Question Answering over Curated and Open Web Sources

Rishiraj Saha Roy

Max Planck Institute for Informatics, Germany

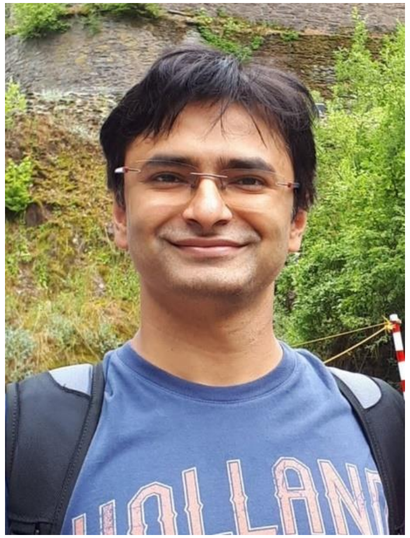
Avishek Anand

L3S Research Center, Germany

SIGIR 2020 Tutorial

26 July 2020

Who are we?



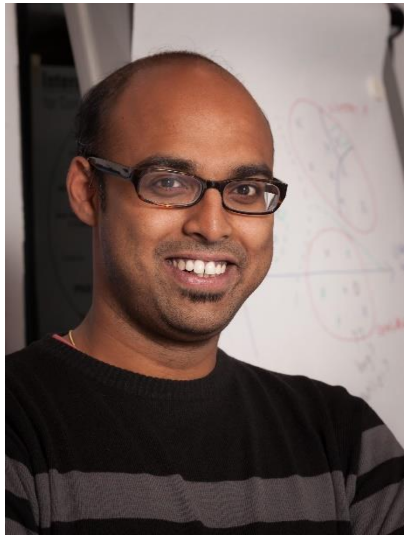
Rishiraj Saha Roy

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
https://www.avishekanand.com/talk/sigir20-tute/

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

Focus on a few instantiations for each method

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

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

Hey Siri

Hi. I'm Cortana.
Ask me a question!

amazon alexa

Question Answering: Vital for Search

What are some films directed by Nolan?

Google Assistant

Hey Siri

Hi. I'm Cortana.
Ask me a question!

amazon alexa

Christopher Nolan / Films directed

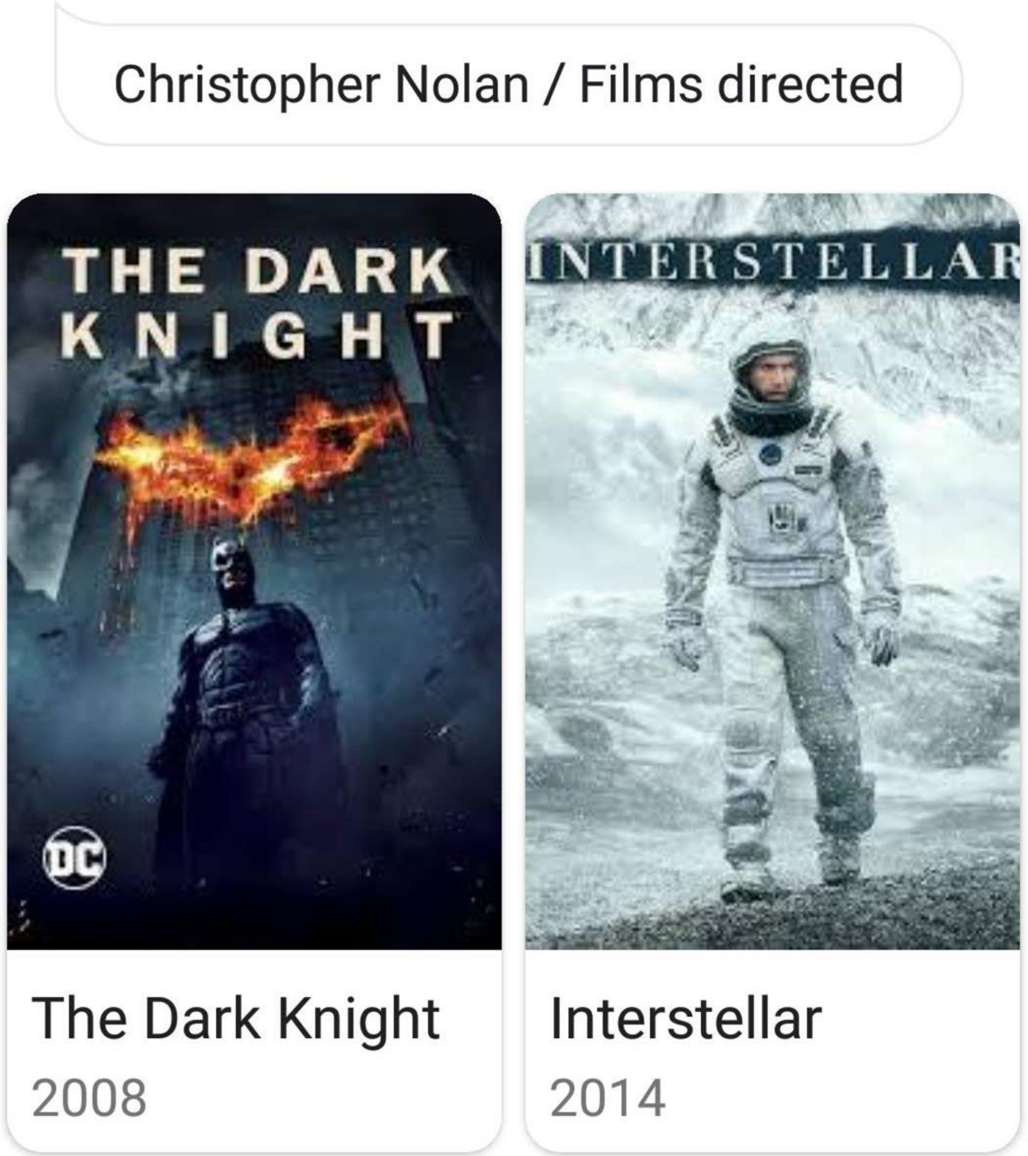
The Dark Knight
2008

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



what is the capital of Germany

which club does Neymar play for

NEYMAR / CURRENT TEAM

Paris Saint-Germain F.

what is the currency of Denmark

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 SHOPPIN

SIGIR 2020 / Location

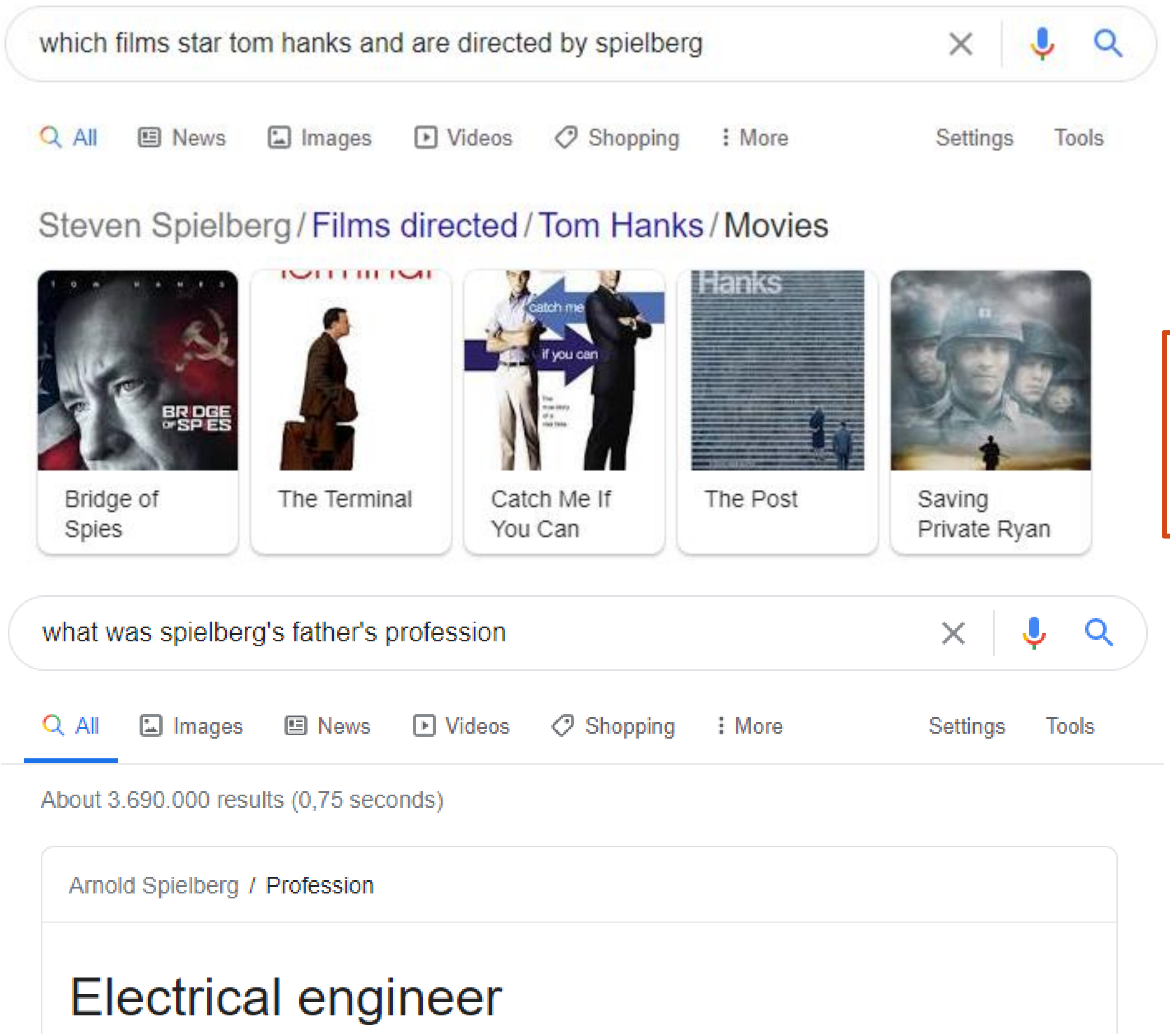
SIGIR 2020
XI'AN·CHINA

SIGIR
ecial Interest Grou
Information Retrie

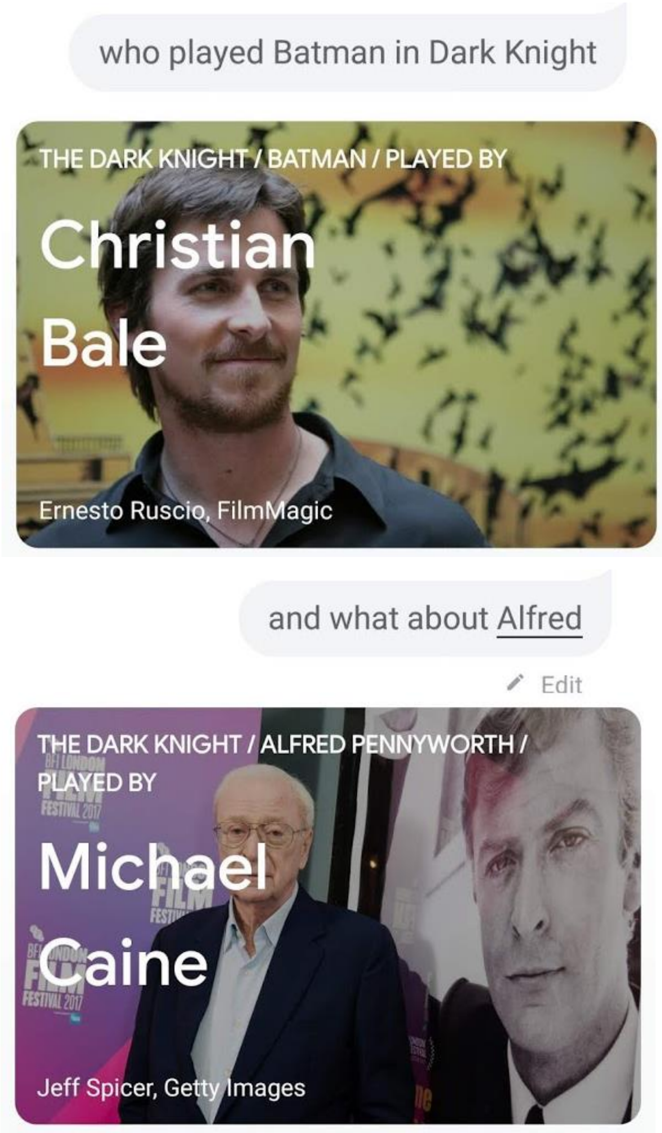
Xi'an, China

Simple questions involving one entity and relation

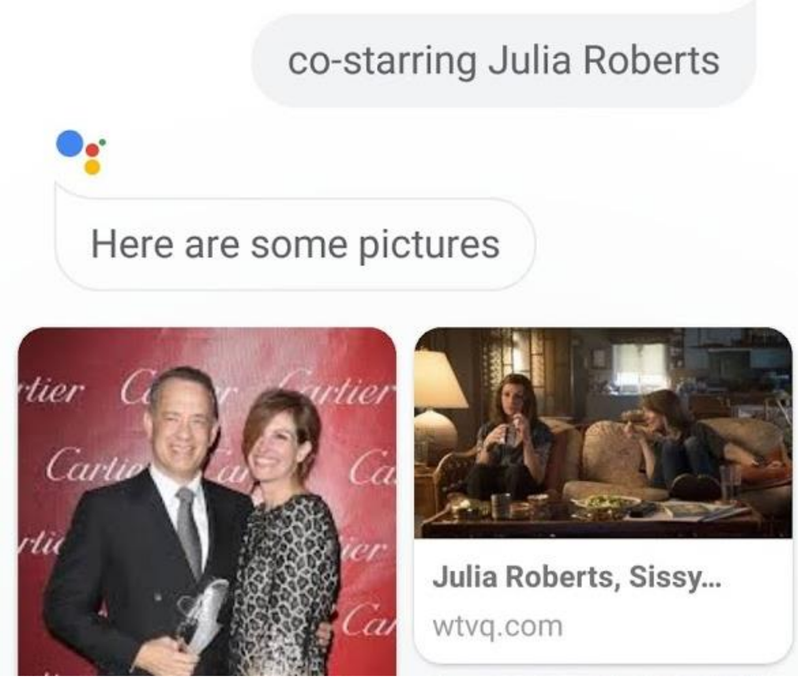
Complex questions involving multiple entities and relations



Conversational questions with implicit context



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



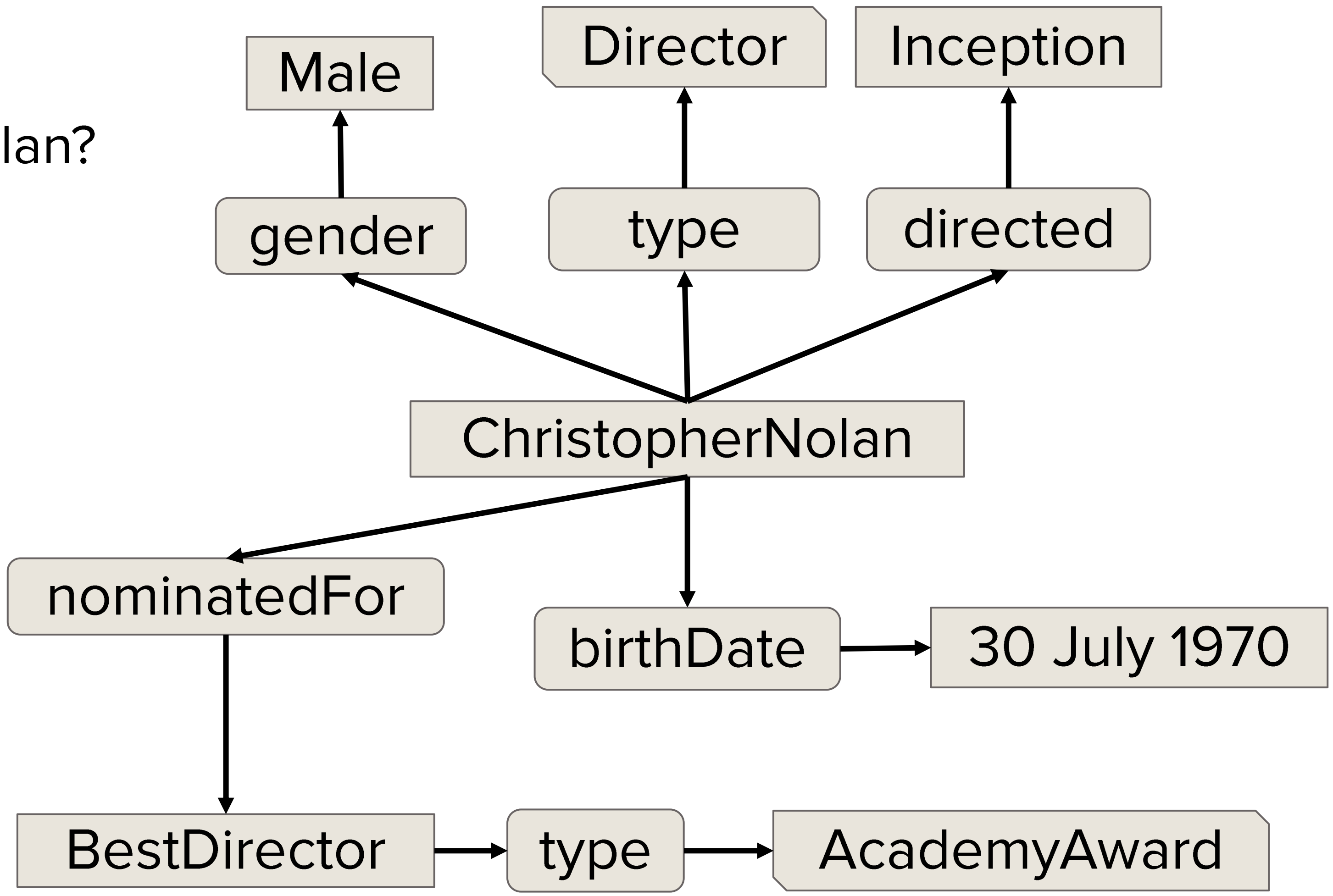
Significant progress has been made on **knowledge base contruction** over the last fifteen years or so; but for **question answering**, which is one of the most valuable applications of KBs, we are still at the **tip of iceberg**!

QA over knowledge graphs (KG-QA)

What are the Oscar nominations of Nolan?

QA over knowledge graphs (KG-QA)

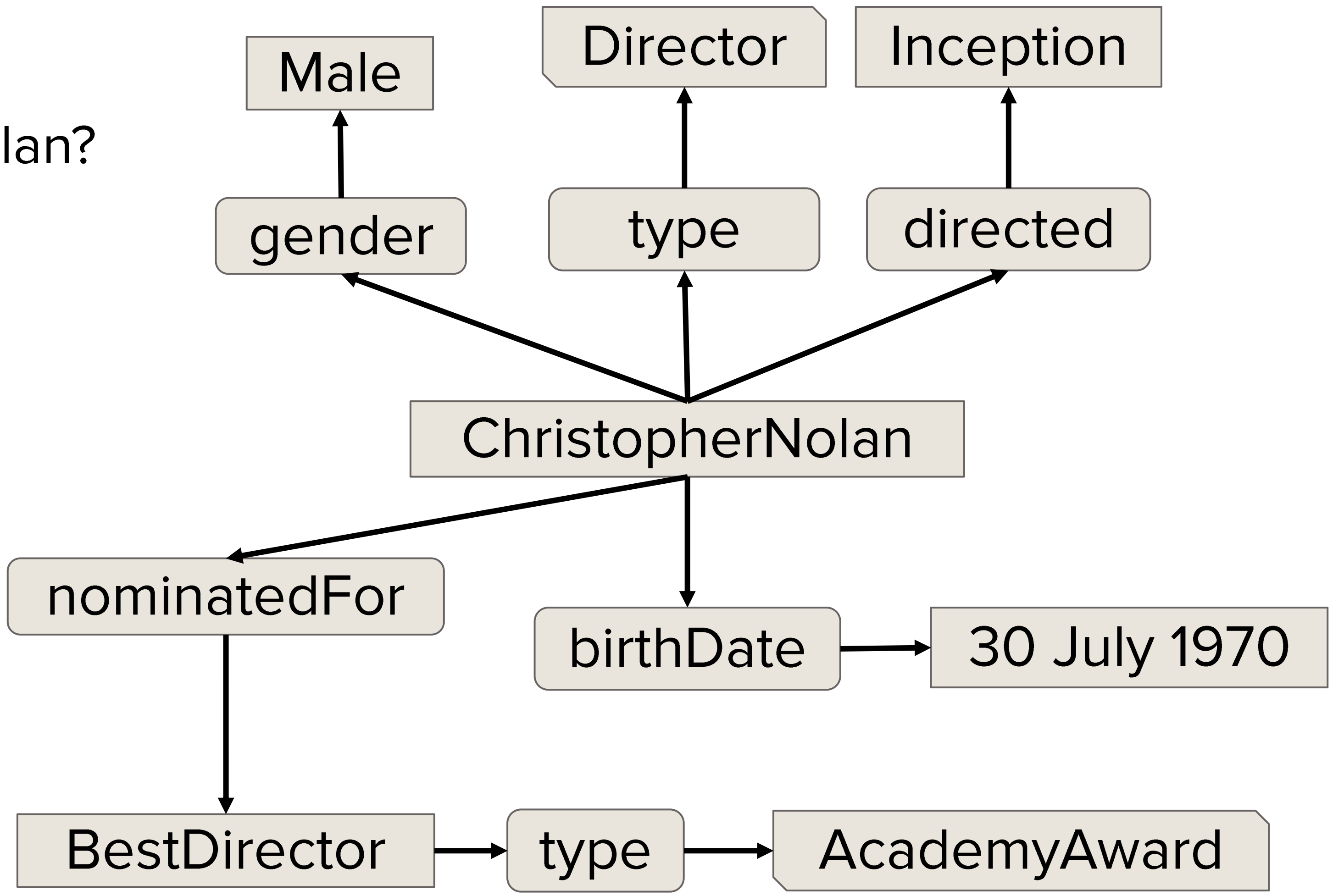
What are the Oscar nominations of Nolan?



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]



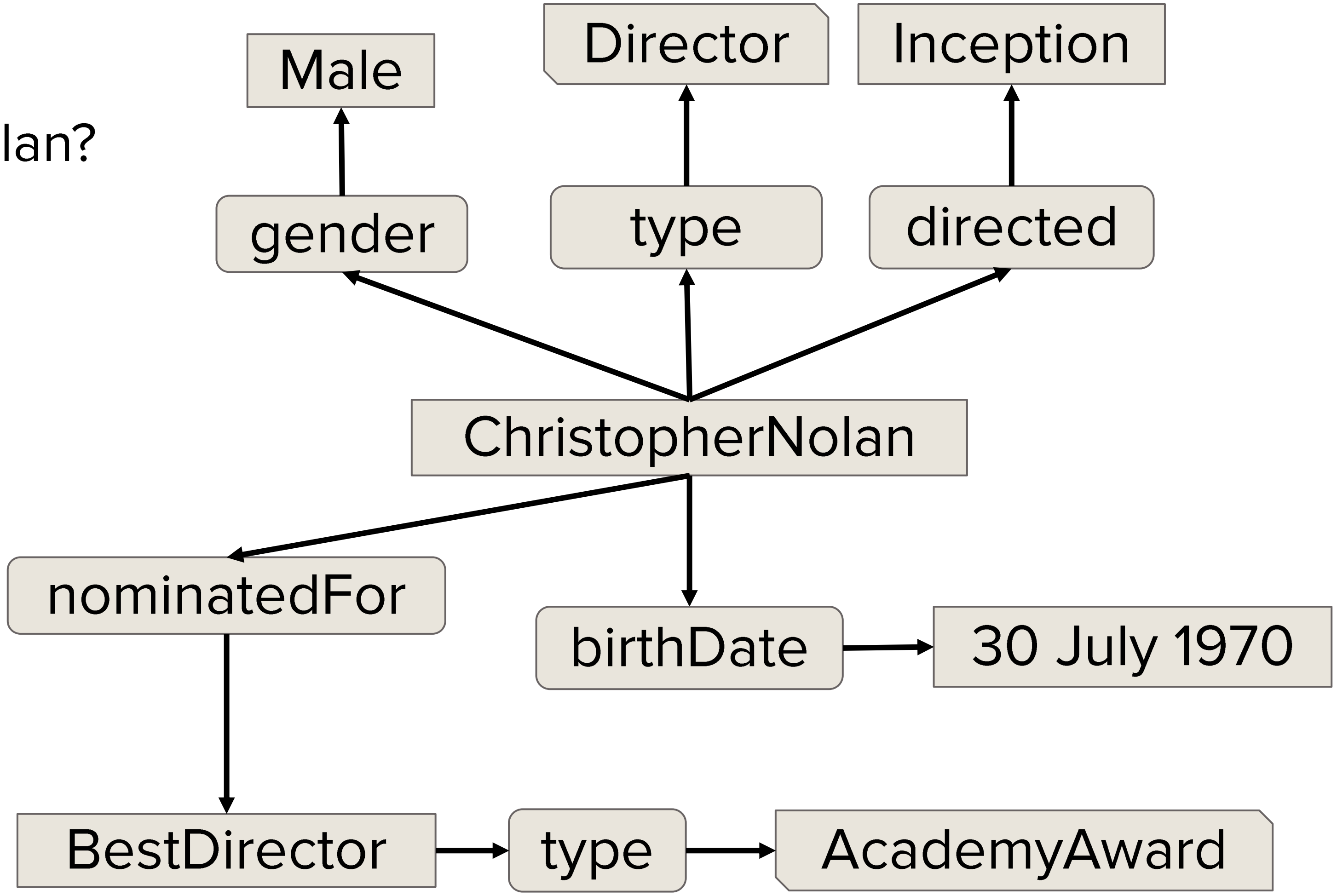
QA over knowledge graphs (KG-QA)

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

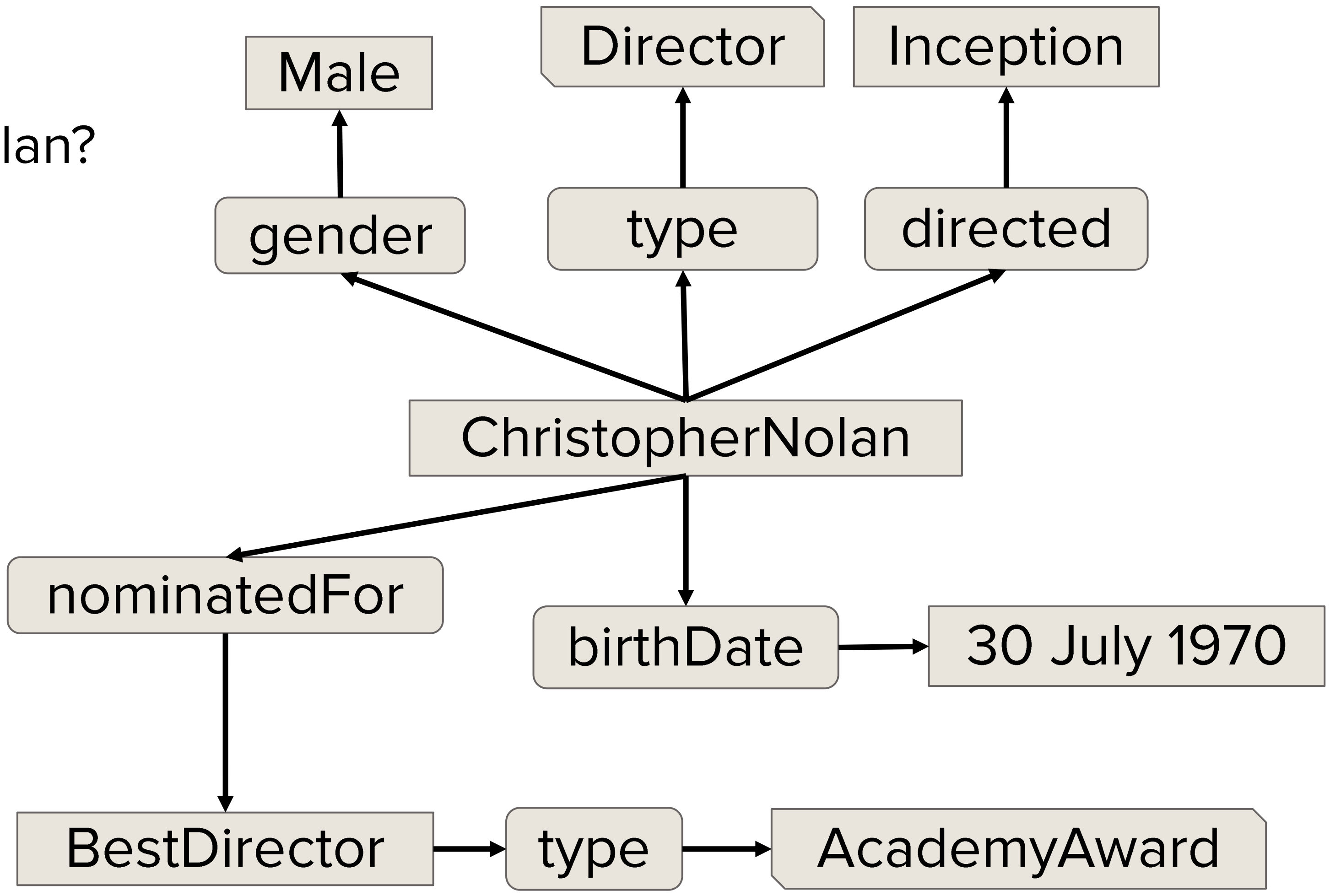
Here: Entities, predicates, types, literals



