



# Question Answering over Curated and Open Web Sources

Rishiraj Saha Roy

Max Planck Institute for Informatics, Germany

Avishek Anand

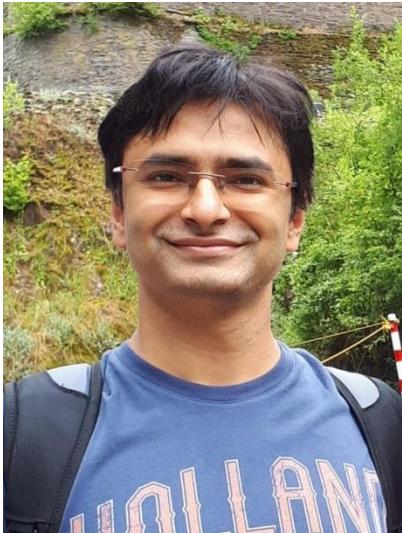
L3S Research Center, Germany



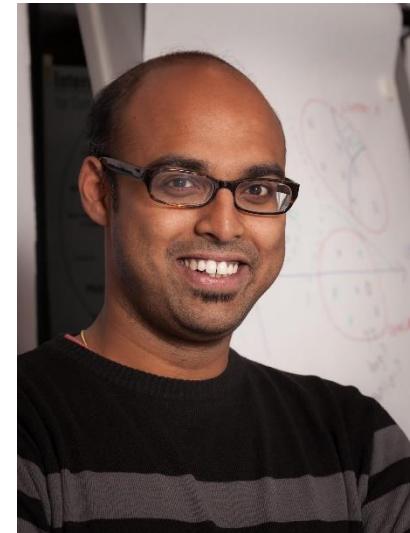
SIGIR 2020 Tutorial

26 July 2020

# Who are we?



Rishiraj Saha Roy  
Senior Researcher  
Max Planck Institute for Informatics  
Saarbrücken, Germany  
<http://people.mpi-inf.mpg.de/~rsaharo/>



Avishek Anand  
Assistant Professor  
L3S Research Center  
Hannover, Germany  
<https://www.l3s.de/~anand/>

# Tutorial outline

- What do we cover? (2016+)
  - QA over knowledge graphs
  - QA over textual sources
- What is out of scope?
  - Visual question answering
  - Domain-specific question answering
- Philosophy: Representative tasks and methods

**Read more at**

<https://arxiv.org/pdf/2004.11980.pdf>

**These slides are available at**

[http://people.mpi-inf.mpg.de/~rsaharo/sigir20slides\\_rsr.pdf](http://people.mpi-inf.mpg.de/~rsaharo/sigir20slides_rsr.pdf)

**Prerequisites:**

Basic IR, NLP, ML, DB

Understanding of core neural techniques

Interactivity ☺

# Outline: QA over knowledge graphs

- **Background:** Setup, benchmarks, metrics
- **Simple QA:** Templates and embeddings
- **Complex QA:** Multiple entities and predicates
- **Heterogeneous sources:** Handling KGs and text
- **Conversational QA:** Implicit context in multi-turn setup
- **Take-home:** Summary and insights

Representative methods from each task

Families of algorithms to build up repertoire for approaching KG-QA

Focus on methods (and not evaluation)

Understand how to go from question to answer

# Methodology

- Templates
- Graph embeddings
- Subgraph computations
- Belief propagation
- Graph traversal
- Sequence-to-sequence models

Focus on a few instantiations  
for each method

# QA over knowledge graphs

- **Background:** Setup, benchmarks, metrics
- **Simple QA:** Templates and embeddings
- **Complex QA:** Multiple entities and predicates
- **Heterogeneous sources:** Handling KGs and text
- **Conversational QA:** Implicit context in multi-turn setup
- **Take-home:** Summary and insights

Representative methods from each task

Families of algorithms to build up repertoire for approaching KG-QA

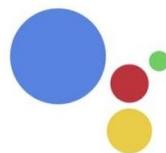
Focus on methods (and not evaluation)

Understand how to go from question to answer

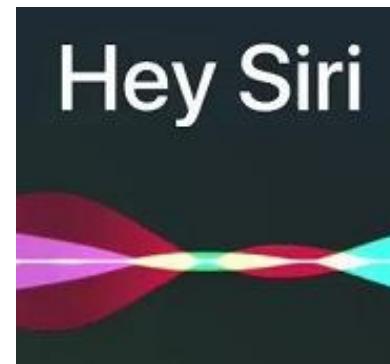
# What is question answering over knowledge graphs all about?

# Question Answering: Vital for Search

What are some films directed  
by Nolan?



Google Assistant

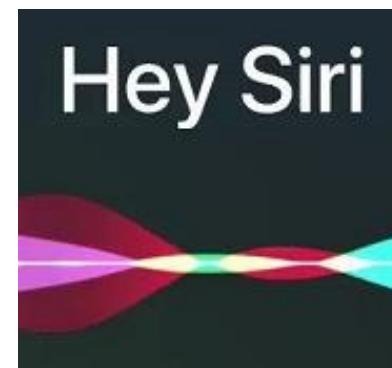


# Question Answering: Vital for Search

What are some films directed by Nolan?



Google Assistant



Question Answering over Curated and Open Web Sources

Christopher Nolan / Films directed



The Dark Knight  
2008



Interstellar  
2014

# Question Answering: Vital for Search

What are some films directed by Nolan?

- Direct answers to questions
- Enabled by knowledge graphs
- Saves time and effort
- Natural in voice UI

Christopher Nolan / Films directed



The Dark Knight  
2008



Interstellar  
2014

# Play with QA

GERMANY / CAPITAL

Berlin

Krzysztof Baranowski



what is the currency of Denmark



DENMARK / CURRENCY

Danish krone

hayatikayhan

what is the capital of Germany



which club does Neymar play for



NEYMAR / CURRENT TEAM

Paris  
Saint-G  
erman F.



what is the dwarf language called in lord of the rings

Khuzdul

Khuzdul was the **language** of the **Dwarves**, written in the 50-letter Cirth script (Runes). It appears to be structured, like real-world Semitic **languages**, around the triconsonantal roots: kh-z-d, b-n-d, z-g-l.



where is sigir 2020



ALL

IMAGES

NEWS

VIDEOS

SHOPPING

SIGIR 2020 / Location



**SIGIR**  
Special Interest Group  
Information Retrieval

**Xi'an, China**

Simple questions involving one entity and relation

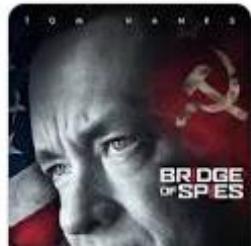
which films star tom hanks and are directed by spielberg

# Play with QA

All News Images Videos Shopping More

Settings Tools

Steven Spielberg / Films directed / Tom Hanks / Movies



Bridge of Spies



The Terminal



Catch Me If You Can



The Post



Saving Private Ryan

Complex questions involving multiple entities and relations

who played Batman in Dark Knight



THE DARK KNIGHT / BATMAN / PLAYED BY

Christian  
Bale

Ernesto Ruscio, FilmMagic

what was spielberg's father's profession

All Images News Videos Shopping More

Settings Tools

About 3.690.000 results (0,75 seconds)

Arnold Spielberg / Profession

Electrical engineer

Conversational  
questions with  
implicit context

and what about Alfred



THE DARK KNIGHT / ALFRED PENNYWORTH /  
PLAYED BY

Michael  
Caine

Jeff Spicer, Getty Images

where was the father of messi born

All Maps Images News

About 4.850.000 results (0,88 seconds)

Jorge Messi / Born

1958

age 62 years

what was Nolan's first film  
with Christian Bale



Edit

Christian Bale first movie

Born in Haverfordwest, Wales, to English parents, Bale had his first starring role at age 13 in Steven Spielberg's war film **Empire of the Sun** (1987).

which Spielberg films won more than three Oscars

<https://en.m.wikipedia.org/wiki>

List of awards and nominations received by Steven Spielberg - Wikipedia

movies with Tom Hanks

Tom Hanks

Actor



VIEW QUOTES MOVIES PEOPLE ALSO ASK FOR

co-starring Julia Roberts



Here are some pictures



Julia Roberts, Sissy...  
wtvq.com

**Play with QA: Try out different formulations, entities, domains, complexities, assistants, sources, languages.... to expose brittleness of SoTA and take community forward!**

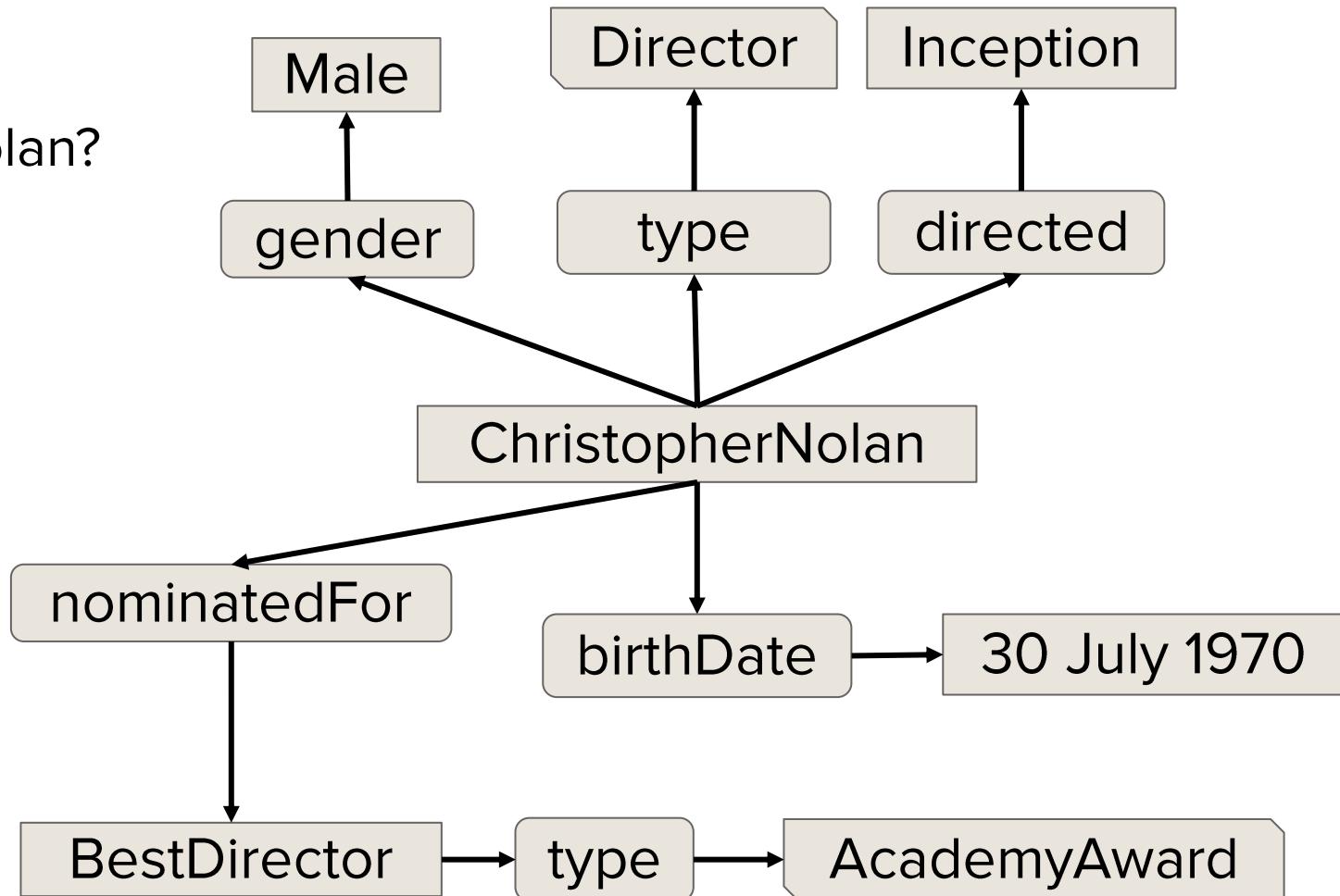
**Significant** progress has been made on **knowledge base construction** over the last fifteen years or so; but for **question answering**, which is one of the most valuable applications of KBs, we are still at the **tip of iceberg!**

# QA over knowledge graphs (KG-QA)

What are the Oscar nominations of Nolan?

# QA over knowledge graphs (KG-QA)

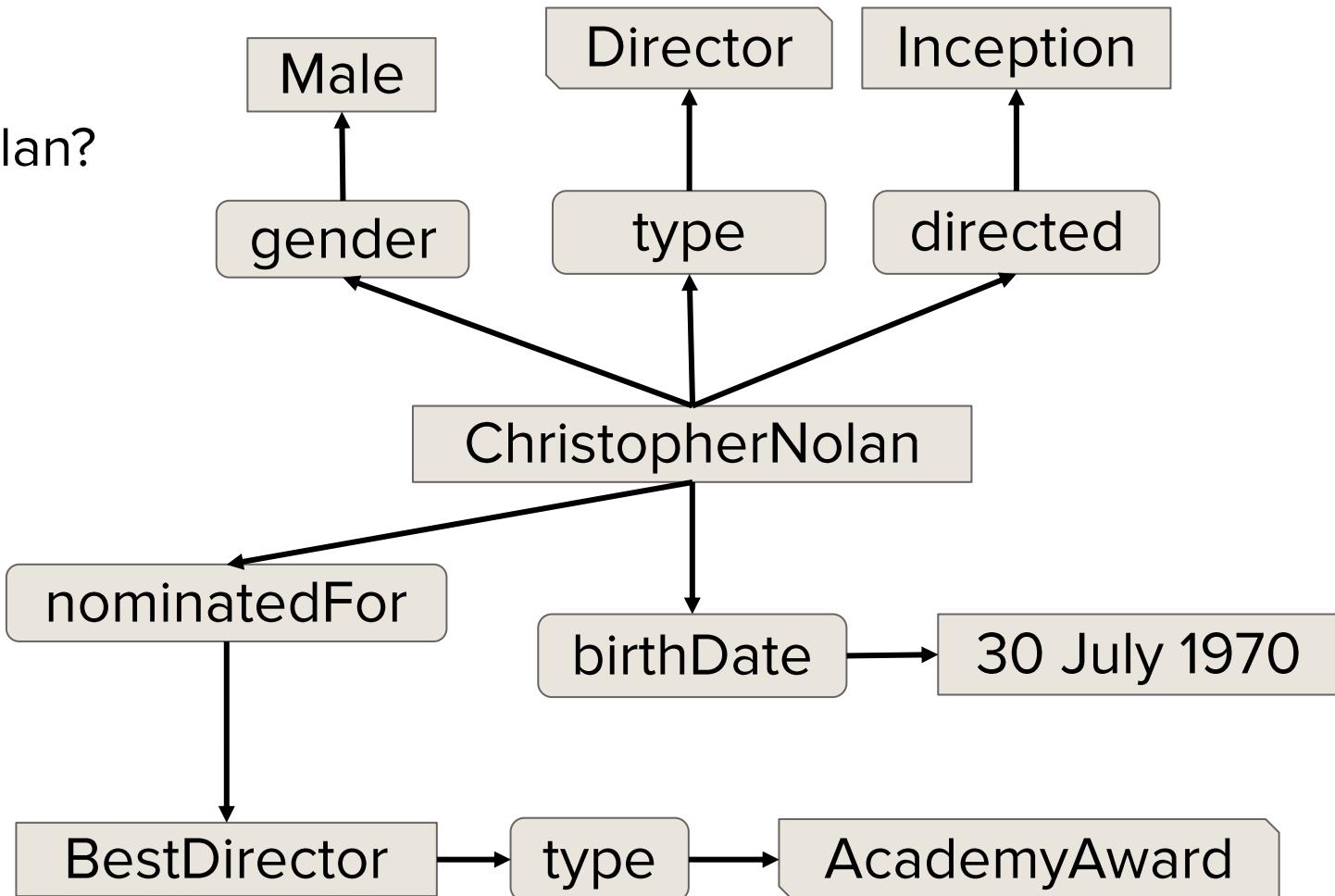
What are the Oscar nominations of Nolan?



# QA over knowledge graphs (KG-QA)

What are the Oscar nominations of Nolan?

- YAGO [[Suchanek et al. 2007](#)]
- DBpedia [[Auer et al. 2007](#)]
- Freebase [[Bollacker et al. 2008](#)]
- Wikidata [[Vrandečić and Krötzsch 2014](#)]



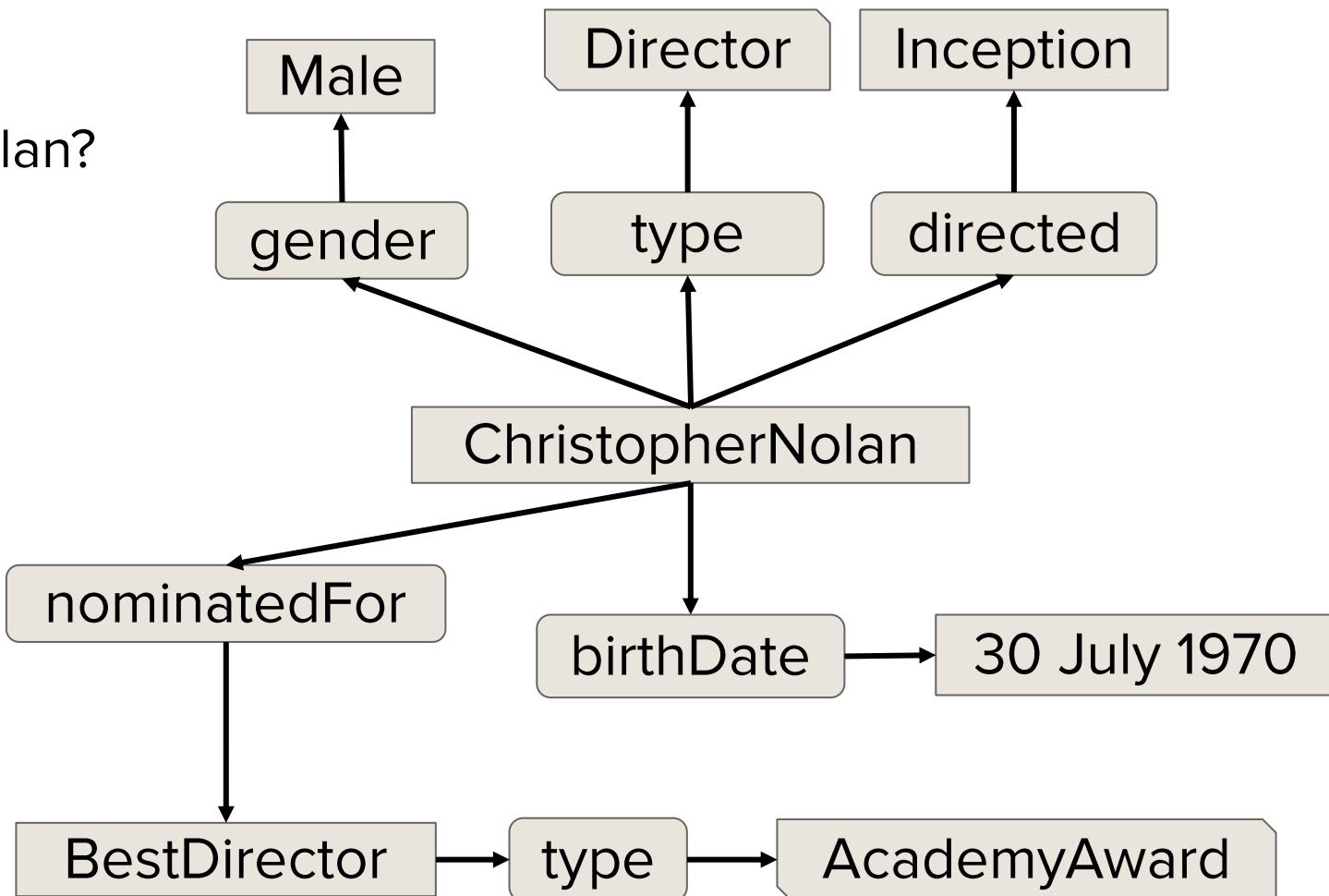
# QA over knowledge graphs (KG-QA)

What are the Oscar nominations of Nolan?

- YAGO [Suchanek et al. 2007]
- DBpedia [Auer et al. 2007]
- Freebase [Bollacker et al. 2008]
- Wikidata [Vrandečić and Krötzsch 2014]

**Terminology varies across KGs**

**Here:** Entities, predicates, types, literals

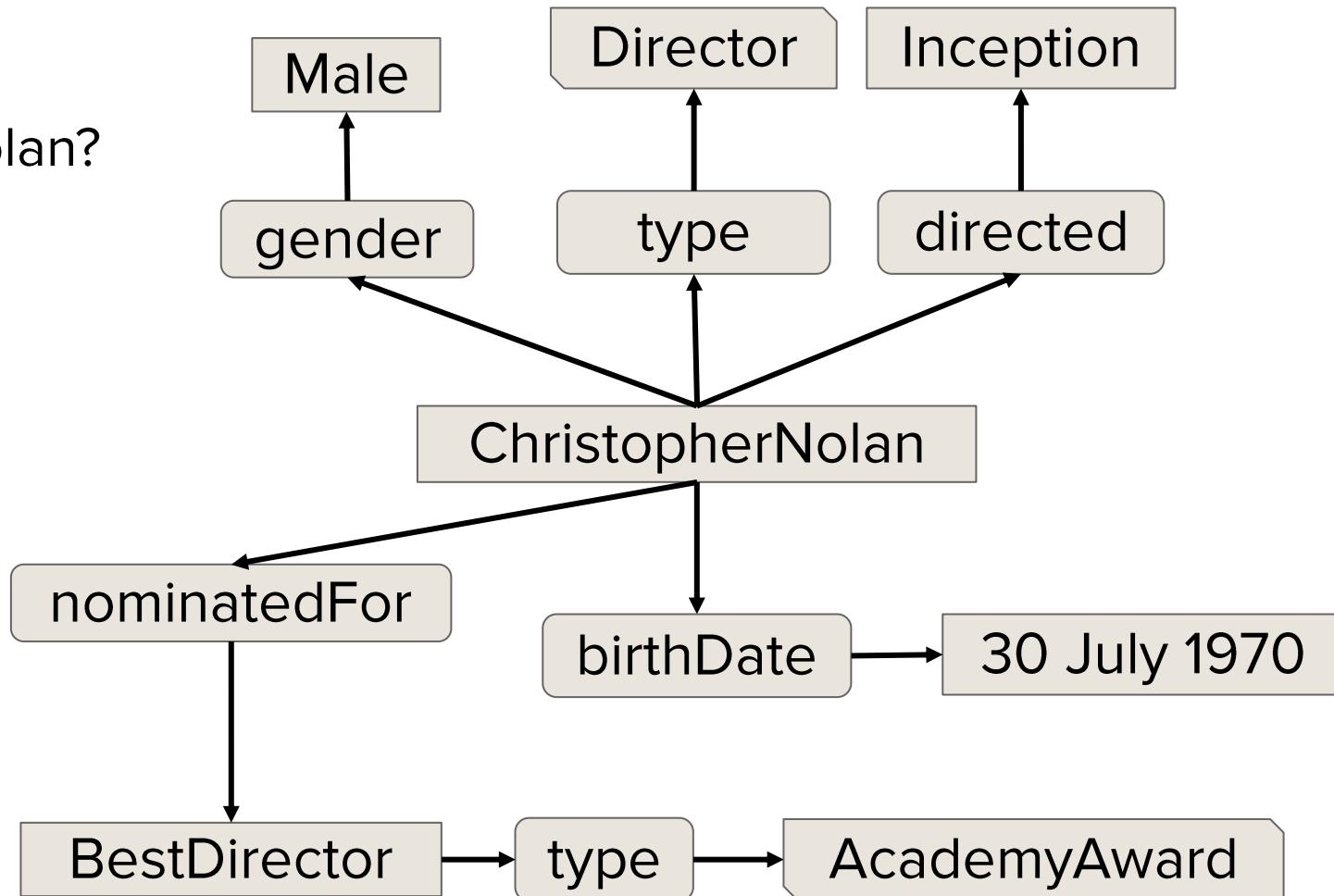


# QA over knowledge graphs (KG-QA)

What are the Oscar nominations of Nolan?

- YAGO [Suchanek et al. 2007]
- DBpedia [Auer et al. 2007]
- Freebase [Bollacker et al. 2008]
- Wikidata [Vrandečić and Krötzsch 2014]

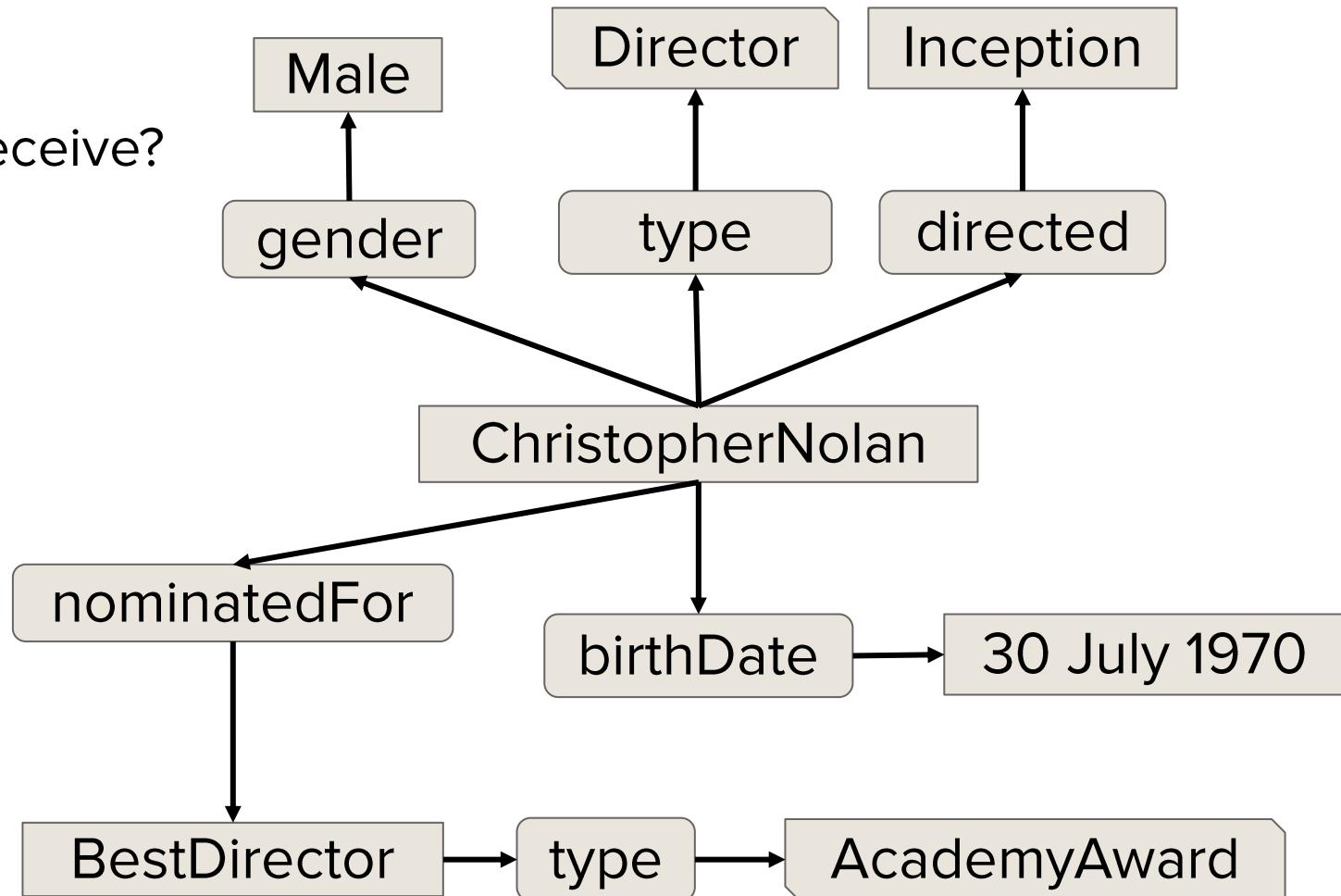
**Wikidata: 12B facts, 84M entities,  
7k predicates, 69k types**



# KGs and KBs are equivalent

Which Oscar nominations did Nolan receive?

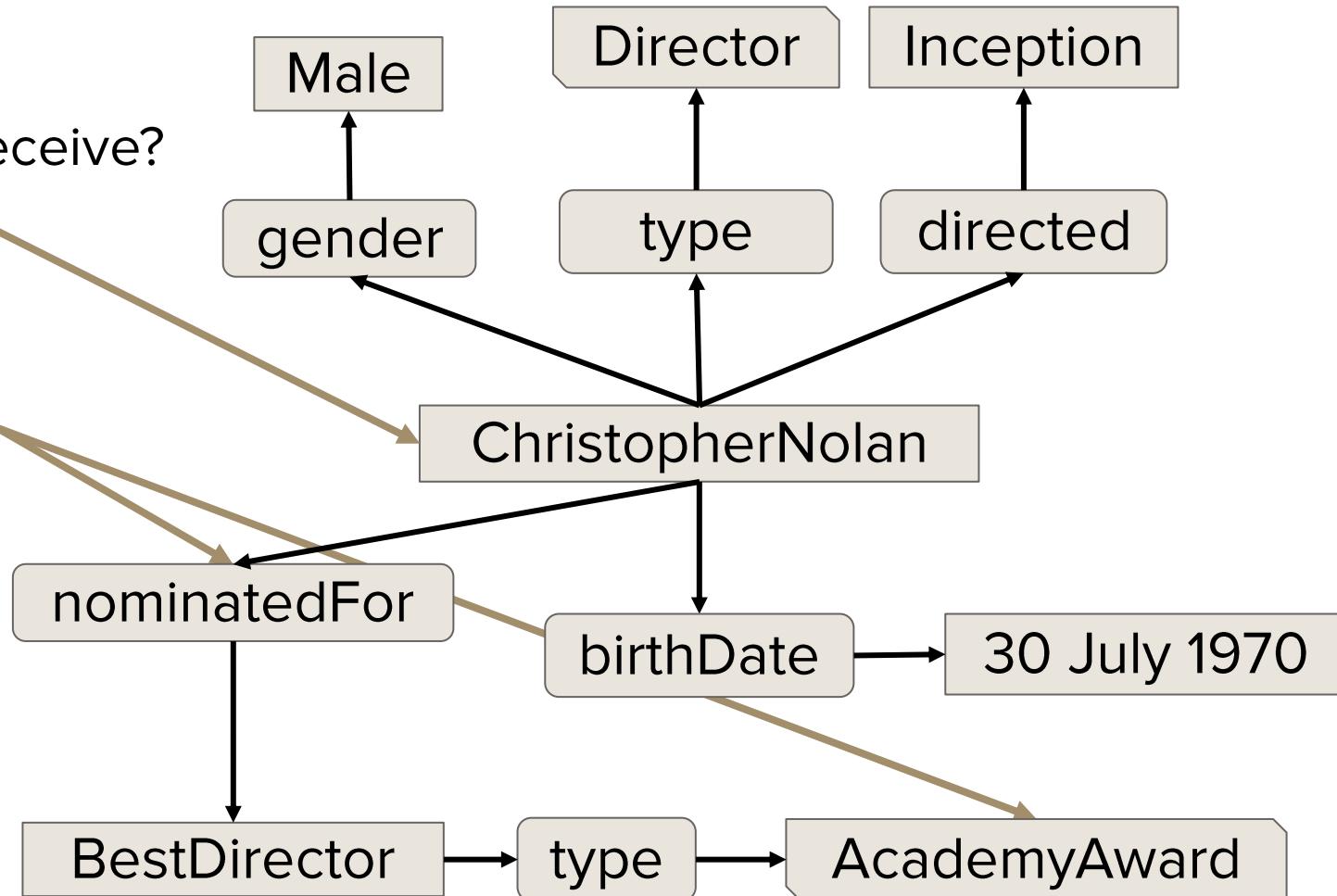
```
<ChristopherNolan, gender, Male>
<ChristopherNolan, type, Director>
<ChristopherNolan, directed, Inception>
<ChristopherNolan, nominatedFor, BestDirector>
<BestDirector, type, AcademyAward>
<ChristopherNolan, birthDate, 30 July 1970>
```



# KG-QA Challenge 1: Bridge vocabulary gap

Which Oscar nominations did Nolan receive?

- <ChristopherNolan, gender, Male>
- <ChristopherNolan, type, Director>
- <ChristopherNolan, directed, Inception>
- <ChristopherNolan, nominatedFor, BestDirector>
- <BestDirector, type, AcademyAward>
- <ChristopherNolan, birthDate, 30 July 1970>



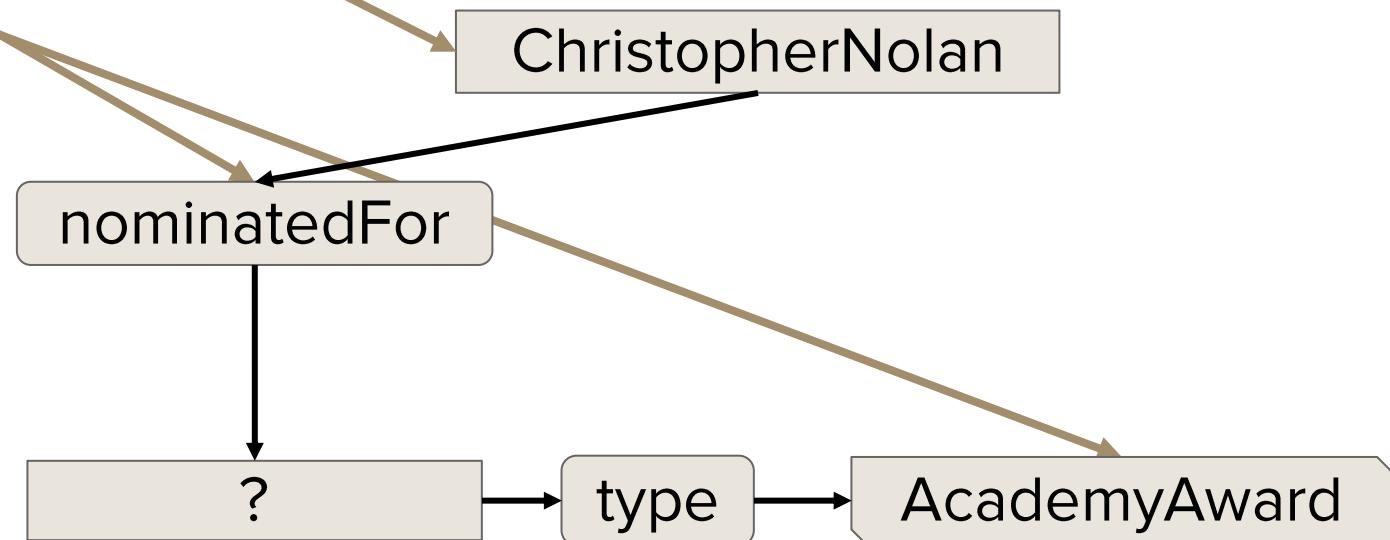
# KG-QA Challenge 2: Query formulation

Which Oscar nominations did Nolan receive?

<ChristopherNolan, gender, Male>  
<ChristopherNolan, type, Director>  
<ChristopherNolan, directed, Inception>  
<ChristopherNolan, nominatedFor, BestDirector>  
<BestDirector, type, AcademyAward>  
<ChristopherNolan, birthDate, 30 July 1970>

SELECT ?ANS  
WHERE {  
 ChristopherNolan nominatedFor ?ANS .  
 ?ANS type AcademyAward }

**SPARQL**



# KG-QA Challenge 2: Query formulation

Which Oscar nominations did Nolan receive?

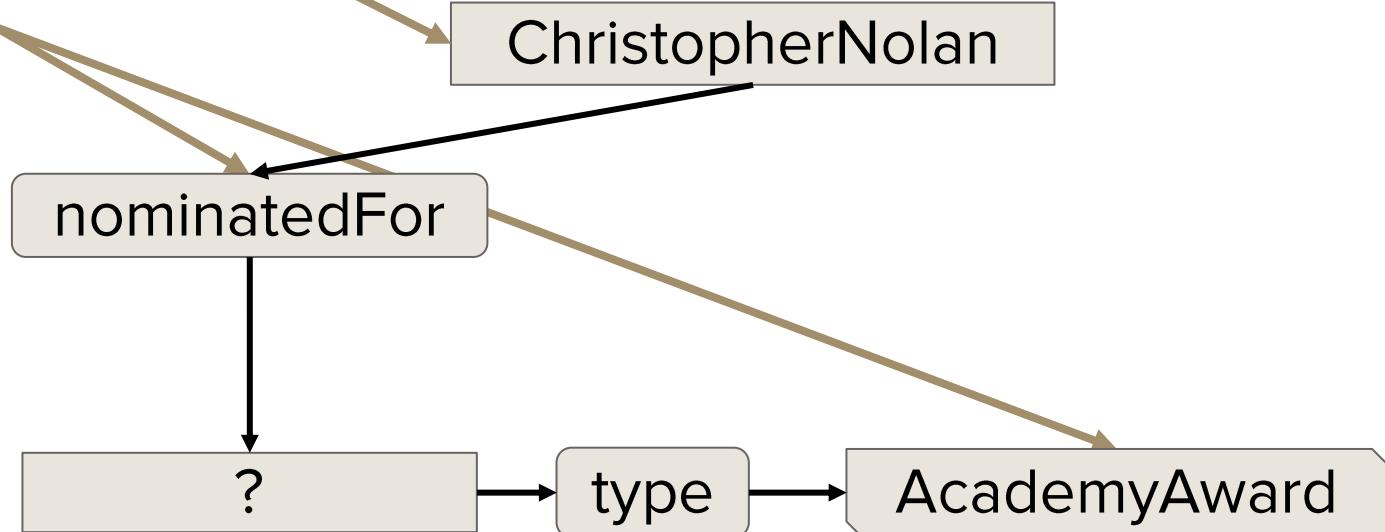
<ChristopherNolan, gender, Male>  
<ChristopherNolan, type, Director>  
<ChristopherNolan, directed, Inception>  
<ChristopherNolan, nominatedFor, BestDirector>  
<BestDirector, type, AcademyAward>  
<ChristopherNolan, birthDate, 30 July 1970>

SELECT ?ANS  
WHERE {  
 ChristopherNolan nominatedFor ?ANS .  
 ?ANS type AcademyAward }

SPARQL

**Named Entity Recognition and Disambiguation (NERD)** systems (aka Entity Detection and Linking):  
[TagME](#), [AIDA](#), [Dandelion](#), [Google NL API](#), [MS Text Analytics](#), [IBM NLU](#)

**Named Entity Recognition (NER):** [Stanford NER](#), [spaCy](#)



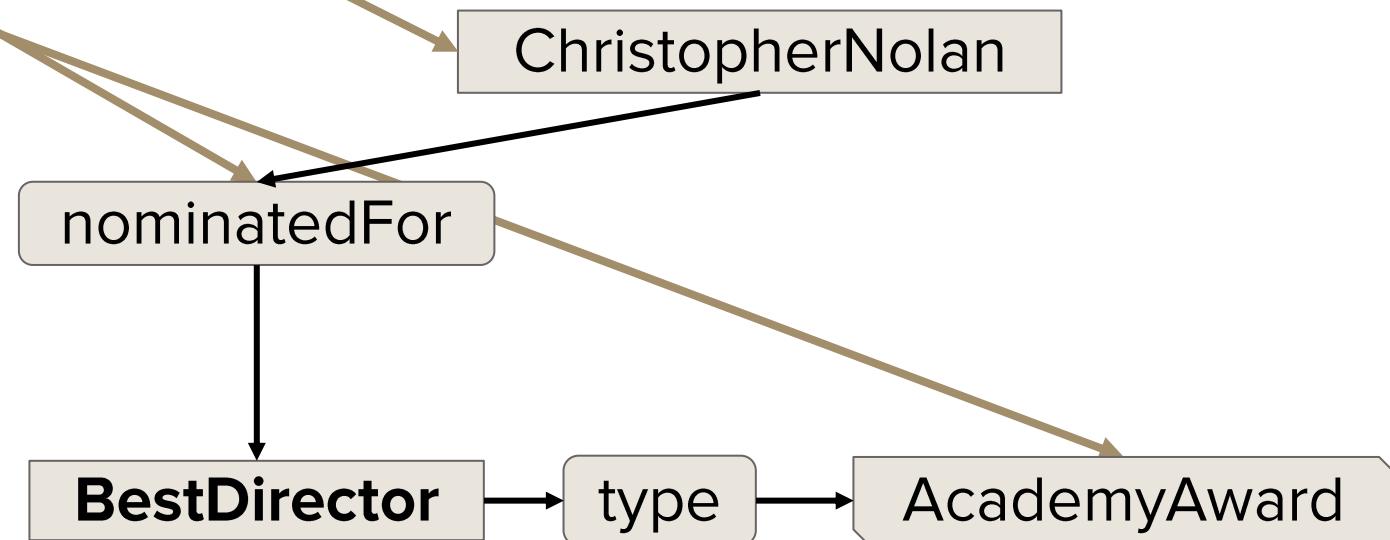
# Answering with query

Which Oscar nominations did Nolan receive?

<ChristopherNolan, gender, Male>  
<ChristopherNolan, type, Director>  
<ChristopherNolan, directed, Inception>  
**<ChristopherNolan, nominatedFor, BestDirector>**  
**<BestDirector, type, AcademyAward>**  
<ChristopherNolan, birthDate, 30 July 1970>

SELECT ?ANS  
WHERE {  
 ChristopherNolan nominatedFor ?ANS .  
 ?ANS type AcademyAward }

**BestDirector**



# Structured queries and logical forms

Which Oscar nominations did Nolan receive?

## Neo4j CYPHER Graph QL

```
MATCH  
(subj {name: 'ChristopherNolan' })-[:nominatedFor]-  
>(obj:AcademyAward) RETURN obj.name
```

## SPARQL BGP

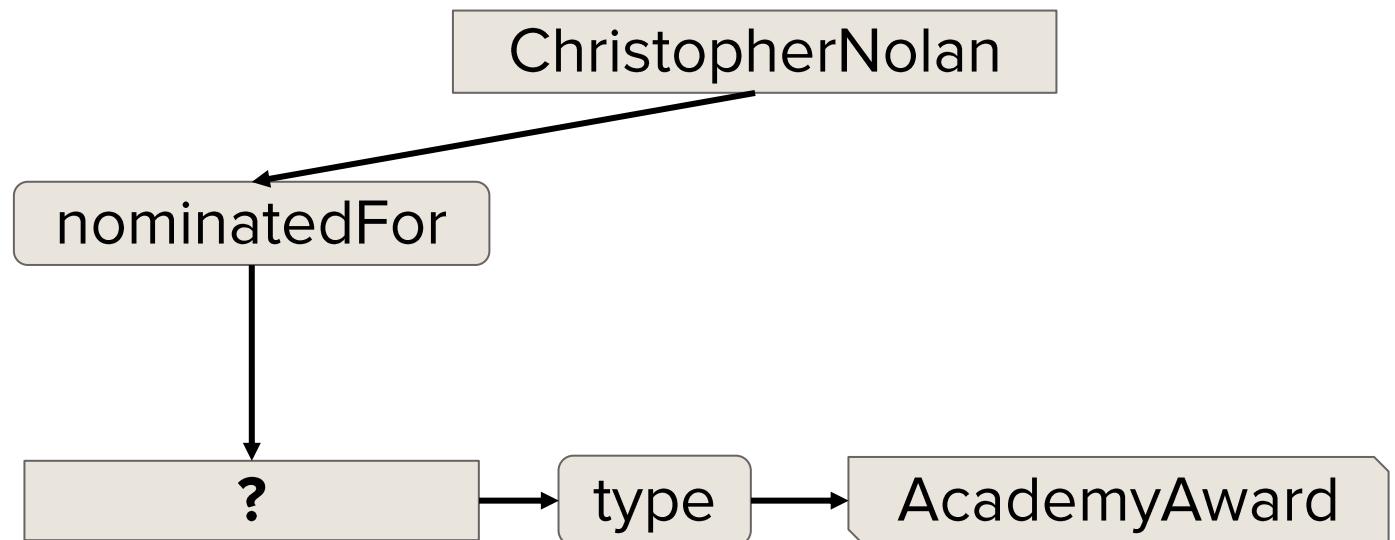
```
SELECT ?ANS  
WHERE {  
  ChristopherNolan nominatedFor ?ANS .  
  ?ANS type AcademyAward }
```

## BestDirector

### Lambda-calculus

$$\lambda x. \text{nominatedFor}(\text{ChristopherNolan}, x) \wedge \text{Type}(x, \text{AcademyAward})$$

### Lambda-DCS

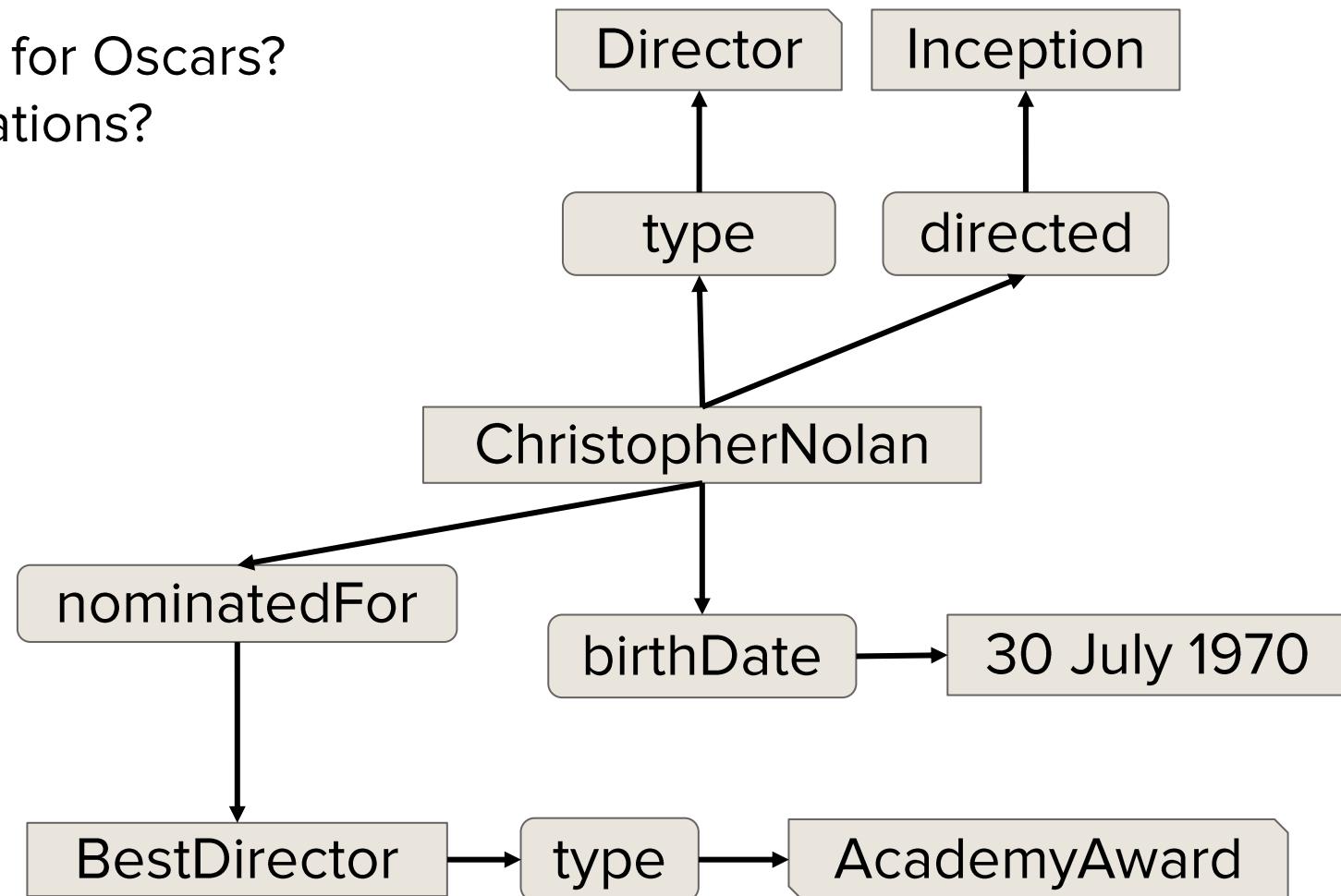
$$\text{nominatedFor}.\text{ChristopherNolan} \sqcap \text{type}.\text{AcademyAward}$$


# Reification: n-ary information in KGs

For which films was Nolan nominated for Oscars?

When did Nolan get his Oscar nominations?

```
<ChristopherNolan, gender, Male>
<ChristopherNolan, type, Director>
<ChristopherNolan, directed, Inception>
<ChristopherNolan, nominatedFor, BestDirector>
<BestDirector, type, AcademyAward>
<ChristopherNolan, birthDate, 30 July 1970>
```

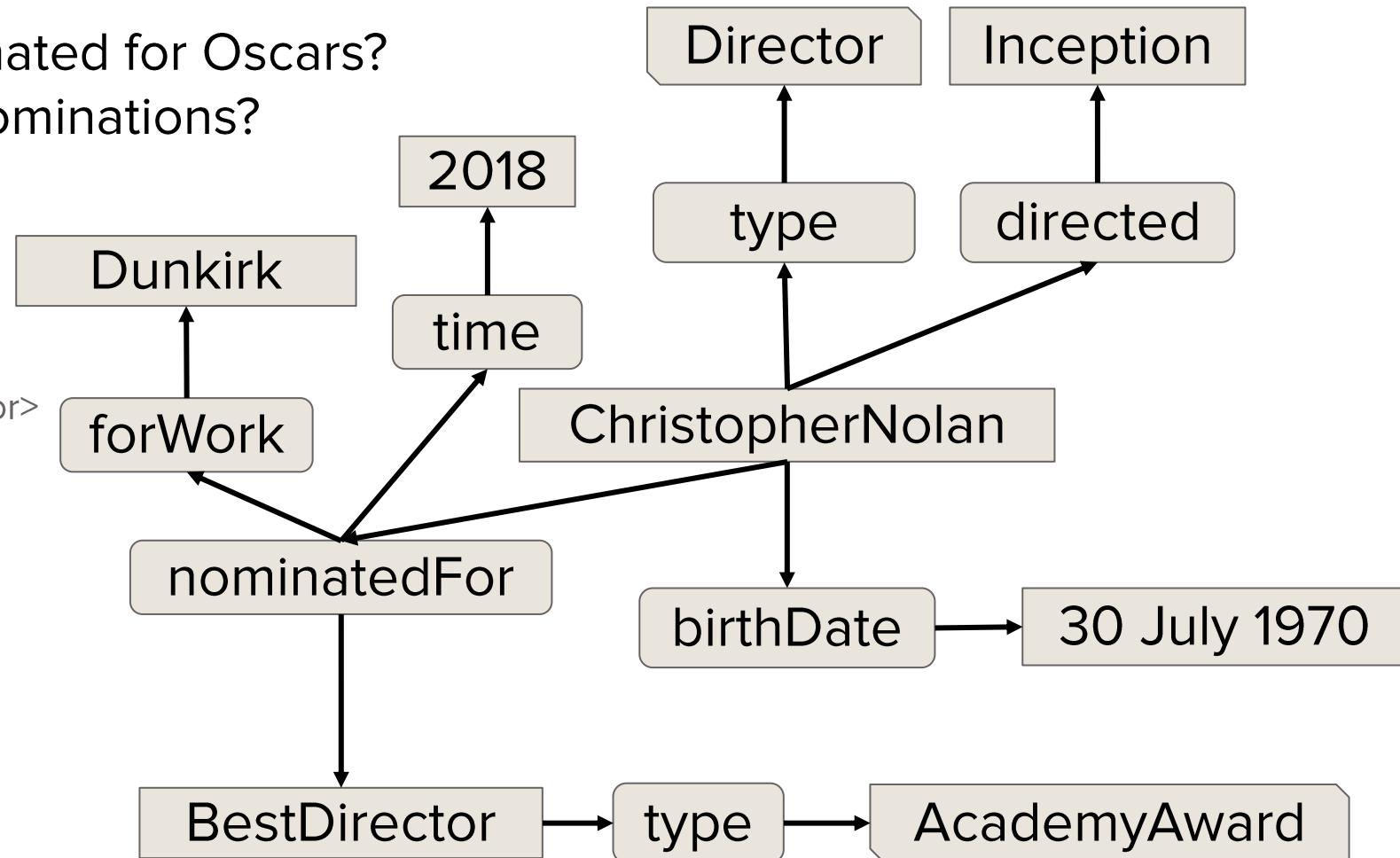


# Reification: n-ary information in KGs

For which films was Nolan nominated for Oscars?

When did Nolan get his Oscar nominations?

```
<ChristopherNolan, gender, Male>
<ChristopherNolan, type, Director>
<ChristopherNolan, directed, Inception>
<ChristopherNolan, nominatedFor, BestDirector>
<BestDirector, type, AcademyAward>
<ChristopherNolan, birthDate, 30 July 1970>
```



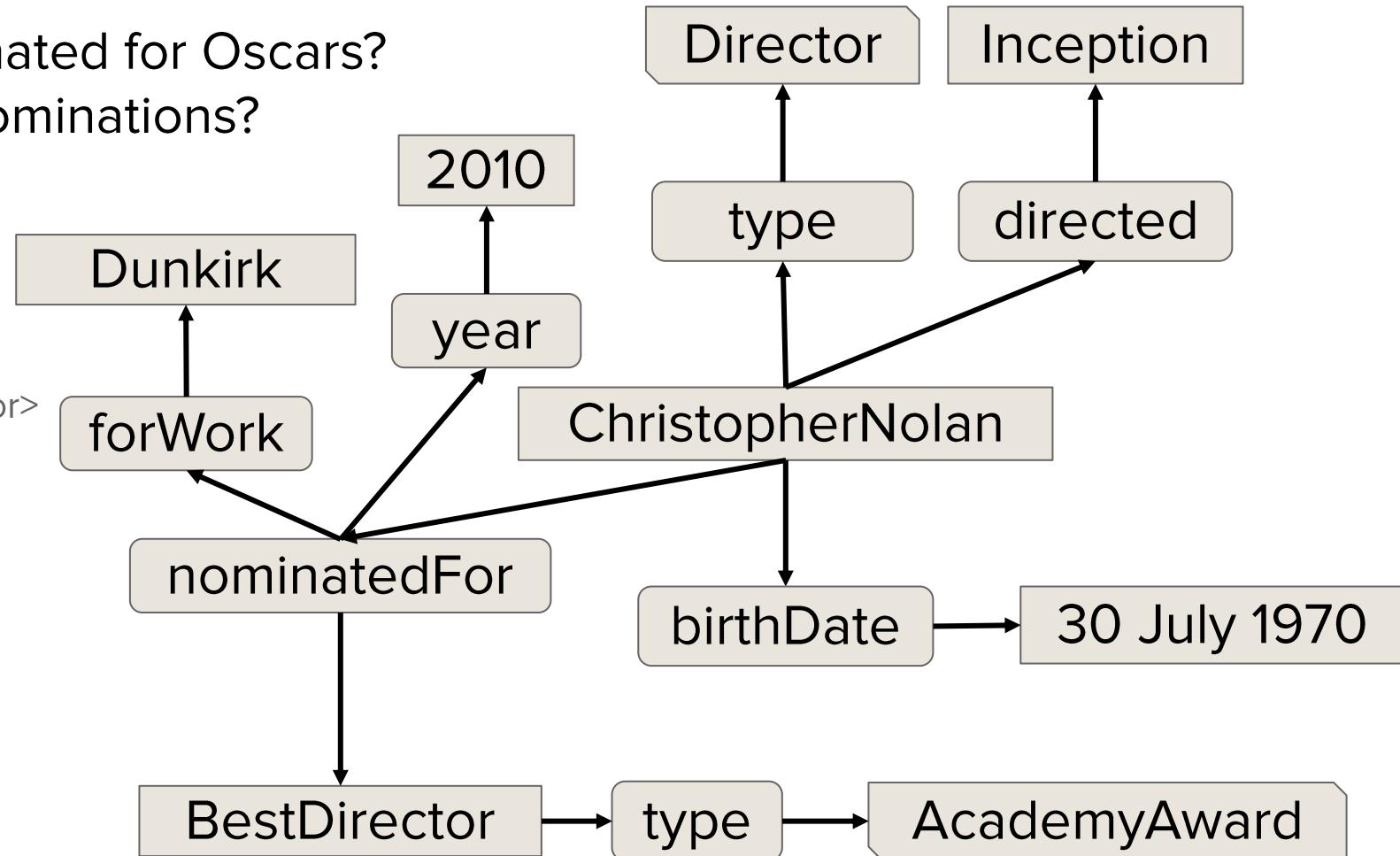
# Reification: n-ary information in KGs

For which films was Nolan nominated for Oscars?

When did Nolan get his Oscar nominations?

```
<ChristopherNolan, gender, Male>
<ChristopherNolan, type, Director>
<ChristopherNolan, directed, Inception>
<ChristopherNolan, nominatedFor, BestDirector>
<BestDirector, type, AcademyAward>
<ChristopherNolan, birthDate, 30 July 1970>
```

```
<ChristopherNolan, nominatedFor, 123>
<123, nominatedFor, BestDirector>
<123, forWork, Dunkirk>
<123, year, 2010>
```



# Qualifiers are a huge part of Wikidata

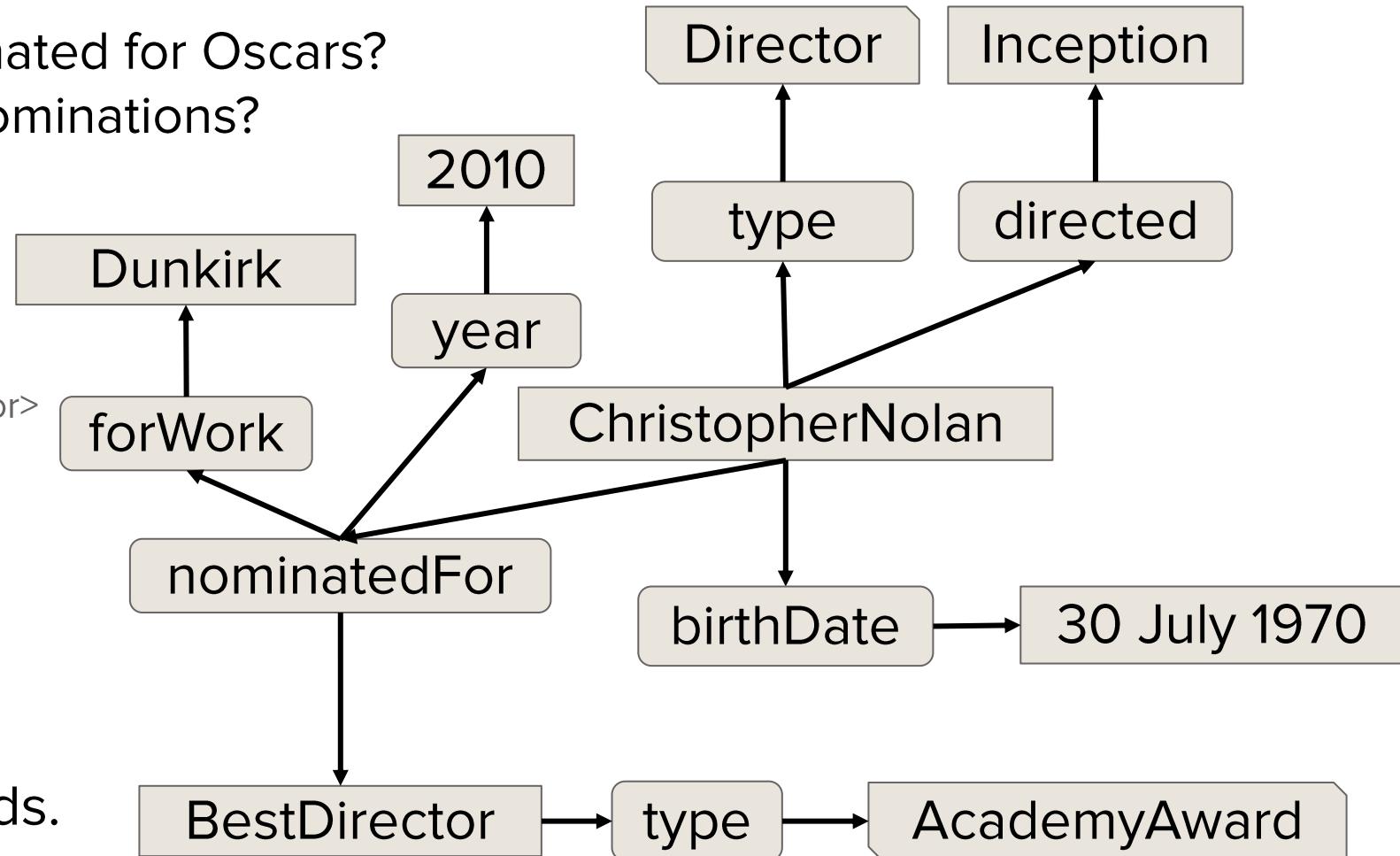
For which films was Nolan nominated for Oscars?

When did Nolan get his Oscar nominations?

```
<ChristopherNolan, gender, Male>
<ChristopherNolan, type, Director>
<ChristopherNolan, directed, Inception>
<ChristopherNolan, nominatedFor, BestDirector>
<BestDirector, type, AcademyAward>
<ChristopherNolan, birthDate, 30 July 1970>
```

```
<ChristopherNolan, nominatedFor, 123>
<123, nominatedFor, BestDirector>
<123, forWork, Dunkirk>
<123, year, 2010>
```

Wikidata: Qualifiers, Statement-Ids.  
**6B** triples part of reified facts!!



# Questions that need reified triples

Who played Cobb in Inception?

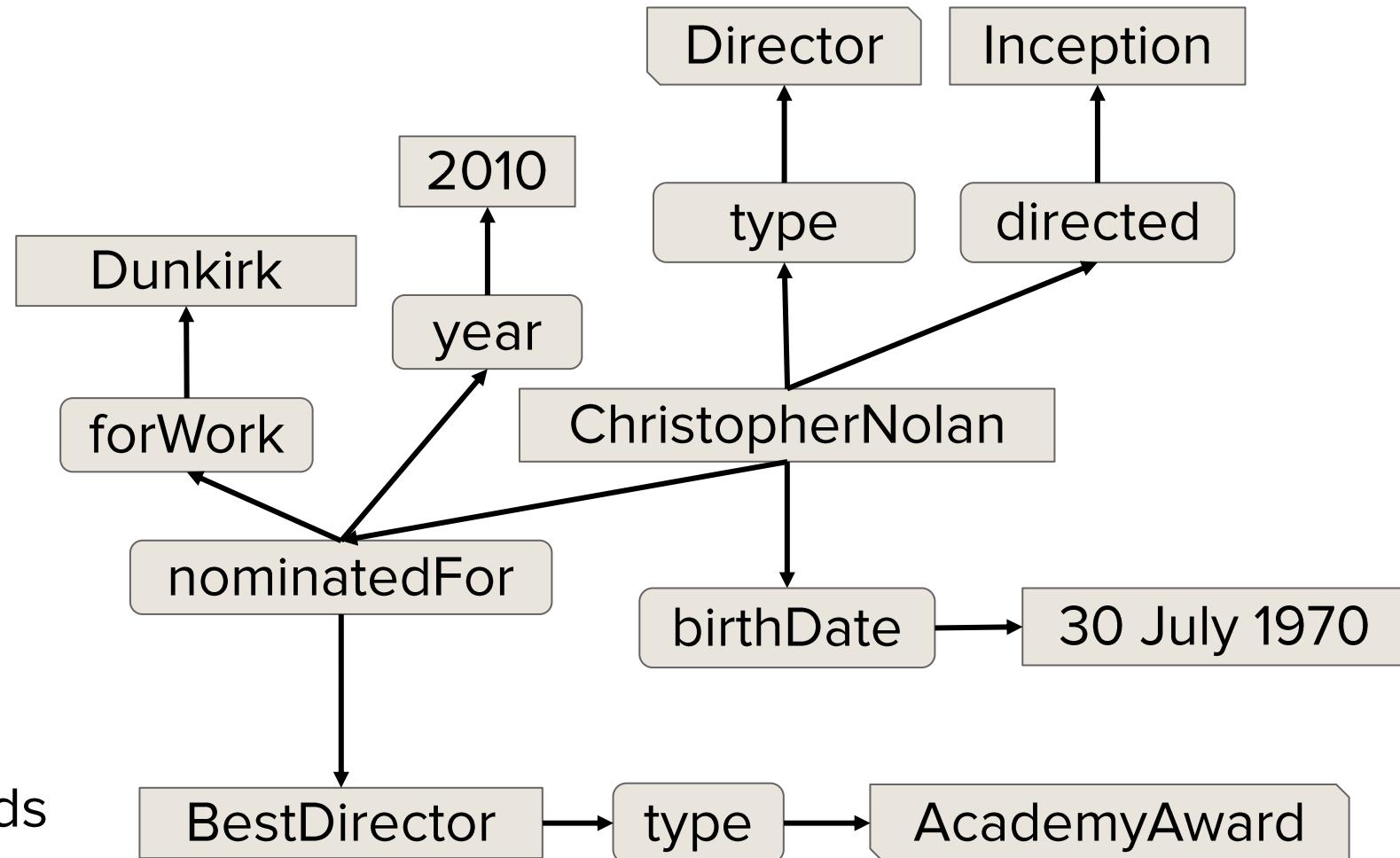
Who did Leo play in Inception?

When did Neymar join PSG?

Who was Trump's first wife?

US president in 2016?

...



Wikidata: Qualifiers, Statement-Ids

**6B** triples part of reified facts!!

# Explore Wikidata

Entity name / Subject

Entity id

(Q163872)

Entity desc

The Dark Knight

2008 British-American superhero film directed by Christopher Nolan

TDK | Dark Knight

Entity aliases

Type

Type predicate

instance of

film

Predicate

genre

action film

nominated for

Academy Award for Best Supporting Actor

statement is subject of

81st Academy Awards

nominee

Heath Ledger

point in time

22 February 2009

Object

part of the series

The Dark Knight Trilogy

follows

Batman Begins

followed by

The Dark Knight Rises

series ordinal

2

cast member

Christian Bale

character role

Bruce Wayne

▶ 11 references

Michael Caine

character role

Alfred Pennyworth

▶ 4 references

Heath Ledger

character role

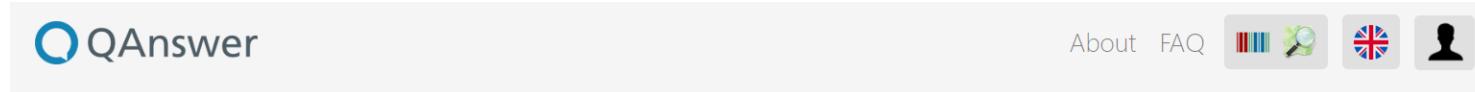
Joker

▶ 9 references

# Explore Wikidata like a pro

- Wikidata: [https://www.wikidata.org/wiki/Wikidata:Main\\_Page](https://www.wikidata.org/wiki/Wikidata:Main_Page)
- Wikidata data model: <https://www.mediawiki.org/wiki/Wikibase/DataModel/Primer>
- Wikidata dumps: [https://www.wikidata.org/wiki/Wikidata:Database\\_download](https://www.wikidata.org/wiki/Wikidata:Database_download)
- Download latest **n-triples** dump: <https://dumps.wikimedia.org/wikidatawiki/entities/>
- Wikidata SPARQL Endpoint: <https://query.wikidata.org/>
- Wikidata statistics: <https://stats.wikimedia.org/#/wikidata.org>
- More stats: <https://www.wikidata.org/wiki/Wikidata:Statistics>

# Play with QA (over Wikidata)



Enter your question...

Go

Who is Bach? Who are the Beatles's members? What is the music genre of Bob Marley? In which countries are the alps?

When was D-Day? post boxes in munich Where is the inventor of dynamite born? Give me songs of Pink Floyd.

Give me actors starring in the Lord of the Rings. Sherlock Holmes What is the surface of Liechtenstein? Who is Tom Cruise?

Who is the prime minister of France? atomic number of polonium bars in borgomasino

Who are the members of Green Day? museums in berlin brands of soft drinks What are the borders of Mexico?

Diefenbach et al., QAnswer: A Question answering prototype bridging the gap between a considerable part of the LOD cloud and end-users, WWW 2019. Available at <https://qanswer-frontend.univ-st-etienne.fr/>

# Benchmarks

- **Simple questions**
  - [WebQuestions \(Berant et al. 2013\)](#) over Freebase
  - [SimpleQuestions \(Bordes et al. 2015\)](#) over Freebase
- **Complex questions**
  - [LC-QuAD 2.0 \(Dubey et al 2018\)](#) over Wikidata + DBpedia
  - [MetaQA \(Zhang et al. 2018\)](#) over Freebase
- **Conversational questions**
  - [ConvQuestions \(Christmann et al. 2019\)](#) over Wikidata
  - [CSQA \(Saha et al. 2018\)](#) over Wikidata

# Benchmarks

- **Simple questions**

- [WebQuestions\\*](#) (Berant et al. 2013) over Freebase 
- [SimpleQuestions](#) (Bordes et al. 2015) over Freebase

Recent benchmarks over Wikidata

More realistic benchmarks are smaller but harder

Much higher numbers on semi-synthetic benchmarks

“Vulnerable” to neural methods

\* Need reified triples for answering

- **Complex questions**

- [LC-QuAD 2.0\\*](#) (Dubey et al 2018) over Wikidata + DBpedia 
- [MetaQA](#) (Zhang et al. 2018) over Freebase

- **Conversational questions**

- [ConvQuestions\\*](#) (Christmann et al. 2019) over Wikidata 
- [CSQA](#) (Saha et al. 2018) over Wikidata

# Benchmarks

- **Simple questions**
  - [WebQuestions\\*](#) (Berant et al. 2013) over Freebase 
  - [SimpleQuestions](#) (Bordes et al. 2015) over Freebase
- **Complex questions**
  - [LC-QuAD 2.0\\*](#) (Dubey et al 2018) over Wikidata + DBpedia 
  - [MetaQA](#) (Zhang et al. 2018) over Freebase
- **Conversational questions**
  - [ConvQuestions\\*](#) (Christmann et al. 2019) over Wikidata 
  - [CSQA](#) (Saha et al. 2018) over Wikidata

**Many, many more:**

LC-QuAD  
([Trivedi et al. 2017](#))

ComQA  
([Abujabal et al. 2019](#))

GraphQuestions  
([Su et al. 2016](#))

QALD  
([Usbeck et al. 2018](#))

TempQuestions  
([Jia et al. 2018](#))

ComplexWebQuestions  
([Talmor and Berant 2018](#))

WikiMovies  
([Miller et al. 2016](#))

ComplexQuestions  
([Bao et al. 2016](#))

# Benchmarks: WebQuestions

- Real questions: Collected using the Google Suggest API
- Mostly simple questions using one fact or reified triple
- 3778 train, 2032 test questions
- Available at: <https://nlp.stanford.edu/software/sempre/>

who was richard nixon married to?  
what high school did harper lee go to?  
what was the capital city of the east roman empire?  
who plays ken barlow in coronation street?  
where is the fukushima daiichi nuclear plant located?

