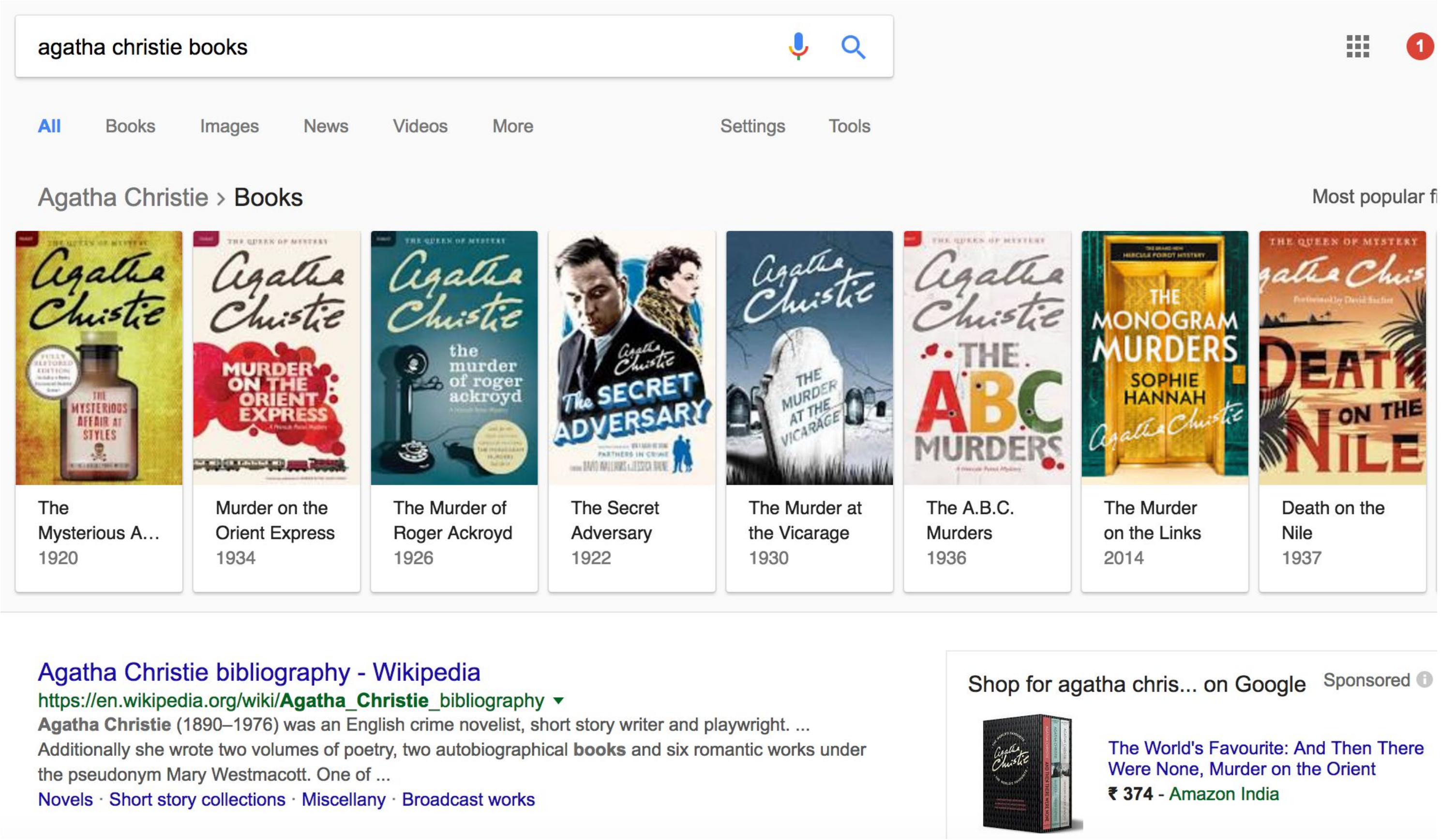
An introduction to Entity Search

Uma Sawant

IIT Bombay, LinkedIn

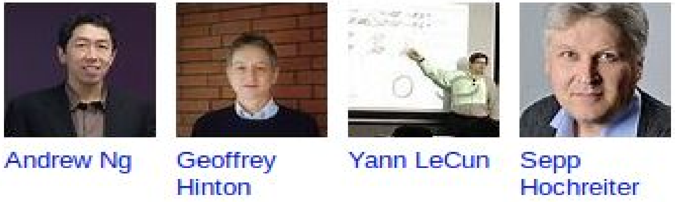
February 2017

Query : agatha christie books

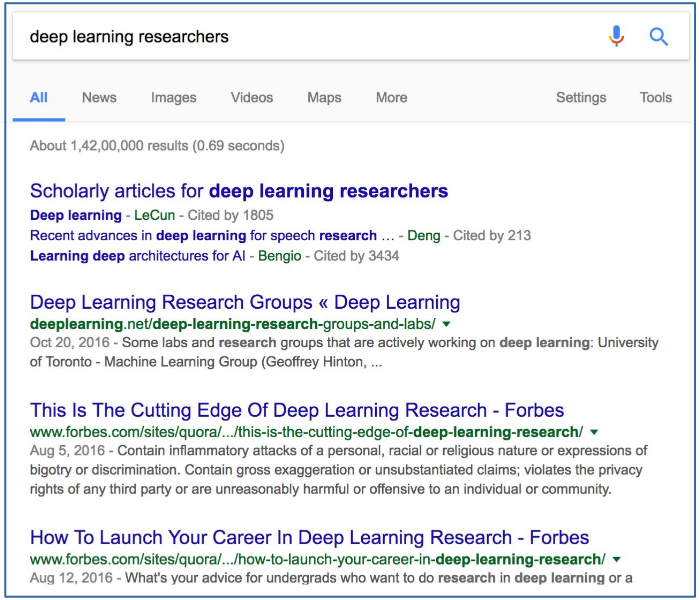


Query : deep learning researchers

What I want



What I get


deep learning researchers

All News Images Videos Maps More Settings Tools

About 1,42,00,000 results (0.69 seconds)

Scholarly articles for **deep learning researchers**

Deep learning - LeCun - Cited by 1805

Recent advances in **deep learning** for speech **research** ... - Deng - Cited by 213

Learning deep architectures for AI - Bengio - Cited by 3434

Deep Learning Research Groups « Deep Learning

deeplearning.net/**deep-learning-research**-groups-and-labs/

Oct 20, 2016 - Some labs and **research** groups that are actively working on **deep learning**: University of Toronto - Machine Learning Group (Geoffrey Hinton, ...

This Is The Cutting Edge Of Deep Learning Research - Forbes

www.forbes.com/sites/quora/.../this-is-the-cutting-edge-of-**deep-learning-research**/

Aug 5, 2016 - Contain inflammatory attacks of a personal, racial or religious nature or expressions of bigotry or discrimination. Contain gross exaggeration or unsubstantiated claims; violates the privacy rights of any third party or are unreasonably harmful or offensive to an individual or community.

How To Launch Your Career In Deep Learning Research - Forbes

www.forbes.com/sites/quora/.../how-to-launch-your-career-in-**deep-learning-research**/

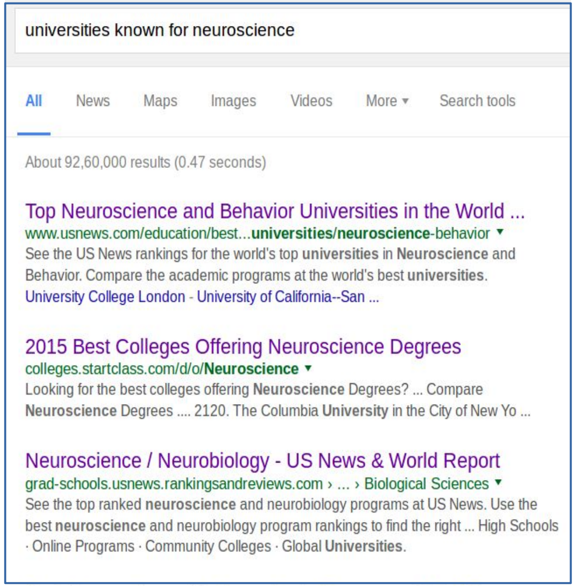
Aug 12, 2016 - What's your advice for undergrads who want to do **research** in **deep learning** or a