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]



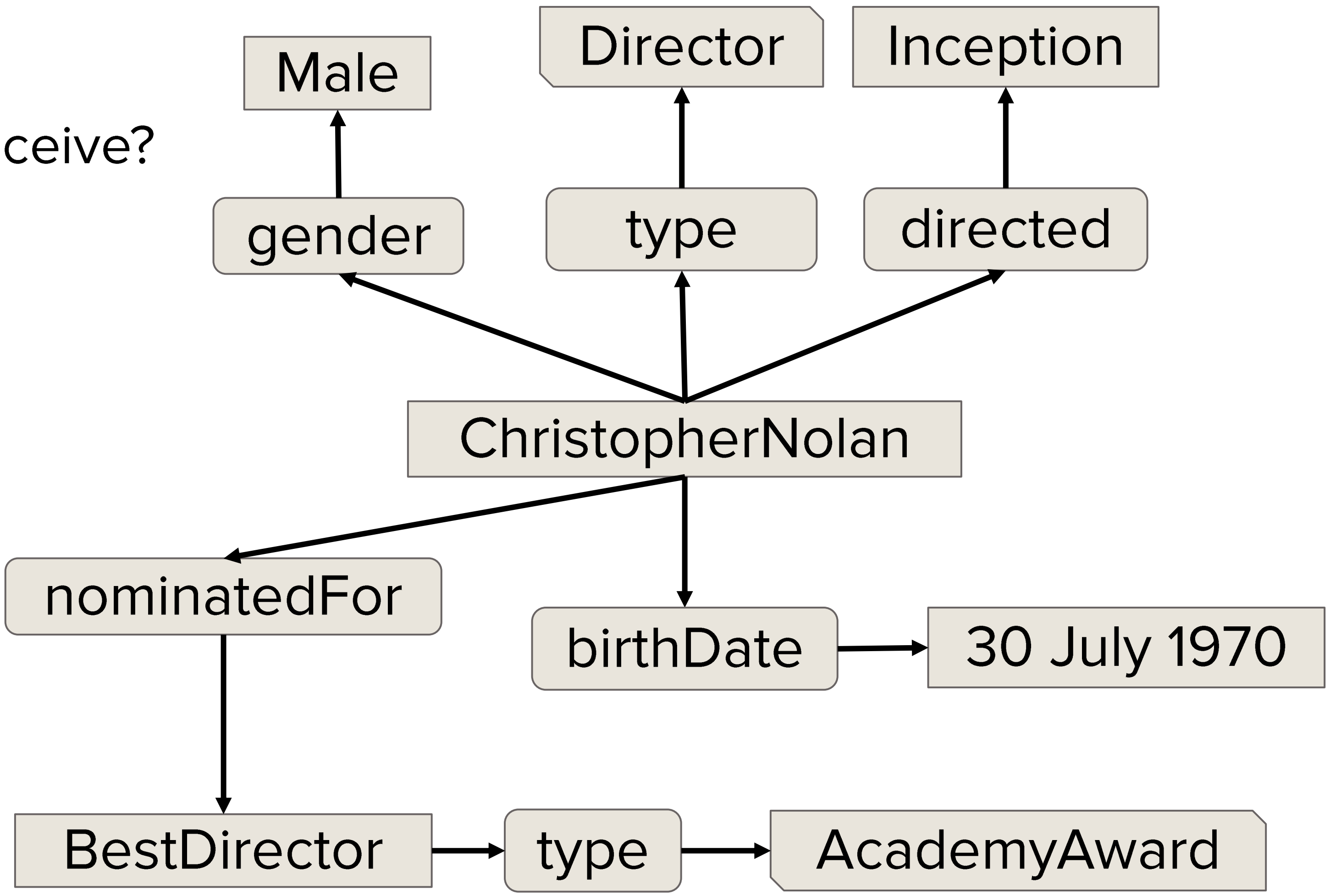
Wikidata: 12B facts, 84M entities, 7k predicates, 69k types

WIKIDATA

KGs and KBs are equivalent

Which Oscar nominations did Nolan receive?

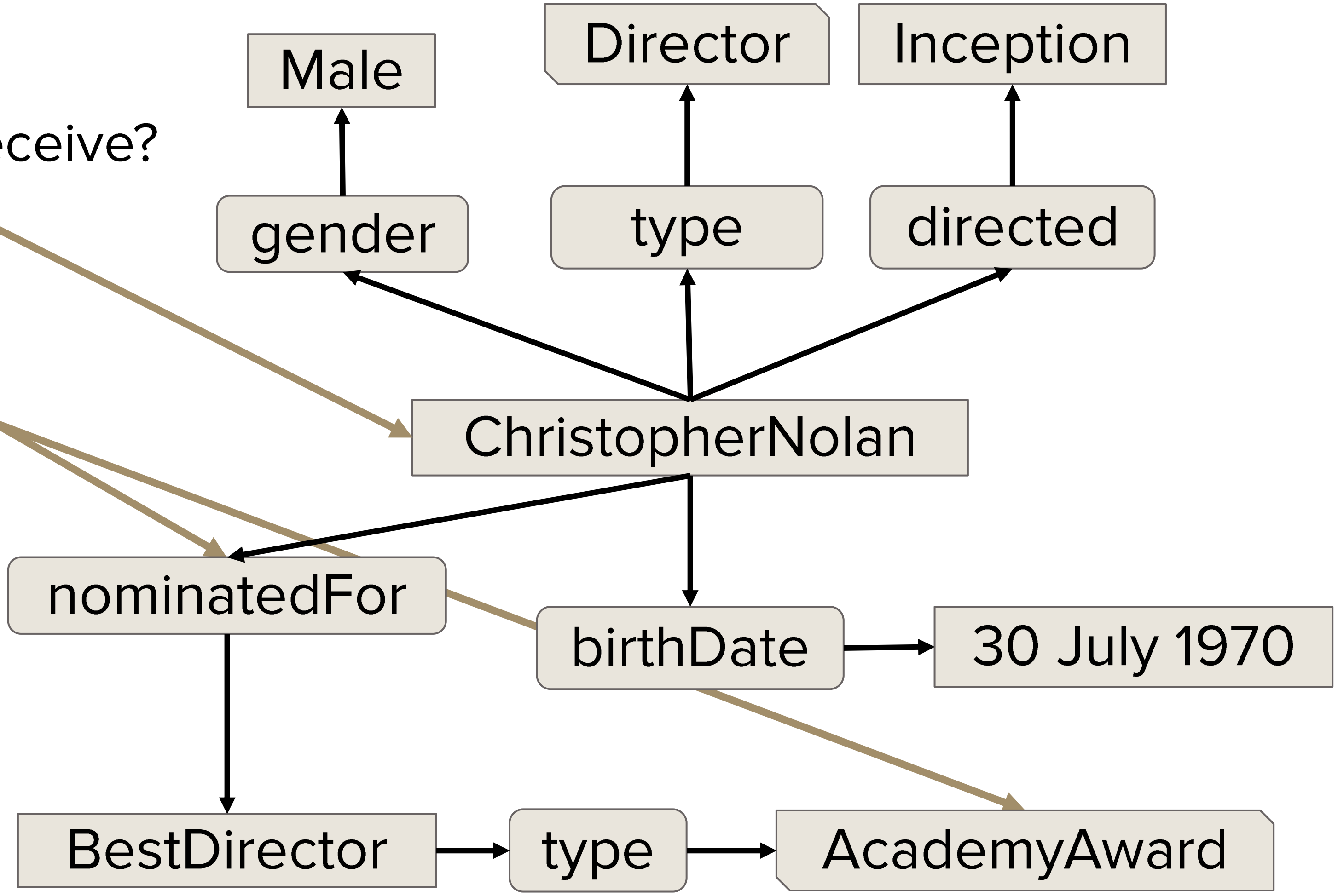
<ChristopherNolan, gender, Male>
<ChristopherNolan, type, Director>
<ChristopherNolan, directed, Inception>
<ChristopherNolan, nominatedFor, BestDirector>
<BestDirector, type, AcademyAward>
<ChristopherNolan, birthDate, 30 July 1970>



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>



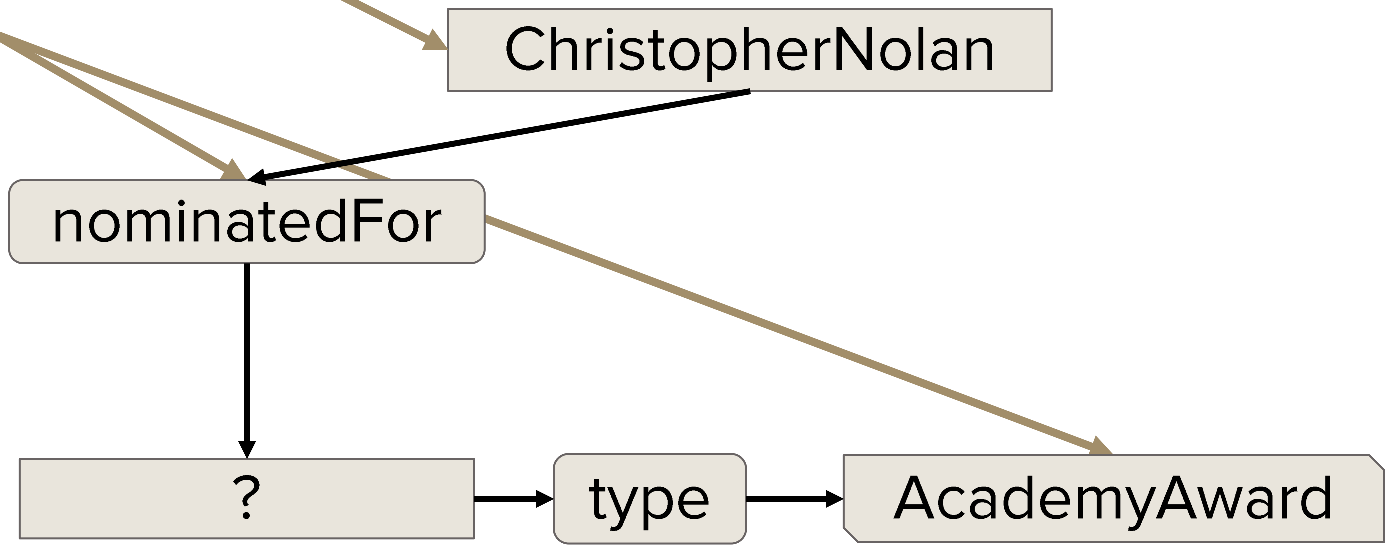
KG-QA Challenge 2: Query formulation

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

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

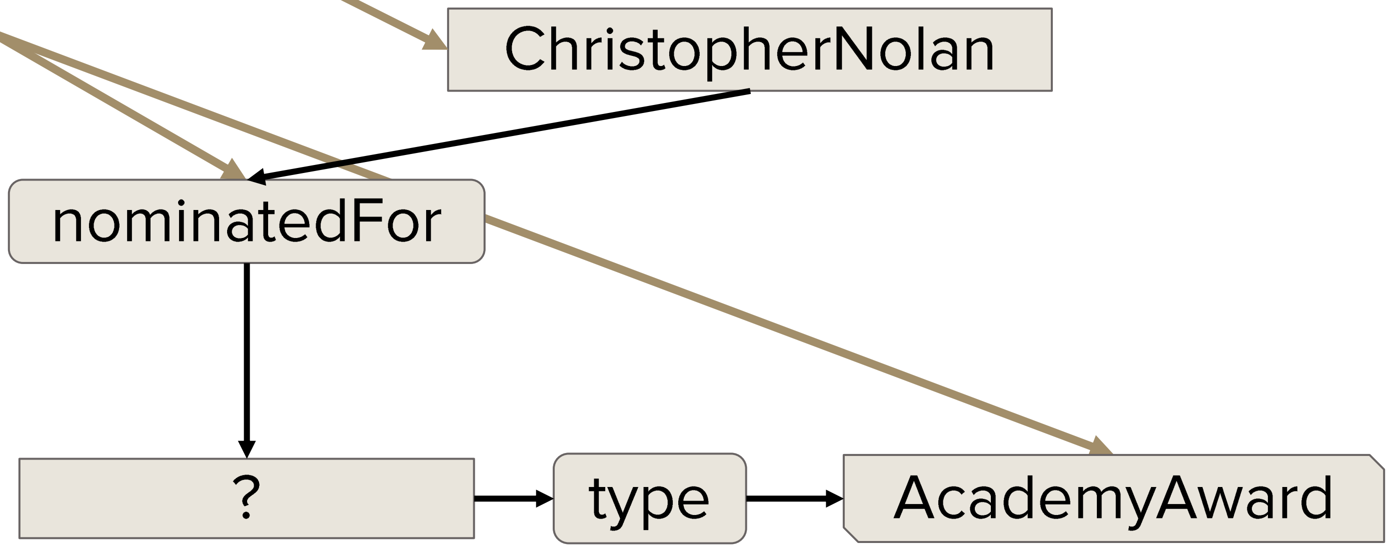
Named Entity Recognition (NER): Stanford NER, spaCy

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>

SPARQL

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



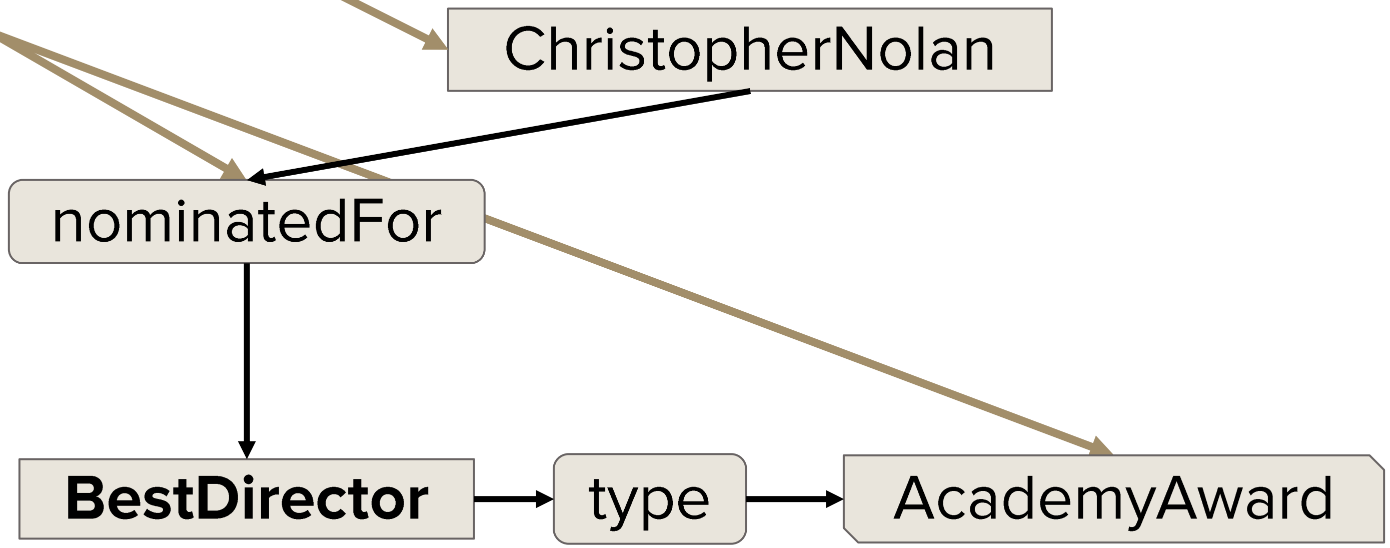
Answering with query

Which Oscar nominations did Nolan receive?

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

SPARQL

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



BestDirector

Structured queries and logical forms

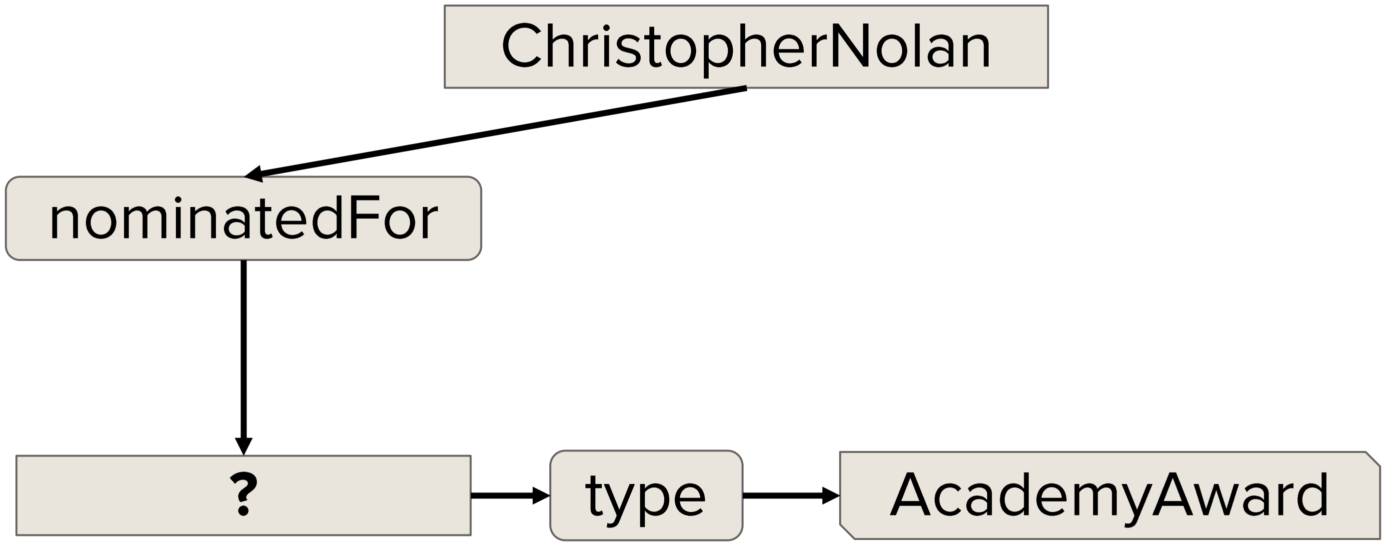
Which Oscar nominations did Nolan receive?

Lambda-calculus $\lambda x. nominatedFor(ChristopherNolan, x) \wedge Type(x, AcademyAward)$

Neo4j CYPHER Graph QL

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

Lambda-DCS $nominatedFor.ChristopherNolan \sqcap type.AcademyAward$



SPARQL BGP

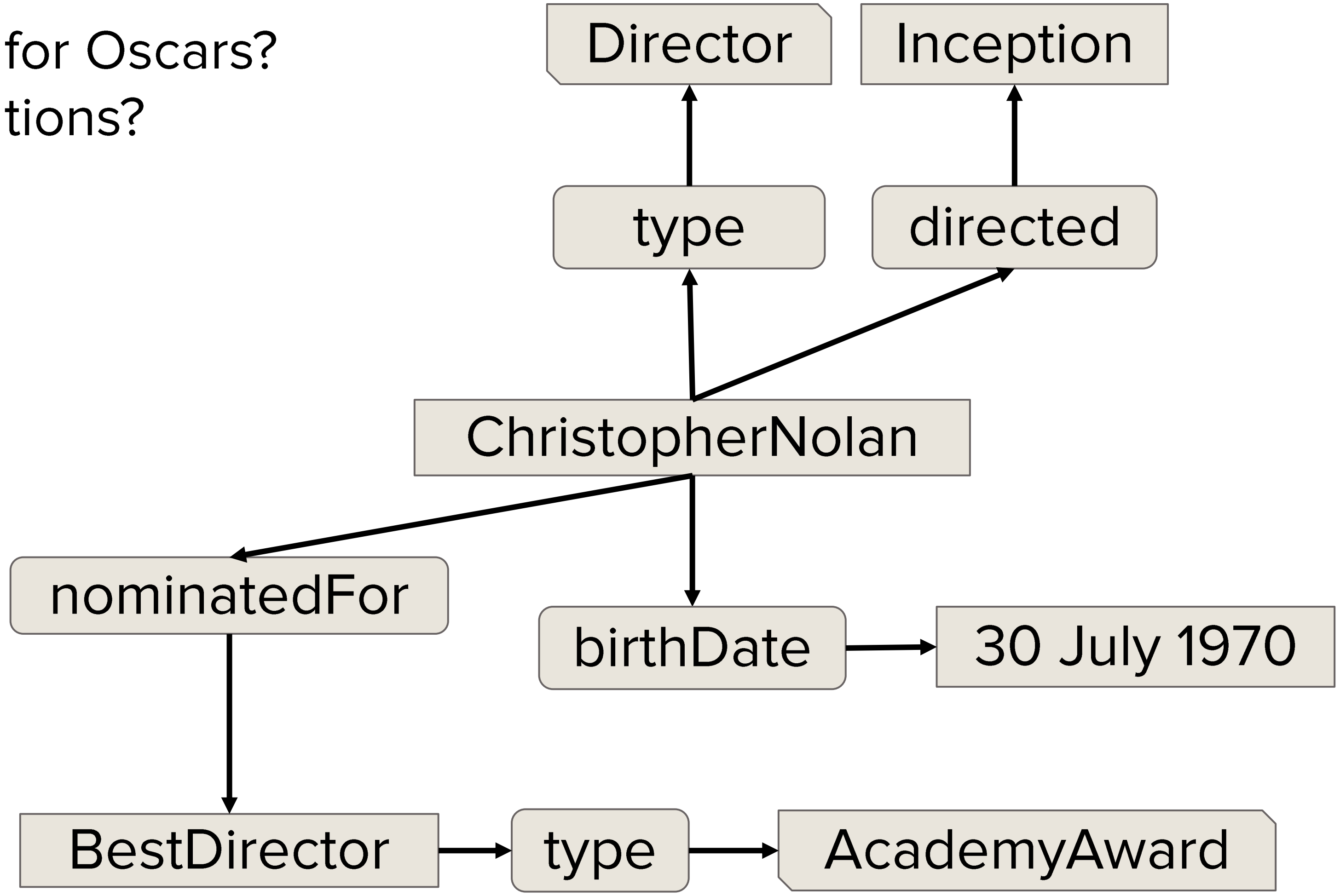
SELECT ?ANS
WHERE {
ChristopherNolan nominatedFor ?ANS .
?ANS type AcademyAward }

BestDirector

Reification: n-ary information in KGs

For which films was Nolan nominated for Oscars?
When did Nolan get his Oscar nominations?

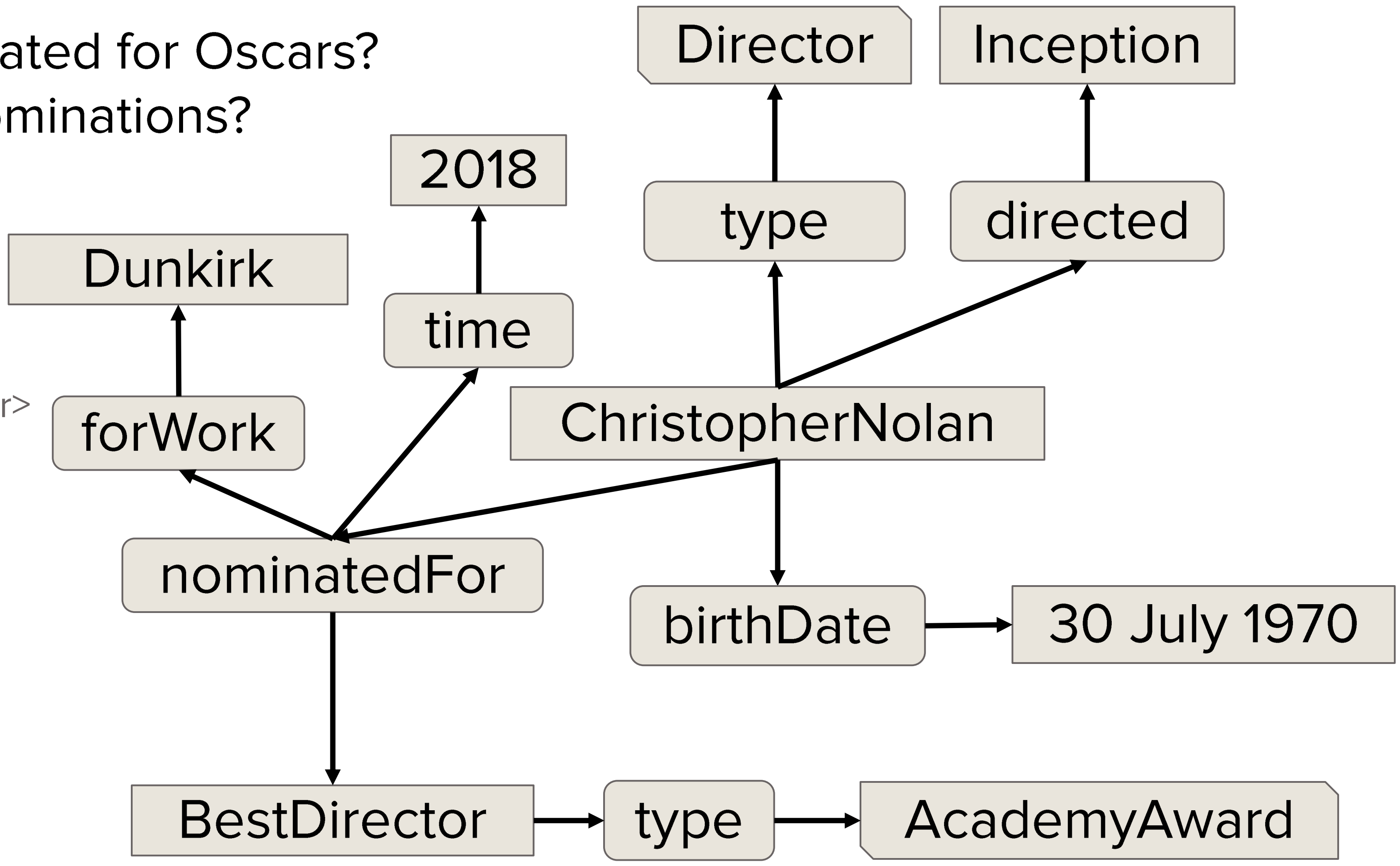
<ChristopherNolan, gender, Male>
<ChristopherNolan, type, Director>
<ChristopherNolan, directed, Inception>
<ChristopherNolan, nominatedFor, BestDirector>
<BestDirector, type, AcademyAward>
<ChristopherNolan, birthDate, 30 July 1970>