# Benchmarks: LC-QuAD 2.0

- Sampled SPARQL queries via templates, verbalized by crowdworkers
- Complex (and simple) questions involving multiple entities and relations
- 23954 train, 6046 test questions
- Available at: <http://lc-quad.sda.tech/>

What city is the twin city of Oslo and also the setting for “A Tree Grows in Brooklyn”?

What Empire used to have Istanbul as its capital?

How long was Shirley Temple the United States Ambassador to Ghana?

Were Dutch and Hungarian the official languages of the Holy Roman Empire?

Who replaced Albus Dumbledore as headmaster of Hogwarts?

# Benchmarks: ConvQuestions

- Natural conversations by crowdworkers after choosing topic
- Both simple and complex
- Five domains
- 6720 train, 2240 dev, 2240 test conversations
- Available at:  
<https://convex.mpi-inf.mpg.de/>

Books	Movies	Soccer	Music	TV series
When was the first book of the book series The Dwarves published ?	Who played the joker in The Dark Knight?	Which European team did Diego Costa represent in the year 2018?	Led Zeppelin had how many band members?	Who is the actor of James Gordon in Gotham?
2003	Heath Ledger	Atlético Madrid	4	Ben McKenzie
What is the <b>name of the second book?</b>	When did he die?	Did they win the Super Cup the previous year?	Which was released first: Houses of the Holy or Physical Graffiti?	<b>What about Bullock?</b>
The War of the Dwarves	22 January 2008	No	Houses of the Holy	Donal Logue
Who is the author ?	<b>Batman actor?</b>	<b>Which club was the winner?</b>	<b>Is the rain song and immigrant song there?</b>	Creator?
Markus Heitz	Christian Bale	Real Madrid C.F.	No	Bruno Heller
In which city was he born ?	Director?	Which English club did Costa play for before returning to Atlético Madrid?	Who wrote those songs?	<b>Married to in 2017?</b>
Homburg	Christopher Nolan	Chelsea F.C.	Jimmy Page	Miranda Phillips Cowley
When was he born ?	<b>Sequel name?</b>	<b>Which stadium is this club's home ground?</b>	<b>Name of his previous band?</b>	<b>Wedding date first wife?</b>
10 October 1971	The Dark Knight Rises	Stamford Bridge Stadium	The Yardbirds	19 June 1993

# Metrics

- Answers as sets (for systems using explicit structured queries)
  - Precision, Recall, F1-Score
- Answers as ranked lists (systems w/o explicit queries: approx. graph search)
  - Precision@1, MRR, MAP
  - Hit@5
- Single answer
  - Accuracy

break duration ?x .  
?x measuredIn minutes .

# Outline: QA over knowledge graphs

- **Background:** Setup, benchmarks, metrics
- **Simple QA:** Templates and embeddings
- **Complex QA:** Multiple entities and predicates
- **Heterogeneous sources:** Handling KGs and text
- **Conversational QA:** Implicit context in multi-turn setup
- **Take-home:** Summary and insights

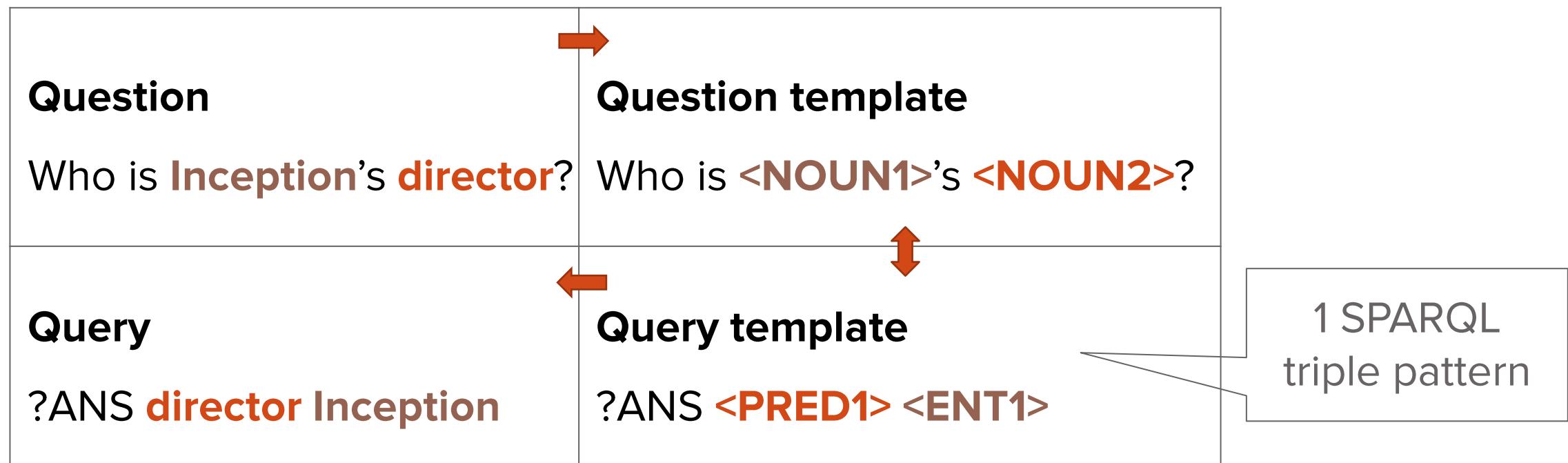
# Getting started: Templates and embeddings

# Foundational work in KG-QA

- Templates over RDF ([Unger et al. 2012](#))
- DEANNA ([Yahya et al. 2012, 2013](#))
- SEMPRE ([Berant et al. 2013](#))
- PARALEX + OQA ([Fader et al. 2013, 2014](#))
- Subgraph embeddings ([Bordes et al. 2014](#))
- STAGG ([Yih et al. 2015](#))
- AQQU ([Bast and Haussman 2015](#))

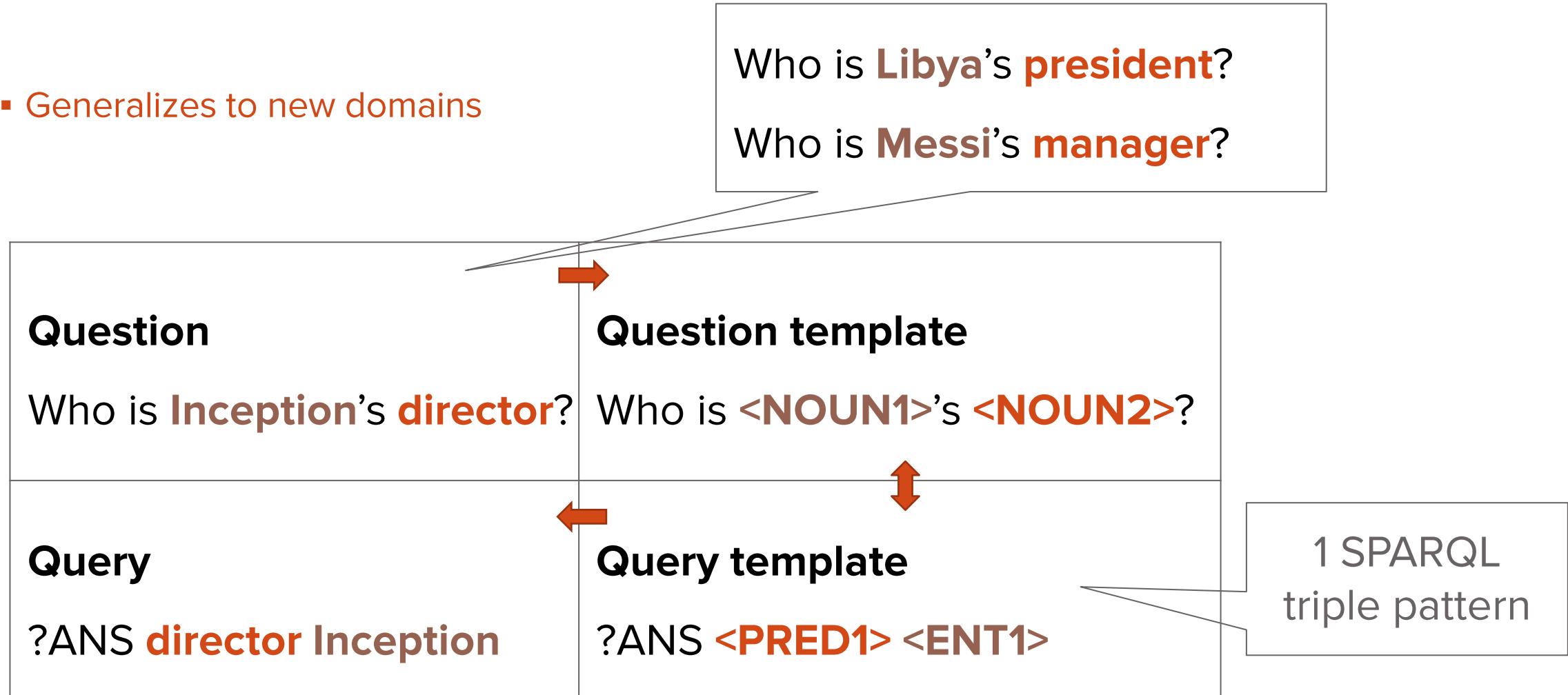
# Templates for KG-QA

- Interpretable



# Templates for KG-QA

- Generalizes to new domains



# Templates for KG-QA

## Question

Who plays the **role of Cobb** in  
**Inception?**

## Query

**Inception** castMember ?VAR  
?VAR castMember ?ANS  
?VAR **characterRole Cobb**



## Question template

Who <VERB> <DT> <**NOUN1**> <**PREP1**>  
<**NOUN2**> <**PREP2**> <**NOUN3**>?



## Query template

<**ENT1**> <**PRED1**> ?VAR  
?VAR <**PRED1**> ?ANS  
?VAR <**PRED2**> <**ENT2**>

Multiple SPARQL  
triple patterns

# Limitations of templates

- Hand-crafted by experts (Fader et al. 2013, 2014; Unger et al. 2012)
- Restricted coverage
- **Solution:** Learn templates
  - Question templates
  - Query templates
  - Slot alignments
- Proposed in the QUINT+NEQA framework (Abujabal et al. 2017, 2018)

Abujabal et al., Automated template generation for question answering over knowledge graphs, WWW 2017.

Abujabal et al., Never-ending learning for open-domain question answering over knowledge bases, WWW 2018.

# Distant supervision from QA pairs

**Question:** Which Oscar award nomination did Nolan get for the film Dunkirk?  
**Answer:** Best Director

NERD system

Dunkirk

NERD system

ChristopherNolan

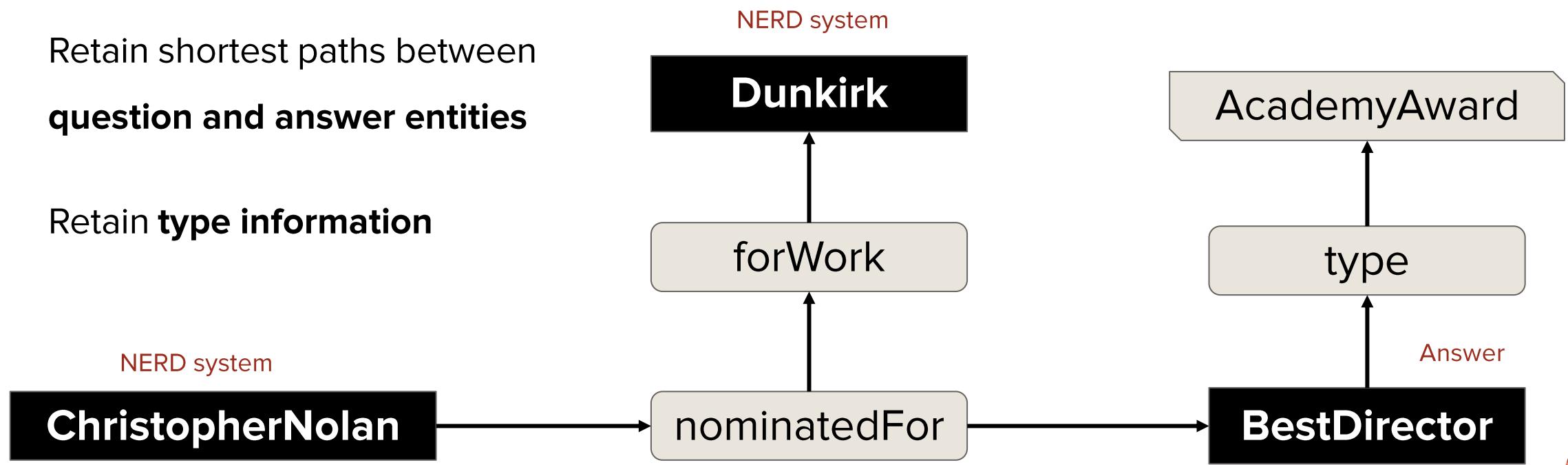
Answer

BestDirector

# Distant supervision from QA pairs

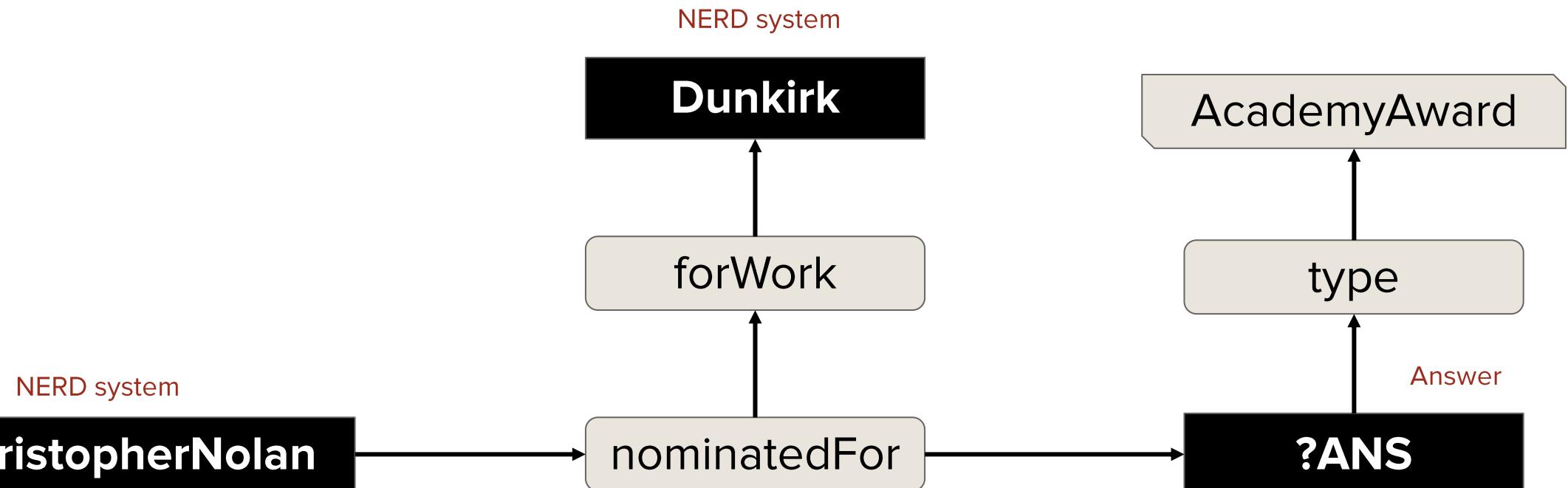
**Question:** Which Oscar award nomination did Nolan get for the film Dunkirk?  
**Answer:** Best Director

- Retain shortest paths between **question and answer entities**
- Retain **type information**

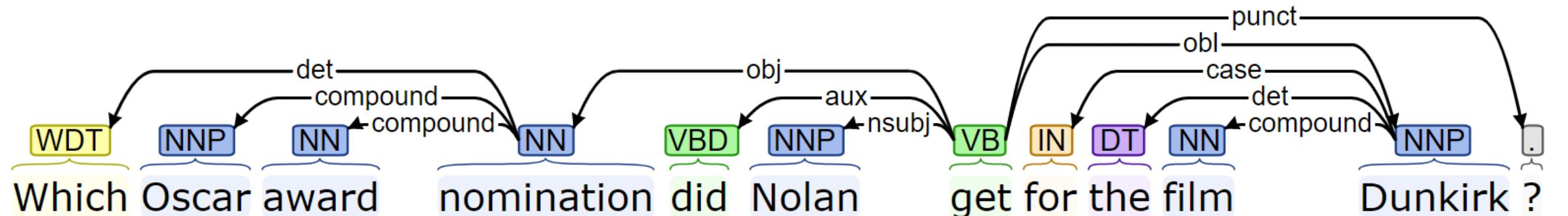


# Distant supervision from QA pairs

**Question:** Which Oscar award nomination did Nolan get for the film Dunkirk?  
**Answer:** Best Director  
**Query:** SELECT ?x WHERE { ChristopherNolan nominatedFor ?VAR . ?VAR nominatedFor ?ANS . ?VAR forWork Dunkirk . ?VAR type AcademyAward . }



# Extract question phrases



which nomination

oscar nomination

get nomination

for film

oscar

get

which

did get

did

oscar award

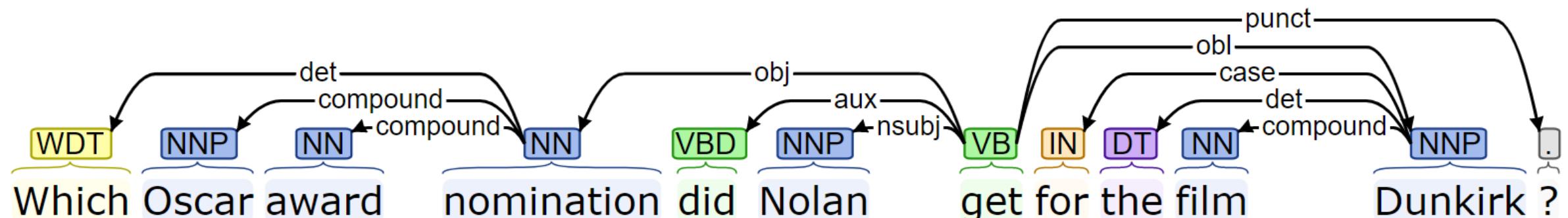
nomination

award

award nomination

Dependency parsing: <https://web.stanford.edu/~jurafsky/slp3/15.pdf>

# Extract query items



which nomination

oscar nomination

get nomination

for film

oscar

get

which

did get

did

oscar award

nomination

award

award nomination

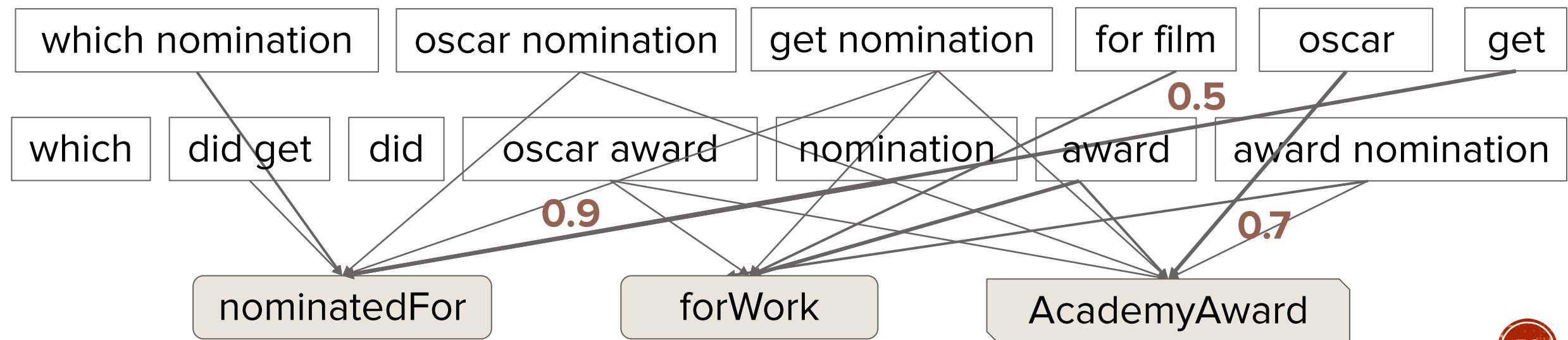
nominatedFor

forWork

AcademyAward

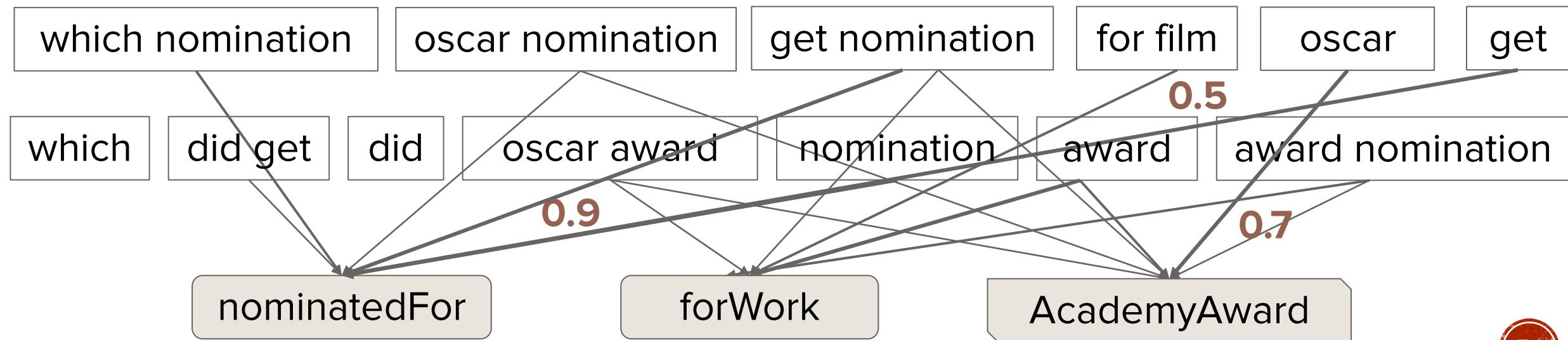
# Create candidate alignments

- **Bipartite graph** with edge weights ([Yahya et al. 2012](#))
- **Weights** from lexicons  $L_P$  and  $L_T$  ([Abujabal et al. 2017](#), [Berant and Liang 2013](#))



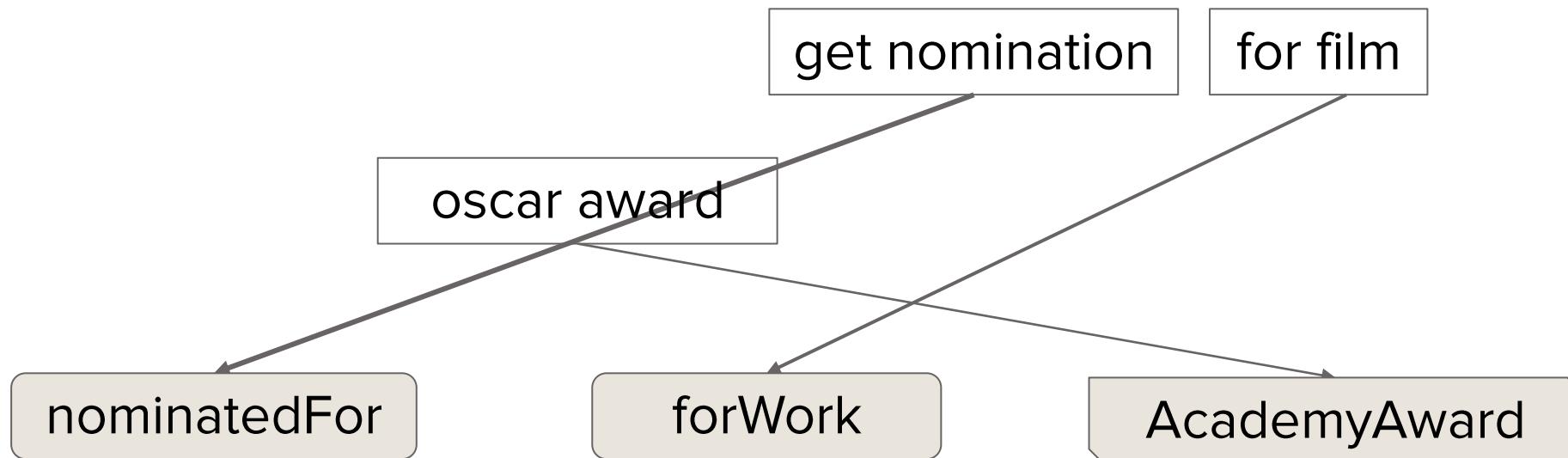
# Optimal mapping via Integer Linear Program (ILP)

- Best alignment of items with **Integer Linear Program (ILP)**
- **Constraint 1:** Each KG item obtained from at most one phrase
- **Constraint 2:** Token contributing to entity cannot contribute to any other phrase
- **Constraint 3:** One phrase can map to at most one type

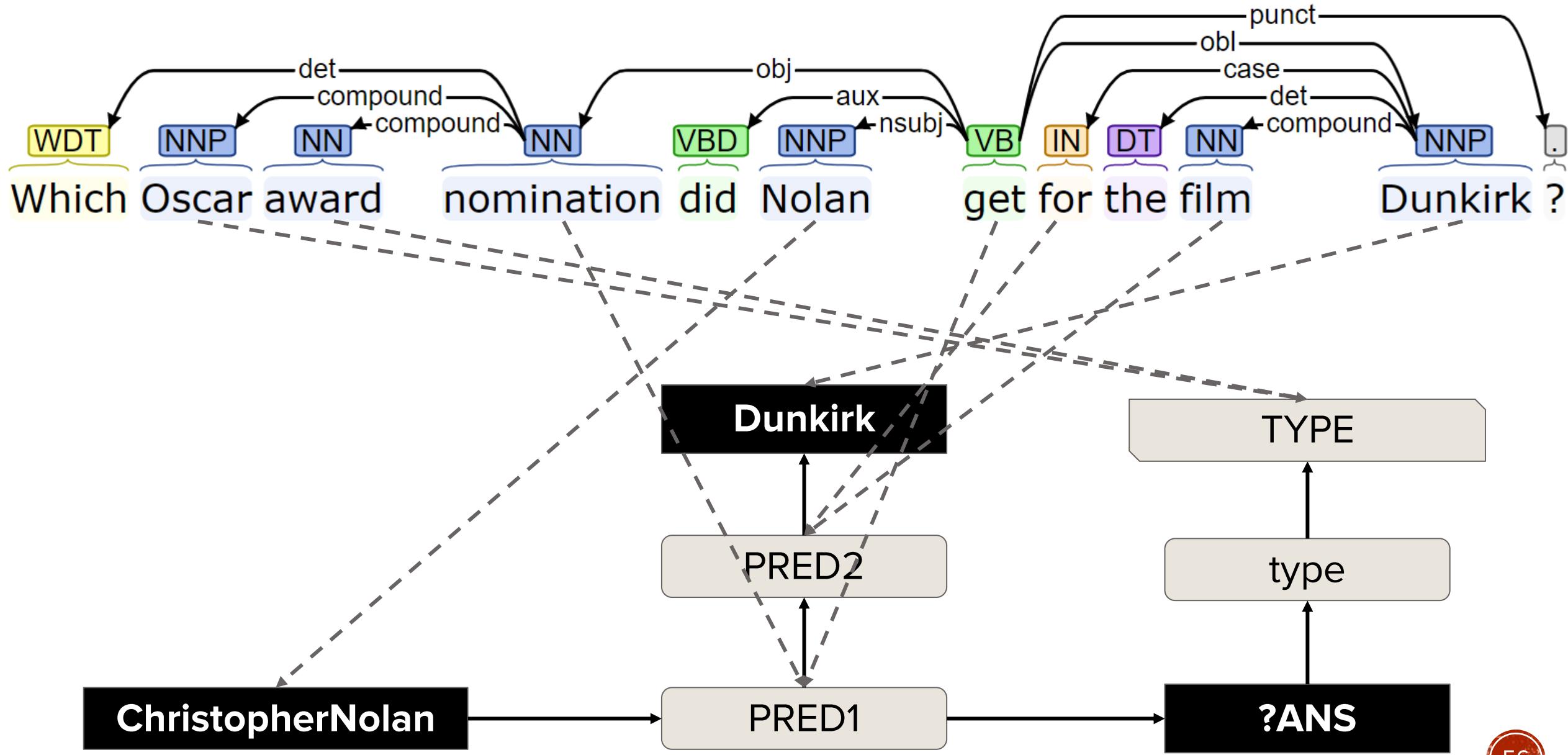


# Optimal mapping via Integer Linear Program (ILP)

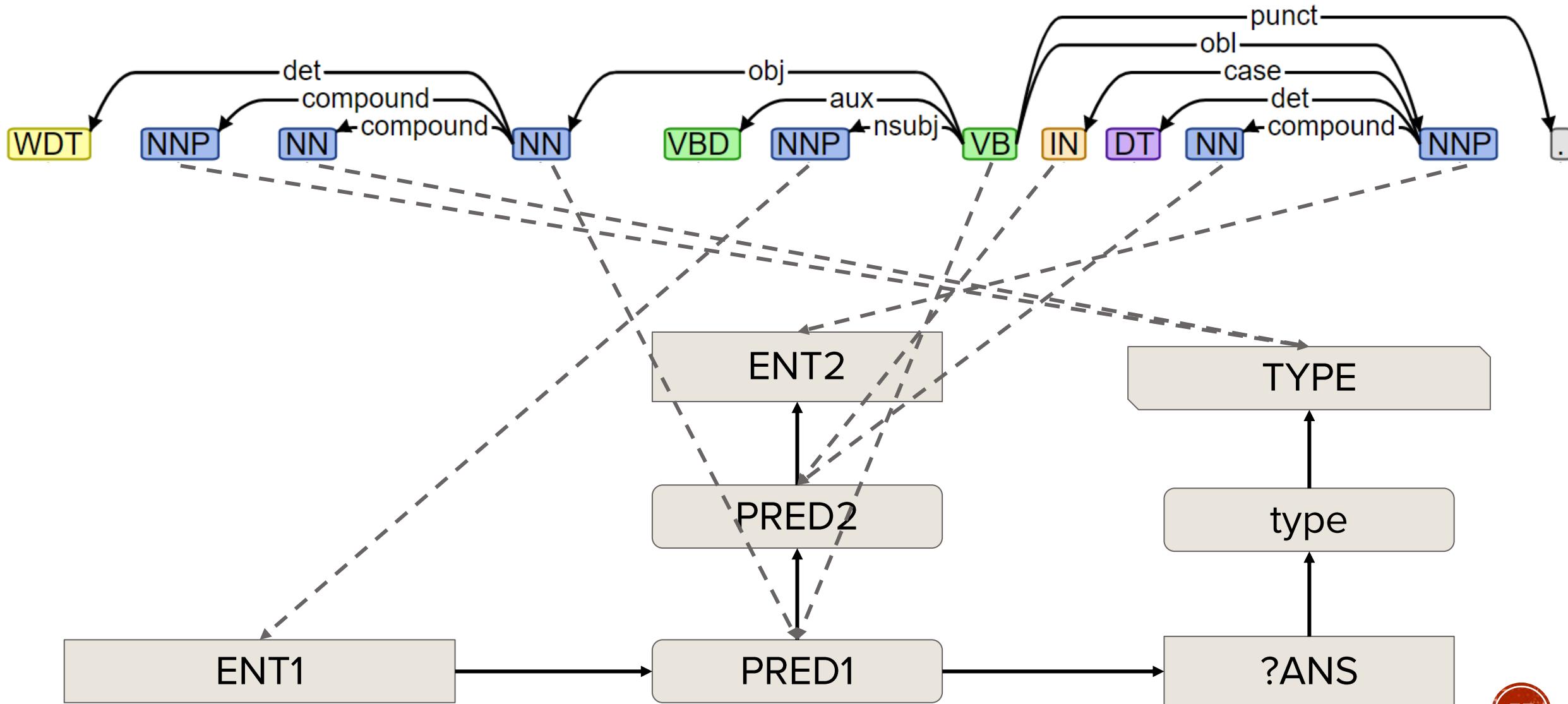
- Best alignment of items with **Integer Linear Program (ILP)**
- **Constraint 1:** Each KG item obtained from at most one phrase
- **Constraint 2:** Token contributing to entity cannot contribute to any other phrase
- **Constraint 3:** One phrase can map to at most one type



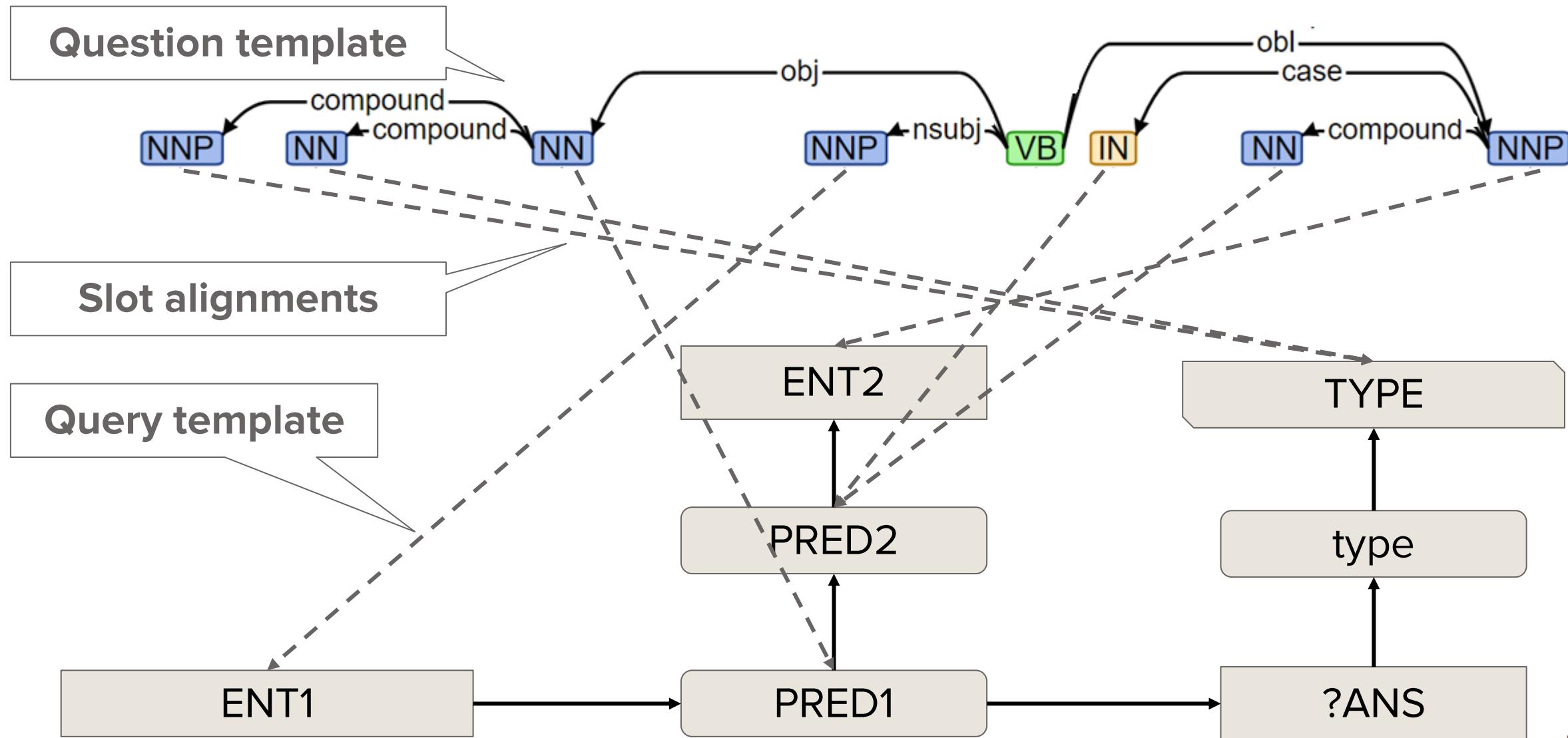
# Apply alignment to question-query



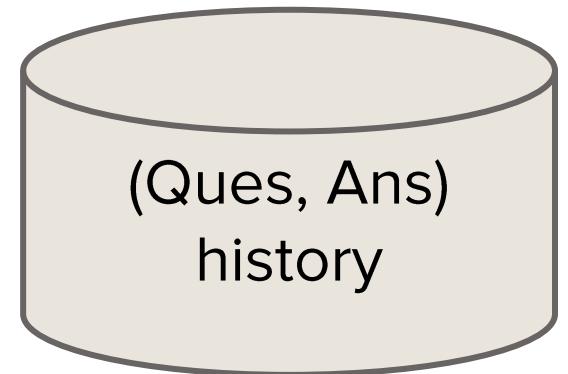
# Replace concrete items by roles



# Drop unnecessary question words



# The QUINT+NEQA framework



Abujabal et al., Automated template generation for question answering over knowledge graphs, WWW 2017.

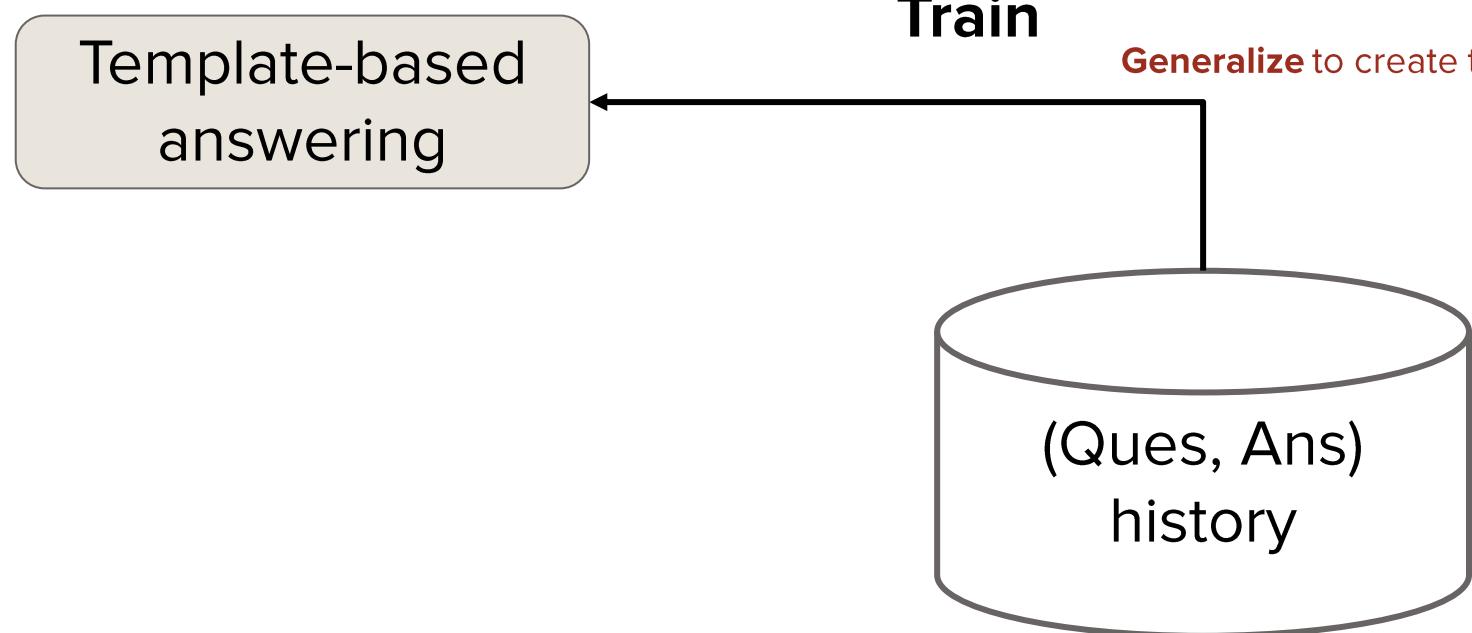
Abujabal et al., Never-ending learning for open-domain question answering over knowledge bases, WWW 2018.

# Train system using (Q, A) pairs

Distant supervision to go from (Ques, Ans) to (Ques, query)

Align ques tokens to query tokens via lexicons and ILP

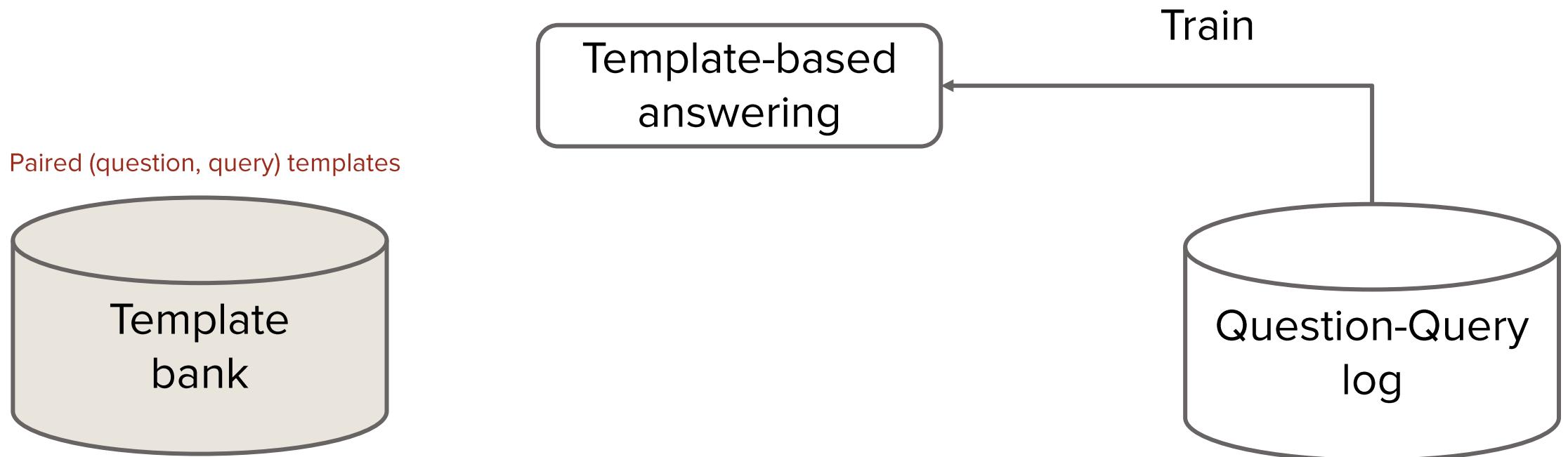
Generalize to create template



Abujabal et al., Automated template generation for question answering over knowledge graphs, WWW 2017.

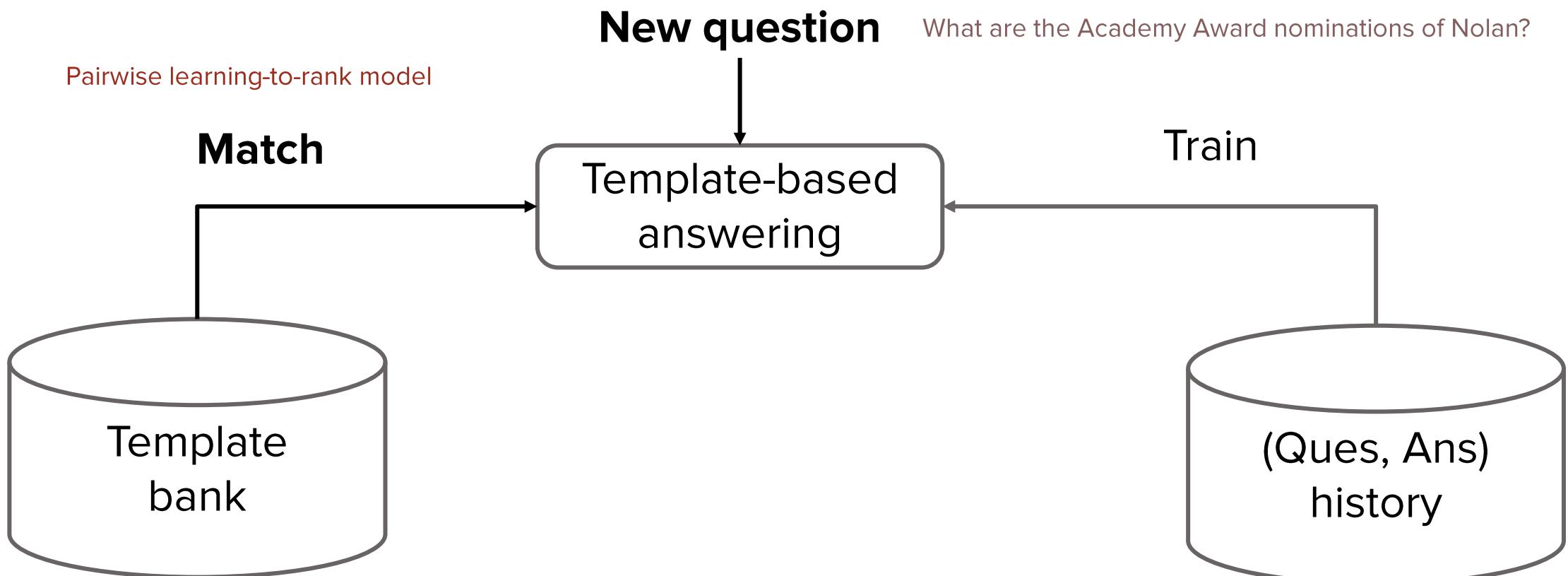
Abujabal et al., Never-ending learning for open-domain question answering over knowledge bases, WWW 2018.

# Learn a template repository



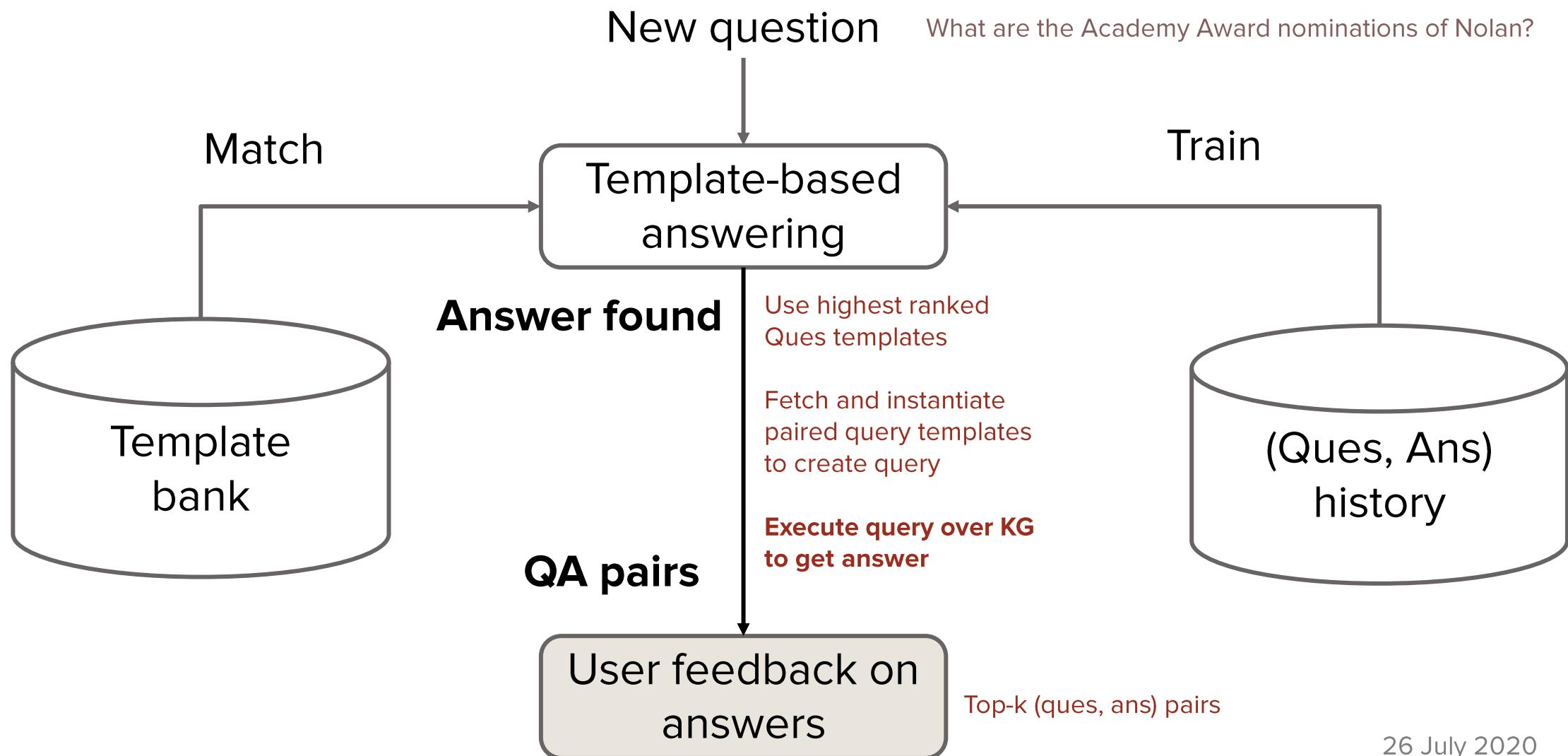
Abujabal et al., Automated template generation for question answering over knowledge graphs, WWW 2017.  
Abujabal et al., Never-ending learning for open-domain question answering over knowledge bases, WWW 2018.

# Answering with templates

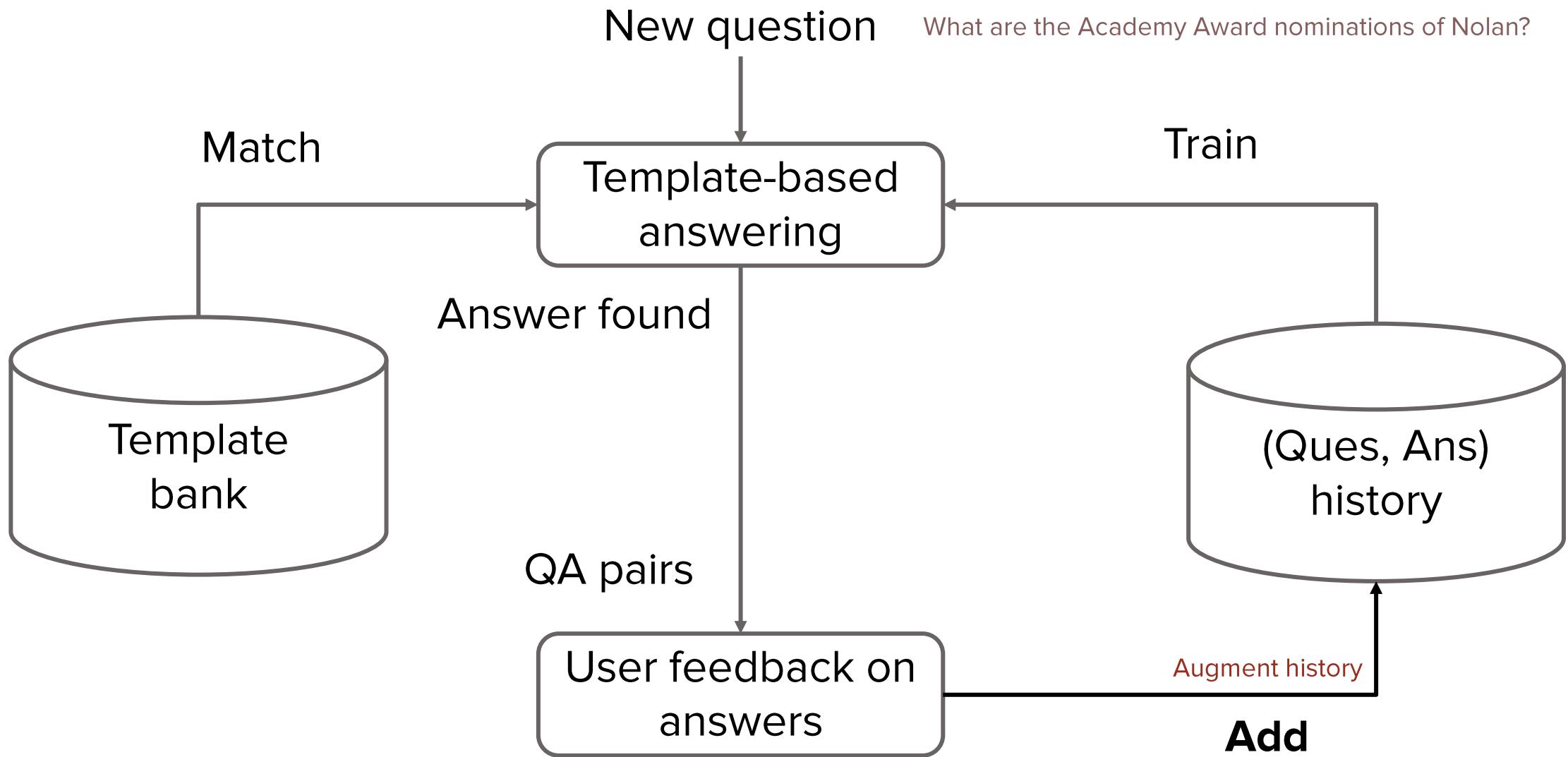


Abujabal et al., Automated template generation for question answering over knowledge graphs, WWW 2017.  
Abujabal et al., Never-ending learning for open-domain question answering over knowledge bases, WWW 2018.

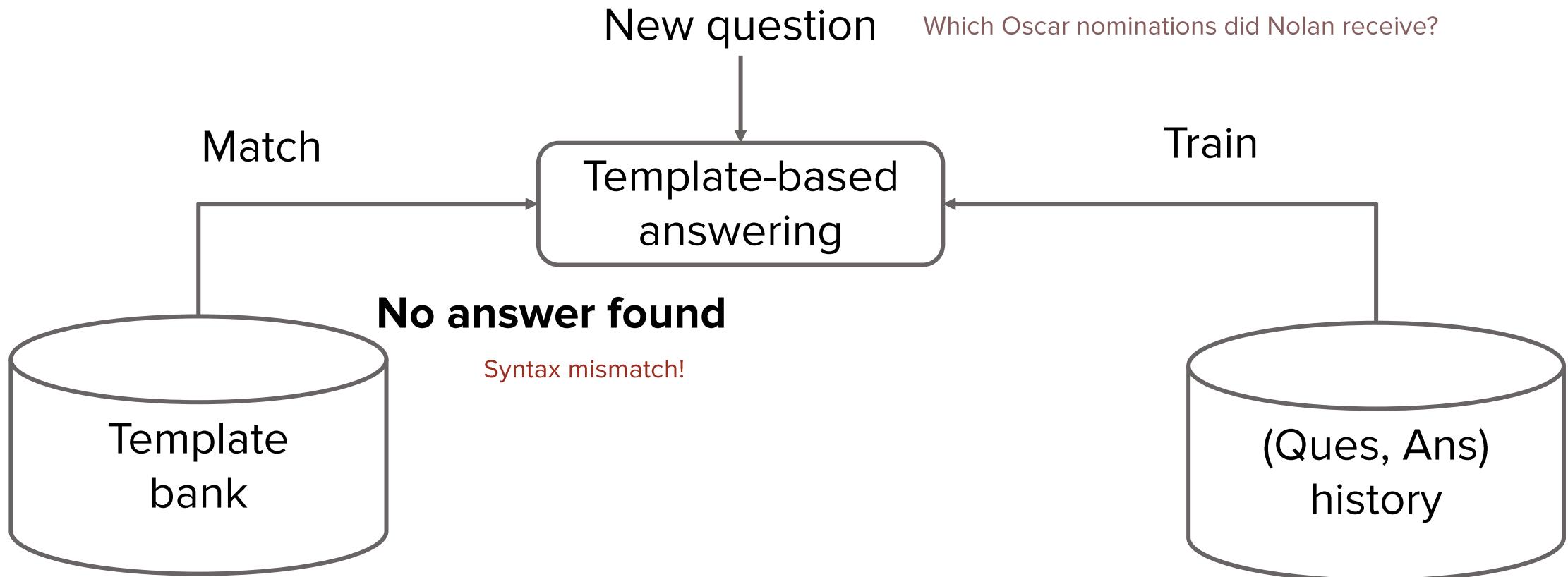
# Close the loop with user feedback



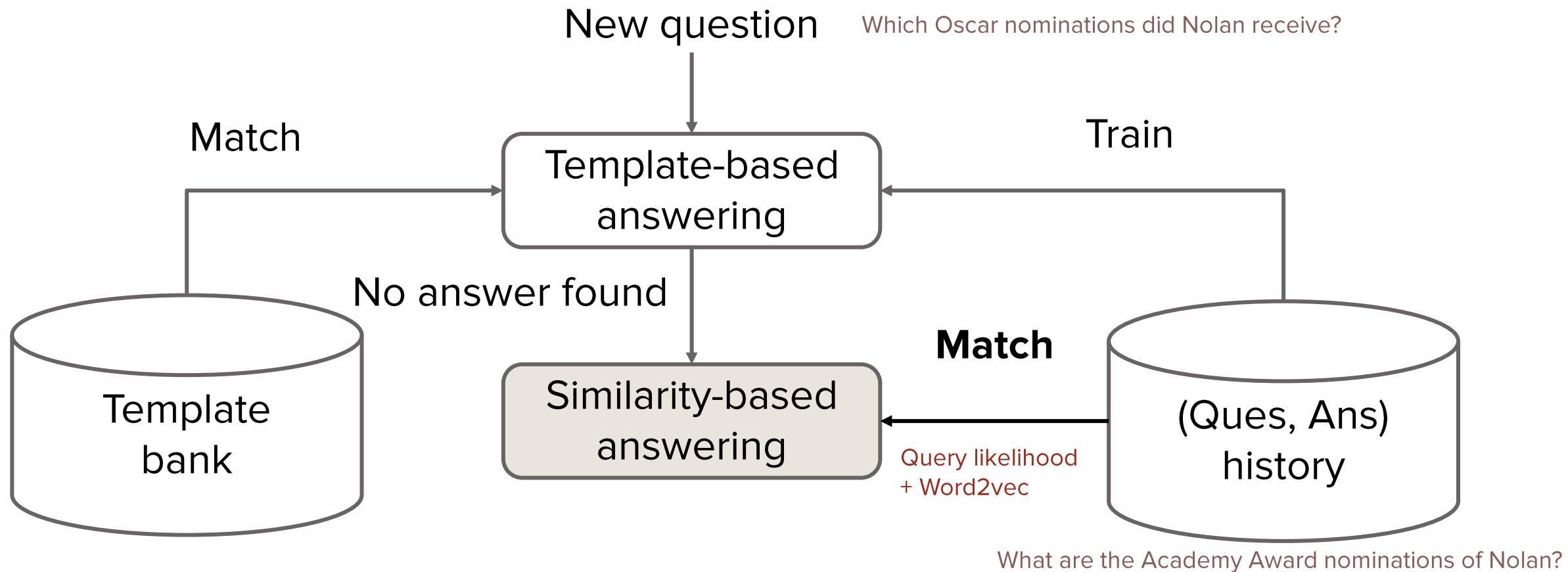
# Augment history on positive feedback



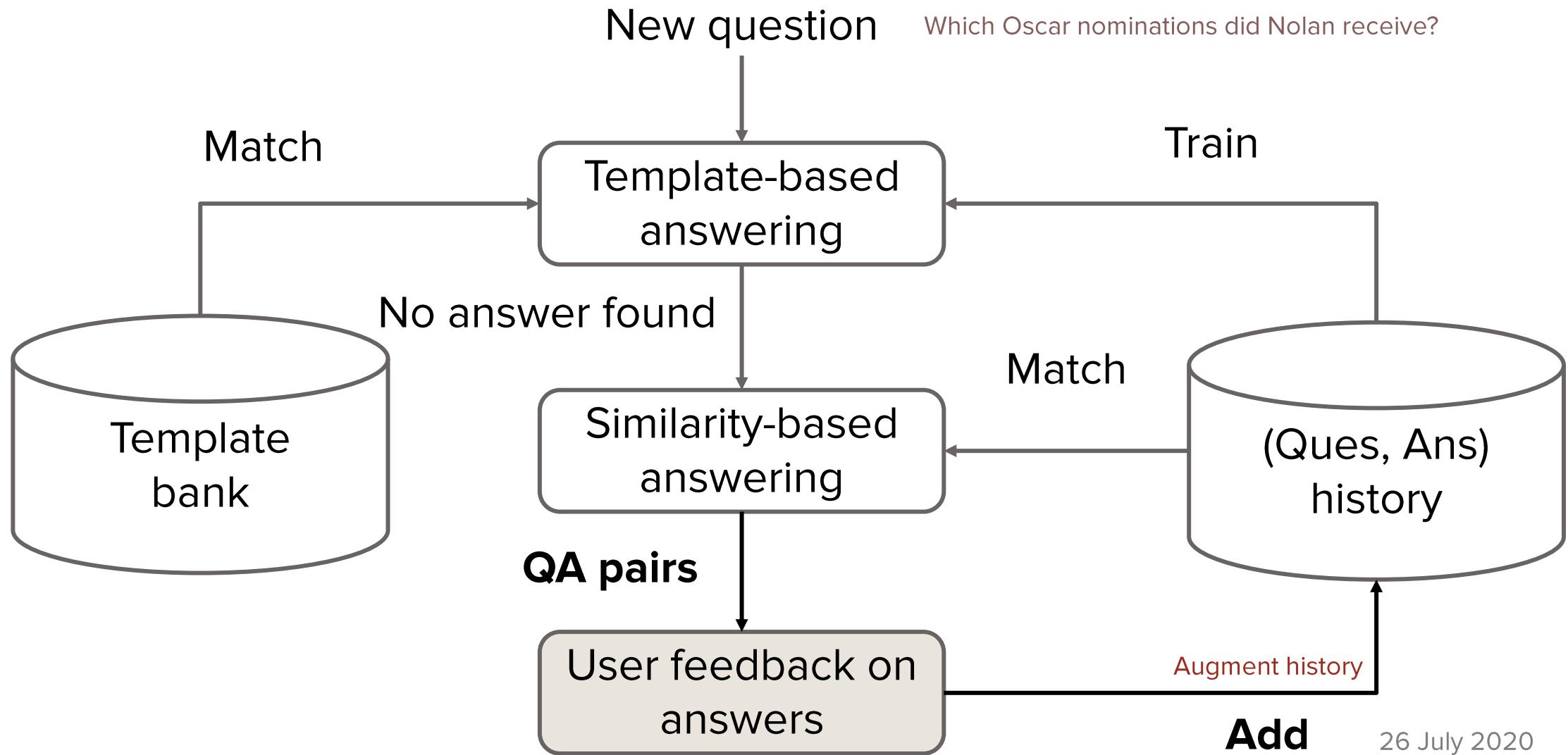
# Templates can fail



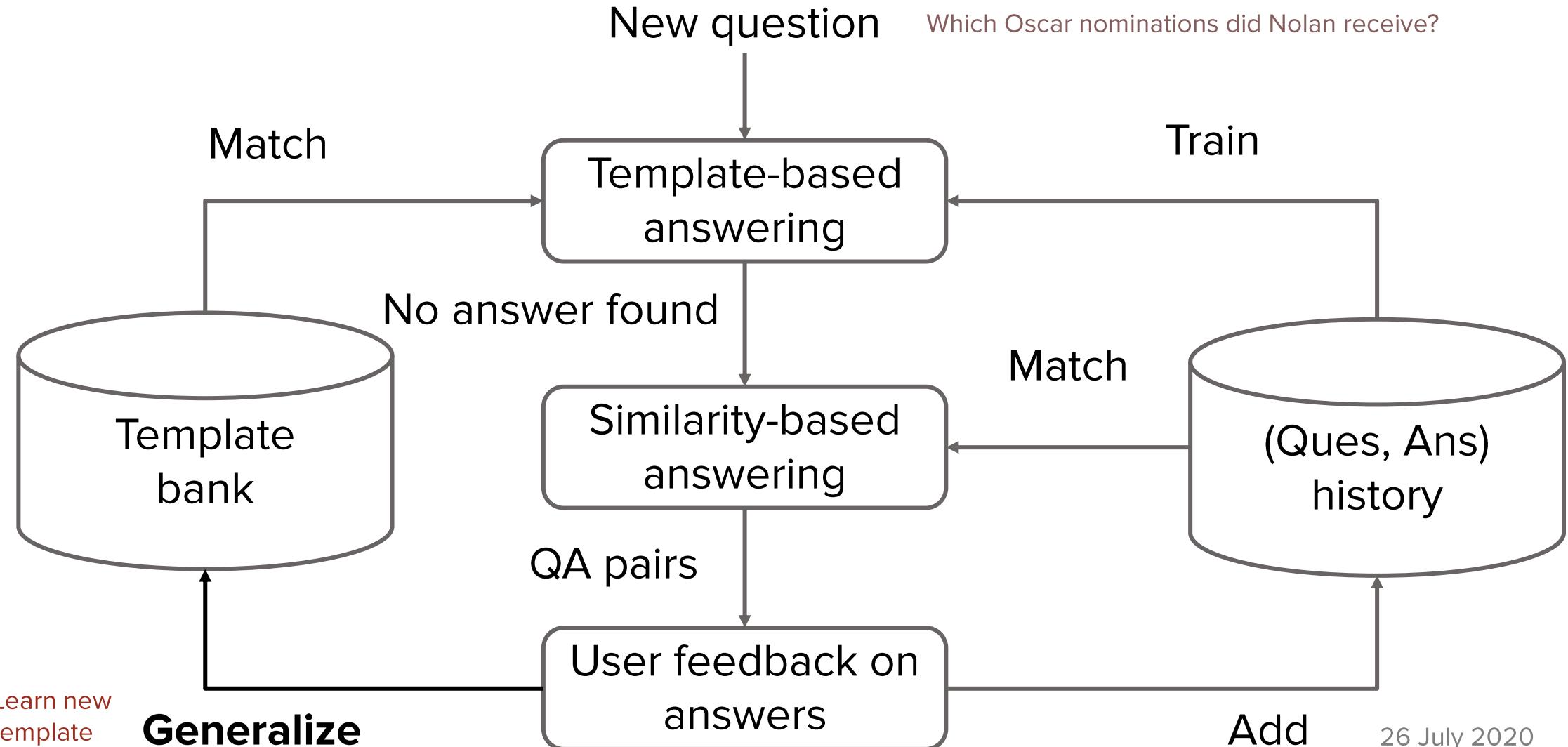
# Invoke similarity-based answering



# Augment history



# Learn new template



Learn new  
template

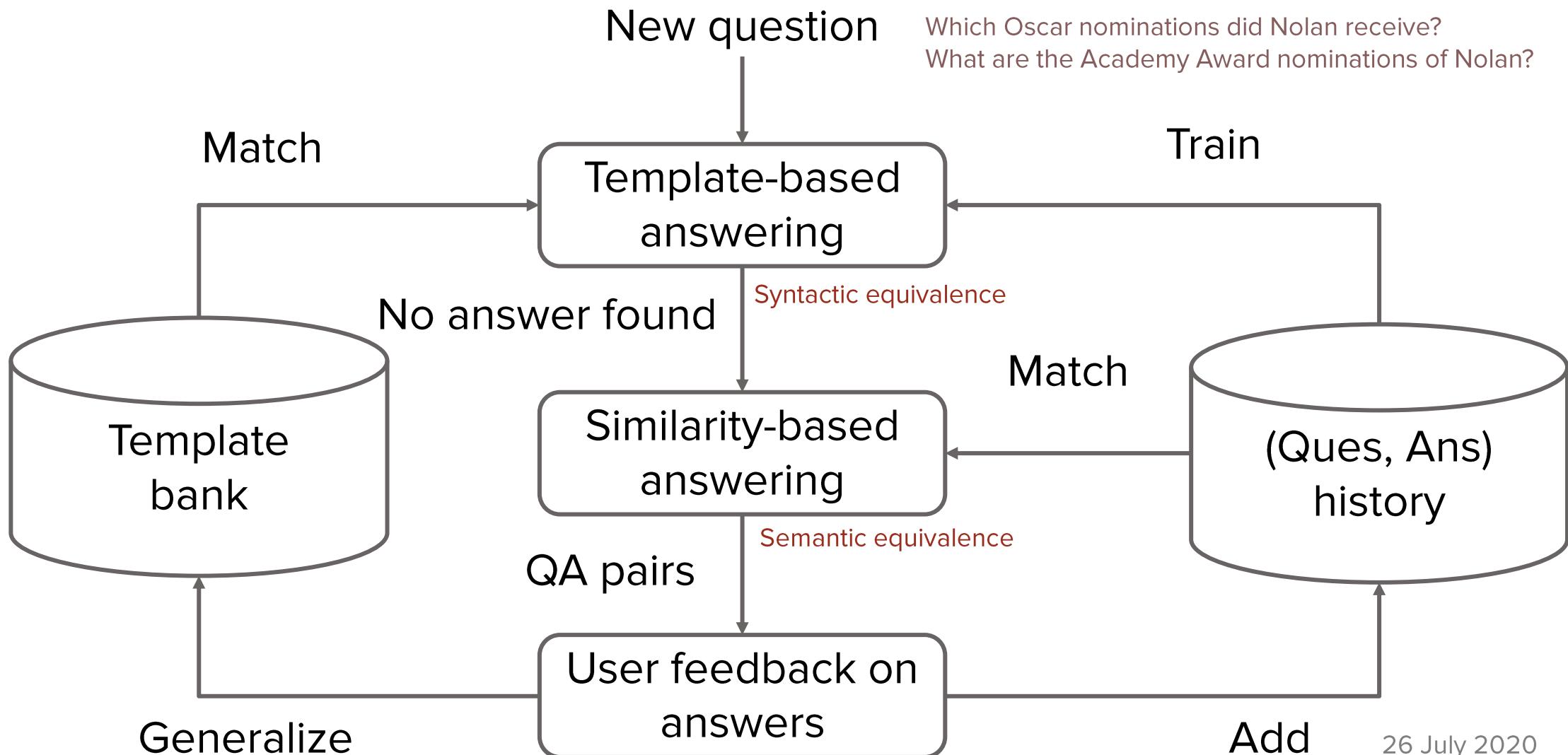
**Generalize**

Which Oscar nominations did Nolan receive?

