Web's natural language text. ... Recently researchers have developed multi instance learning algorithms to combat the noisy training data that can come from heuristic labeling, but their models assume relations are disjoint -for example they cannot extract the pair Founded(Jobs, Apple) and CEO-of(Jobs, Apple). This paper presents a novel approach for multi-instance learning with overlapping relations that combines a sentence-level extraction model with a simple, corpus-level component for aggregating the individual facts. We apply our model to learn extractors for NY Times text using weak supervision from Freebase. Experiments show that the approach runs quickly and yields surprising gains in accuracy, at both the aggregate and sentence level.

Knowledge-Based Weak Supervision for Information Extraction of Overlapping Relations

Information extraction (IE) holds the promise of generating a large-scale knowledge base from the Web's natural language text. ... Recently researchers have developed multi instance learning algorithms to combat the noisy training data that can come from heuristic labeling, but their models assume relations are disjoint -for example they cannot extract the pair Founded(Jobs, Apple) and CEO-of(Jobs, Apple). This paper presents a novel approach for multi-instance learning with overlapping relations that combines a sentence-level extraction model with a simple, corpus-level component for aggregating the individual facts. We apply our model to learn extractors for NY Times text using weak supervision from Freebase. Experiments show that the approach runs quickly and yields surprising gains in accuracy, at both the aggregate and sentence level.





Scientific Knowledge

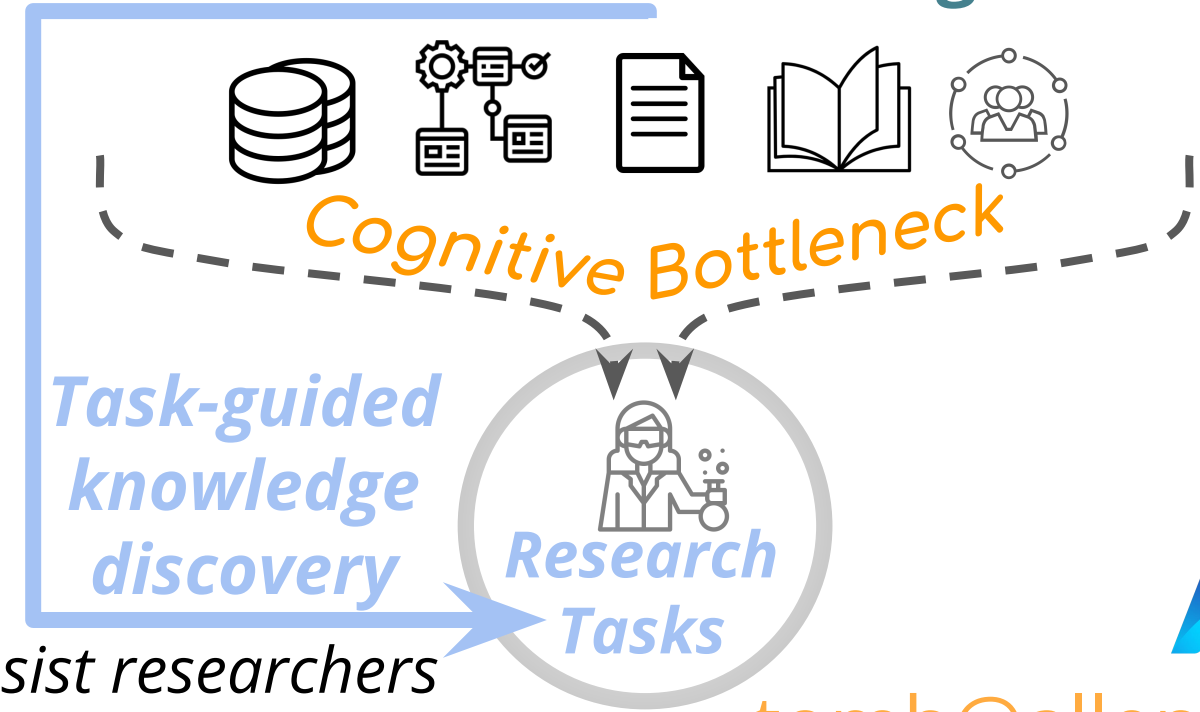
Cognitive Bottleneck

Task-guided knowledge discovery

Research Tasks

Assist researchers ***throughout their tasks***

Scientific Knowledge



Cognitive Bottleneck

Task-guided knowledge discovery

Research Tasks

Assist researchers ***throughout their tasks***

tomh@allenai.org

cs.huji.ac.il/~tomhope/