Query : universities known for neuroscience

What I want

Stanford John Hopkins Yale U. Chicago

What I get


universities known for neuroscience

All News Maps Images Videos More ▾ Search tools

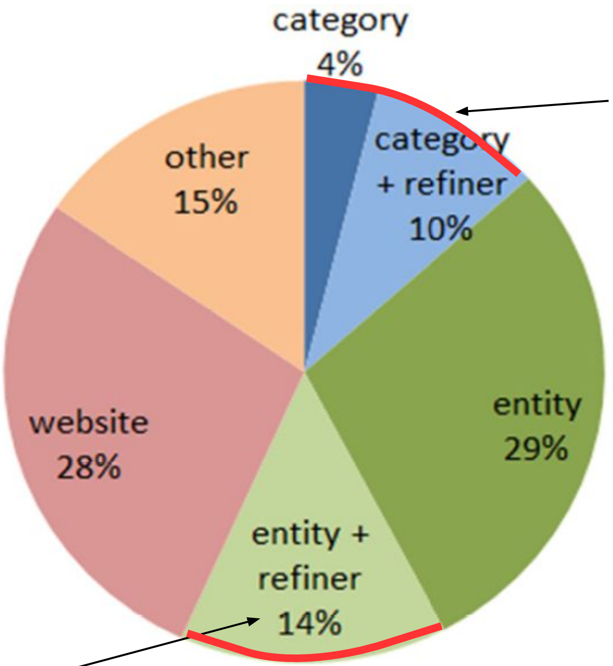
About 92,60,000 results (0.47 seconds)

Top Neuroscience and Behavior Universities in the World ...
www.usnews.com/education/best...**universities/neuroscience**-behavior ▾
See the US News rankings for the world's top **universities** in **Neuroscience** and Behavior. Compare the academic programs at the world's best **universities**.
University College London - University of California--San ...

2015 Best Colleges Offering Neuroscience Degrees
colleges.startclass.com/d/o/**Neuroscience** ▾
Looking for the best colleges offering **Neuroscience** Degrees? ... Compare **Neuroscience** Degrees .... 2120. The Columbia **University** in the City of New Yo ...

Neuroscience / Neurobiology - US News & World Report
grad-schools.usnews.rankingsandreviews.com › ... › Biological Sciences ▾
See the top ranked **neuroscience** and neurobiology programs at US News. Use the best **neuroscience** and neurobiology program rankings to find the right ... High Schools · Online Programs · Community Colleges · Global **Universities**.

~28% of Web search queries



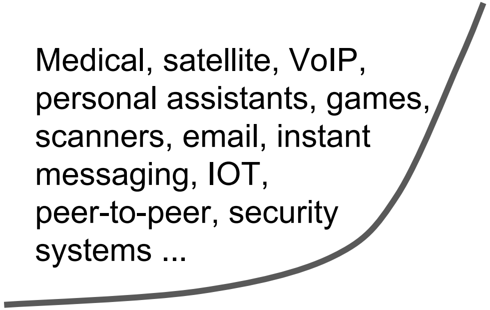
"deep learning researchers"

"Universities known for neuroscience"

Lin et al., WWW 2012

The big picture

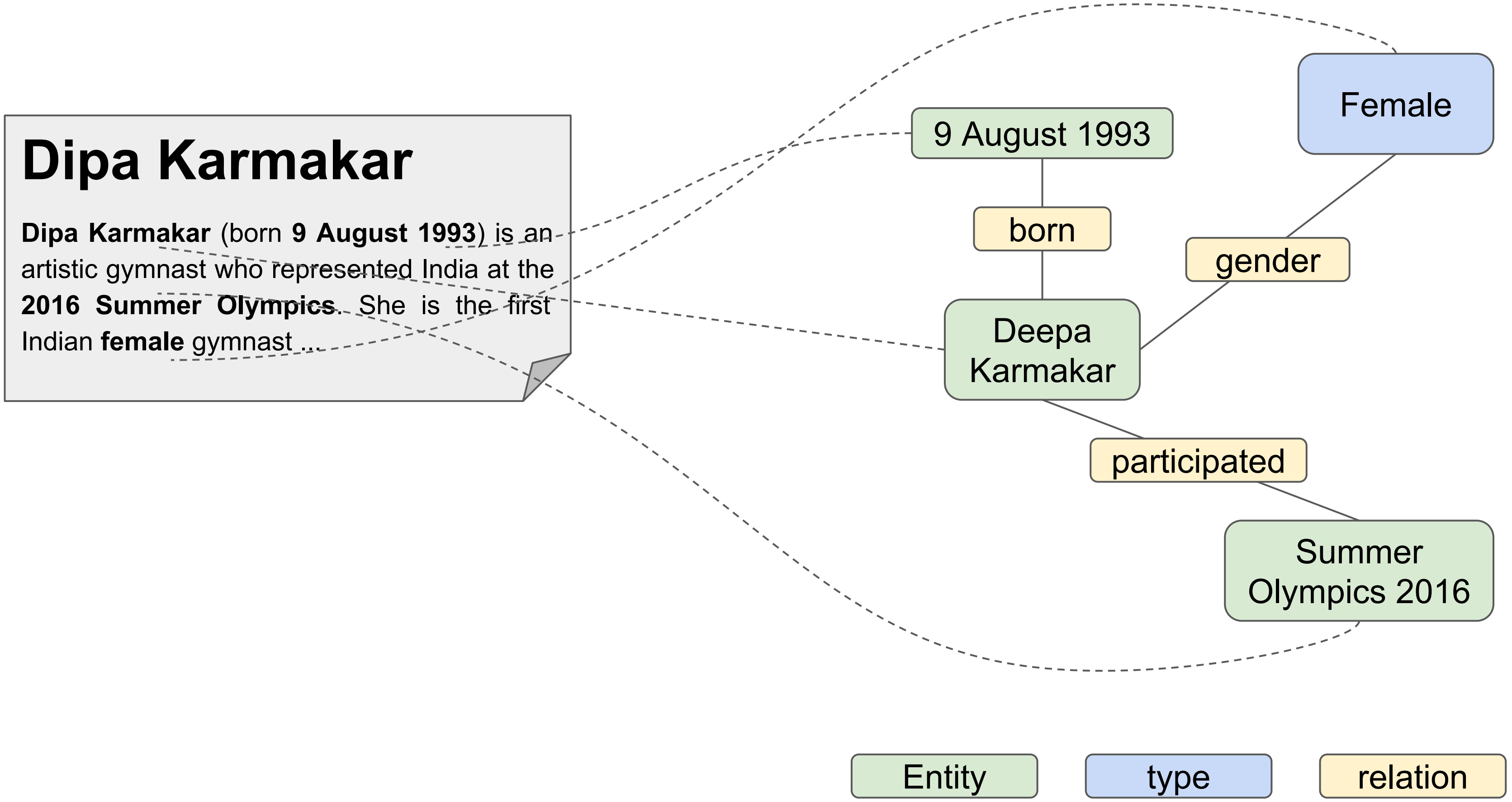
How to organize and search this big data?



Information explosion

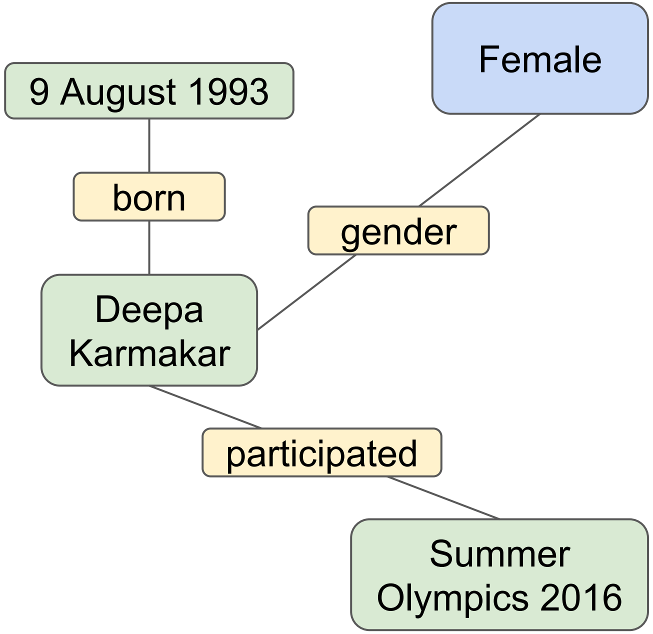
Users want direct answers

Documents vs. entities (dual view)



Knowledge graph

1. High precision (subject, relation, object) fact triplets
2. Not all information from Web is present in KG
3. Extracted using natural language resources and tools e.g. pos tagger, dictionaries, rule based systems …
4. Example : Wikipedia (infobox), Freebase, dbpedia



Knowledge graph (KG) of entities, types, relations

Problem statement : KG-driven entity search

Given structured information in a knowledge graph, how to answer any query?

Problem statement : KG-driven entity search

Given structured information in a knowledge graph, how to answer ~~any~~ entity-seeking query?