**Add**

26 July 2020

# Never-ending learning with NEQA



# Templates: Wrap-up

- Key ideas: **Distant supervision** via shortest paths to go from (Question, Answer) to (Question, query) pair, **joint disambiguation via Integer Linear Program**
- **Template learning** also explored by [Cui et al. \(2017\)](#) and [Hu et al. \(2017\)](#)
- Works well for simple questions, but limited for **complex questions** (initial ideas in Abujabal et al. 2017, Cui et al. 2017, Hu et al. 2017)
- **Distant supervision** gets harder for complex cases
- **Similarity functions** and feedback extending scope of templates useful beyond QA?
- **Feedback in QA** subsequently investigated in QApedia ([Kratzwald and Feuerriegel 2019](#)) and IMPROVE-QA ([Zhang et al. 2019](#))

# QA with embeddings: The KEQA model

- Leverages knowledge graph embeddings (+ word embeddings)
- Uses the TransE Model (or TransE-like ...)
- From Baidu Research
- Simple questions, no qualifiers
- Seminal work on neural QA in [Bordes et al. \(2014\)](#), [Yih et al. \(2015\)](#)

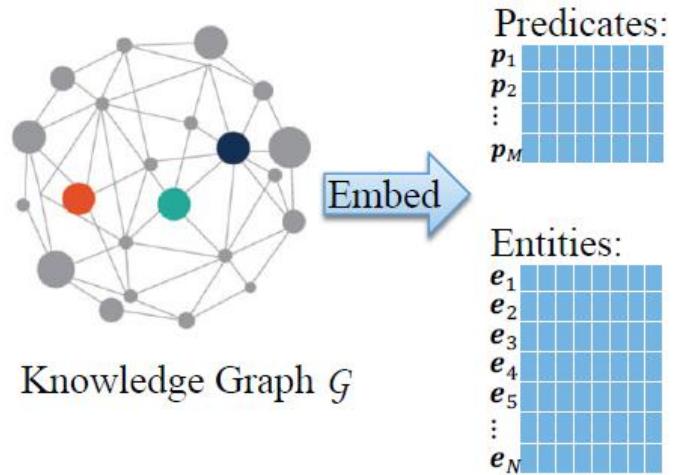
# KEQA: Outline



Knowledge Graph

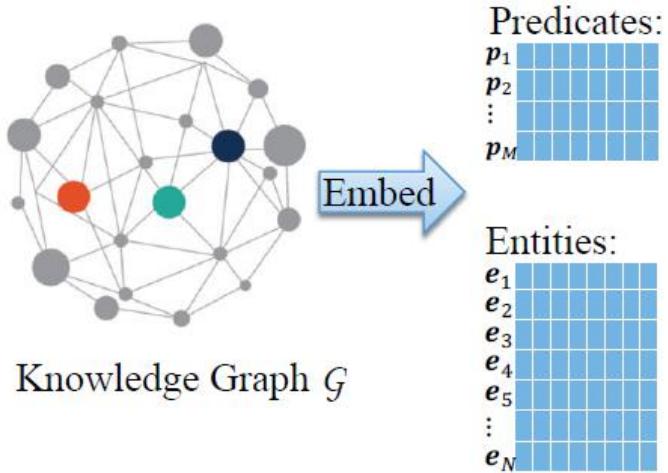
Huang et al., Knowledge graph embedding based question answering, WSDM 2019.

# KEQA: Learn KG embeddings



Huang et al., Knowledge graph embedding based question answering, WSDM 2019.

# KEQA: Using TransE



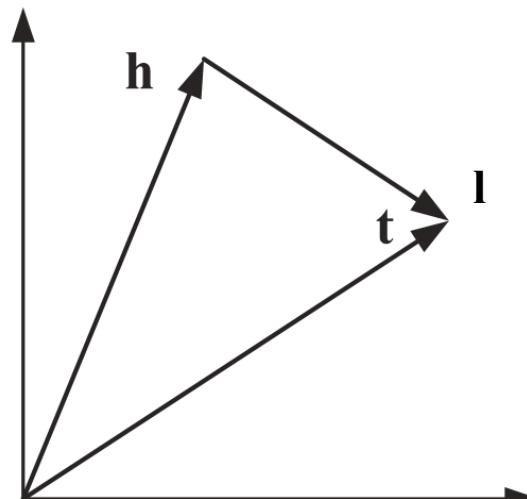
**TransE** (or TransE-like model) ([Bordes et al. 2013](#))

- Head entity, predicate, tail entity
- Loss function using correct and corrupted triples

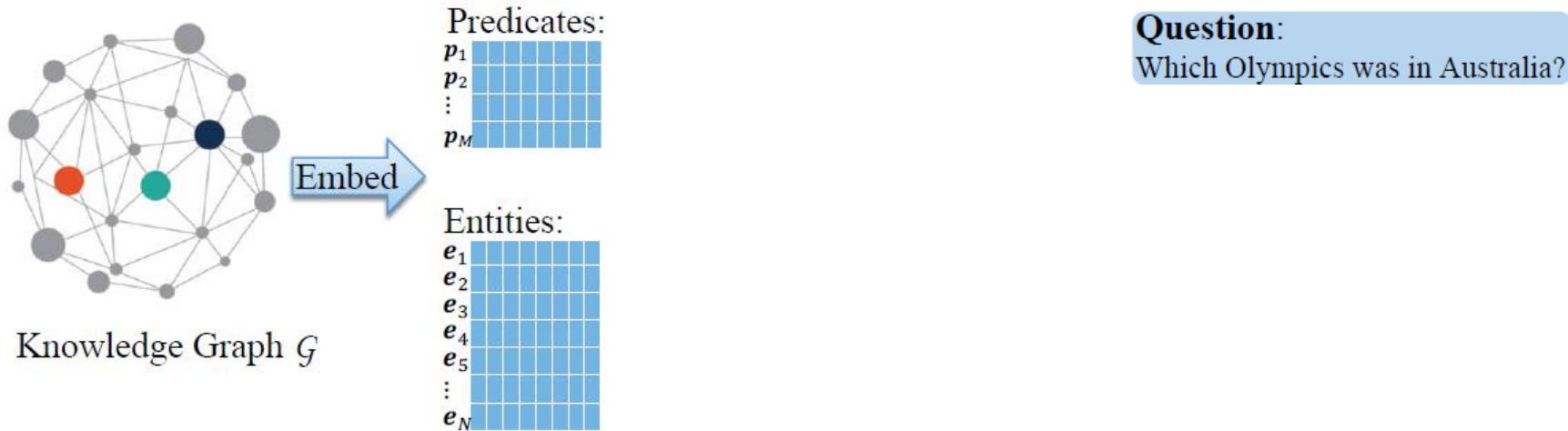
$$\mathcal{L} = \sum_{(h, \ell, t) \in S} \sum_{(h', \ell, t') \in S'_{(h, \ell, t)}} [\gamma + d(\mathbf{h} + \mathbf{\ell}, \mathbf{t}) - d(\mathbf{h}' + \mathbf{\ell}, \mathbf{t}')]_+$$

$$S'_{(h, \ell, t)} = \{(h', \ell, t) | h' \in E\} \cup \{(h, \ell, t') | t' \in E\}$$

- L2-norm of entity embeddings 1, predicates unconstrained

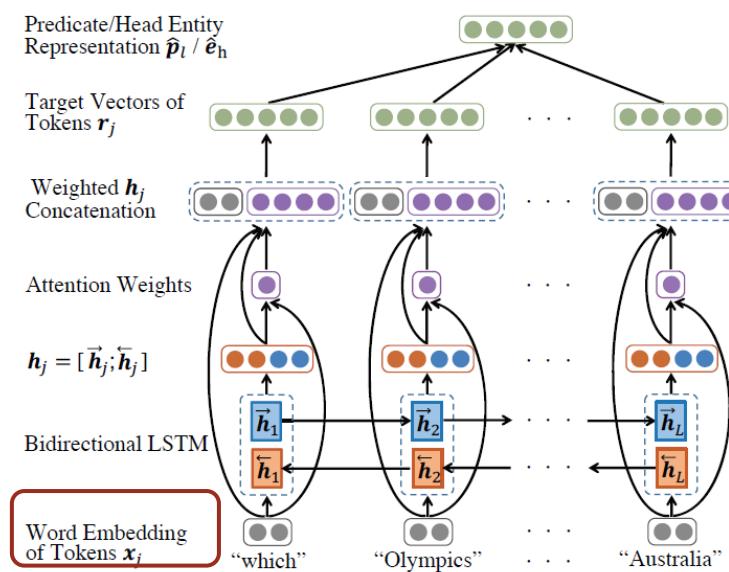
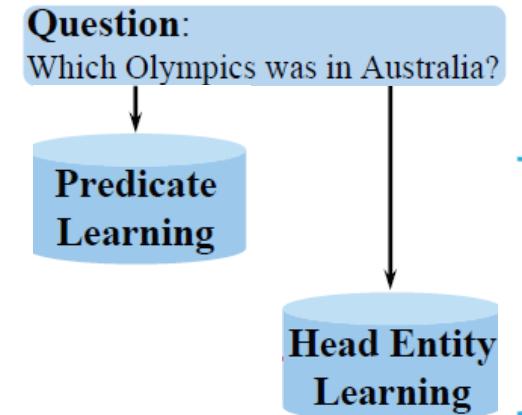
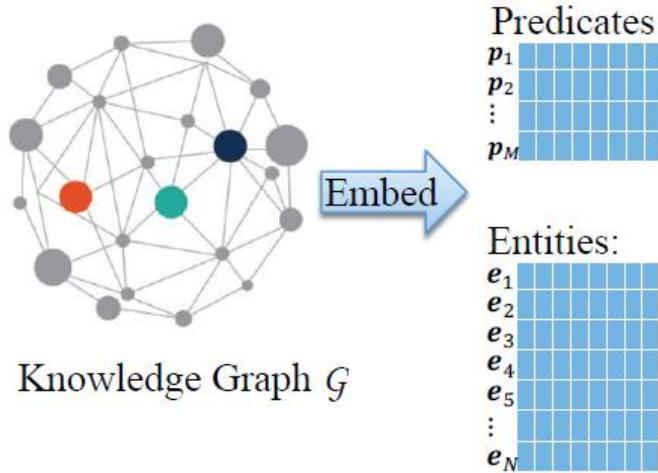


# KEQA: Input question



Huang et al., Knowledge graph embedding based question answering, WSDM 2019.

# KEQA: Learn to predict head and body

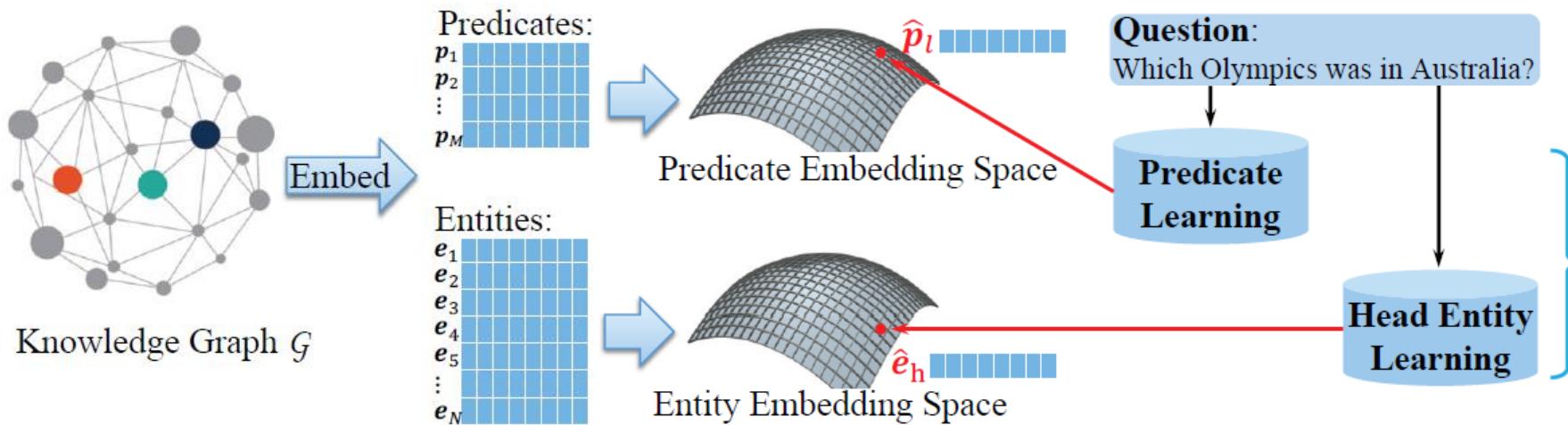


Bi-LSTM for word order

Attention for word importance

Learning representations generalization to  
unseen predicates at test time

# KEQA: Use learnt models for prediction

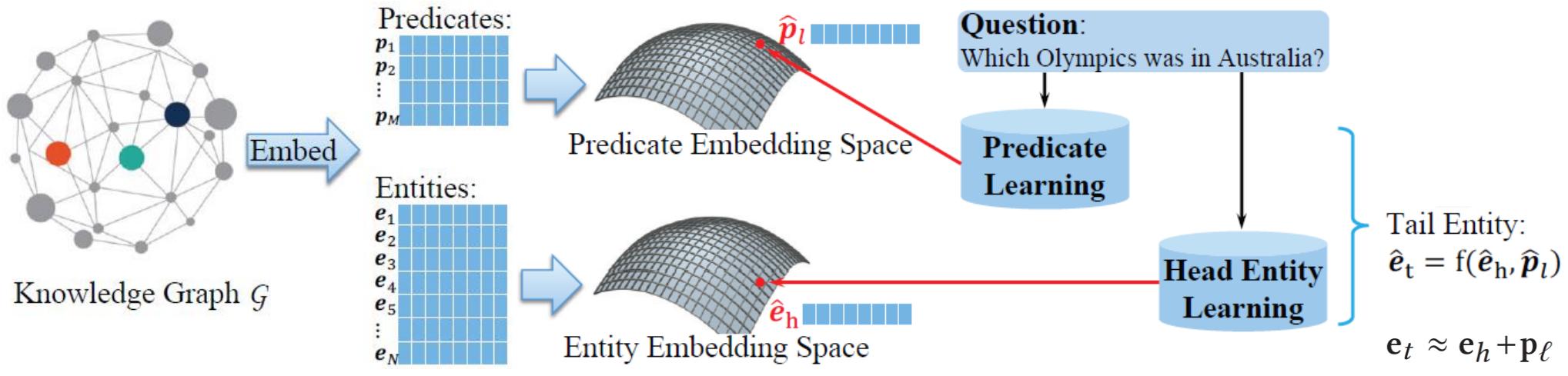


TransE (or TransE-like model) ([Bordes et al. 2013](#))

- Head entity, predicate, tail entity
- L2-norm of entity embeddings 1, predicates unconstrained

Huang et al., Knowledge graph embedding based question answering, WSDM 2019.

# KEQA: Obtain tail from head and body

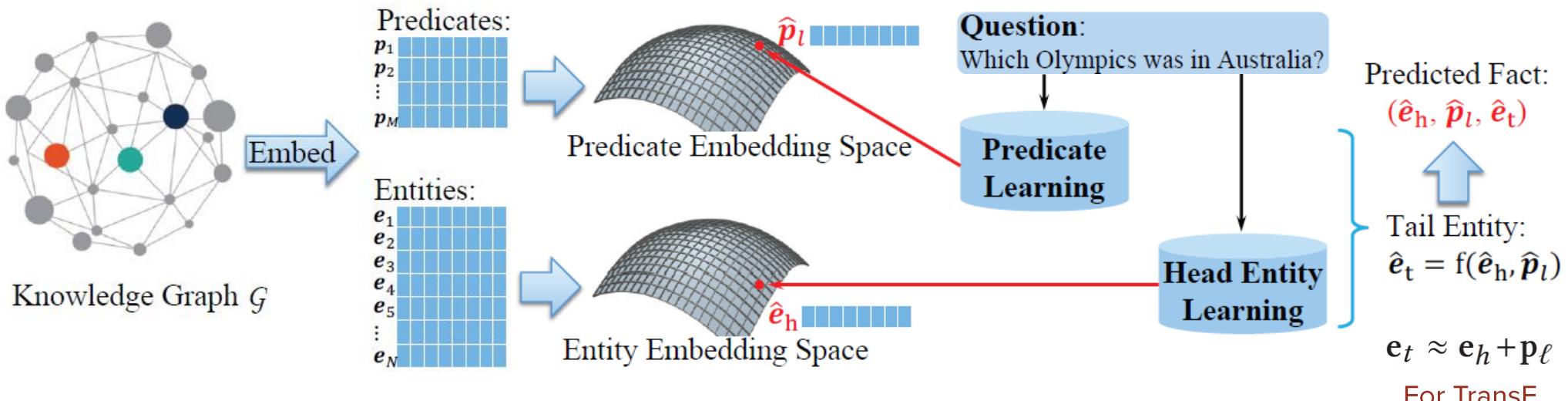


TransE (or TransE-like model) ([Bordes et al. 2013](#))

- Head entity, predicate, tail entity
- L2-norm of entity embeddings 1, predicates unconstrained

Huang et al., Knowledge graph embedding based question answering, WSDM 2019.

# KEQA: Put (head, body, tail) together

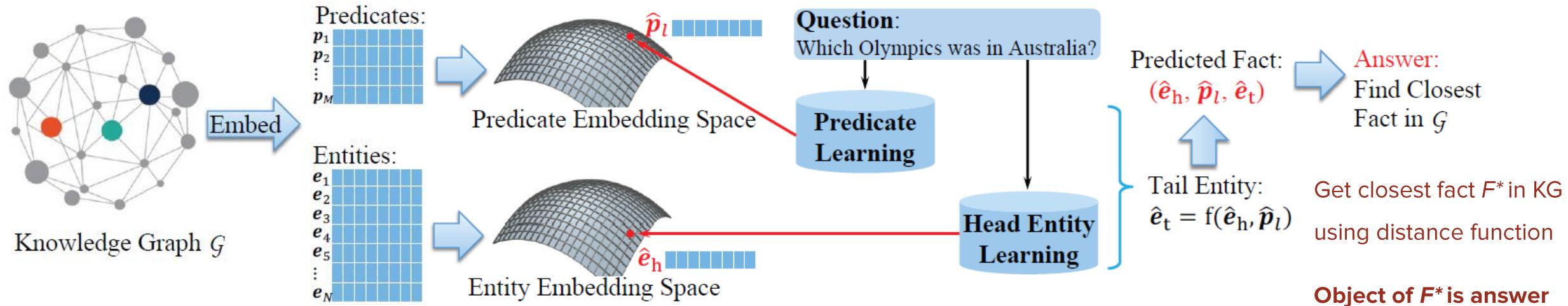


TransE (or TransE-like model) (Bordes et al. 2013)

- Head entity, predicate, tail entity
- L2-norm of entity embeddings 1, predicates unconstrained

Huang et al., Knowledge graph embedding based question answering, WSDM 2019.

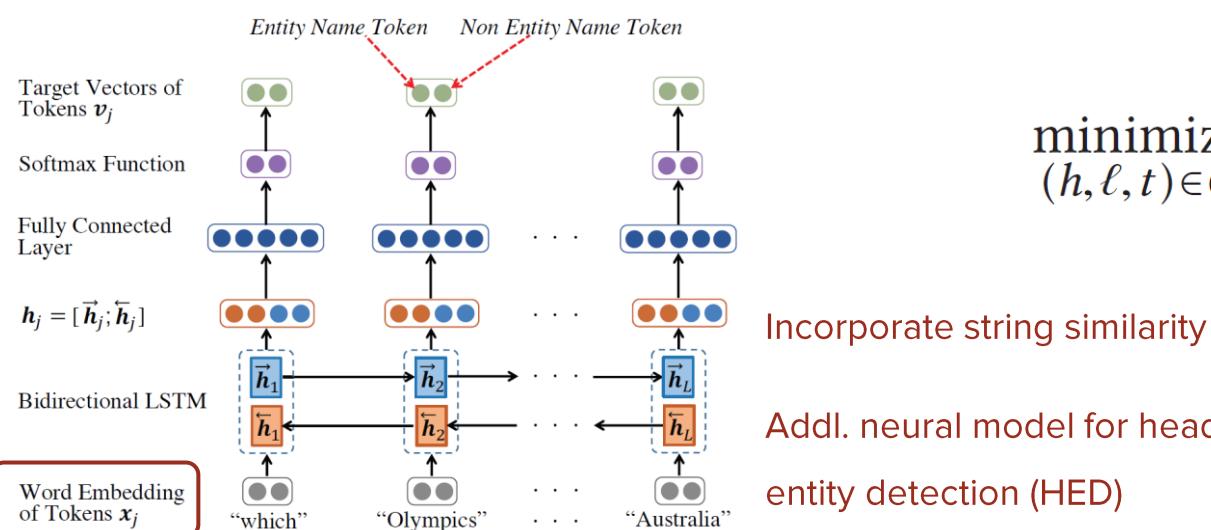
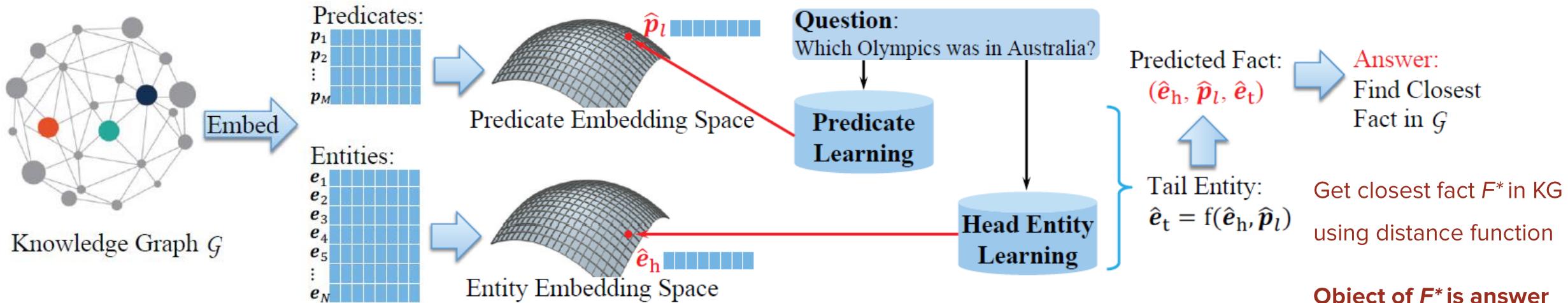
# KEQA: Search for closest fact in KG



$$\underset{(h, \ell, t) \in C}{\text{minimize}} \quad \|\mathbf{p}_\ell - \hat{\mathbf{p}}_\ell\|_2 + \beta_1 \|\mathbf{e}_h - \hat{\mathbf{e}}_h\|_2 + \beta_2 \|f(\mathbf{e}_h, \mathbf{p}_\ell) - \hat{\mathbf{e}}_t\|_2$$

Huang et al., Knowledge graph embedding based question answering, WSDM 2019.

# KEQA: Closest fact to answer



$$\begin{aligned} & \underset{(h, \ell, t) \in C}{\text{minimize}} \quad \|p_\ell - \hat{p}_\ell\|_2 + \beta_1 \|e_h - \hat{e}_h\|_2 + \beta_2 \|f(e_h, p_\ell) - \hat{e}_t\|_2 \\ & \quad - \beta_3 sim[n(h), \text{HED}_{\text{entity}}] - \beta_4 sim[n(\ell), \text{HED}_{\text{non}}], \end{aligned}$$

Incorporate string similarity

Addl. neural model for head

entity detection (HED)

# Embeddings: Wrap-up

- Graph embeddings useful for simple questions, not clear for complex cases
- Embeddings and neural methods are ubiquitous now
- Much more than using pre-trained embeddings
- Leveraging sequence models (Bi-LSTMs, transformers) with attention

break duration ?x .  
?x measuredIn minutes .

# Outline: QA over knowledge graphs

- **Background:** Setup, benchmarks, metrics
- **Simple QA:** Templates and embeddings
- **Complex QA:** Multiple entities and predicates
- **Heterogeneous sources:** Handling KGs and text
- **Conversational QA:** Implicit context in multi-turn setup
- **Take-home:** Summary and insights

How can we answer more complex questions  
with multiple entities and predicates?

# Complex questions

- Two basic types
  - Star joins
  - Who played for Barcelona and Real Madrid?
- Chain joins
  - What is the profession of Messi's father?

```
SELECT ?x WHERE  
?x playedFor Barcelona .  
?x playedFor RealMadrid .
```

**Single variable**

```
SELECT ?y WHERE  
?x fatherOf Messi .  
?x profession ?y .
```

**Two or more variables**

# Complex questions

- **Much more:** Star joins, chain joins, aggregations, comparatives, superlatives, reasoning, existential, temporal, ....
- Focus on **substructures** in questions and queries  
([Bhutani et al. 2019](#), [Ding et al. 2019](#), [Sun et al. 2020](#))
- Often rely on question **decomposition** ([Bao et al. 2016](#), [Talmor and Berant 2018](#), [Sun et al. 2020](#))
- **Joint disambiguation** of question concepts ([Yahya et al. 2012](#), [Lu et al. 2019](#))

Which female **actor** played in Casablanca and is married to a writer who was born in Rome?

Where is the **founder of Tesla** born?

Who was the **second wife** of Tom Cruise?

Which Portuguese speaking countries import fish from Brazil?

Who wrote **more books**: Enid Blyton or Agatha Christie?

Which is the **third highest** mountain in Asia?

How many **movies have the same director** as The Shawshank Redemption?

**How many movies** were directed by the graduate of Burbank High School?

**Did any cosmonauts** die in the same place they were born in?

# Complex questions

- Early efforts in [Yahya et al. \(2012\)](#)
- Further explorations in [Bao et al. \(2016\)](#),  
[Abujabal et al. \(2017\)](#) and [Cui et al. \(2017\)](#)
- Dedicated methods for complex questions in  
[Vakulenko et al. \(2019\)](#), [Bhutani et al. \(2019\)](#),  
[Ding et al. \(2019\)](#), [Hu et al. \(2018\)](#), [Luo et al. \(2018\)](#), ...

Which female **actor** played in Casablanca and is married to a writer who was born in Rome?

Where is the **founder of Tesla** born?

Who was the **second wife** of Tom Cruise?

Which Portuguese speaking countries import fish from Brazil?

Who wrote **more books**: Enid Blyton or Agatha Christie?

Which is the **third highest** mountain in Asia?

How many **movies have the same director** as The Shawshank Redemption?

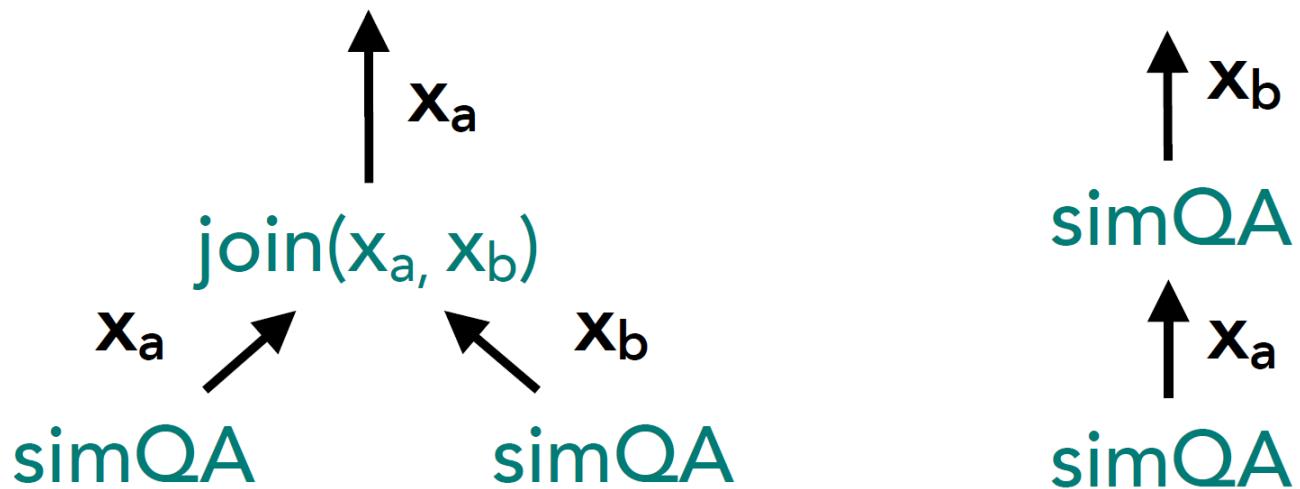
**How many movies** were directed by the graduate of Burbank High School?

**Did any cosmonauts** die in the same place they were born in?

# Complex QA: Structured query generation

- The **TextRay** system (Bhutani et al. CIKM 2019)
- Learning complex query patterns difficult for **data sparsity**
- **Decompose-execute-join** approach to complex questions
- Constructs complex query patterns using **simple queries**
- **Semantic matching** model learns simple queries  
using **distant supervision** from QA pairs

# TextRay: Computation plan



Predict computation plan upfront with supervised method

Or

with linguistic cues

## Single variable

### Star join

```
SELECT ?x WHERE  
?x playedFor Barcelona .  
?x playedFor RealMadrid .
```

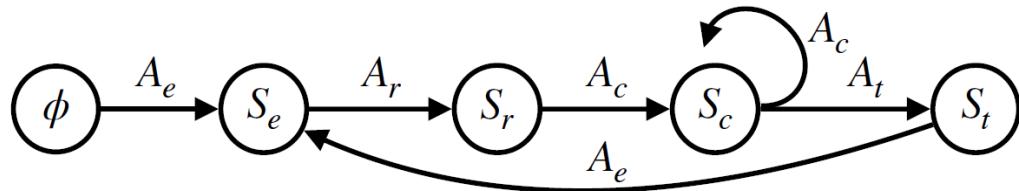
## Two or more variables

### Chain join

```
SELECT ?y WHERE  
?x fatherOf Messi .  
?x profession ?y .
```

# TextRay: Walkthrough

Which Portuguese speaking countries import fish from Brazil?



Staged query graph generation (Yih et al. ACL 2015)

$s_1$   $s_2$    
Brazil Portuguese

$s_3$    
Brazilian Portuguese

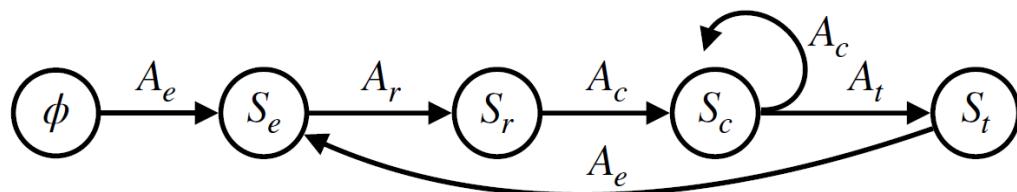
a) Identify seed

Top-k entities

Bhutani et al., Learning to Answer Complex Questions over Knowledge Bases with Query Composition, CIKM 2019.

# TextRay: Partial query graph

Which Portuguese speaking countries import fish from Brazil?



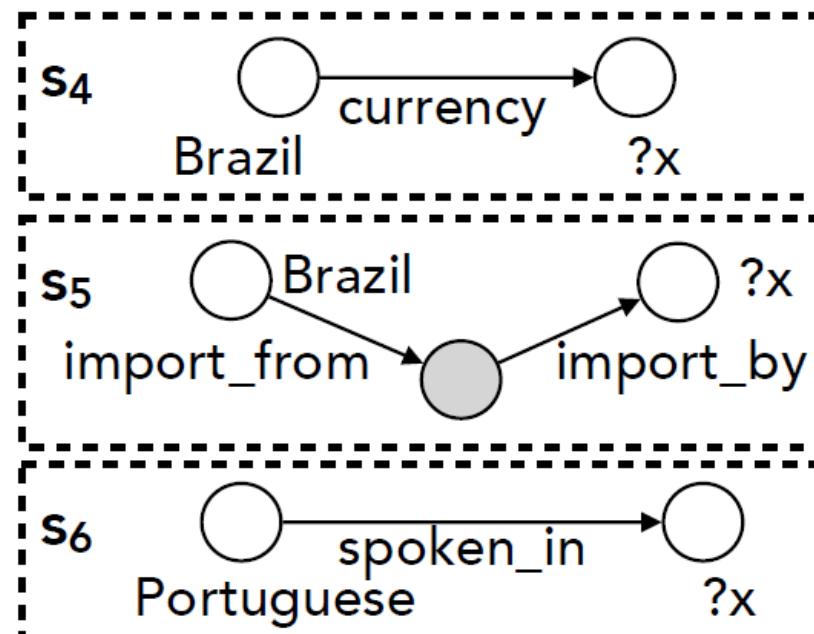
Staged query graph generation (Yih et al. ACL 2015)

$s_1$  Brazil  
 $s_2$  Portuguese

$s_3$  Brazilian Portuguese

a) Identify seed

Top-k entities

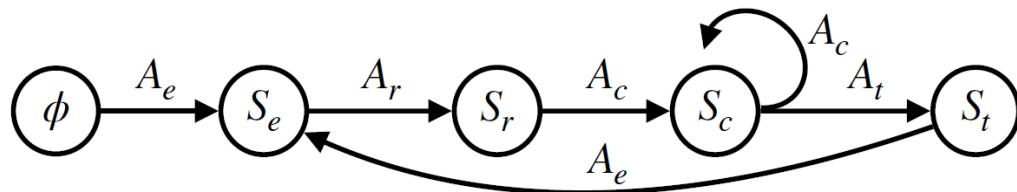


b) Identify main relation path

Bhutani et al., Learning to Answer Complex Questions over Knowledge Bases with Query Composition, CIKM 2019.

# TextRay: Partial query graph

Which Portuguese speaking countries import fish from Brazil?

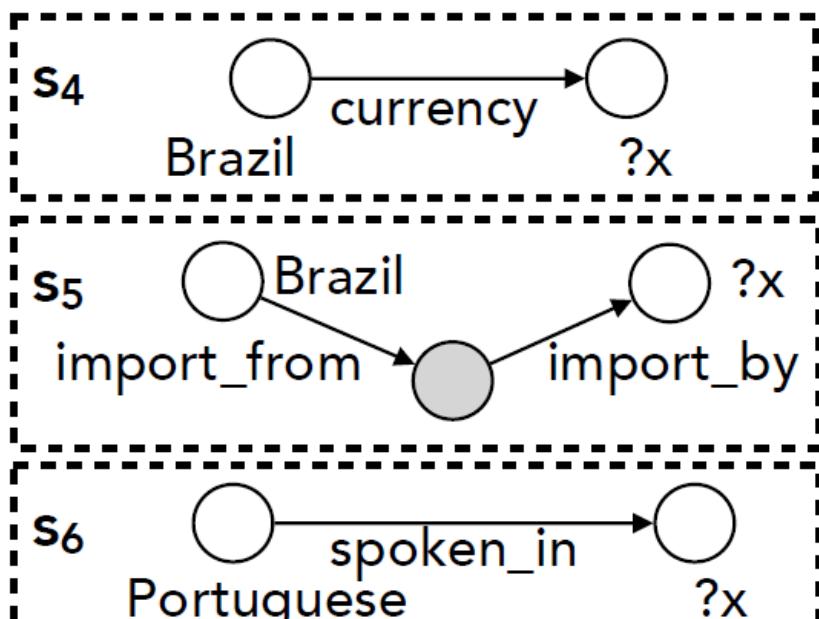


Staged query graph generation (Yih et al. ACL 2015)

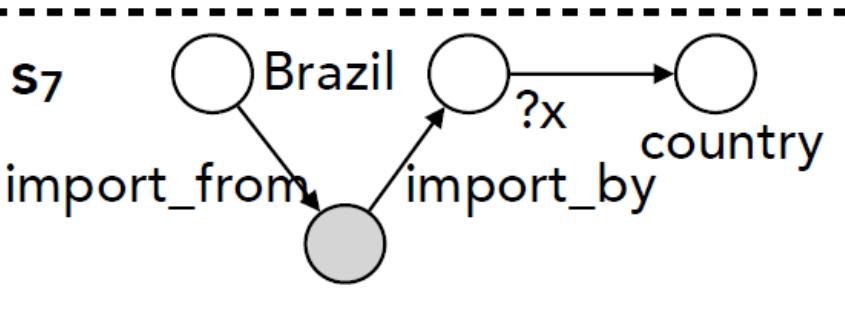
$s_1$  Brazil     $s_2$  Portuguese  
 $s_3$  Brazilian Portuguese

a) Identify seed

Top-k entities



b) Identify main relation path



c) Identify constraints

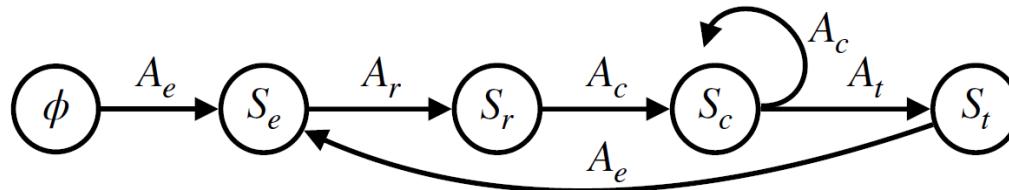
Constraints: Qualifiers, dates, entities

Consult computation plan: Grow parallel branch of partial query

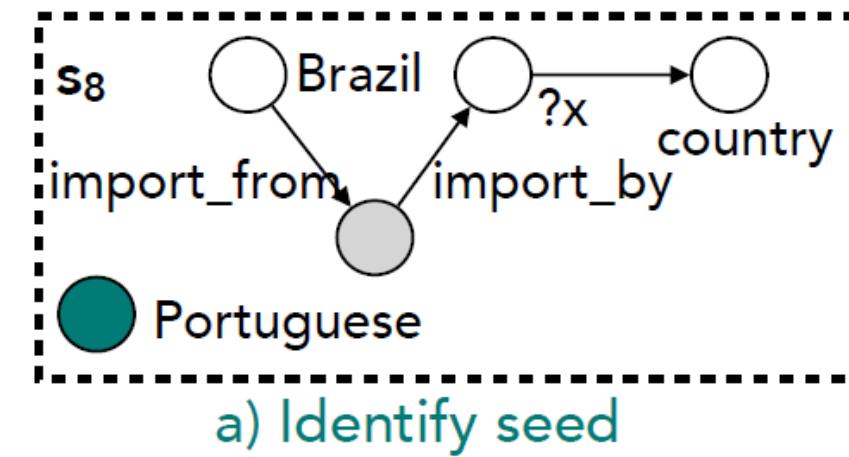
Bhutani et al., Learning to Answer Complex Questions over Knowledge Bases with Query Composition, CIKM 2019.

# TextRay: Partial query graph

Which Portuguese speaking countries import fish from Brazil?



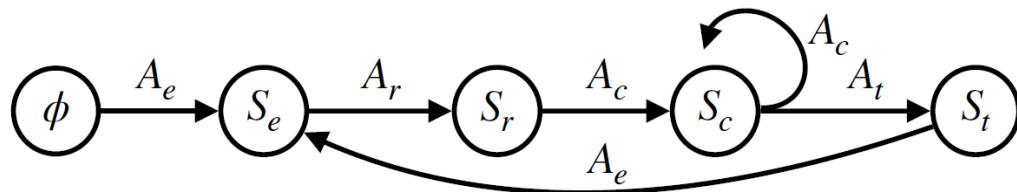
Staged query graph generation (Yih et al. ACL 2015)



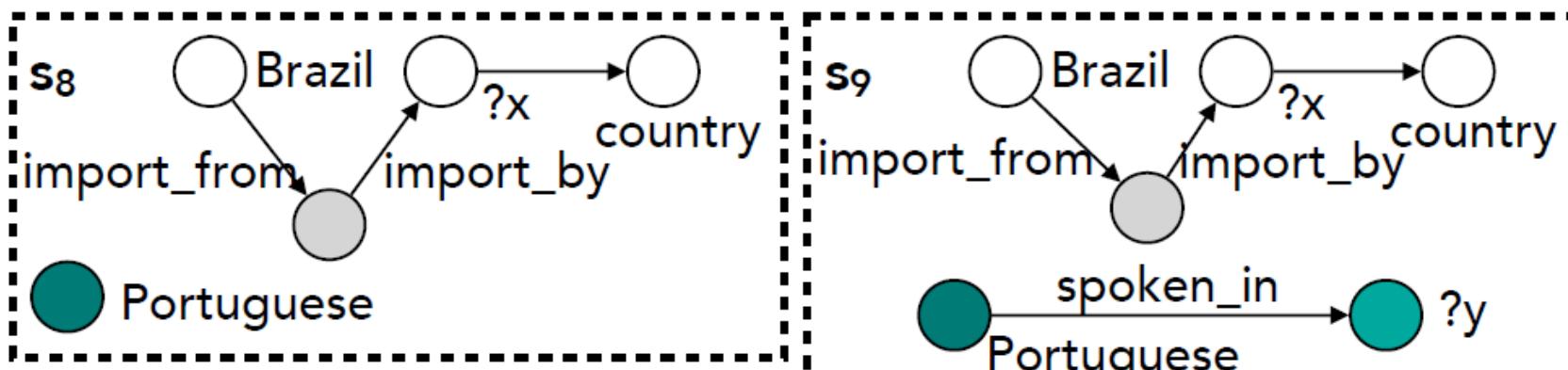
Bhutani et al., Learning to Answer Complex Questions over Knowledge Bases with Query Composition, CIKM 2019.

# TextRay: Partial query graph

Which Portuguese speaking countries import fish from Brazil?



Staged query graph generation (Yih et al. ACL 2015)



a) Identify seed

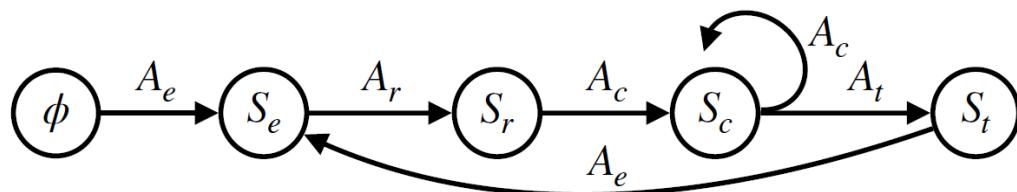
b) Identify main relation path

c) Identify constraints

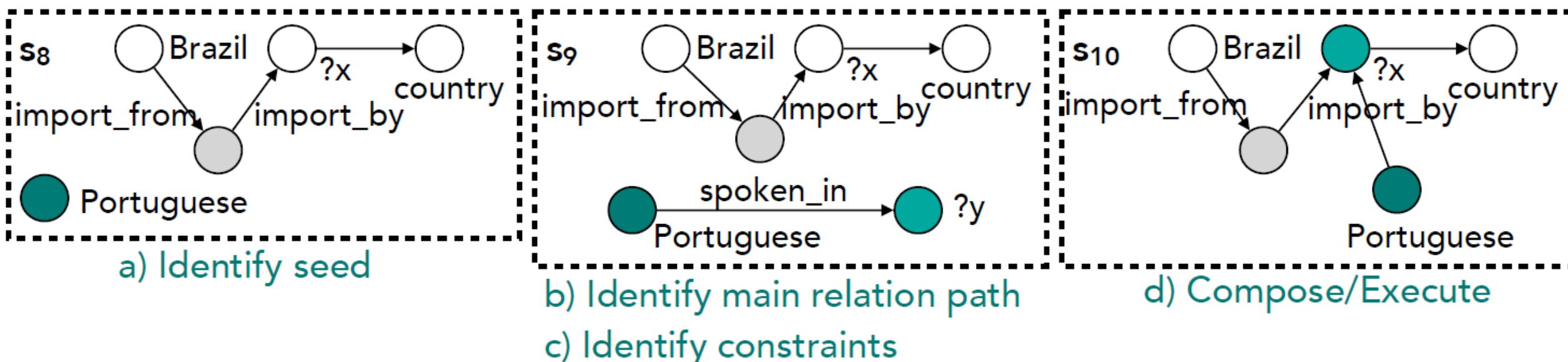
Bhutani et al., Learning to Answer Complex Questions over Knowledge Bases with Query Composition, CIKM 2019.

# TextRay: Partial query graph

Which Portuguese speaking countries import fish from Brazil?



Staged query graph generation (Yih et al. ACL 2015)



Beam search to maintain top-k best derivations + Semantic similarity learned via LSTMs with attention

Bhutani et al., Learning to Answer Complex Questions over Knowledge Bases with Query Composition, CIKM 2019.

# Complex QA: Computing compact subgraphs

- The **QUEST** system ([Lu et al. 2019](#))
- Works over open vocabulary quasi KGs ([Bhutani et al. 2019b](#), [Yin et al. 2015](#), [Fader et al. 2013](#), [2014](#))
- Augment quasi KGs with alignments and types
- Spot question cornerstones
- **Unsupervised compact subgraph computation:** Compute Group Steiner Tree (GST) with cornerstones as terminals for joint disambiguation of question concepts

# Creating a quasi KG

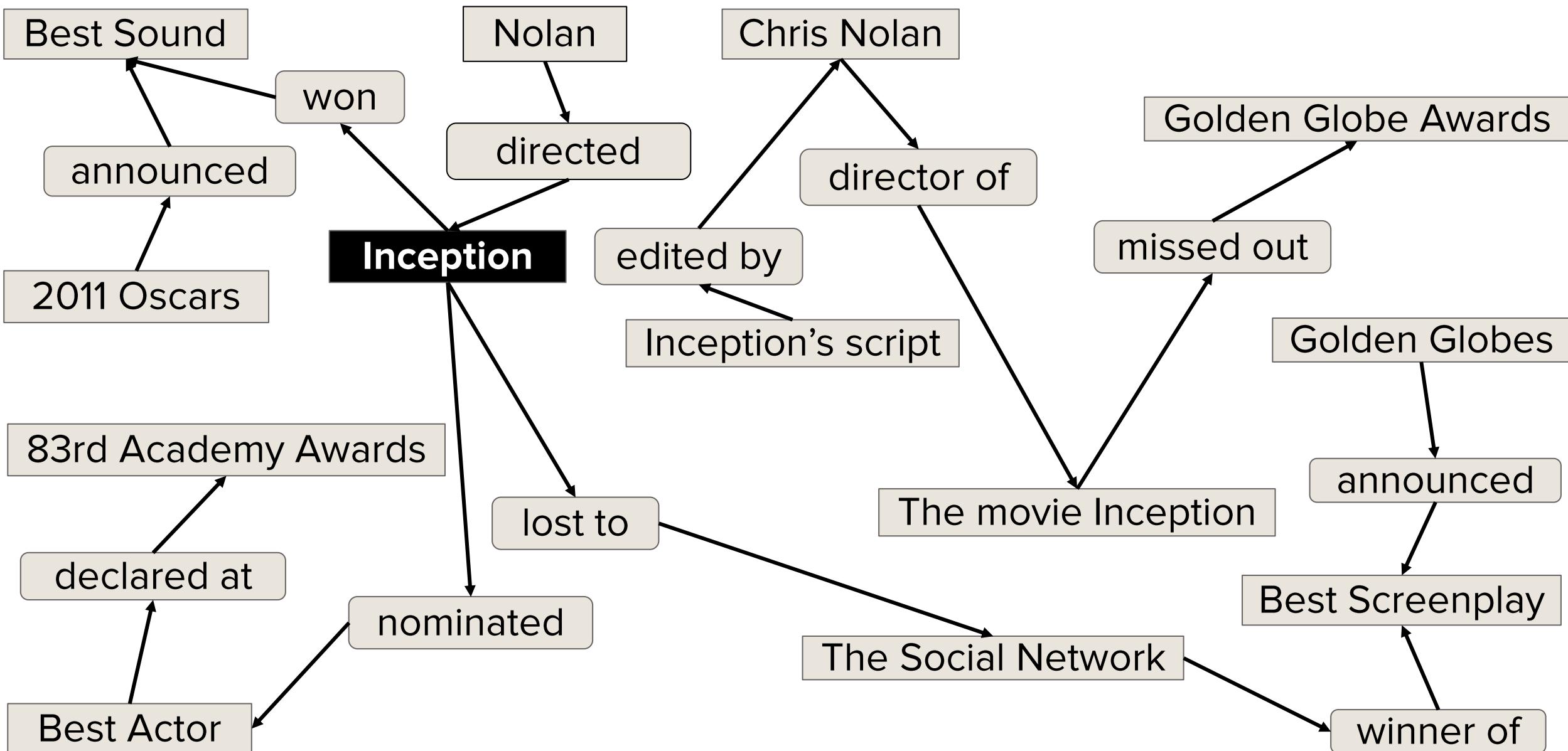
<Nolan, directed, Inception>  
<Inception, won, Best Sound>  
<2011 Oscars, announced, Best Sound>  
<Inception, nominated, Best Actor>  
<The movie Inception, missed out, Golden Globe Awards>  
<Chris Nolan, director of, The movie Inception>  
<Inception's script, edited by, Chris Nolan>  
<Inception, lost to, The Social Network>  
<Best Actor, declared at, 83rd Academy Awards>  
<The Social Network, winner of, Best Screenplay>  
<Golden Globes, announced, Best Screenplay>

Compile an open-vocabulary triple store

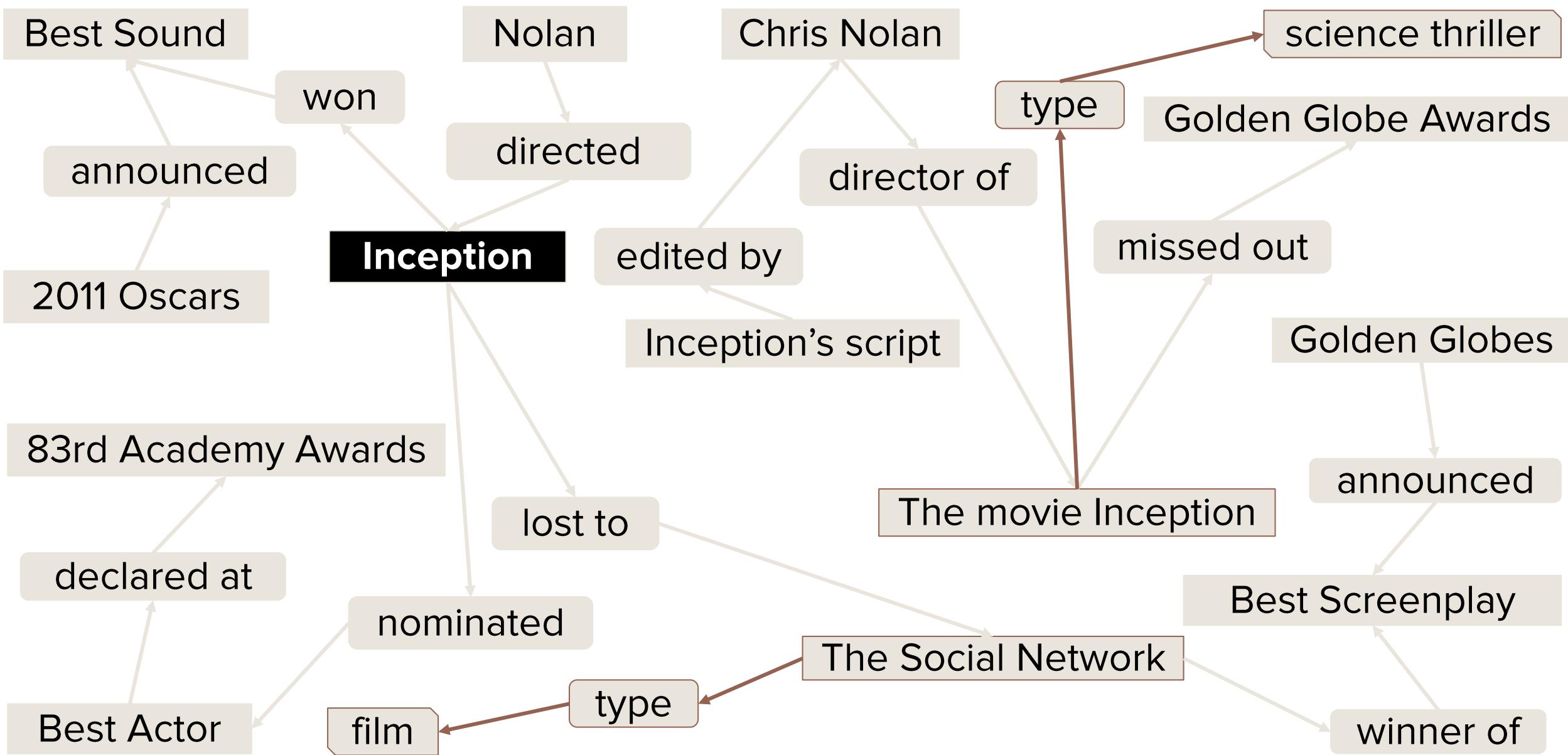
Triples can ideally come from text (via Open IE), KG, or both

Open IE extracts KG-style triples by running pattern extraction over raw text:  
Stanford Open IE, ClausIE, OpenIE 5.0, ...

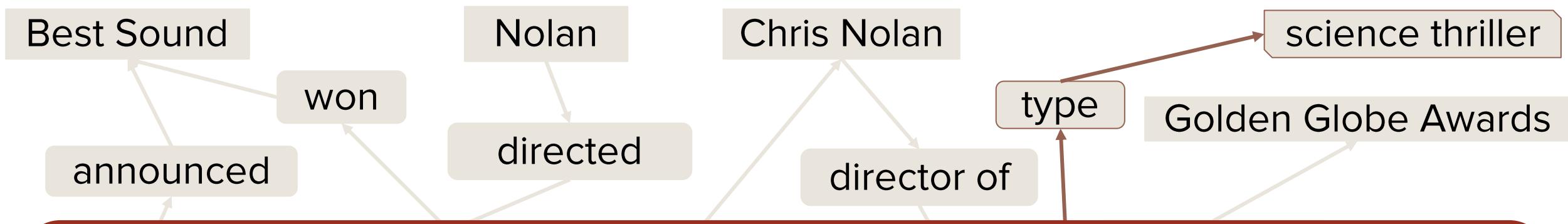
**Question:** Which Nolan films won an Oscar but missed a Golden Globe?



**Question:** Which Nolan films won an Oscar but missed a Golden Globe?

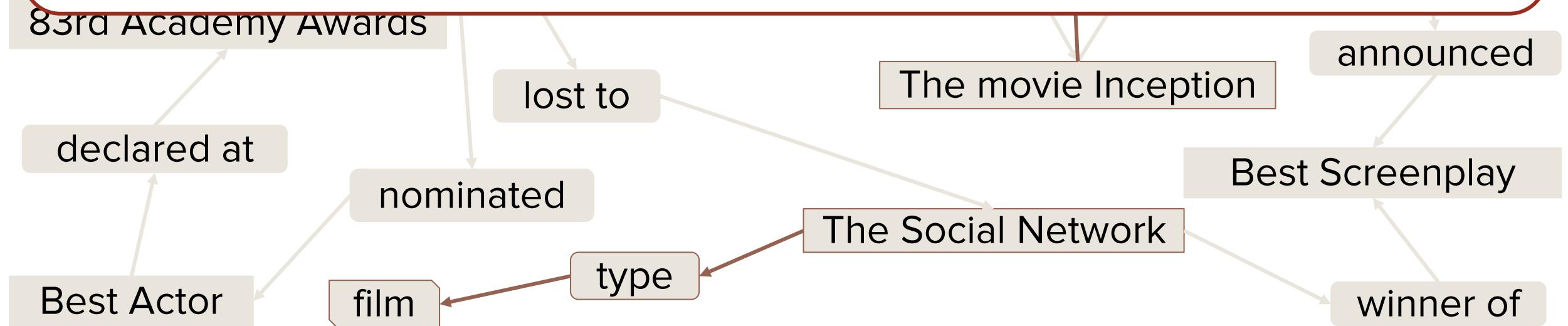


**Question:** Which Nolan films won an Oscar but missed a Golden Globe?

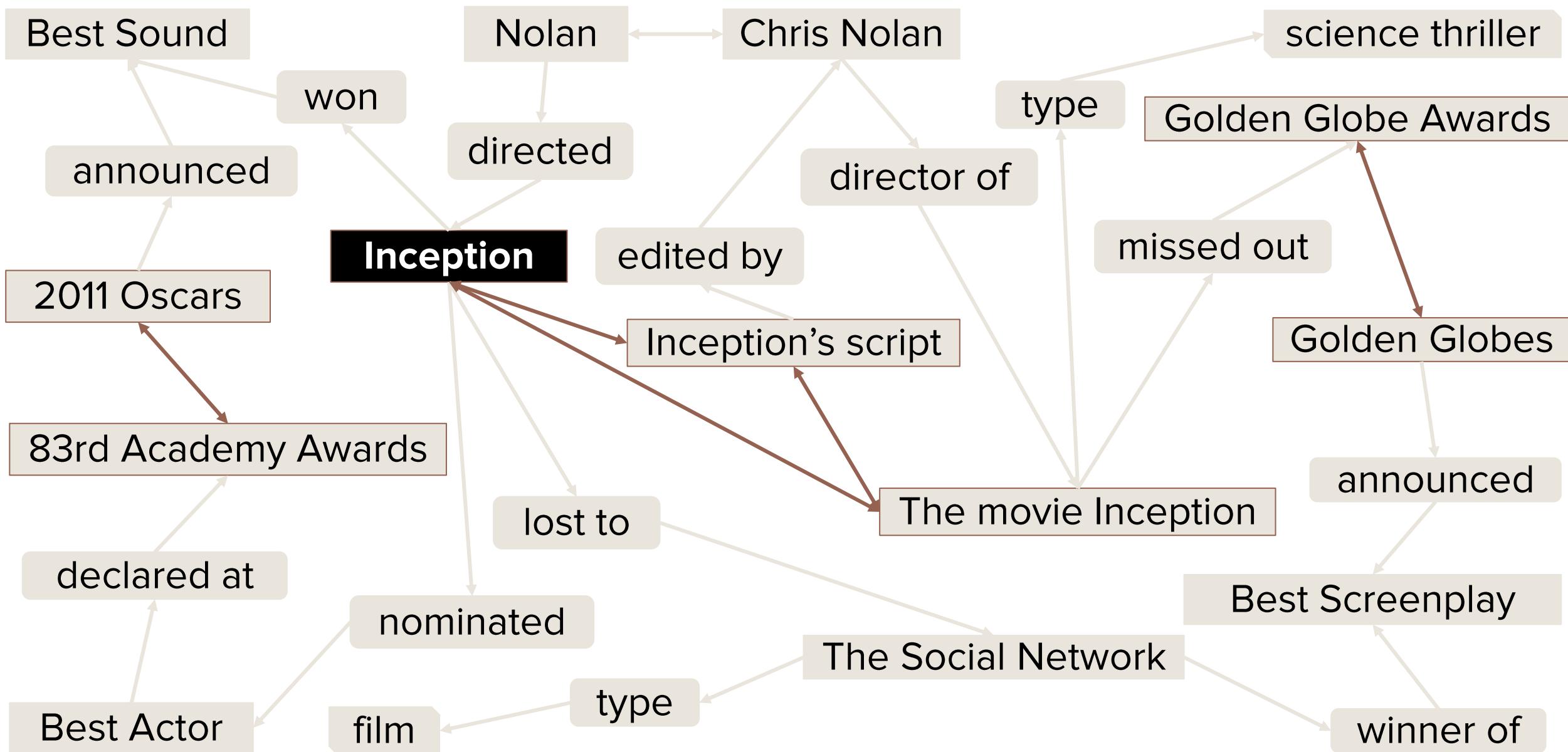


**Entity types are useful for QA:**

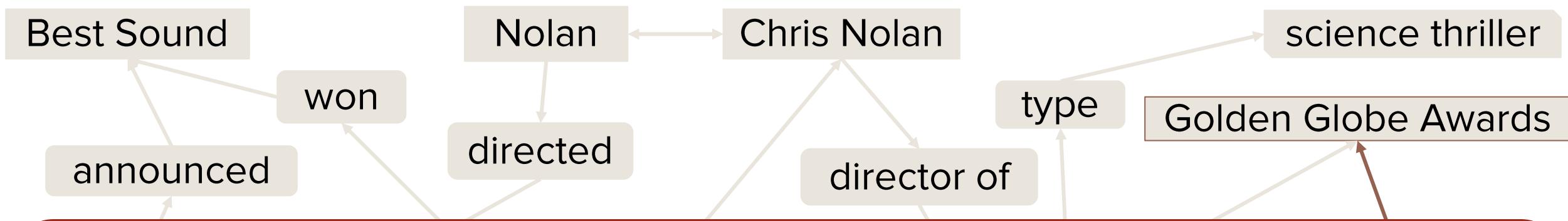
**Add types from KG or running Hearst patterns over text**



**Question:** Which Nolan films won an Oscar but missed a Golden Globe?

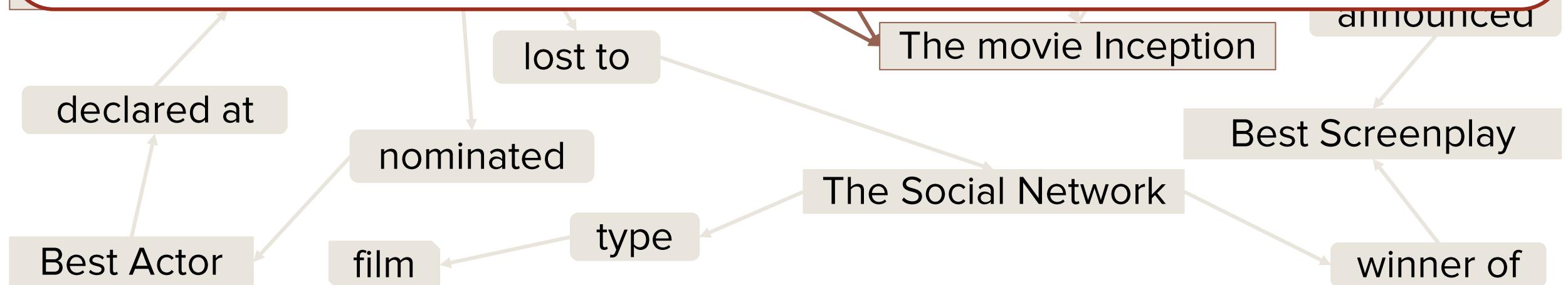


**Question:** Which Nolan films won an Oscar but missed a Golden Globe?

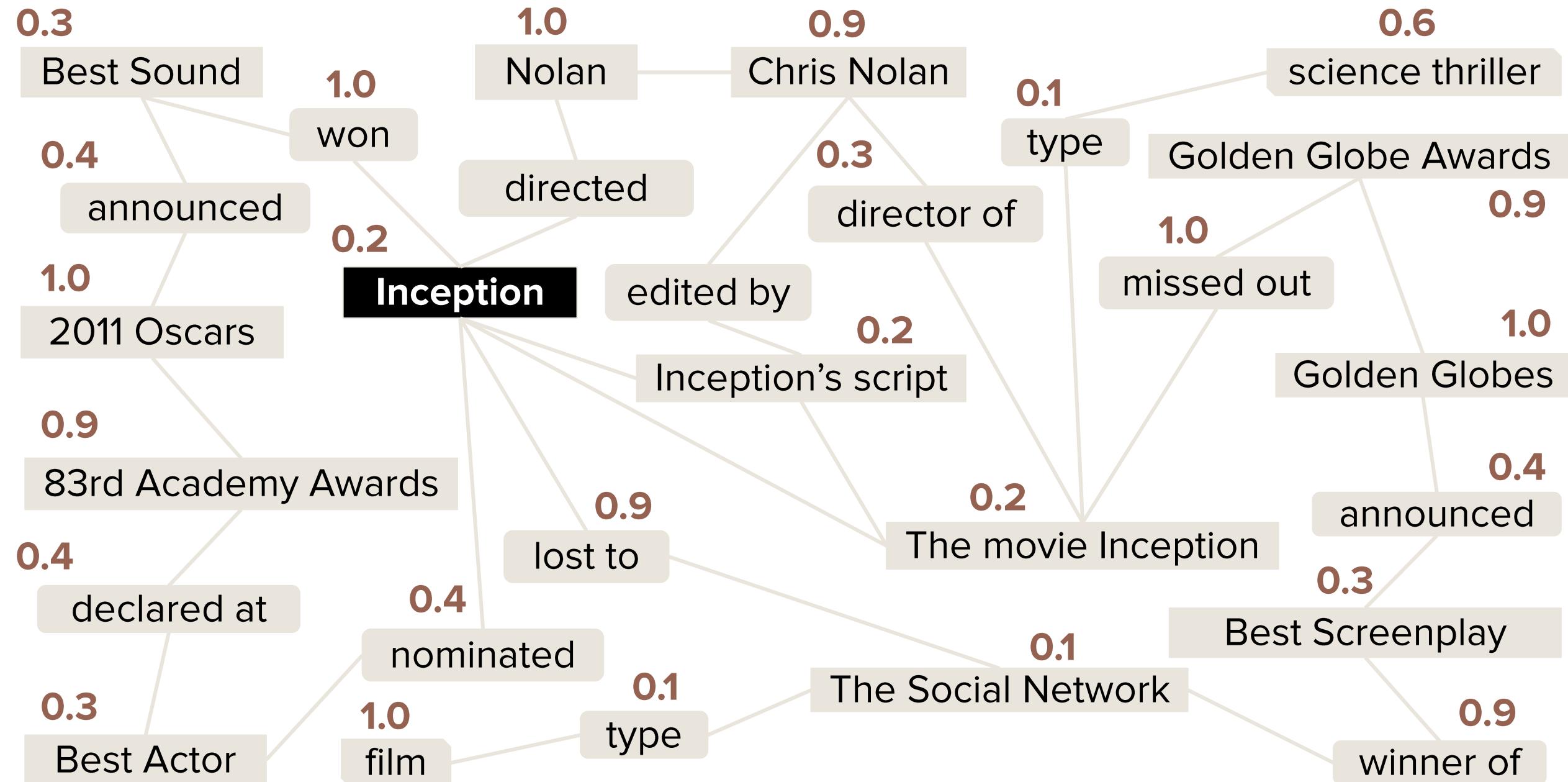


**Lightweight canonicalization helps:**

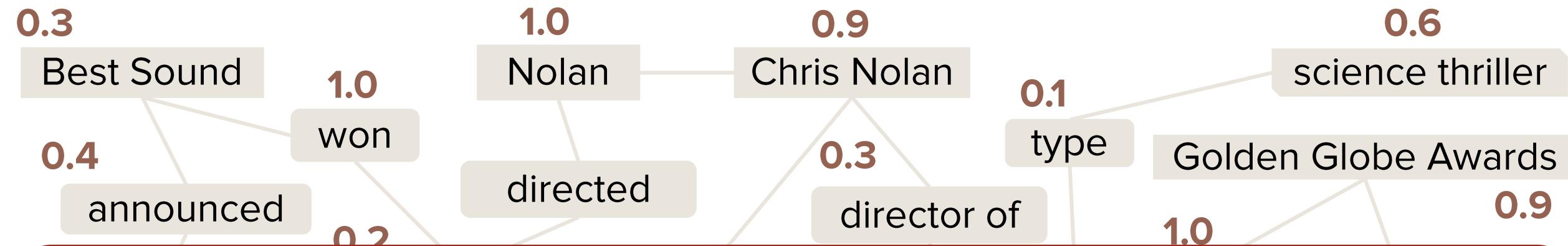
**Insert alignment edges using lexicons and similarity**



**Question:** Which Nolan films won an Oscar but missed a Golden Globe?

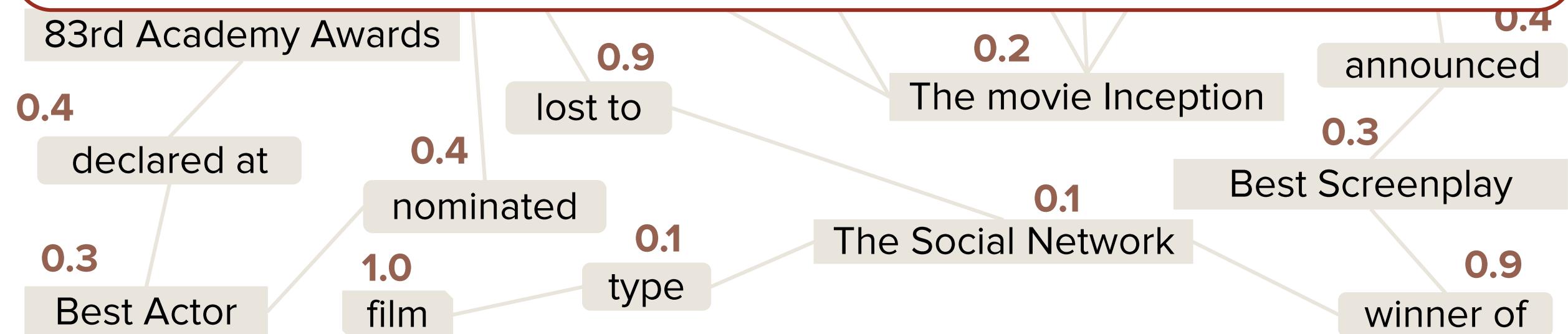


**Question:** Which Nolan films won an Oscar but missed a Golden Globe?

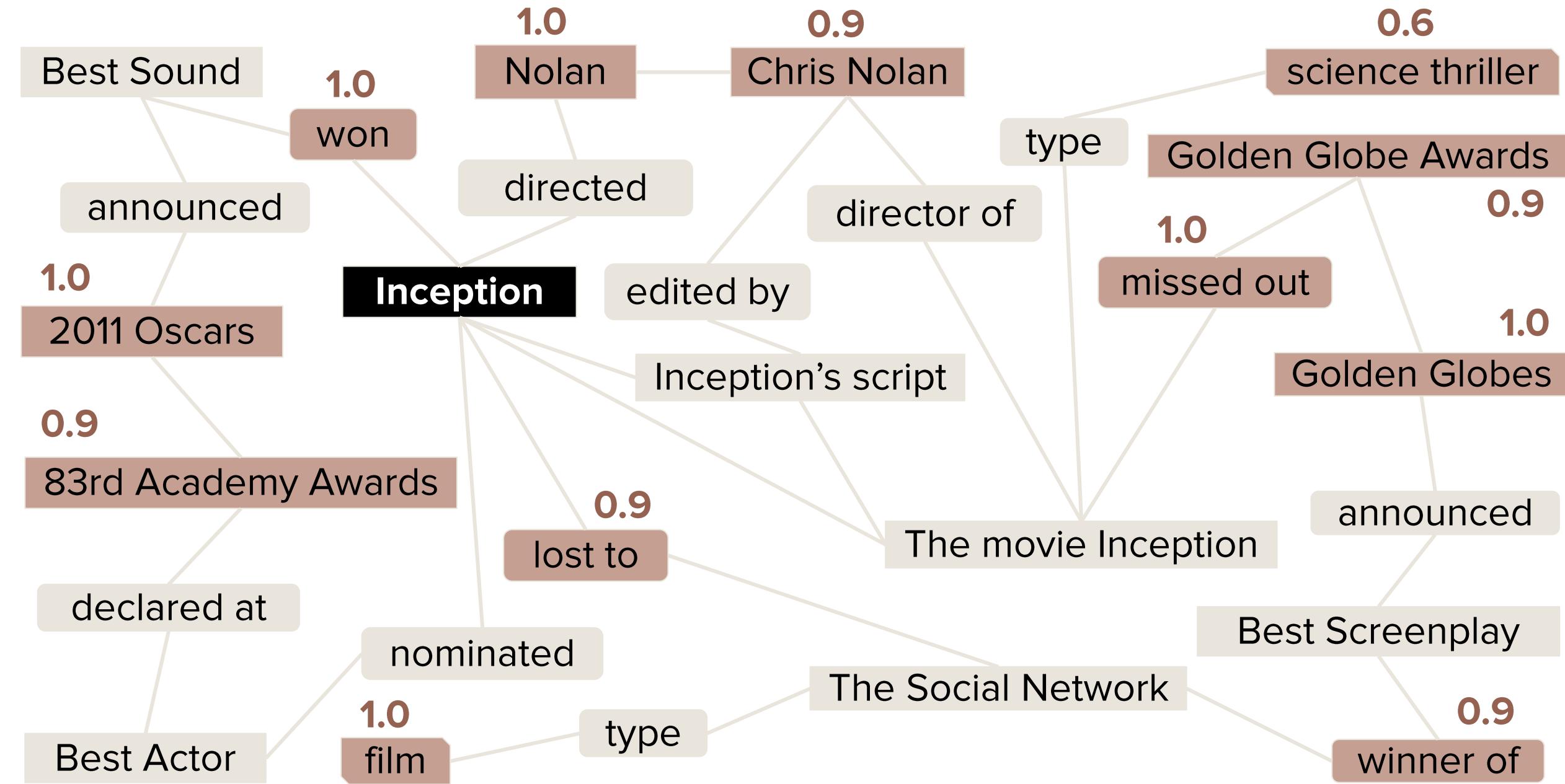


Towards compact subgraph:

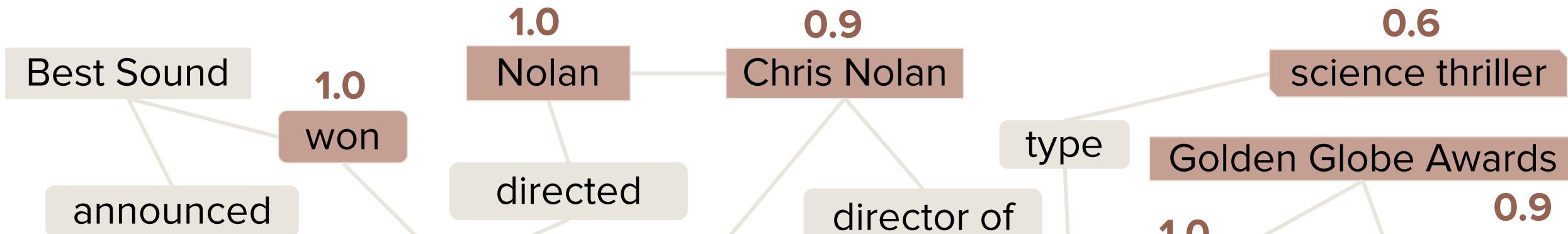
Compute node weights using similarity with question words



**Question:** Which Nolan films won an Oscar but missed a Golden Globe?

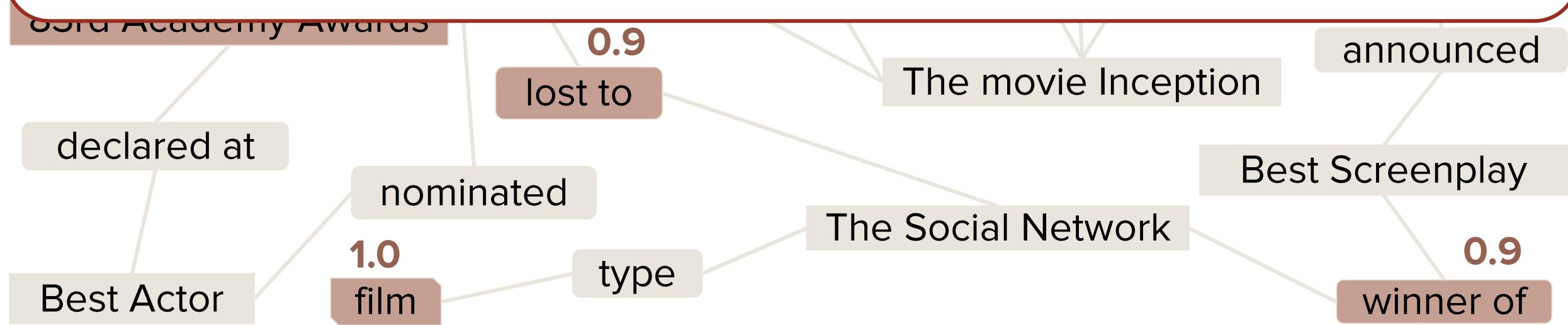


**Question:** Which Nolan films won an Oscar but missed a Golden Globe?

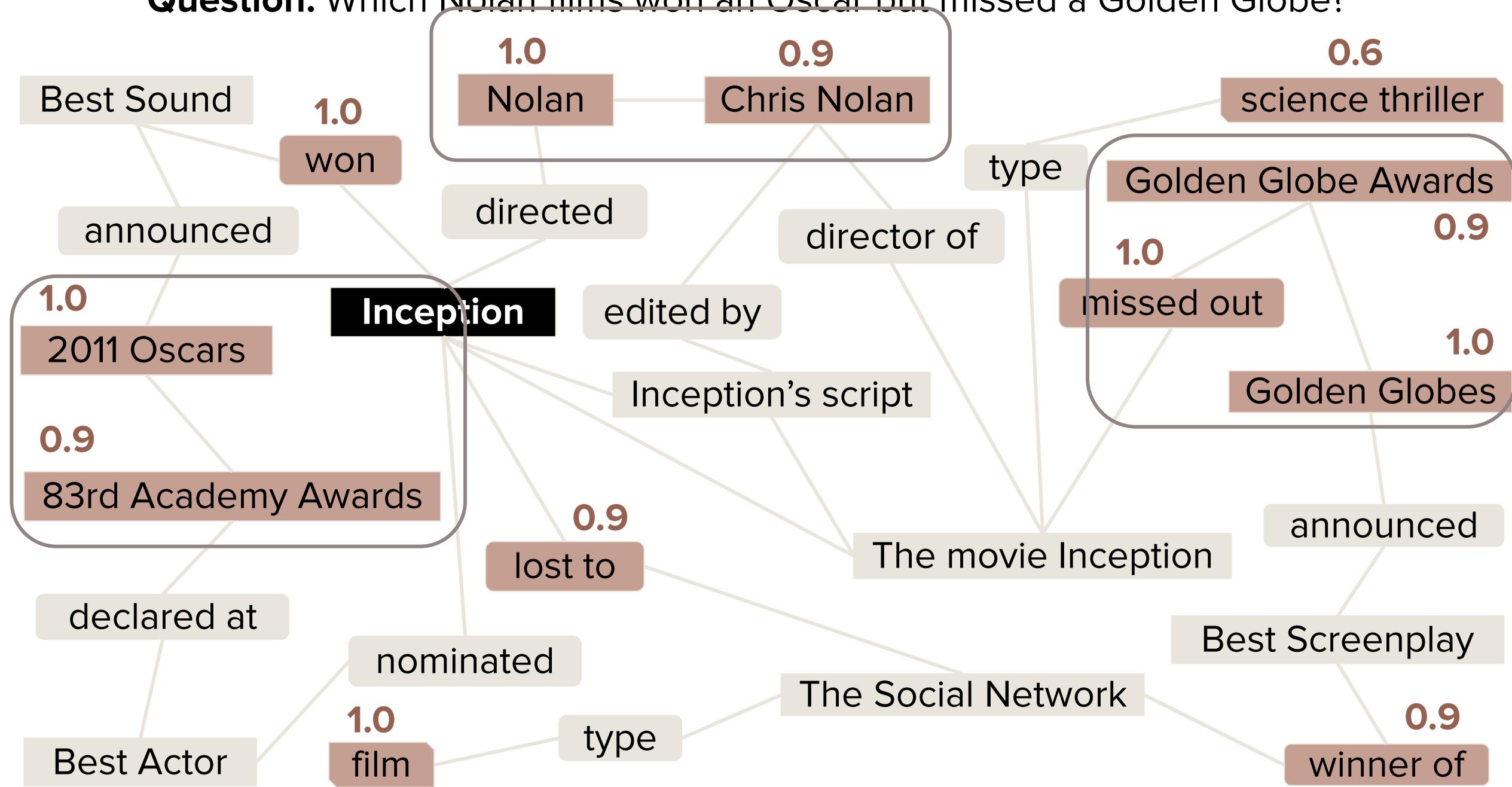


**Find question relevant part of subgraph:**

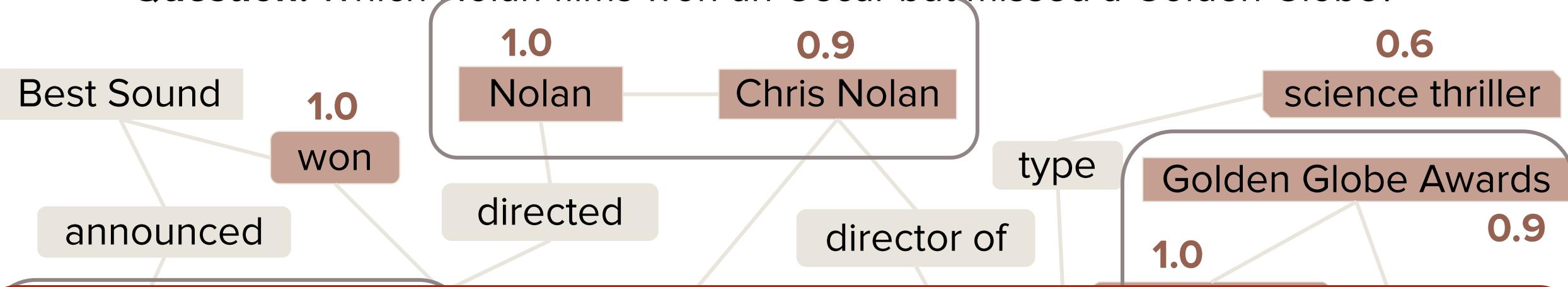
**Identify cornerstones by thresholding node weights**



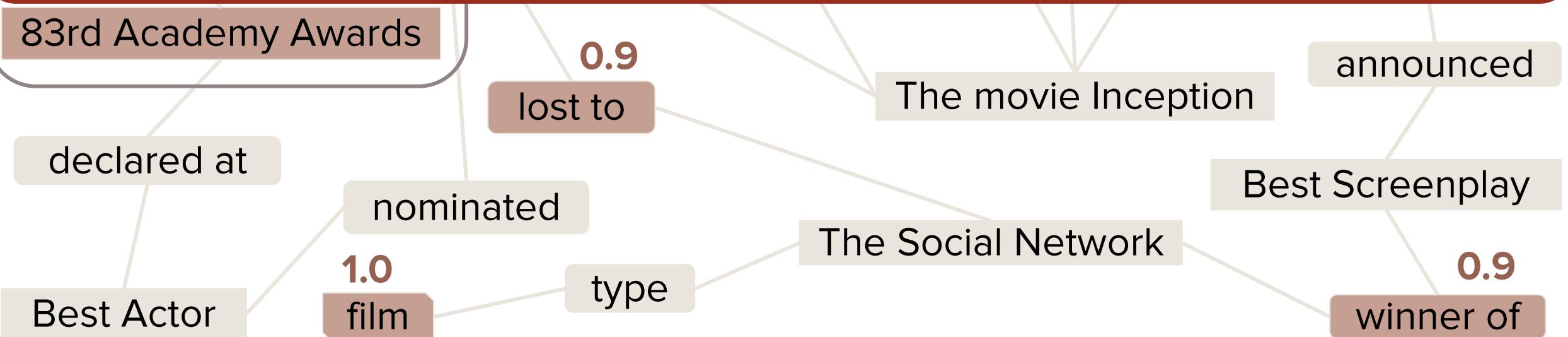
**Question:** Which Nolan films won an Oscar but missed a Golden Globe?



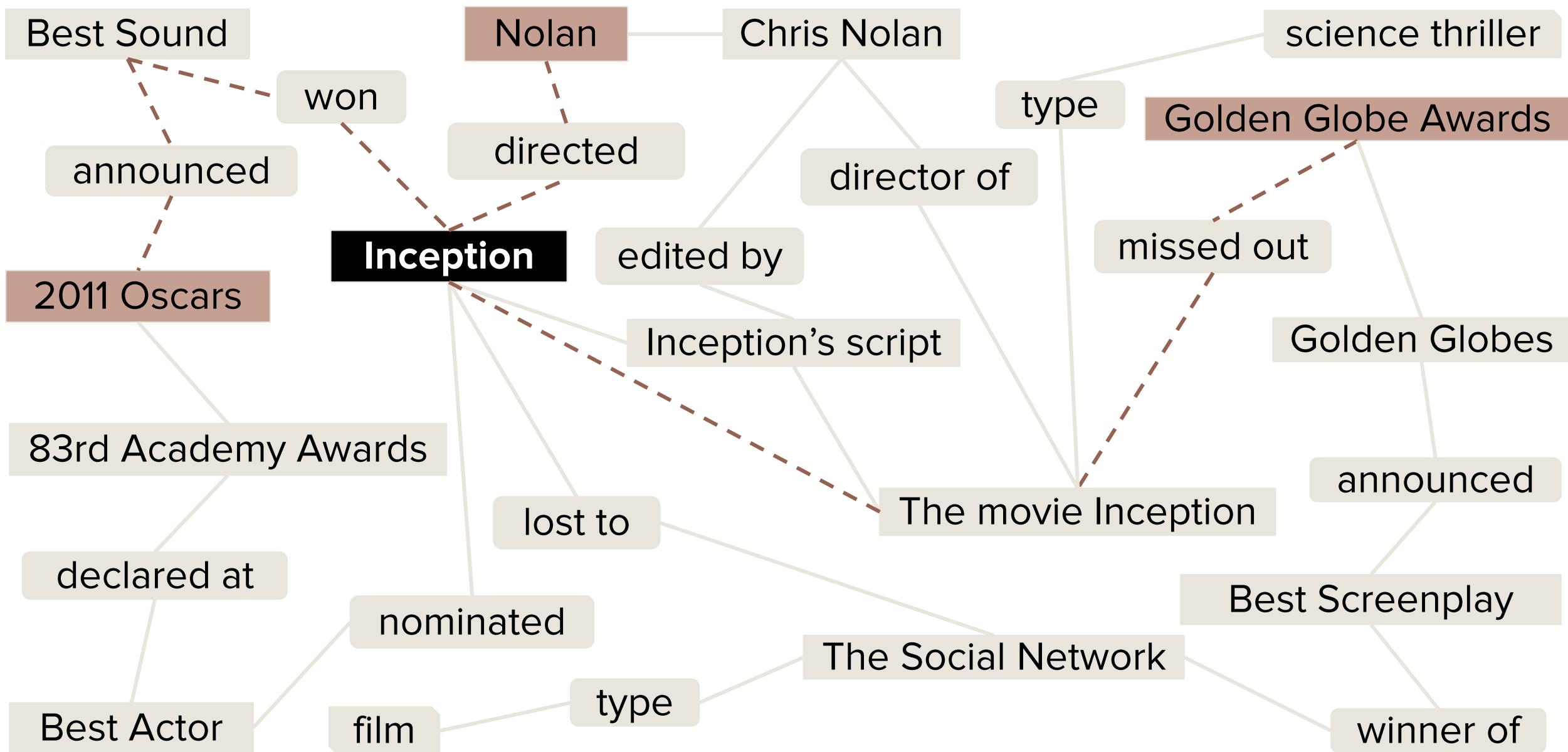
**Question:** Which Nolan films won an Oscar but missed a Golden Globe?



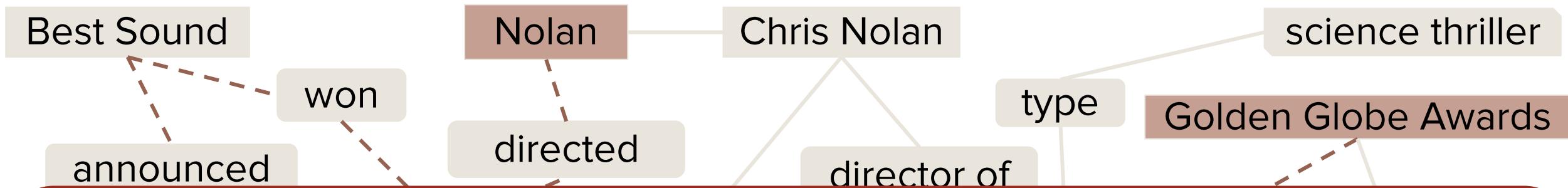
## Cornerstones appear in groups



**Question:** Which Nolan films won an Oscar but missed a Golden Globe?

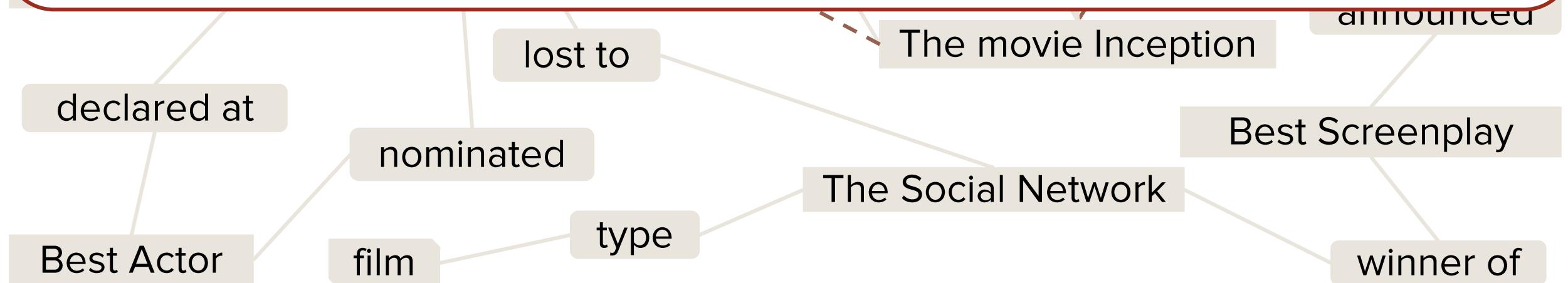


**Question:** Which Nolan films won an Oscar but missed a Golden Globe?

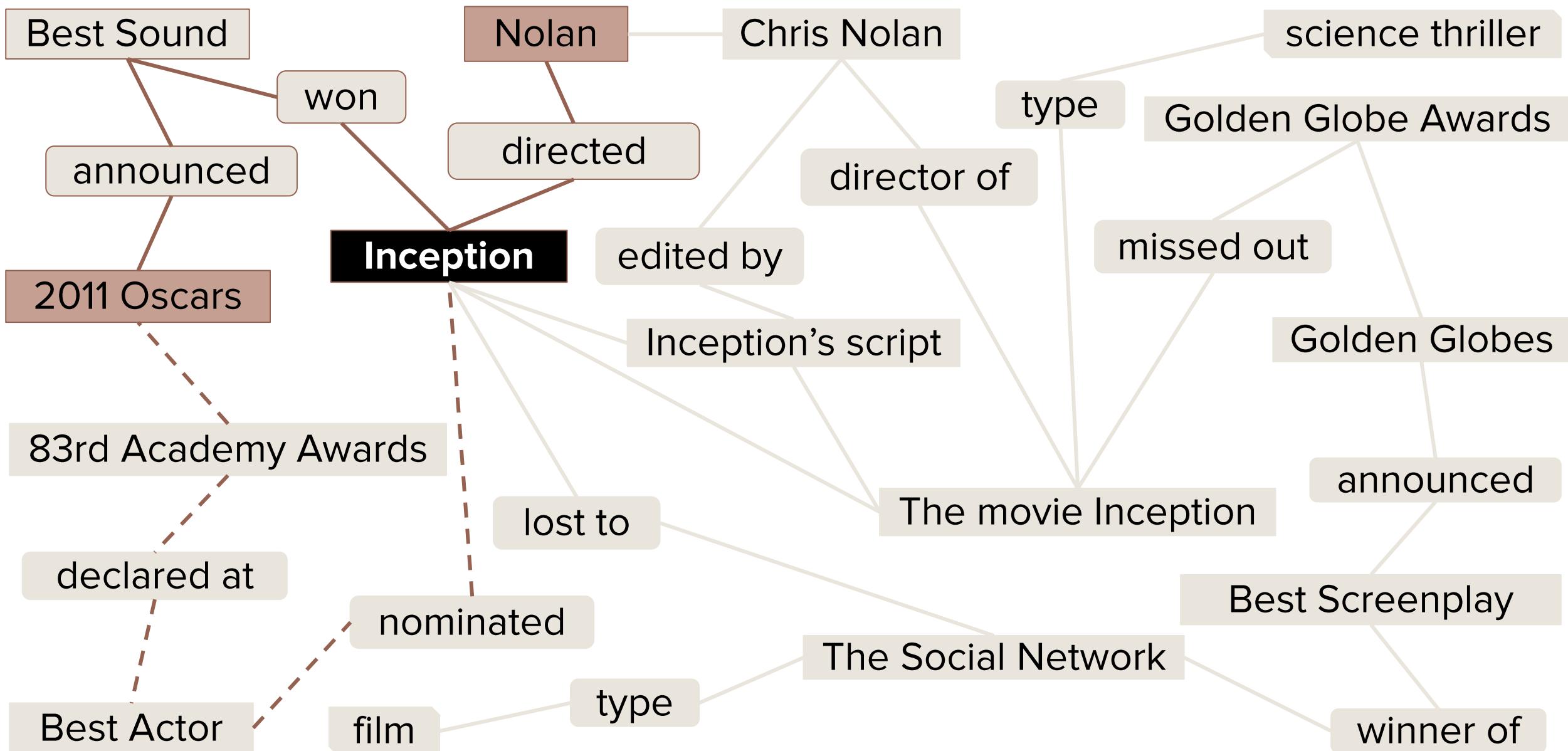


**Answers lie on paths connecting cornerstones:**

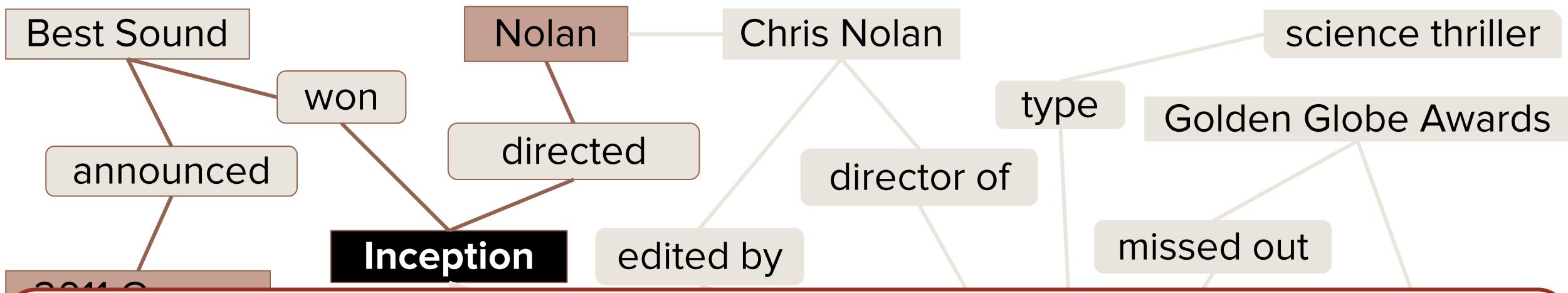
**Internal nodes on paths are answer candidates**



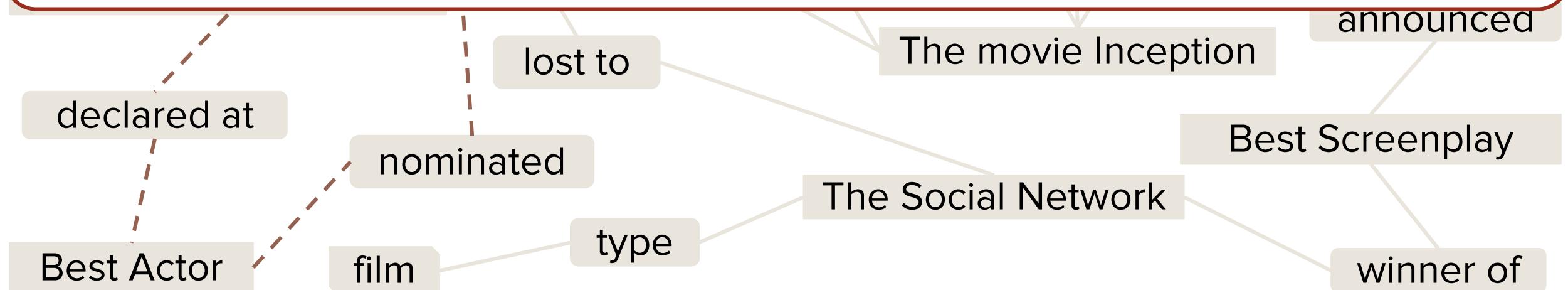
**Question:** Which Nolan films won an Oscar but missed a Golden Globe?



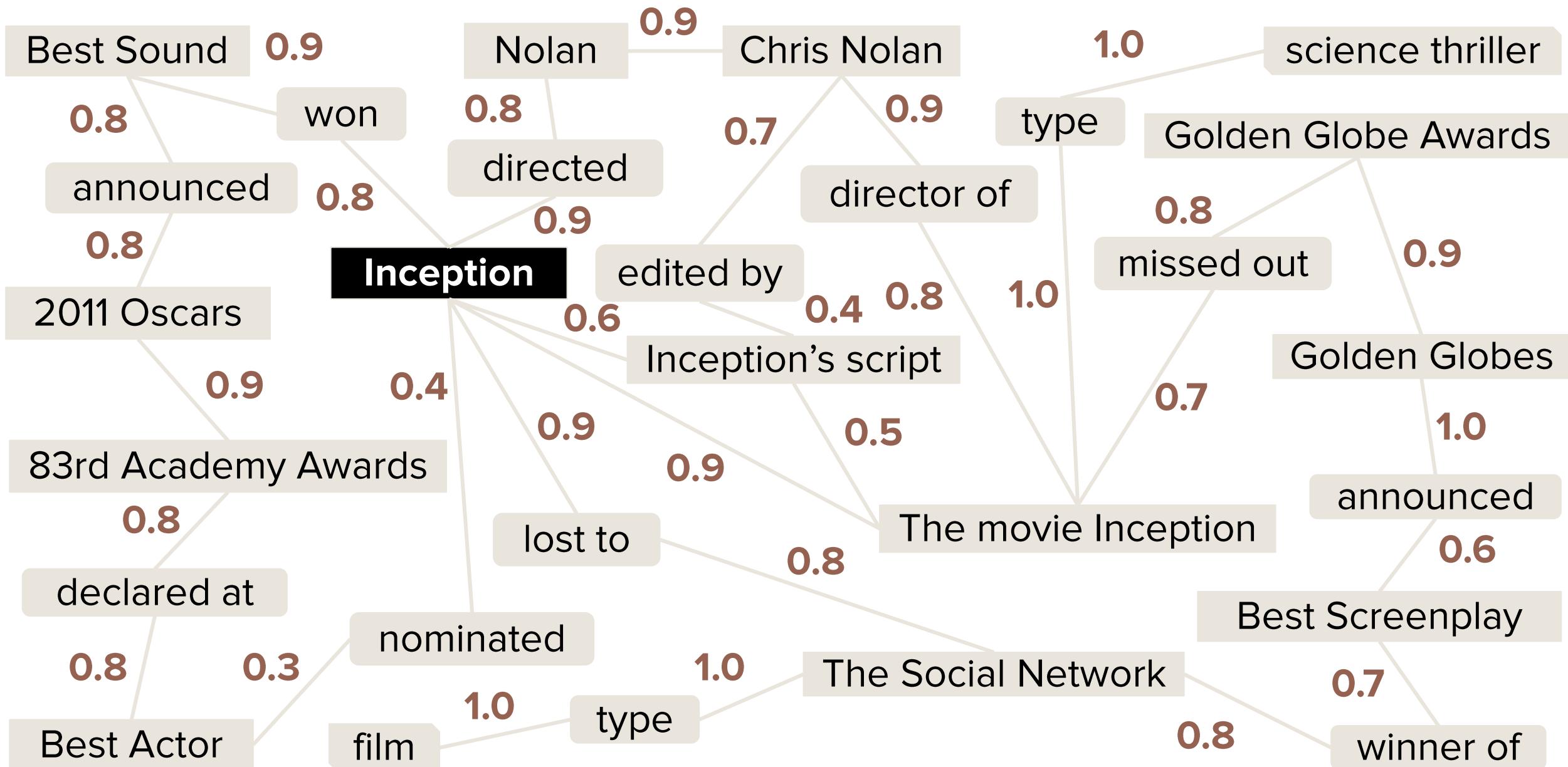
**Question:** Which Nolan films won an Oscar but missed a Golden Globe?



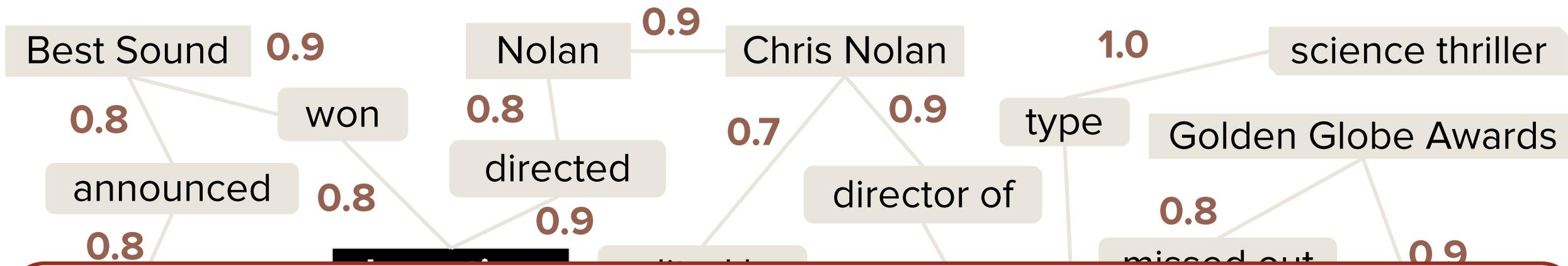
**Shorter paths cleaner for finding answer candidates**



## Question: Which Nolan films won an Oscar but missed a Golden Globe?

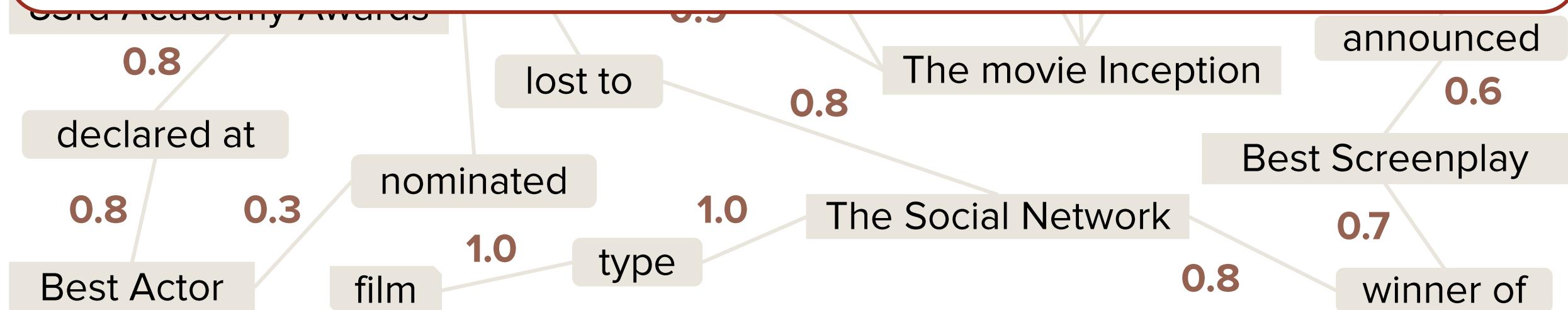


**Question:** Which Nolan films won an Oscar but missed a Golden Globe?

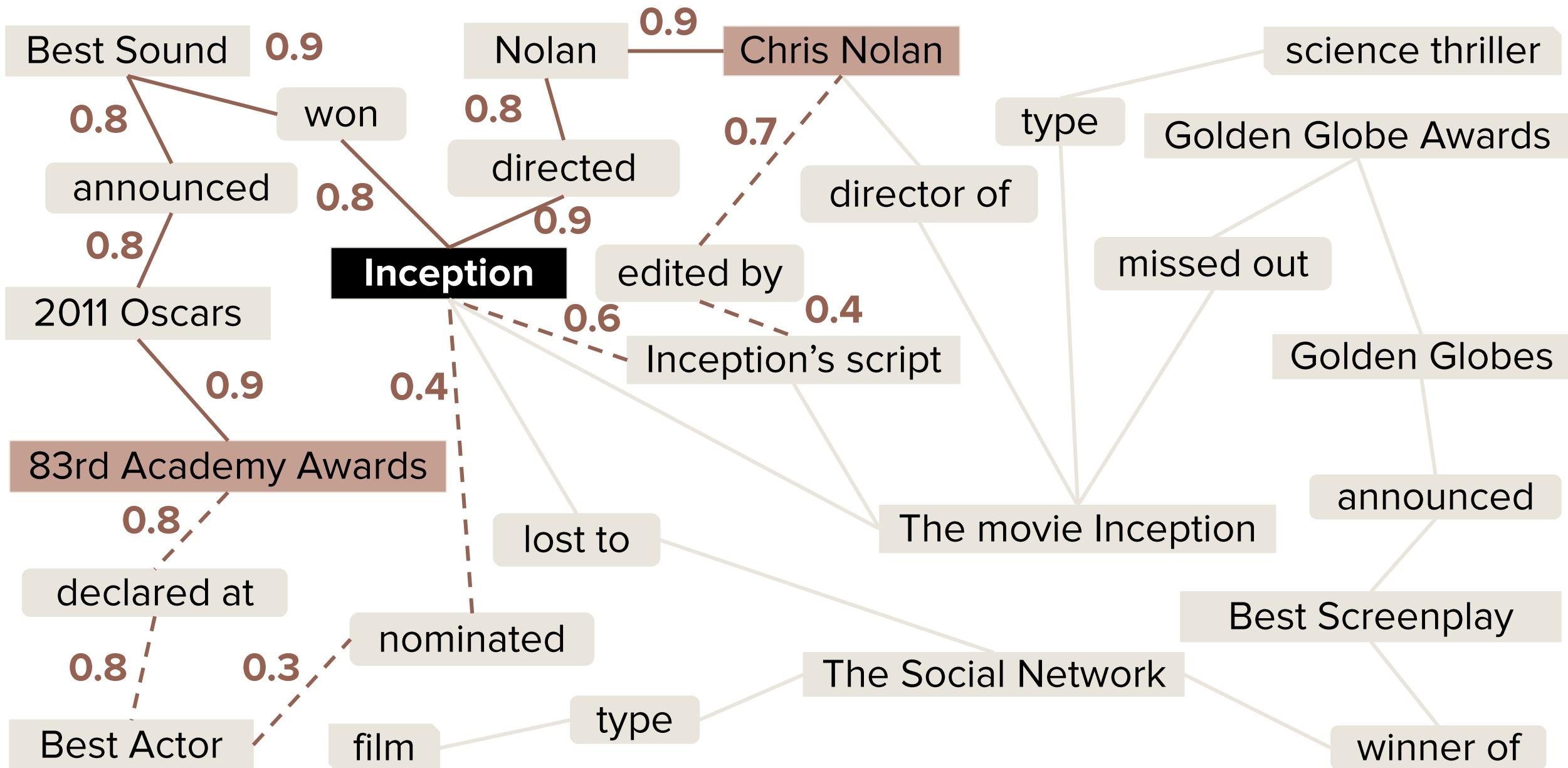


**Weighted edges more meaningful for distance computations:**

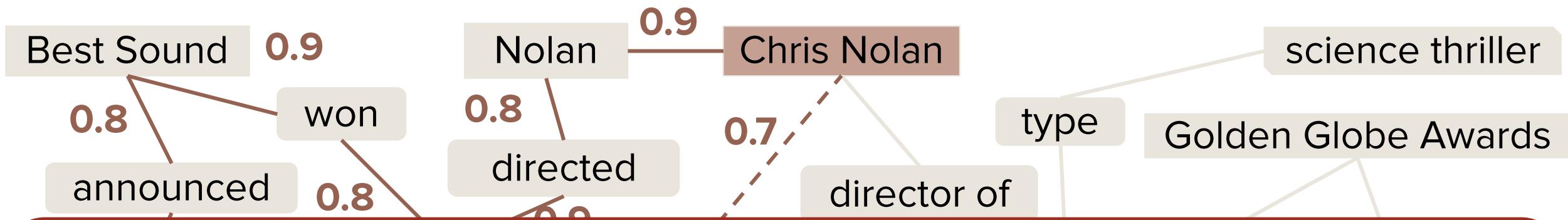
**Insert edge weights with term proximities and lexicon similarities**



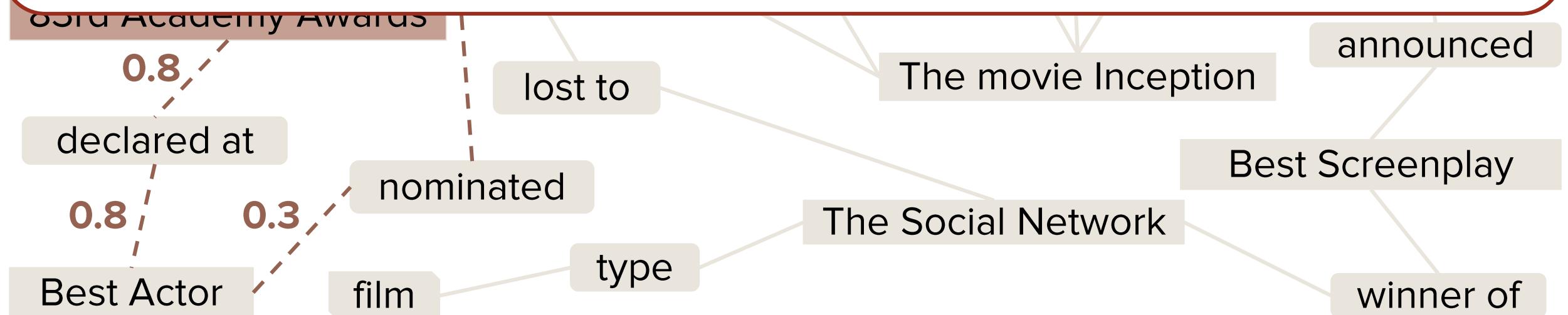
**Question:** Which Nolan films won an Oscar but missed a Golden Globe?



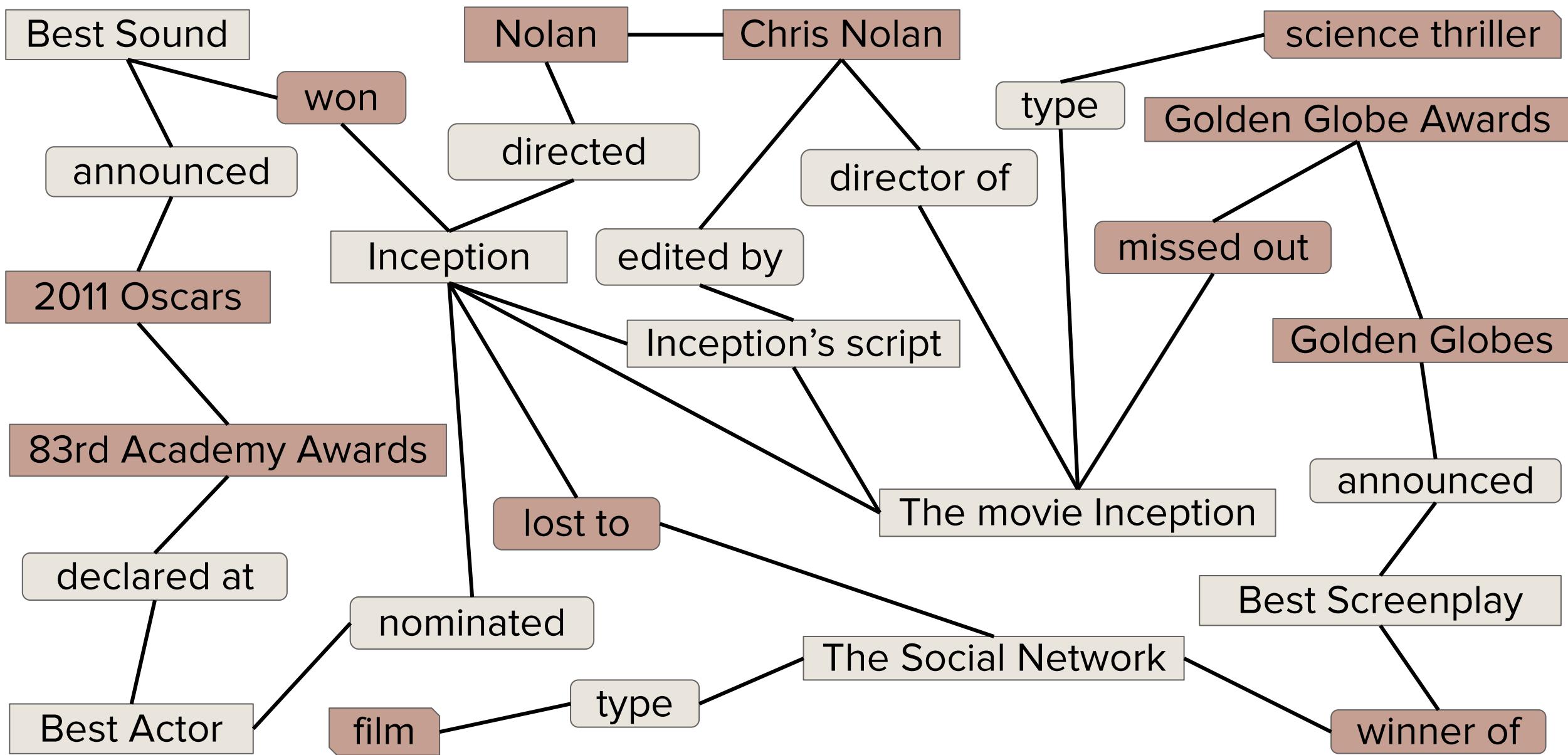
**Question:** Which Nolan films won an Oscar but missed a Golden Globe?



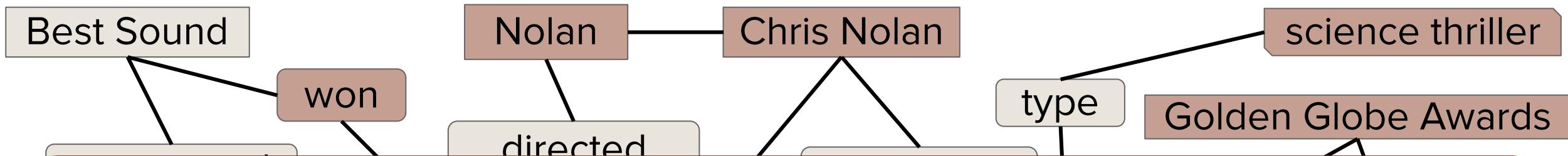
**Higher weight paths cleaner for finding answer candidates**



**Question:** Which Nolan films won an Oscar but missed a Golden Globe?

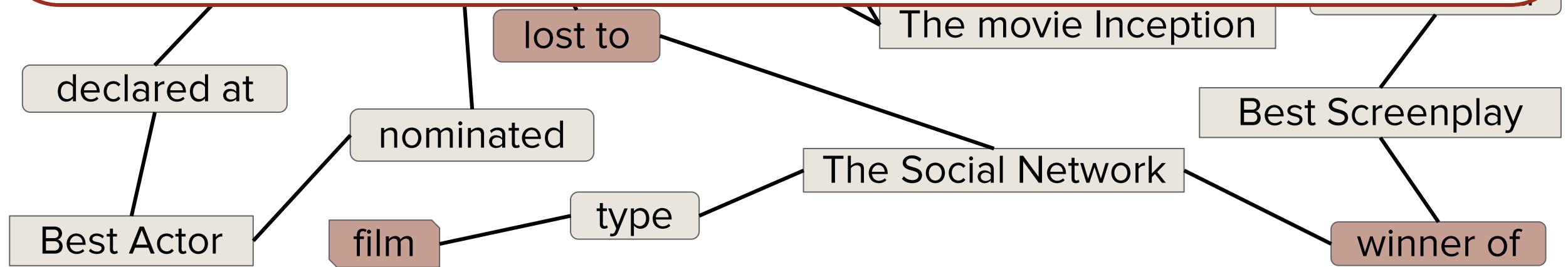


**Question:** Which Nolan films won an Oscar but missed a Golden Globe?

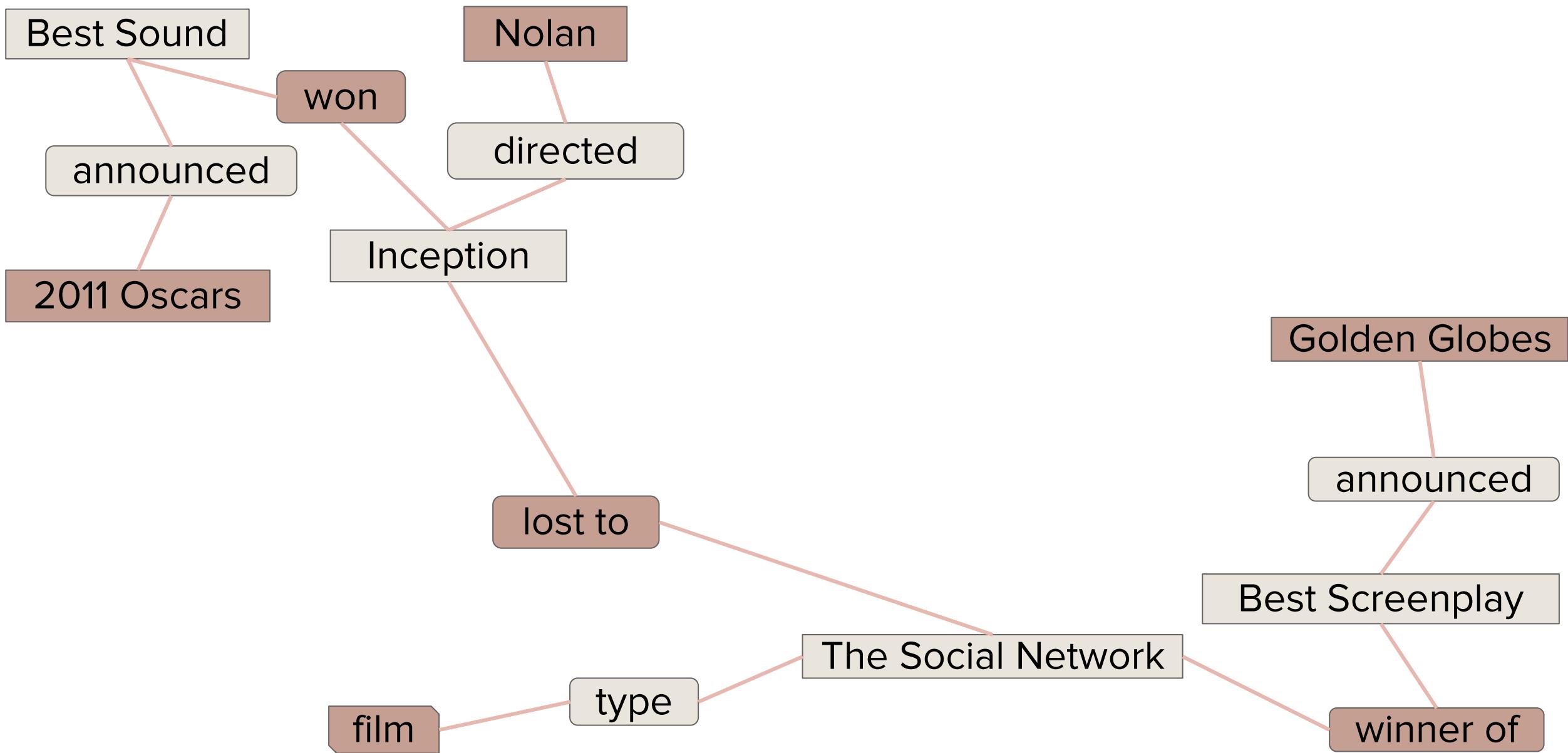


**Unsupervised computation of dense and compact subgraph:**

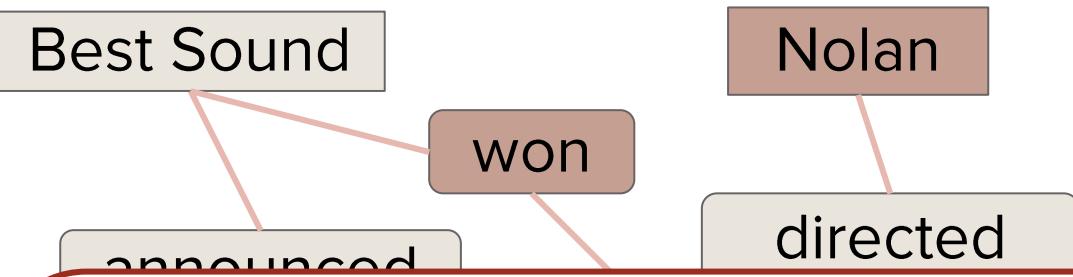
**Joint disambiguation of question concepts**



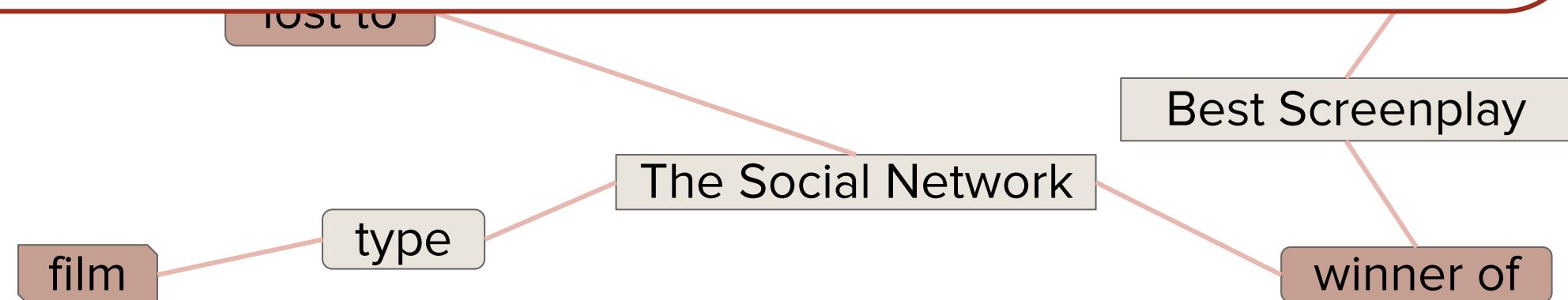
**Question:** Which Nolan films won an Oscar but missed a Golden Globe?



**Question:** Which Nolan films won an Oscar but missed a Golden Globe?



**Compute Group Steiner Tree (GST) on quasi-KG with  
cornerstones as terminals: Optimal connections between  
question concepts for faithful answering**



## Question: Which Nolan films won an Oscar but missed a Golden Globe?

Best Sound

Nolan

won

announced

directed

2011 Oscars

Inception

- **Answers are nodes**

- Cornerstones are not answers
- Only entities
- Must respect type constraints

Golden Globes

announced

lost to

Best Screenplay

The Social Network

film

type

winner of

**Question:** Which Nolan films won an Oscar but missed a Golden Globe?

Best Sound

won

announced

directed

Inception

- Answers are nodes
- **Cornerstones are not answers**
- Only entities
- Must respect type constraints

lost to

announced

type

The Social Network

Best Screenplay

winner of

**Question:** Which Nolan films won an Oscar but missed a Golden Globe?

Best Sound

Inception

- Answers are nodes
- Cornerstones are not answers
- **Only entities**
- Must respect type constraints

Best Screenplay

The Social Network

## Question: Which Nolan films won an Oscar but missed a Golden Globe?

Inception

- Answers are nodes
- Cornerstones are not answers
- Only entities
- **Must respect type constraints**

The Social Network

# **Question:** Which Nolan films won an Oscar but missed a Golden Globe?

- **Answers ranked by aggregation**
- Best answer chosen

Inception

Number of GSTs	5
----------------	---

Number of GSTs

2

The Social Network

# Question: Which Nolan films won an Oscar but missed a Golden Globe?

- Answers ranked by multiple criteria
- **Best answer chosen**

Inception

# Complex QA: Graph-based belief propagation

- The **QAmp** system (Vakulenko et al. CIKM 2019)
- Interpretation
  - Parsing
  - Matching
- Reasoning
  - **Message passing**
  - Score aggregation

API access possible by appending the text of a question to <https://kbqa-api.ai.wu.ac.at/ask?question=>

For example, for “Name the municipality of Roberto Clemente Bridge?”, use:

<https://kbqa-api.ai.wu.ac.at/ask?question=Name%20the%20municipality%20of%20Roberto%20Clemente%20Bridge%20?>

# Interpretation: Parsing by sequence labeling

Where is the founder of Tesla born?

P1

E1

P2

Use CRFs trained on labeled data

Predict question type: **SELECT**, COUNT, ASK

# Interpretation: Matching

Where is the founder of Tesla born?

P1

E1

P2

Use CRFs trained on labeled data

Predict question type: **SELECT**, COUNT, ASK

P1

founder

1

founded

0.8

P2

bornIn

0.8

E1

Tesla

1

Nicola Tesla

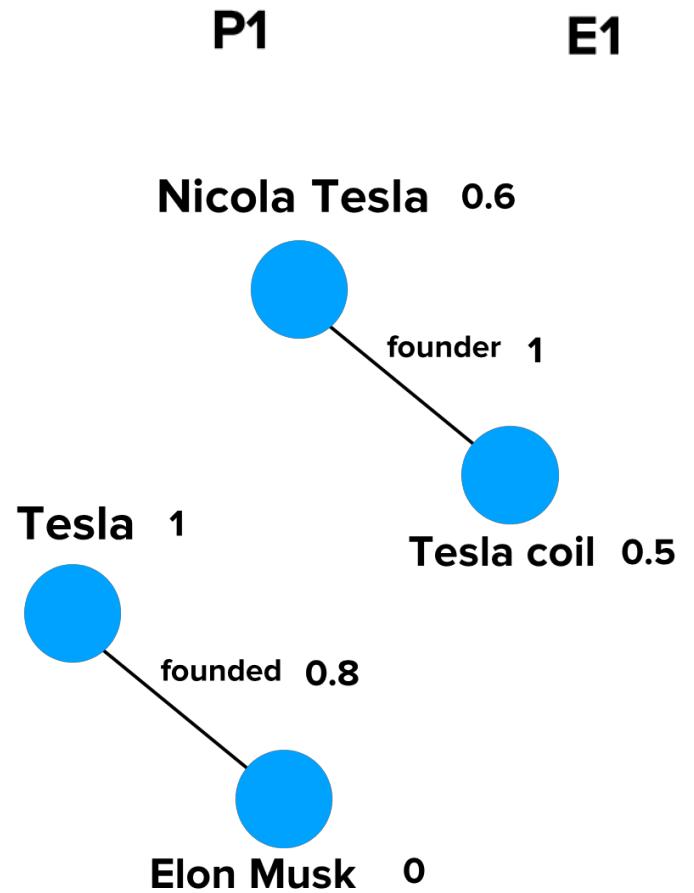
0.6

Tesla coil

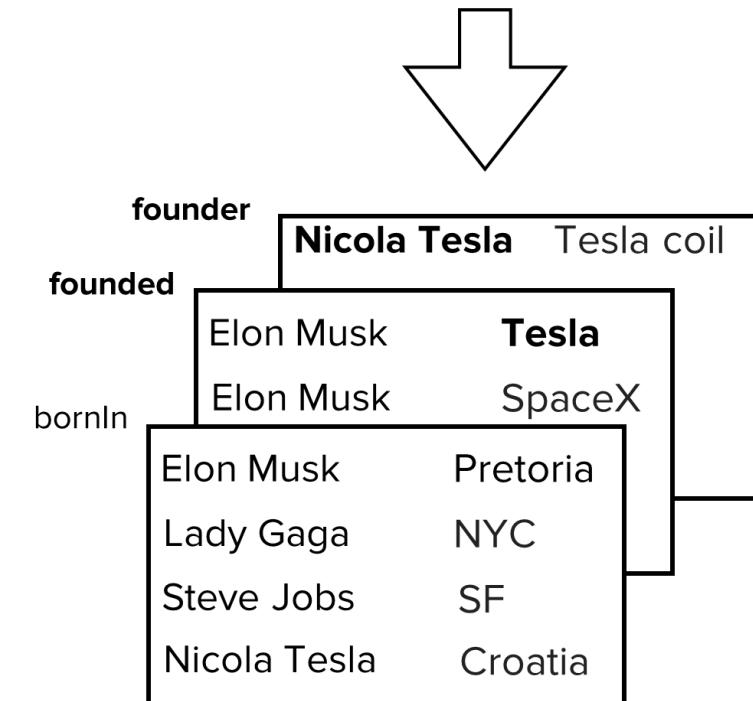
0.5

# Reasoning: Message passing Hop 1

Where is the founder of Tesla born?

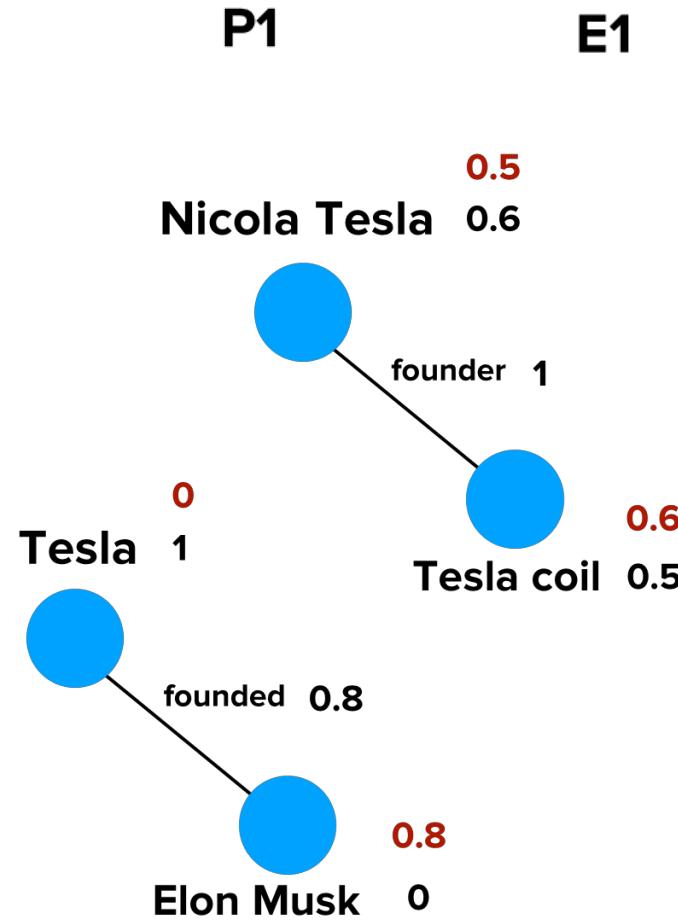


P1	E1	P2		P1	E1	
founder	1			Tesla	1	
founded	0.8			Nicola Tesla	0.6	
				Tesla coil	0.5	

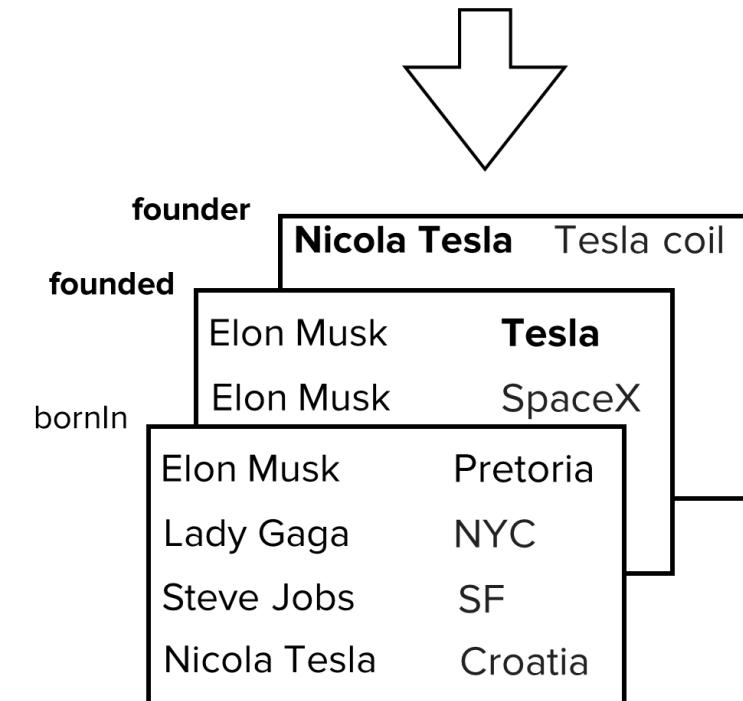


# Reasoning: Message passing Hop 1

Where is the founder of Tesla born?

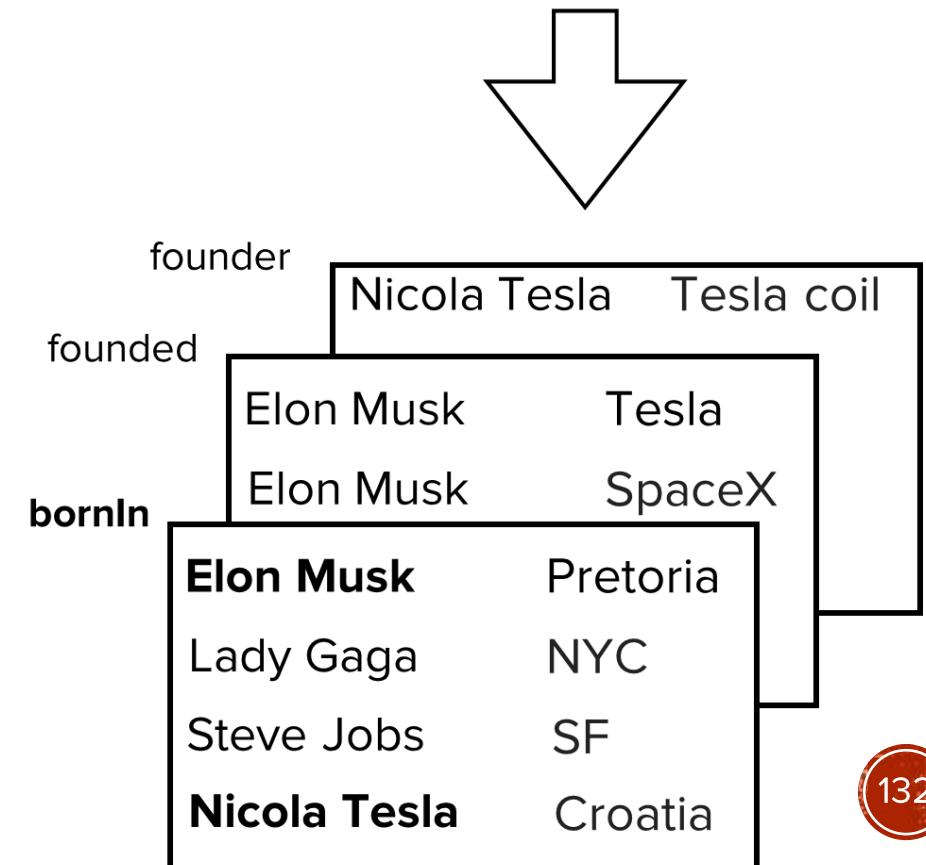


P1	E1
founder	1
founded	0.8

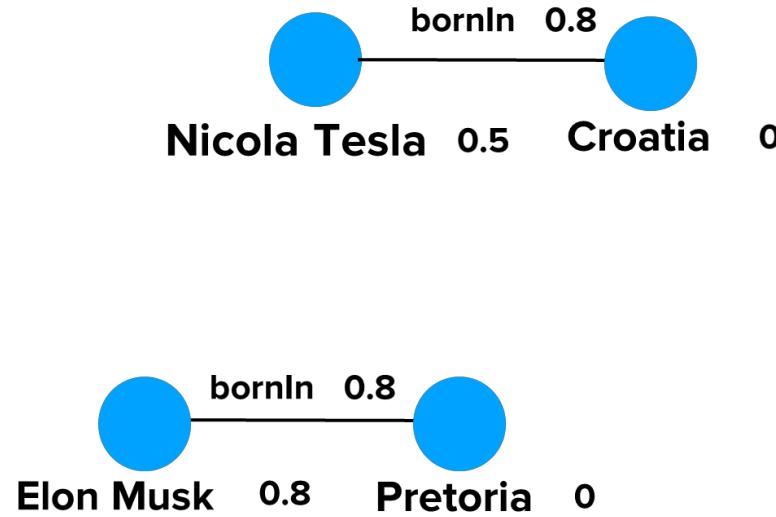


# Reasoning: Message passing Hop 2

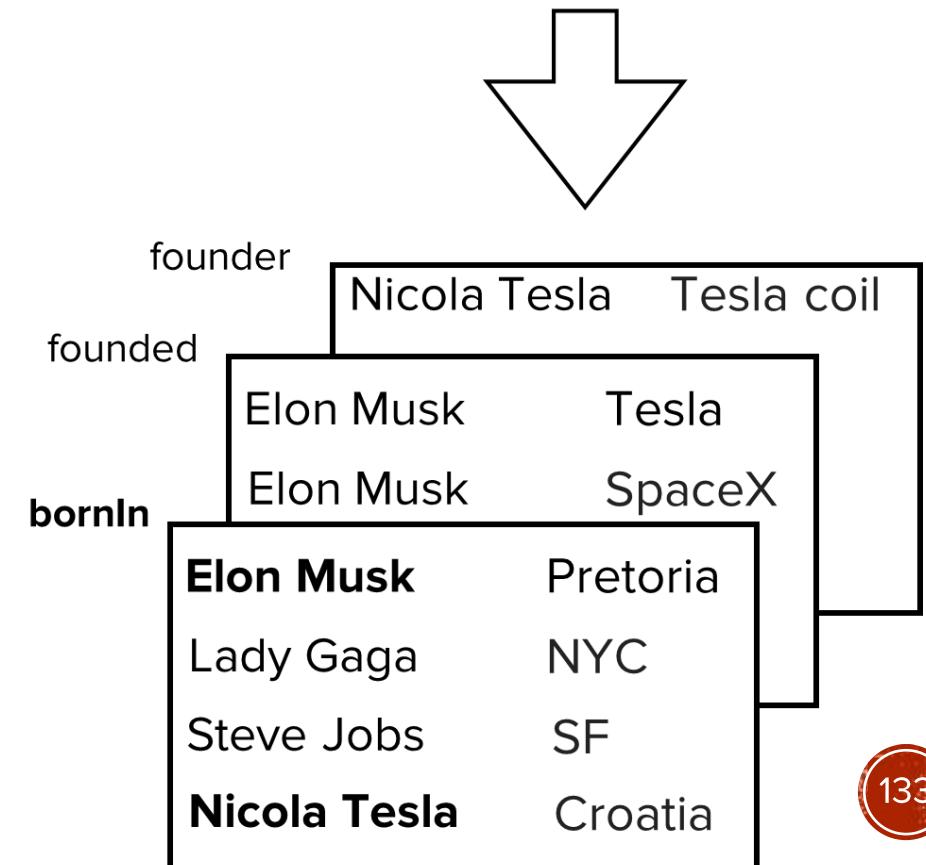
P2	E2
bornIn 0.8	Elon Musk 0.8
	Tesla coil 0.6
	Nicola Tesla 0.5



# Reasoning: Message passing Hop 2

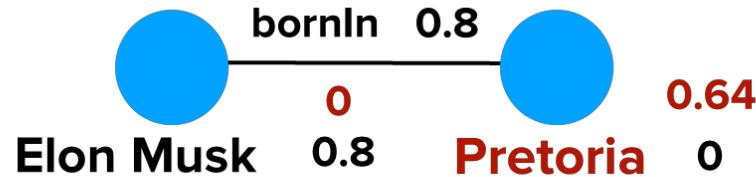
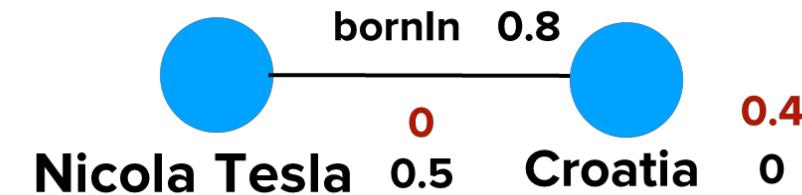


P2	E2		
bornIn	0.8	Elon Musk	0.8
		Tesla coil	0.6
		Nicola Tesla	0.5

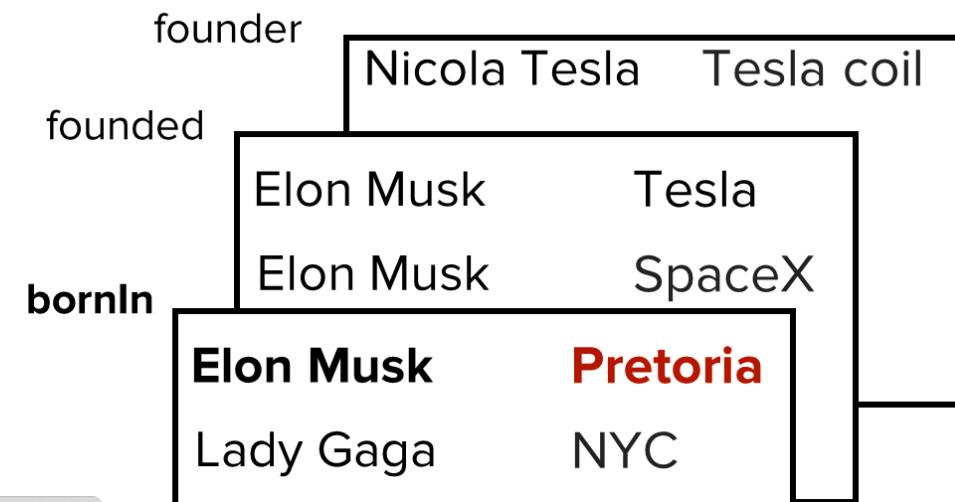


# Reasoning: Message passing Hop 2

Where is the founder of Tesla born?