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>

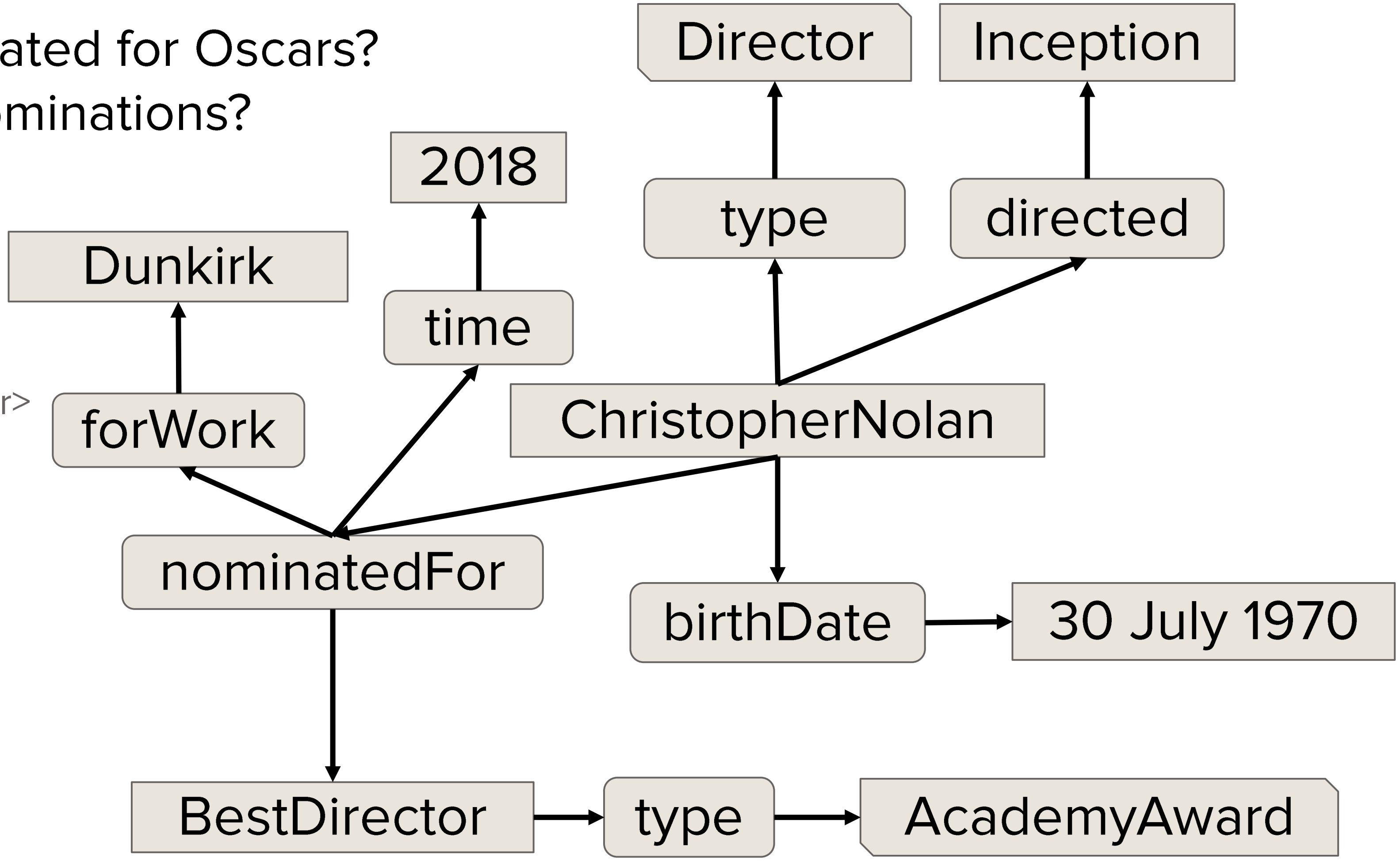


Reification: n-ary information in KGs

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, 2018>

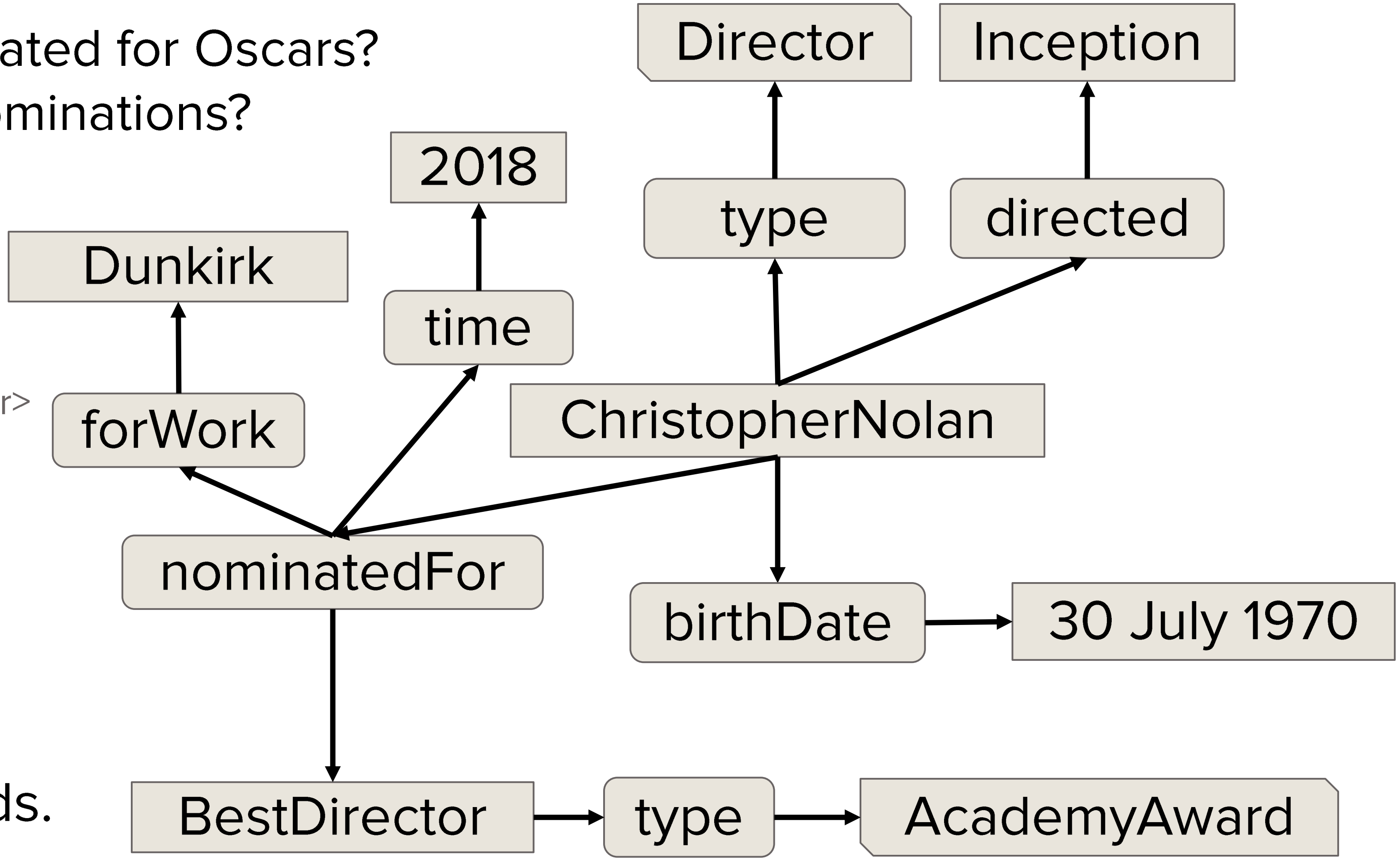


Qualifiers are a huge part of Wikidata

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>

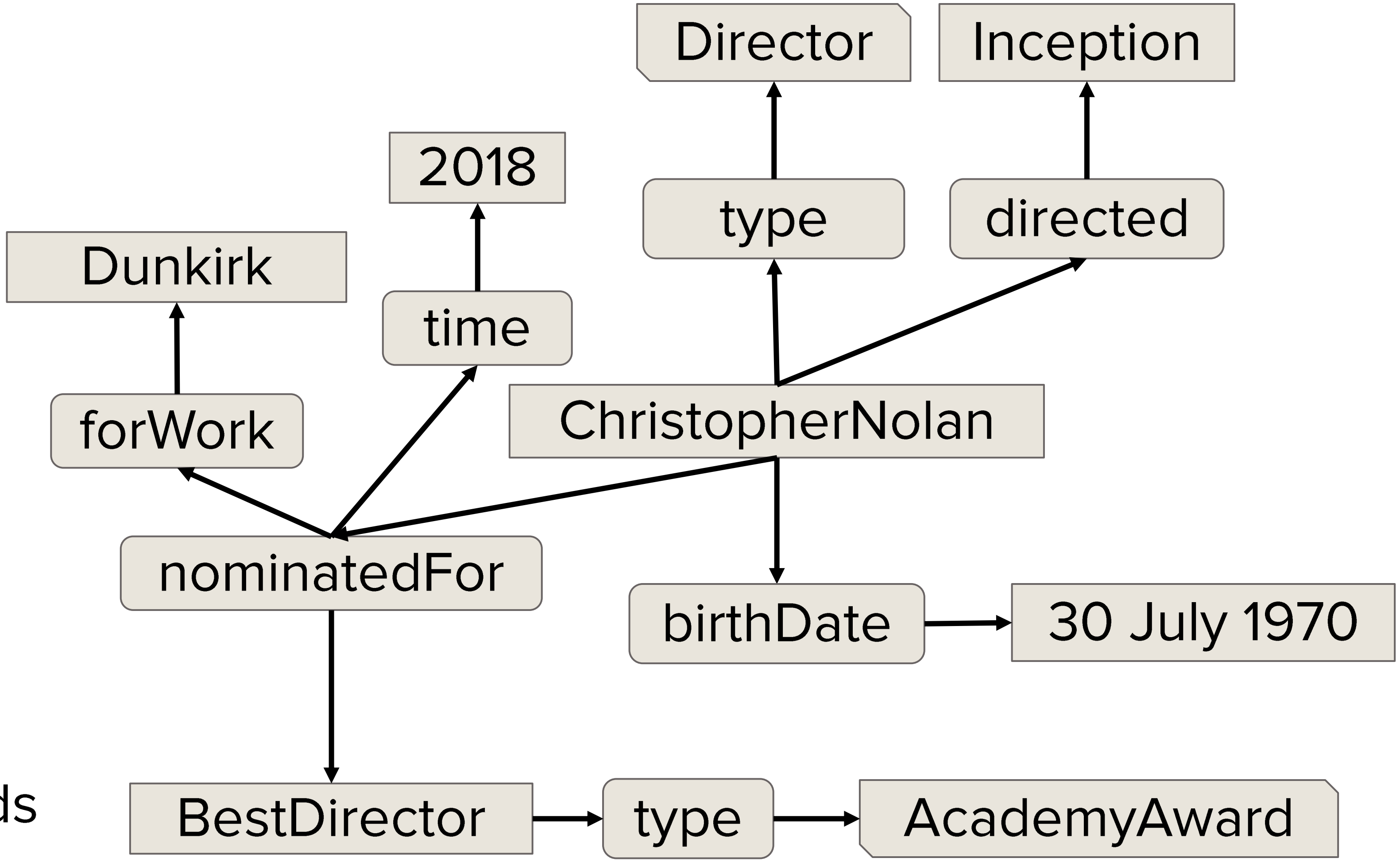
<ChristopherNolan, nominatedFor, 123>
<123, nominatedFor, BestDirector>
<123, forWork, Dunkirk>
<123, time, 2018>



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

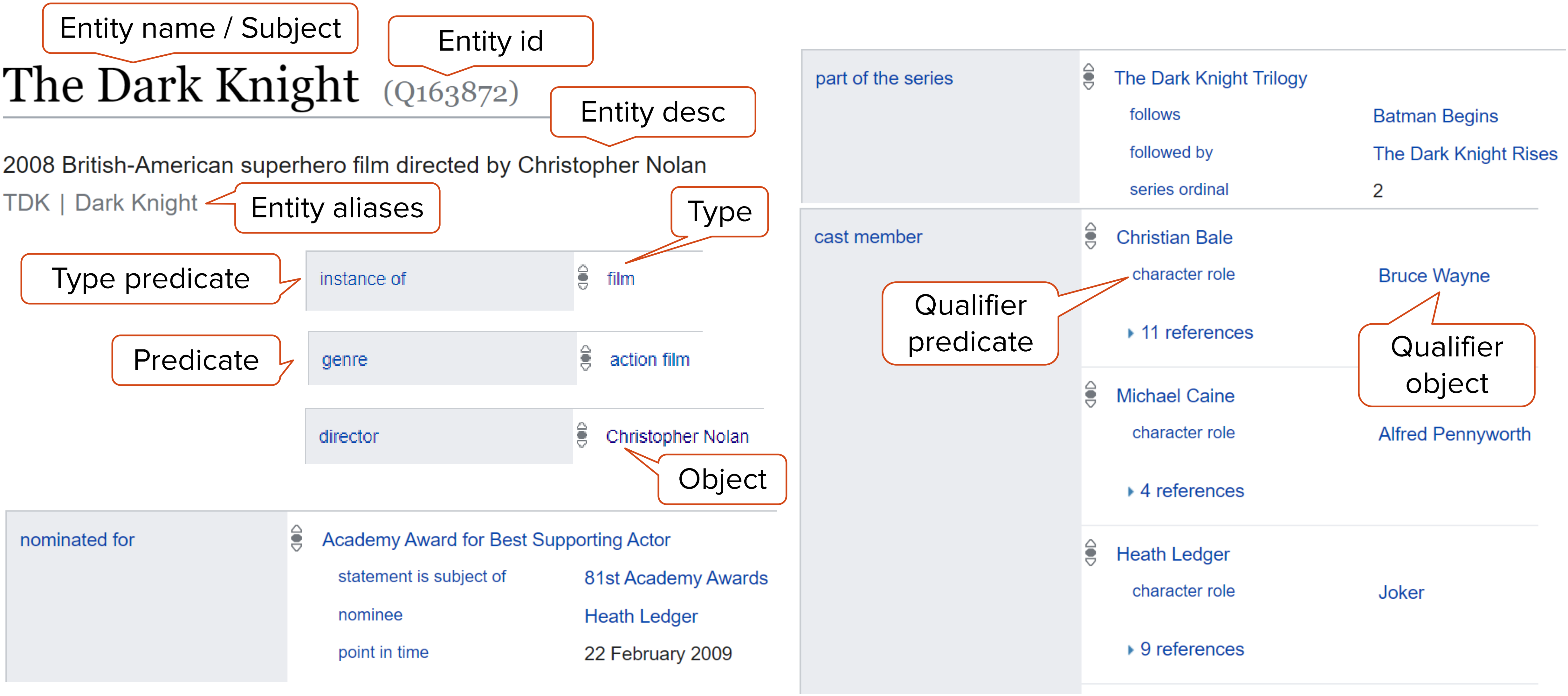
Questions that need reified triples

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!!

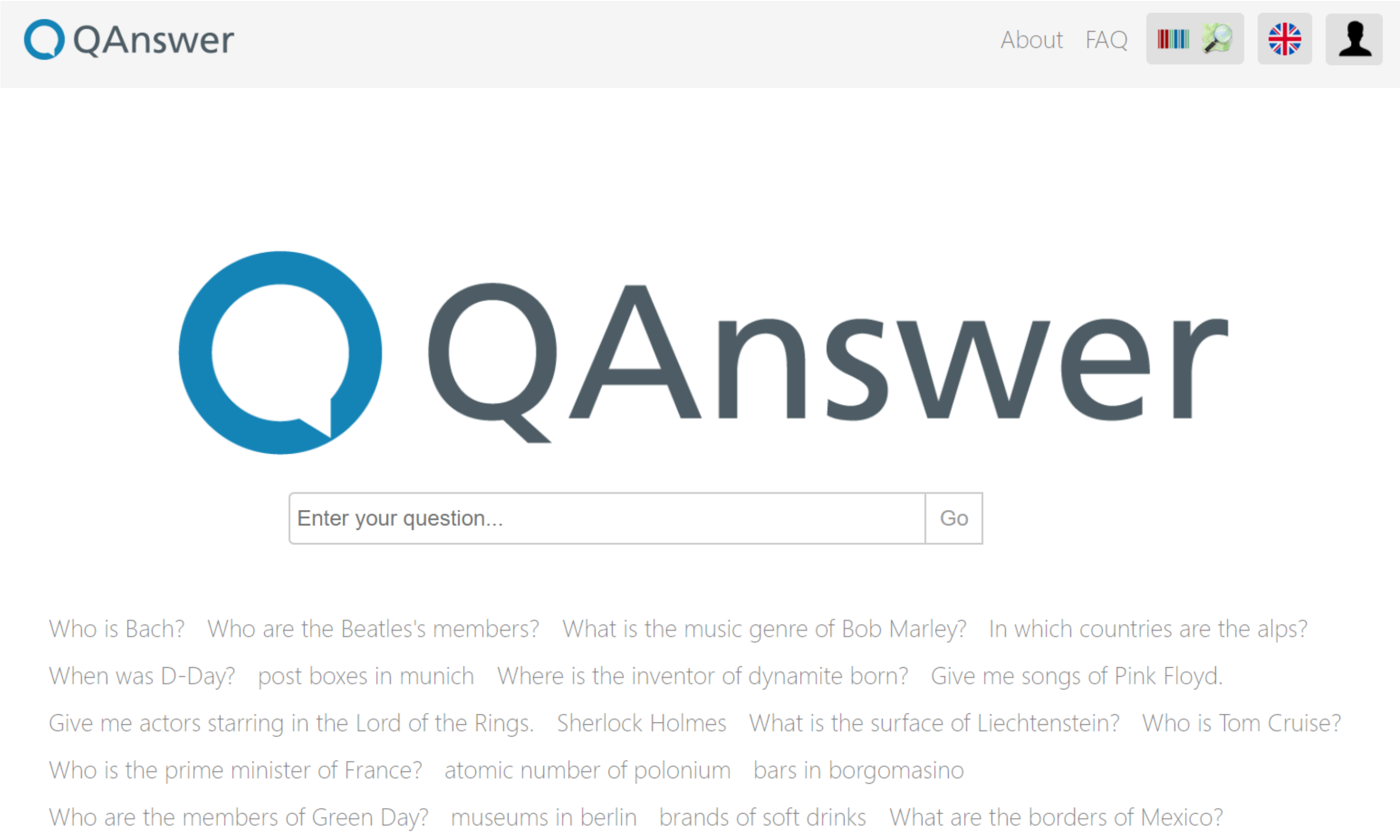
Explore Wikidata



Explore Wikidata like a pro

- Wikidata: 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
- 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)



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
- **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

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

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. AAAI 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 **name of the second book?**	When did he die?	Did they win the Super Cup the previous year?	Which was released first: Houses of the Holy or Physical Graffiti?	**What about Bullock?**
The War of the Dwarves	22 January 2008	No	Houses of the Holy	Donal Logue
Who is the author ?	**Batman actor?**	**Which club was the winner?**	**Is the rain song and immigrant song there?**	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?	**Married to in 2017?**
Homburg	Christopher Nolan	Chelsea F.C.	Jimmy Page	Miranda Phillips Cowley
When was he born ?	**Sequel name?**	**Which stadium is this club's home ground?**	**Name of his previous band?**	**Wedding date first wife?**
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

Question Who is **Inception**'s **director**?	**Question template** Who is **<NOUN1>**'s **<NOUN2>**?
Query ?ANS **director** **Inception**	**Query template** ?ANS **<PRED1>** **<ENT1>**

1 SPARQL triple pattern

Templates for KG-QA

- Generalizes to new domains

Who is **Libya**’s **president**?

Who is **Messi**’s **manager**?

Question Who is **Inception**’s **director**?	**Question template** Who is **<NOUN1>**’s **<NOUN2>**?
Query ?ANS **director Inception**	**Query template** ?ANS **<PRED1> <ENT1>**

1 SPARQL triple pattern

Templates for KG-QA

Question Who **plays** the **role of Cobb in** **Inception**?	**Question template** Who <**VERB**> <DT> **<NOUN1> <PREP1>** **<NOUN2>** <**PREP2**> **<NOUN3>**?
Query **Inception castMember** ?VAR ?VAR **castMember** ?ANS ?VAR **characterRole Cobb**	**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)

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

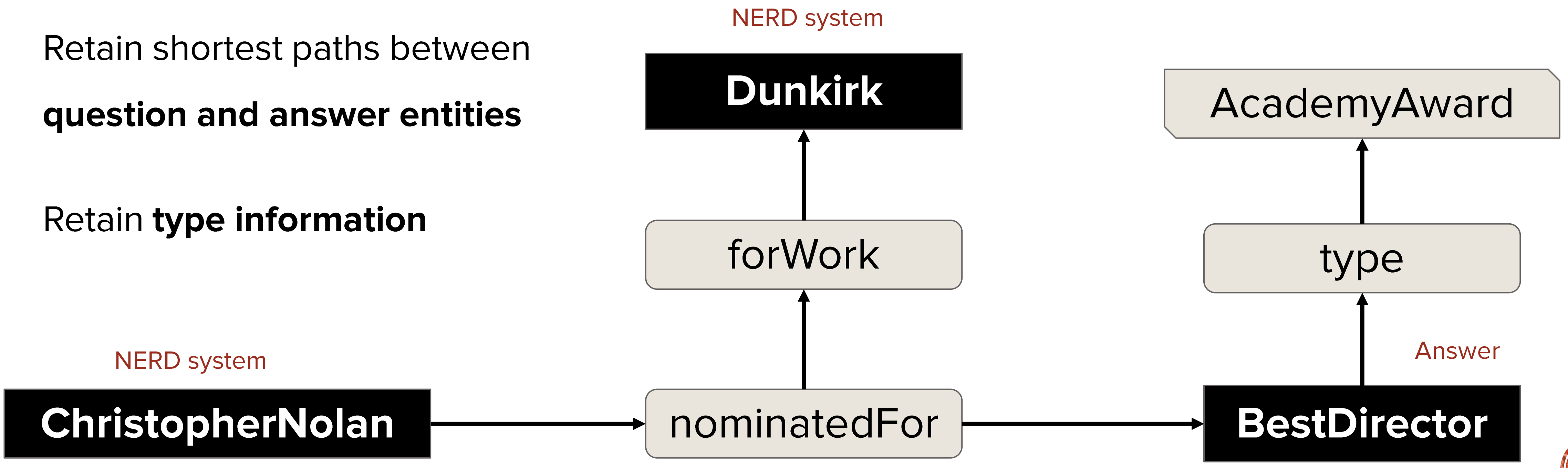
Answer

BestDirector

Distant supervision from QA pairs

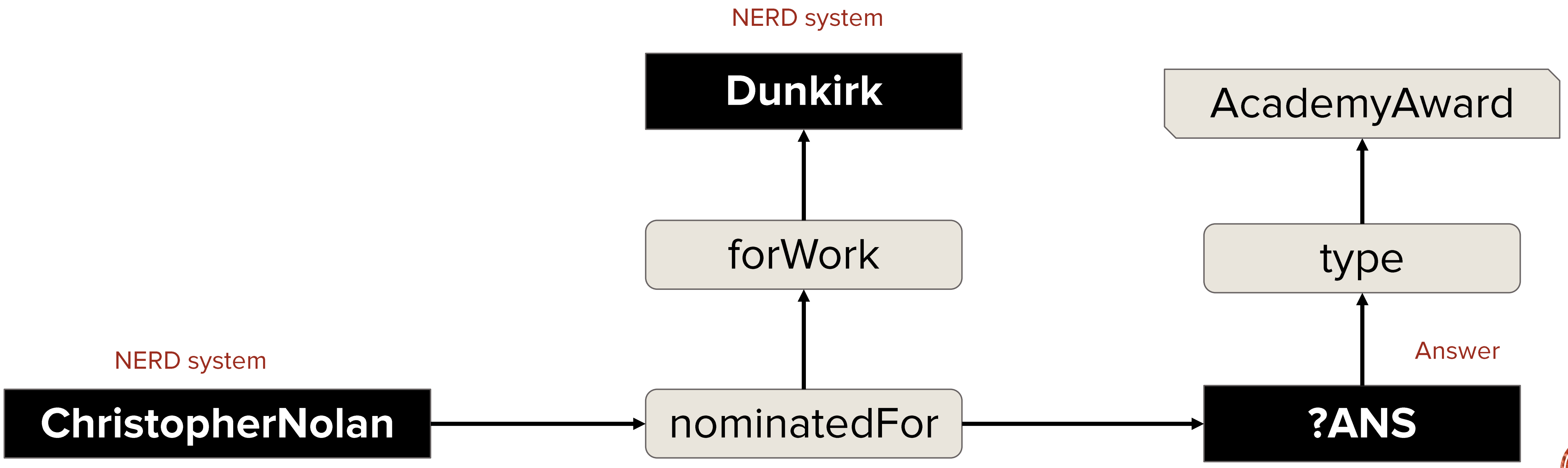
Question: Which Oscar award nomination did Nolan get for the film Dunkirk?
Answer: Best Director

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

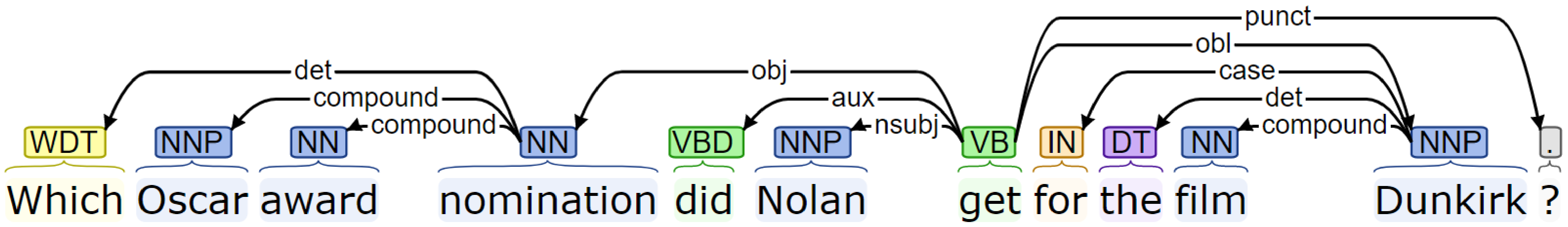


Distant supervision from QA pairs

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

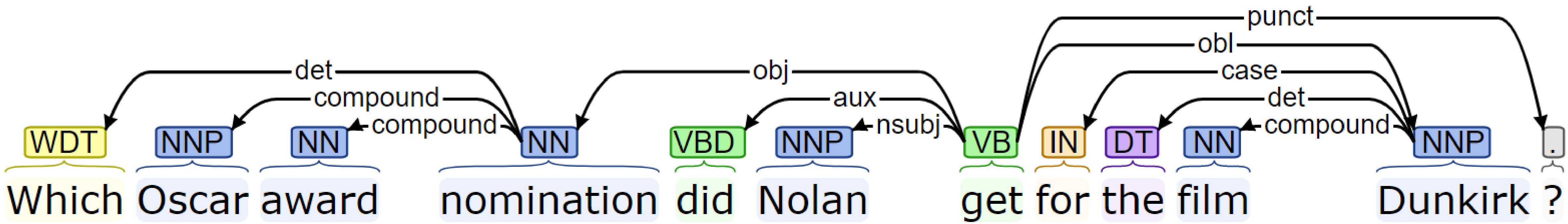


which nomination | oscar nomination | get nomination | for film | oscar | get

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

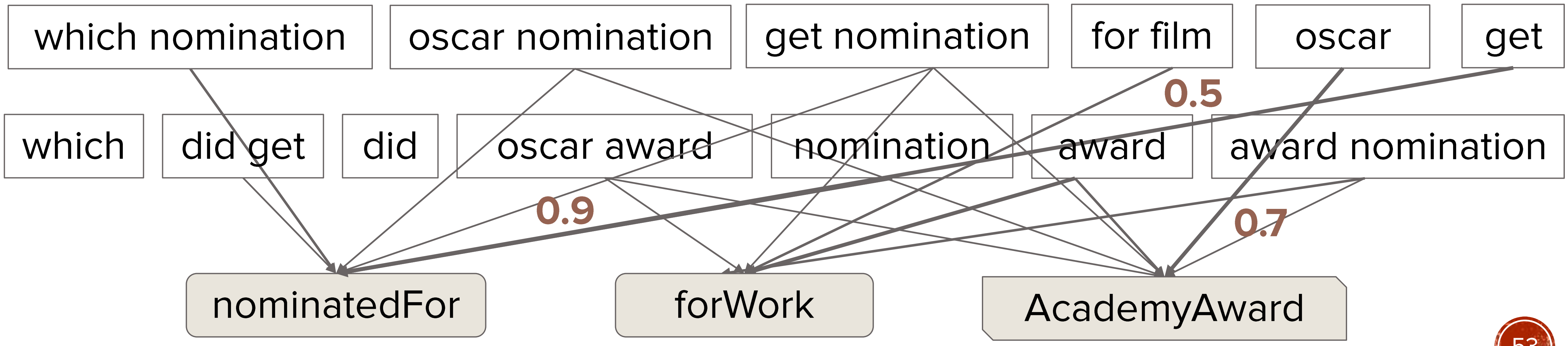
nominatedFor

forWork

AcademyAward

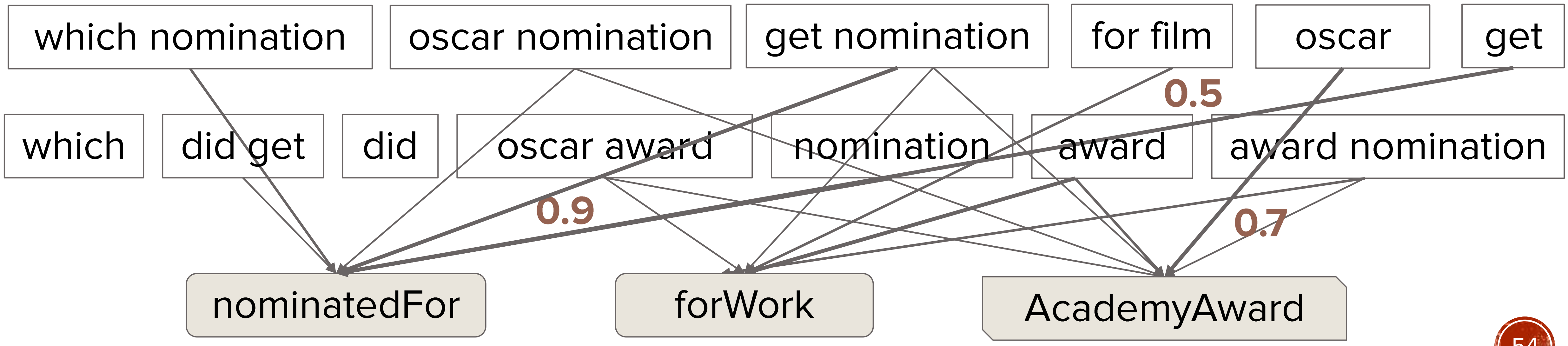
Create candidate alignments

- **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)