Entity - seeking queries	**Other queries**
Who is the lead singer of Euphoria band?	How did world war 2 enfold?
olympics most award winning country	If a = 2, b = 5, what is a * b ?
Name the deepest lake in the world.	How to make vanilla icing?

spanish poet died civil war
Which spanish poet died in the civil war?

→

Federico_Garcia_Lorca

Talk outline

- Overview of entity search
- Challenges in building an entity search system
- Query interpretation and ranking for entity search
  - Discriminative and Generative models for joint QI and ranking
  - Deep learning
- Experiments and results

How does an entity search engine work?

Query to answer

spanish poet died civil war
Which spanish poet died in the civil war?

→

Federico_Garcia_Lorca

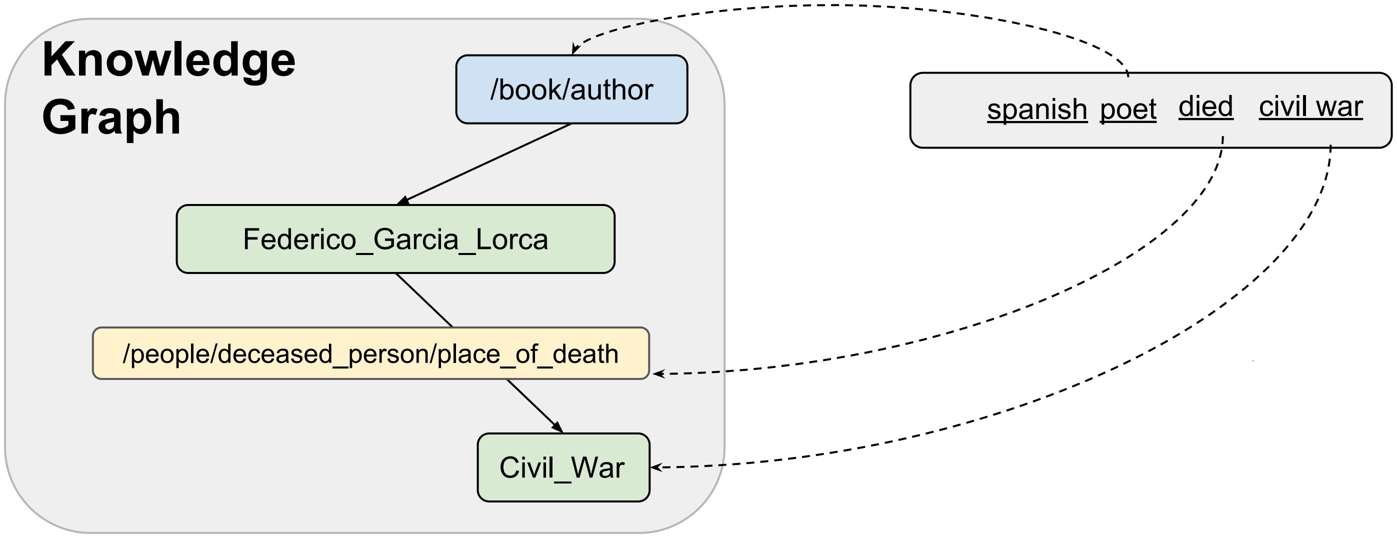
Recipe

1. Find a structured interpretation of the query by recognizing ‘semantic hints’
    a. Entities
    b. Types
    c. Relations
2. Execute the structured query on the knowledge graph.

Query to answer

spanish poet died civil war
Which spanish poet died in the civil war?

?x /people/deceased_person/place_of_death Civil_War .
?x /type/object/type /book/author

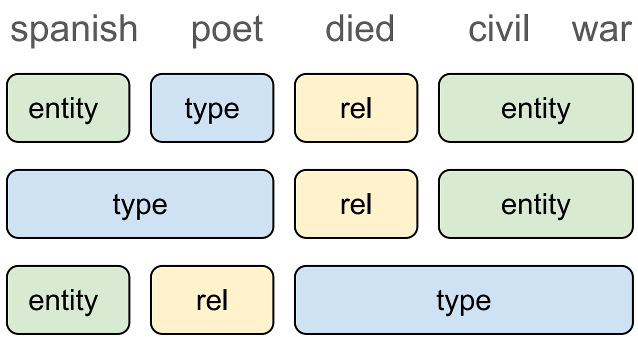


What is the difficulty?

But … there is a wall between query and answer!

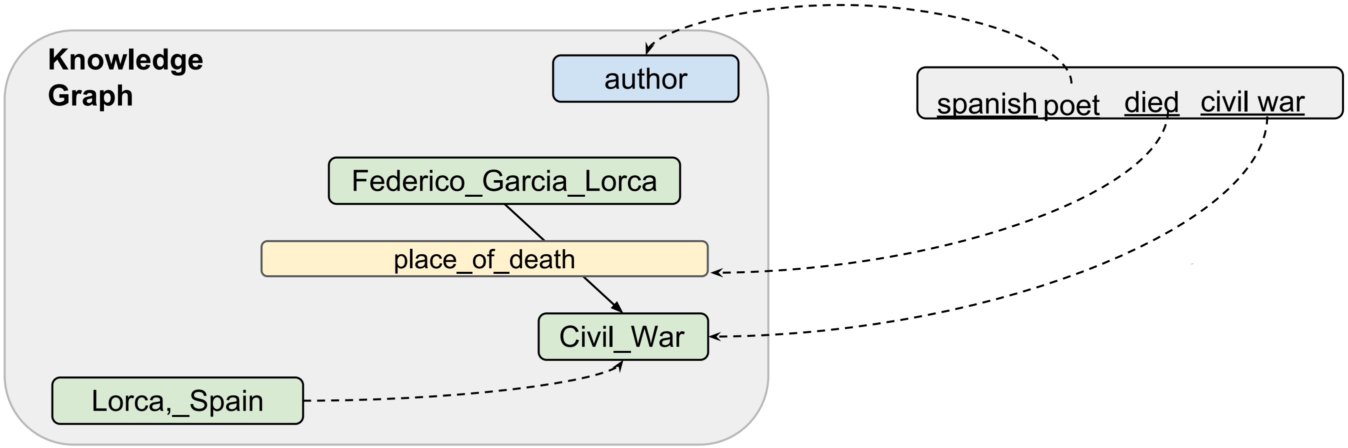
- Query understanding is difficult
    a. Many correct / incorrect interpretations
    b. Query syntax cannot always be depended on (keyword queries have no syntax)

Example :



But … there is a wall between query and answer!

- Incomplete / noisy information sources
  a. Missing KG links
  b. Incorrect KG links
  c. Information needed to answer a query may be scattered in multiple places



But … there is a wall between query and answer!

- Other challenges such as Web-scale data, index design, distributed processing, parallelization … (not in focus for this talk)

How do I solve this problem?

Our method

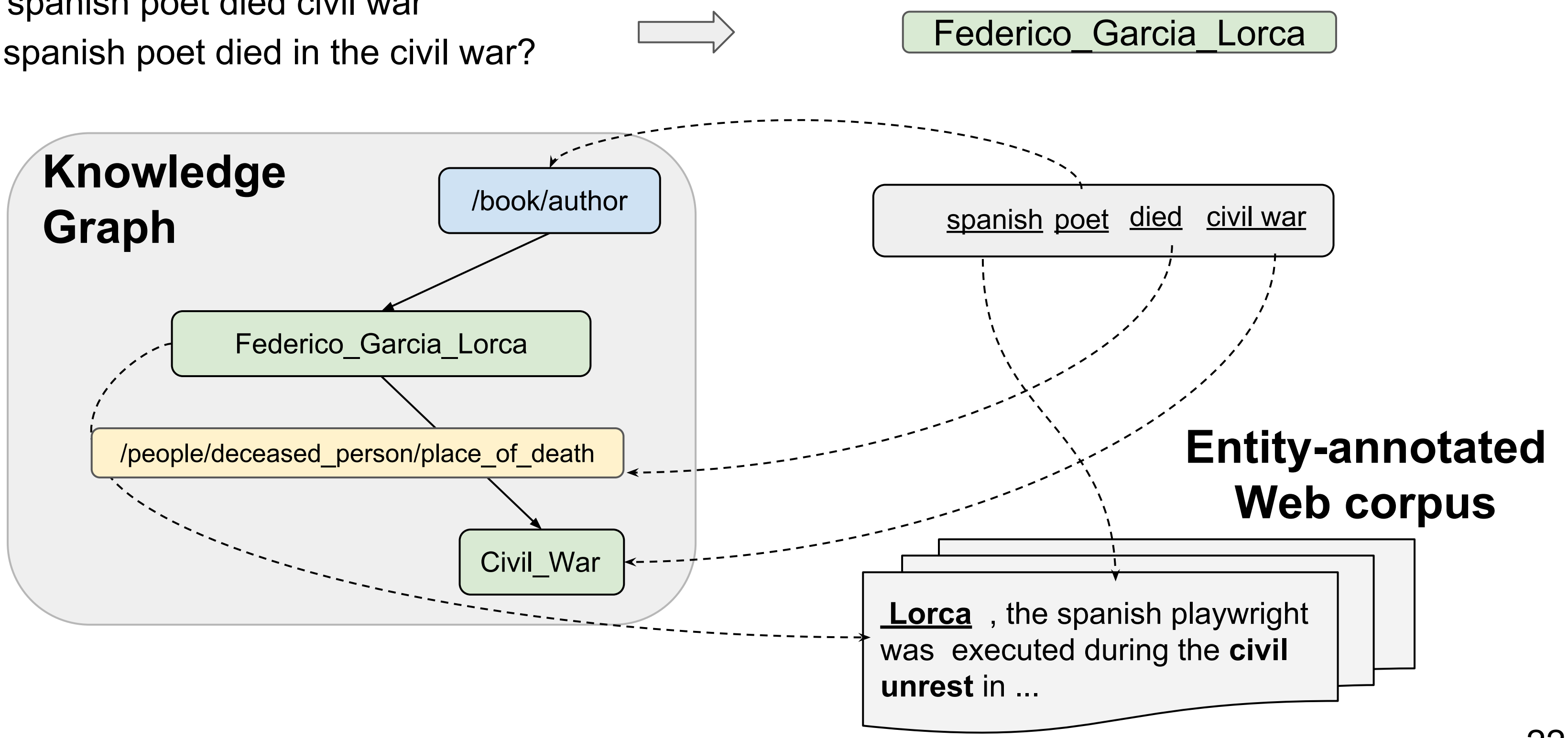
1. Entity ranking problem (instead of graph query identification problem)
    a. For each input query q, generate output ranking over entities using any number of information sources