P2	E2
bornIn 0.8	Elon Musk 0.8
	Tesla coil 0.6
	Nicola Tesla 0.5



Constructs explanatory evidence

Efficient (2 sec/ques) for use of HDT (recall QAnswer) + matrix operations

# Complex questions: Wrap-up

- Complex KG-QA the sub-topic with the **highest attention**
- **Efficiency** generally an open issue: several partial queries executed in TextRay, several similarity computations in QUEST, ...
- Bias in SoTA towards **certain classes**: QAmp (**chains**), QUEST (stars), ...
- How to reduce **large neighborhood sizes**? KGs are dense: considering full 2-hop neighborhoods often intractable due to popular entities or general types
- But complex QA is **much, much more**: Star joins, temporal intent, aggregations, comparatives, superlatives, reasoning, existential, ...

break duration ?x .  
?x measuredIn minutes .

# Outline: QA over knowledge graphs

- **Background:** Setup, benchmarks, metrics
- **Simple QA:** Templates and embeddings
- **Complex QA:** Multiple entities and predicates
- **Heterogeneous sources:** Handling KGs and text
- **Conversational QA:** Implicit context in multi-turn setup
- **Take-home:** Summary and insights

# How can we answer questions over heterogeneous sources?

# QA over heterogeneous sources

- **Heterogeneous source:** System should tap into multiple KGs, or KG + Text
- **Why fuse?** Each source has its advantages and disadvantages
- **Early fusion:** GRAFT-Net ([Sun et al. 2018](#)), **PullNet** ([Sun et al. 2019](#))
- **Late fusion:** [Ferrucci et al. \(2010\)](#), [Baudis \(2015\)](#), [Sun et al. \(2015\)](#), Xu et al. ([2016a](#), [2016b](#)), [Savenkov and Agichtein \(2016\)](#)
- **Unified representations:** OQA ([Fader et al. 2014](#)), TriniT ([Yahya et al. 2016](#)), UniSchema ([Das et al. 2017](#)), Nestique ([Bhutani et al. 2019b](#)), **QAnswer** ([Diefenbach et al. 2019](#))

# Heterogeneous QA: Early fusion

- The **PullNet** system ([Sun et al. 2019](#))
- Fusion via **KG facts** and **KG-entity linked sentences**
- Built for **multi-hop** questions
- Uses **question-focused subgraph**
- **Judiciously expands context** subgraph
- Uses classifiers for **expansion points** and answers

Sun et al., PullNet: Open Domain Question Answering with Iterative Retrieval on Knowledge Bases and Text, EMNLP 2019.

# PullNet: Handling heterogeneity

KG facts

<ChristopherNolan, birthplace, London>

<Memento, director, Nolan>

<Interstellar, castMember, AnneHathaway>

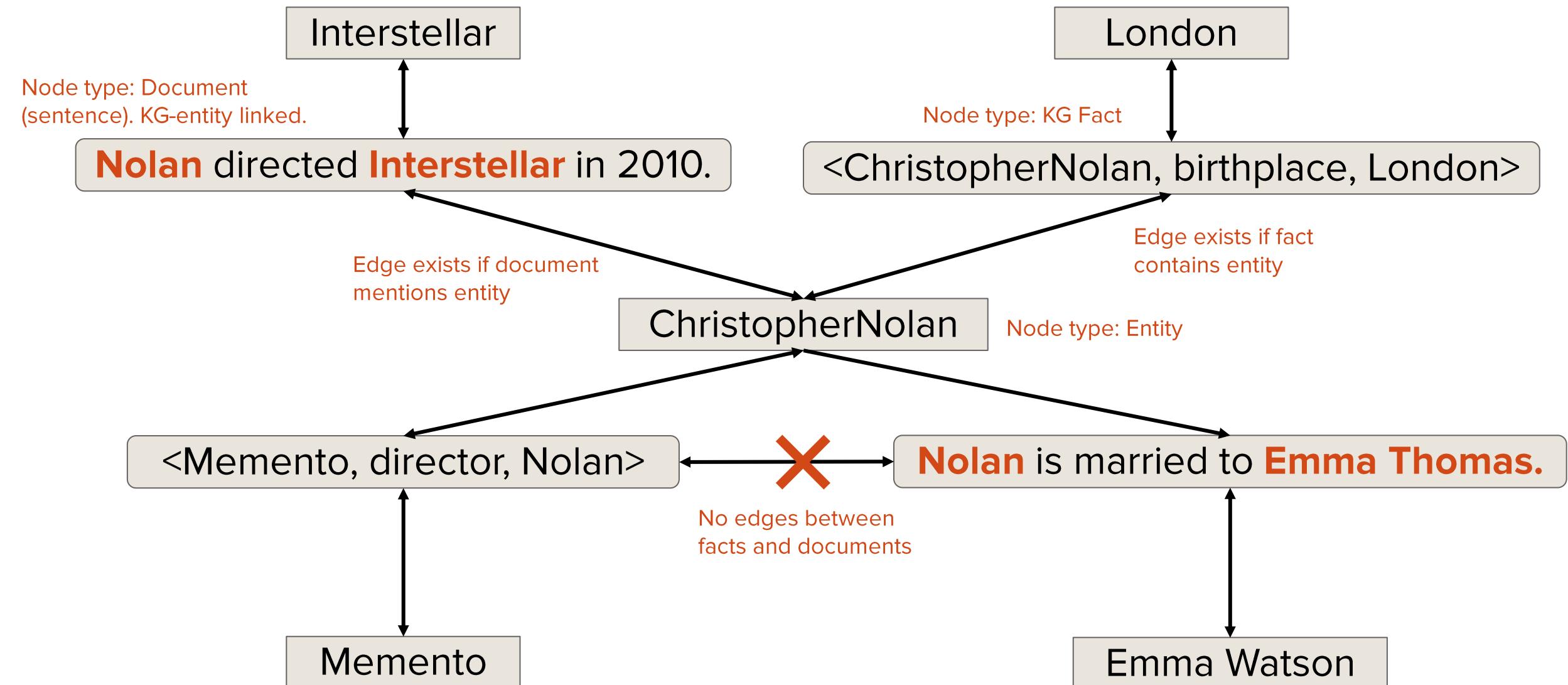
Entity-linked sentences

**Nolan** is married to **Emma Thomas**.

**Nolan** directed **Interstellar** in 2010.

**Guy Pearce** was in **Memento** and **Flynn**.

# PullNet: Graph model



**Question:** Who are the actors in movies directed by Nolan?

Sun et al., PullNet: Open Domain Question Answering with Iterative Retrieval on Knowledge Bases and Text, EMNLP 2019.

**Question:** Who are the actors in movies directed by Nolan?

NERD system

**ChristopherNolan**

Sun et al., PullNet: Open Domain Question Answering with Iterative Retrieval on Knowledge Bases and Text, EMNLP 2019.

**Question:** Who are the actors in movies directed by Nolan?

Node type: Document (sentence). KG-entity linked.

**Nolan** directed **Interstellar** in 2010.

Node type: Fact

<ChristopherNolan, birthplace, London>

Edge exists if document  
mentions entity

ChristopherNolan

Edge exists if fact  
contains entity

Node type: Entity

<Memento, director, Nolan>

**Nolan** is married to **Emma Thomas**.

## Early fusion

1. Pull sentences with linked entity from corpus (Lucene)
2. Pull facts of entity from the KG (using predicate similarity learned via LSTMs)

**Question:** Who are the actors in movies directed by Nolan?

**Nolan** directed **Interstellar** in 2010.

<ChristopherNolan, birthplace, London>

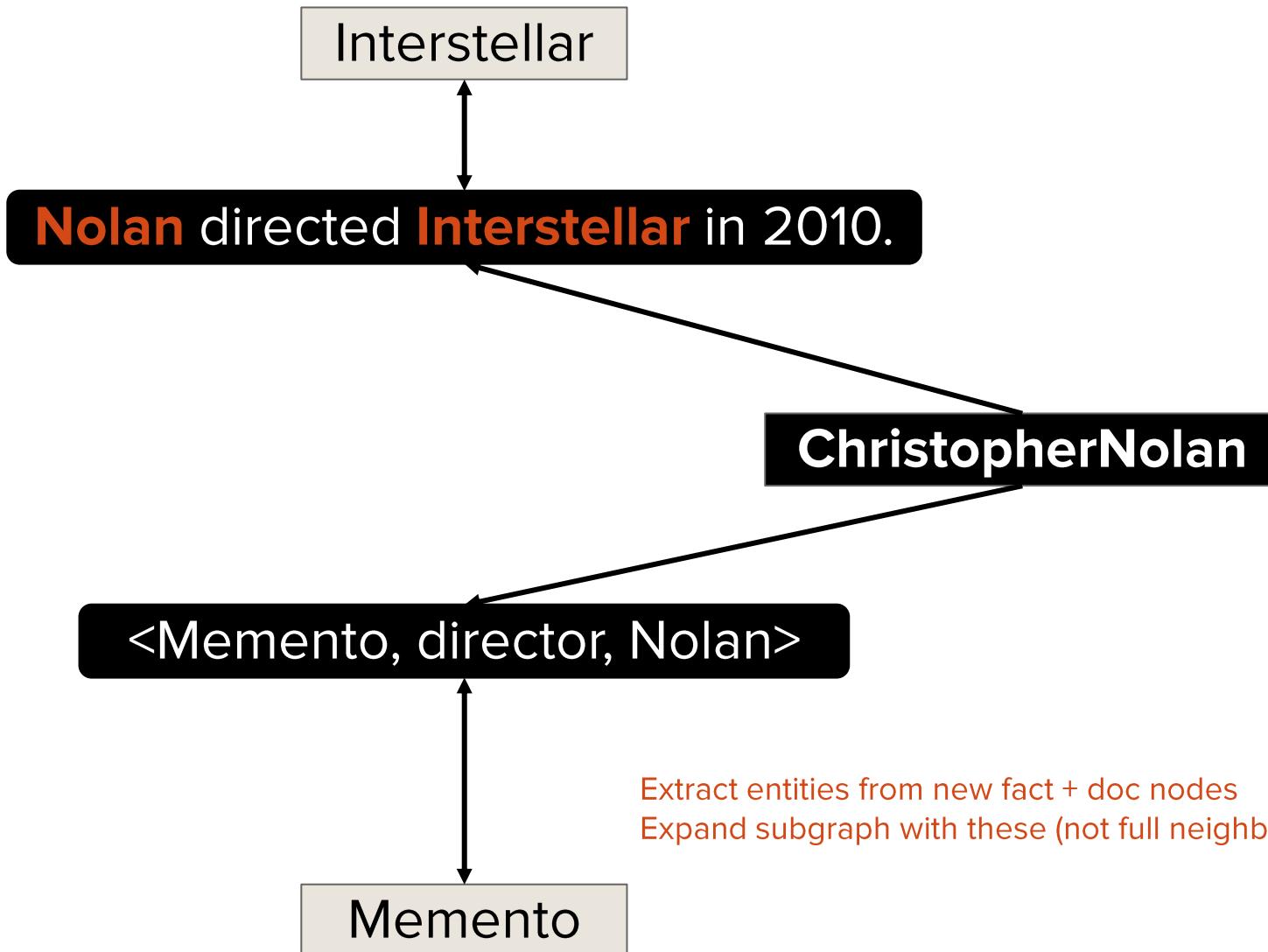
**ChristopherNolan**

<Memento, director, Nolan>

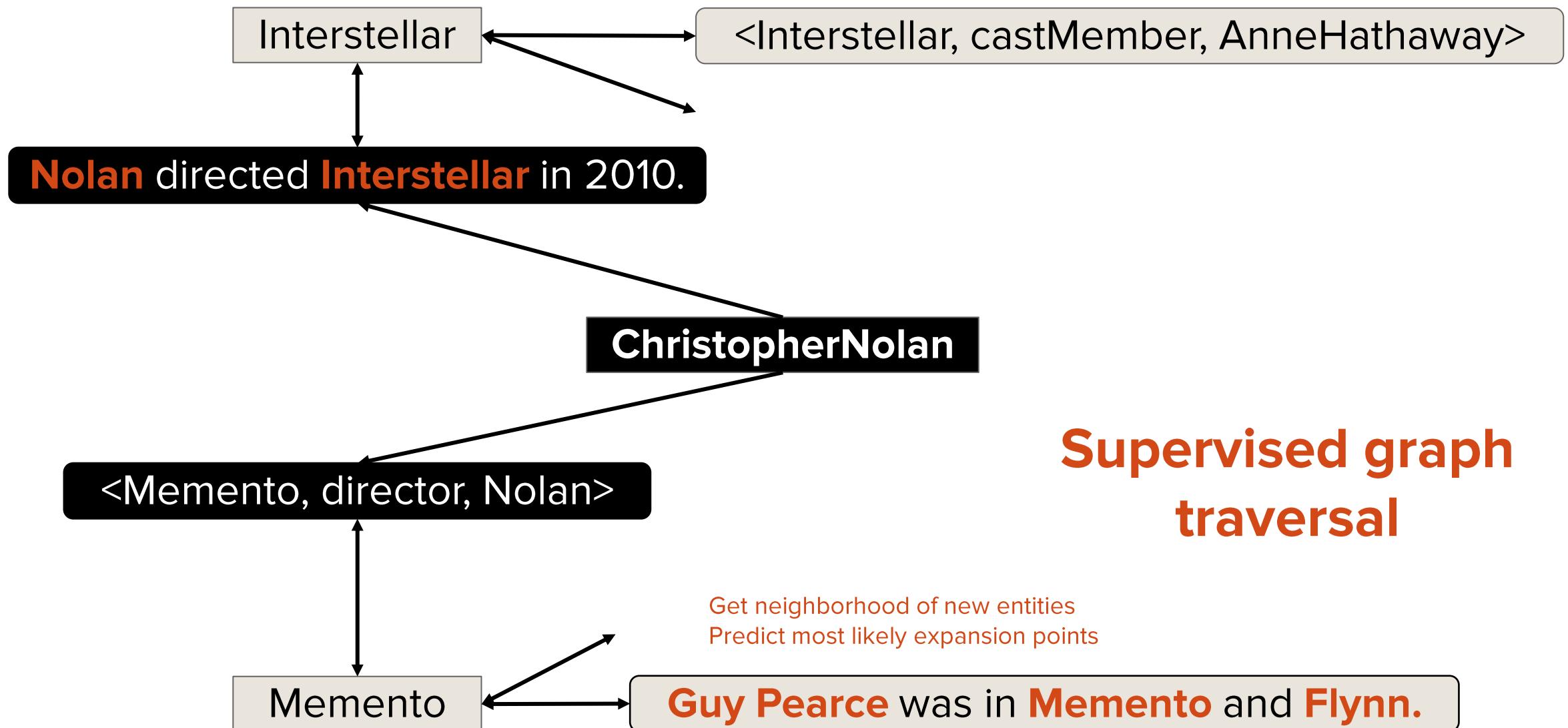
**Nolan** is married to **Emma Thomas**.

Predict most likely expansion points for next hop

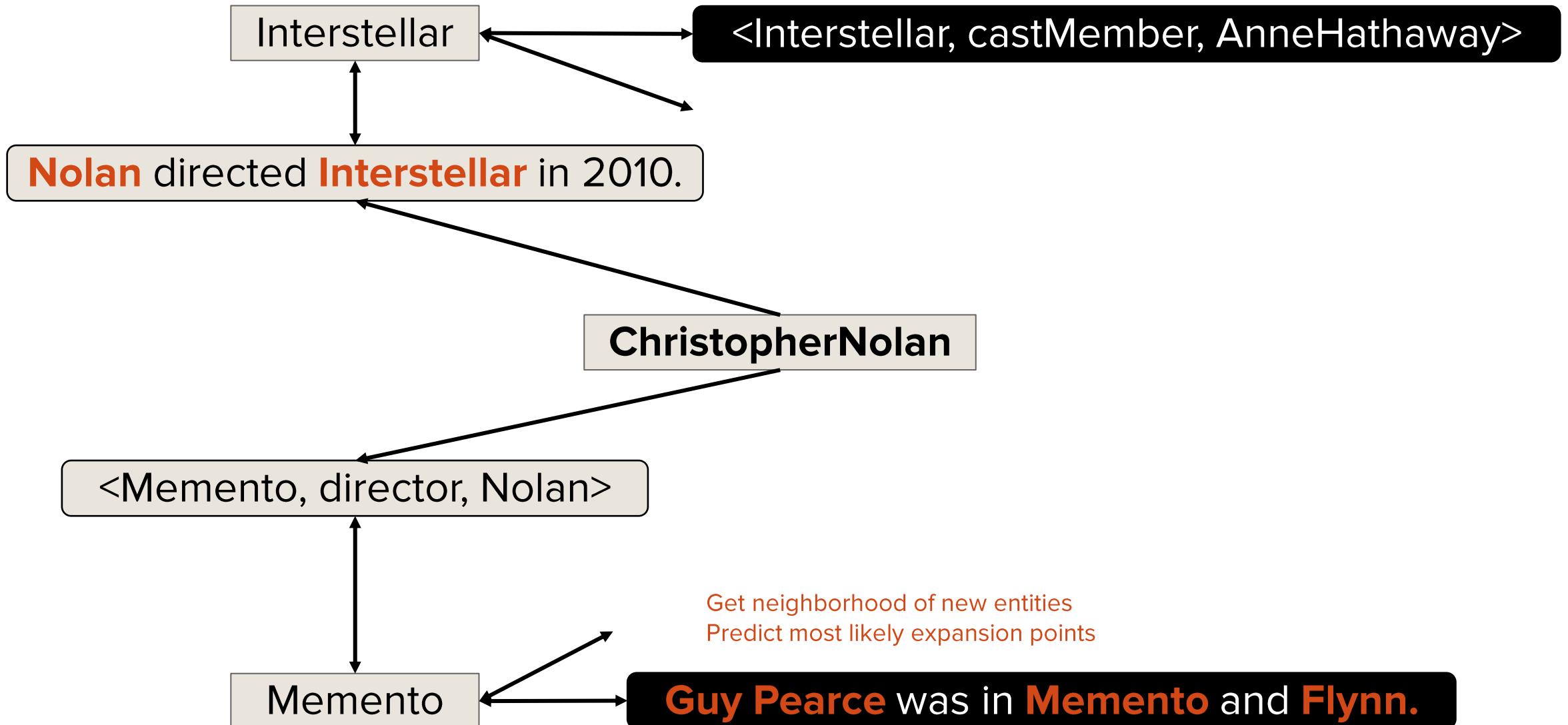
**Question:** Who are the actors in movies directed by Nolan?



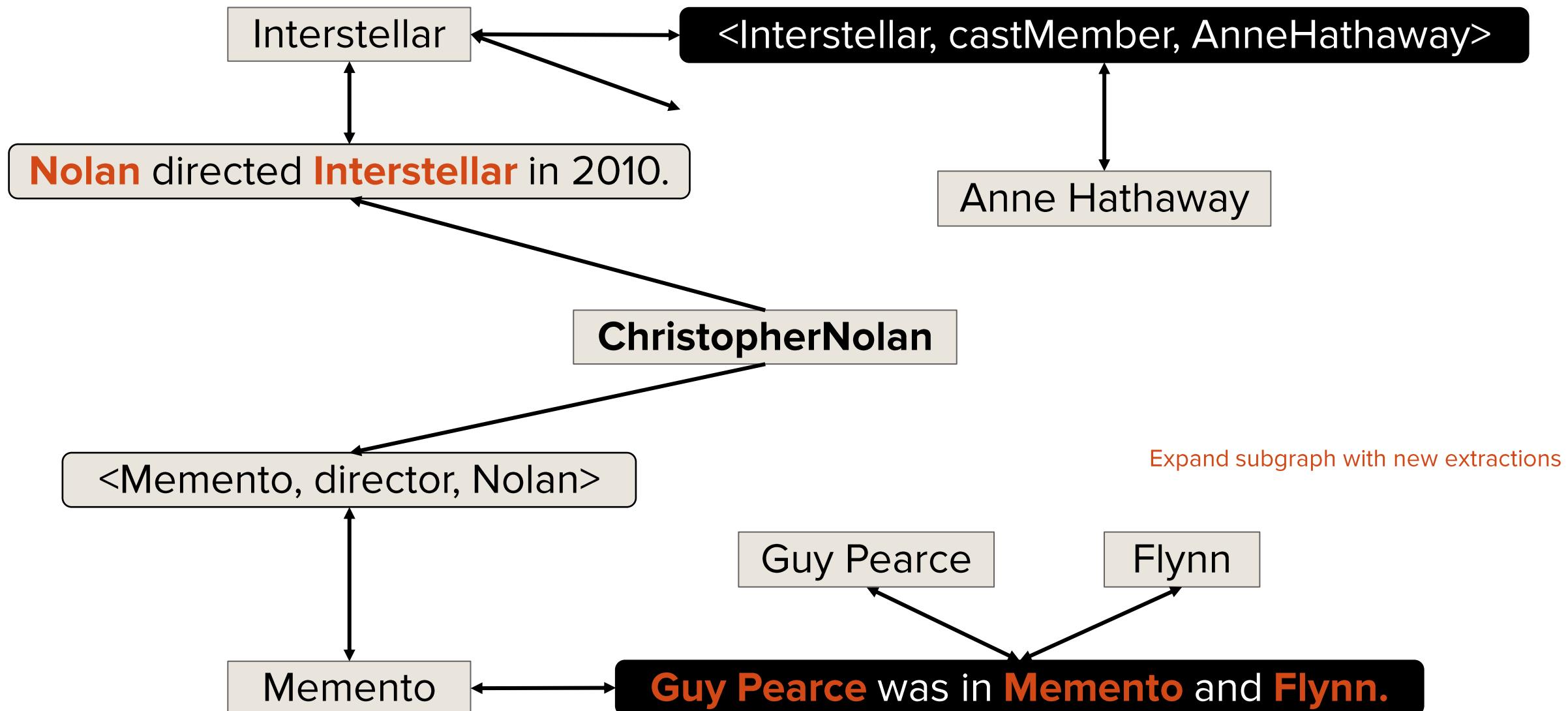
**Question:** Who are the actors in movies directed by Nolan?



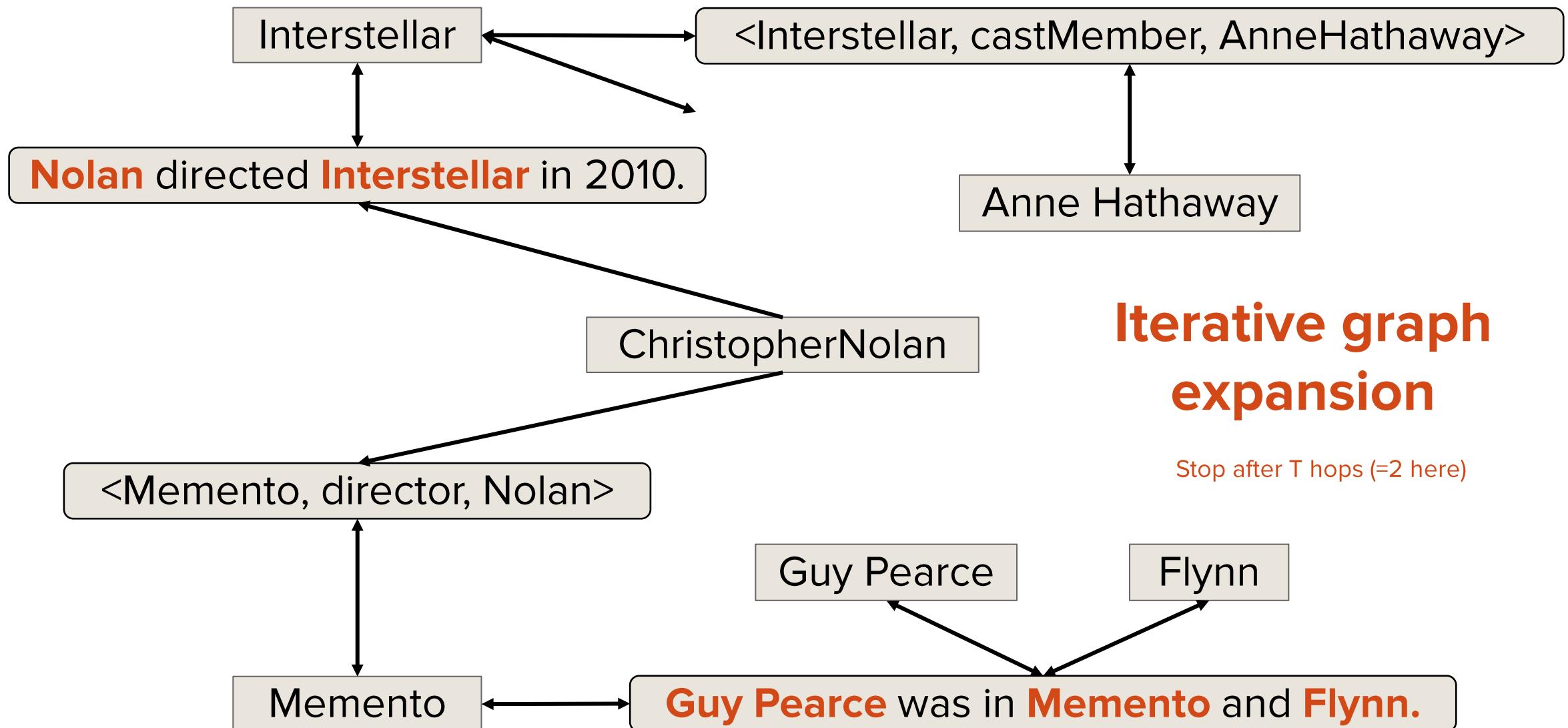
**Question:** Who are the actors in movies directed by Nolan?



**Question:** Who are the actors in movies directed by Nolan?



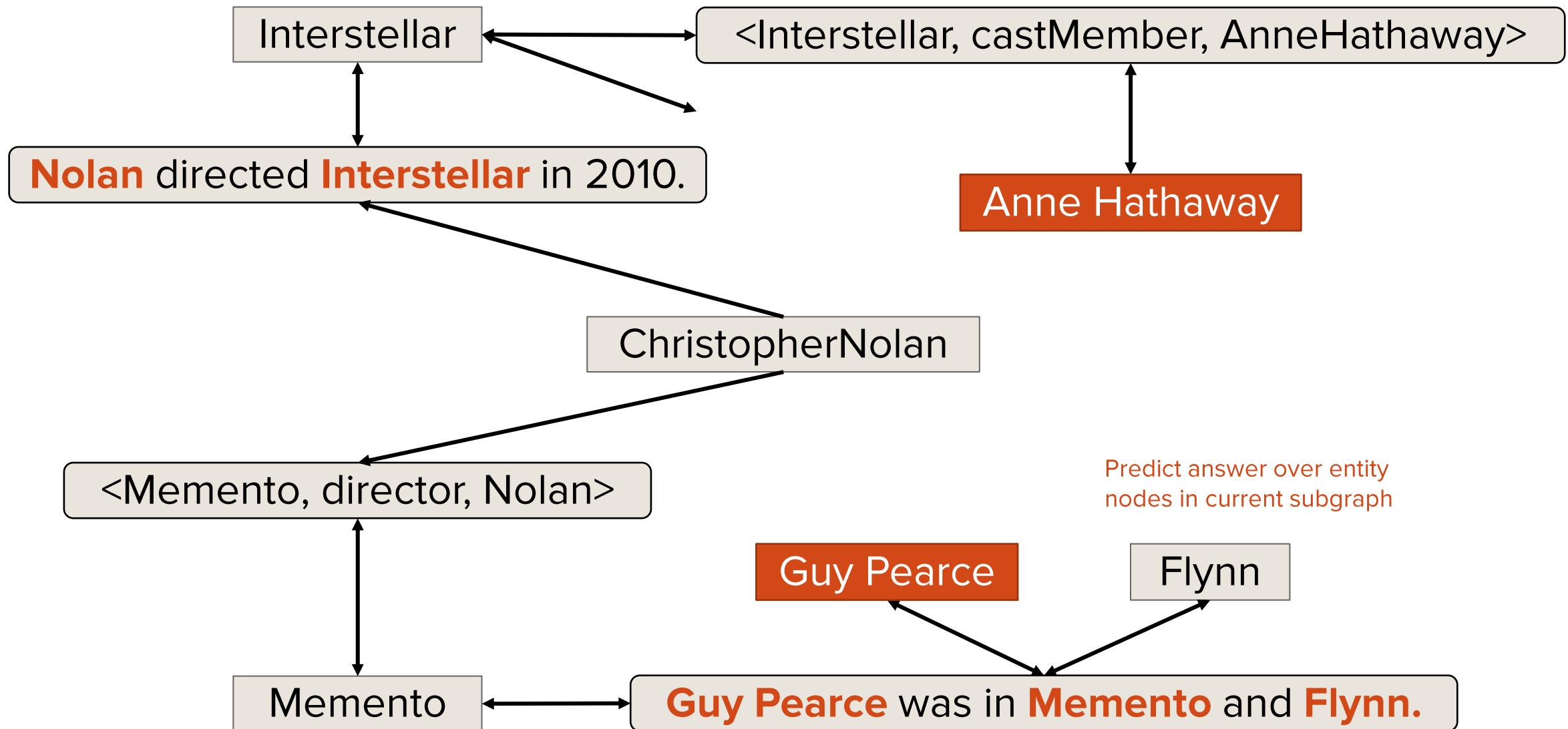
**Question:** Who are the actors in movies directed by Nolan?



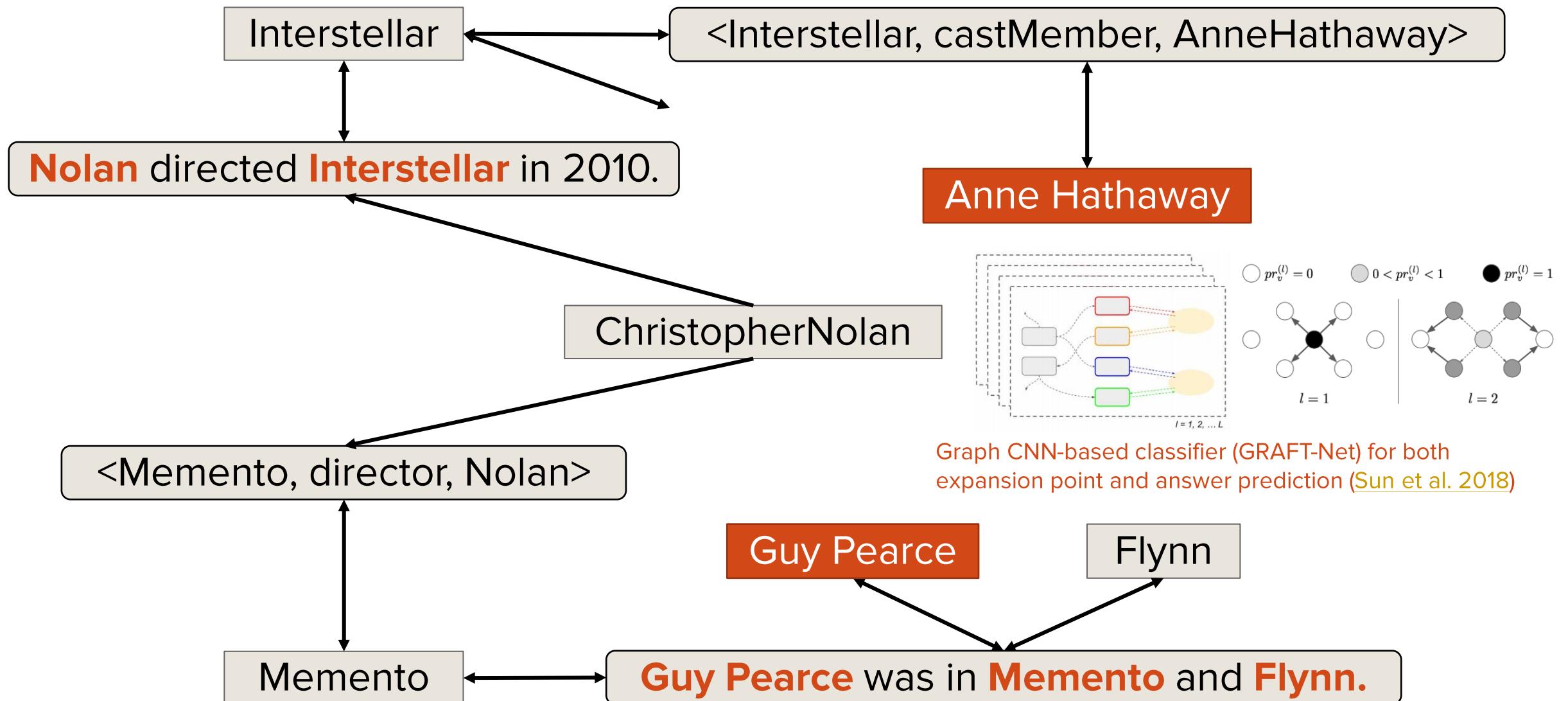
## Iterative graph expansion

Stop after T hops (=2 here)

**Question:** Who are the actors in movies directed by Nolan?



**Question:** Who are the actors in movies directed by Nolan?



# PullNet: Training

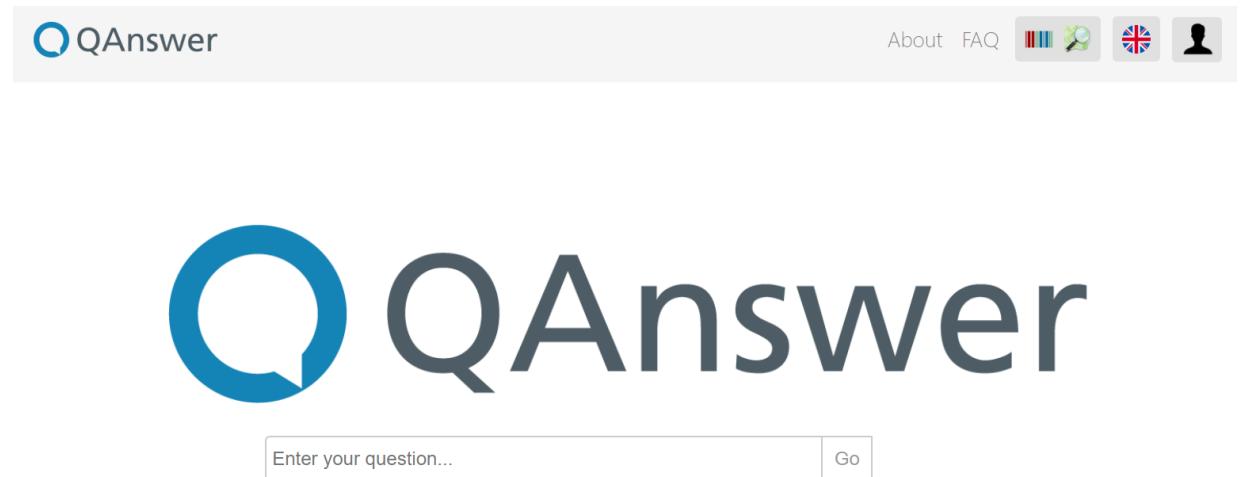
- Distant supervision with QA pairs
- Uses shortest paths between Q and A entities **in KG**
- Gold expansion points: Intermediate nodes on shortest paths
- Uses teacher forcing
- Gold answers: From benchmark

# Closely related to multi-hop KGR

- Multi-hop knowledge graph reasoning (KGR) and knowledge graph completion (KGC) closely associated with multi-hop QA
- Bridge between neural and symbolic space
- MINERVA ([Das et al. 2018](#)) [Reinforcement learning]
- SRN ([Qiu et al. 2020](#)) [Reinforcement learning]
- DrKIT ([Dhingra et al. 2020](#))
- Similar ideas explored for multi-hop MRC ([Asai et al. 2020](#))

# Heterogeneous QA: Unified resource

- **QAnswer** ([Diefenbach et al. 2019](#))
- Multiple KGs as unified triple store
- KG-agnostic approach for QA



Diefenbach et al., QAnswer: A Question answering prototype bridging the gap between a considerable part of the LOD cloud and end-users, WWW 2019.

# Structured querying over multiple KGs

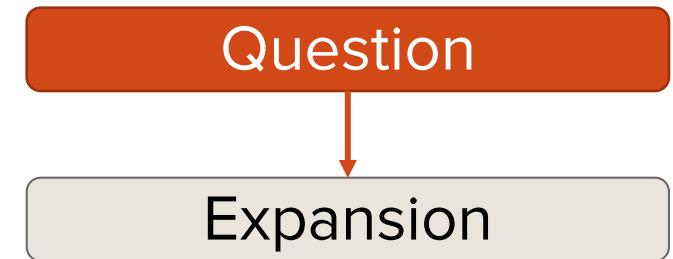
- QAnswer (Diefenbach et al. 2019)  
*Give me actors born in Berlin.*
- Multiple KGs as unified triple store
- KG-agnostic approach for QA

Question

Diefenbach et al., QAnswer: A Question answering prototype bridging the gap between a considerable part of the LOD cloud and end-users, WWW 2019.

# Expand question with candidate KG concepts

- QAnswer (Diefenbach et al. 2019)  
Give me actors born in Berlin.
- Multiple KGs as unified triple store  
 $R = \{\text{actor}, \text{TV actor}, \text{bornIn}, \text{Born}, \text{Berlin}, \text{Berlin Univ}, \text{West Berlin}\}$
- KG-agnostic approach for QA  
Lucene-based lookup



Diefenbach et al., QAnswer: A Question answering prototype bridging the gap between a considerable part of the LOD cloud and end-users, WWW 2019.

# Generation of SPARQL queries with candidates

- QAnswer (Diefenbach et al. 2019)
- Multiple KGs as unified triple store
- KG-agnostic approach for QA

Efficient construction of SPARQL queries using a BFS of depth 2 on the KG (**exhaustive but valid**)

Enabled by **HDT** + additional indexing of KG (distances between object pairs)

Give me actors born in Berlin.

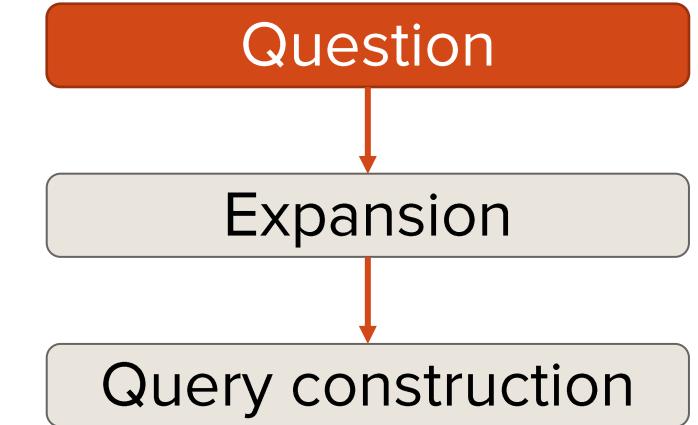
R = {actor, TVActor, bornIn, Born, Berlin, BerlinUniv, WestBerlin}

SELECT ?x  
WHERE {s1 s2 s3}

SELECT / ASK ?x  
WHERE {s1 s2 s3 . s4 s5 s6 .}

SELECT ?x  
WHERE { ?x bornIn Berlin .  
            ?x ?y actor }

SELECT ?x  
WHERE { ?x ?y BerlinUniv .  
            ?x ?y TVActor .}



# Structured querying over multiple KGs

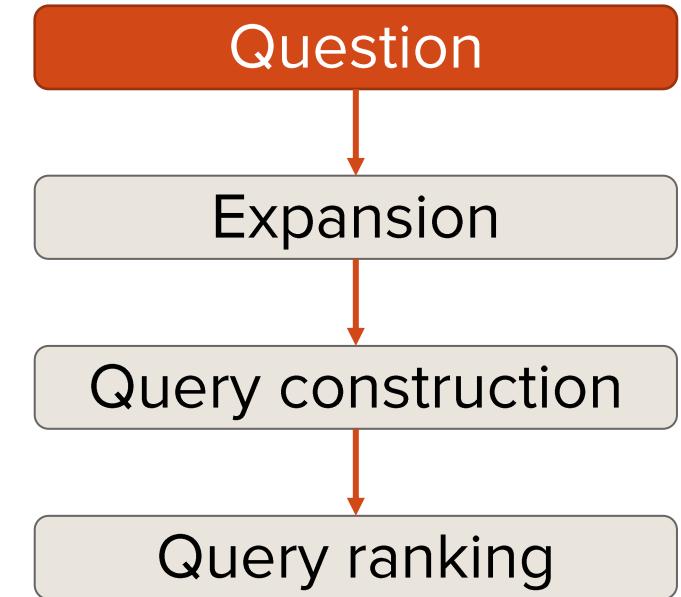
- QAnswer (Diefenbach et al. 2019)
- Multiple KGs as unified triple store
- KG-agnostic approach for QA

Give me actors born in Berlin.

R = {actor, TV actor, bornIn, Born,  
Berlin, Berlin Univ, West Berlin}

SELECT / ASK ?x  
WHERE {s1 s2 s3 . s4 s5 s6 .}

LTR (RankLib + coordinate ascent)



# Structured querying over multiple KGs

- QAnswer (Diefenbach et al. 2019)
- Multiple KGs as unified triple store
- KG-agnostic approach for QA

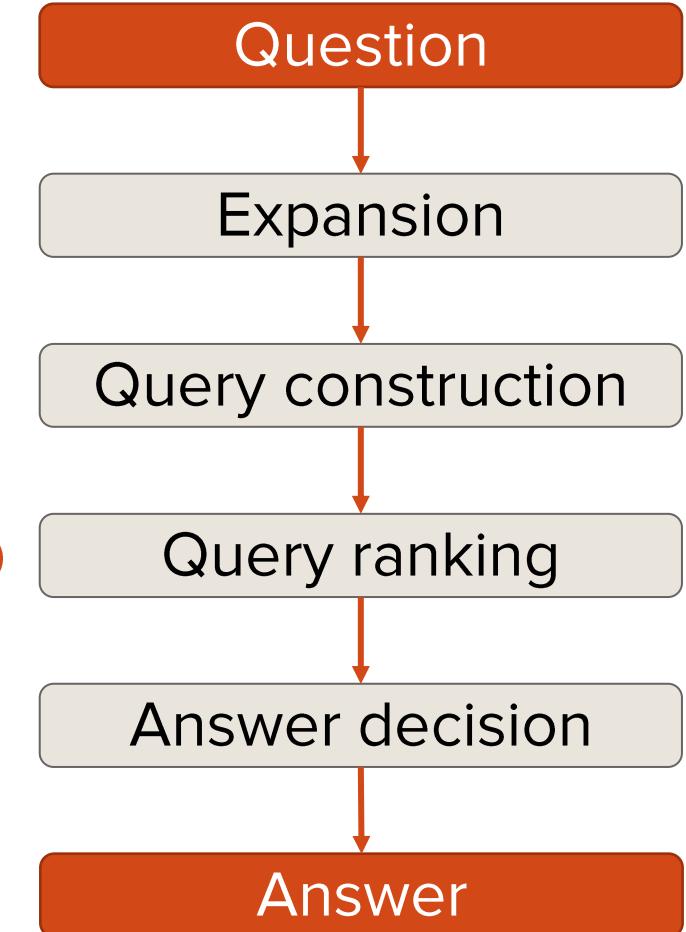
Give me actors born in Berlin.

R = {actor, TV actor, bornIn, Born,  
Berlin, Berlin Univ, West Berlin}

SELECT / ASK ?x  
WHERE {s1 s2 s3 . s4 s5 s6 .}

LTR (RankLib + coordinate ascent)

Top query score > threshold



# Structured querying over multiple KGs

- QAnswer (Diefenbach et al. 2019)
- Multiple KGs as unified triple store
- KG-agnostic approach for QA
- Extremely efficient due to HDT and additional KG indexing ☺
- Syntax agnostic ☹

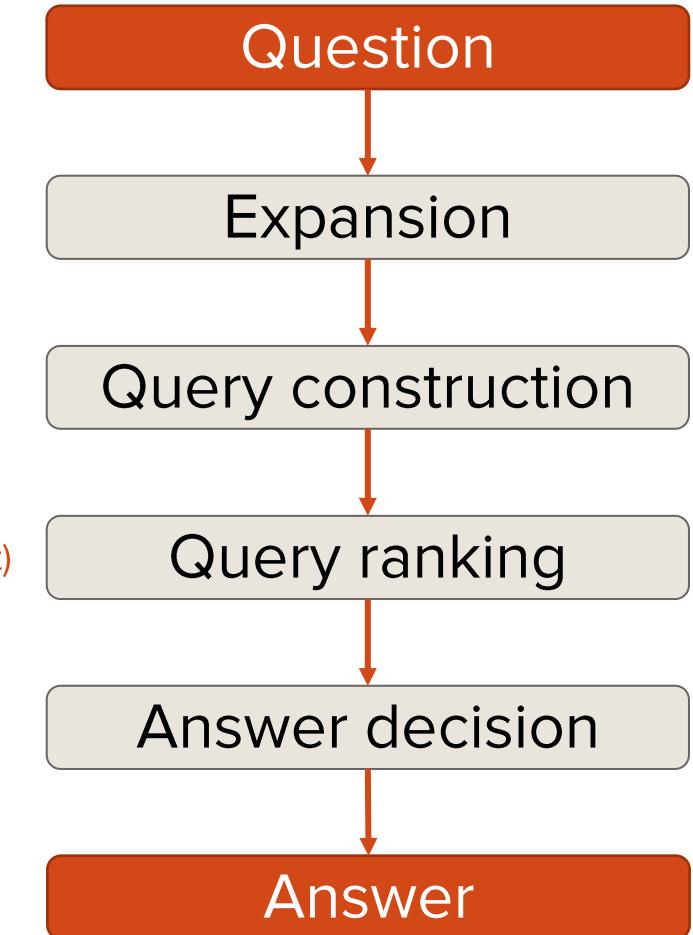
Give me actors born in Berlin.

$R = \{\text{actor, TV actor, bornIn, Born, Berlin, Berlin Univ, West Berlin}\}$

`SELECT / ASK ?x  
WHERE {s1 s2 s3 . s4 s5 s6 .}`

LTR (RankLib + coordinate ascent)

Top query score > threshold



# Heterogeneous QA: Wrap-up

- Early fusion and unified representation truer to spirit of heterogeneous QA than late fusion
- PullNet uses early fusion 😊 and deals with complex questions 😊
- But in principle only for chain joins 😞
- Efficiency is an open issue (too many predictions), no. of hops assumed to be known 😟
- QAnswer is efficient 😊 and works over multiple KGs (largely unexplored) 😊
- But works mostly for relatively simple questions 😞
- Current systems still **not truly unified**: reliance on **KG entities** for linking and distant supervision, and a **triplified view** of knowledge

break duration ?x .  
?x measuredIn minutes .

# Outline: QA over knowledge graphs

- **Background:** Setup, benchmarks, metrics
- **Simple QA:** Templates and embeddings
- **Complex QA:** Multiple entities and predicates
- **Heterogeneous sources:** Handling KG and text
- **Conversational QA:** Implicit context in multi-turn setup
- **Take-home:** Summary and insights

How can we deal with information needs spread  
over multi-turn conversations?



# Conversational KG-QA

Mia Farrow

Which actor voiced the character Unicorn in The Last Unicorn?

Schmendrick

Which role was voiced by Alan Arkin in the Last Unicorn?

America

Who performed the songs in the movie The Last Unicorn?

Folk rock

What is the genre of the band that performed the songs in The Last Unicorn?

Jules Bass

Who was the director of the movie The Last Unicorn?



# Conversational KG-QA

Mia Farrow

Which actor voiced the character Unicorn in The Last Unicorn?

Schmendrick

And Alan Arkin was behind ...?

America

The songs were by...?

Folk rock

Genre of this band?

Jules Bass

By the way, who directed the movie?

# Conversational KG-QA

- Information needs **rarely one-off**
- Sequence of **follow-up questions** on a topic
- Analogous to **search sessions** and **interactive retrieval**
- Users want to simulate **natural experience** with assistant
- **Leave context unspecified** in follow-ups

# Conversational KG-QA

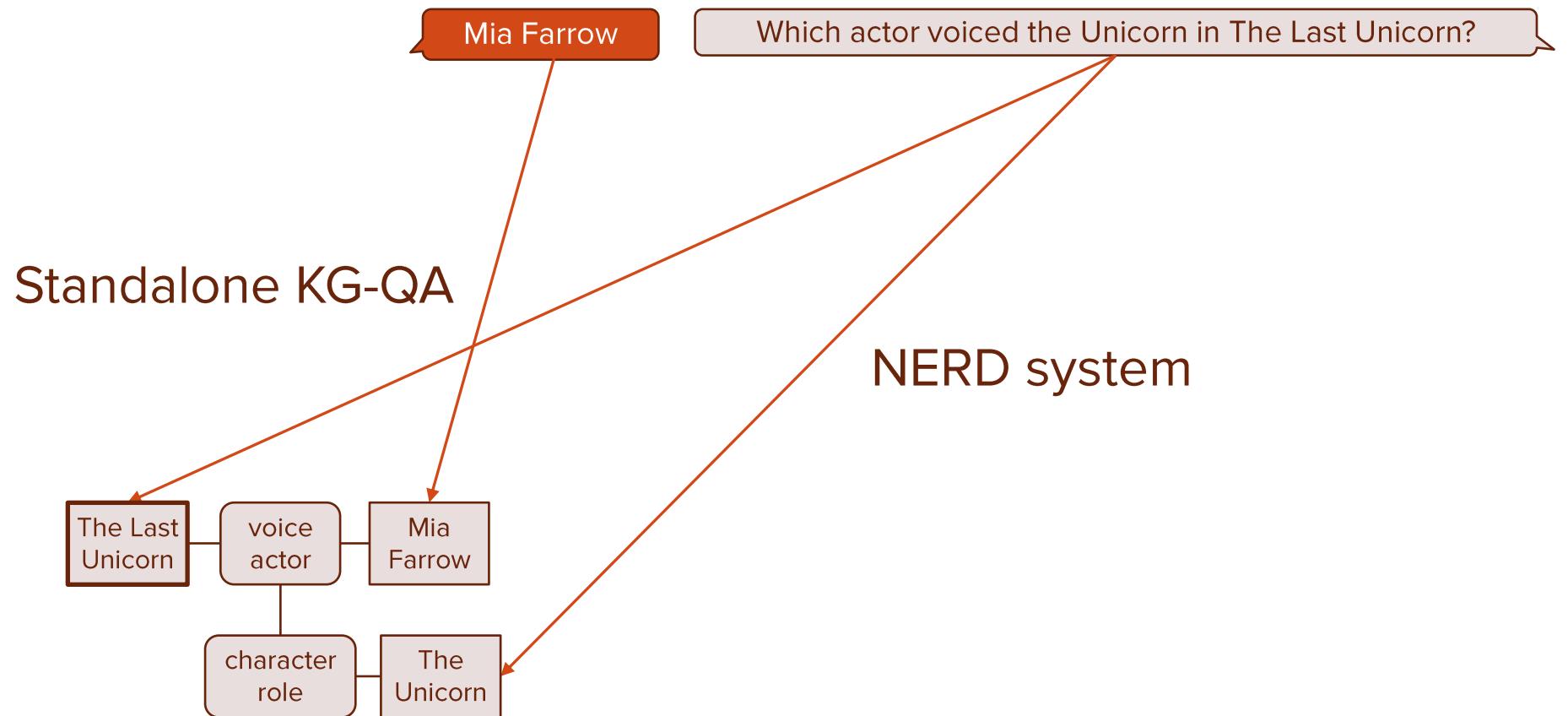
- **Key challenges** in conversational (KG-)QA
  - Infer implicit context
  - Handle ad hoc formulations
- Initially explored over **small tables** as sequential QA ([Iyyer et al. 2017](#))
- Key direction for KG-QA now ([Saha et al. 2018](#), [Guo et al. 2018](#),  
[\*\*Christmann et al. 2019\*\*](#), [\*\*Shen et al. 2019\*\*](#))

# Conversational QA: Graph traversal

- The **CONVEX** system ([Christmann et al. 2019](#))
- Large topic jumps in conversations are rare: establish **localized KG context**
- Harness **KG-connectivity**: No need to complete/rewrite questions
- **Expand context judiciously** with relevant entities and predicates in neighborhood
- **Unsupervised** iterative graph traversal (c.f. supervised graph traversal in PullNet)
- CONVEX works on top of **any KG-QA system** to handle conversations

Christmann et al., Look before you Hop: Conversational Question Answering over Knowledge Graphs Using Judicious Context Expansion, CIKM 2019.

# Initial context

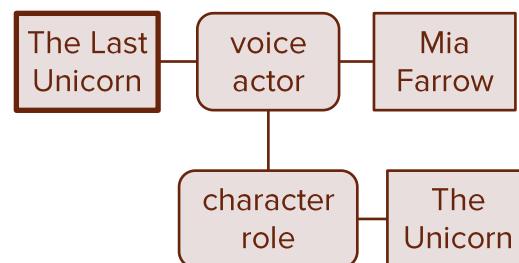


# Initial context

Mia Farrow

Which actor voiced the Unicorn in The Last Unicorn?

And Alan Arkin was behind . . .?



How to expand the context?

**Neighborhood of  
Mia Farrow**

**Neighborhood of  
The Last Un'**

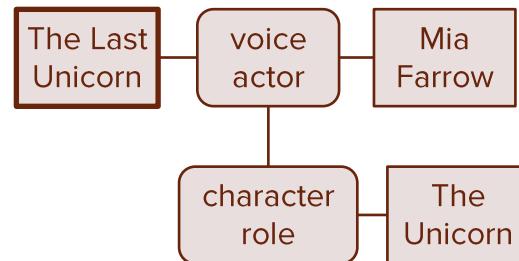
**Neighborhood of  
The Unicorn**

# Judicious context expansion

Mia Farrow

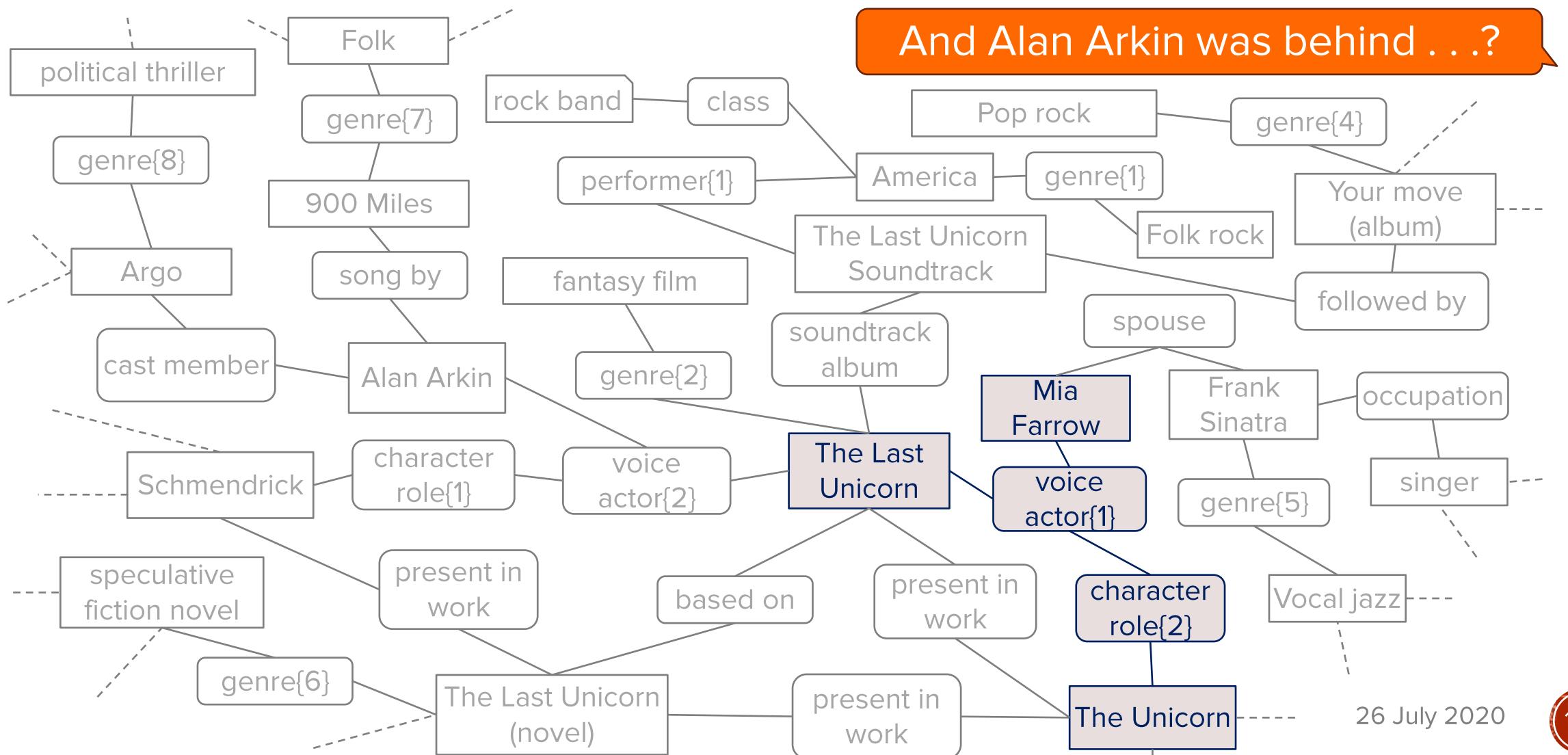
Which actor voiced the Unicorn in The Last Unicorn?

And Alan Arkin was behind . . . ?

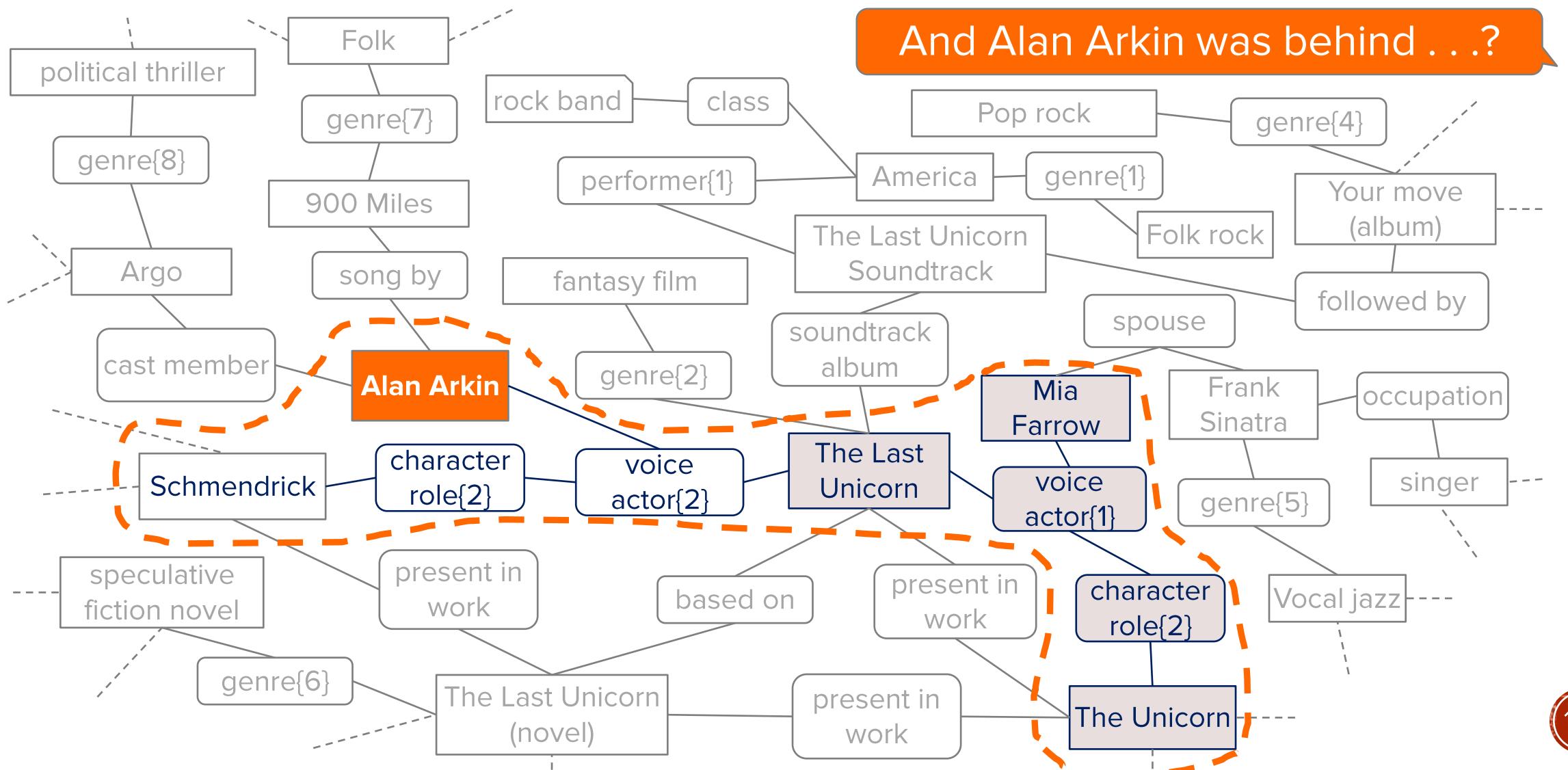


Do not expand with the complete neighborhood!

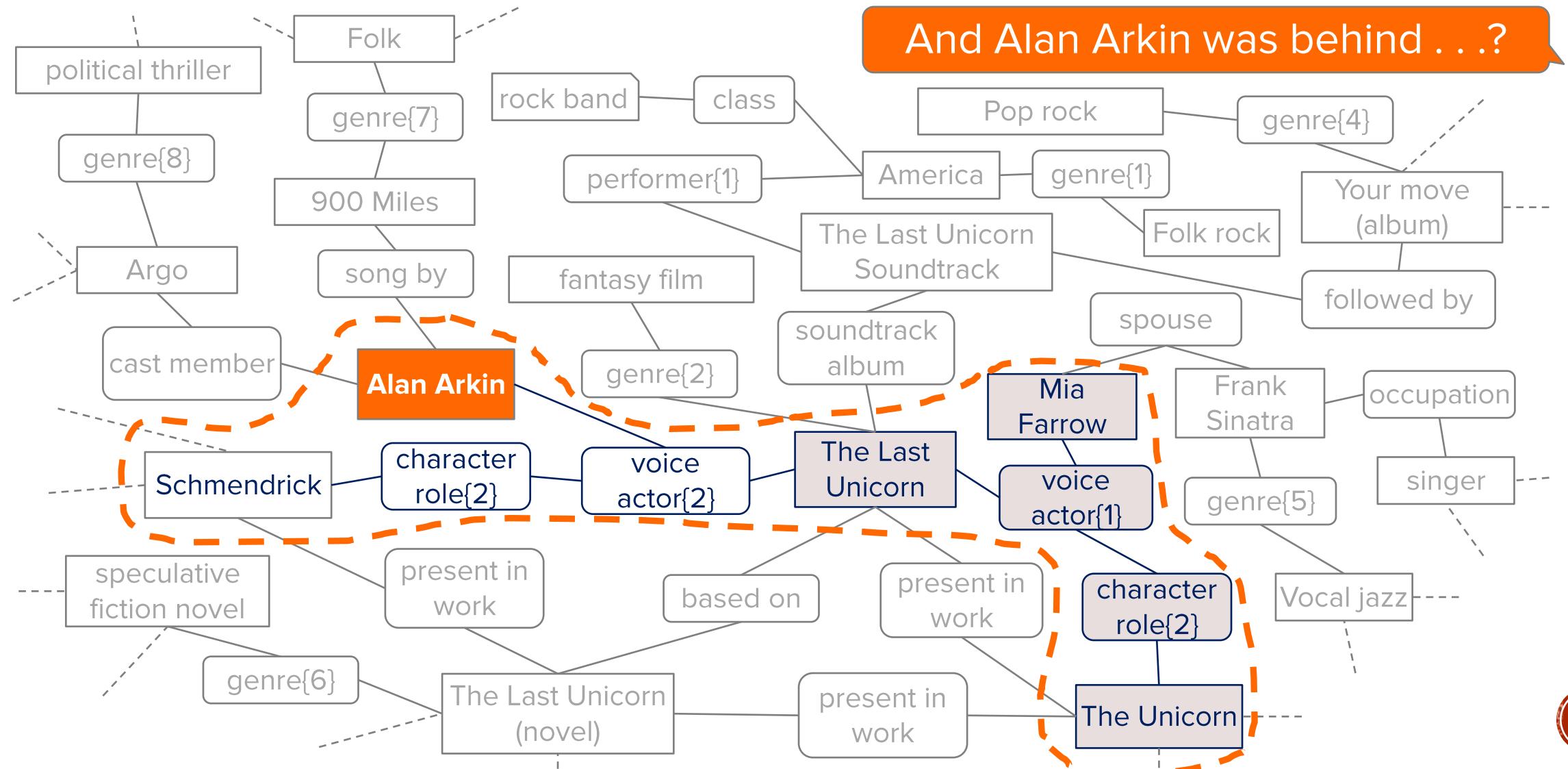
# Exploring context neighborhood



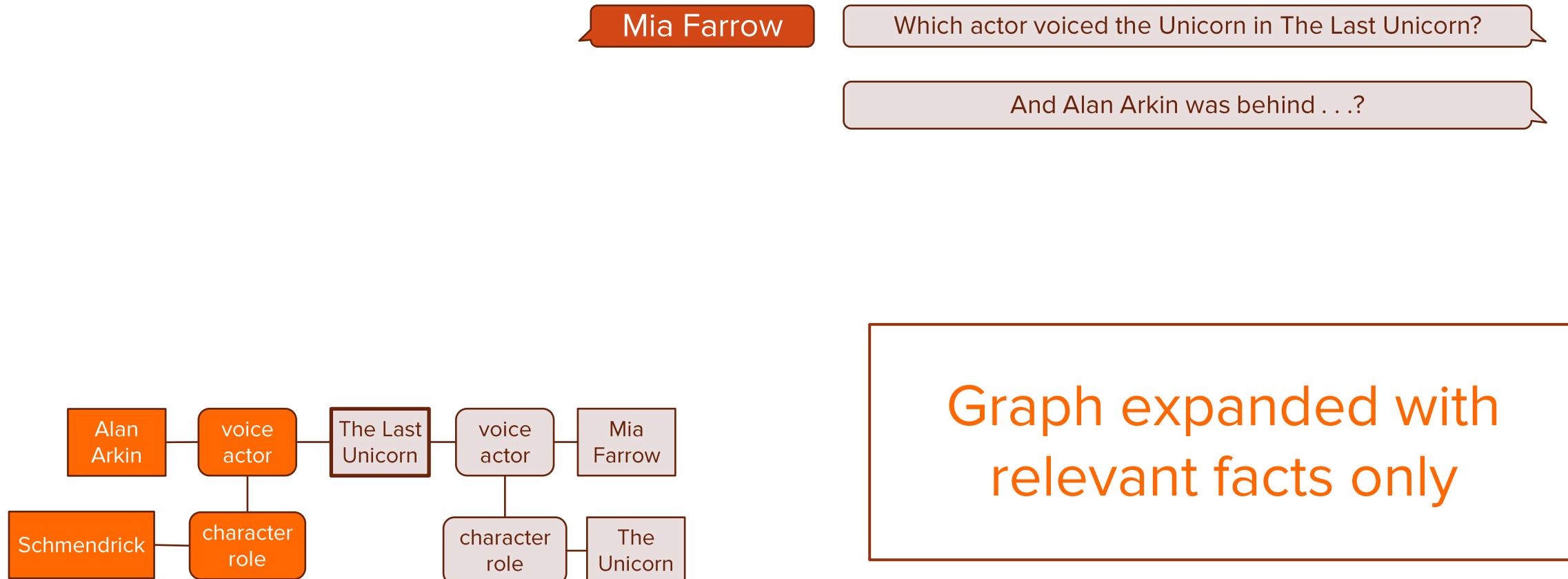
# Exploring context neighborhood



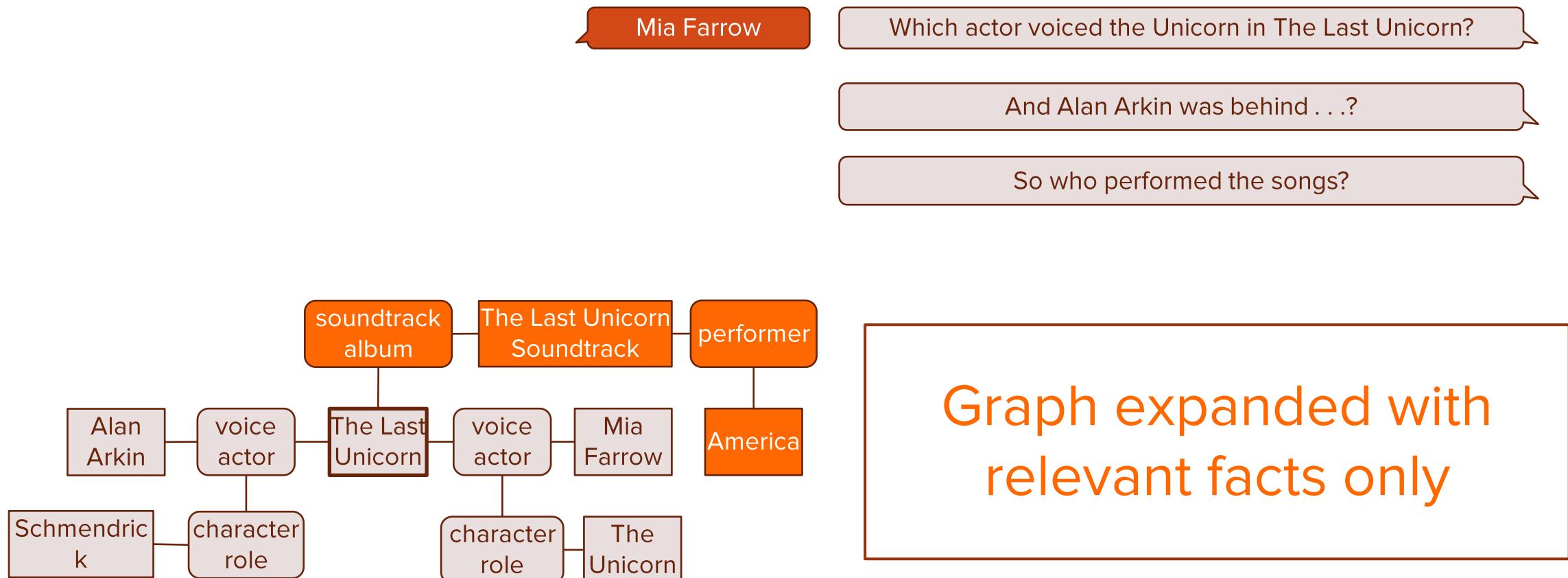
# Find frontier nodes to define expansion border



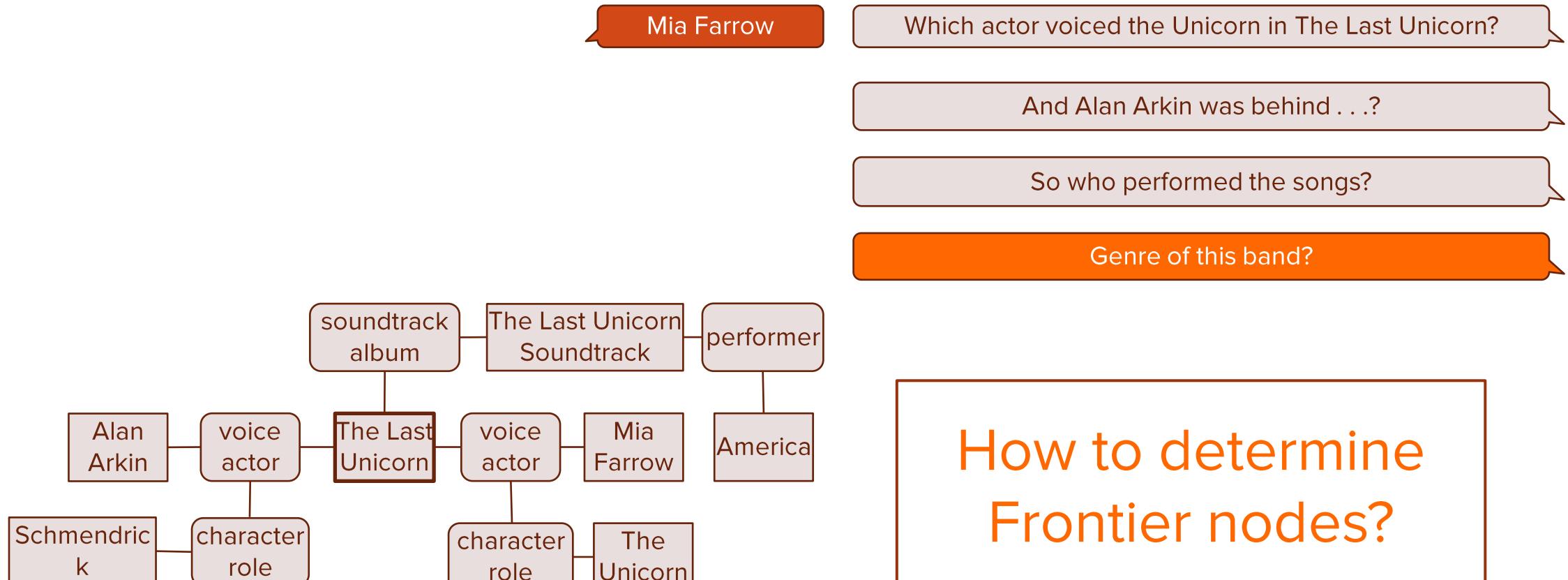
# Context graph



# Context graph



# Context graph



# Frontier score

**Matching similarity**

*match (candidate c)*

**Context relevance**

*prox (candidate c)*

**KG priors**

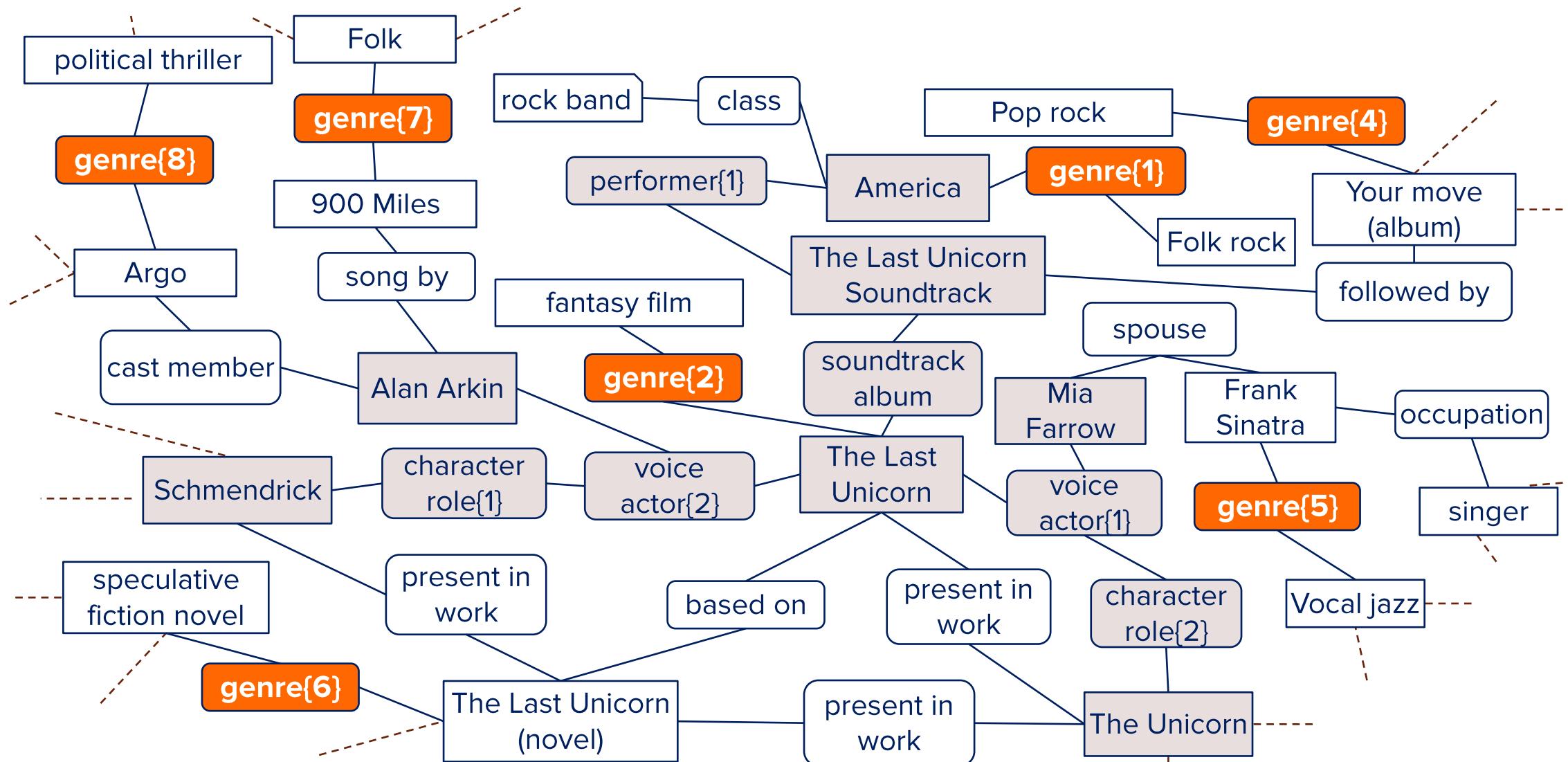
*prior (candidate c)*

$$\text{frontier\_score}(\text{candidate } c) = h_1 \cdot \text{match}(c) + h_2 \cdot \text{prox}(c) + h_3 \cdot \text{prior}(c)$$

With hyperparameters  $h_1, h_2, h_3$

# The great disambiguation

Genre of this band?