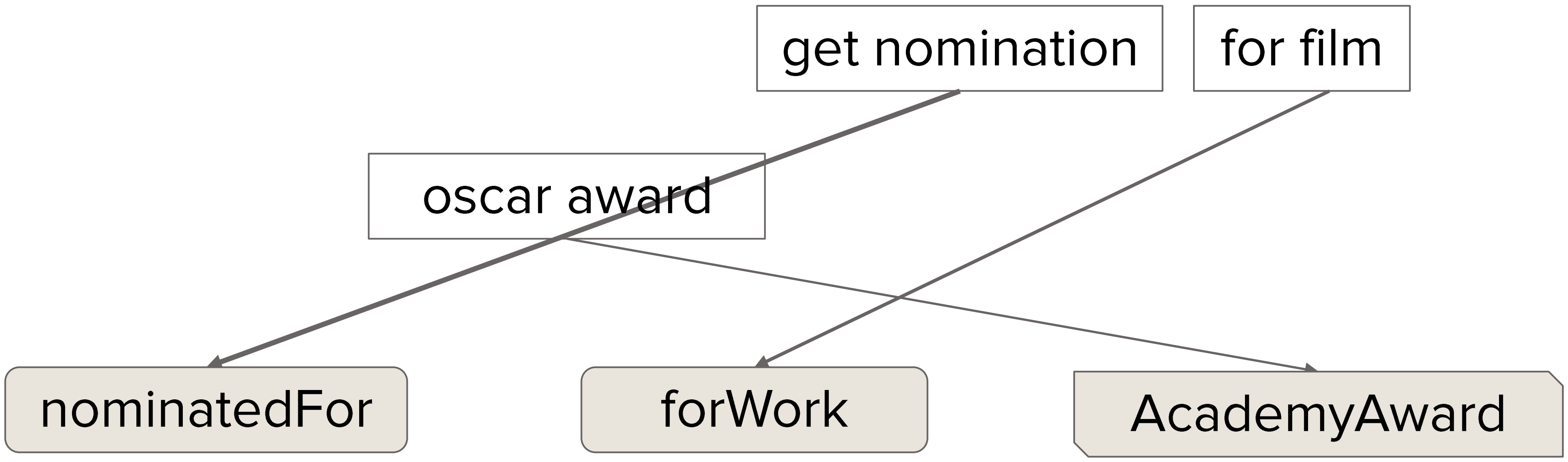
Optimal mapping via Integer Linear Program (ILP)

- 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

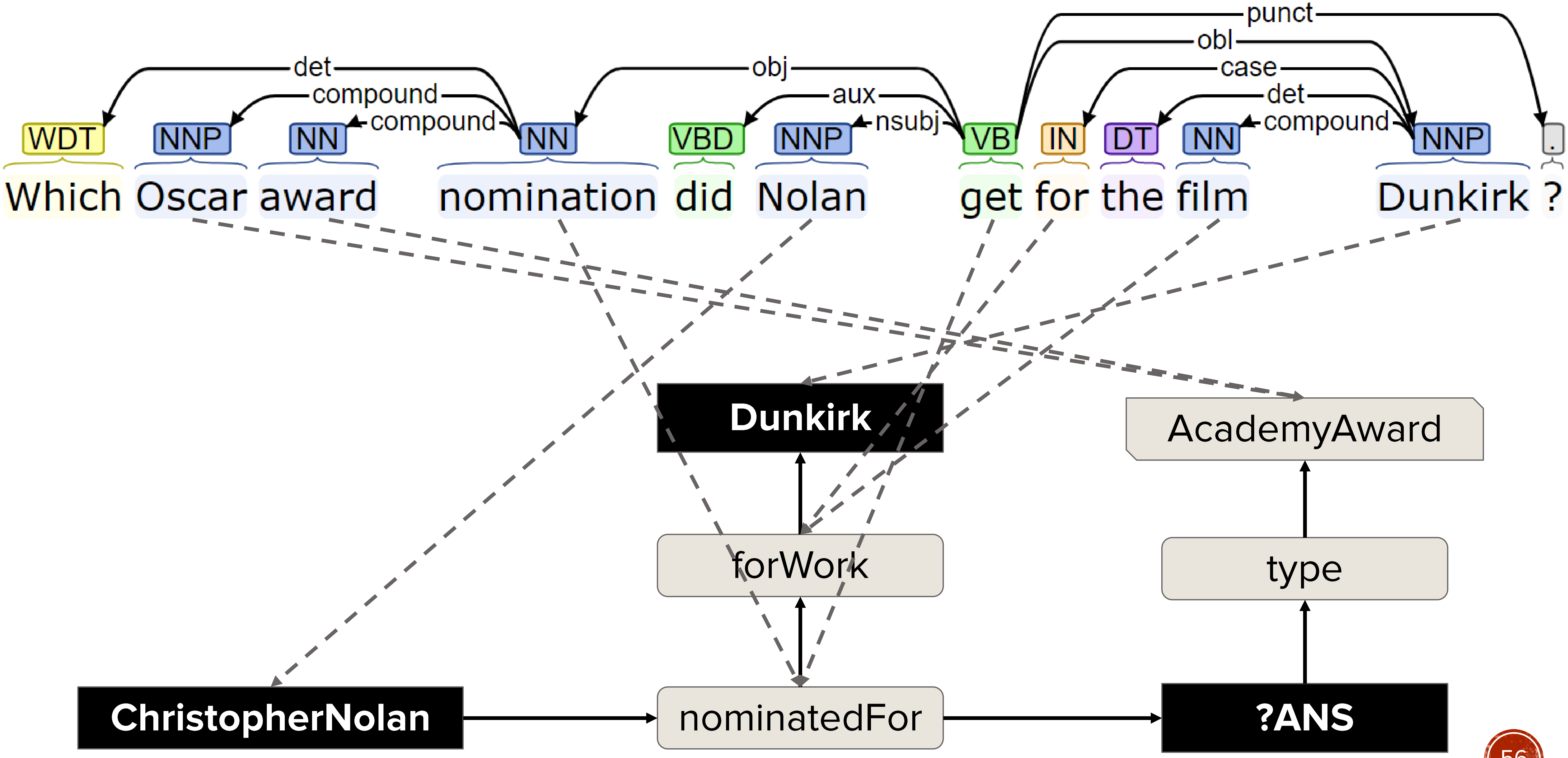


Optimal mapping via Integer Linear Program (ILP)

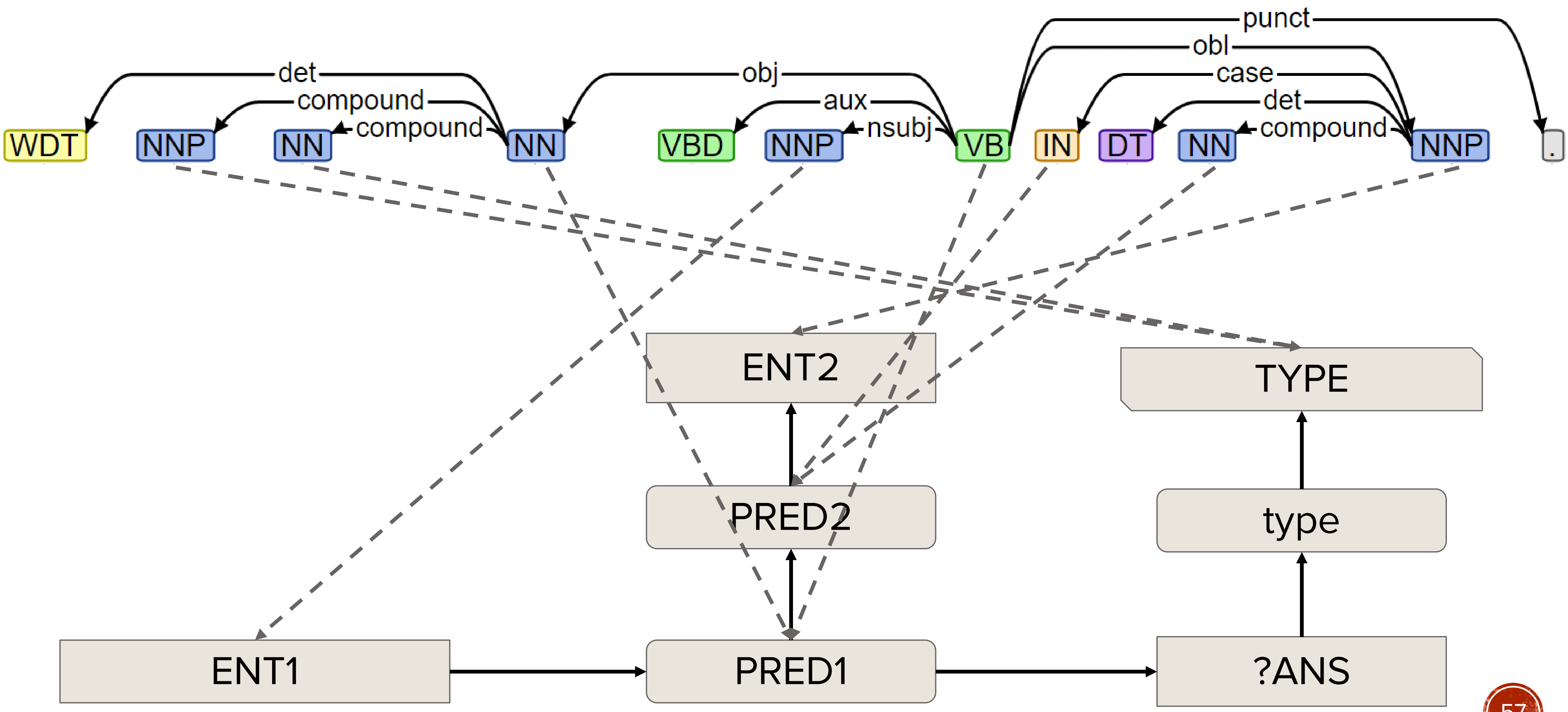
- Best alignment of items with **Integer Linear Program (ILP)**
- **Constraint 1:** Each KG item obtained from at most one phrase
- **Constraint 2:** Token contributing to entity cannot contribute to any other phrase
- **Constraint 3:** One phrase can map to at most one type



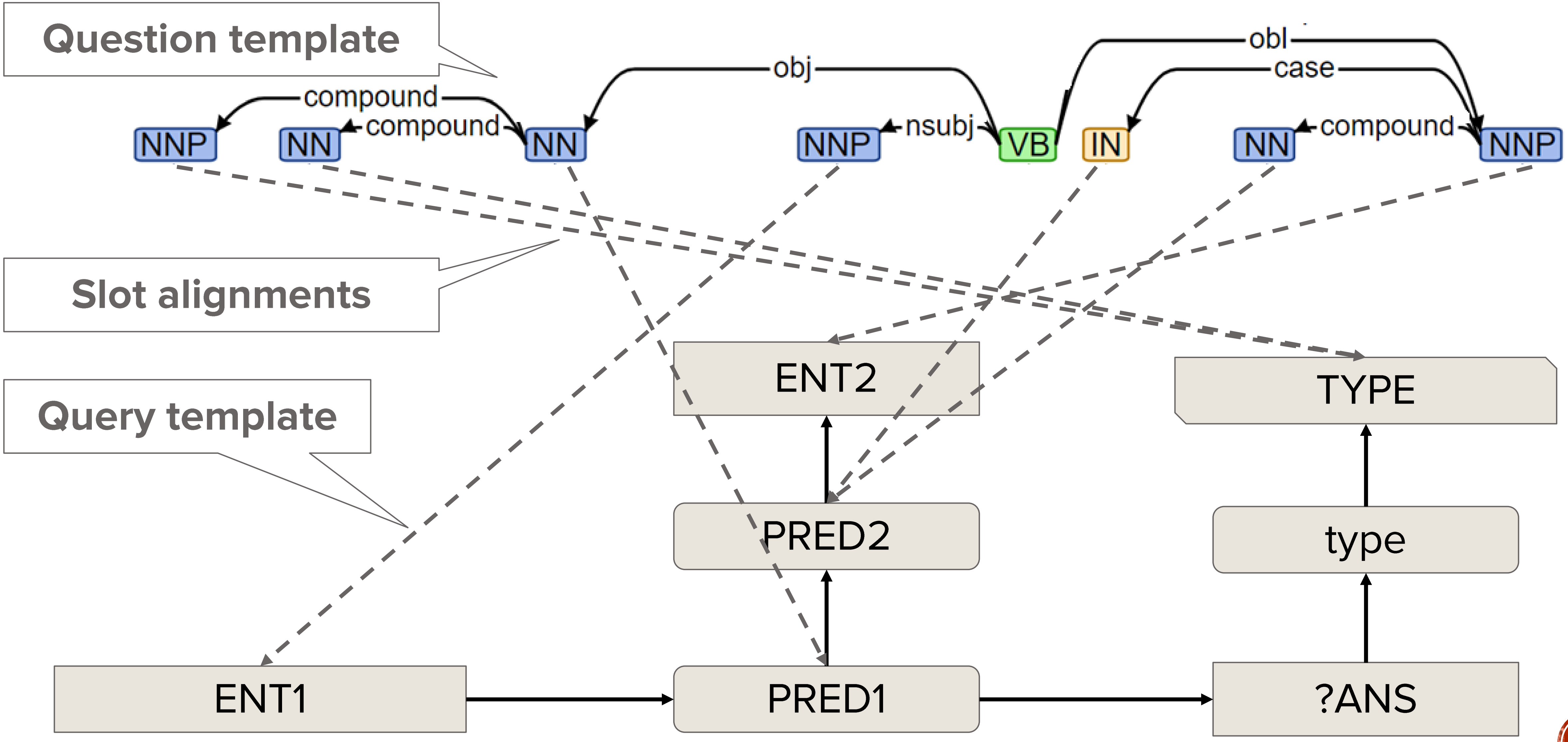
Apply alignment to question-query



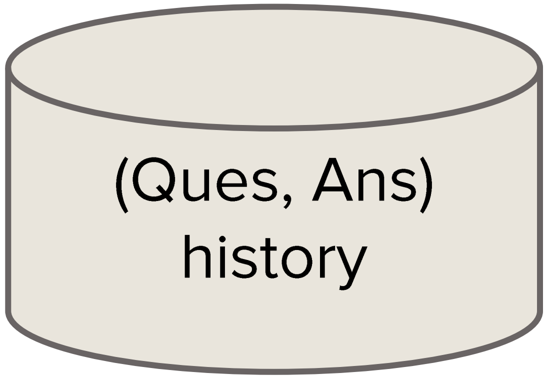
Replace concrete items by roles



Drop unnecessary question words



A continuous learning framework



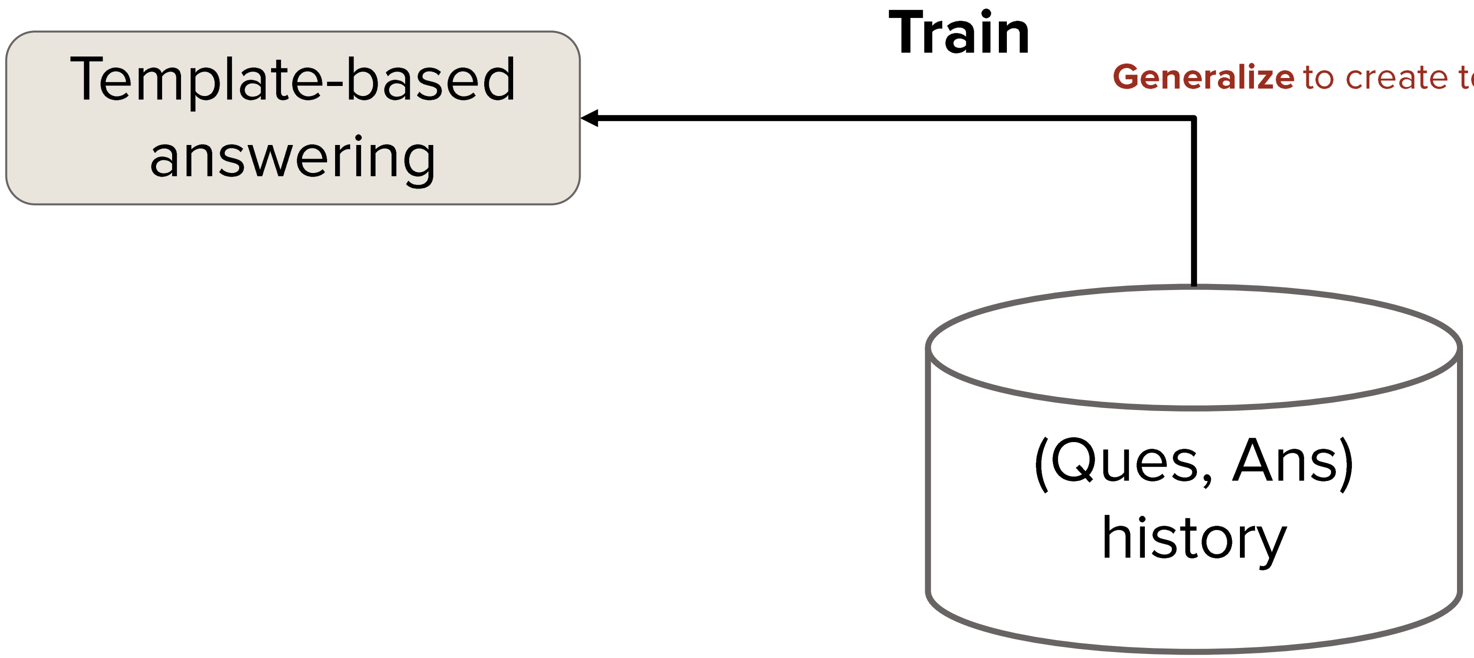
Abujabal et al., Automated template generation for question answering over knowledge graphs, WWW 2017.
Abujabal et al., Never-ending learning for open-domain question answering over knowledge bases, WWW 2018.

Train system using (Q, A) pairs

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

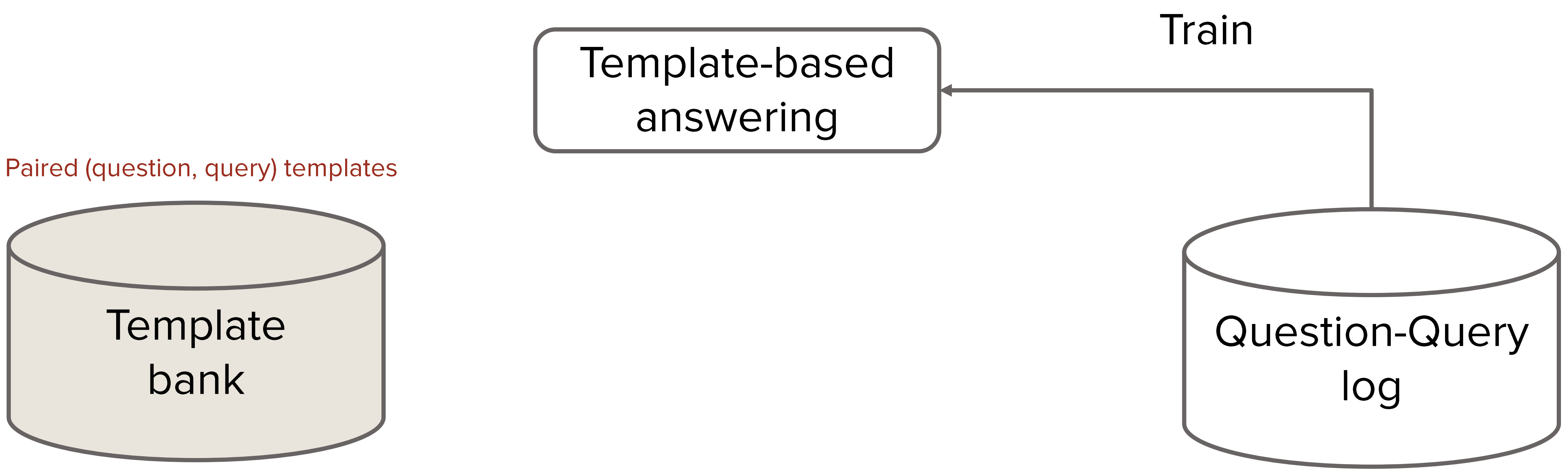
Align ques tokens to query tokens via lexicons and ILP

Generalize to create template



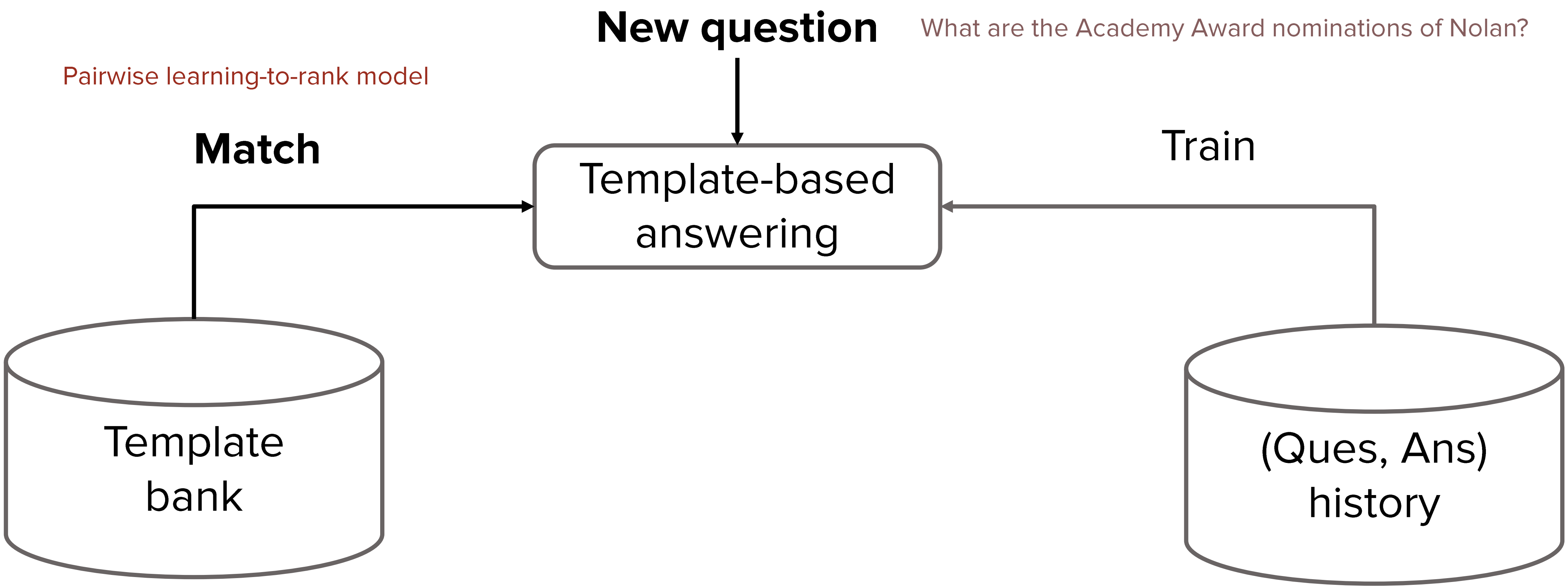
Abujabal et al., Automated template generation for question answering over knowledge graphs, WWW 2017.
Abujabal et al., Never-ending learning for open-domain question answering over knowledge bases, WWW 2018.

Learn a template repository



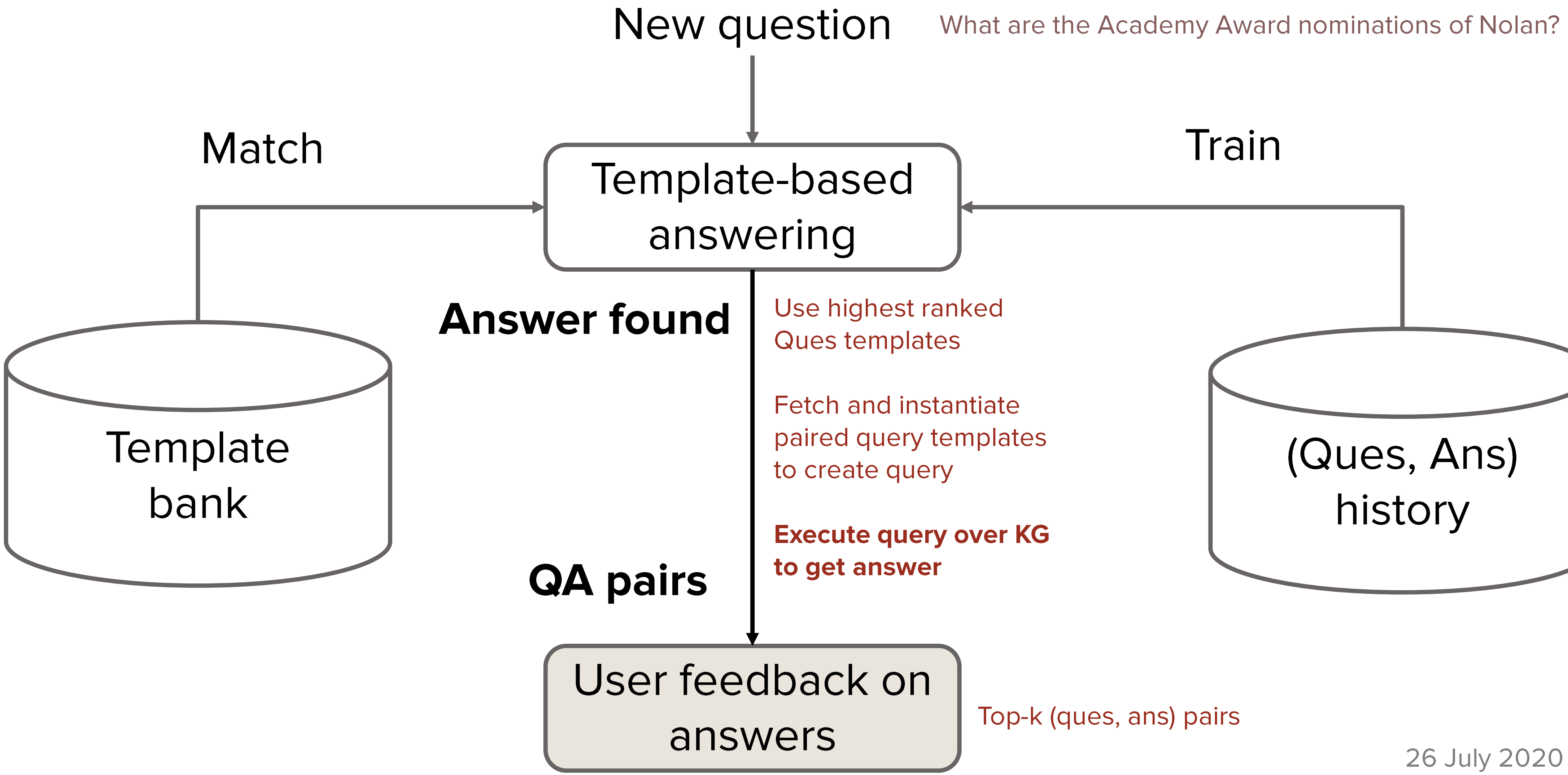
Abujabal et al., Automated template generation for question answering over knowledge graphs, WWW 2017.
Abujabal et al., Never-ending learning for open-domain question answering over knowledge bases, WWW 2018.