Our method

1. Entity ranking problem
   a. For each input query q, generate output ranking over entities
2. Incomplete / noisy information sources
   a. Use both annotated corpus and KG as information sources

Query to answer

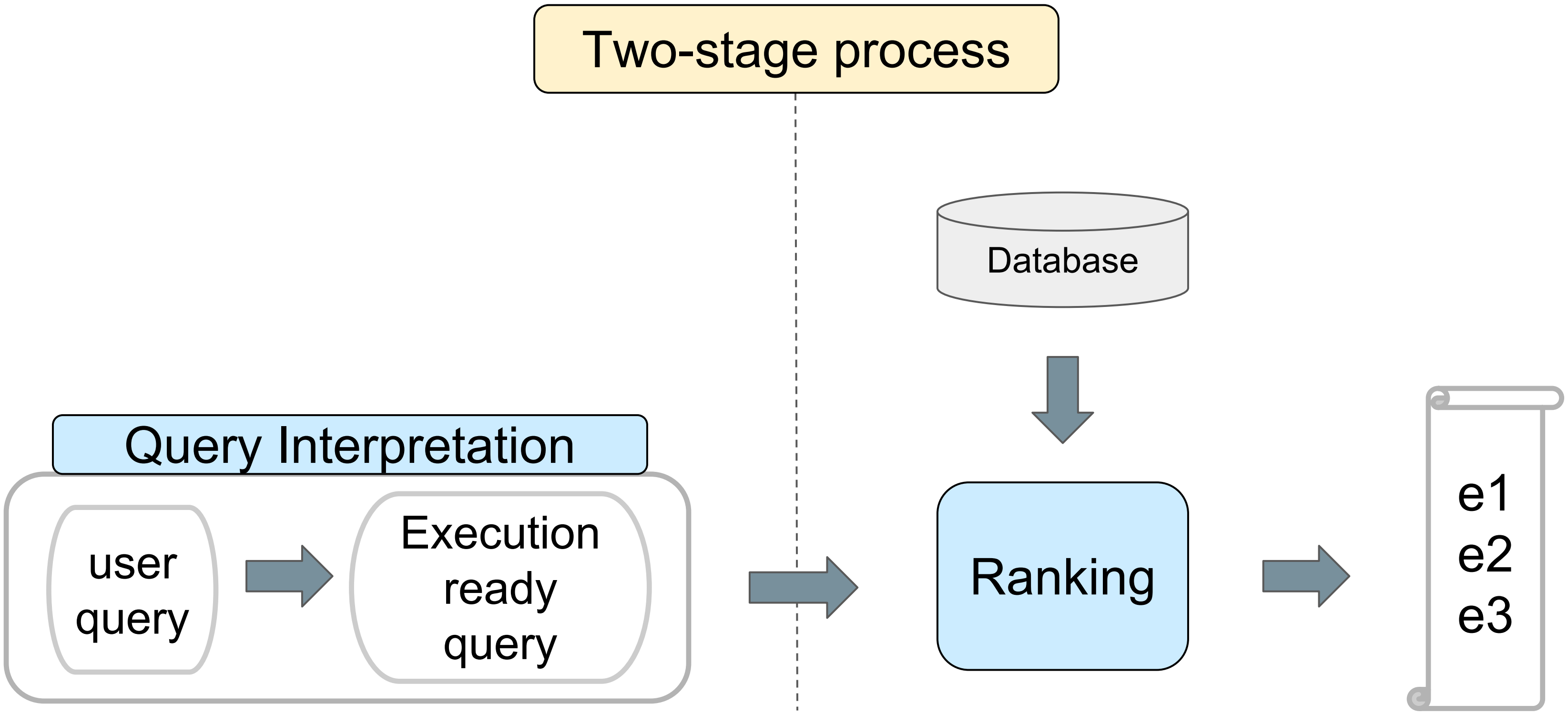
spanish poet died civil war
Which spanish poet died in the civil war?



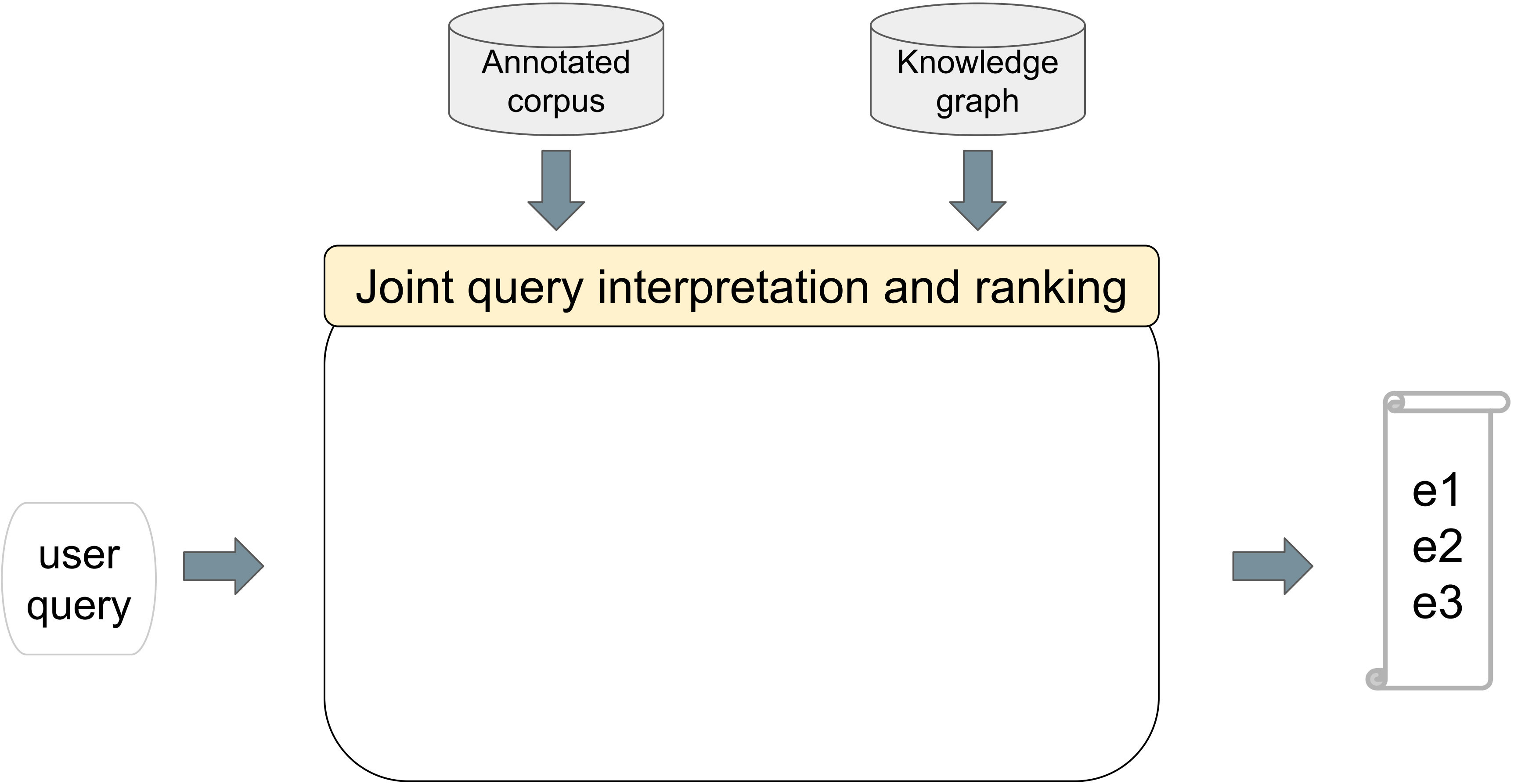
Our method

1. Entity ranking problem
   a. For each input query q, generate output ranking over entities
2. Incomplete / noisy information sources
   a. Use both annotated corpus and KG as information sources
3. Query interpretation is difficult
   a. Ideal query interpretation as a latent variable
   b. Consider many possible interpretations and jointly solve the interpretation and ranking problem