# Frontier nodes

## Matching similarity

<i>Candidate</i>	<i>Match</i>
genre{1}	1.00
genre{2}	1.00
...	...
folk rock band	0.89
RSH-Gold for Cult Band	0.87
fantasy film	0.36
...	...

## Context relevance

<i>Candidate</i>	<i>Prox</i>
genre{1}	0.91
folk rock band	0.86
RSH-Gold for Cult Band	0.86
...	...
genre{2}	0.34
fantasy film	0.36
...	...

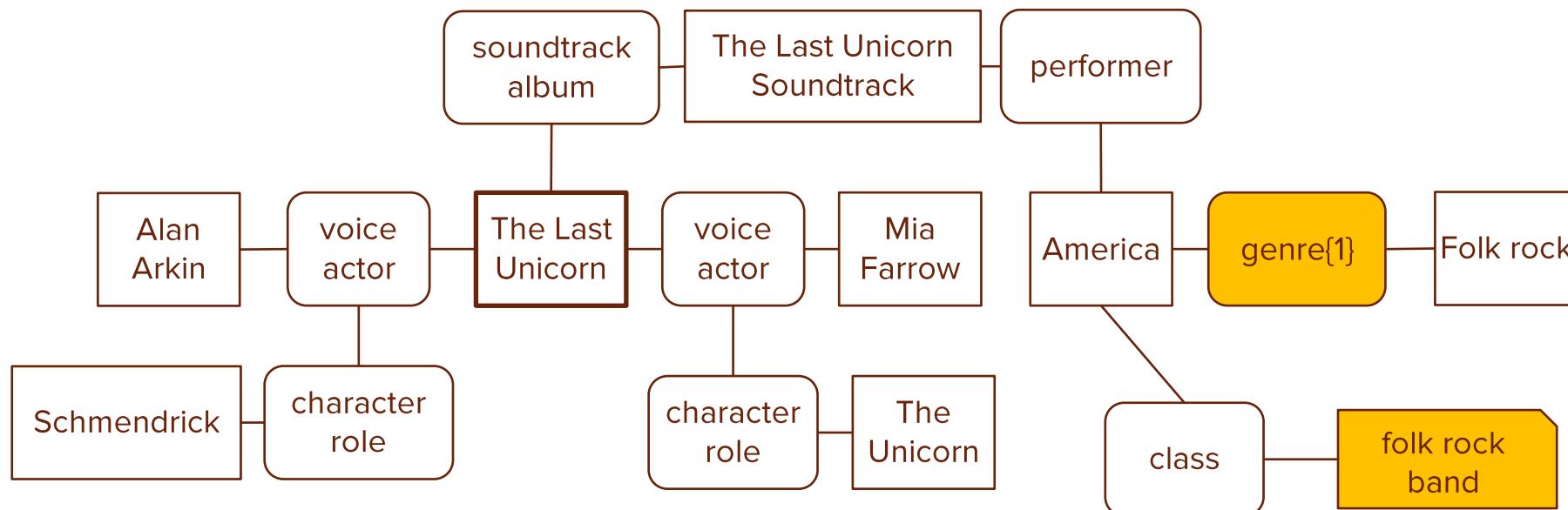
## KG priors

<i>Candidate</i>	<i>KG priors</i>
...	...
genre{1}	0.56
genre{2}	0.56
...	...
folk rock band	0.34
...	...
RSH-Gold for Cult Band	0.01

Fagin's Threshold Algorithm (FTA) to retrieve top- $k$  ranked nodes according to frontier score

# Frontier nodes

Genre of this band?



Frontier nodes

# Answer to the question

Genre of this band?

- Distance to Frontier nodes
  - Weighted by the frontier score
  - $distance\_F$

=> Explicit part
- Distance to all nodes in context graph X
  - Weighted by the turn they occurred in
  - $distance\_X$

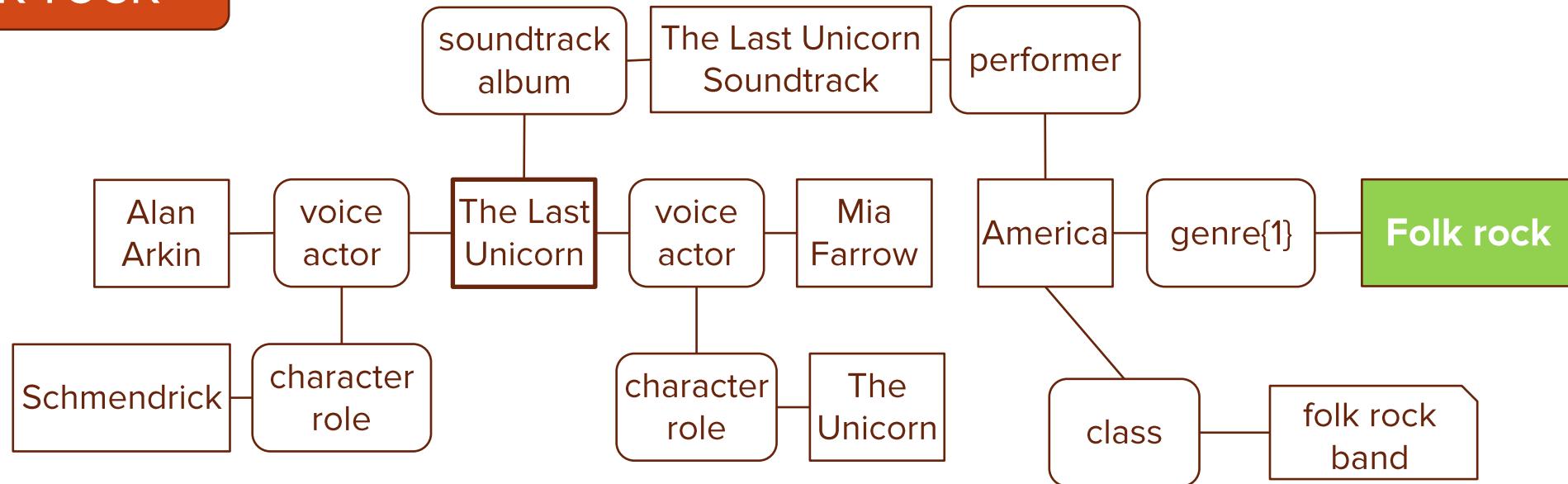
=> Implicit part

$$answer\_score(candidate\ c) = h_4 \cdot distance\_F + h_5 \cdot distance\_X$$

# Answer detection

Genre of this band?

Folk rock



Top-ranked node according to *answer\_score*

# Conversational QA: Sequence-to-sequence modeling

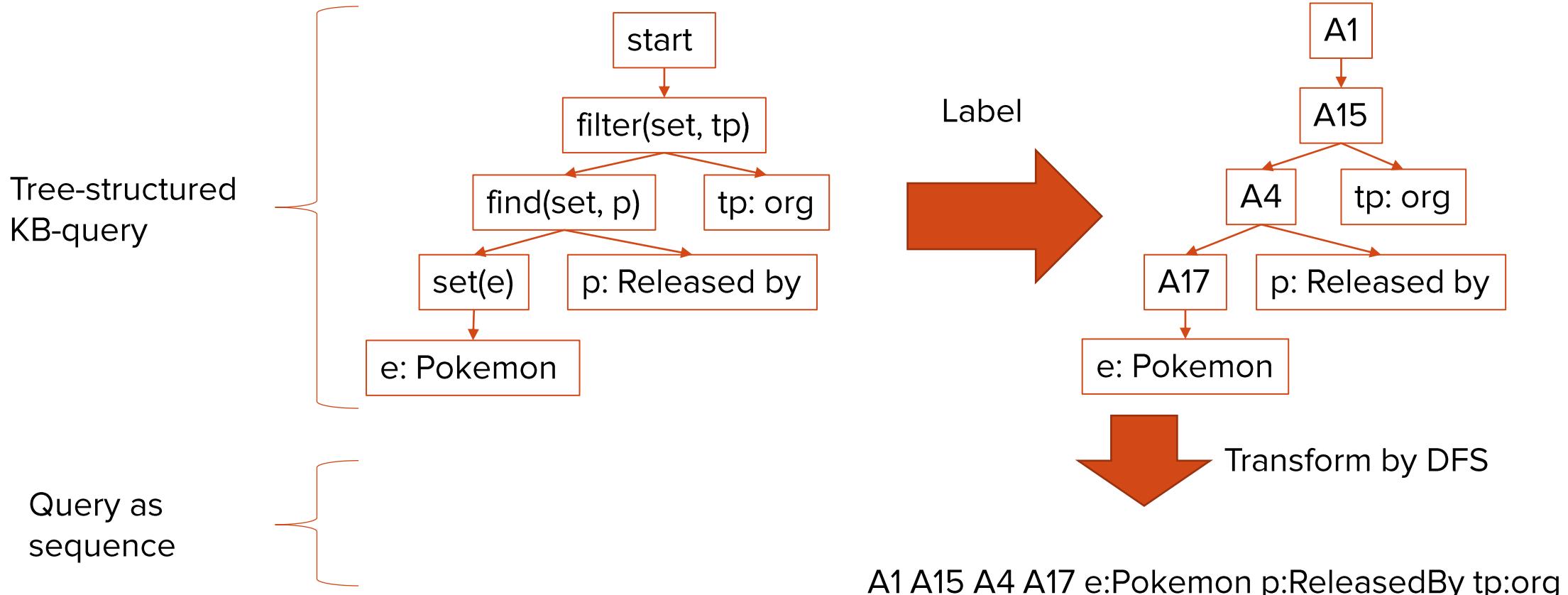
- The MaSP model (Shen et al. EMNLP 2019)
- Shared supervision for tasks: Entity detection and answering
- Grammar-based semantic parsing model
- Designed to resolve coreference in conversations
- Type-aware entity detection
- Uses transformers for sequence encoding

Shen et al., Multi-Task Learning for Conversational Question Answering over a Large-Scale Knowledge Base, EMNLP 2019.

# Sequence-to-sequence model

NL question as sequence

**Question:** Who released Pokemon?



Shen et al., Multi-Task Learning for Conversational Question Answering over a Large-Scale Knowledge Base, EMNLP 2019.

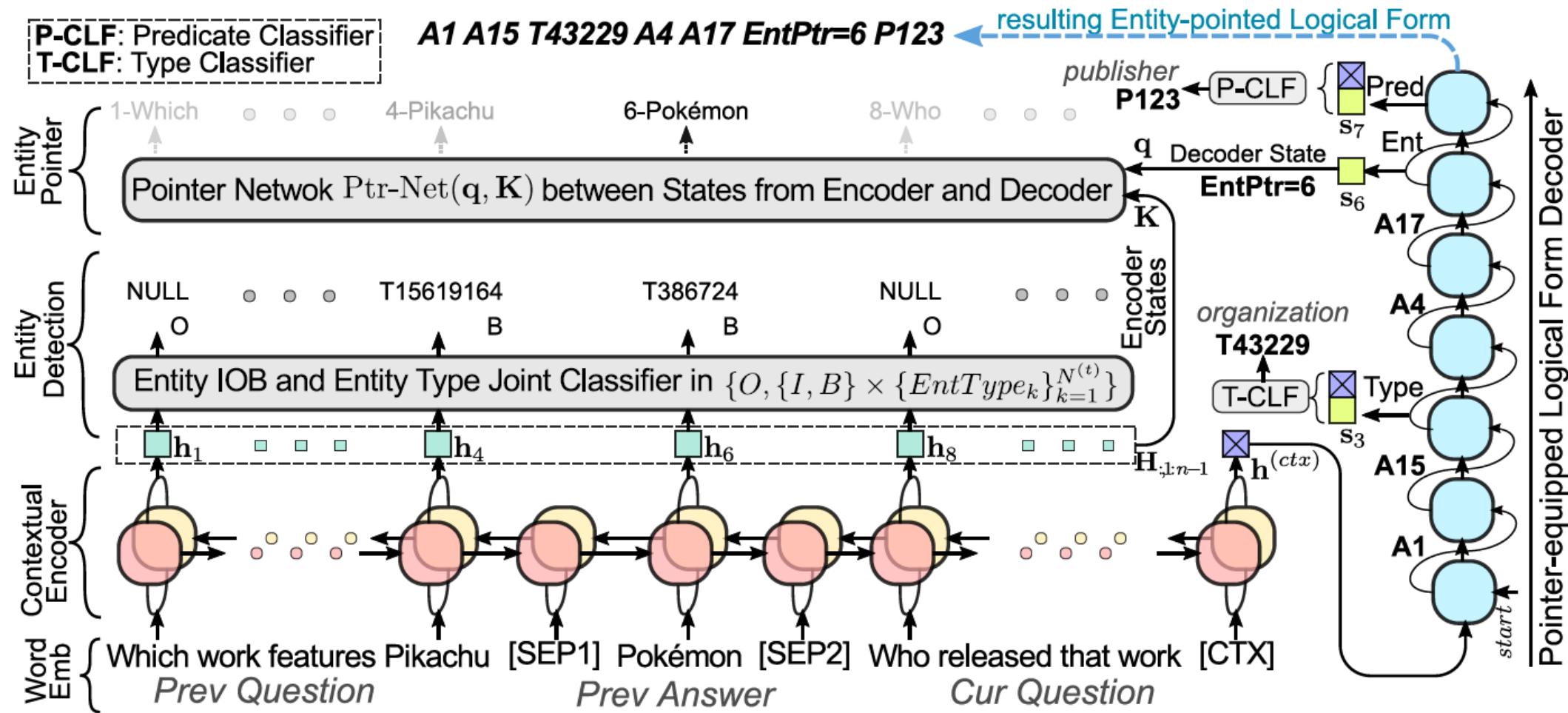
# Sequence-to-sequence model

NL question as sequence

**Question: [CONTEXT]** Who released that work?

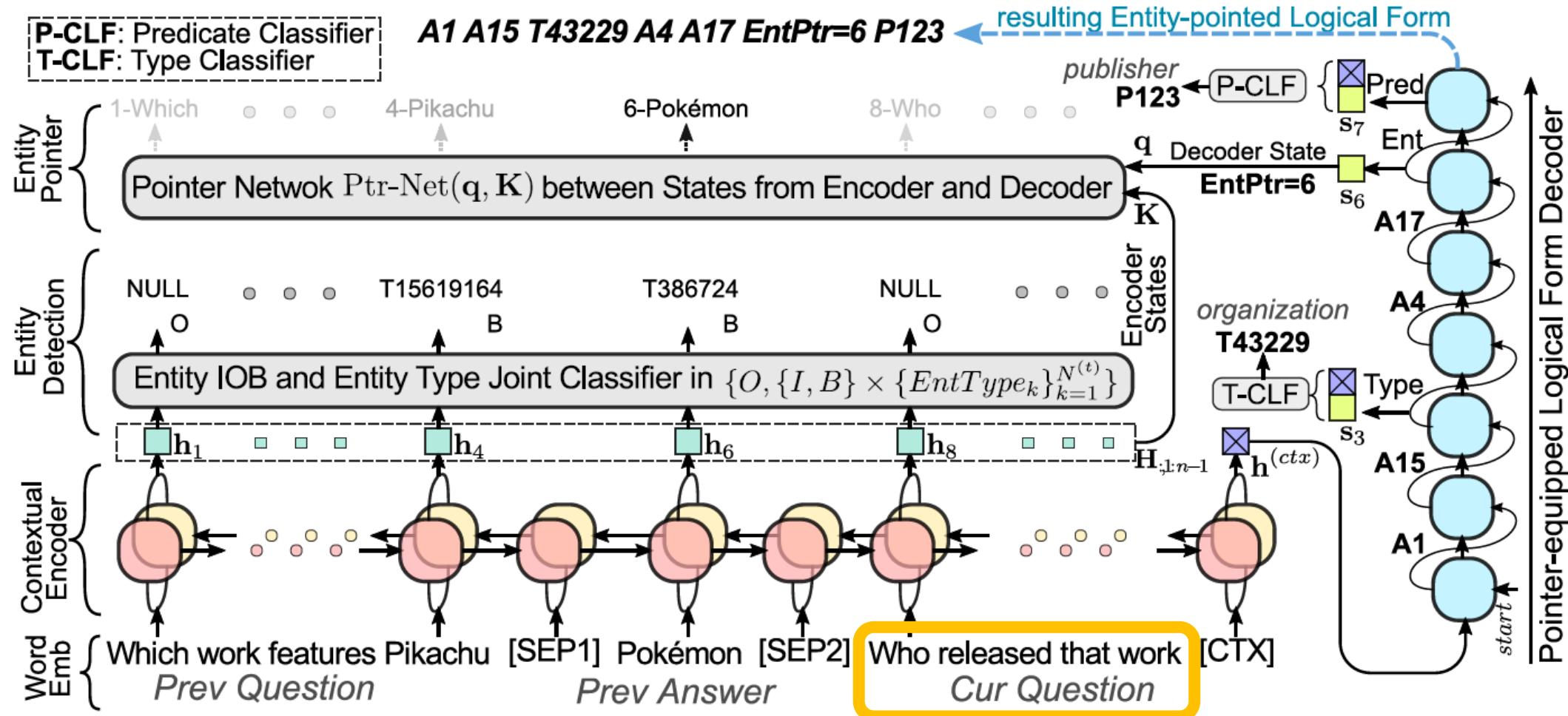
Shen et al., Multi-Task Learning for Conversational Question Answering over a Large-Scale Knowledge Base, EMNLP 2019.

# Seq-to-seq: The MaSP model



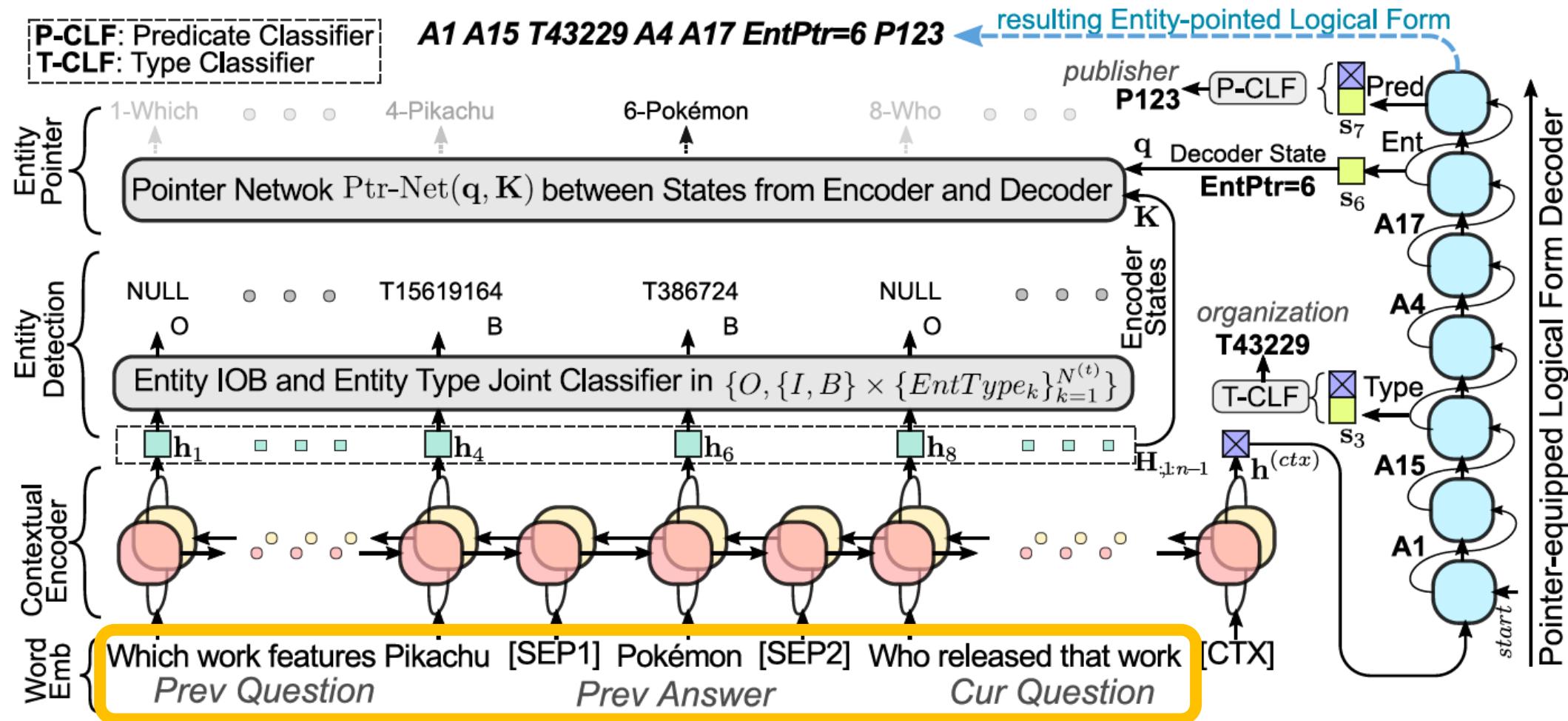
Shen et al., Multi-Task Learning for Conversational Question Answering over a Large-Scale Knowledge Base, EMNLP 2019.

# MaSP: Step-by-step



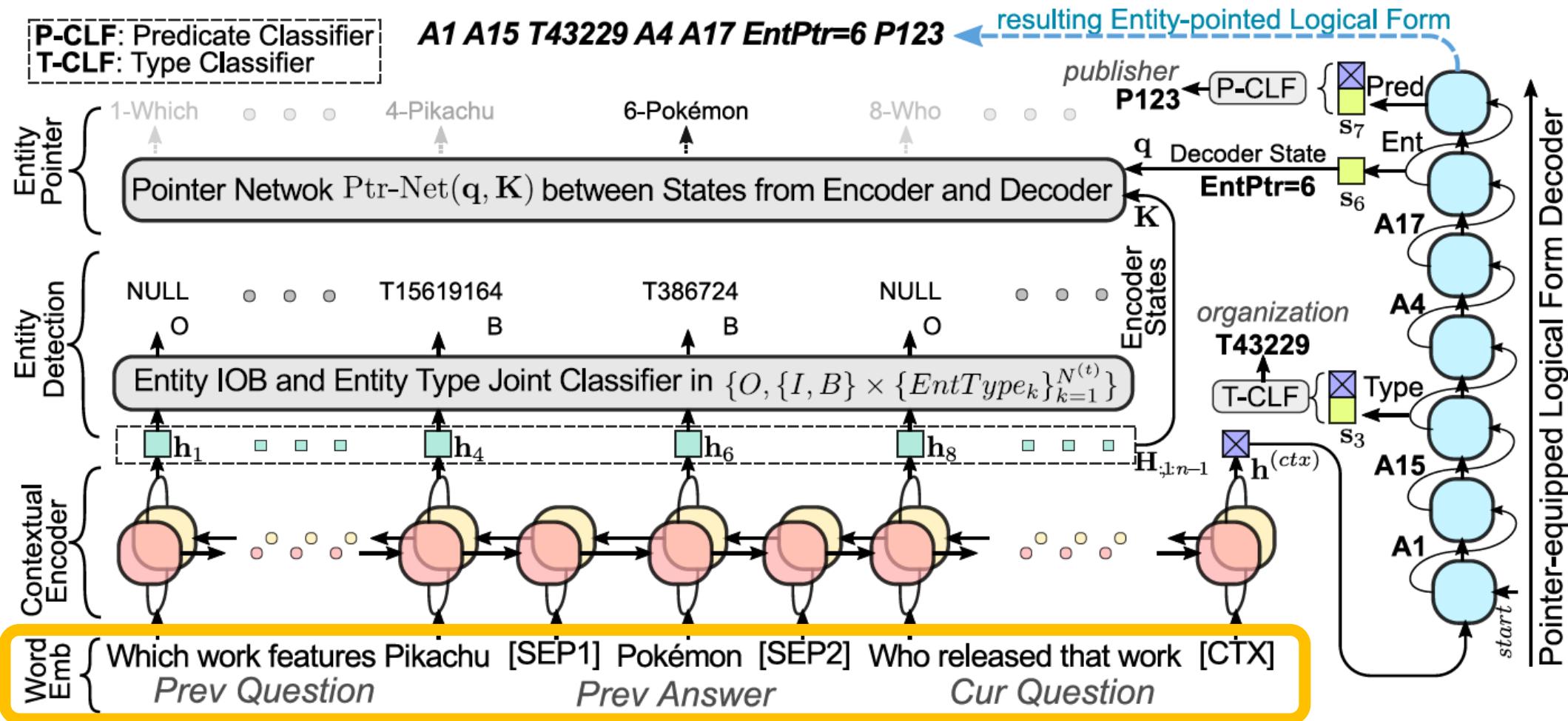
Shen et al., Multi-Task Learning for Conversational Question Answering over a Large-Scale Knowledge Base, EMNLP 2019.

# MaSP: Step-by-step



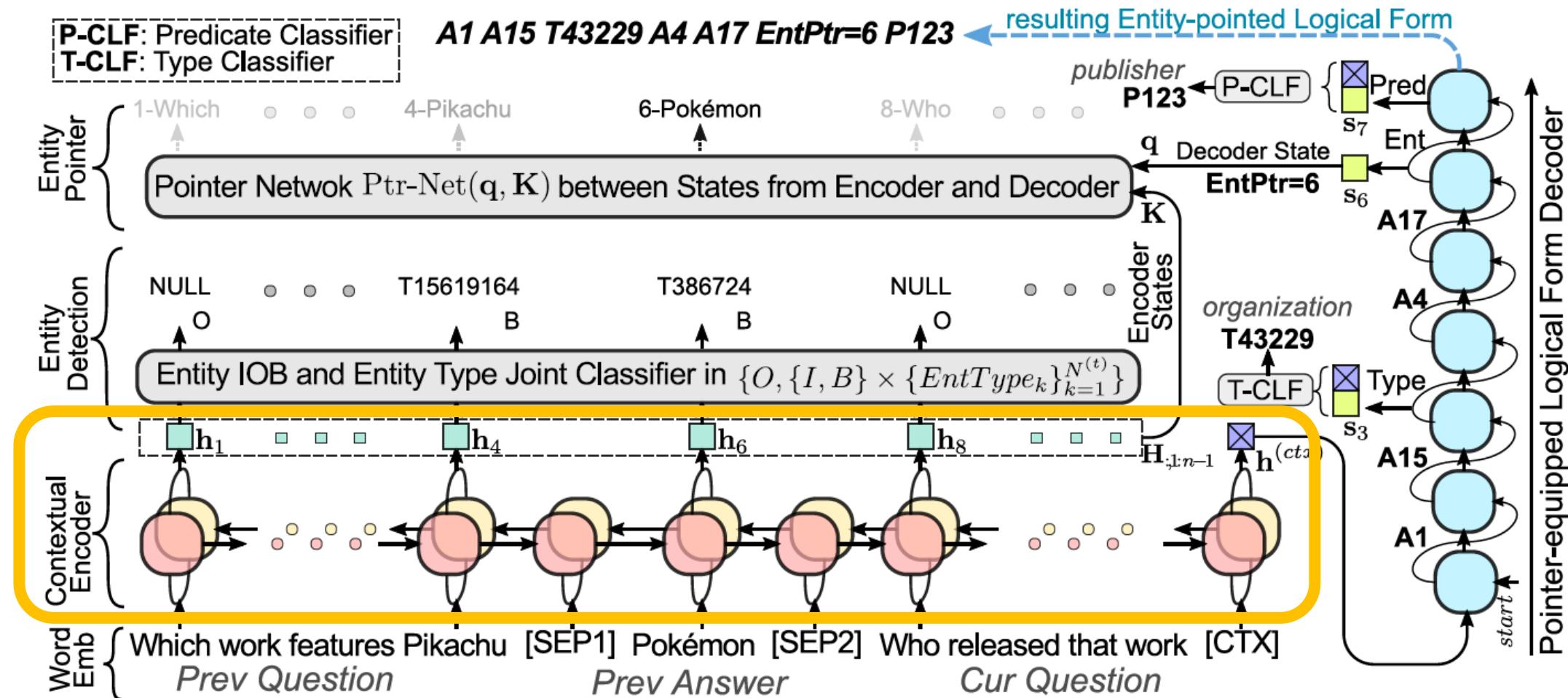
Shen et al., Multi-Task Learning for Conversational Question Answering over a Large-Scale Knowledge Base, EMNLP 2019.

# MaSP: Step-by-step



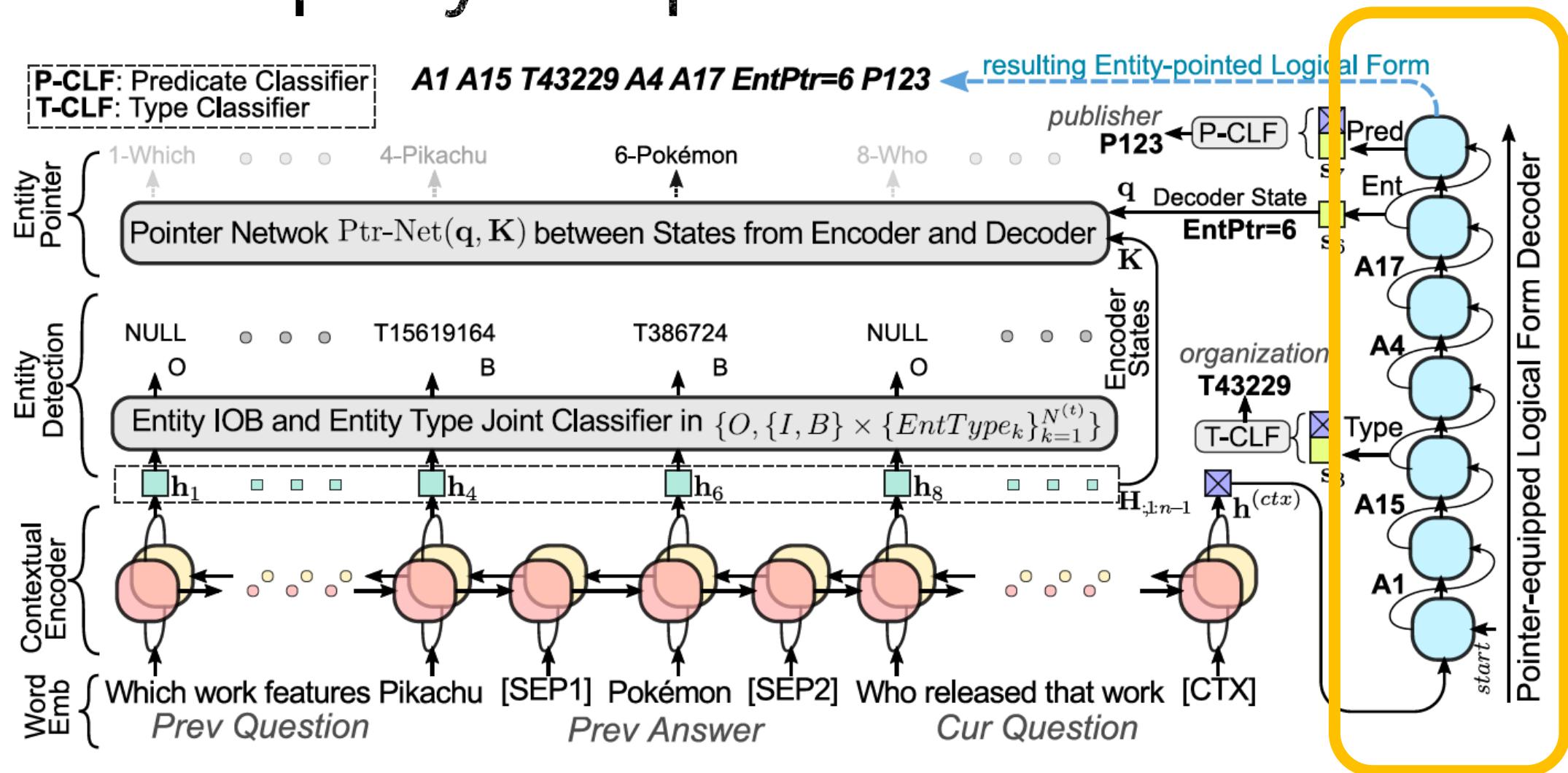
Shen et al., Multi-Task Learning for Conversational Question Answering over a Large-Scale Knowledge Base, EMNLP 2019.

# MaSP: Step-by-step



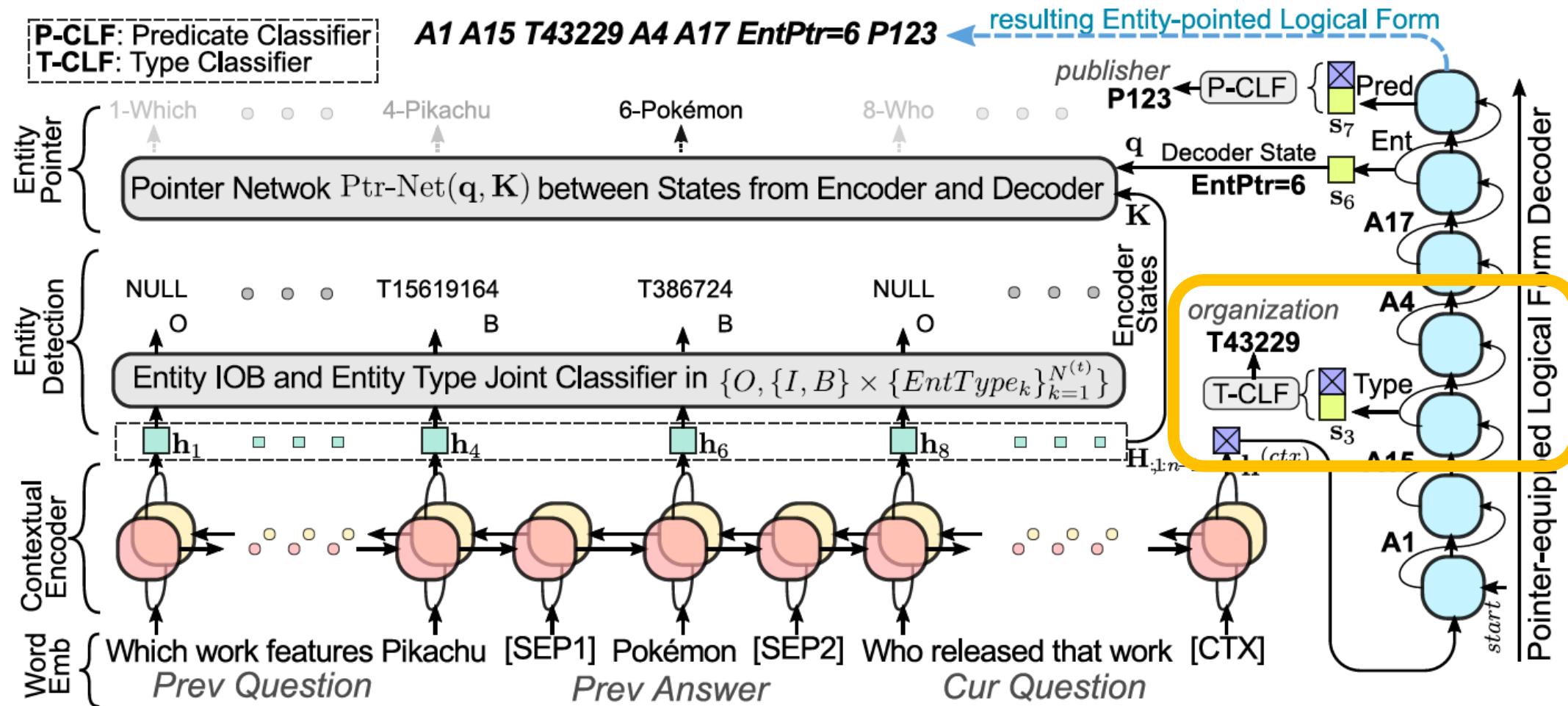
Shen et al., Multi-Task Learning for Conversational Question Answering over a Large-Scale Knowledge Base, EMNLP 2019.

# MaSP: Step-by-step



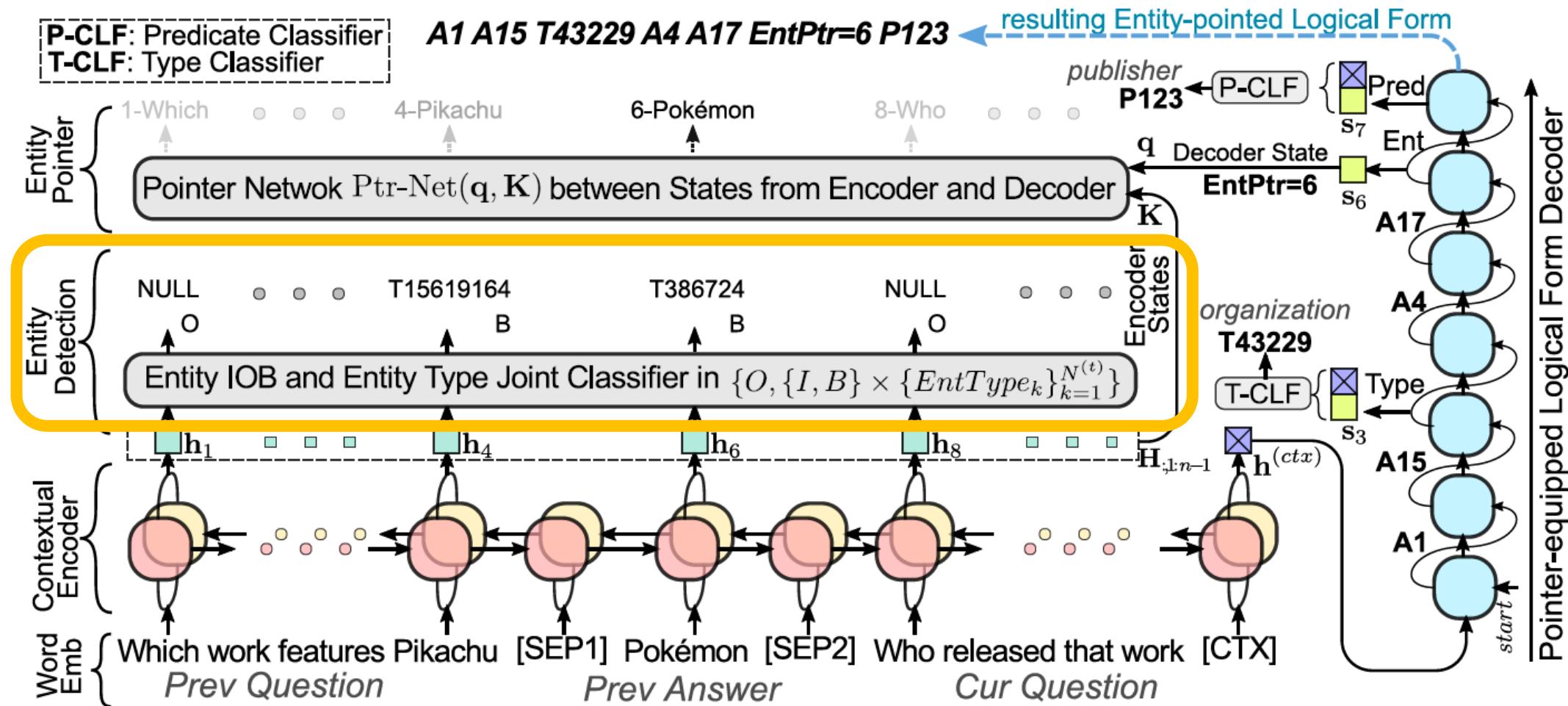
Shen et al., Multi-Task Learning for Conversational Question Answering over a Large-Scale Knowledge Base, EMNLP 2019.

# MaSP: Step-by-step



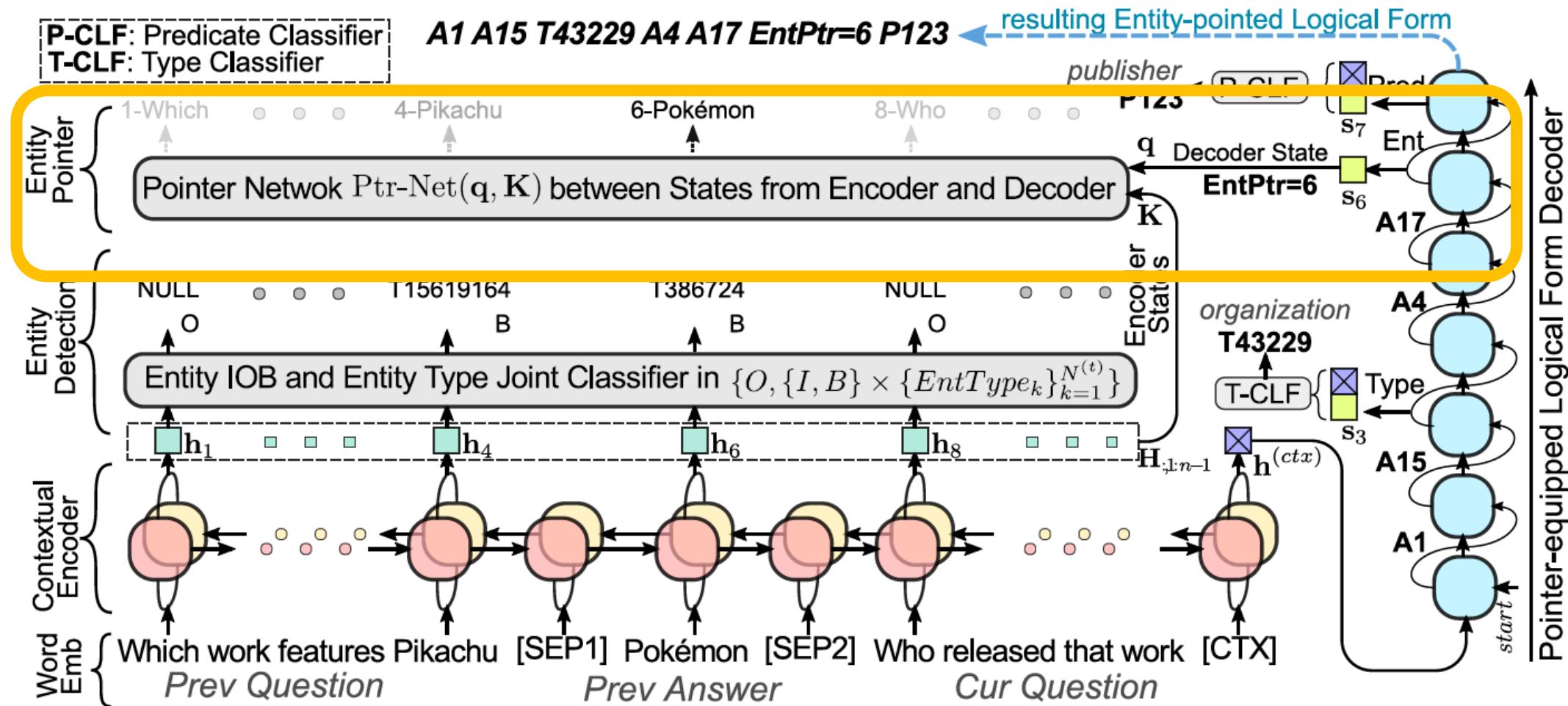
Shen et al., Multi-Task Learning for Conversational Question Answering over a Large-Scale Knowledge Base, EMNLP 2019.

# MaSP: Step-by-step



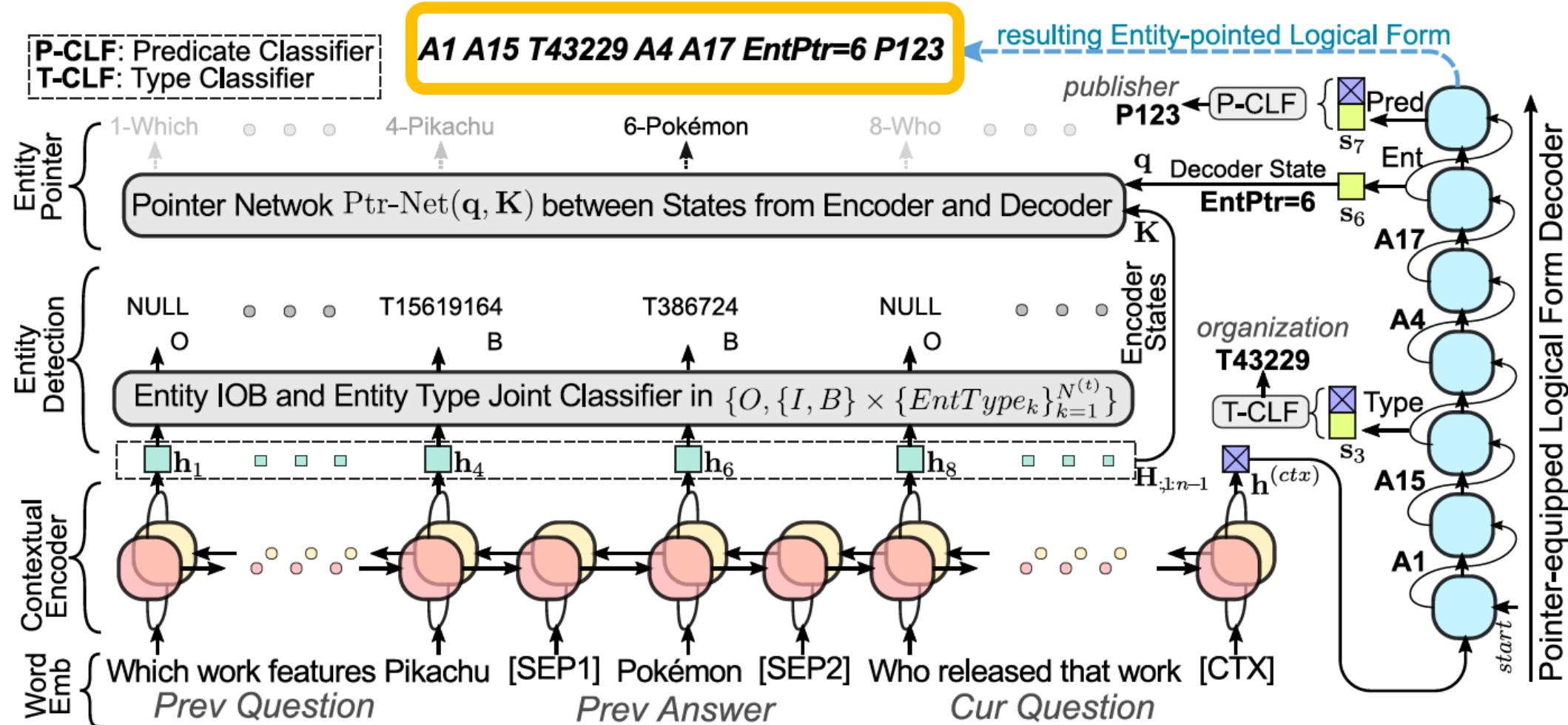
Shen et al., Multi-Task Learning for Conversational Question Answering over a Large-Scale Knowledge Base, EMNLP 2019.

# MaSP: Step-by-step



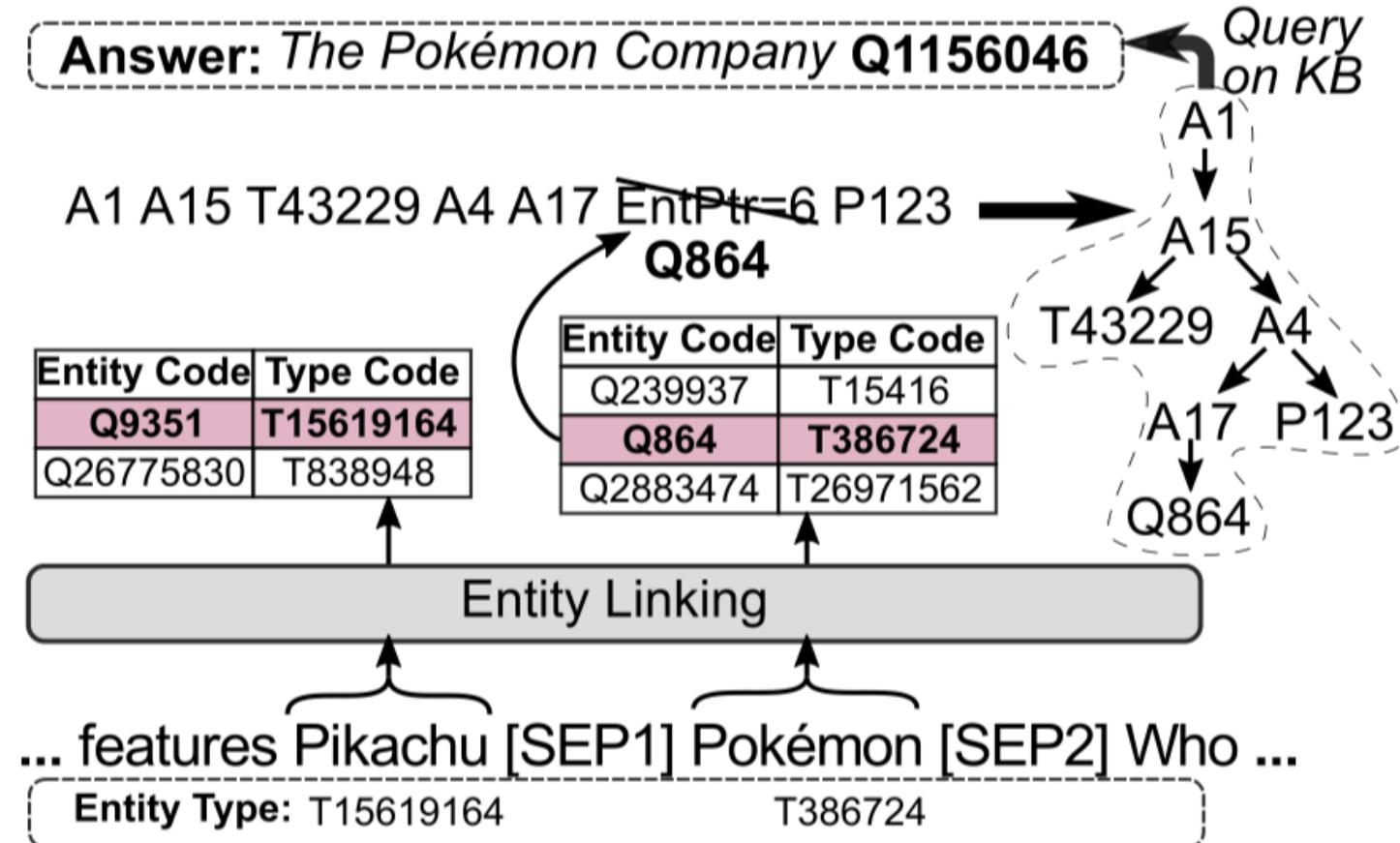
Shen et al., Multi-Task Learning for Conversational Question Answering over a Large-Scale Knowledge Base, EMNLP 2019.

# MaSP: Step-by-step



Shen et al., Multi-Task Learning for Conversational Question Answering over a Large-Scale Knowledge Base, EMNLP 2019.

# Execute query obtained via sequence decoding to get answer



# Conversational QA: Wrap-up

- Unsupervised graph traversal is promising
- KG connections offer vital clues for initializing and expanding context
- But limited to relatively simple information needs in utterances
- Sequence-sequence models can capture context well
- But ConvQA is much more than coreference and ellipsis resolution
- Zero-coreference / zero-anaphora utterances common (“batman actor?”)
- Question completion may be intractable + overkill

# Side glance: Table-QA

break duration ?x .  
?x measuredIn minutes .

- Web tables also constitute a huge volume of the curated Web
- Represent canonical challenges of querying a large-scale KB
- Selected references below for the interested reader

Chakrabarti et al., Open Domain Question Answering Using Web Tables, arXiv 2020.

Zhang, CFGNN: Cross Flow Graph Neural Networks for Question Answering on Complex Tables, AAAI 2020.

Wang et al., A Neural Question Answering Model Based on Semi-Structured Tables, COLING 2018.

Iyyer et al., Search-based Neural Structured Learning for Sequential Question Answering, ACL 2017.

Jauhar et al., Tables as Semi-structured Knowledge for Question Answering, ACL 2016.

Khashabi et al, Question Answering via Integer Programming over Semi-Structured Knowledge, IJCAI 2016.

Sun et al., Table Cell Search for Question Answering, WWW 2016.

Pasupat and Liang, Compositional Semantic Parsing on Semi-Structured Tables, ACL 2015.

# Outline: QA over knowledge graphs

- **Background:** Setup, benchmarks, metrics
- **Simple QA:** Templates and embeddings
- **Complex QA:** Multiple entities and predicates
- **Heterogeneous sources:** Handling KGs and text
- **Conversational QA:** Implicit context in multi-turn setup
- **Take-home:** Summary and insights

# Summary and insights

# Take-home messages

- Methodology
- Deployable system
- Open problems

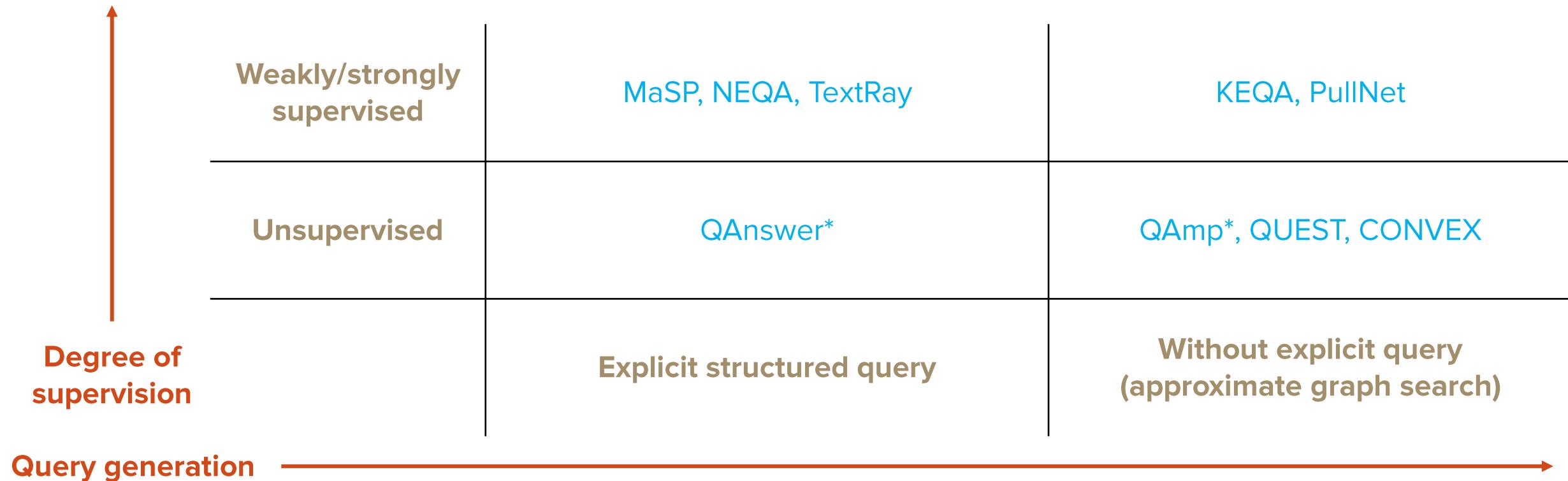
# Methodology

- Templates
- Graph embeddings
- Subgraph computations
- Belief propagation
- Graph traversal
- Sequence-to-sequence models

# Methodology

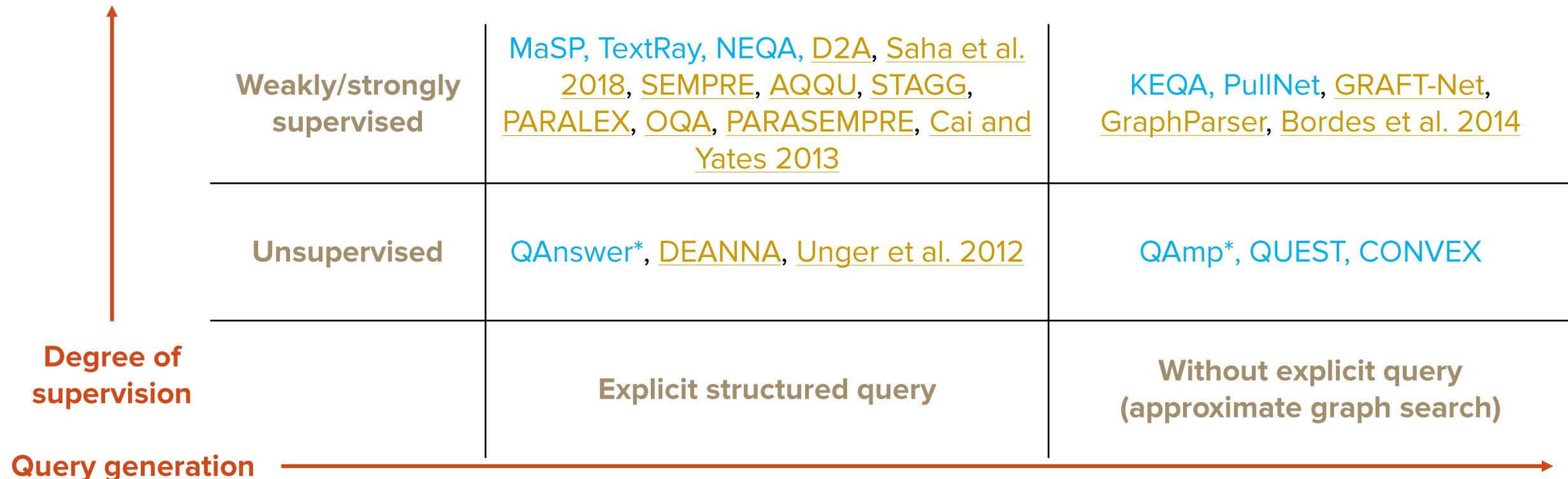
- Templates: NEQA
- Graph embeddings: KEQA
- Subgraph computations: TextRay, QUEST, QAnswer
- Belief propagation: QAmp
- Graph traversal: PullNet, CONVEX
- Sequence-to-sequence models: MaSP

# Methodology quad chart



\* Ranker/labeler supervised

# Methodology quad chart



\* Ranker/labeled supervised

# Methodology quad chart

Strongly supervised with $(Q, q)$	Cai and Yates 2013	-	-
Weakly supervised with $(Q, A)$	MaSP, NEQA, TextRay, D2A, Saha et al. 2018, SEMPRE, AQQU, STAGG	KEQA, PullNet, GRAFT-Net, GraphParser, Bordes et al. 2014	<p>There is some interplay in current systems but largely open area</p> <p>Unsupervised subgraph computations with small degree of supervised neural learning..?</p>
Weakly supervised with paraphrases	PARASEMPRE, PARALEX, OQA	-	
Unsupervised	QAnswer*, DEANNA, Unger et al. 2012	QAmp*, QUEST, CONVEX	
	Explicit structured query	Without explicit query	Hybrid search methods?

# Methodology: Pros and cons

Aspect	With explicit structured query (SPARQL-like)	Without explicit structured query (approx. graph search)
Simple questions	😊	😊
Single answer	😊	😊
List answer	😊	⁇?
Efficiency	😊	⁇?
Complex questions	⁇?	😊
Conversational questions	⁇?	😊
Heterogenous sources	⁇?	😊
Handling reified triples	⁇?	😊

😊 Preferable

⁇? Less preferable but scope for improvement

# Methodology: Lessons learnt

- Templates good for simple questions, but hits hurdles for complex questions, and useless for conversational 😞
- Graph embeddings effective for simple questions 😊 , not yet clear for complex scenarios...
- Sequence models (LSTM with attention) with pre-trained word embeddings very common
- Graph models generally more flexible (scope for node/edge types/weights)

# Open problems

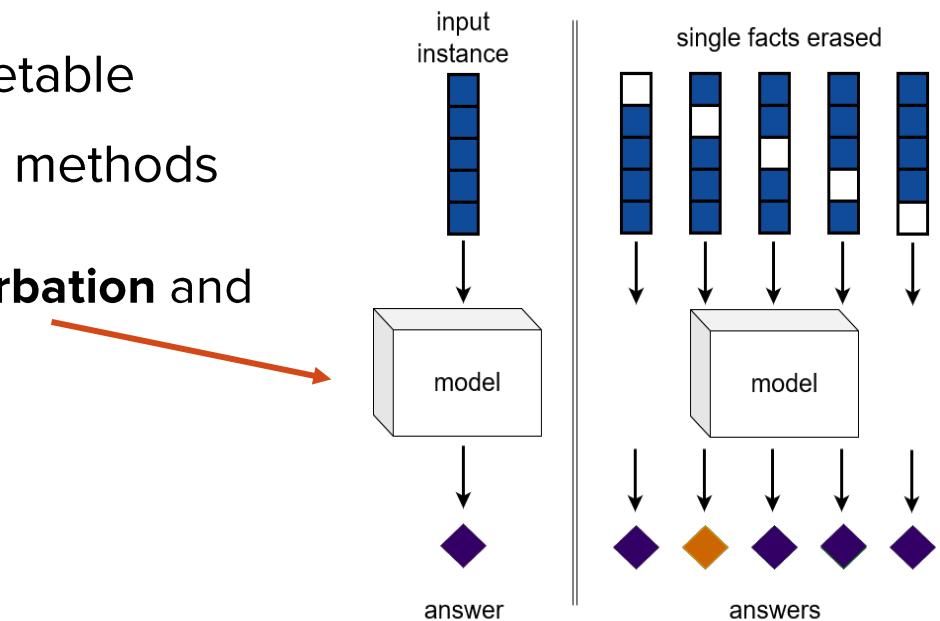
- Unanswerability
- Interpretability
- Interactivity
- Efficiency
- Robustness

# Open problems: Unanswerability

- Learn when to stay quiet and prevent embarrassment ☺
  - *Where was Messi's father born?*
  - *Who was the first man on Mars?*
- Knowing when answer is:
  - Not confident
  - Not in KG
  - Null
- Open and closed world assumptions
- Learn when to consult text

# Open problems: Interpretability

- Are your system's answers explainable? To the developer? What about the end user?
- Interpretability increases trust and guides user in case of mistakes
- Template- and graph-based methods construct interpretable evidence for answers - an unsolved concern for neural methods
- [Sydorova et al. \(2019\)](#) provide insights with **input perturbation** and **evaluation** of interpretability
- But very much an open problem!



# Open problems: Interactivity

- Towards mixed initiative systems ([Radlinski and Craswell 2017](#))
- Can your system absorb feedback?
- Positive and negative feedback?
- What kinds of feedback?
- Can your system ask clarifications?

# Open problems: Efficiency

- Critical component of QA systems
- Largely unexplored
- Identify bottlenecks
- Measure trade-offs

# Open problems: Robustness

- Think out of the ~~box~~ benchmark
- What is **open-domain** question answering?
- What happens for entities not seen during training?
- What about unseen predicates and vocabulary?

# Deploying a QA system

- Templates and unsupervised graph methods great way to get off the blocks with limited complexity
- Preferably with NER/NERD systems and pre-trained word embeddings
- Need seed data + domain knowledge
- Continuous learning with similarity function and feedback vital cogs
- Level of structure and heterogeneity in data and questions indicators of follow-up modeling

# Take-home messages

- Overview of state-of-the-art in KG-QA and their positioning
- Families of algorithms with a few specific instantiations
- Several open problems in the key areas of focus

**Simple / complex / heterogeneous / conversational questions for me 😊 ?**

# QA@MPII-D5: Visit [qa.mpi-inf.mpg.de](https://qa.mpi-inf.mpg.de)

- **Course** on QA systems: <https://www.mpi-inf.mpg.de/question-answering-systems/>
- **CONVEX**: Conversational QA over KGs [CIKM 2019]: <https://convex.mpi-inf.mpg.de/>
- **CROWN**: Conversational QA over passages [SIGIR 2020]: <https://crown.mpi-inf.mpg.de/>
- **QUEST**: Complex question answering [SIGIR 2019]: <https://quest.mpi-inf.mpg.de/>
- **ComQA**: QA benchmark with paraphrase [NAACL 2019]: <http://qa.mpi-inf.mpg.de/comqa/>
- **TEQUILA**: Temporal question answering [CIKM 2018]: <https://tequila.mpi-inf.mpg.de/>
- **QUINT**: Template-based question answering [EMNLP 2017]: <https://quint.mpi-inf.mpg.de/>
- Send an email to [rishiraj@mpii.de](mailto:rishiraj@mpii.de) in case of any issues!