Answering with templates

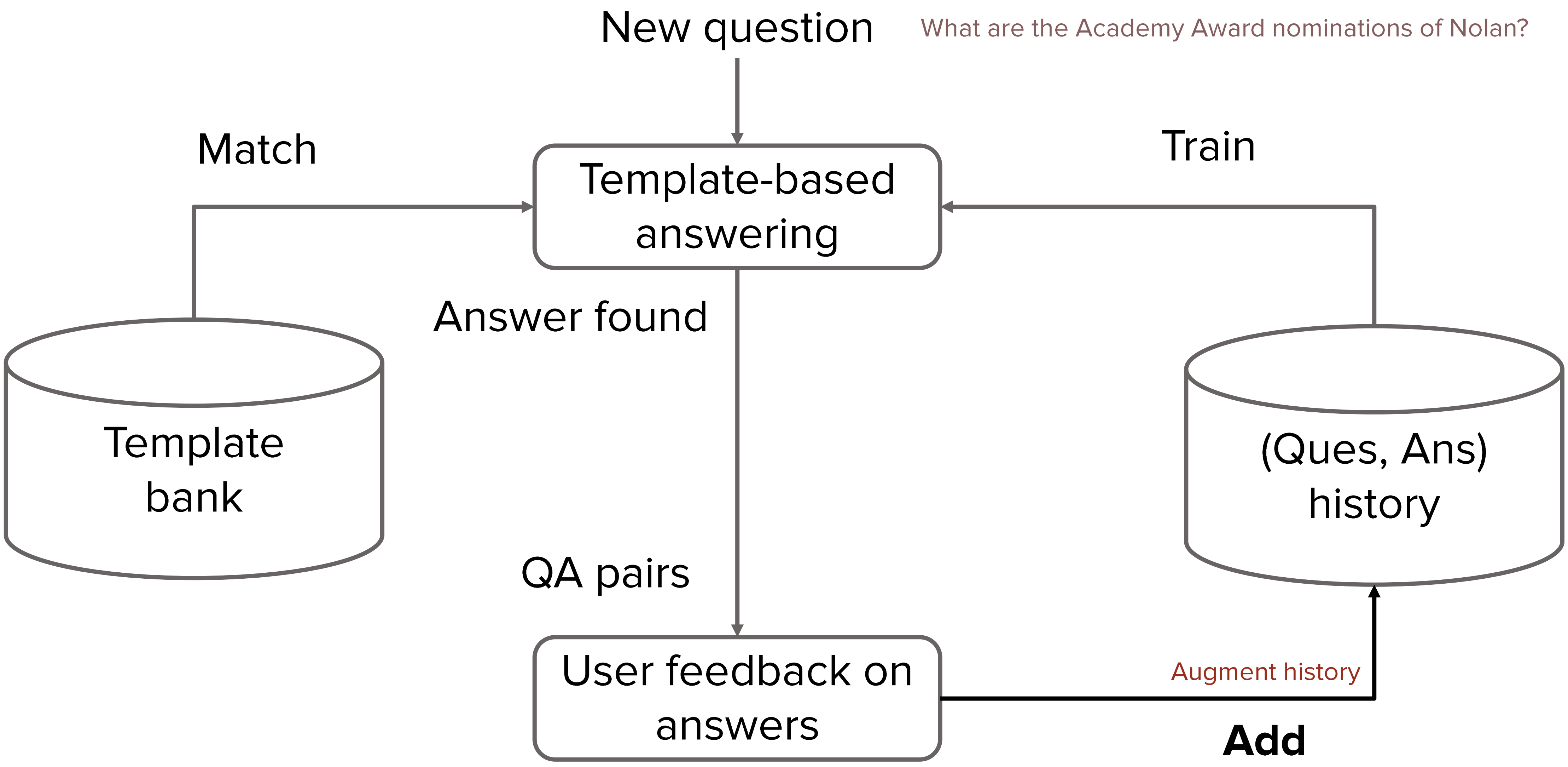


Abujabal et al., Automated template generation for question answering over knowledge graphs, WWW 2017.
Abujabal et al., Never-ending learning for open-domain question answering over knowledge bases, WWW 2018.

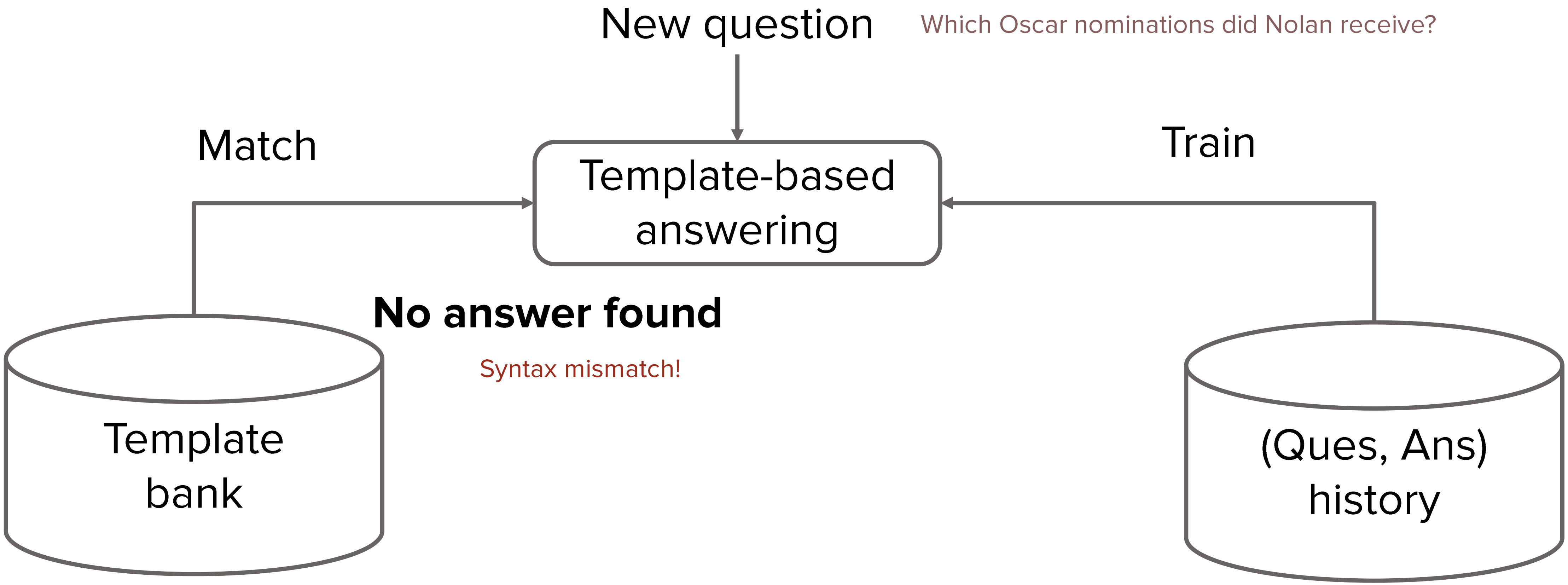
Close the loop with user feedback



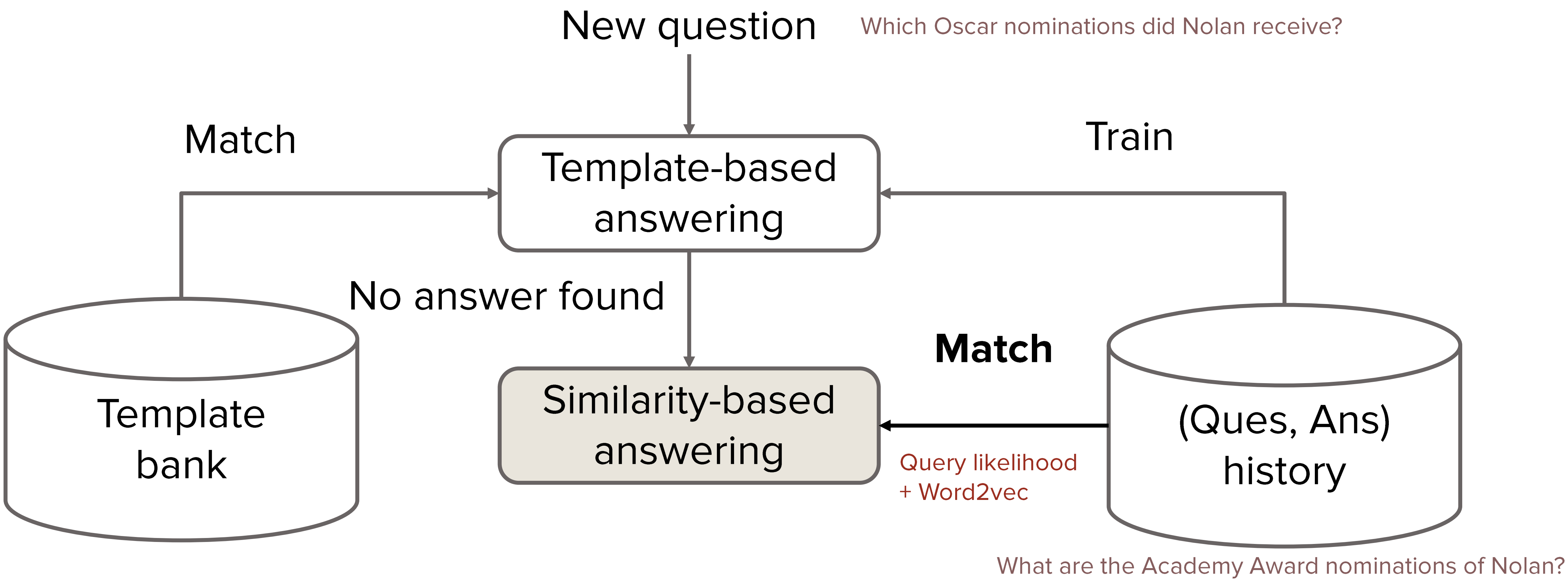
Augment history on positive feedback



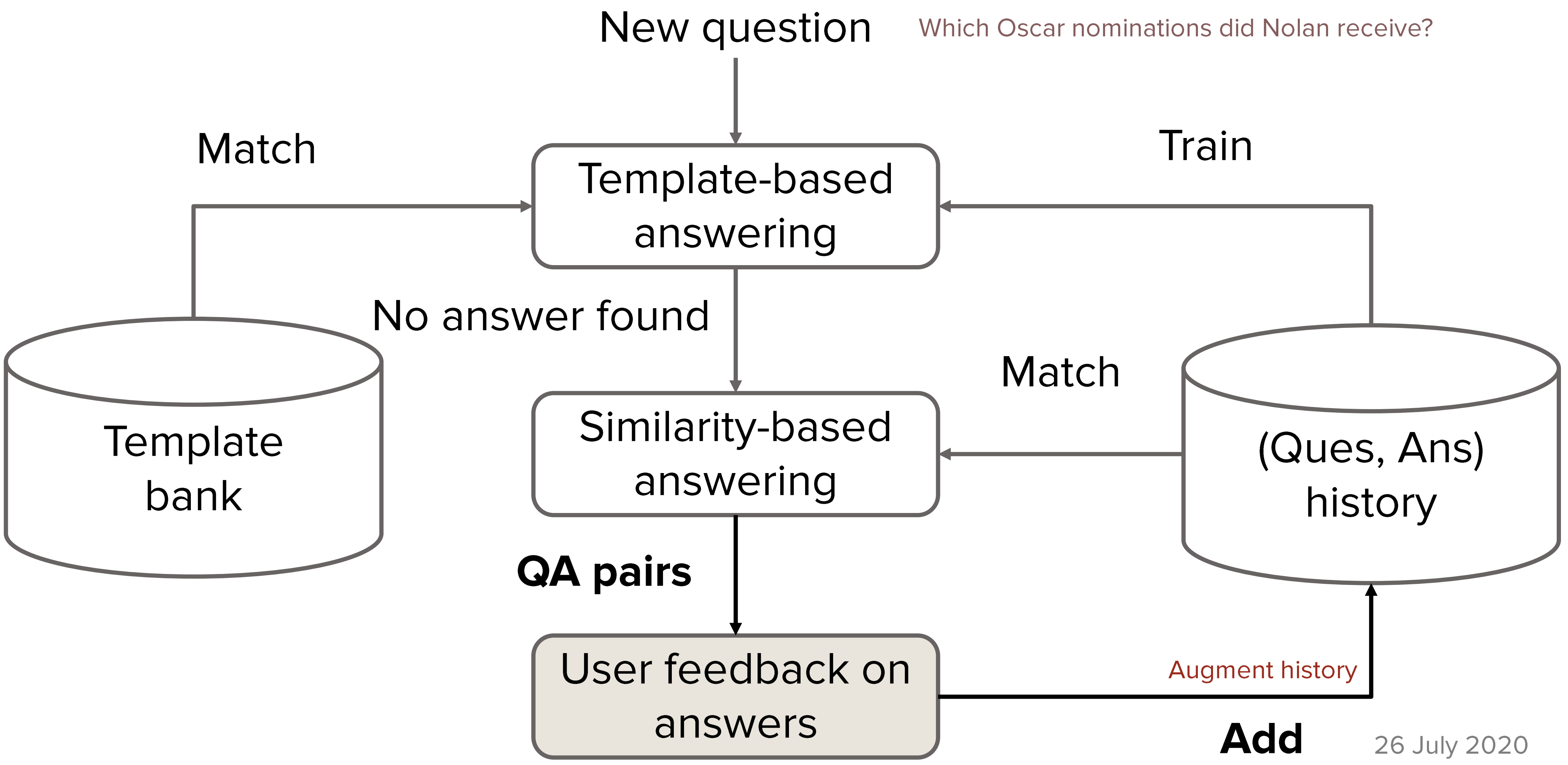
Templates can fail



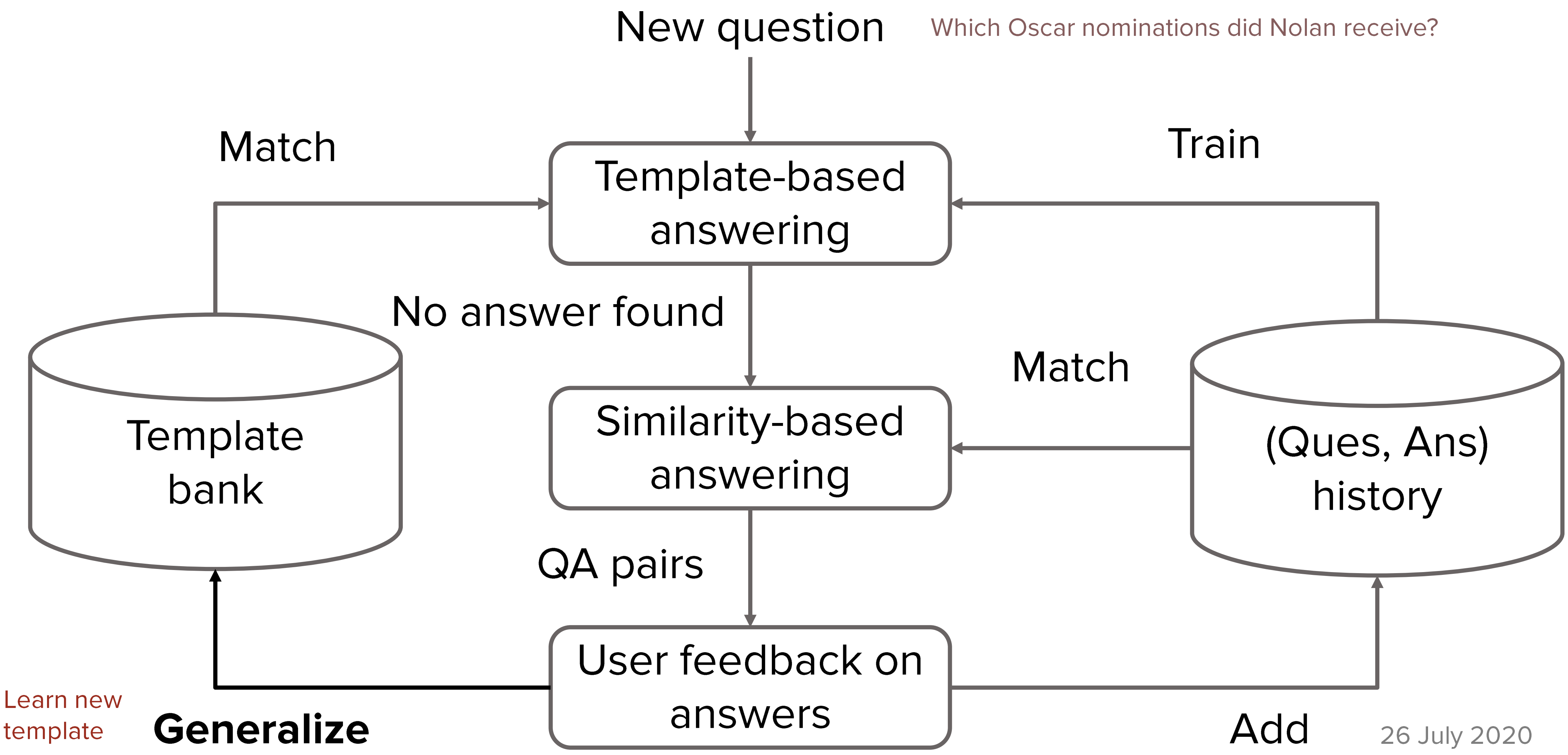
Invoke similarity-based answering



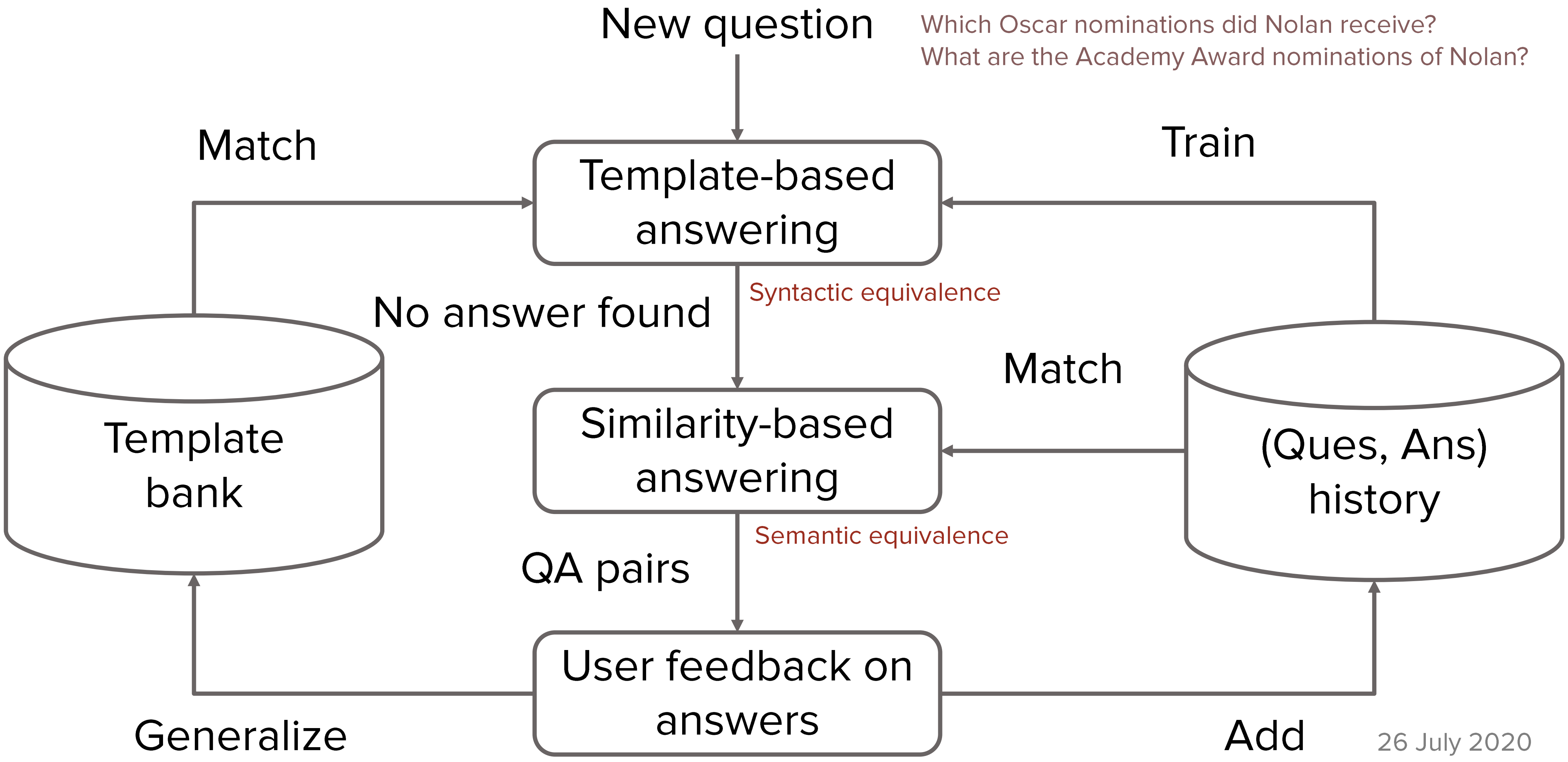
Augment history



Learn new template



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 graph embeddings

- The **KEQA** model (Huang et al. 2019)
- 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)

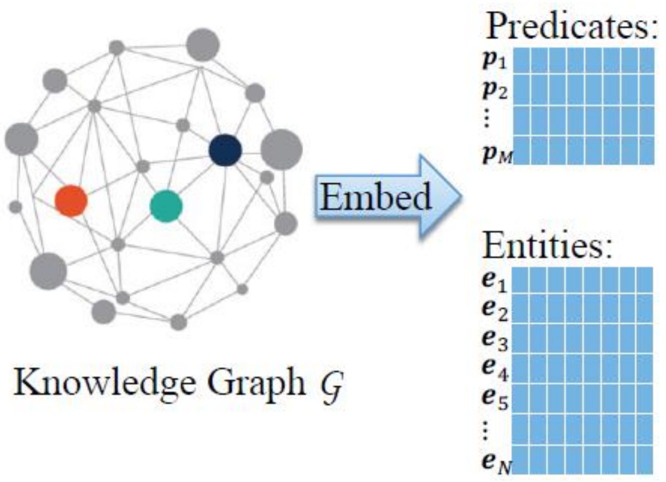
Huang et al., Knowledge graph embedding based question answering, WSDM 2019.

KEQA: Outline

Knowledge Graph

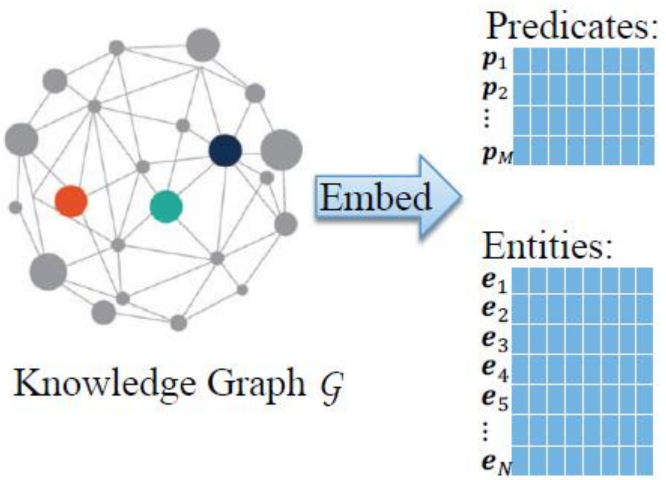
Huang et al., Knowledge graph embedding based question answering, WSDM 2019.

KEQA: Learn KG embeddings



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

KEQA: Using TransE



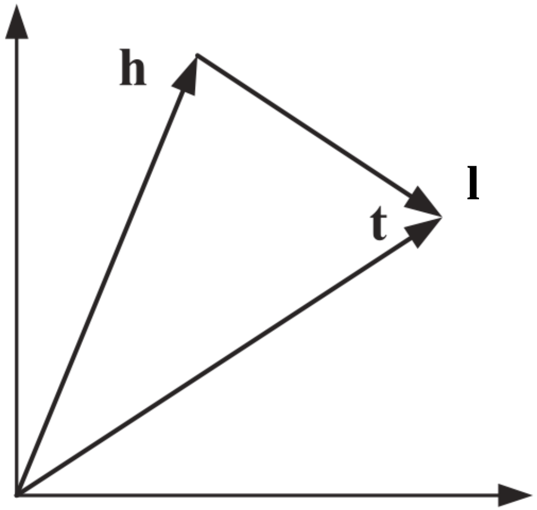
TransE (or TransE-like model) (Bordes et al. 2013)

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

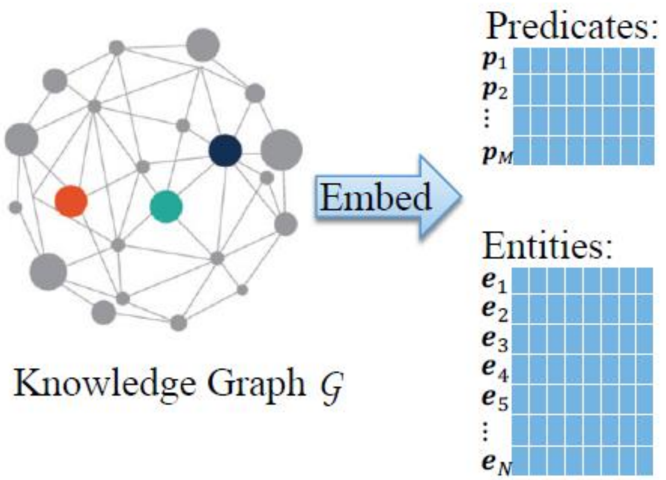
$$\mathcal{L} = \sum_{(h,\ell,t)\in S} \sum_{(h',\ell,t')\in S'_{(h,\ell,t)}} \left[\gamma + d(\boldsymbol{h}+\boldsymbol{\ell},\boldsymbol{t}) - d(\boldsymbol{h'}+\boldsymbol{\ell},\boldsymbol{t'})\right]_+$$

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

- L2-norm of entity embeddings 1, predicates unconstrained



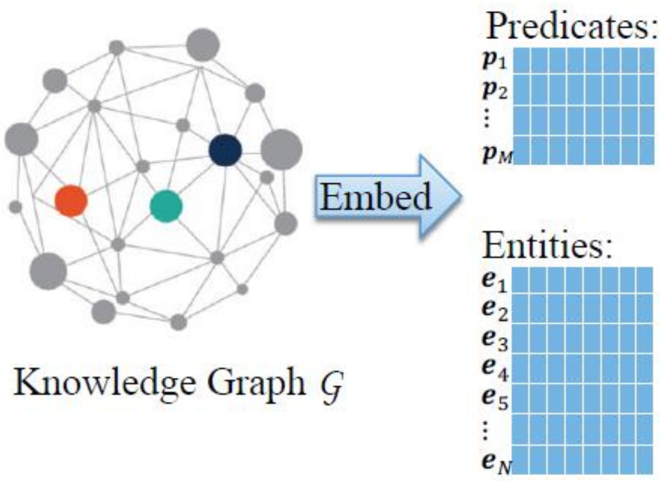
KEQA: Input question

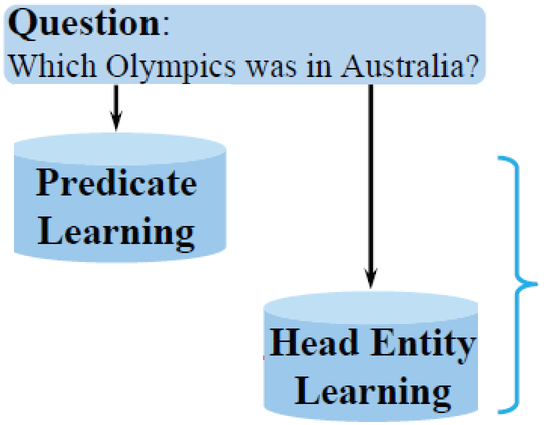


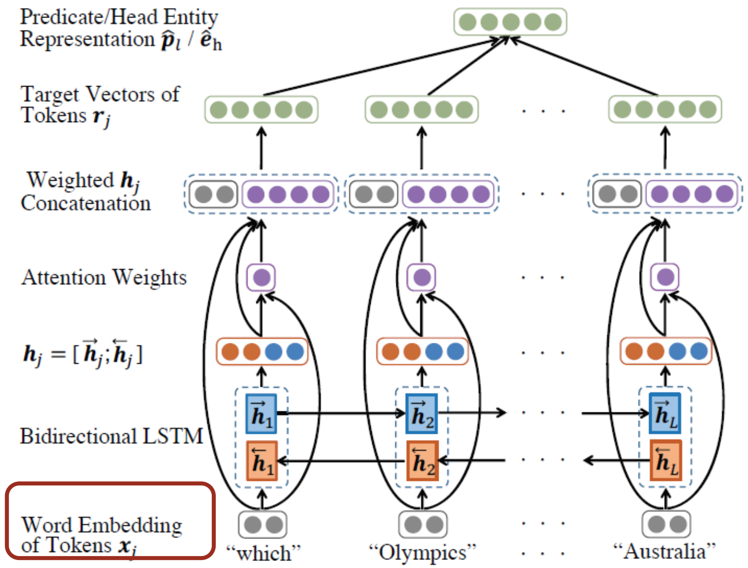
Question:
Which Olympics was in Australia?