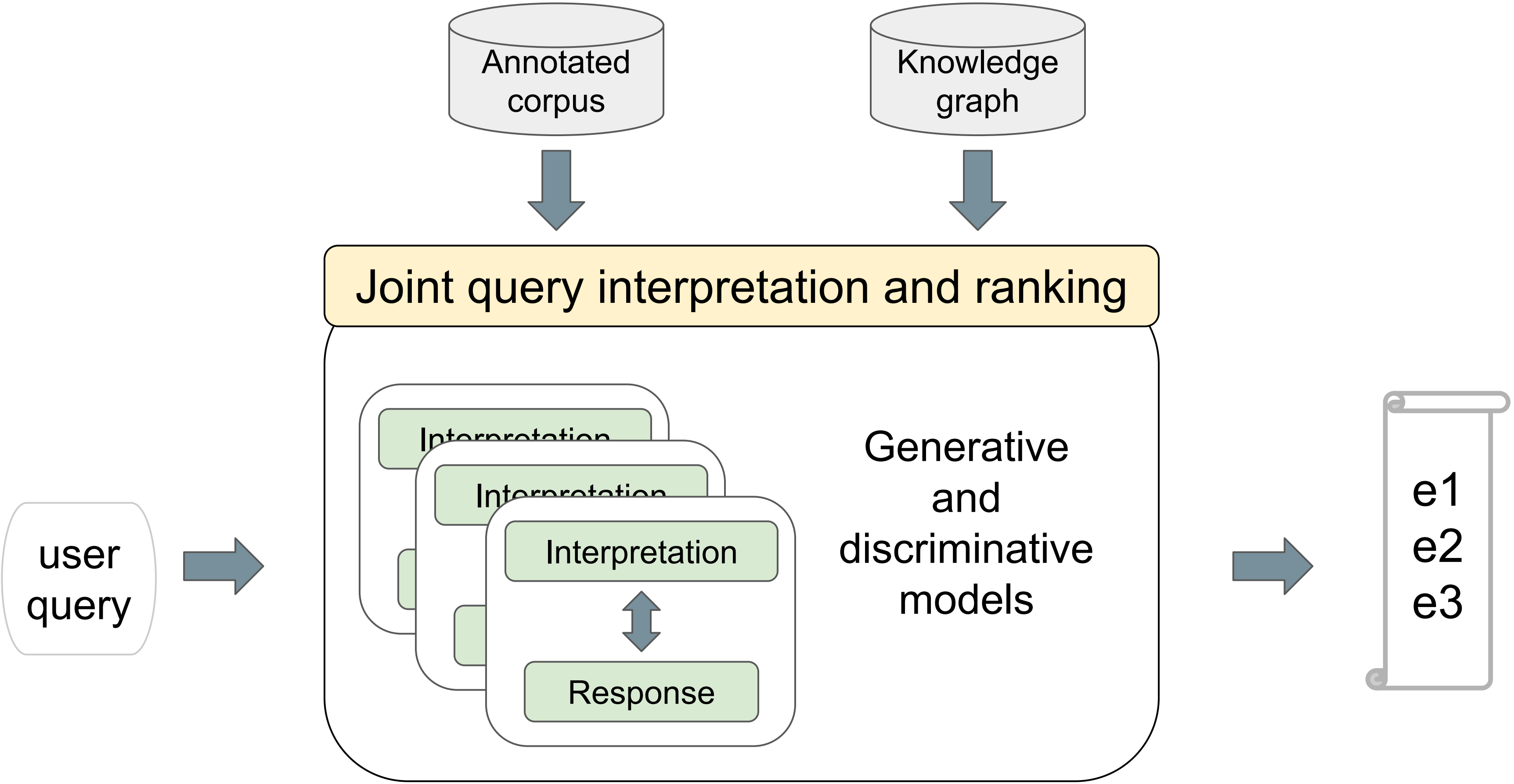
Simplified view of related work



Our approach



Our approach



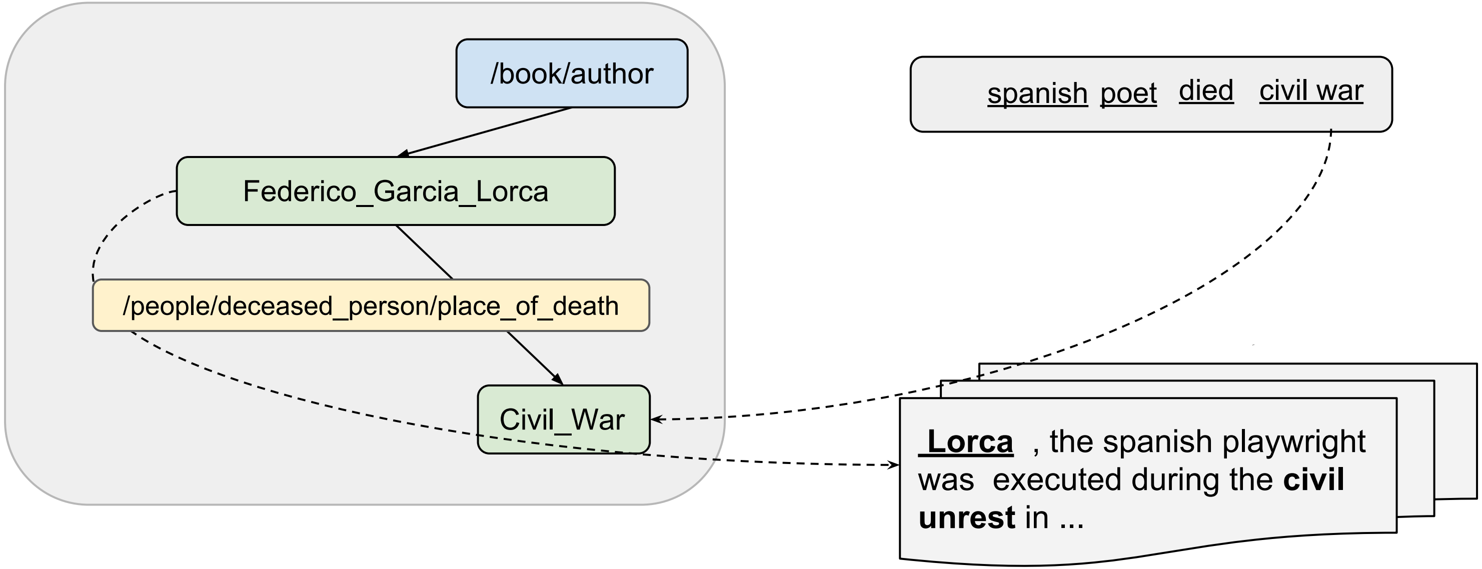
Our approach (recipe)

1. Generate candidate interpretations and hence candidate answer entities
2. Gather supporting evidence / features from KG and corpus
3. Run discriminative / generative models to perform joint interpretation and ranking

Candidate generation

Input : Query q

1. Identify in-query entities E_1
2. Gather text snippets containing query words and an entity
3. Identify answer entity set E_2
   a. Neighbours of E_1 in KG
   b. Entities that occur in corpus snippets
4. All the KG paths between E_1 and E_2 , and corpus snippets are candidate query interpretations I


/book/author
Federico_Garcia_Lorca
/people/deceased_person/place_of_death
Civil_War
spanish poet died civil war
Lorca, the spanish playwright was executed during the civil unrest in ...

Feature generation

Goal : Generate a feature vector to describe the match between query q, candidate interpretation I and candidate answer entity e

Features :

1. Entity tagger score for query entity
2. Match score for (q, t)
3. Match score for (q, r)
4. Corpus snippet score for q
5. Deep neural networks ! (a.k.a. The magic wand)
6. ...