# Question Answering over Curated and Open Web Sources

Rishiraj Saha Roy

Max Planck Institute for Informatics, Germany

Avishek Anand

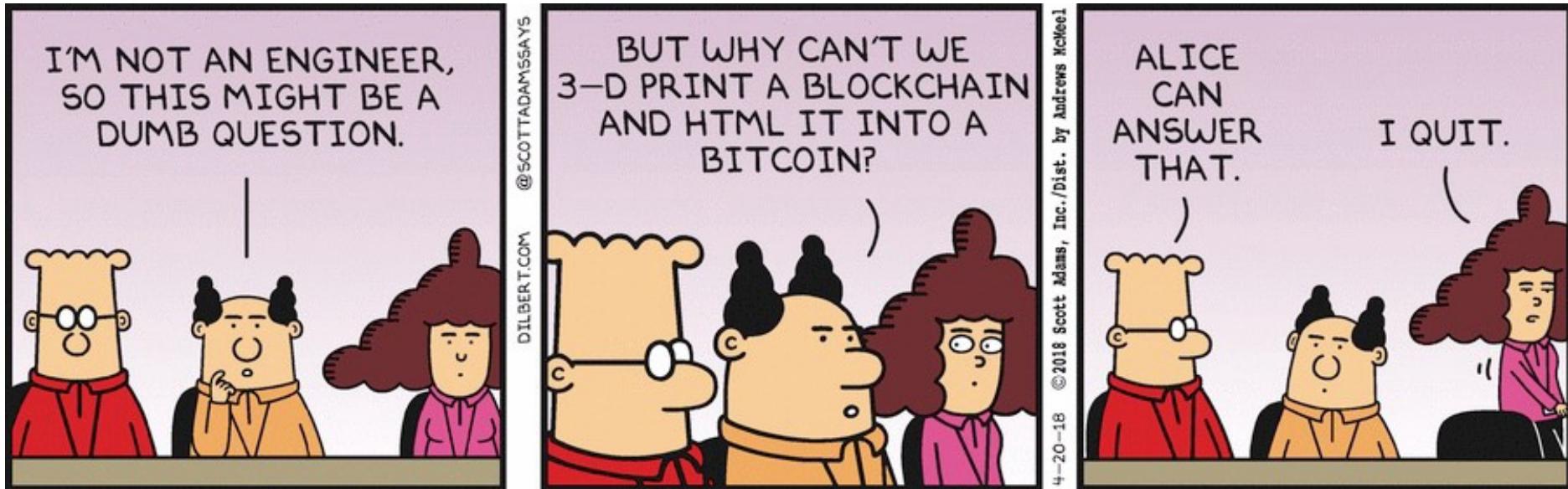
L3S Research Center, Germany



SIGIR 2020 Tutorial

26 July 2020

# Question Answering is AI complete



- To perfectly solve QA
  - Involves world-knowledge
  - Linguistics: co-ref, pragmatics, high-level reasoning tasks such as natural language inference
  - Common sense reasoning

# Question Answering is AI complete

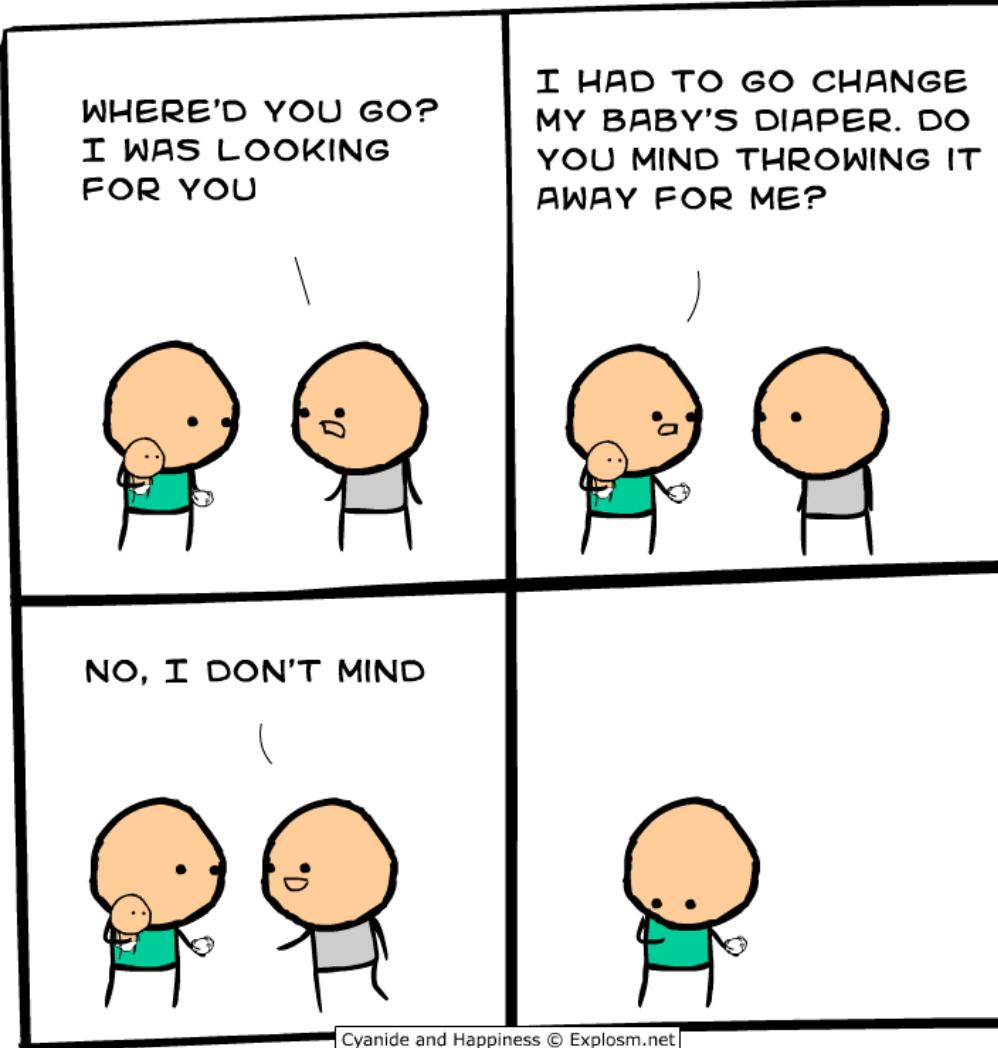
**Passage:** *Anita was stung by a bee and left the garden.*

**Question:** Why did Anita leave the garden ?

- A) Because she was in pain
- B) Because it was time for a TV show she didn't want to miss
- C) Because its common practice to leave the garden after being stung by a bee
- D) Because the bee needed some peace

- Simple questions that are hard for the machine
- Need Pragmatics

# Question Answering is AI complete



# Question Answering is AI complete

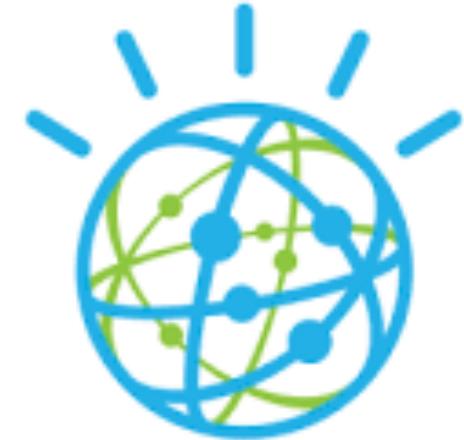
- In its full glory, it is indeed hard
- But there has been a lot of progress
  - Knowledge-driven Question Answering
  - Reading comprehension

*The AI in this sci-fi movie owed its intelligence to a massive cache of search engine data.*

*This movie has the plot adapted from this famous play..and has 3 of its main characters named after all biblical characters*

*— Eve, Nathan and Caleb*

# Lots of Success



**IBM Watson**

# Lots of Applications

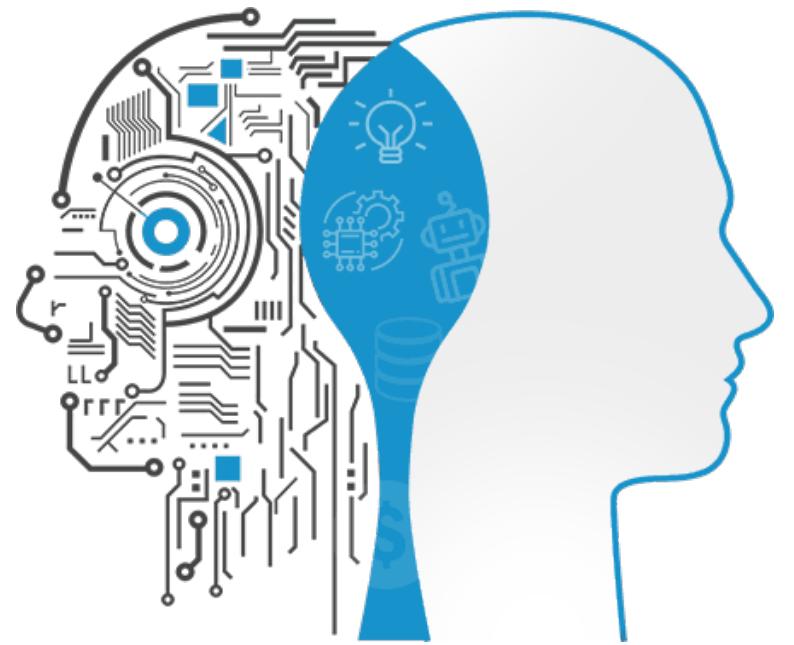
*It is automated extracting structured information, defining objects, their relations, and characteristics in documents in natural language, extract required info can extract from text events, terminology, emotions, organizations, locations) and other data.*



**Problems in NLP, Dialog and Search can be formulated as QA**

# In this part of the tutorial...

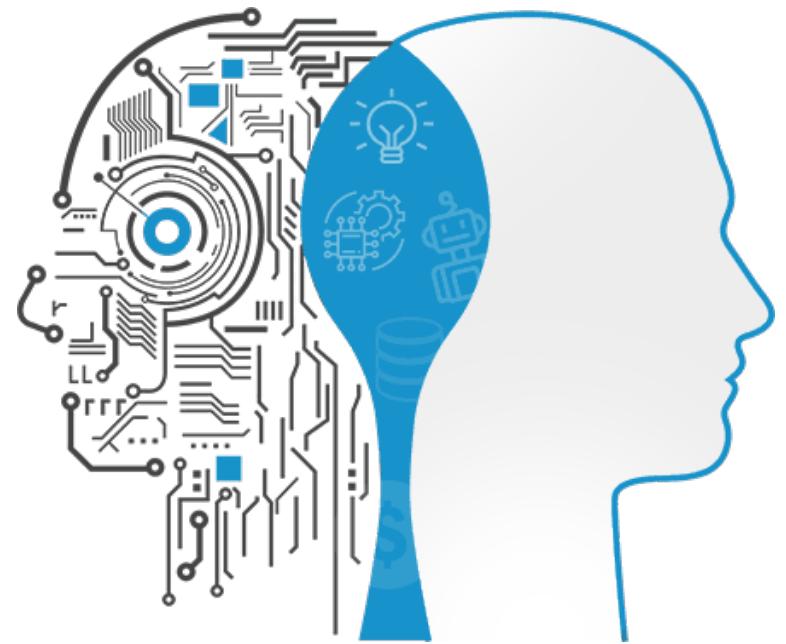
- We focus on where progress has been made
- What tasks are out there
  - Reading comprehension, Open-domain QA, Conversational QA
- What models are usually used to solve them
  - Neural, neural, neural
- Challenges and design decision



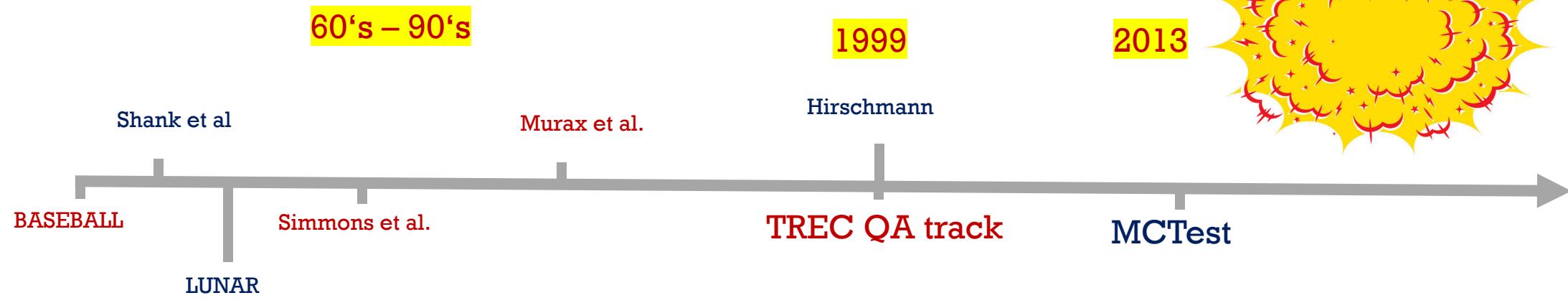
# What we do not cover

- Approaches pre 2016
- Other related QA tasks
  - MCQ
  - Visual QA
  - Complex QA requiring selection, aggregation ...
- Model details

We will miss many QA approaches and many QA tasks....



# Brief History of QA



## NLP interfaces to databases

- Precursors to the modern open-domain QA
- Mostly structured knowledge and limited domain

## Story Comprehension

- Precursors to Modern RC tasks
- Shank et al. (1977) – Yale AI Project
- Hirschman (1999)

# QA @ TREC

- **Goal**

- Encourage research in information retrieval based on large-scale collections

- **Types of Questions:**

**Fact-based, short answers**

- How many feet in a mile ?
- Name a food high in zinc.
- When was the first stamp issued ?

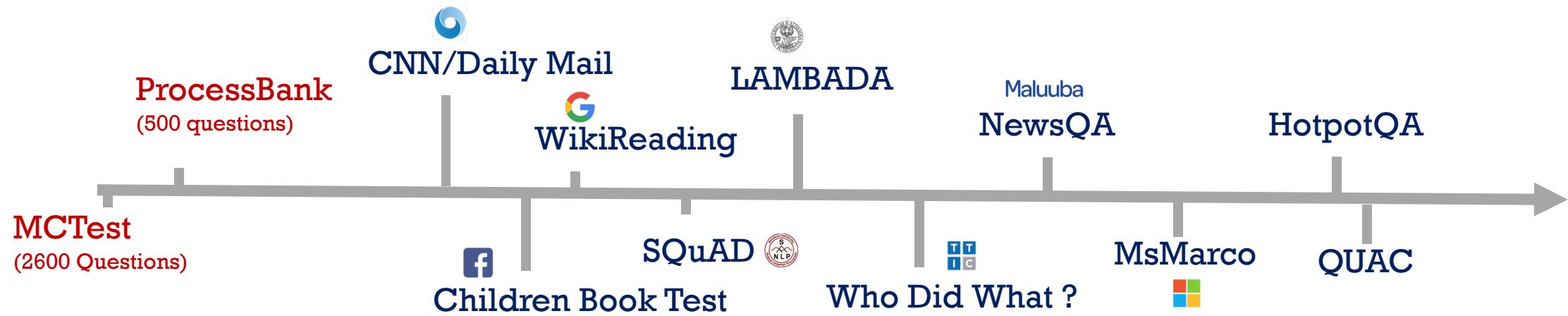
**Definition questions**

- Who was Galileo ?
- What is an atom ?
- What is lymphosarcoma ?

**Reformulation questions**

- What attracts tourists in Reims ?
- What are tourist attractions in Reims ?

# Modern History of Text QA



- Started with CNN/Daily Mail and popularized with SQuAD
- Benchmarks ranging from
  - Simple to complex questions
  - Realistic to synthetically generated questions
  - Crowdsourced to extractive answers

# SQUAD

**Question:** Which team won Super Bowl 50?

## Passage

Super Bowl 50 was an American football game to determine the champion of the National Football League (NFL) for the 2015 season. The American Football Conference (AFC) champion Denver Broncos defeated the National Football Conference (NFC) champion Carolina Panthers 24–10 to earn their third Super Bowl title. The game was played on February 7, 2016, at Levi's Stadium in the San Francisco Bay Area at Santa Clara, California.

- Extractive answers, span extraction
- > 100k examples
- Deep learning wave

# SQuAD v1, v2

**Question:** Which team won Super Bowl 50?

## Passage

Super Bowl 50 was an American football game to determine the champion of the National Football League (NFL) for the 2015 season. The American Football Conference (AFC) champion **Denver Broncos** defeated the National Football Conference (NFC) champion Carolina Panthers 24–10 to earn their third Super Bowl title. The game was played on February 7, 2016, at Levi's Stadium in the San Francisco Bay Area at Santa Clara, California.

**Question:** Which team won Quiz Bowl 50 ?

### Version 1

- All answers in the context
- Evaluation – Exact Match
- Evaluation – F1
  - Partial match assuming BoW

### Version 2

- Open world assumption
- 1/3 training instances have no answer
- ½ dev/test instances have no answer

# Cloze tests



CNN/Daily Mail

Context:

the *ent381* producer allegedly struck by *ent212* will not press charges against the “*ent153*” host, his lawyer said Friday. *ent212*, who hosted one of the most-watched television shows in the world, was dropped by the *ent381* Wednesday after an internal investigation by the *ent180* broadcaster found he had subjected producer *ent193* “to an unprovoked physical and verbal attack.”

Question: producer X will not press charges against *ent212*, his lawyer says.

Answer: *ent193*

# Span Extraction

- Context: Computational complexity theory is a branch of the theory of computation in theoretical computer science that focuses on classifying computational problems according to their **inherent difficulty**, and relating those classes to each other. A computational problem is understood to be a task that is in principle amenable to being solved by a computer, which is equivalent to stating that the problem may be solved by mechanical application of mathematical steps, such as an algorithm.
- 
- Question: By what main attribute are computational problems classified using computational complexity theory?
- 
- Answer: inherent difficulty



SQuAD

# Multiple Choice Answers

Context: If you have a cold or flu, you must always deal with used tissues carefully. Do not leave dirty tissues on your desk or on the floor. Someone else must pick these up and viruses could be passed on.

---

Question: Dealing with used tissues properly is important because \_\_\_\_\_.

---

Options:

- A. it helps keep your classroom tidy
- B. people hate picking up dirty tissues
- C. it prevents the spread of colds and flu
- D. picking up lots of tissues is hard work

---

Answer: C

# Free form Answers

- 
- Context 1: Rachel Carson's essay on The Obligation to Endure, is a very convincing argument about the harmful uses of chemical, pesticides, herbicides and fertilizers on the environment.  
.....
- Context 5: Carson believes that as man tries to eliminate unwanted insects and weeds; however he is actually causing more problems by polluting the . environment with, for example, DDT and harming living things  
.....
- Context 10: Carson subtly defers her writing in just the right writing for it to not be subject to an induction run rampant style which grabs the readers interest without biasing the whole article.
- 
- Question: Why did Rachel Carson write an obligation to endure?
- 
- Answer: Rachel Carson writes The Obligation to Endure because believes that as man tries to eliminate unwanted insects and weeds; however he is actu -ally causing more problems by polluting the environment.

# Outline: QA over Text

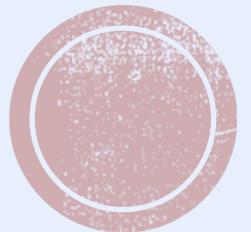
- **Background:** History, Tasks
- **Machine Comprehension:** Neural models, attention
- **Open Domain QA:** QA over a text corpus
- **Feedback and Interpretability**
- **Conversational QA:** Implicit context in multi-turn setup
- **Take-home:** Summary and insights

Representative methods from each task

Families of algorithms to build up repertoire for text

Focus on methods (and not evaluation)

Design decisions and challenges



# MACHINE COMPREHENSION

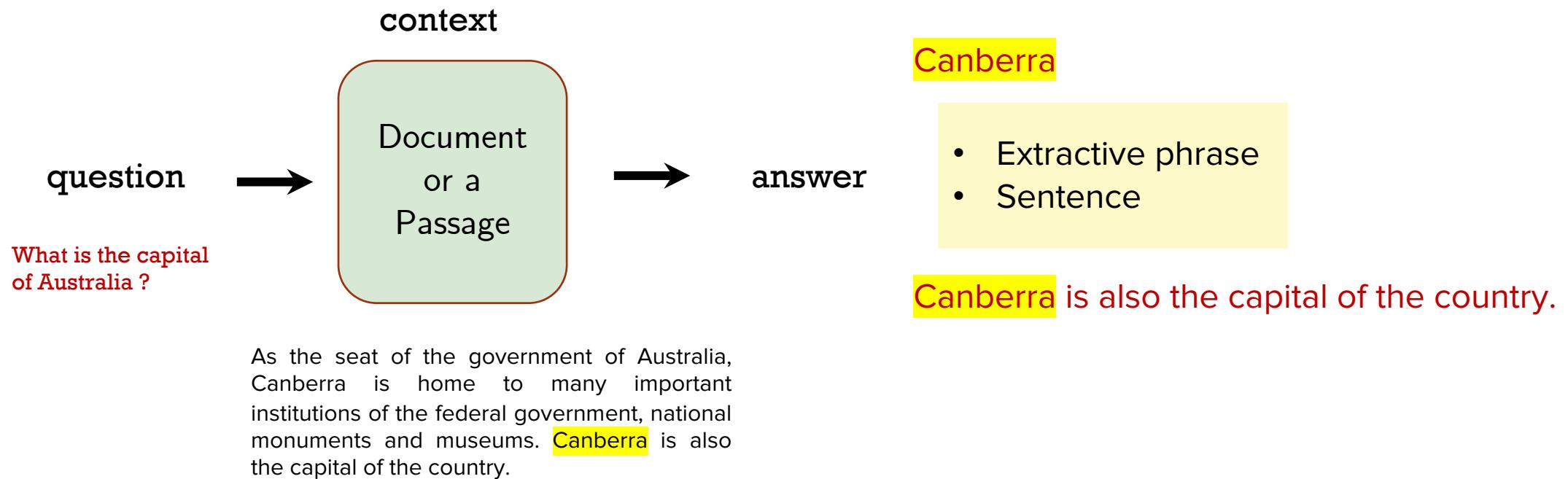


# Machine Comprehension

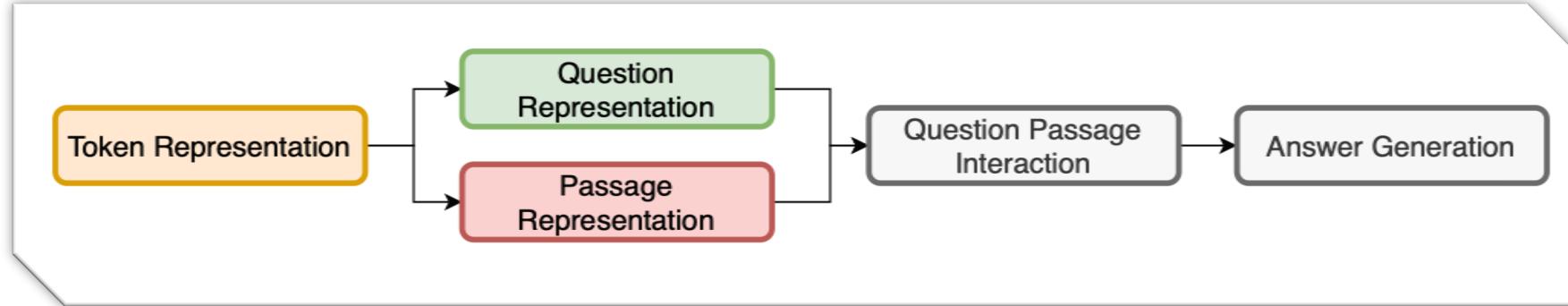
Chris Burges 2013

“A machine **comprehends** a passage of **text** if, for any **question** regarding that text that can be **answered** correctly by a majority of native speakers, that machine can provide a string which those speakers would agree both answers that question, and does not contain information irrelevant to that question.”

# Problem Setting

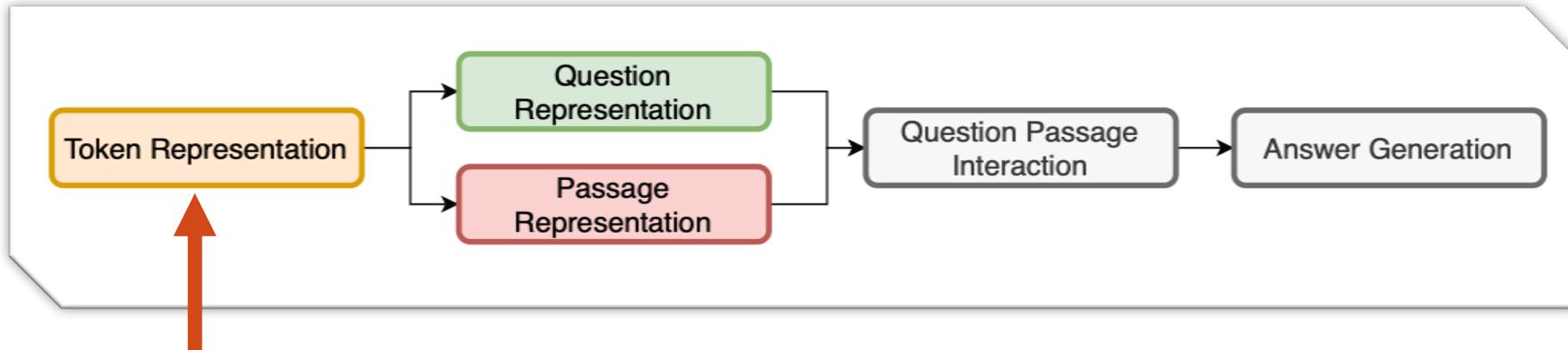


# The MRC Pipeline



- Words, characters, sub-words embeddings
- Contextual Embeddings
- Other features – Matching, Alignment, Language structure
- Sequential representation
- Contextual representation
- Attentive reading
- Attentive reading
- Attention flows
- Multiple input passes inputs
- Re-representation of question and passages
- Token prediction
- Span prediction
- Free-form generation

# Token Representation



## Conventional

- One Hot
- Word Embeddings
- Sub-word, Character Embeddings

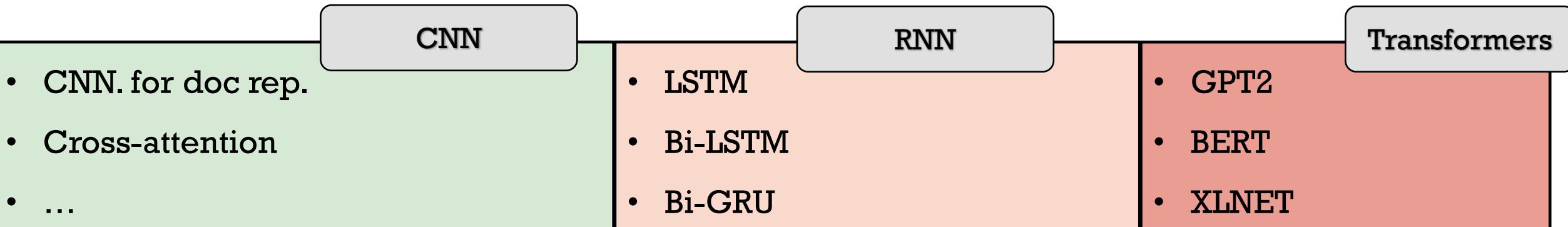
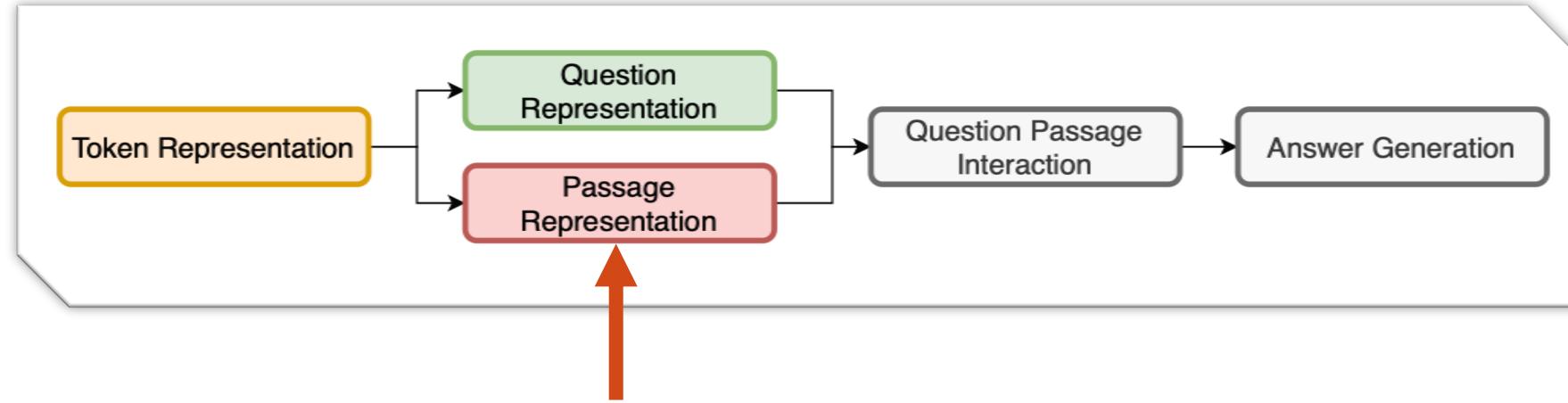
## Linguistic

- POS
- Named Entity
- Query Category

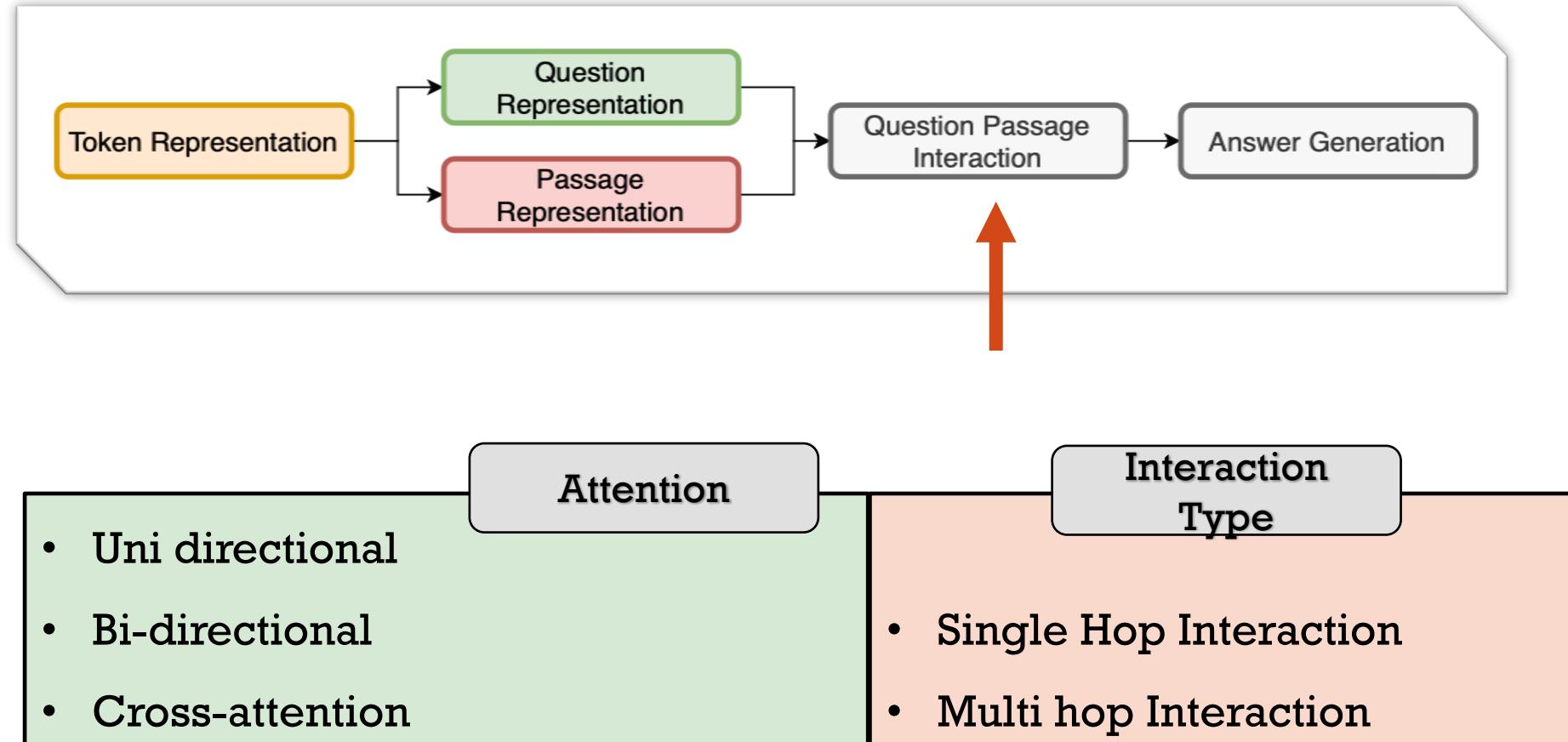
## Contextual

- Bi-LSTM
- BERT
- ELMO

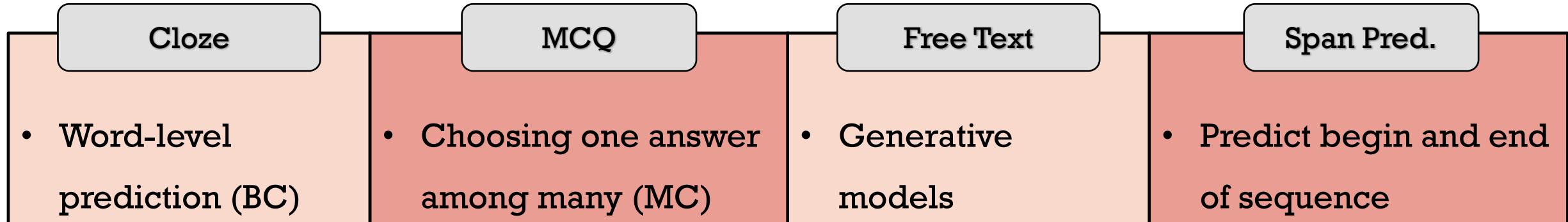
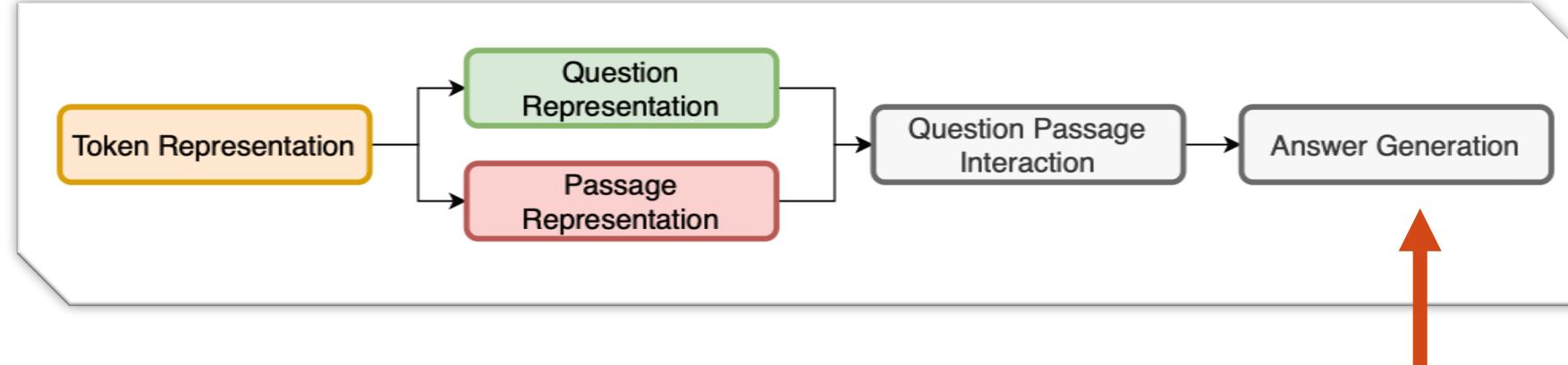
# Question/Passage Representation



# Question And Passage Interaction



# Answer Generation



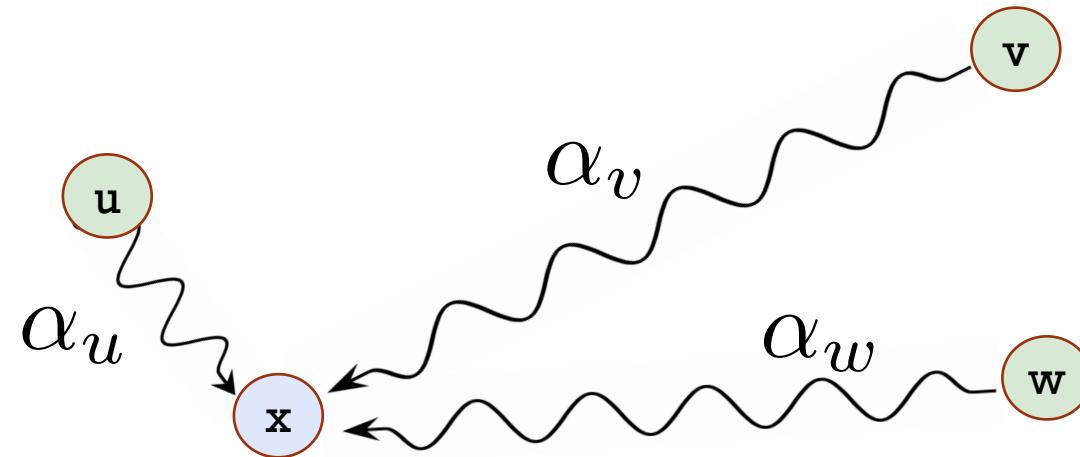
# Attention Mechanism

- Attention is used to represent tokens, question and passages
  - How do we re-represent otherwise independent token representations ?
  - How do we leverage contextualization ?
- Hard attention
- **Soft Attention**
- Co-attention
- Self-attention

# Attention – Influence Point Of View

Attention encodes how much influence the context  $u$  has on  $x$

$$\alpha_u = \left( \frac{e^{\mathbf{x} \cdot \mathbf{u}}}{e^{\mathbf{x} \cdot \mathbf{u}} + e^{\mathbf{x} \cdot \mathbf{v}} + e^{\mathbf{x} \cdot \mathbf{w}}} \right)$$



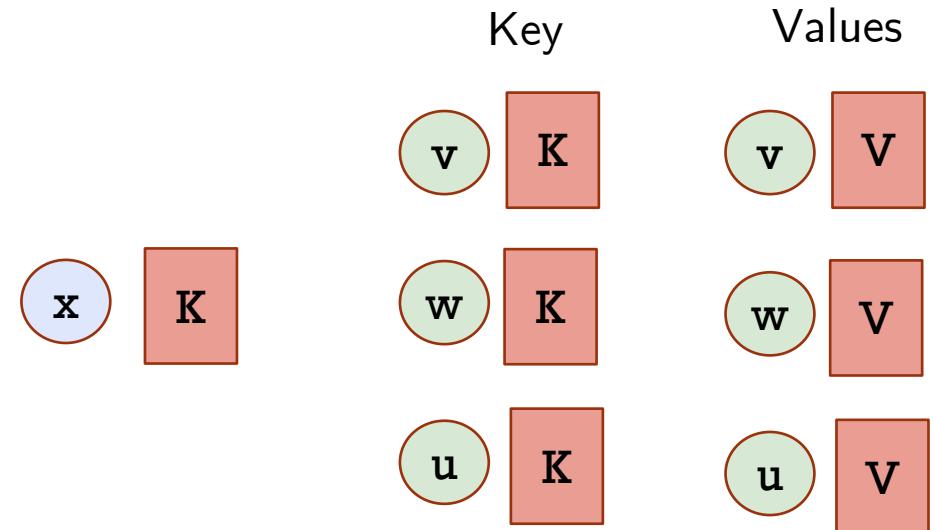
$$\mathbf{x}' = \alpha_u \mathbf{u} + \alpha_v \mathbf{v} + \alpha_w \mathbf{w}$$

- Typically  $x$  and context vectors are first projected through a learnable matrix  $W$

# Attention Mechanism – Memory Point Of View

Attention retrieves values from a continuous memory using fuzzy matching

- Assume vectors are stored in memory referenced by Key matrix K
- Thought expt: for 1-hot vectors = hashmaps
- Instead  $Kx$  retrieves from this continuous memory as a weighted sum over all values



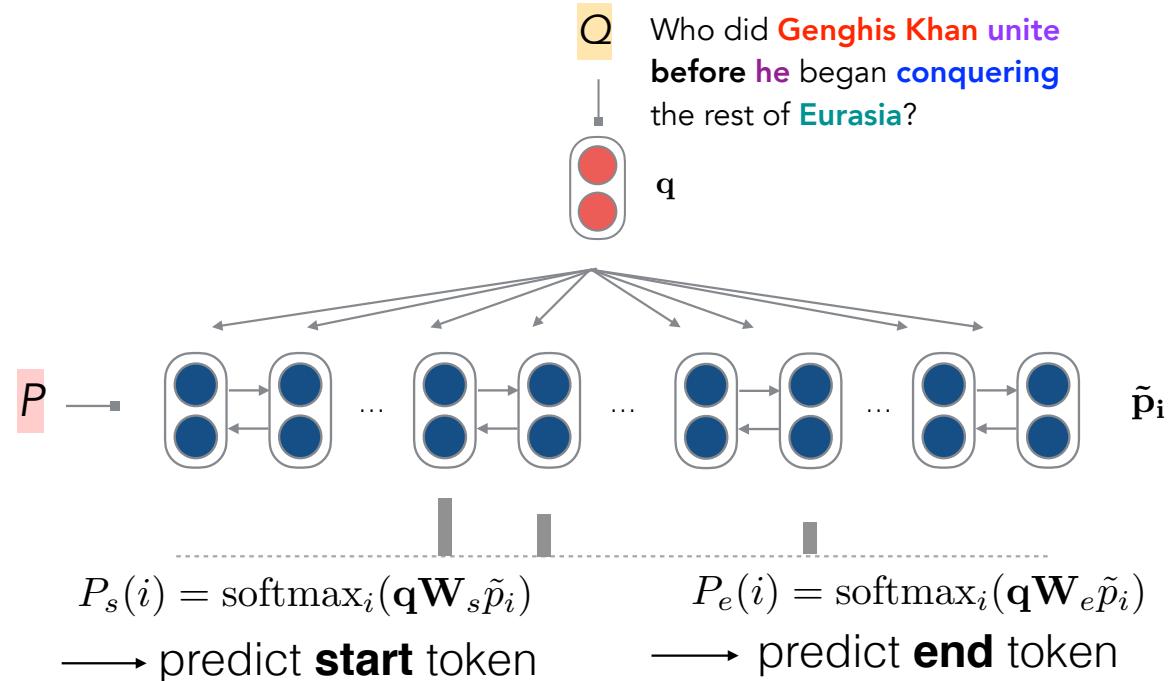
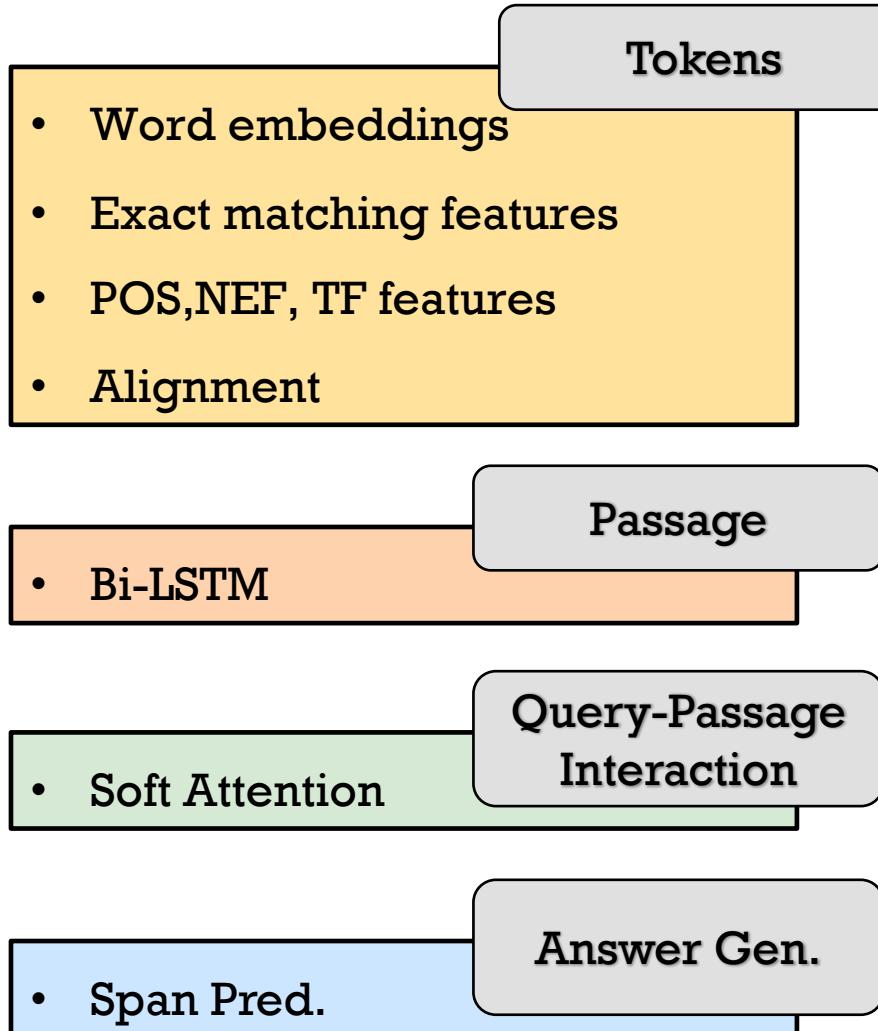
Attention weight

$$\alpha_u = \frac{e^{K\mathbf{x} \cdot K\mathbf{u}}}{e^{K\mathbf{x} \cdot K\mathbf{u}} + e^{K\mathbf{x} \cdot K\mathbf{v}} + e^{K\mathbf{x} \cdot K\mathbf{w}}}$$

$$\mathbf{x}' = \alpha_u \mathbf{u} + \alpha_v \mathbf{v} + \alpha_w \mathbf{w}$$

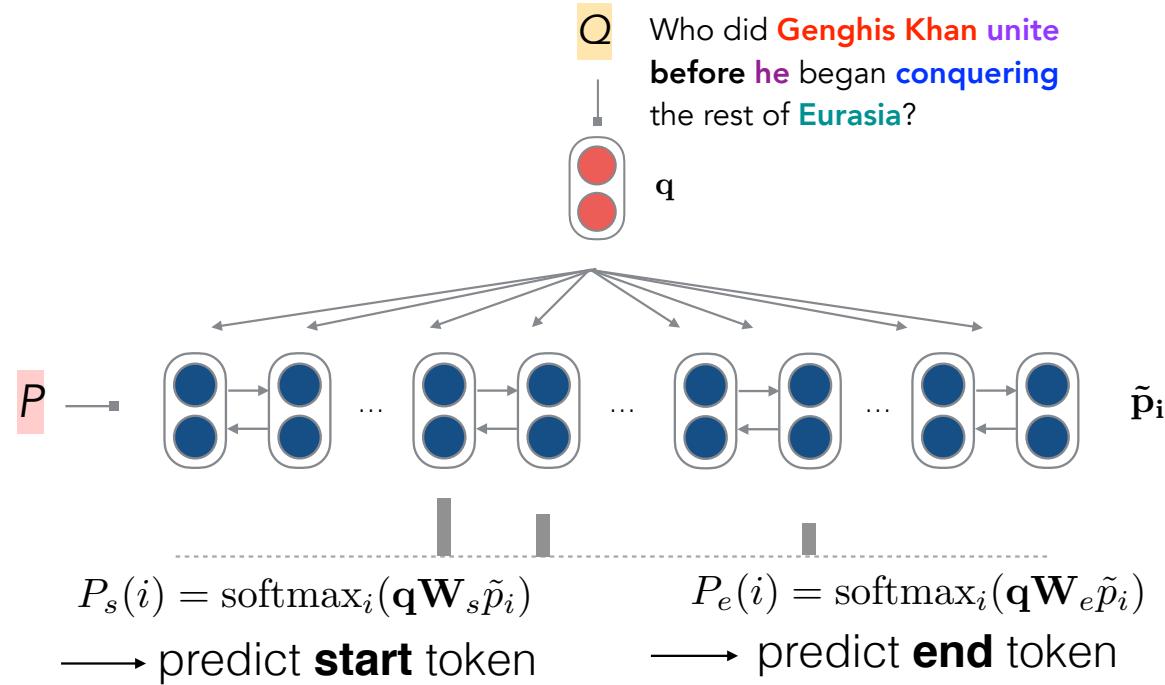
# Attentive Reader

[Chen '16] Attentive reader



# Attentive Reader

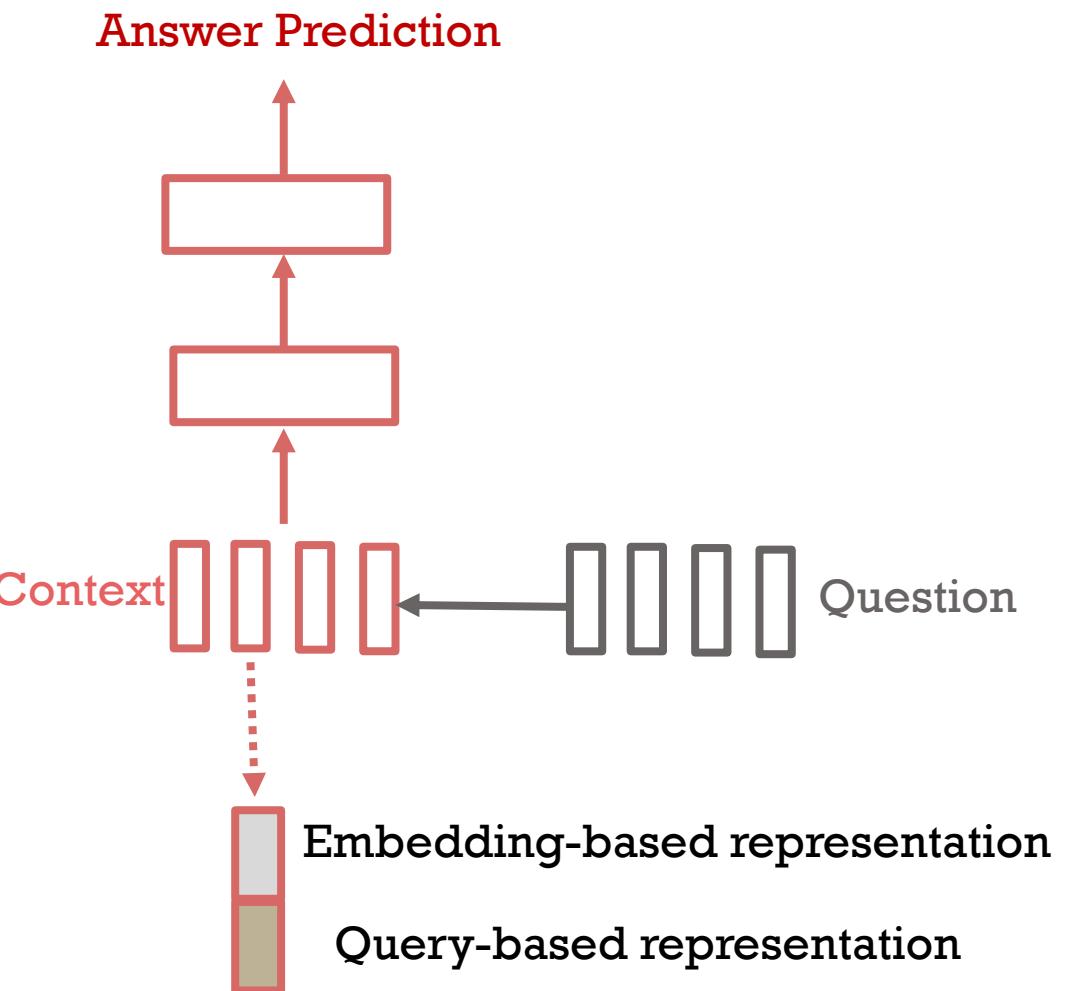
[Chen '16]



$$\Pr(a|q, p_i) = P_s(a_s)P_e(a_e)$$

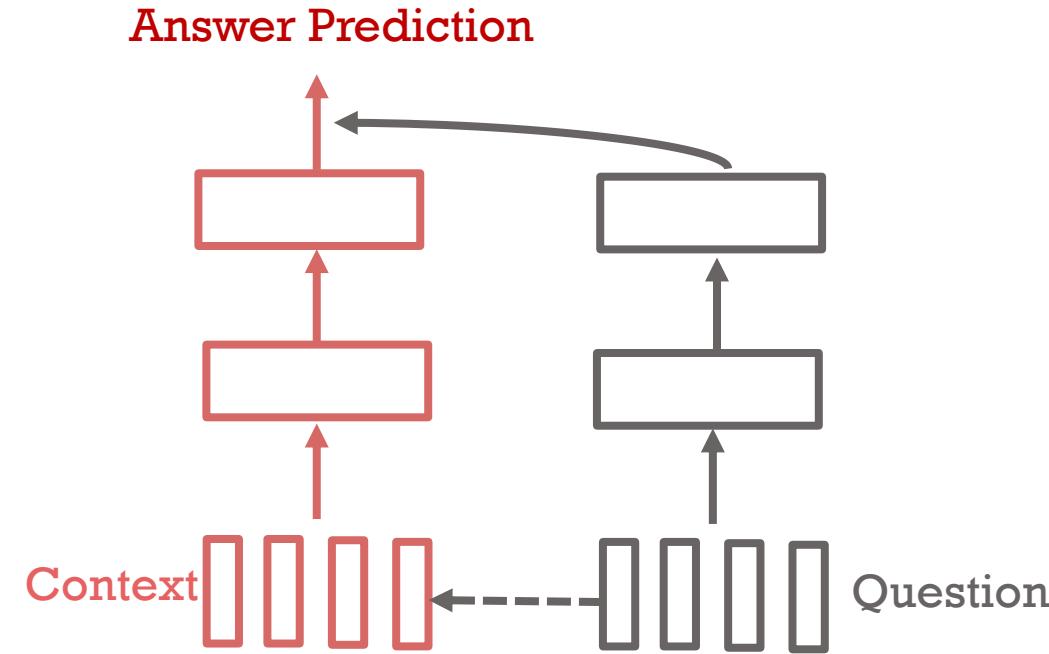
# Input Representation

- Context words are represented based on similarity with the query
- Semantic similarity
  - Word embeddings
- Matching similarity
  - Direct word-level matching
  - Weighted matching
  - Attention mechanism

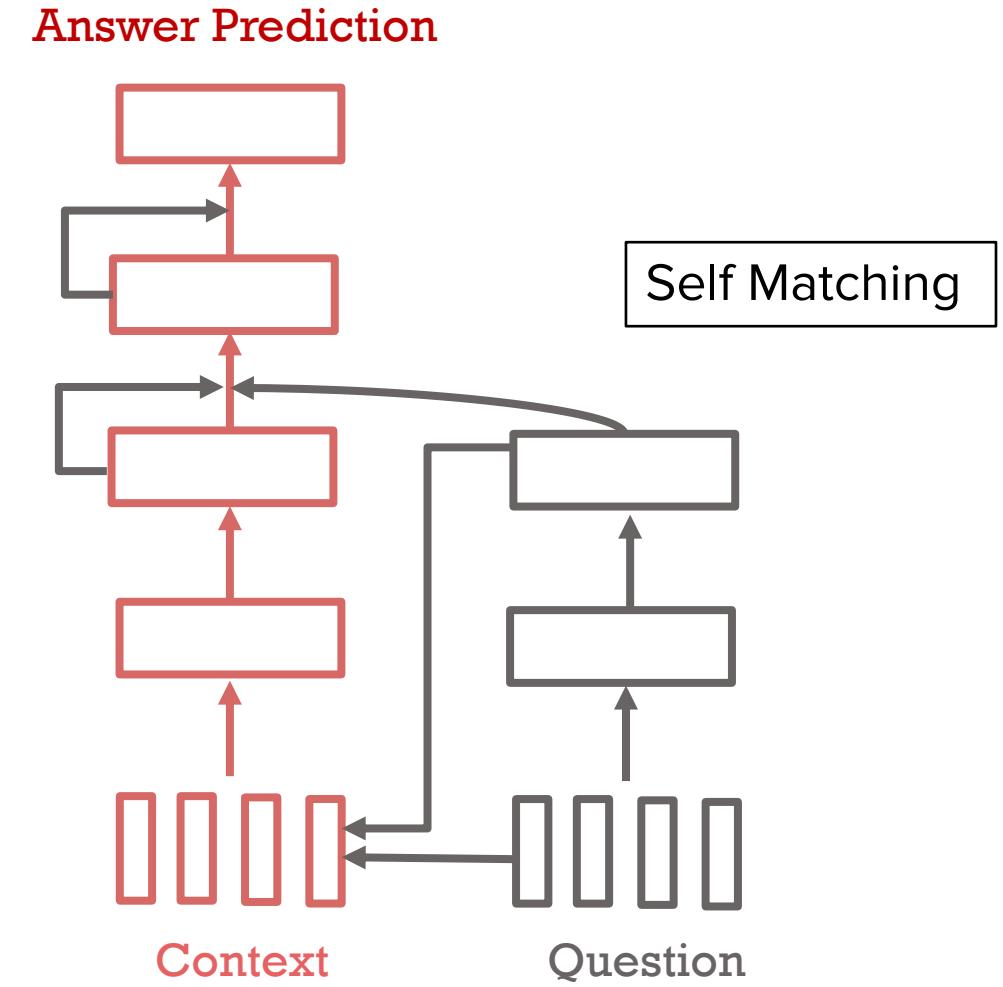
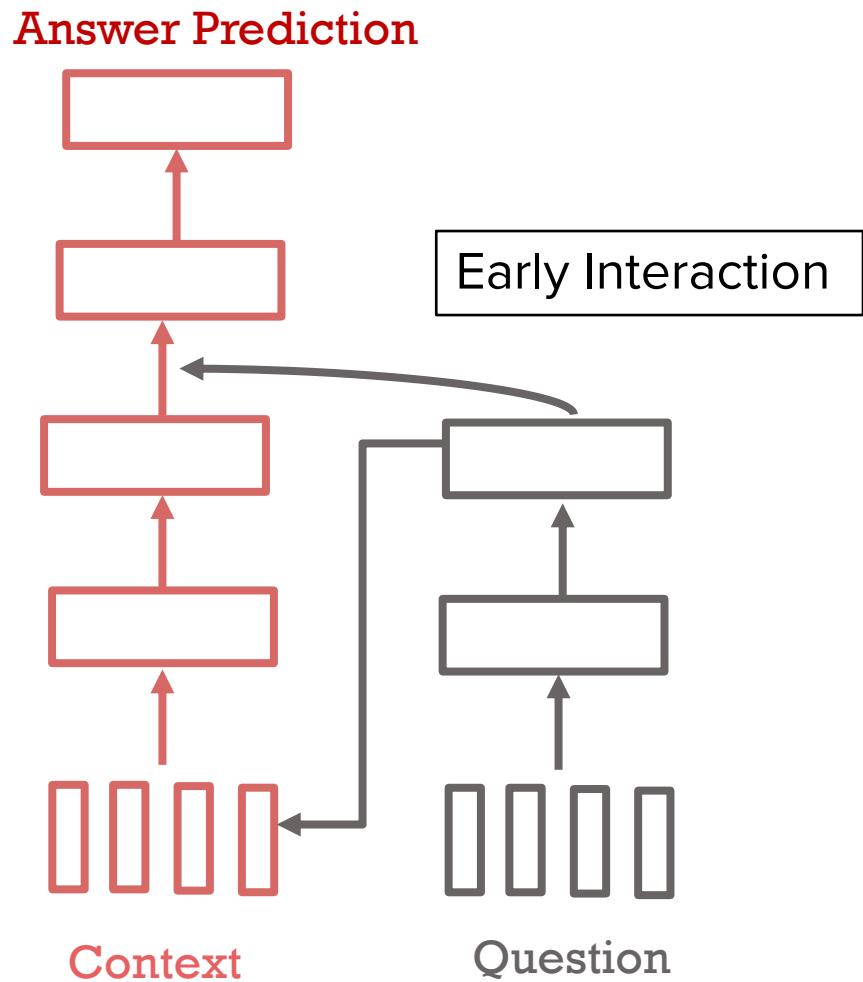


# Late Interaction

- First encode question and context sufficiently
- Choice of encoders
  - Bi-LSTMs
  - Conv Nets
- Most popular Model
  - Bi-directional attention flow [Seo '17]



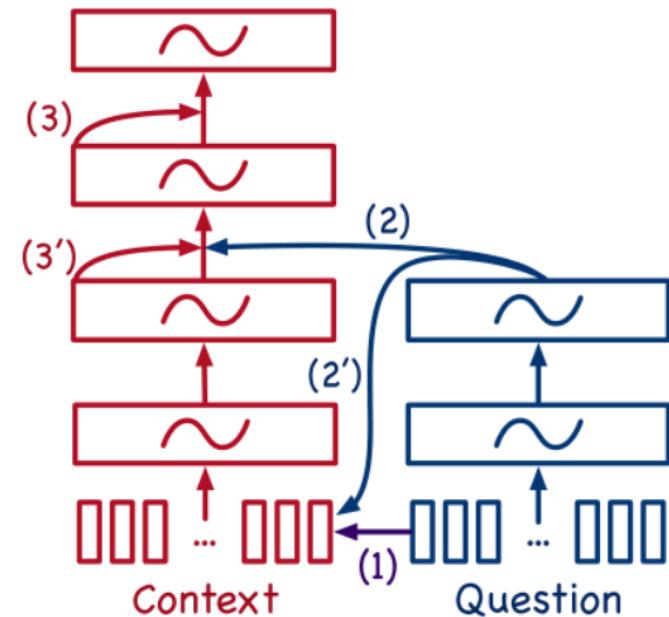
# Other Variants



# Attention Based Architectures

- 2016 – 2017 – Multitude of attention based architecture

Architectures	(1)	(2)	(2')	(3)	(3')
Match-LSTM (Wang and Jiang, 2016)	✓				
DCN (Xiong et al., 2017)		✓		✓	
FastQA (Weissenborn et al., 2017)	✓				
FastQAExt (Weissenborn et al., 2017)	✓	✓		✓	✓
BiDAF (Seo et al., 2017)		✓			
RaSoR (Lee et al., 2016)	✓		✓		
DrQA (Chen et al., 2017)	✓				
MPCM (Wang et al., 2016)	✓	✓			
Mnemonic Reader (Hu et al., 2017)	✓	✓		✓	
R-net (Wang et al., 2017b)		✓		✓	



(1) Word-level fusion, (2) high-level fusion, (2') high-level fusion (alternative), (3) self-boosted fusion, and (3') self-boosted fusion (alternative).

# Contextual Language Models

- BERT – No Recurrence, only attention
- Re-representing each token based on the context
- Shows the most promising performance

Answer Prediction

Transformer block

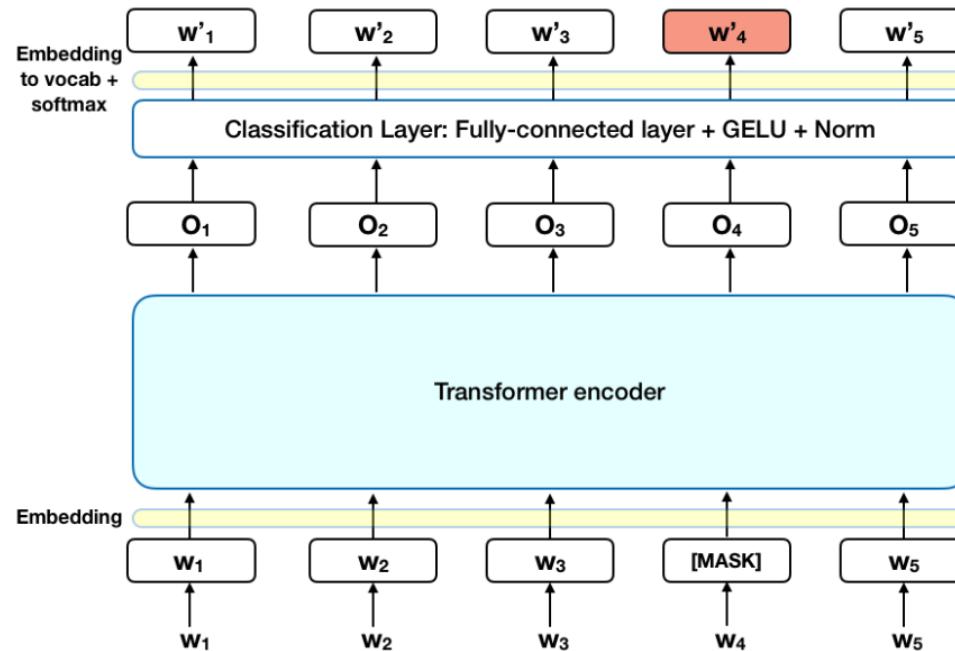
Transformer block

Transformer block

Context      Question

# BERT

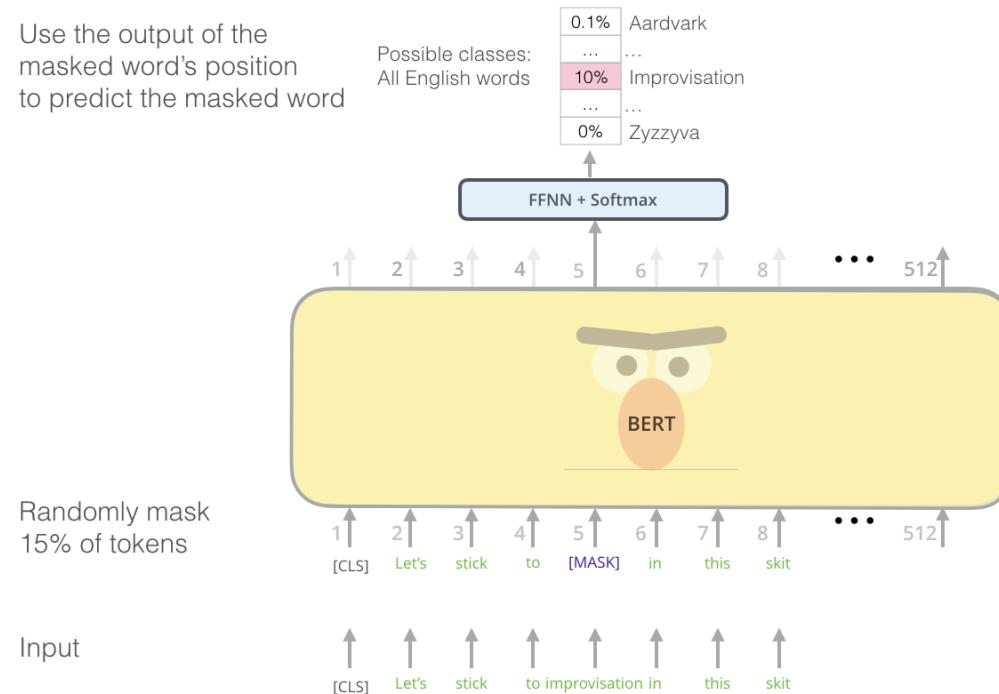
- Bi-directional : Transformer encoder reads the entire sequence of words at once.
  - Learns the context of a word based on all of its surroundings (left and right of the word).



# BERT– Masked Language Model

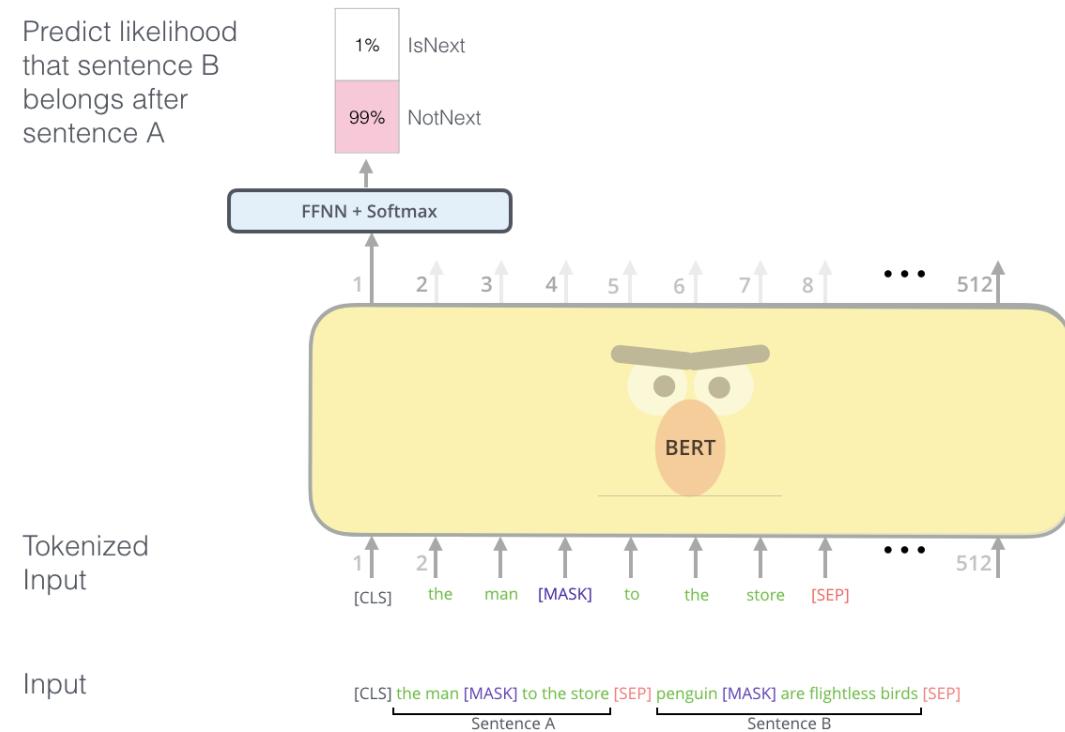
## Masked word prediction

- Given a sentence with some words masked at random, can we predict them?
- Randomly select 15% of tokens to be replaced with “<MASK>”



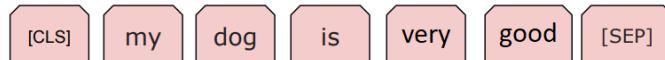
# Next Sentence Prediction

- Given two sentences, does the first follow the second? Teaches BERT about relationship between two sentences
- 50% of the time the actual next sentence, 50% random



# BERT Fine Tuning

- Inputs to BERT – [CLS] <token embeddings> [SEP] ...