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

KEQA: Learn to predict head and body





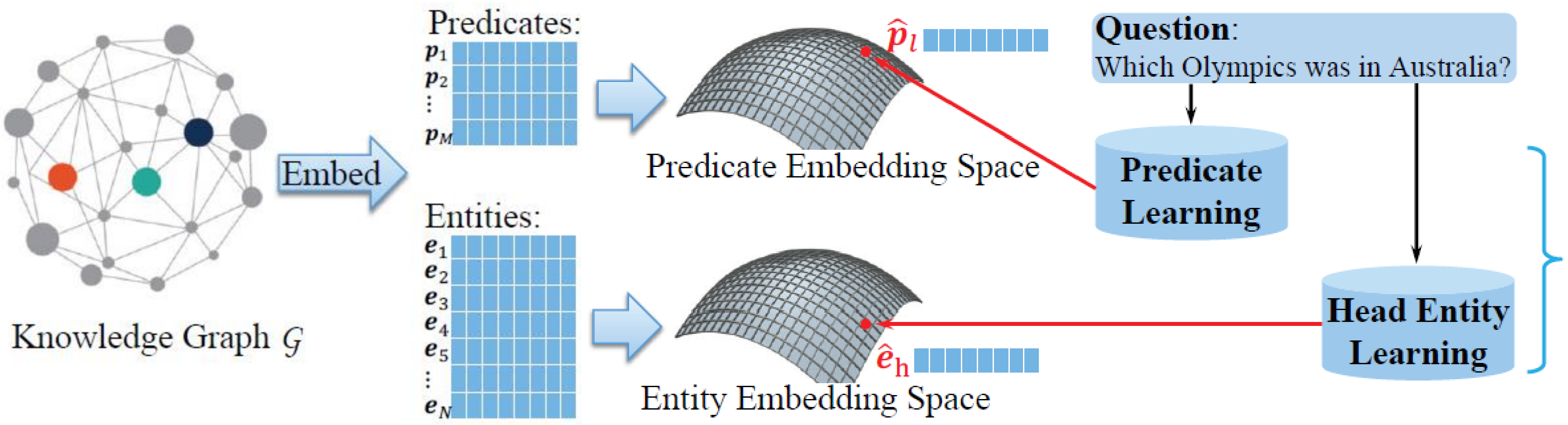


Bi-LSTM for word order

Attention for word importance

Learning representations generalization to unseen predicates at test time

KEQA: Use learnt models for prediction

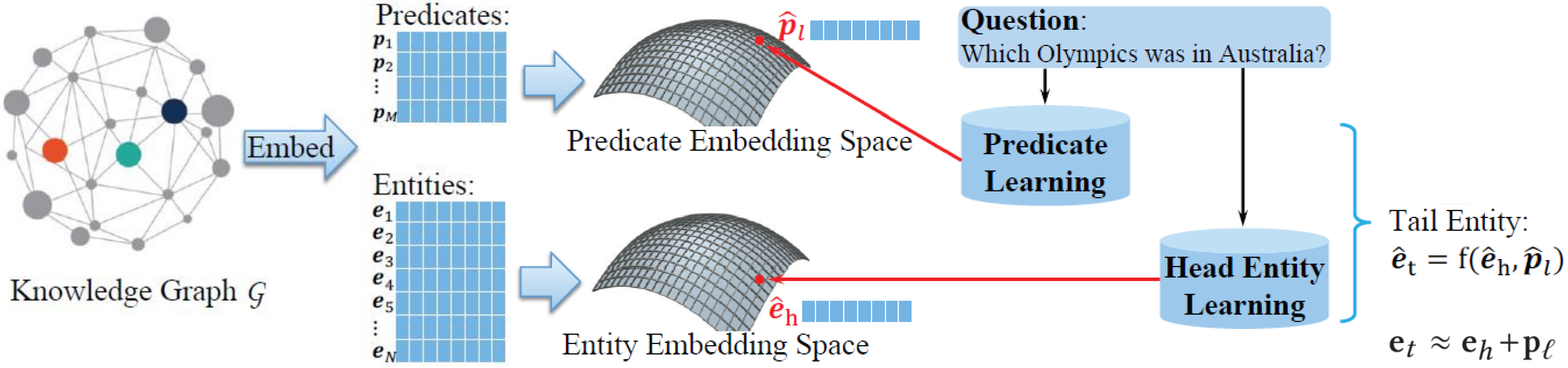


TransE (or TransE-like model) (Bordes et al. 2013)

- 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



$\mathbf{e}_t \approx \mathbf{e}_h + \mathbf{p}_\ell$

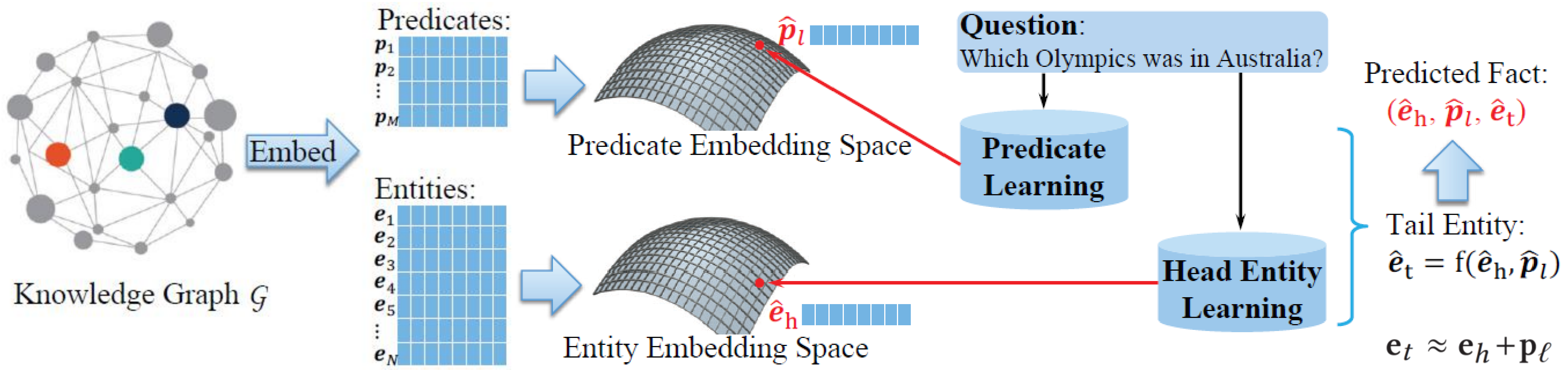
For TransE

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



$\mathbf{e}_t \approx \mathbf{e}_h + \mathbf{p}_\ell$

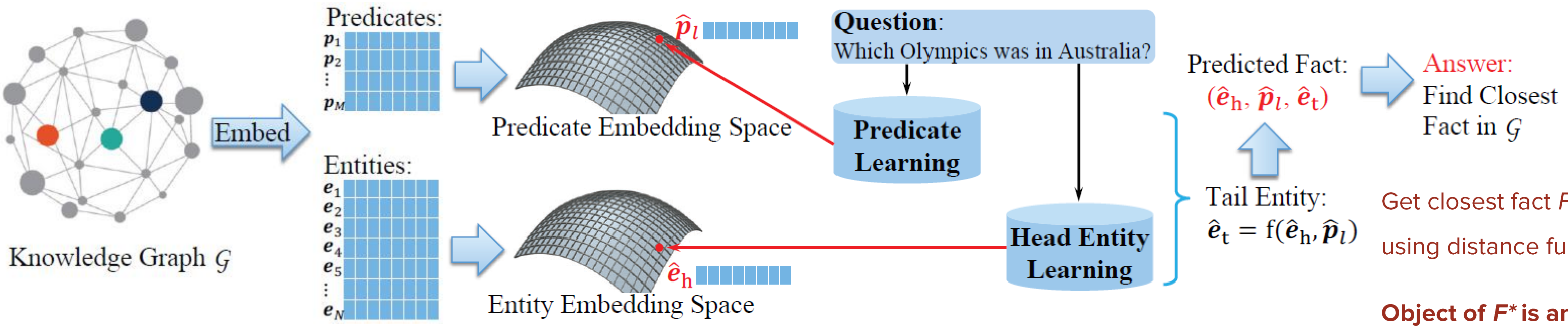
For TransE

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



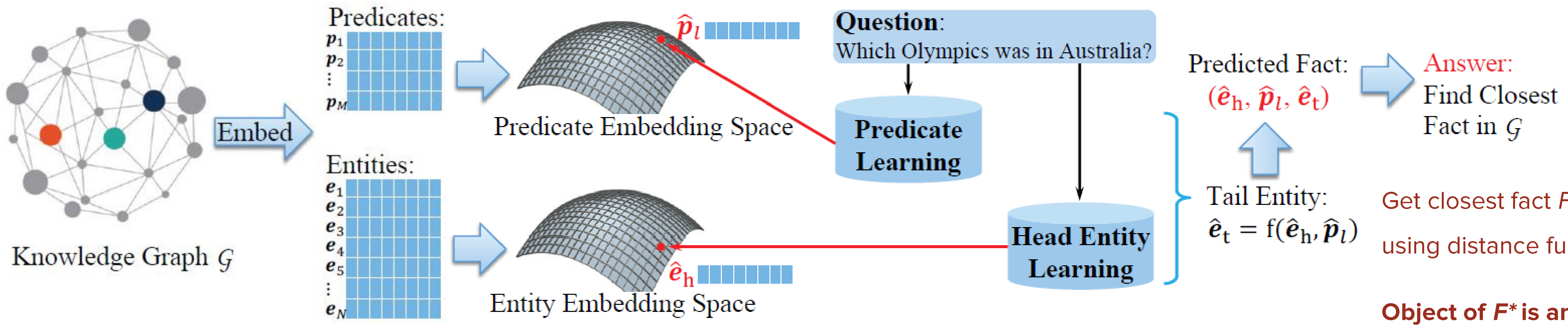
Get closest fact F^* in KG using distance function

Object of F^* is answer

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

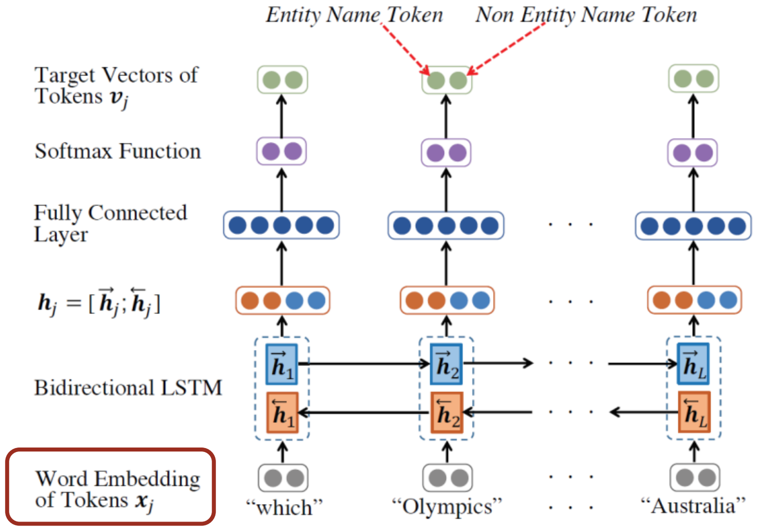
Huang et al., Knowledge graph embedding based question answering, WSDM 2019.

KEQA: Closest fact to answer



Get closest fact F^* in KG using distance function

Object of F^* is answer



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

$$- \beta_3 sim[n(h), \text{HED}_{\text{entity}}] - \beta_4 sim[n(\ell), \text{HED}_{\text{non}}],$$

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:** 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 Ding et al. (2019), Hu et al. (2018), Luo et al. (2018), **Bhutani et al. (2019)**, **Lu et al. (2019)**, **Vakulenko 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?

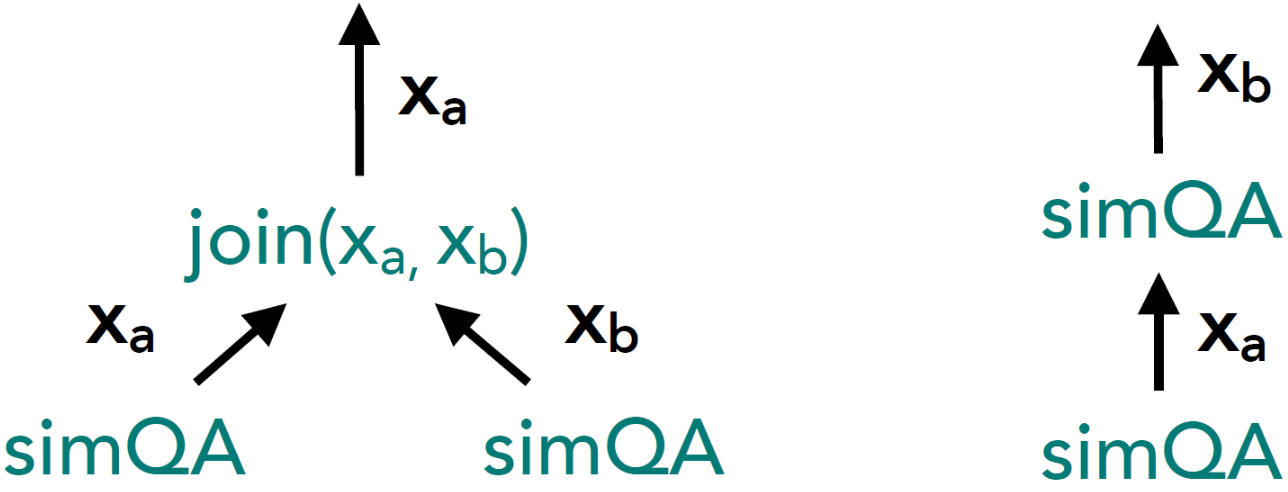
Complex QA: Structured query generation

- 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

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

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

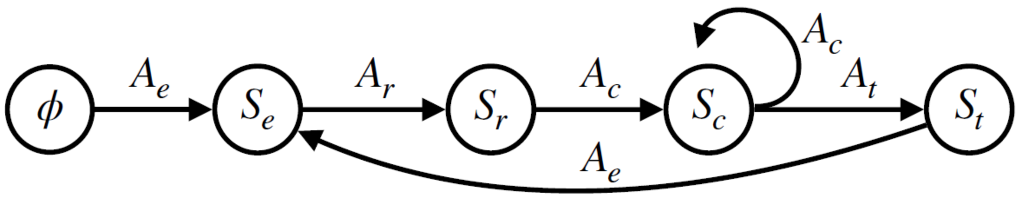
Chain join

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

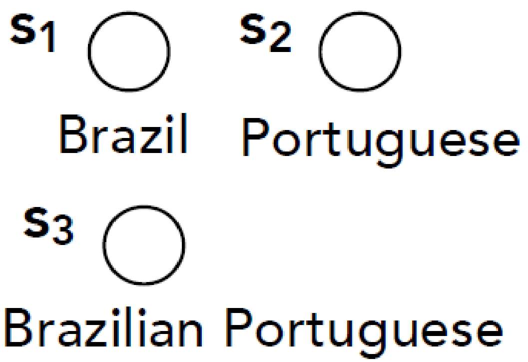
Bhutani et al., Learning to Answer Complex Questions over Knowledge Bases with Query Composition, CIKM 2019.

TextRay: Walkthrough

Which Portuguese speaking countries import fish from Brazil?



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



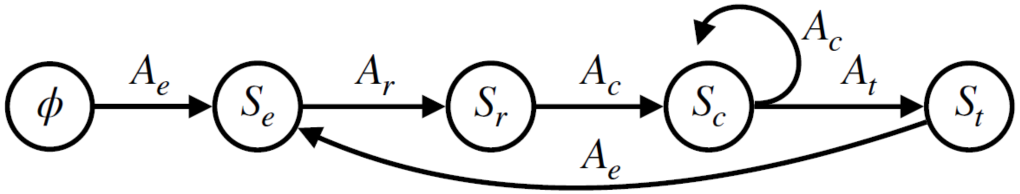
a) Identify seed

Top-k entities

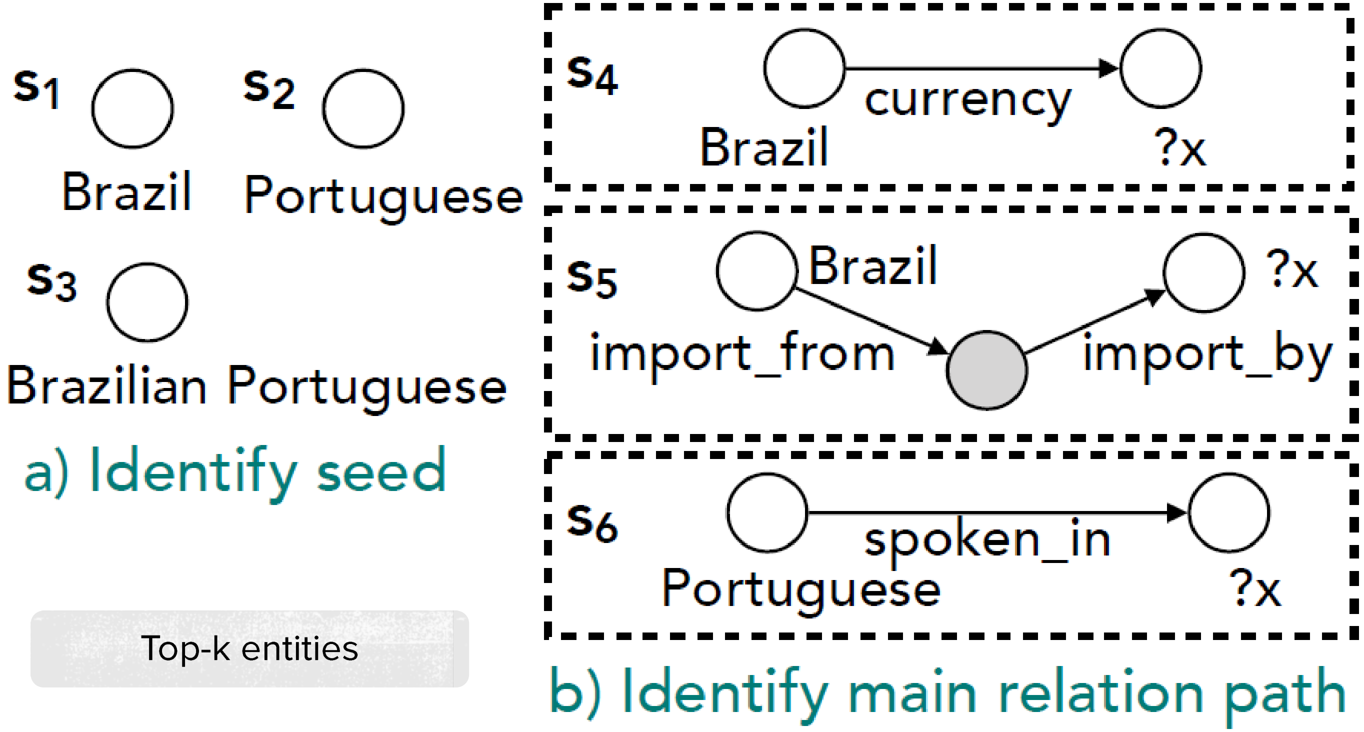
Bhutani et al., Learning to Answer Complex Questions over Knowledge Bases with Query Composition, CIKM 2019.

TextRay: Partial query graph

Which Portuguese speaking countries import fish from Brazil?



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

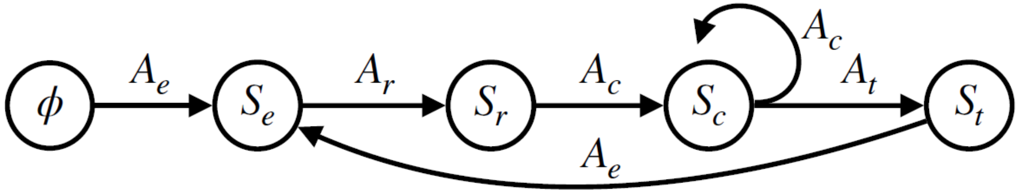


Top-k entities

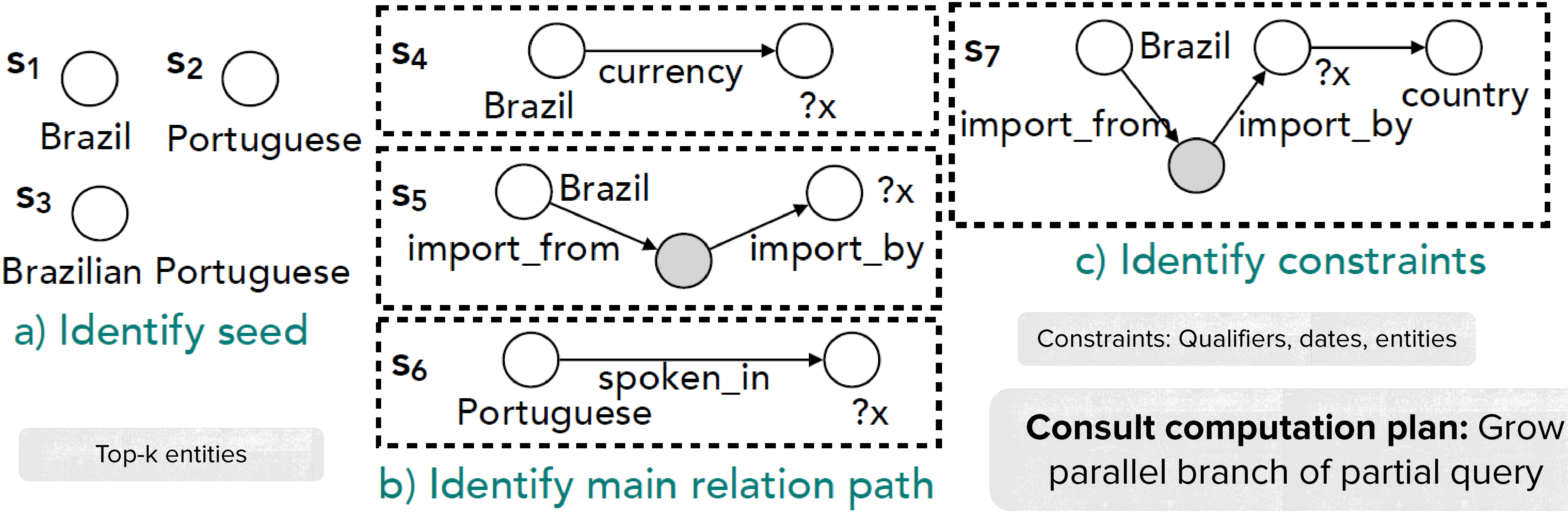
Bhutani et al., Learning to Answer Complex Questions over Knowledge Bases with Query Composition, CIKM 2019.

TextRay: Partial query graph

Which Portuguese speaking countries import fish from Brazil?



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



Top-k entities

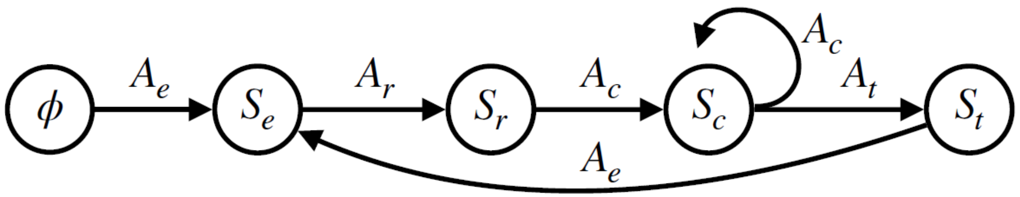
Constraints: Qualifiers, dates, entities

Consult computation plan: Grow parallel branch of partial query

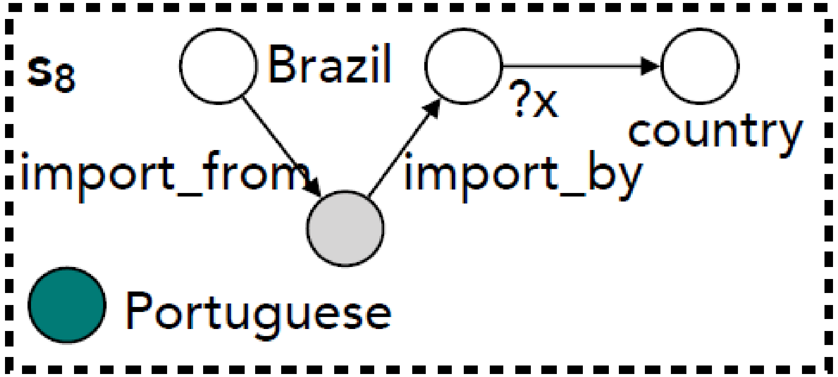
Bhutani et al., Learning to Answer Complex Questions over Knowledge Bases with Query Composition, CIKM 2019.

TextRay: Partial query graph

Which Portuguese speaking countries import fish from Brazil?



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

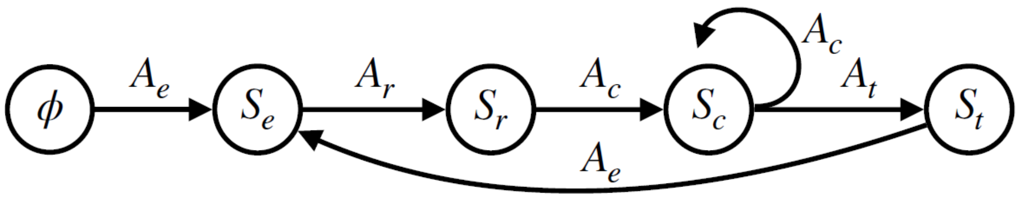


a) Identify seed

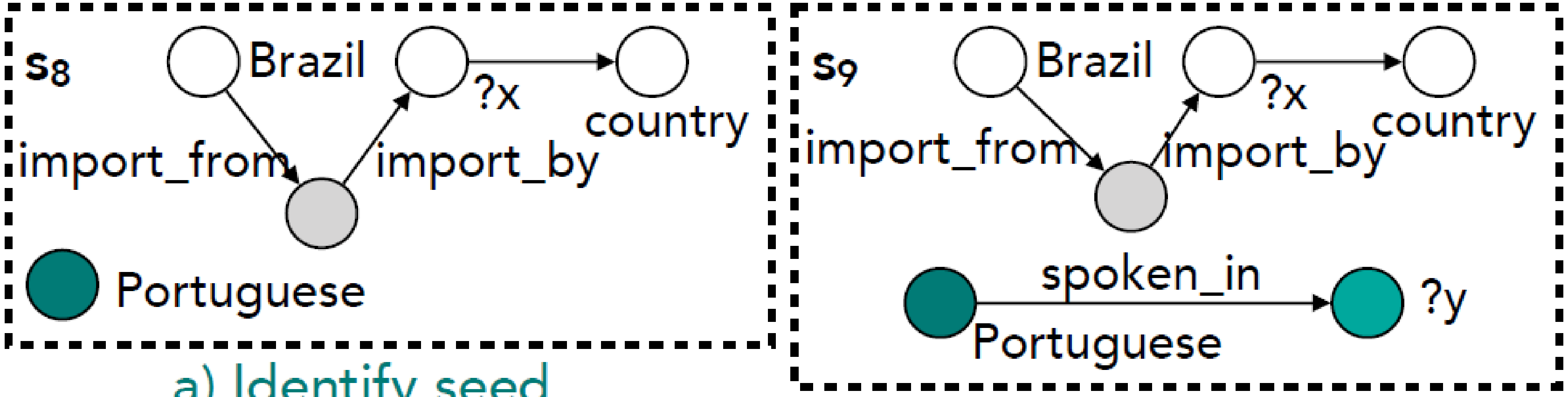
Bhutani et al., Learning to Answer Complex Questions over Knowledge Bases with Query Composition, CIKM 2019.

TextRay: Partial query graph

Which Portuguese speaking countries import fish from Brazil?



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



a) Identify seed

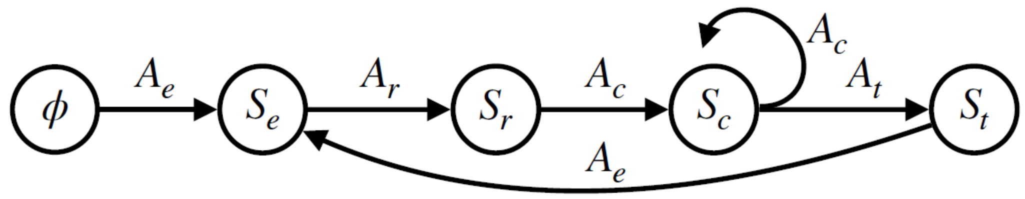
b) Identify main relation path

c) Identify constraints

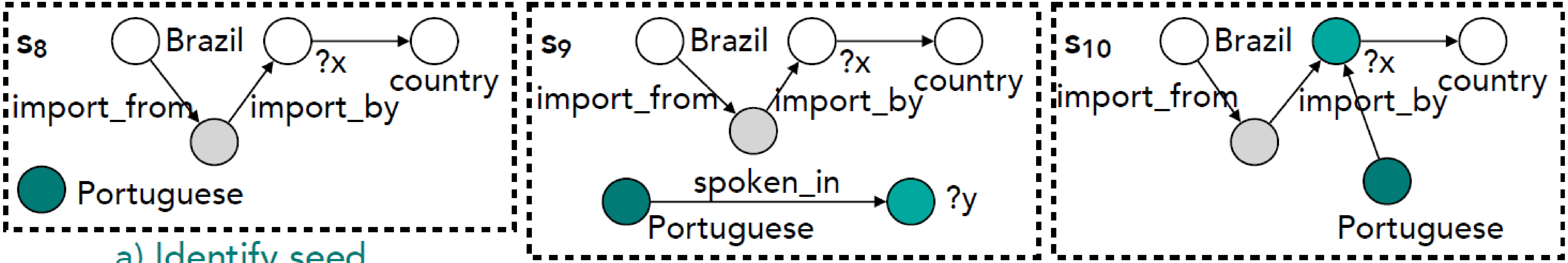
Bhutani et al., Learning to Answer Complex Questions over Knowledge Bases with Query Composition, CIKM 2019.

TextRay: Partial query graph

Which Portuguese speaking countries import fish from Brazil?