Models for joint QI and Ranking

1. Goal : Correct entity should score higher than incorrect entity
2. Constraint : Ideal interpretation unknown
3. Models :
    a. Latent Variable Discriminative Model (LVDT)
    b. Graphical model

Model 1 : Latent Variable Discriminative Model

q : spanish poet died civil war

\+ w −

Federico_García_Lorca

Spanish poet died civil

Spanish poet died civil

spanish poet died civil war
(*t:poets_from_spain*)

Spanish_Civil_War

Spanish poet died

Spanish poet died

Spanish poet died civil war
(*t = civil_wars*)

LVDT formulation

- Constraints based on best scoring interpretation
  - Find weight vector s.t. Best scoring positive entity interpretation scores higher than best scoring negative entity interpretation
- Non convex formulation, solved via alternative optimization

best scoring

to learn

Query interpretation connected to positive entity

$$\max_{q,z,e_1,t_2,r} w \cdot \phi(q, z, e_1, t_2, r, e_2^{+}) + \xi$$

$$\geq 1 + \max_{q,z,e_1,t_2,r} w \cdot \phi(q, z, e_1, t_2, r, e_2^{-})$$

SVM margin

best scoring

Query interpretation connected to negative entity

LVDT complete formulation

$$\min_{w,\xi,u} \ \frac{1}{2}\|w\|^2 + \frac{C}{|\mathcal{Q}|}\sum_{q\in\mathcal{Q}} \frac{1}{|\mathcal{E}_q^+|\,|\mathcal{E}_q^-|} \sum_{e_2+\in\mathcal{E}_q^+, e_2^-\in\mathcal{E}_q^-} \xi_{q,e_2^+,e_2^-}$$

$$\forall q, e_2^+, e_2^-, e_1', t_2', r' :$$

$$\sum_{z,e_1,t_2,r} u(q,z,e_1,t_2,r,e_2^+) w \cdot \phi(q,z,e_1,t_2,r,e_2^+)$$

$$\geq 1 - \xi_{q,e_2^+,e_2^-} + w \cdot \phi(q,z,e_1',t_2',r',e_2^-)$$

$$u(q,z,e_1,t_2,r,e_2^+) \in \{0,1\}$$

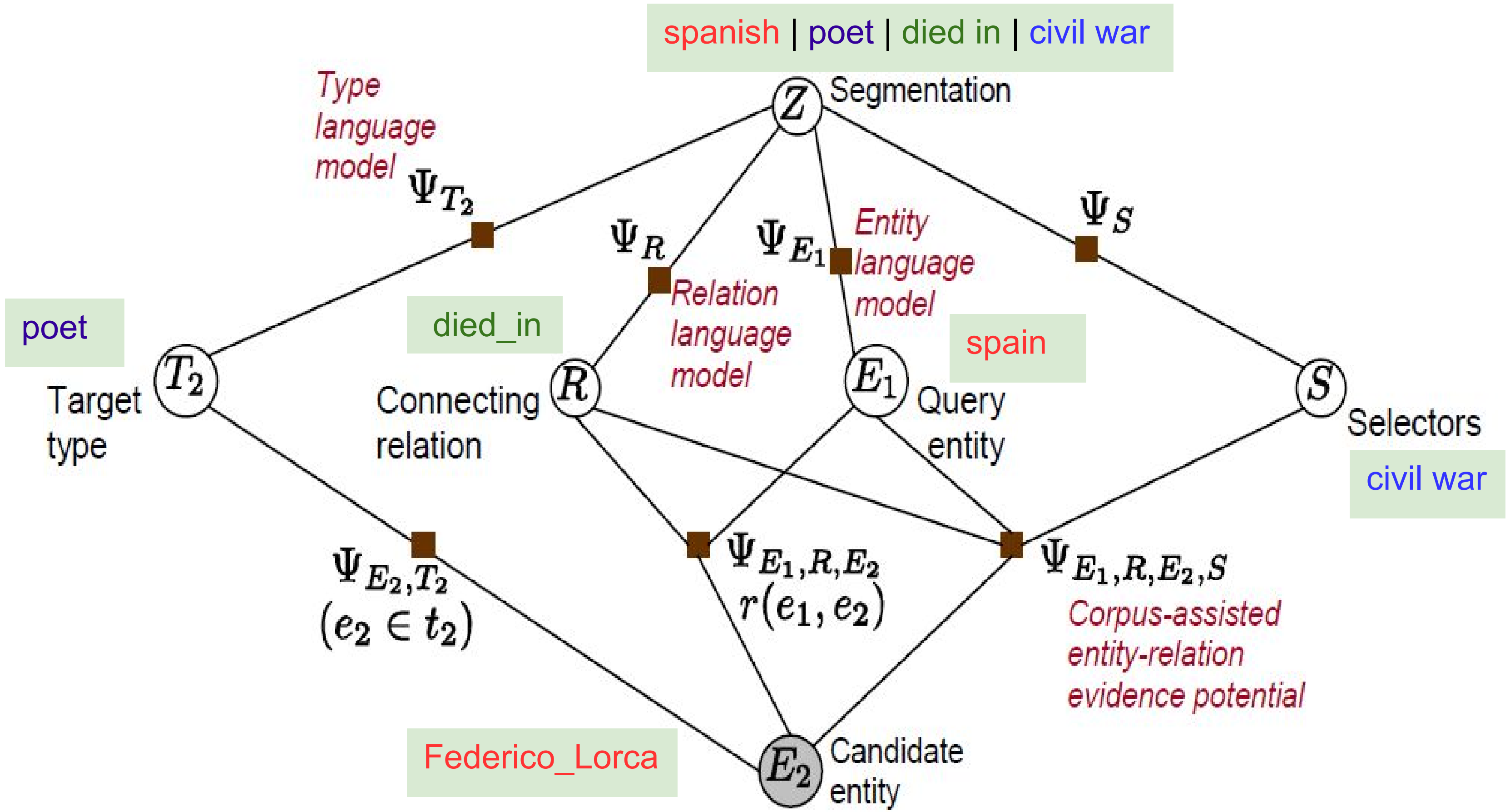
$$\forall q, e_2^+ : \sum_{z,e_1,t_2,r} u(q,z,e_1,t_2,r,e_2^+) = 1$$

$$\forall q, e_2^+, e_2^- : \xi_{q,e_2^+,e_2^-} \geq 0$$

Model 2 : Graphical model

- Generative model represented as a graph
- Nodes = variables (observed evidence or hidden parameters)
- Edges = dependencies between variables
- Potentials = Unnormalized weights on the edges, indicate connection strength
- Inference = Assign best values to nodes

Model 2 : Graphical model



Experiment setup

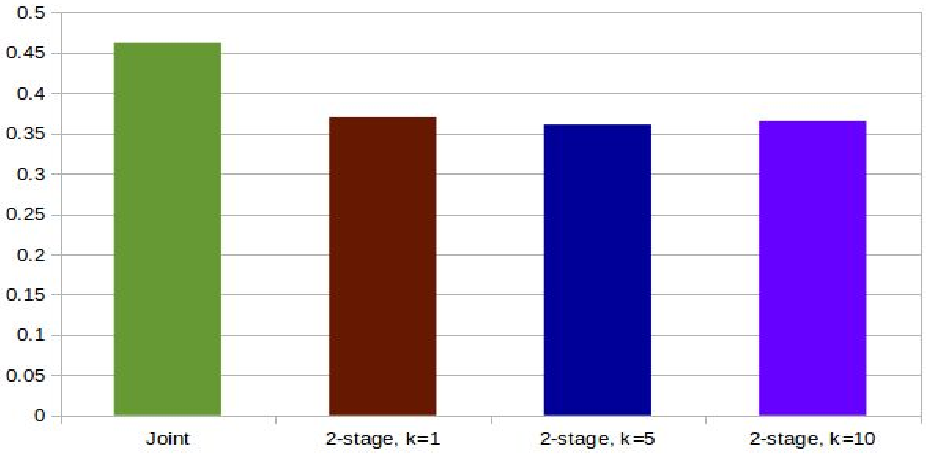
- Freebase knowledge graph
  - ~29 million entities, 14K types, ~4.6K relation types
- FACC1/ClueWeb09B entity-annotated corpus :
  - 50 million pages, ~13 annotations per page
- Querysets

Source	Queryset	#queries	Type
TREC-INEX	TI-KW	704	Keyword
	TI-NLQ	704	Well-formed
WebQuestions	WQ-KW	803	Keyword
	WQ-NLQ	5810	Well-formed

Does joint query interpretation and ranking work better than two-stage?