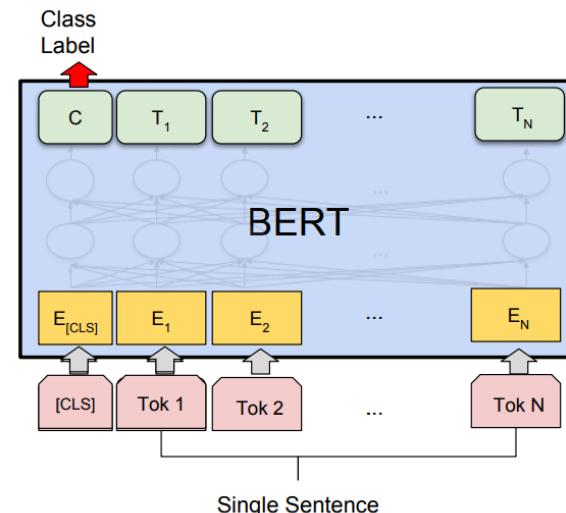


Single sentence input



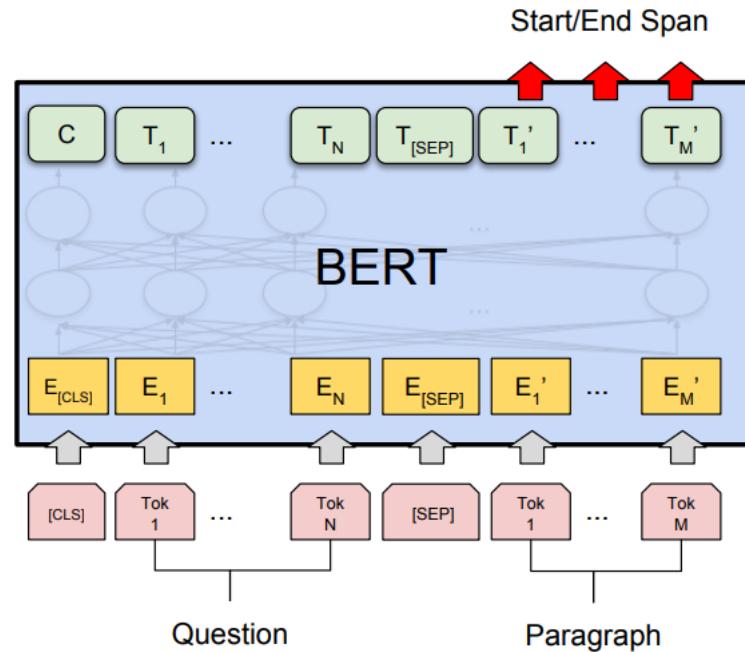
Single sentence input

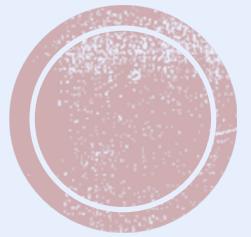
- Classification tasks such as sentiment analysis are done similarly to Next Sentence classification, by adding a classification layer on top of the Transformer output for the [CLS] token.



# BERT Fine Tuning

Q&A model can be trained by learning two extra vectors that mark the beginning and the end of the answer.

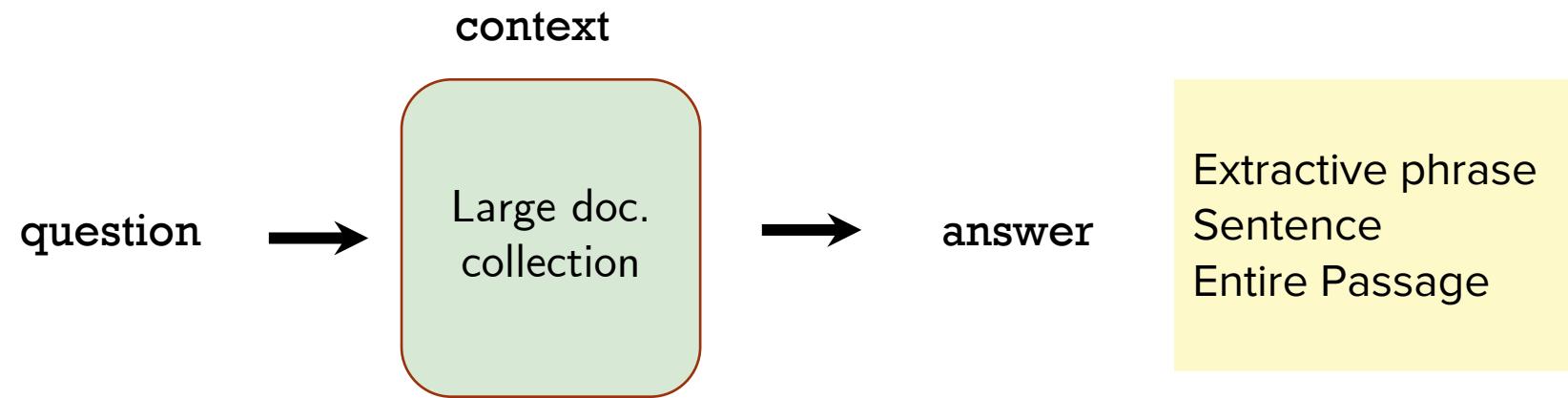




# OPEN-DOMAIN QA

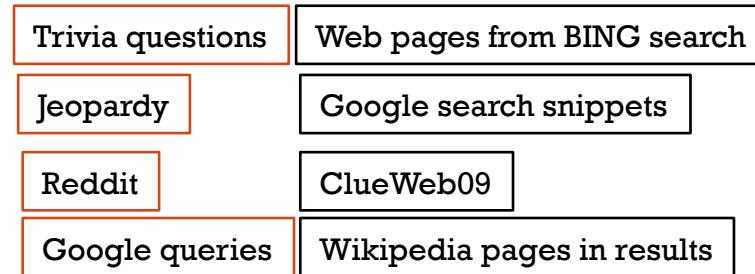


# Problem Setting



# Datasets Commonly Used

- TriviaQA [Joshi et al., 2017]
- SearchQA [Dunn et al., 2017]
- Quasar-T [Dhingra et al., 2017]
- Natural Questions [Kwiatkowski et al., 2019]



Dataset	Train	Val	Test
NQ	79,168	8,757	3,610
WebQ	3,417	361	2,032
TREC	1,353	133	694
TriviaQA	78,785	8,837	11,313
SQuAD	78,713	8,886	10,570

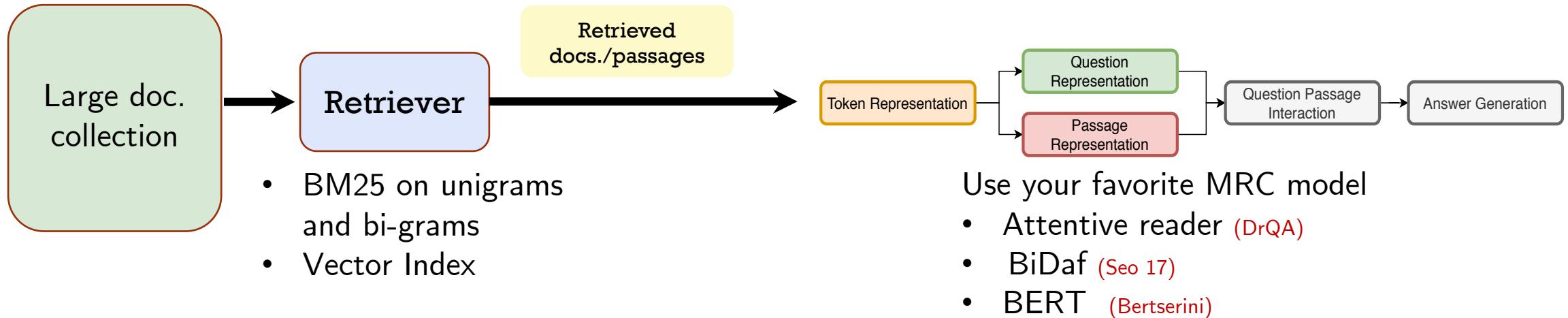
## Repurposed for ODQA

- SQuAD [Rajpurkar et al., 2016]
- CuratedTREC [Baudis & Sedivy, 2015]
- WebQuestions [Berant et al., 2013]
- WikiMovies [Miller et al., 2016]

# Metric Used

- **Exact Match:** measures whether the two strings, after preprocessing, are equal or not.
- **F1 Measure:** measures the overlap between the two bags of tokens in answers, after preprocessing
- **Entity Match**

# Retrieve and Read



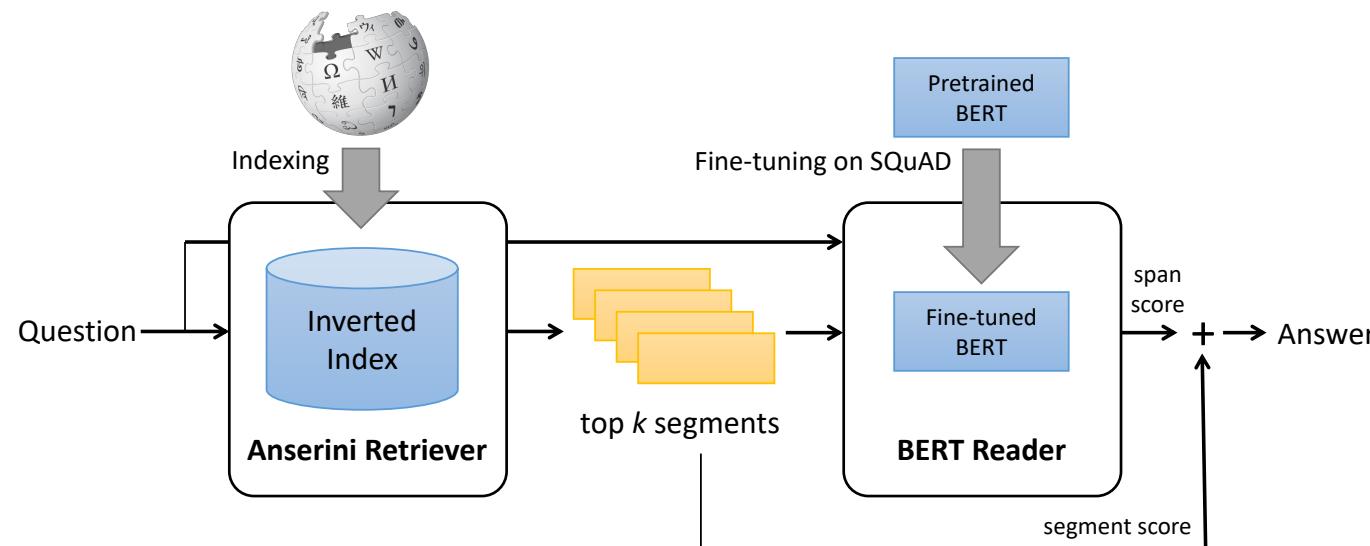
How is the reader model trained ?

Using an existing QA dataset (e.g. SQuAD)

How does it answer questions ?

Independently find answers for token passage and return the most “probable” span

# BERTserini



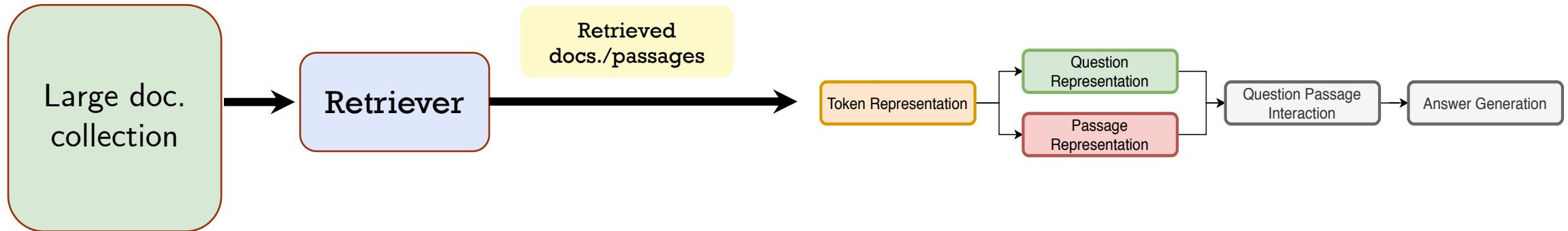
## Retriever

- Using Anserini (based on Lucene)
- Segments = sentence/passage are indexed
- Retrieved sentences are scored using BM25

## Reader

- Fine-tuned BERT on SQUAD
- Final score is interpolation of
  - Span score
  - BM25(segment)

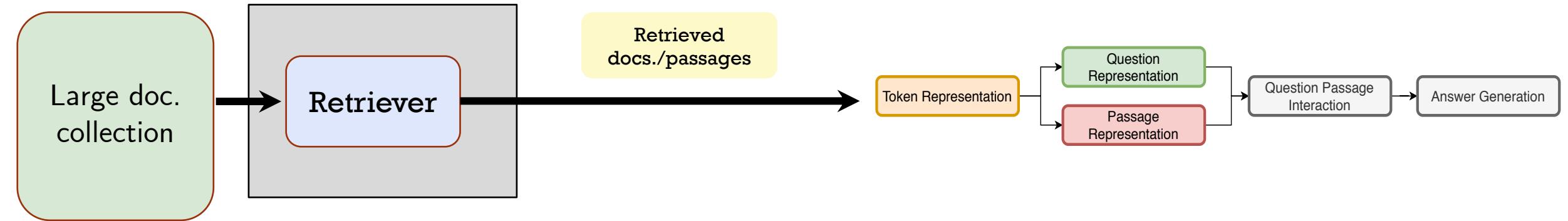
# Design Questions



How do we aggregate evidence in retrieved passages ?

How do exploit the collection for a better reader model ?

How do we exploit reader state to re-retrieve more relevant passages ?



# How Do We Aggregate Evidence In Retrieved Passages ?

# Support

## Question1: What is the more popular name for the londonderry air?

### A1: tune from county

P1: the best known title for this melody is londonderry air - lrb- sometimes also called the **tune from county** derry -rrb- .

### A2: danny boy

P1: londonderry air words : this melody is more commonly known with the words `` **danny boy** ''

P2: londonderry air **danny boy** music ftse london i love you .

P3: **danny boy** limavady is most famous for the tune londonderry air collected by jane ross in the mid-19th century from a local fiddle player .

P4: it was here that jane ross noted down the famous londonderry air -lrb- ` **danny boy** ' -rrb- from a passing fiddler .

# Coverage

**Question2:** Which physicist, mathematician and astronomer discovered the first 4 moons of Jupiter

**A1: Isaac Newton**

**P1:** Sir Isaac Newton was an English physicist , mathematician , astronomer , natural philosopher , alchemist and theologian ...

**P2:** Sir Isaac Newton was an English mathematician, astronomer, and physicist who is widely recognized as one of the most influential scientists ...

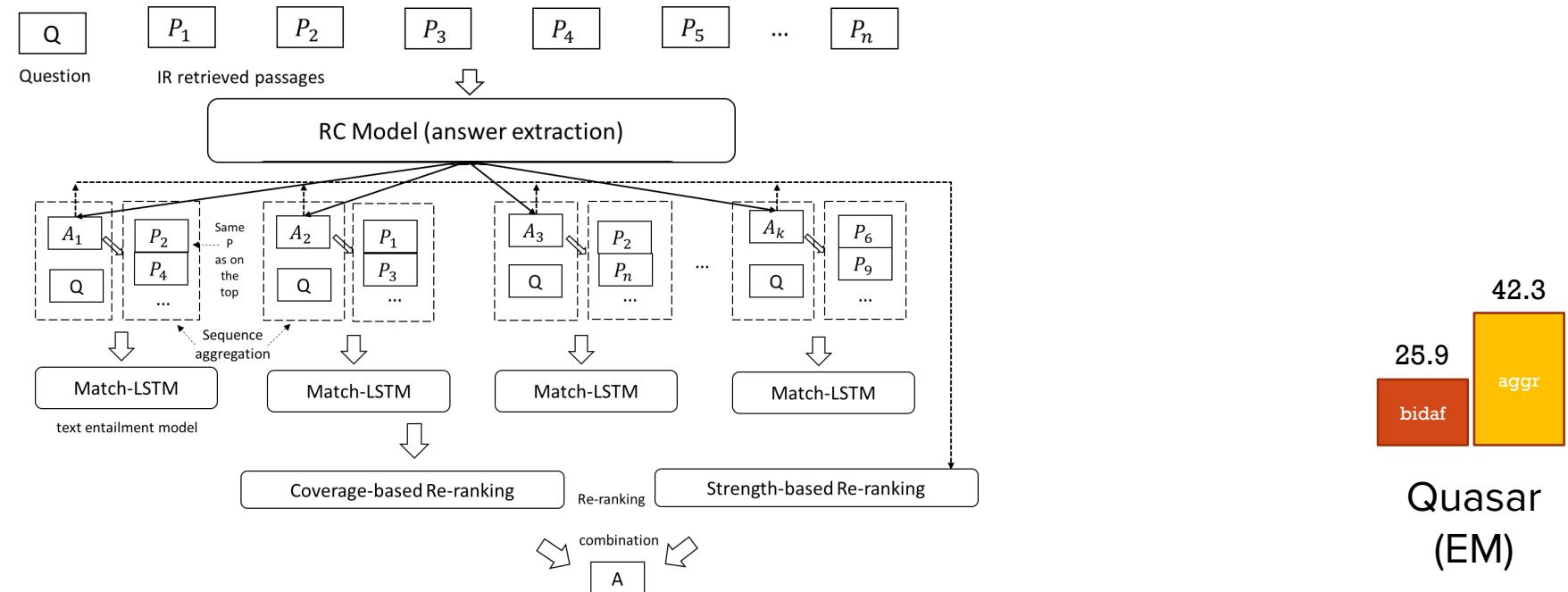
**A2: Galileo Galilei**

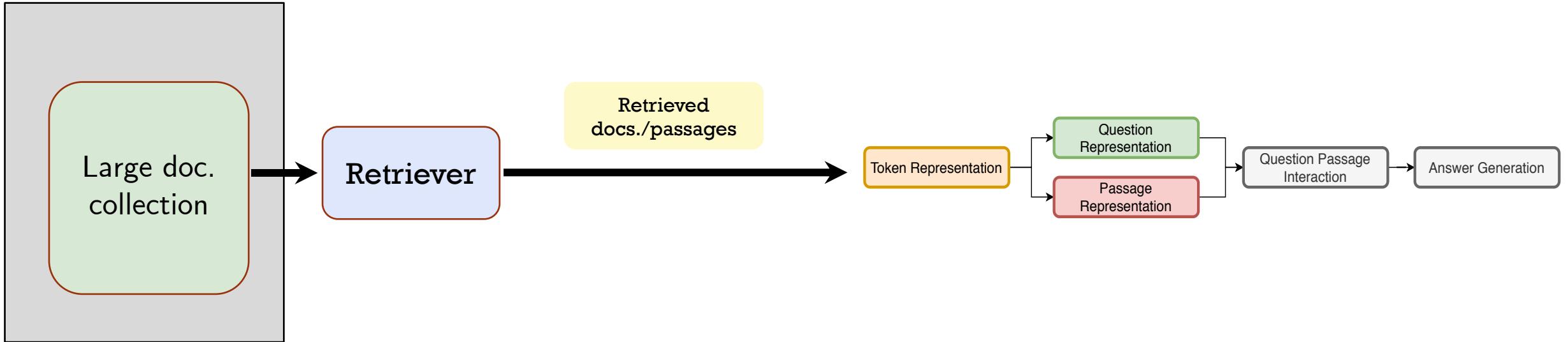
**P1:** Galileo Galilei was an Italian physicist , mathematician , astronomer , and philosopher who played a major role in the Scientific Revolution .

**P2:** Galileo Galilei is credited with discovering the first four moons of Jupiter .

# Support And Coverage

- For each candidate answer, re-rank retrieved passages based on
  - Support – counts
  - Coverage – attention mechanism



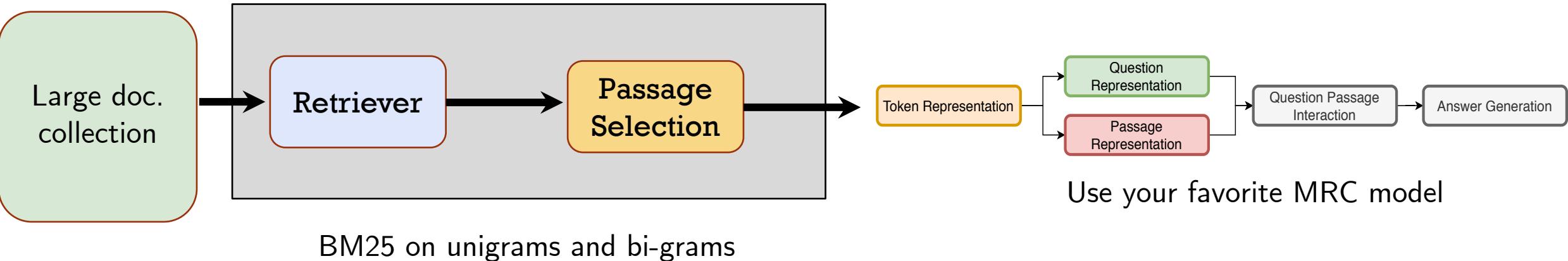


# How do we Exploit Evidence from the collection ?

Extract **Answers** to a given **Question** In the large-scale un-labeled Corpus.

# Distant Supervision

Exploit information about the question that is ignored in retrieved passages



- In MRC training data – (question, passage, answer)
- Distance Supervision [Chen et al. '17]
  - Create extra (question, **passage**, answer) triples
  - Simple Idea: Add all retrieved passages that mention the answer

# Distant Supervision

- Add all retrieved passages that mention the answer
- Which passages to learn from ?
  - **Liberal addition**
    - All passages in the corpus containing answer added
    - All retrieved passages
  - **Restrictive addition**
    - Named entities constraints, passage length limits
- Noise in vanilla DS
  - Noise due to indiscriminate addition [DSQA Model \[Lin et al, '18\]](#)
  - Information loss due to filtered paragraphs [DRQA \[Chen '17\]](#)
  - Noise due to increasing collection sizes and retrieval depth [\[Kratzwald & Feuerriegel '18\]](#)

# Distractors

**Question:** What is the capital of Ireland?

**A: Dublin**

**P1:** As the **capital** of Ireland, **Dublin** is ...

**P2:** **Ireland** is an island in the North Atlantic...

**P3:** **Dublin** is the capital of Ireland. Besides, Ottawa is one of famous tourist cities in Ireland and ...

- Key Idea: Select passages judiciously from the retrieved docs/passages

# Selecting Passages

[Wang et al. '18]

$$\Pr(a|q, P) = \sum_{p_i \in P} \Pr(a|q, p_i) \Pr(p_i|q, P)$$

Likelihood of the answer given a cand. passage

Likelihood of the passage containing the answer

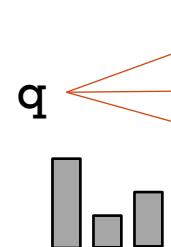
# Selecting Passages

$$\Pr(a|q, p_i) = P_s(a_s)P_e(a_e)$$

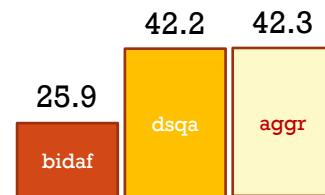
$$\Pr(a|q, P) = \sum_{p_i \in P} \Pr(a|q, p_i) \Pr(p_i|q, P)$$

**Question:** What is the capital of Ireland?

**A: Dublin**



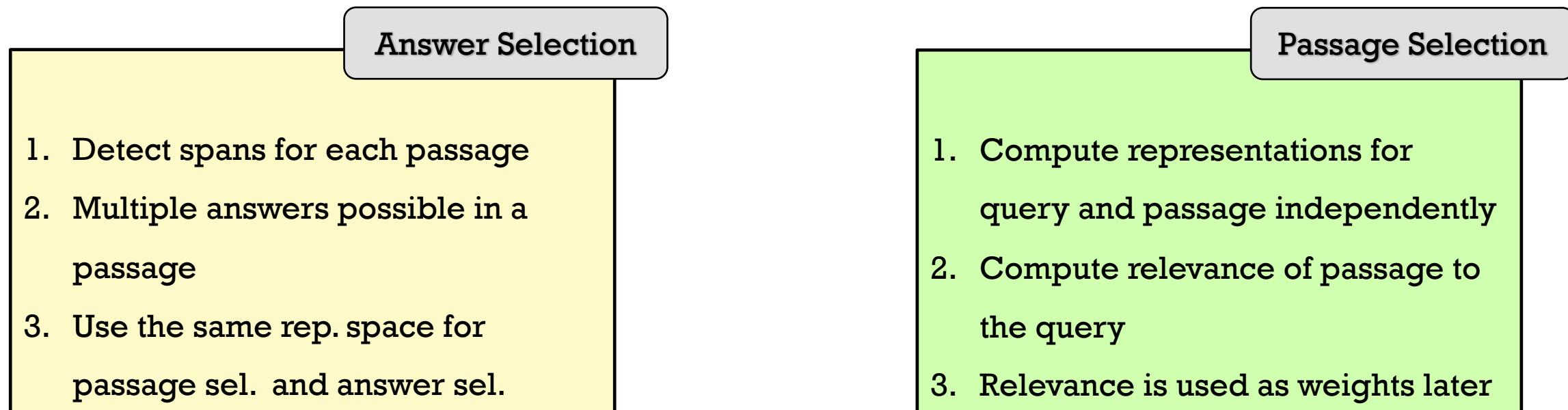
- P1:** As the **capital** of Ireland, **Dublin** is ...
- P2:** **Ireland** is an island in the North Atlantic...
- P3:** **Dublin** is the capital of Ireland. Besides, Ottawa is one of famous tourist cities in Ireland and ...



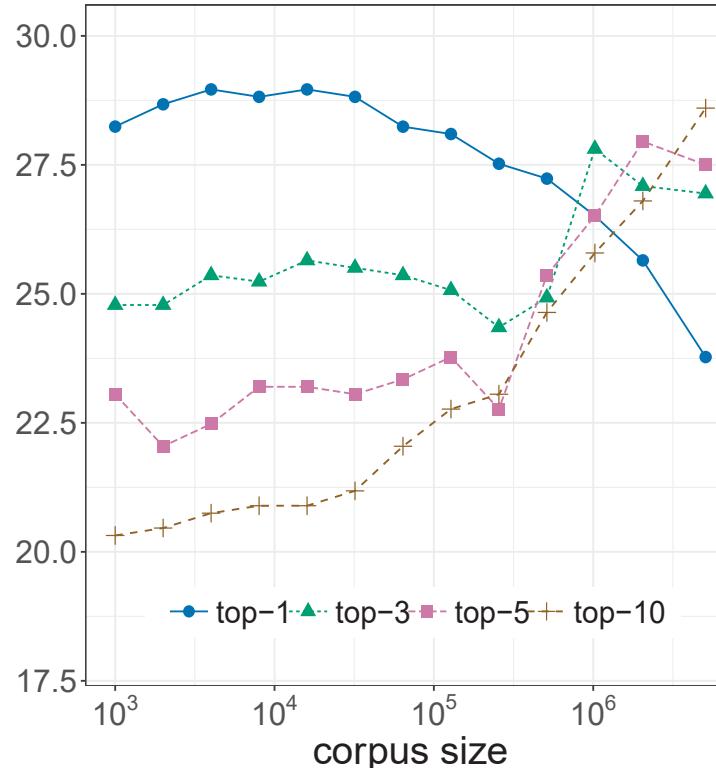
Quasar  
(EM)

# Selecting Passages

$$\Pr(a|q, P) = \sum_{p_i \in P} \Pr(a|q, p_i) \Pr(p_i|q, P)$$



# Retrieval Depth and Collection Size



Large corpus = more noise

Idea: The more confident we are, the less we should retrieve

$$n_i = \max_k \sum_{j=1}^k s_i^{(j)} < \theta$$

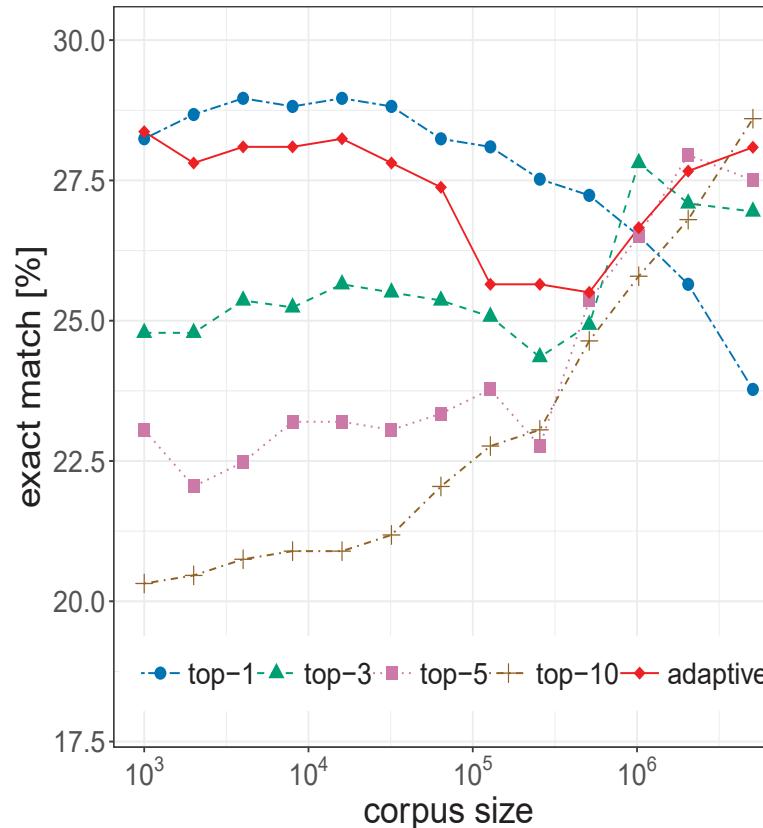
Retrieved doc/passage score

$$s_i = [s_i^{(1)}, \dots, s_i^{(\tau)}]^T$$

$$\sum_j s_i^{(j)} = 1$$

- Choose passages until surpassing a certain confidence threshold
- if document retrieval is certain → selects fewer docs/passages
  - If uncertain → retrieval depth is higher

# Retrieval Depth and Collection Size

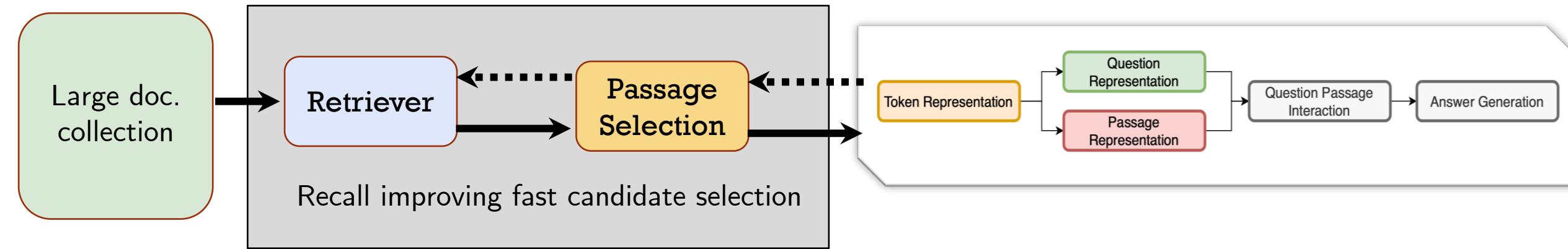


Slightly more involved depth prediction

- Predict the rank of the first relevant document
  - With a small tolerance

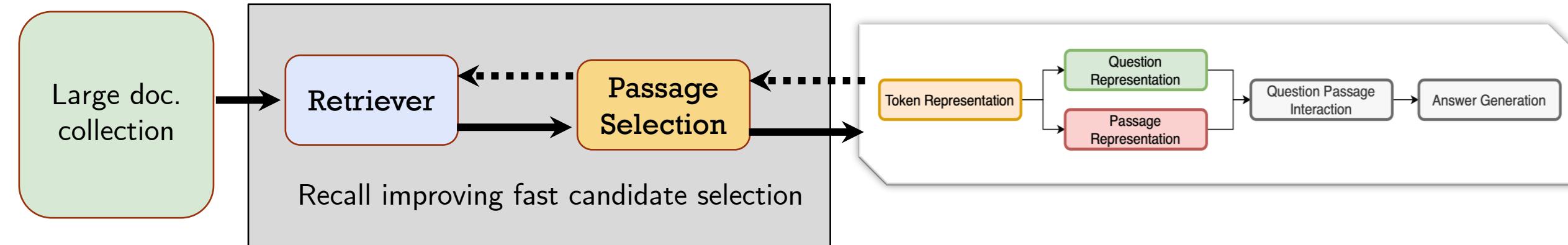
$$n_i = \lceil s_i^T \beta \rceil + b$$

Ret. depth                                    Learnable param.                            tolerance



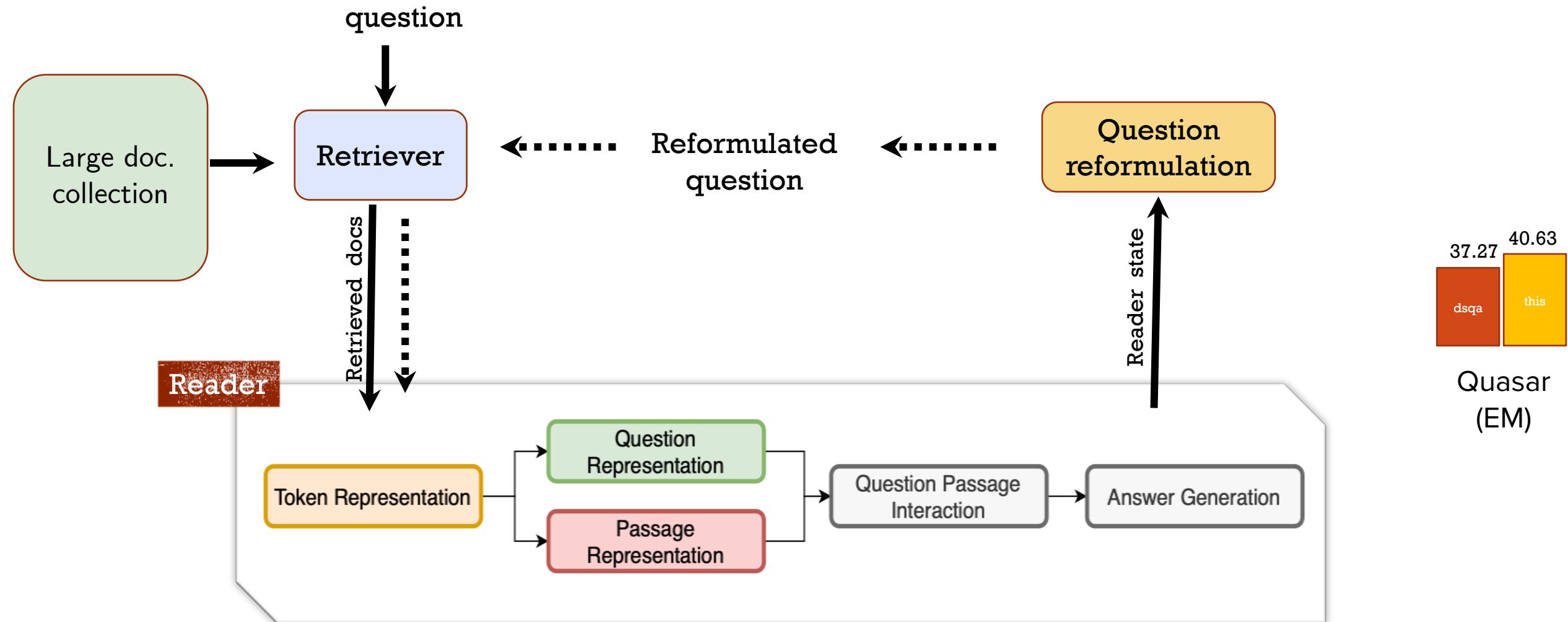
## How we exploit reader feedback for better retrieval ?

# Retriever Reader Interaction



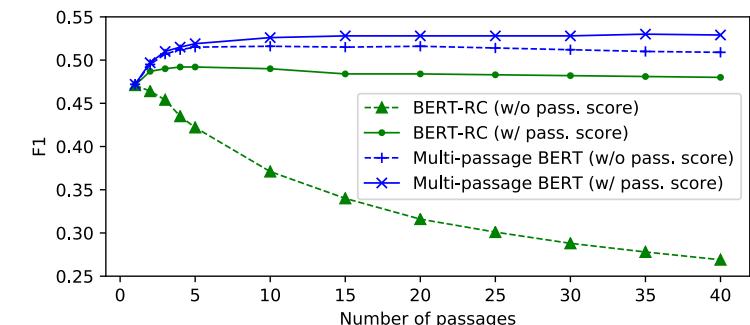
- Single retrieve and read step is limiting – vocabulary gap between question and corpus passages
- How can we enable multi-stage retriever-reader interaction ?
  - Akin to Neural Query Expansion
  - Take care about efficiency concerns

# Retriever Reader Interaction



# Other Notable Approaches

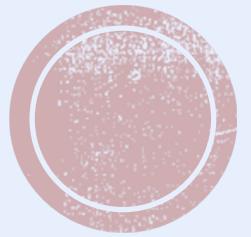
- Document gated reader [Wang et al. '19]
  - Document gating during span prediction
- Tracernet [Dehgani et al '19]
  - Larger contextual models to incorporate reasoning between multiple passages
- R3 [Wang et al '19]
  - Train reader over retrieved docs using the final answer as signal (using REINFORCE)
- Shared Normalization [Clark & Gardner '18, Wang '19]
  - process passages independently, but compute the span probability across spans in all passages in every mini-batch



# Other Notable Approaches

*Instead of an inverted index, use a vector index*

- ORQA [Lee et al '19]
  - Both retriever and reader are learnable (BERT)
- REALM [Wang et al '19]
  - Train reader over retrieved docs using the final answer as signal (using REINFORCE)
- DenSPI [Seo '19]
  - Turns the QA problem into a retrieval problem why sparse encoding of docs and dense indexing of phrases



# INTERPRETABILITY AND FEEDBACK

# Interpretability Landscape

Post-hoc interpretability	
Model Agnostic	By Design
<ul style="list-style-type: none"><li>• LIME</li><li>• SHAP</li><li>• Instance-wise feature selection</li></ul>	<ul style="list-style-type: none"><li>• Linear models</li><li>• Templated models</li><li>• Matching models</li><li>• Analytical sol. possible</li></ul>
<ul style="list-style-type: none"><li>• Simple gradients</li><li>• LRP</li><li>• Integrated gradients</li><li>• Analyzing attention</li></ul>	<ul style="list-style-type: none"><li>• Coarse-to-fine models</li><li>• Generate explanations and task prediction jointly</li></ul>

# Coarse-to-fine Models

## What is the capital of Australia ?

The country's other major metropolitan areas are Melbourne, Brisbane, Perth, and Adelaide. As the seat of the government of Australia, Canberra is home to many important institutions of the federal government, national monuments and museums. Canberra is also the capital of the country.

# Select Sentences as Explanations

**What is the capital of Australia ?**

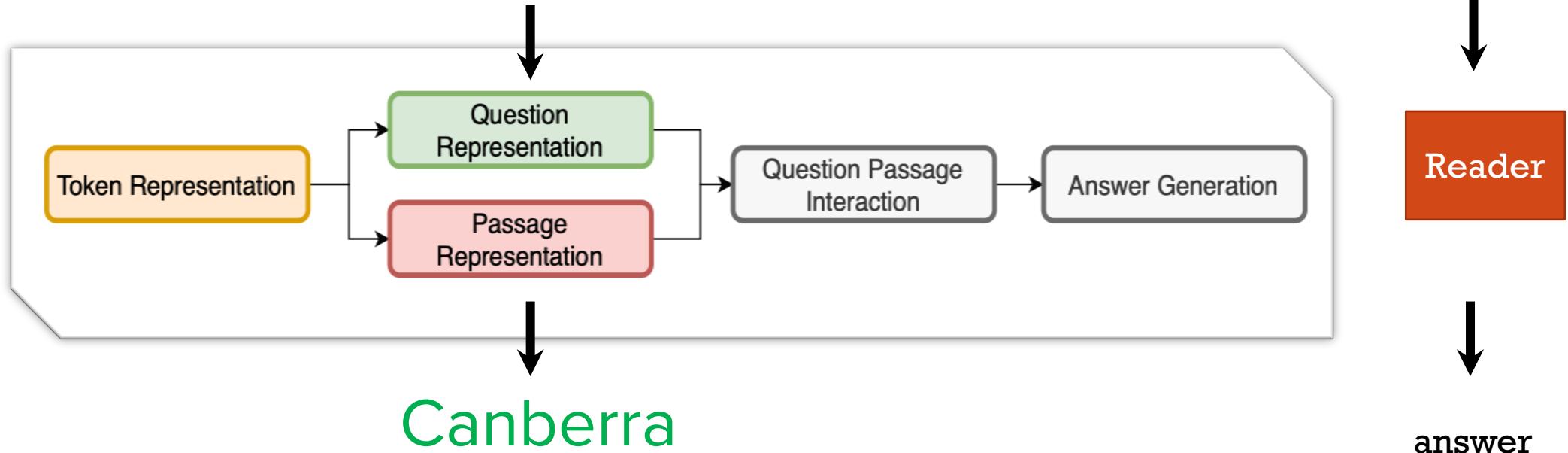
The country's other major metropolitan areas are Melbourne, Brisbane, Perth, and Adelaide. As the seat of the government of Australia, Canberra is home to many important institutions of the federal government, national monuments and museums.

Canberra is also the capital of the country.

# Input to Reader

**What is the capital of Australia ?**

.... Canberra is also the capital of the country.



# Pipelined Models

- Sentence selection and answer predictions are independently trained
- What is the training data for sentence selection ?
  - **Distance supervision**
    - All sentences in the document containing answer is a positive instance
    - First sentence in the document containing the answer
- **Sentence selector** is trained on distantly supervised data
- **Answer predictor** is trained on the actual training data
  - Training data modified to only contain sentences selected from the selection stage

Selector  
network



Reader



answer

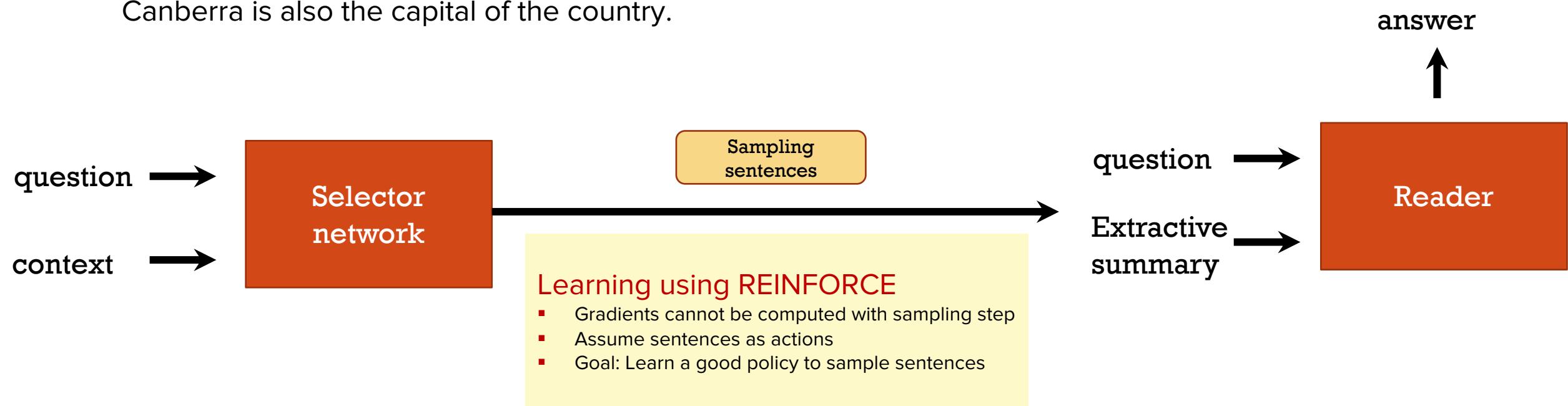
# End-to-end Models

## What is the capital of Australia ?

The country's other major metropolitan areas are Melbourne, Brisbane, Perth, and Adelaide.

As the seat of the government of Australia, Canberra is home to many important institutions of the federal government, national monuments and museums.

Canberra is also the capital of the country.



# User Feedback

- Current systems assume a static collection, static training set
- In an online systems
  - Users continuously issue queries, provide implicit feedback
- How can we construct a continuously learning system from explicit user feedback ?
  - How do we use the feedback to update training set ?
  - Can we reconcile noisy and sometimes erroneous feedback ?

Which vintage rock and roll singer was known as "The Killer"?

Search

## Farrokh Bulsara

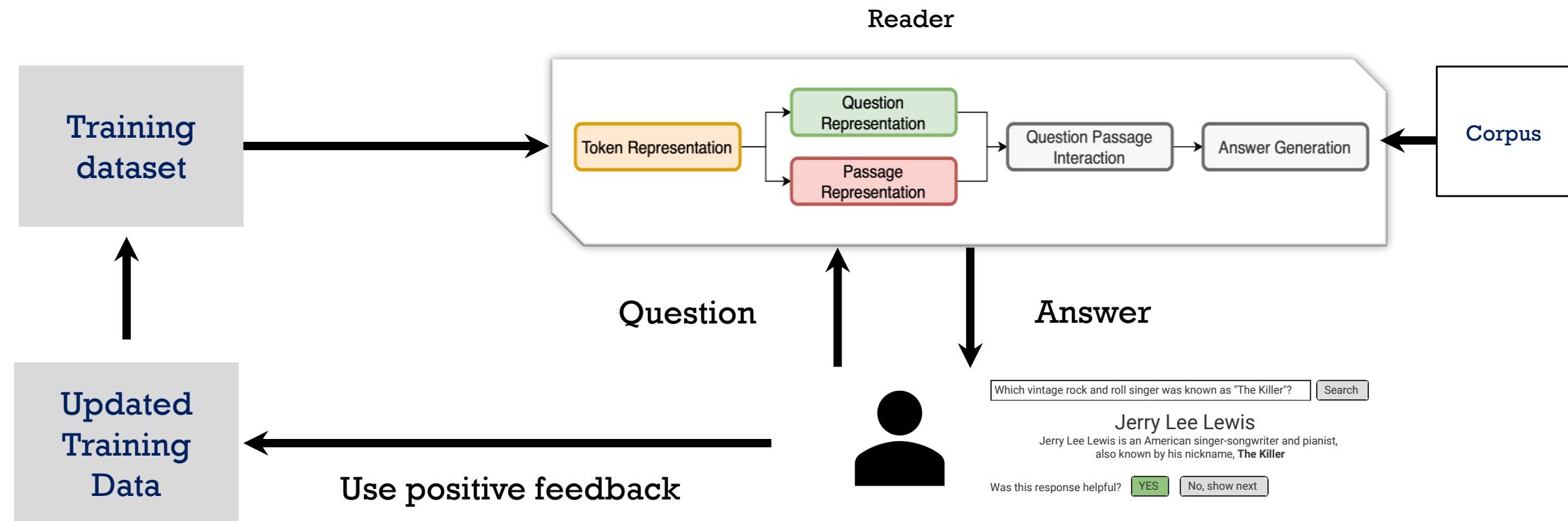
**Farrokh Bulsara**, known professionally as Freddie Mercury,  
was a British singer, songwriter and record producer...

Was this response helpful?

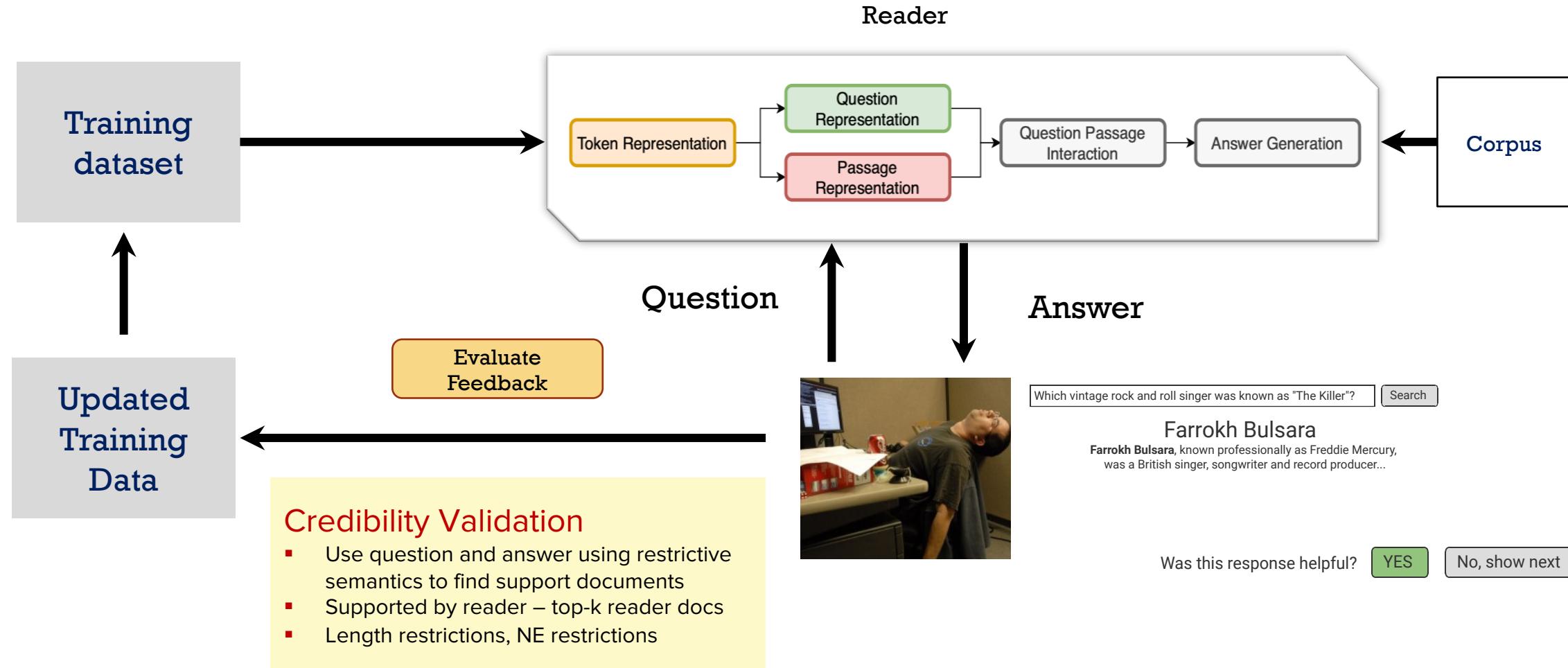
YES

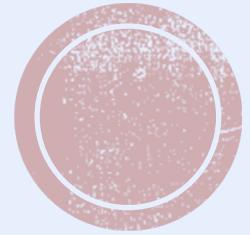
No, show next

# Updating Training Set



# Credibility Validation





# CONVERSATIONAL QA



# Conversational Question Answering

- Questions and answers in free-form text
- Different forms, different challenges
  - Chit Chat
  - Multi-turn QA
  - Clarifications
- Different from MRC:
  - Isolated vs contextual
    - Question lengths: shorter for conversational QA datasets (contextual)

# Conversational Question Answering (CoQA)



- Multi-turn conversation, each turn is a question and an answer
- Questions and answers in free-form text
- Conversation is grounded in Passage
  - Concrete eval unlike chit-chat
- 127,000 questions and answers
  - 8K conversations (avg. 15 turns)
  - 7 diverse domains
    - Children stories, literature, exams, cnn news, Wikipedia
    - Hidden domains : reddit, science

# CoQA Dataset

The Virginia governor's race, billed as the marquee battle of an otherwise anticlimactic 2013 election cycle, is shaping up to be a foregone conclusion. Democrat Terry McAuliffe, the longtime political fixer and moneymen, hasn't trailed in a poll since May. Barring a political miracle, Republican Ken Cuccinelli will be delivering a concession speech on Tuesday evening in Richmond. In recent ...

Q<sub>1</sub>: What are the candidates **running** for?

A<sub>1</sub>: Governor, R<sub>1</sub>: The Virginia governor's race

Q<sub>2</sub>: **Where**?

A<sub>2</sub>: Virginia, R<sub>2</sub>: The Virginia governor's race

Q<sub>3</sub>: Who is the democratic candidate?

A<sub>3</sub>: **Terry McAuliffe**, R<sub>3</sub>: Democrat Terry McAuliffe

Q<sub>4</sub>: Who is **his** opponent?

A<sub>4</sub>: **Ken Cuccinelli**, R<sub>4</sub>: Republican Ken Cuccinelli

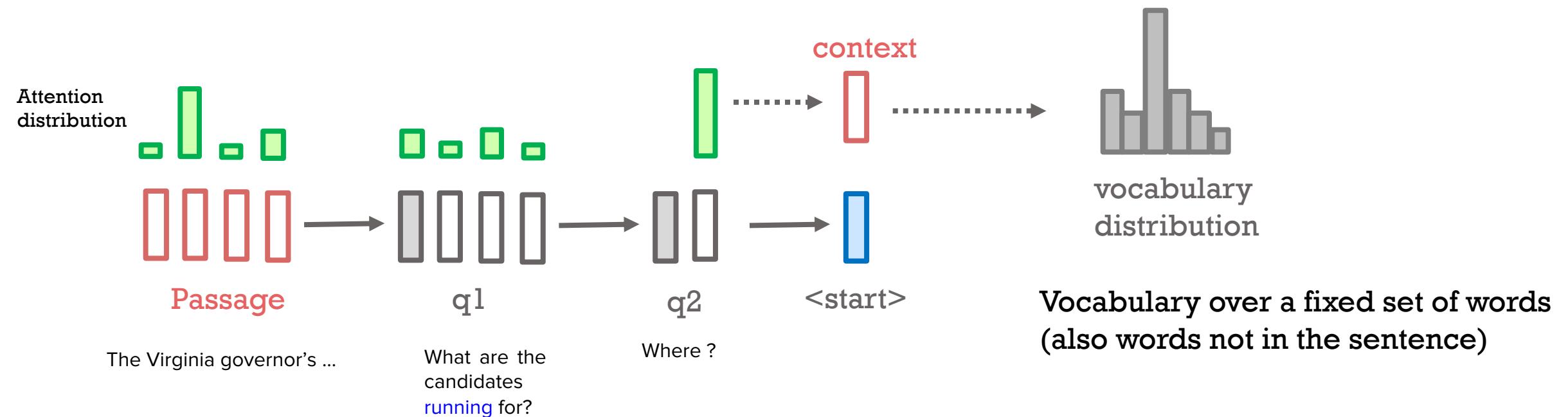
Q<sub>5</sub>: What party does **he** belong to?

A<sub>5</sub>: Republican, R<sub>5</sub>: Republican Ken Cuccinelli

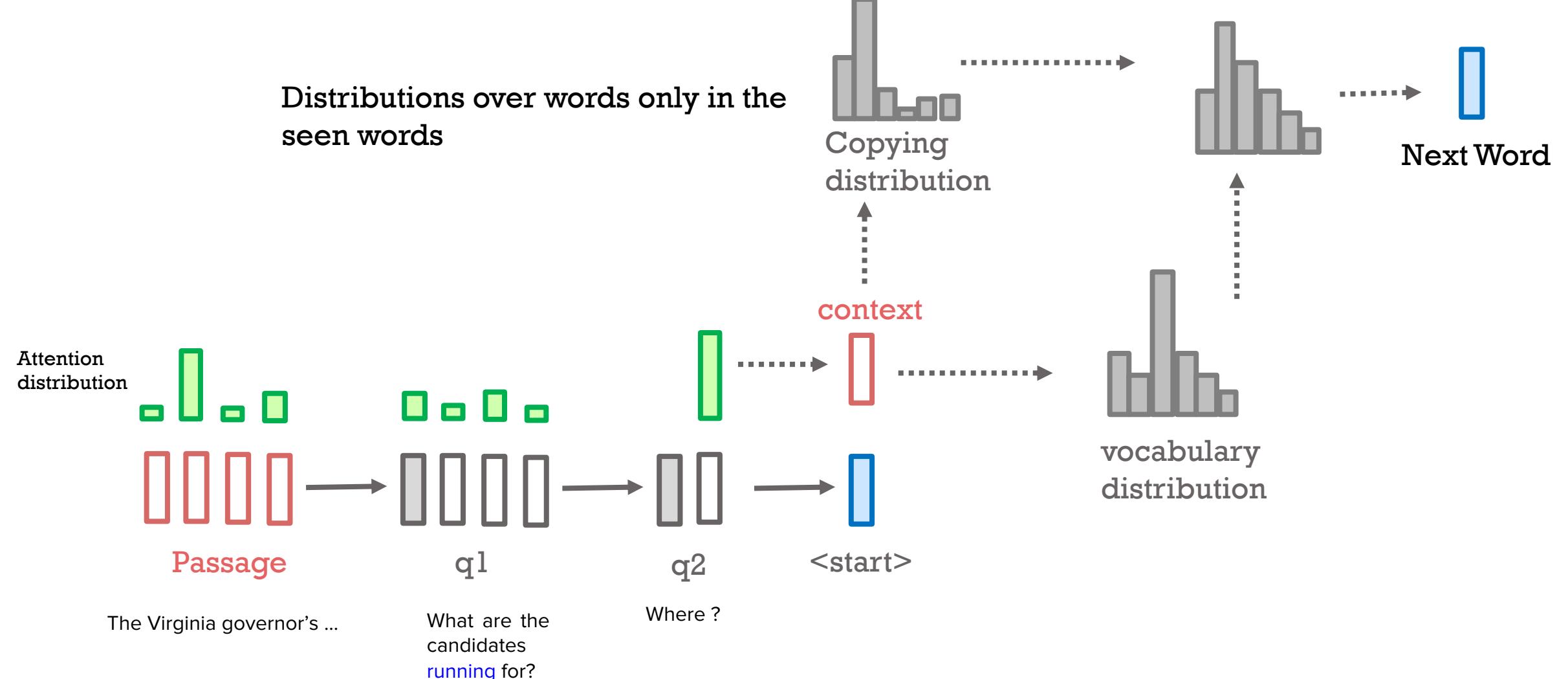
Q<sub>6</sub>: Which of **them** is winning?

A<sub>6</sub>: **Terry McAuliffe**, R<sub>6</sub>: Democrat Terry McAuliffe, the longtime political fixer and moneymen, hasn't trailed in a poll since May

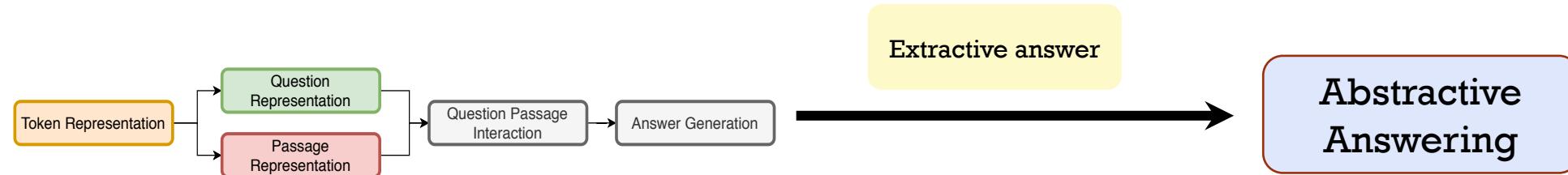
# Seq2Seq Abstractive Response Generation



# PGNet



# Hybrid Model



Type	Seq2seq	PGNet	DrQA	Augmt. DrQA	DrQA+ PGNet	Human PGNet
Answer Type						
Answerable	27.5	45.4	54.7	67.3	66.3	89.9
Unanswerable	33.9	38.2	55.0	49.1	51.2	72.3
Span found	20.2	43.6	69.8	71.0	70.5	91.1
No span found	43.1	49.0	22.7	59.4	57.0	86.8
Named Entity	21.9	43.0	72.6	73.5	72.2	92.2
Noun Phrase	17.2	37.2	64.9	65.3	64.1	88.6
Yes	69.6	69.9	7.9	75.7	72.7	95.6
No	60.2	60.3	18.4	59.6	58.7	95.7
Number	15.0	48.6	66.3	69.0	71.7	91.2
Date/Time	13.7	50.2	79.0	83.3	79.1	91.5
Other	14.1	33.7	53.5	55.6	55.2	80.8

**Seq2seq  
Model**

# Leaderboard Sneakpeak

Rank	Model	In-domain	Out-of-domain	Overall					
	Human Performance <i>Stanford University</i> <b>(Reddy &amp; Chen et al. TACL '19)</b>	89.4	87.4	88.8					
1	RoBERTa + AT + KD (ensemble) <i>Zhuiyi Technology</i> <a href="https://arxiv.org/abs/1909.10772">https://arxiv.org/abs/1909.10772</a>	91.4	<b>89.2</b>	<b>90.7</b>					
1	TR-MT (ensemble) <i>WeChatAI</i>	<b>91.5</b>	88.8	<b>90.7</b>					
2	RoBERTa + AT + KD (single model) <i>Zhuiyi Technology</i> <a href="https://arxiv.org/abs/1909.10772">https://arxiv.org/abs/1909.10772</a>	90.9	<b>89.2</b>	90.4	39	DrQA + seq2seq with copy attention (single model) <i>Stanford University</i> <a href="https://arxiv.org/abs/1808.07042">https://arxiv.org/abs/1808.07042</a>	67.0	60.4	65.1
3	TR-MT (ensemble) <i>WeChatAI</i>	91.1	87.9	90.2					
4	Google SQuAD 2.0 + MMFT (ensemble) <i>MSRA + SDRG</i>	89.9	88.0	89.4					
5	TR-MT (single model) <i>WeChatAI</i>	90.4	86.8	89.3					
6	XLNet + Augmentation (single model) <i>Xiaoming</i> <a href="https://github.com/stevezheng23/xl_net_extension_tf">https://github.com/stevezheng23/xl_net_extension_tf</a>	89.9	86.9	89.0					

# Other Conversational Datasets

## QuAC [Choi '19]

- Simulating info. seeking dialog
  - About a Wikipedia text
- 11k Dialogs, 98K QA Pairs
- Simple evaluation

## QuLAC [Aliannejadi '19]

- Clarifying questions in info. seeking conversations
- Open domain, IR setting
- 198 topics [TREC Web Track]

**Section:**  Daffy Duck, Origin & History

**STUDENT:** What is the origin of Daffy Duck?

**TEACHER:** ↗ first appeared in Porky's Duck Hunt

**STUDENT:** What was he like in that episode?

**TEACHER:** ↗ assertive, unrestrained, combative

**STUDENT:** Was he the star?

**TEACHER:** ↗ No, barely more than an unnamed bit player in this short

**STUDENT:** Who was the star?

**TEACHER:** ↗ No answer

**STUDENT:** Did he change a lot from that first episode in future episodes?

**TEACHER:** ↗ Yes, the only aspects of the character that have remained consistent (...) are his voice characterization by Mel Blanc

**STUDENT:** How has he changed?

**TEACHER:** ↗ Daffy was less anthropomorphic

**STUDENT:** In what other ways did he change?

**TEACHER:** ↗ Daffy's slobbery, exaggerated lisp (...) is barely noticeable in the early cartoons.

**STUDENT:** Why did they add the lisp?

**TEACHER:** ↗ One often-repeated "official" story is that it was modeled after producer Leon Schlesinger's tendency to lisp.

**STUDENT:** Is there an "unofficial" story?

**TEACHER:** ↗ Yes, Mel Blanc (...) contradicts that conventional belief

...



**Information Need (Facet)**

I'm looking for the Discovery Channel's dinosaur site, which has pictures of dinosaurs and games.

 Are you looking for dinosaur books?

No, just the discovery channel website.

 Are you looking for meat-eating or plant-eating dinosaurs?

I'm not sure.

 Would you like to see pictures or videos of dinosaurs?

I'd like to see pictures of dinosaurs on the discovery channels website.

**Others:** CSQA (Saha et al., 2018) CQA (Talmor and Berant, 2018) SQA (Iyyer et al., 2017)

# Some papers this SIGIR...

## Analyzing and Learning from User Interactions for Search Clarification

Hamed Zamani, Bhaskar Mitra, Everest Chen, Gord Lueck, Fernando Diaz, Paul N. Bennett,  
Nick Craswell, and Susan T. Dumais

Microsoft

[hazamani.bmitra.vvchen.gordonl.fdiaz.pnben.nickcr.sdumai@microsoft.com](mailto:hazamani.bmitra.vvchen.gordonl.fdiaz.pnben.nickcr.sdumai@microsoft.com)

## Open-Retrieval Conversational Question Answering

Chen Qu<sup>1</sup> Liu Yang<sup>1</sup> Cen Chen<sup>2</sup> Minghui Qiu<sup>3</sup> W. Bruce Croft<sup>1</sup> Mohit Iyyer<sup>1</sup>

<sup>1</sup>University of Massachusetts Amherst <sup>2</sup>Ant Financial <sup>3</sup>Alibaba Group

{chenqu,lyang,croft,miyyer}@cs.umass.edu,chencen.cc@antfin.com,minghui.qmh@alibaba-inc.com

## Query Resolution for Conversational Search with Limited Supervision

Nikos Voskarides<sup>1</sup> Dan Li<sup>1</sup> Pengjie Ren<sup>1</sup> Evangelos Kanoulas<sup>1</sup> Maarten de Rijke<sup>1,2</sup>

<sup>1</sup>University of Amsterdam, Amsterdam, The Netherlands <sup>2</sup>Ahold Delhaize, Zaandam, The Netherlands

[nickvosk@gmail.com,d.li@uva.nl,p.ren@uva.nl,e.kanoulas@uva.nl,m.derijke@uva.nl](mailto:nickvosk@gmail.com,d.li@uva.nl,p.ren@uva.nl,e.kanoulas@uva.nl,m.derijke@uva.nl)

# Lessons learnt

- Contextual representations for text go a long way
- Using sparse training data in open-domain QA is important
- Understanding your dataset is important
  - Aggregation
  - Multi-step reasoning
- Anecdotal success and failure cases extremely valuable
- Training neural models is an art and science in itself

# How to get started

- Download your dataset of choice SQuAD, MSMarco, COQA
- Implement simplest QA system that you can think of
- Examine failure cases, analyse errors, get to know your datasets
- Reimplement recent method of choice: Is it perfect?
- Time for your own research!
  - Leaderboarding is valuable but not always reflective of true improvements

# Open problems

- Efficiency
  - Open-domain QA at scale – recent advances but lots to discover
- Interpretability
  - How can you go beyond feature attributions, selections
- Interactivity
  - Multiple interaction paradigms – training and inference settings
- Robustness

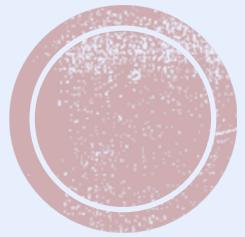
# Conclusions

- QA over text ...
- Text corpora are noisy but have more information coverage and redundancy
- Efficiency and scalability in open-domain QA is a challenge
- “Explainability” is important but often overlooked
- Conversational Search is upcoming and has some crucial challenges

Thank  
you

99

THANK YOU



# REFERENCES

# TEXT REFERENCES – DATASETS

- [31] Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In NAACL-HLT.
- [49] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redeld, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, and Kenton Lee. 2019. Natural questions: A benchmark for question answering research. TACL 7 (2019).
- [63] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In EMNLP.
- [66] Siva Reddy, Danqi Chen, and Christopher Manning. 2019. CoQA: A conversational question answering challenge. TACL 7 (2019).
- [92] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In EMNLP.
- [26] Pradeep Dasigi, Nelson F Liu, Ana Marasovic, Noah A Smith, and Matt Gardner. 2019. Quoref: A reading comprehension dataset with questions requiring coreferential reasoning. In EMNLP-IJCNLP.
- [18] Eunsol Choi, HeHe, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. QuAC: Question answering in context. In EMNLP 2018.



# TEXT REFERENCES – ODQA+ RC

- [65] Deepak Ravichandran and Eduard Hovy. 2002.  
Learning surface text patterns for a question answering system. In ACL
- [75] H. Sun, H. Ma, W. Yih, C. Tsai, J. Liu, and M. Chang. 2015.  
Open domain question answering via semantic enrichment. In WWW.
- [19] Christopher Clark and Matt Gardner. 2018.  
Simple and Effective Multi-Paragraph Reading Comprehension. In ACL.
- [27] Mostafa Dehghani, Hosein Azarbonyad, Jaap Kamps, and Maarten de Rijke. 2019.  
Learning to transform, combine, and reason in open-domain question answering. In WSDM.
- [53] Yankai Lin, Haozhe Ji, Zhiyuan Liu, and Maosong Sun. 2018.  
Denoising distantly supervised open-domain question answering. In ACL.
- [XX] Bingning Wang, Ting Yao, Qi Zhang, Jingfang Xu, Zhixing Tian, Kang Liu, and Jun Zhao. 2019.  
Document Gated Reader for Open-Domain Question Answering. In SIGIR.
- [78] Alon Talmor and Jonathan Berant. 2019.  
MultiQA: An empirical investigation of generalization and transfer in reading comprehension. In ACL.



# FEEDBACK, CLARIFICATION, INTERPRETABILITY

- [14] Pavel Braslavski, Denis Savenkov, Eugene Agichtein, and Alina Dubatovka. 2017.  
What do you mean exactly? Analyzing clarification questions in CQA. In CHIIR.
- [48] Bernhard Kratzwald and Stefan Feuerriegel. 2019.  
Learning from on-line user feedback in neural question answering on the web. In WWW.
- [64] Sudha Rao and Hal Daumé III. 2018. Learning to ask good questions: Ranking clarification questions using neural expected value of perfect information. In ACL.
- [67] Marzieh Saeidi, Max Bartolo, Patrick Lewis, Sameer Singh, Tim Rocktäschel, Mike Sheldon, Guillaume Bouchard, and Sebastian Riedel. 2018.  
Interpretation of Natural Language Rules in Conversational Machine Reading. In EMNLP.
- [87] Jingjing Xu, Yuechen Wang, Duyu Tang, Nan Duan, Pengcheng Yang, Qi Zeng, Ming Zhou, and Xu Sun. 2019.  
Asking Clarification Questions in Knowledge- Based Question Answering. In EMNLP-IJCNLP.
- [96] Xinbo Zhang, Lei Zou, and Sen Hu. 2019.  
An Interactive Mechanism to Improve Question Answering Systems via Feedback. In CIKM.