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



a) Identify seed

b) Identify main relation path

c) Identify constraints

d) Compose/Execute

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 in quasi KG
- **Unsupervised compact subgraph computation:** Compute Group Steiner Tree (GST) with cornerstones as terminals for joint disambiguation of question concepts

Lu et al., Answering Complex Questions by Joining Multi-Document Evidence with Quasi Knowledge Graphs, SIGIR 2019.

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>

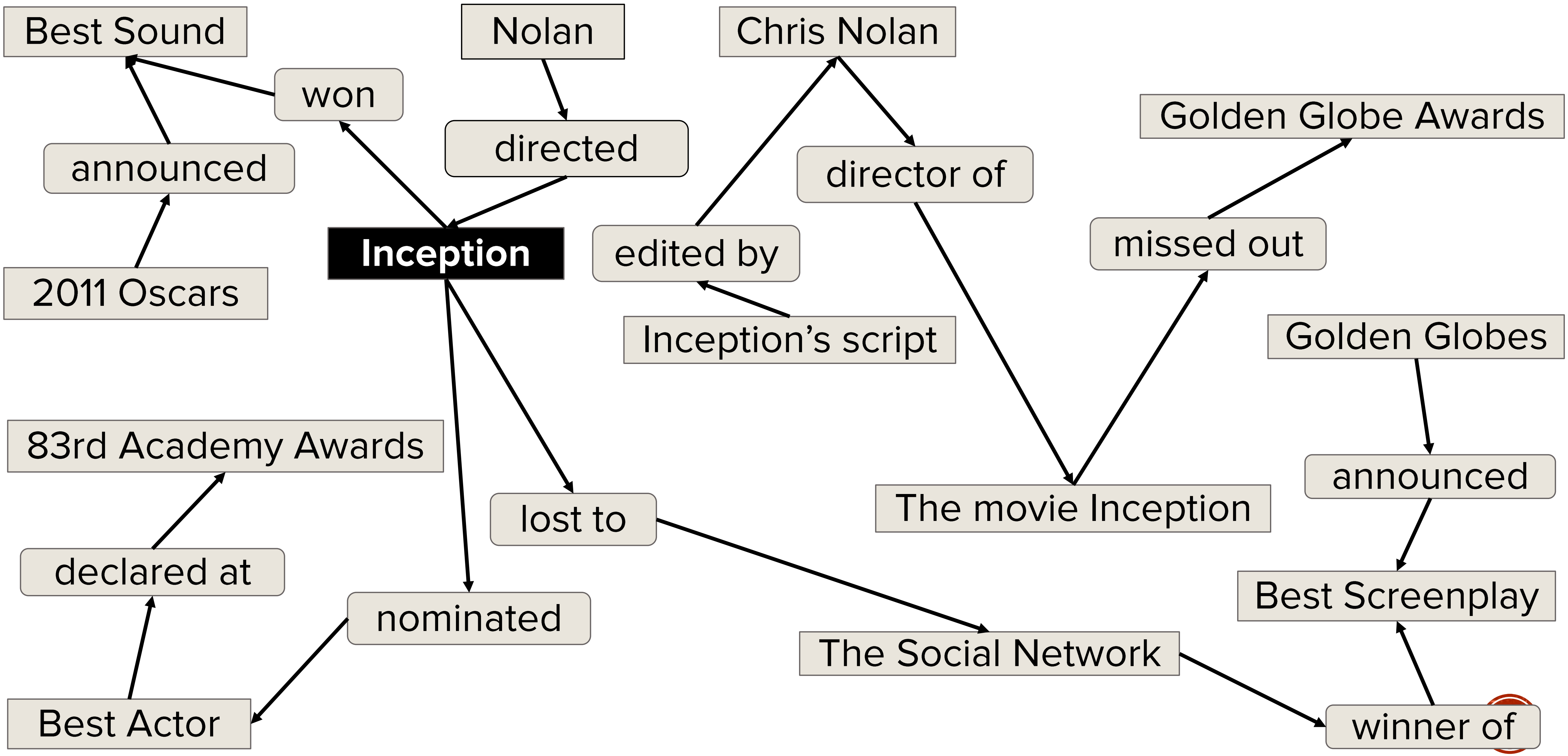
Compile an open-vocabulary triple store

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

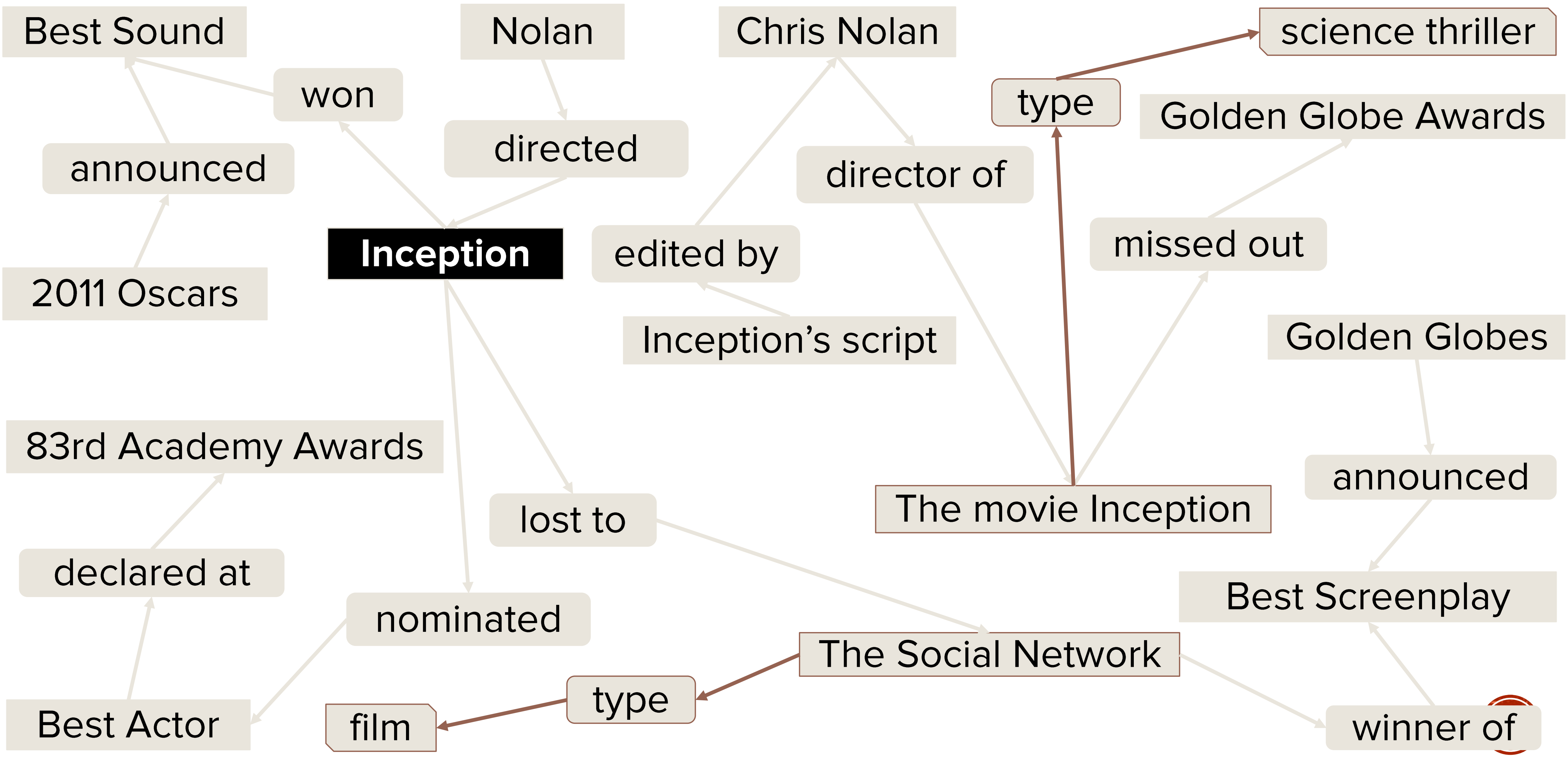
Open IE extracts KG-style triples by running pattern extraction over raw text: Stanford Open IE, ClausIE, OpenIE 5.0, ...

Lu et al., Answering Complex Questions by Joining Multi-Document Evidence with Quasi Knowledge Graphs, SIGIR 2019.

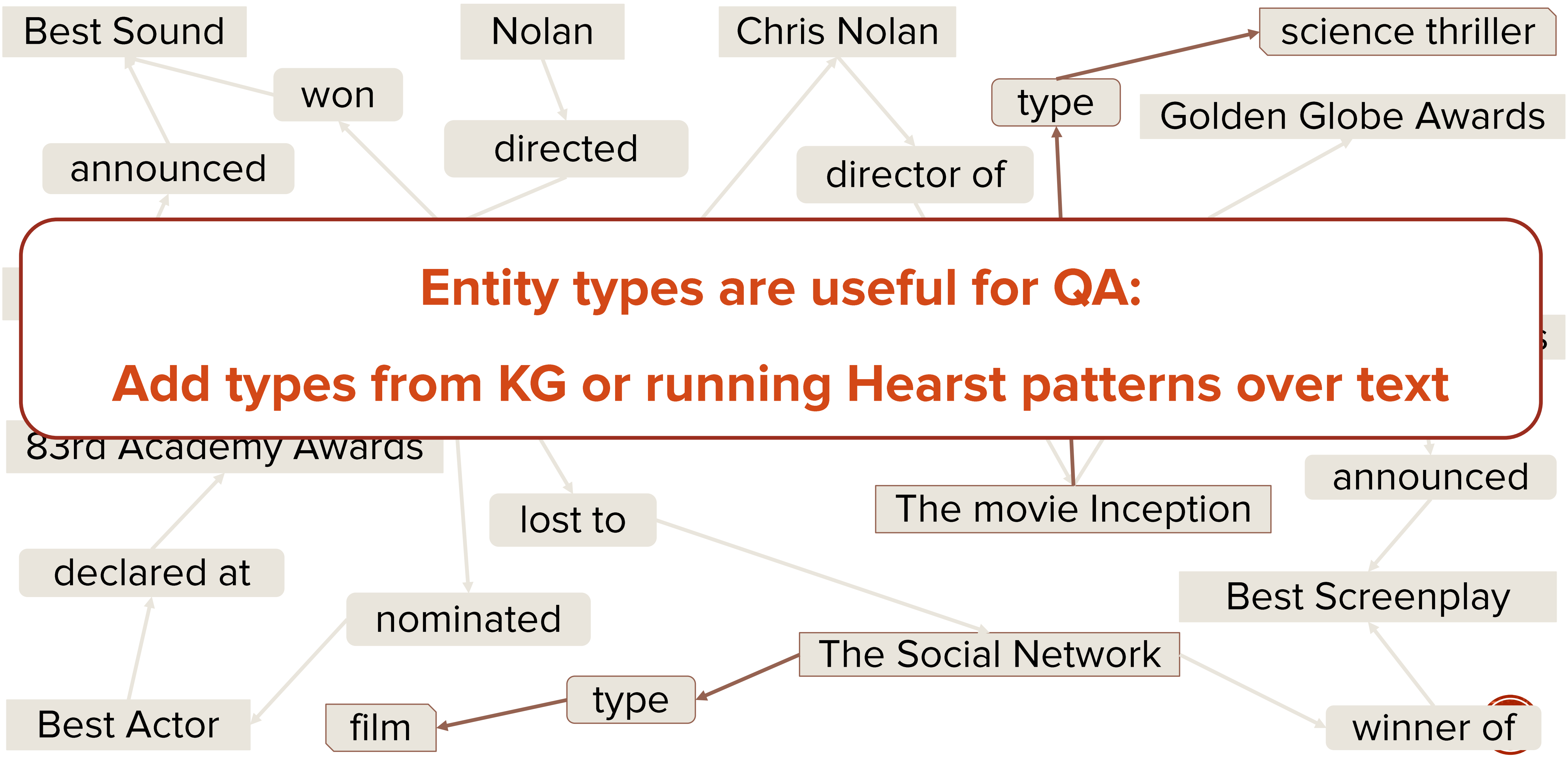
Question: Which Nolan films won an Oscar but missed a Golden Globe?



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



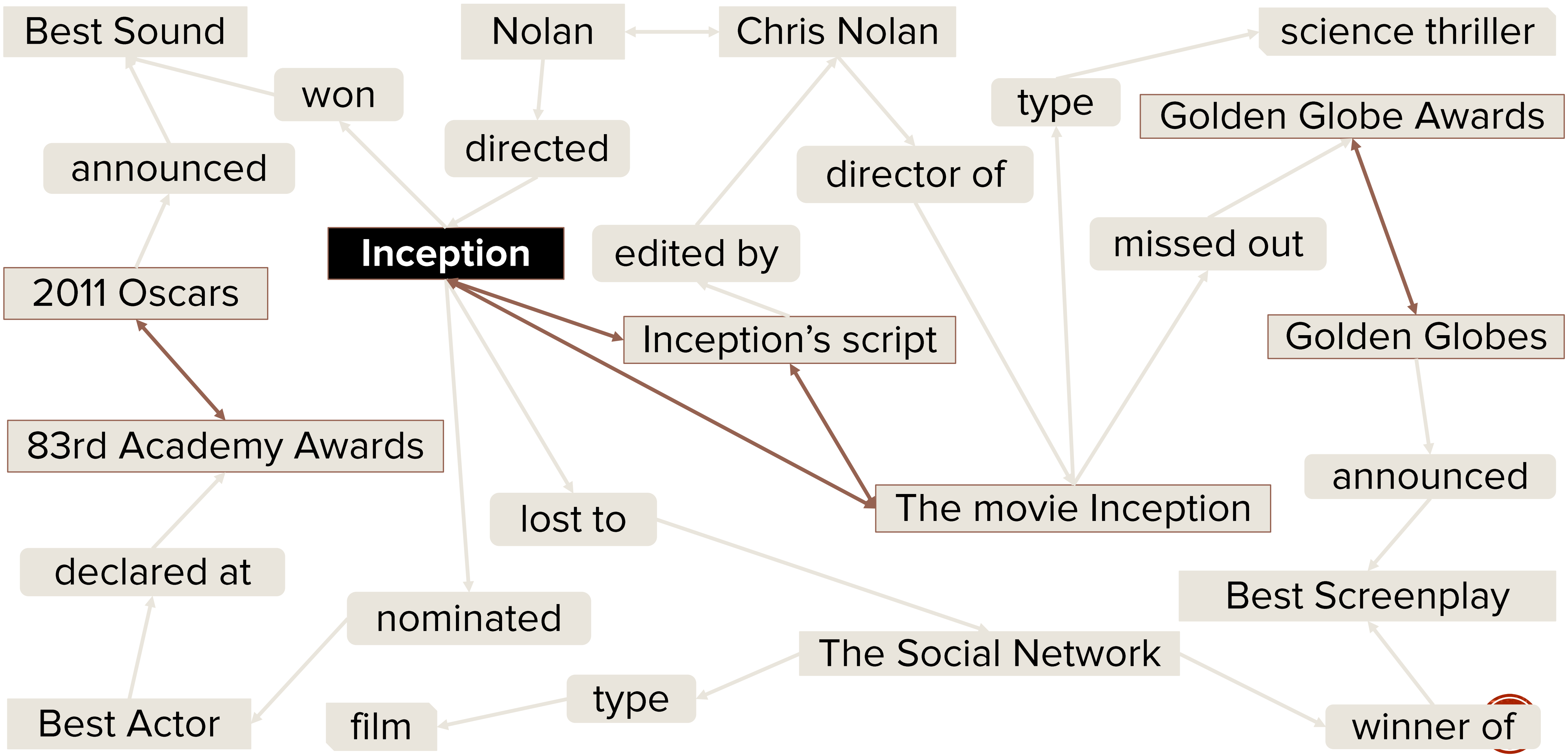
Question: Which Nolan films won an Oscar but missed a Golden Globe?



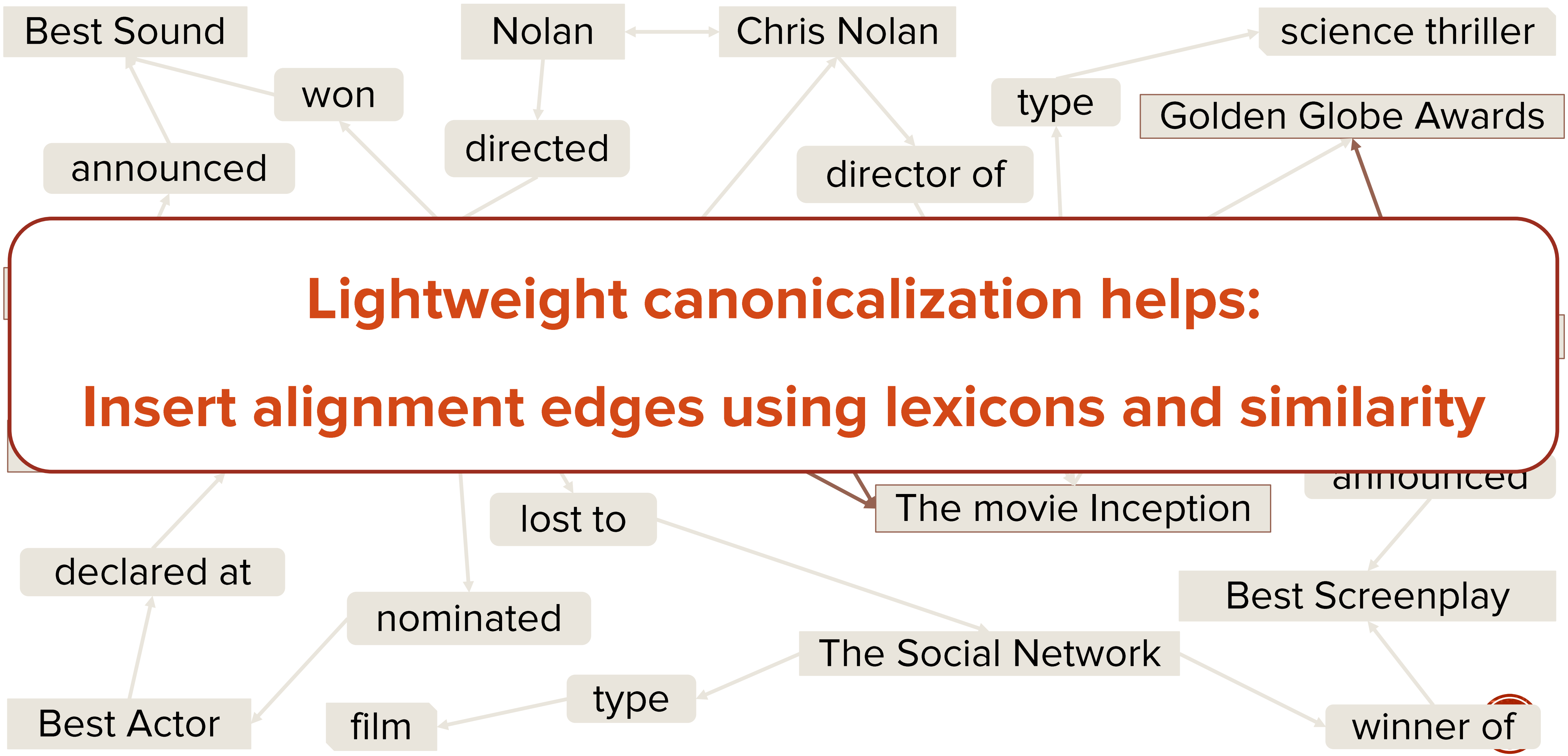
Entity types are useful for QA:

Add types from KG or running Hearst patterns over text

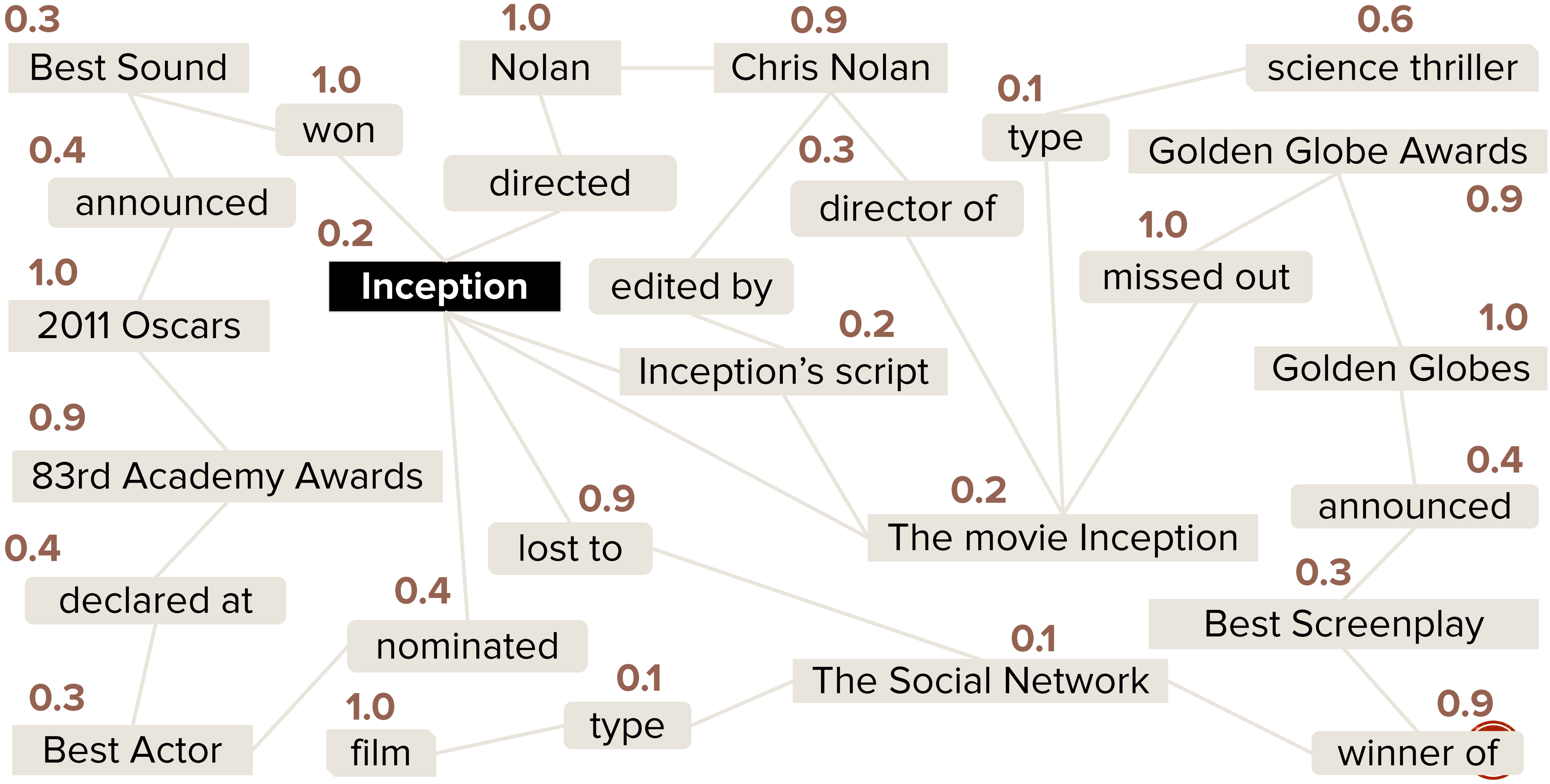
Question: Which Nolan films won an Oscar but missed a Golden Globe?



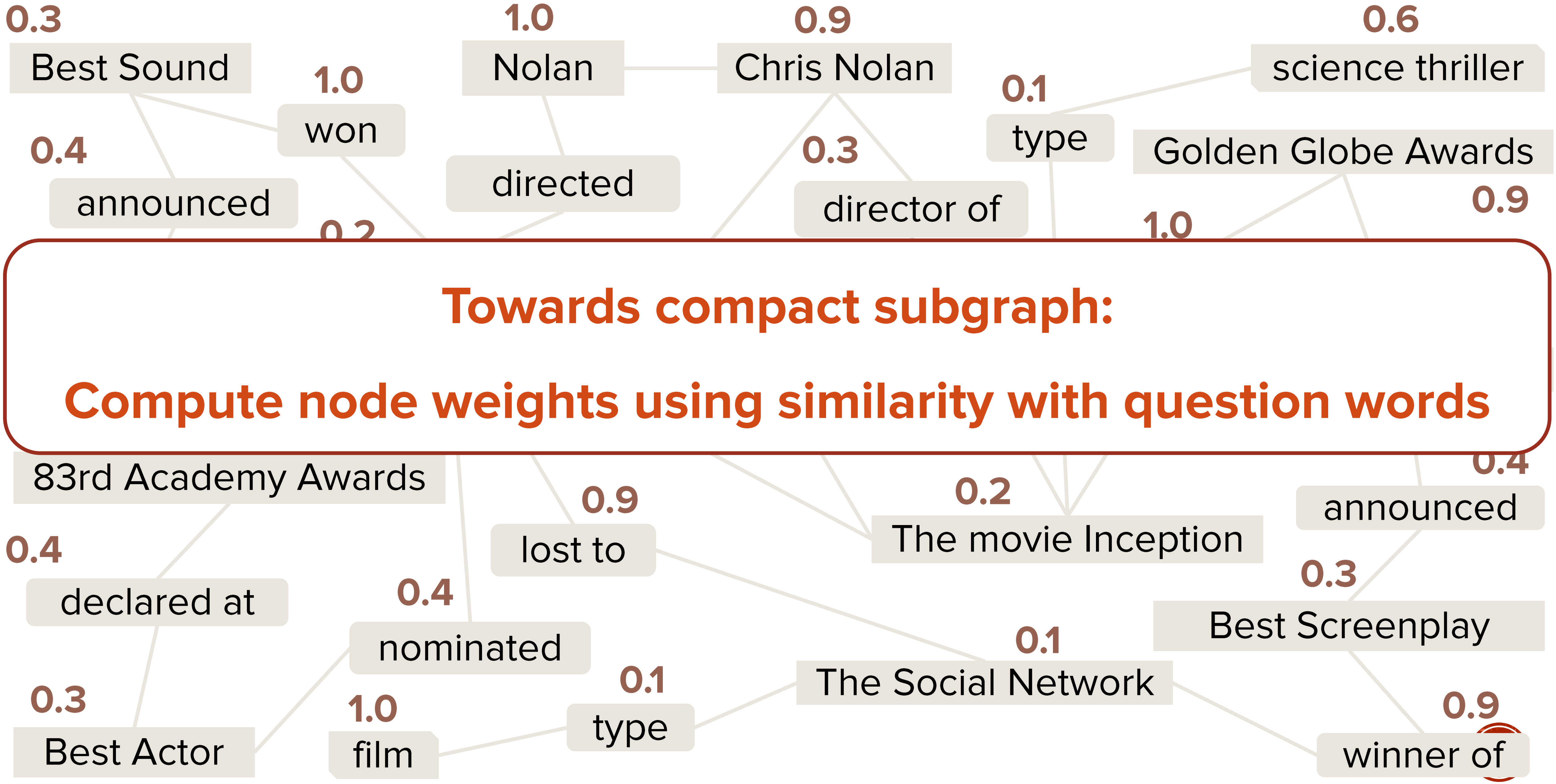
Question: Which Nolan films won an Oscar but missed a Golden Globe?