- Setting : Compare two-stage type-predictor + ranking with our models
- State-of-the-art target type predictor (Balog et. al.)
- Union of k types to improve recall
- Launch type-restricted query on corpus + graph

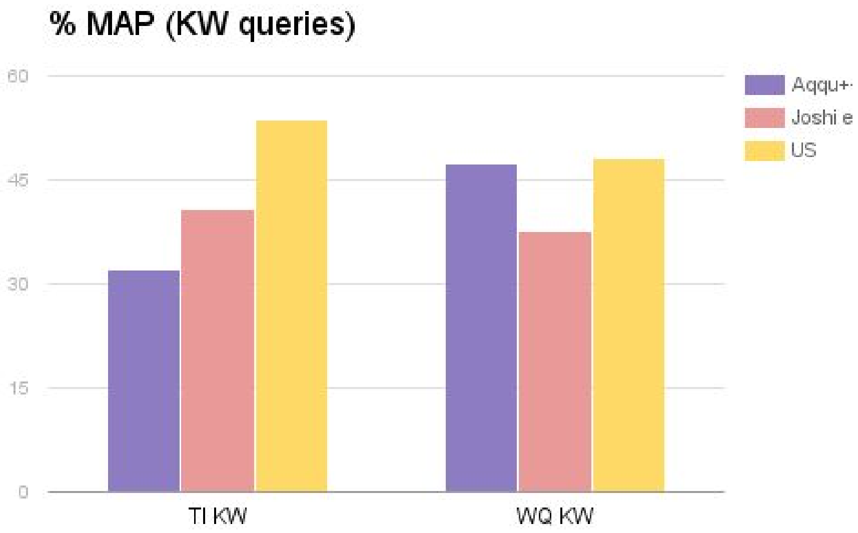
LVDT

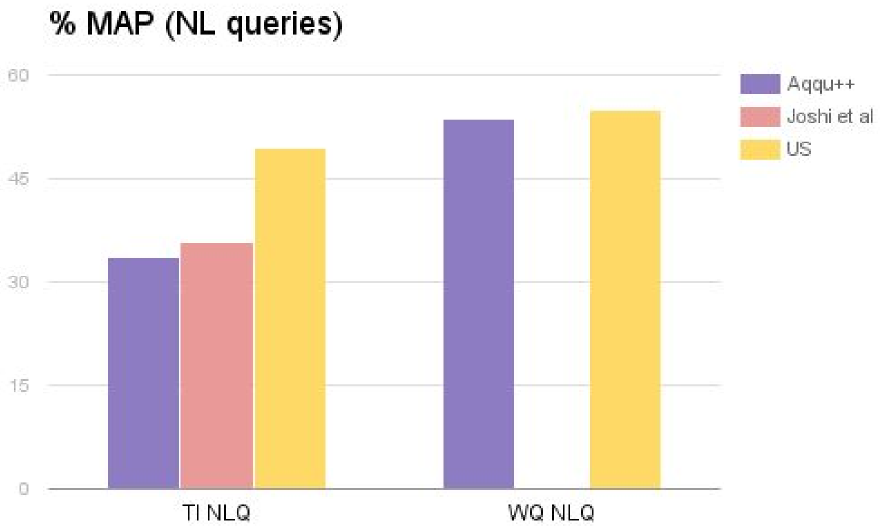


TREC-INEX

Conclusion : Upto 10% absolute gain through joint prediction and ranking

End-to-end comparison with related work





- 1 to 15% absolute MAP gain over Joshi 2014 and Aqqu++

Failure analysis

- Good
  - Queries including qualifiers such as ‘first’, ‘oldest’ (Who was the first U.S. president ever to resign?)
  - Incomplete knowledge graph (president sworn on airplane)
  - No clear query entity e_1 (Which kennedy died first?)
- Bad
  - When to trust which information source?
  - Corpus popularity promotes incorrect entities : Jon_Stewart ranked above Madeleine_Smithberg for “creator of the daily show”
  - Failure of type/relation CNNs

Take-away

1. Entity search is a critical component of Web search, but non-trivial.
2. Knowledge graph and corpus offer complementary benefits.
3. Joint query and interpretation performs better than two-stage approach.

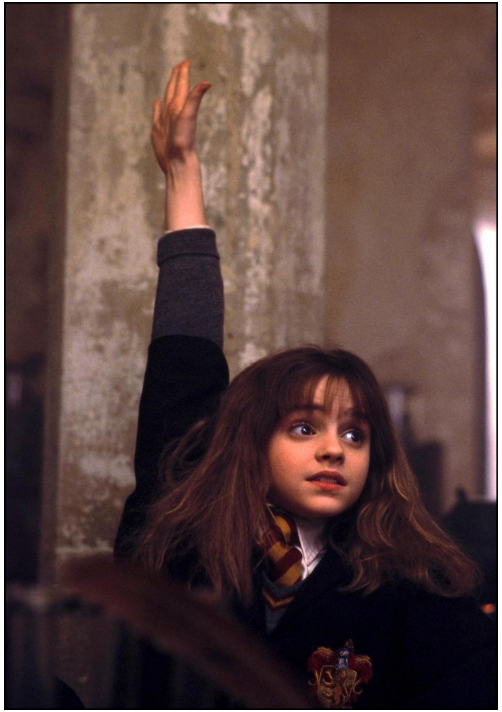
End-to-end entity search systems

1. Aqqu : http://ad-publications.informatik.uni-freiburg.de/CIKM_freebase_qa_BH_2015.materials/
2. Sempre : http://www-nlp.stanford.edu/software/sempre/
3. CSAW : https://www.cse.iitb.ac.in/~soumen/doc/CSAW/
4. Ours (work in progress)

References

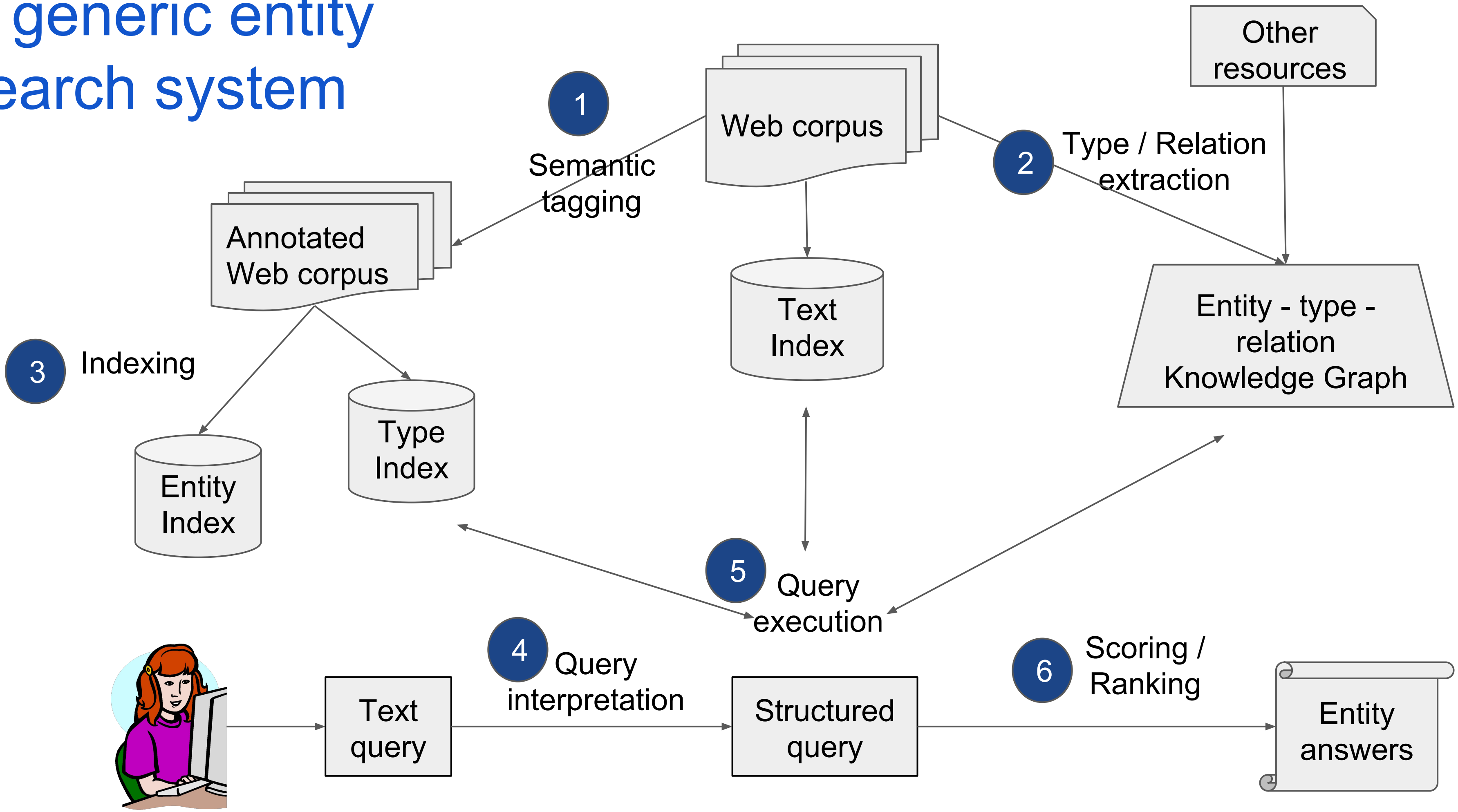
1. Features and aggregators for ranking interpreted entity search queries (Technical report)
2. Joint query (type) interpretation and ranking for entity-seeking queries (WWW 2013)
3. Corpus and knowledge graph driven query segmentation and ranking (EMNLP 2014)
4. Hannah Bast and Elmar Haussmann. More accurate question answering on freebase. (CIKM 2015).
5. Aliaksei Severyn and Alessandro Moschitti. 2015. 829 Learning to rank short text pairs with convolutional 830 deep neural networks. (SIGIR ’15)
6. Antoine Bordes, Sumit Chopra, and Jason Weston. (2014). Question answering with subgraph embeddings

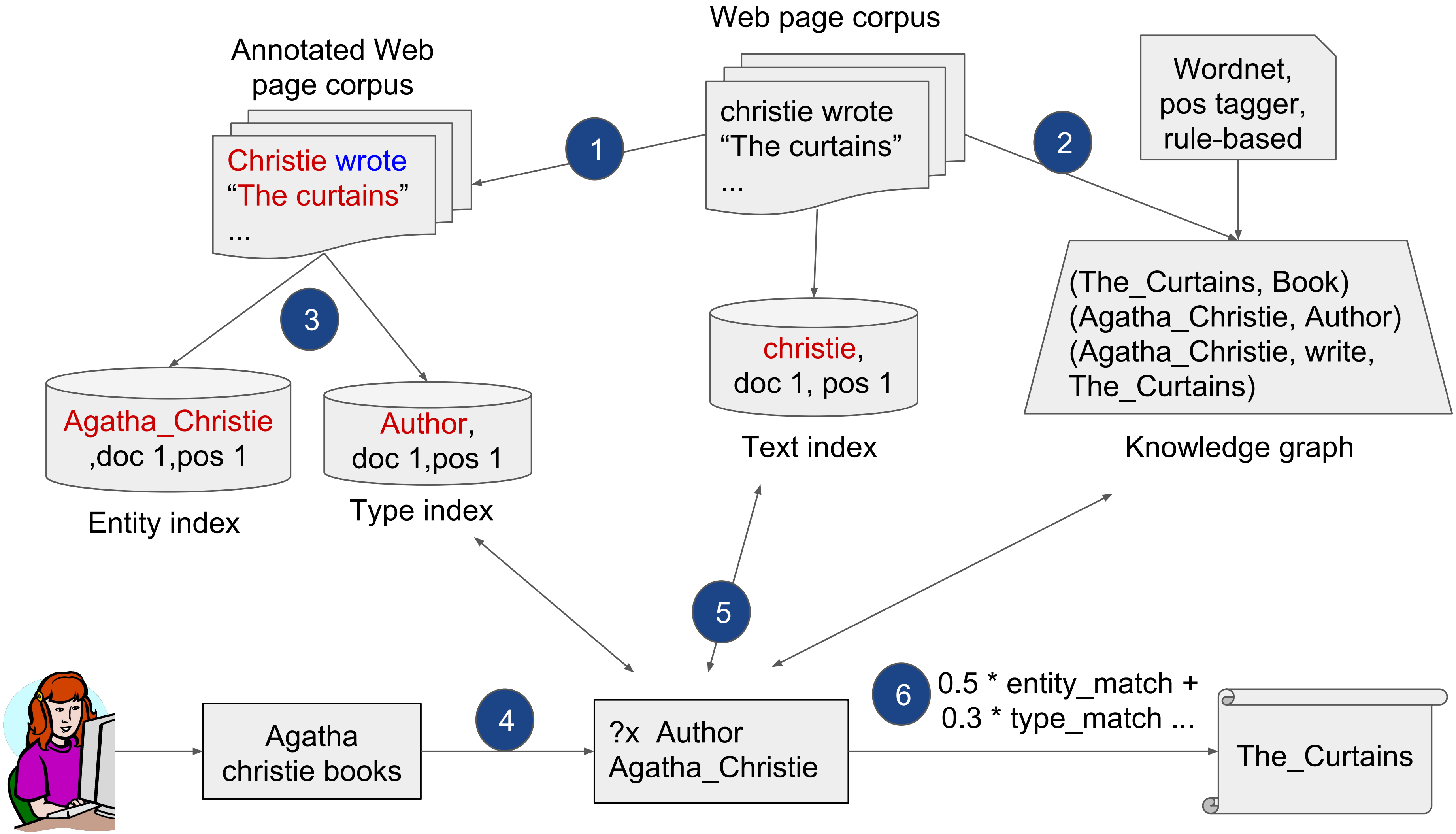
Thank you! Questions? Comments?



Extra slides

A generic entity search system





Web page corpus
christie wrote
“The curtains”
...
Annotated Web page corpus
Christie wrote
“The curtains”
...
Wordnet,
pos tagger,
rule-based
1
2
3
Agatha_Christie
,doc 1,pos 1
Entity index
Author,
doc 1,pos 1
Type index
christie,
doc 1, pos 1
Text index
(The_Curtains, Book)
(Agatha_Christie, Author)
(Agatha_Christie, write,
The_Curtains)
Knowledge graph
5
Agatha
christie books
4
?x Author
Agatha_Christie
6
0.5 * entity_match +
0.3 * type_match ...
The_Curtains

Related work (bridge query to answer gap)

1. Query understanding
    a. Feature engineering using hand-created features (Bast2015) vs. Deep neural networks (Dong2015, Stagg2015, Sawant2017),
    b. Take advantage of natural language syntax e.g. semantic parsers (Berant2013, Berant2014, Berant2016) vs. segmentation based models for keyword queries (Sawant2013, Joshi2014)
    c. Two-staged approach of query interpretation followed by ranking (Berant2013) vs Joint query interpretation and ranking (Sawant2013, Joshi2014)

Related work (bridge query to answer gap)

1. Query understanding
2. Incomplete / noisy information sources
    a. Enrich KG facts with text descriptions (Robust QA)
    b. Add more facts to KG (Renoun, Reverb)
    c. Discover new types and add to KG (Universal schema)
    d. Discover missing entity annotations in the Web corpus (TMI)
    e. Combine information from KG and corpus (Sawant2013, Joshi2014)
    f. Add type annotations to Web corpus (FIGER)

Related work (bridge query to answer gap)

1. Query understanding
2. Incomplete / noisy information sources
3. Getting to the perfect answer
    a. Pose it as a “KG query prediction problem” : Returns an answer set after KG query execution. Know when you don’t know the answer
        i. Berant2013, Berant2014, Dong2015, Stagg2015, Berant2016, ...
        ii. Problem : no order between answer set, need ideal interpretation as labeled data
    b. Pose the problem as a “entity ranking problem” : allow ordering between answer entities.
        i. Sawant2013, Joshi2014,
        ii. Problem : will always have an answer, even for invalid questions.

Tools for annotating and indexing corpus and graph

1. Indexing : Lucene (http://lucene.apache.org/core/), mg4j (http://mg4j.di.unimi.it/)

2. Tagging text with wikipedia entities : tagme (https://tagme.d4science.org/tagme/), wikipedia miner (https://sourceforge.net/projects/wikipedia-miner/)

3. Querying an existing graph : http://ad-publications.informatik.uni-freiburg.de/CIKM_freebase_qa_BH_2015.materials/ This software queries a graph index loaded in virtuoso and performs question answering .

Graphical model toolkit

Keving Murphy has a comprehensive list --
https://www.cs.ubc.ca/~murphyk/Software/bnsoft.html

Datasets / querysets

1. ClueWeb12 and ClueWeb09 Web corpus --
   a. http://lemurproject.org/clueweb12/
   b. http://lemurproject.org/clueweb09/
2. Freebase entity annotations for the above --
   a. http://lemurproject.org/clueweb12/FACC1/,
   b. http://lemurproject.org/clueweb09/FACC1/
3. Question-answer querysets --
   a. https://worksheets.codalab.org/worksheets/0xba659fe363cb46e7a505c5b6a774dc8a/
   b. http://bit.ly/1OCKbVW
4. Linked Open Data : Haven’t used this myself, but recommended by others -- http://linkeddata.org/home