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



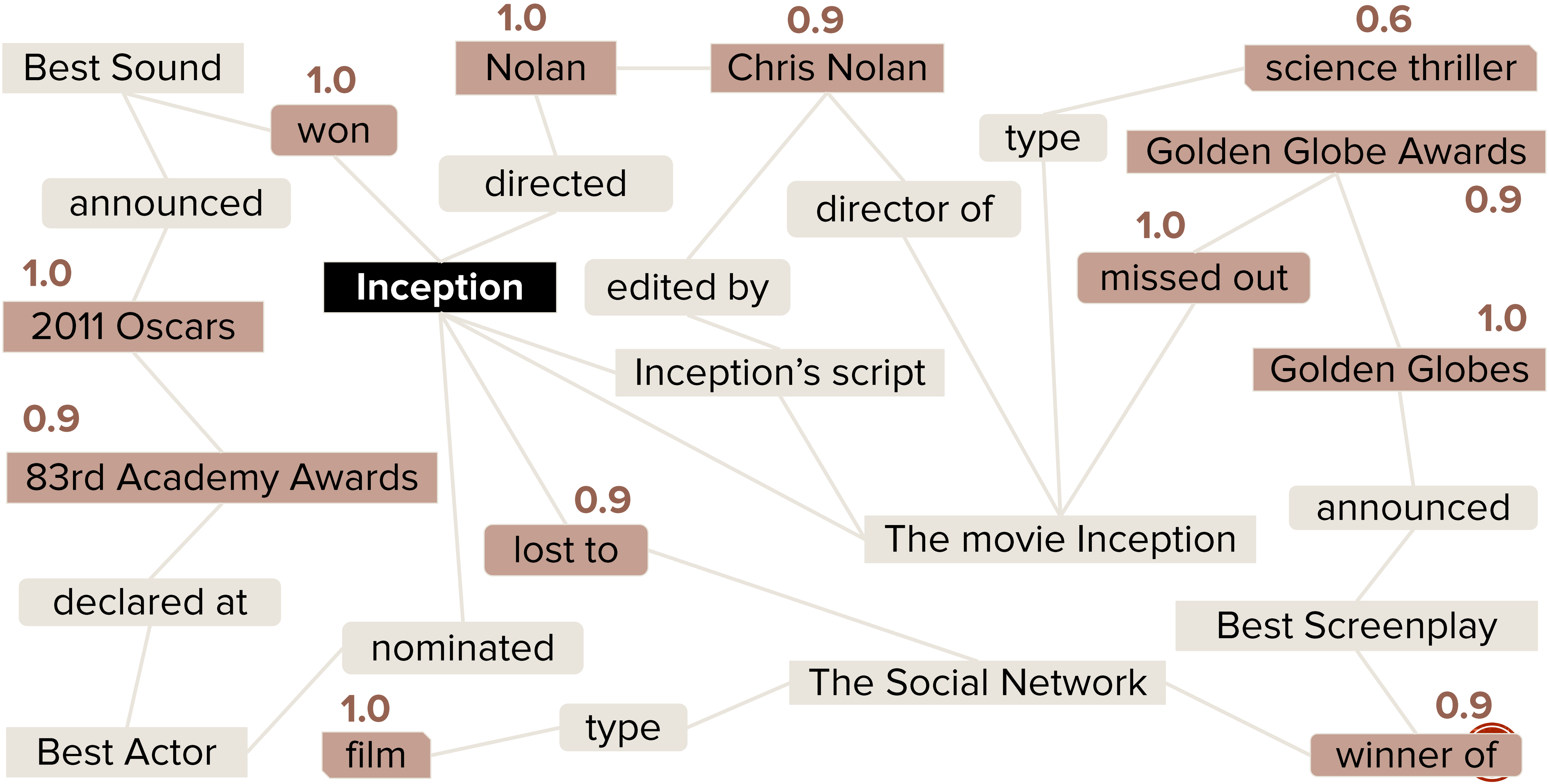
Question: Which Nolan films won an Oscar but missed a Golden Globe?



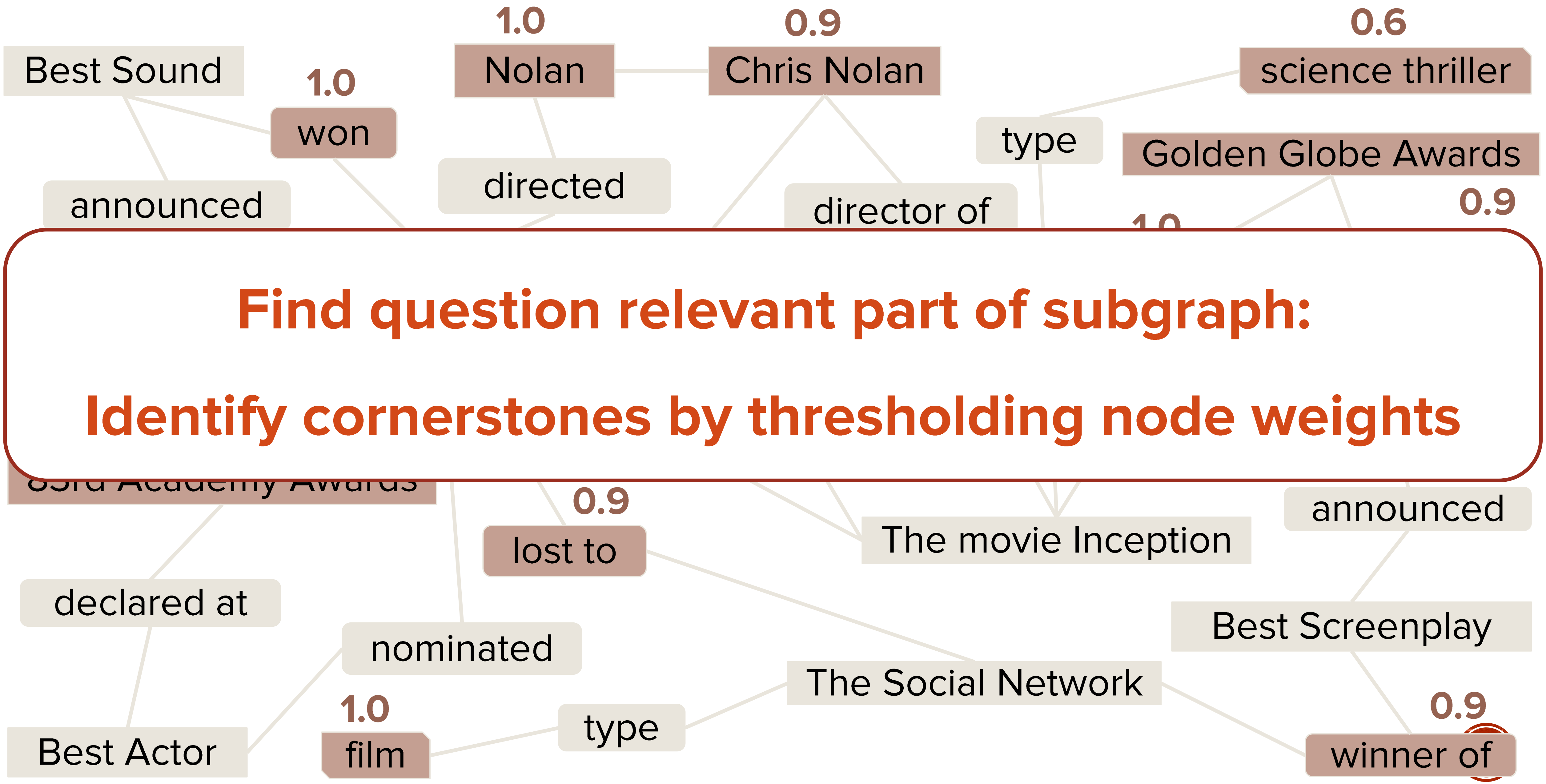
Towards compact subgraph:

Compute node weights using similarity with question words

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



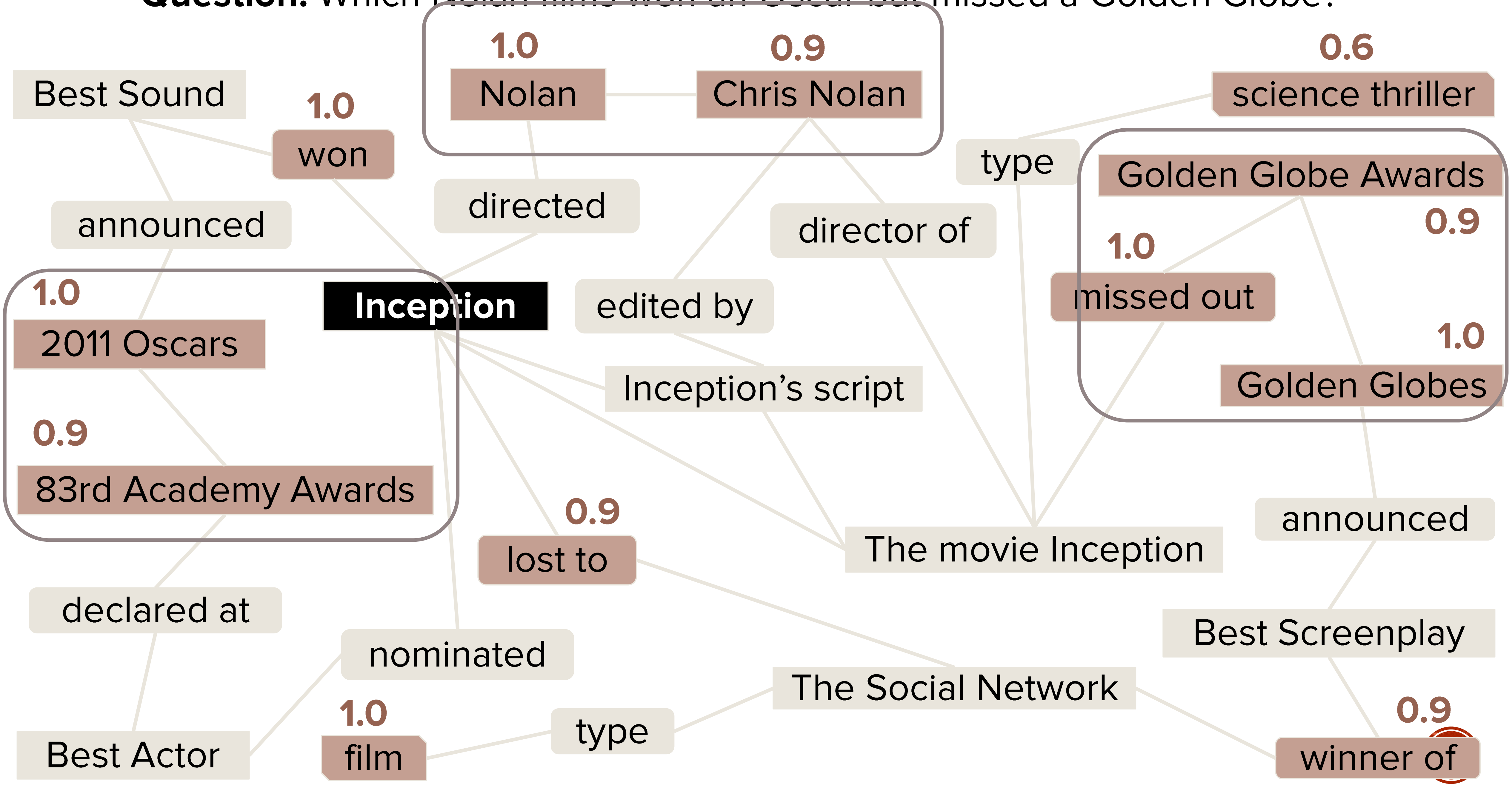
Question: Which Nolan films won an Oscar but missed a Golden Globe?



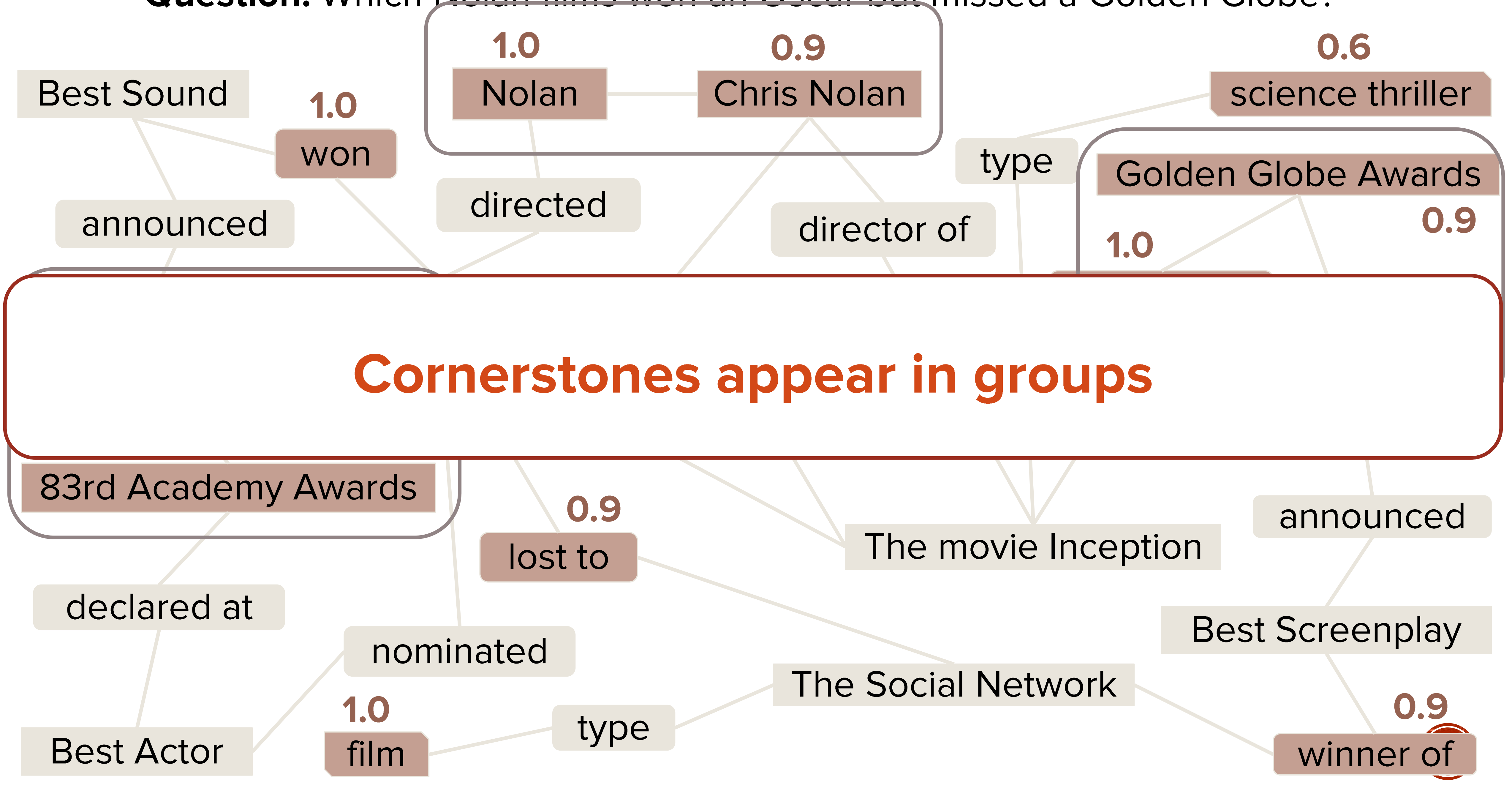
Find question relevant part of subgraph:

Identify cornerstones by thresholding node weights

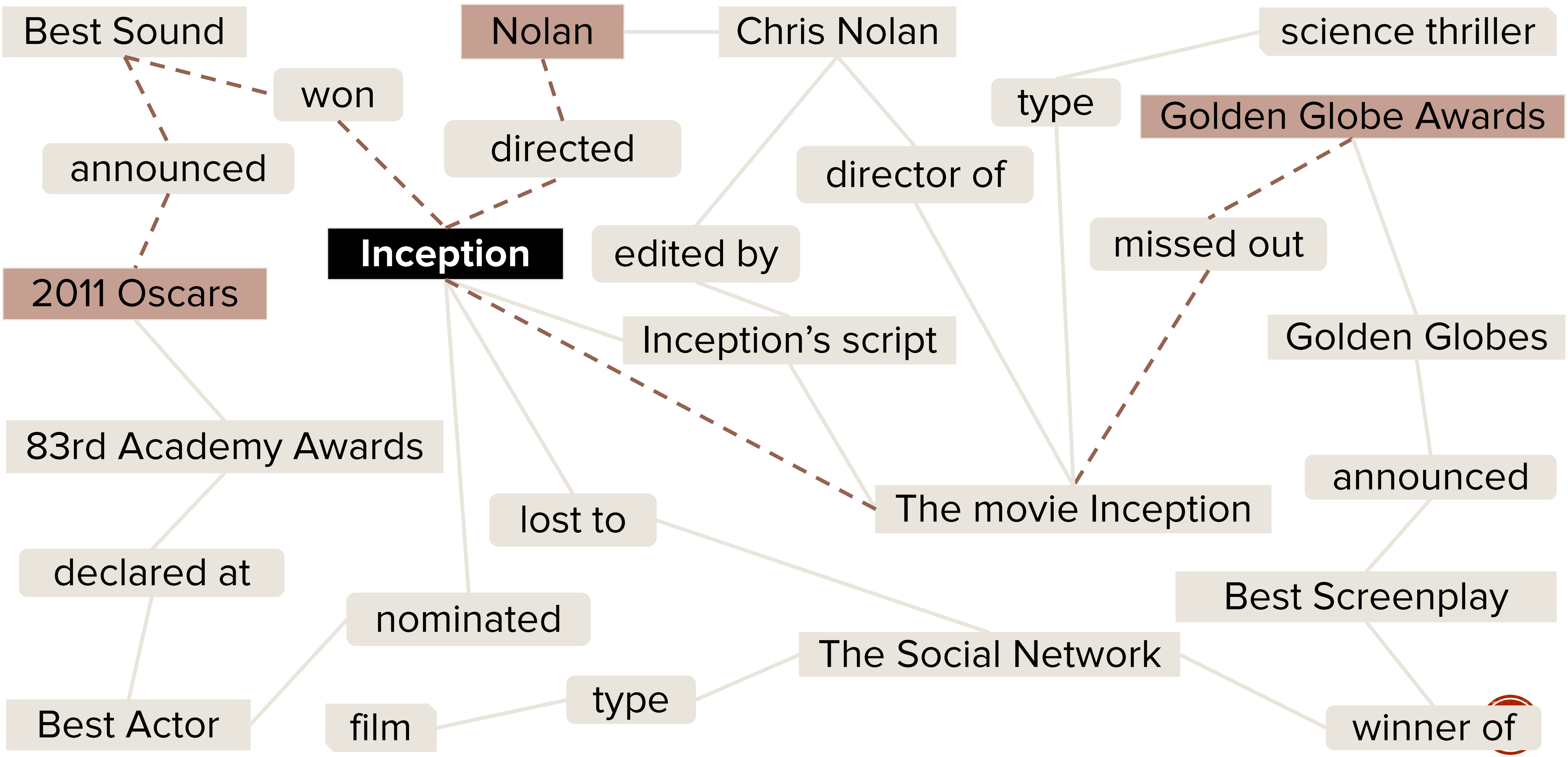
Question: Which Nolan films won an Oscar but missed a Golden Globe?



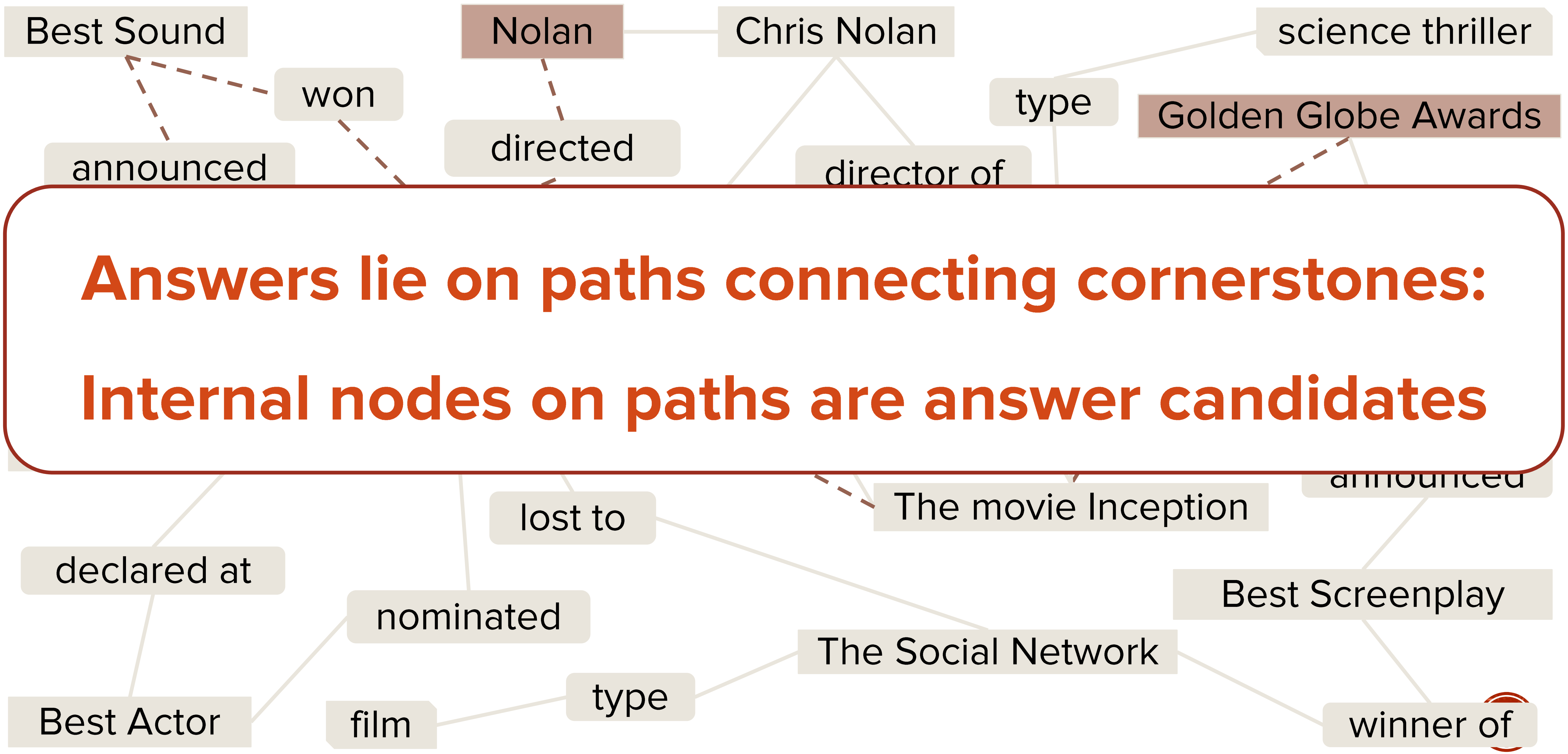
Question: Which Nolan films won an Oscar but missed a Golden Globe?



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



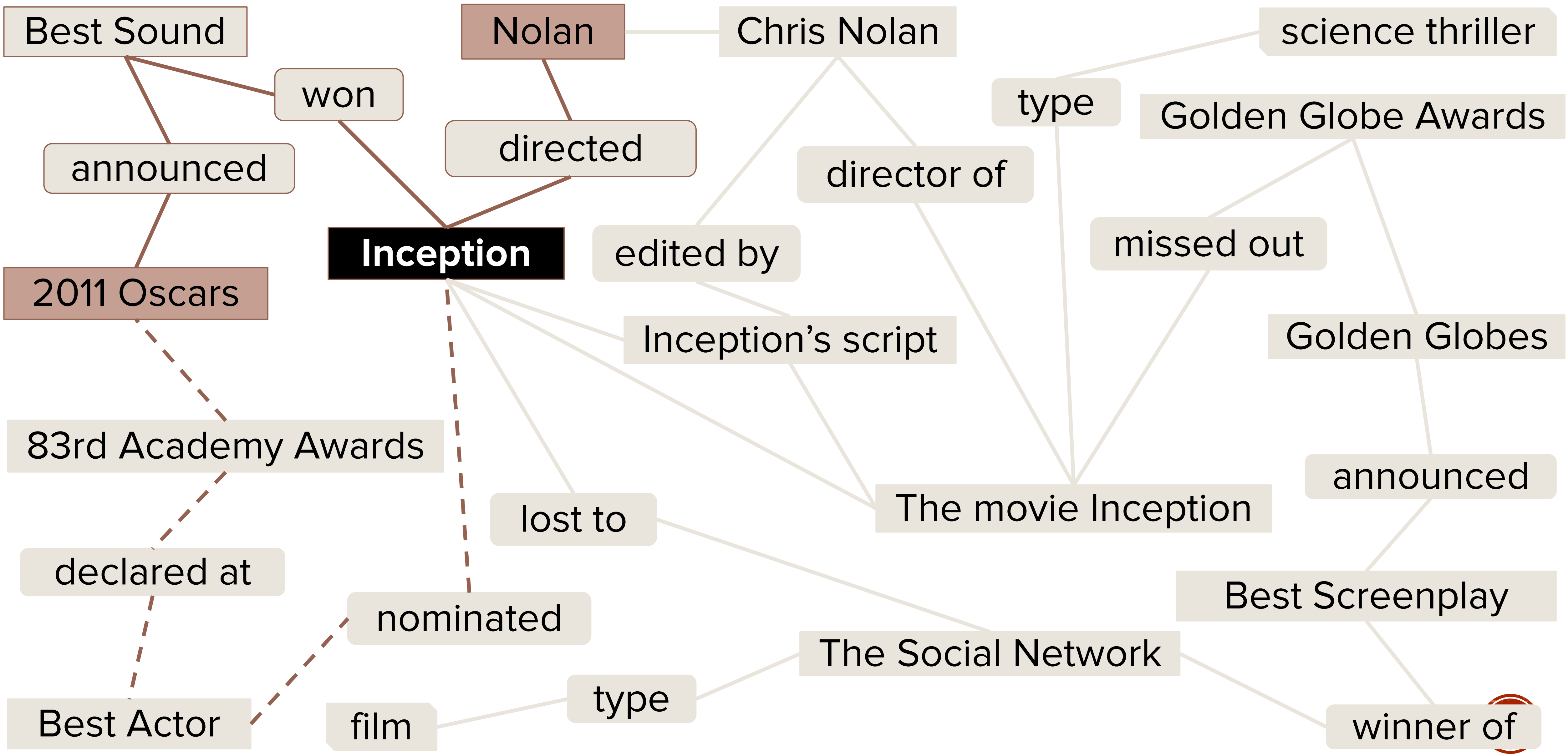
Question: Which Nolan films won an Oscar but missed a Golden Globe?



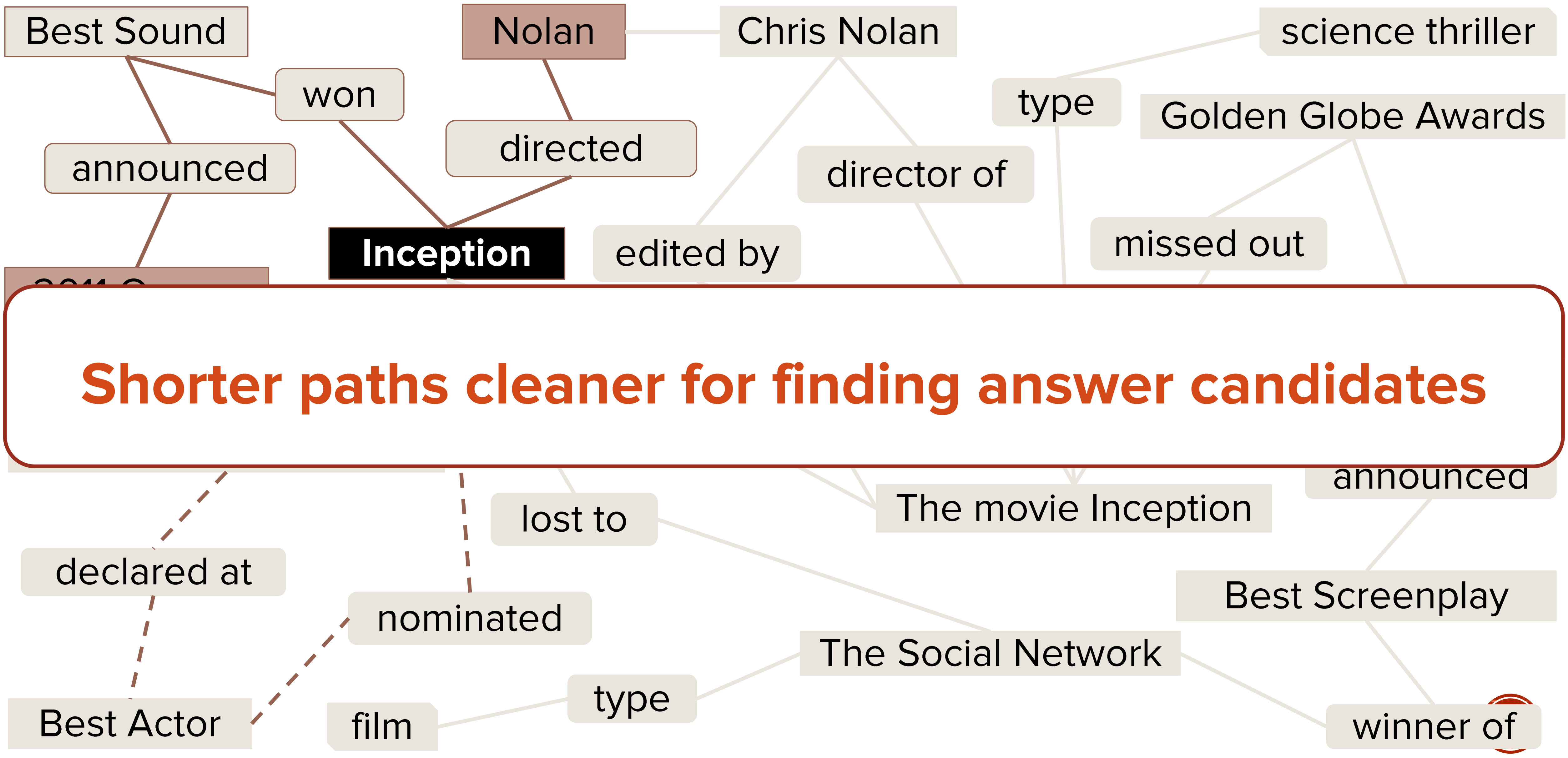
Answers lie on paths connecting cornerstones:

Internal nodes on paths are answer candidates

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

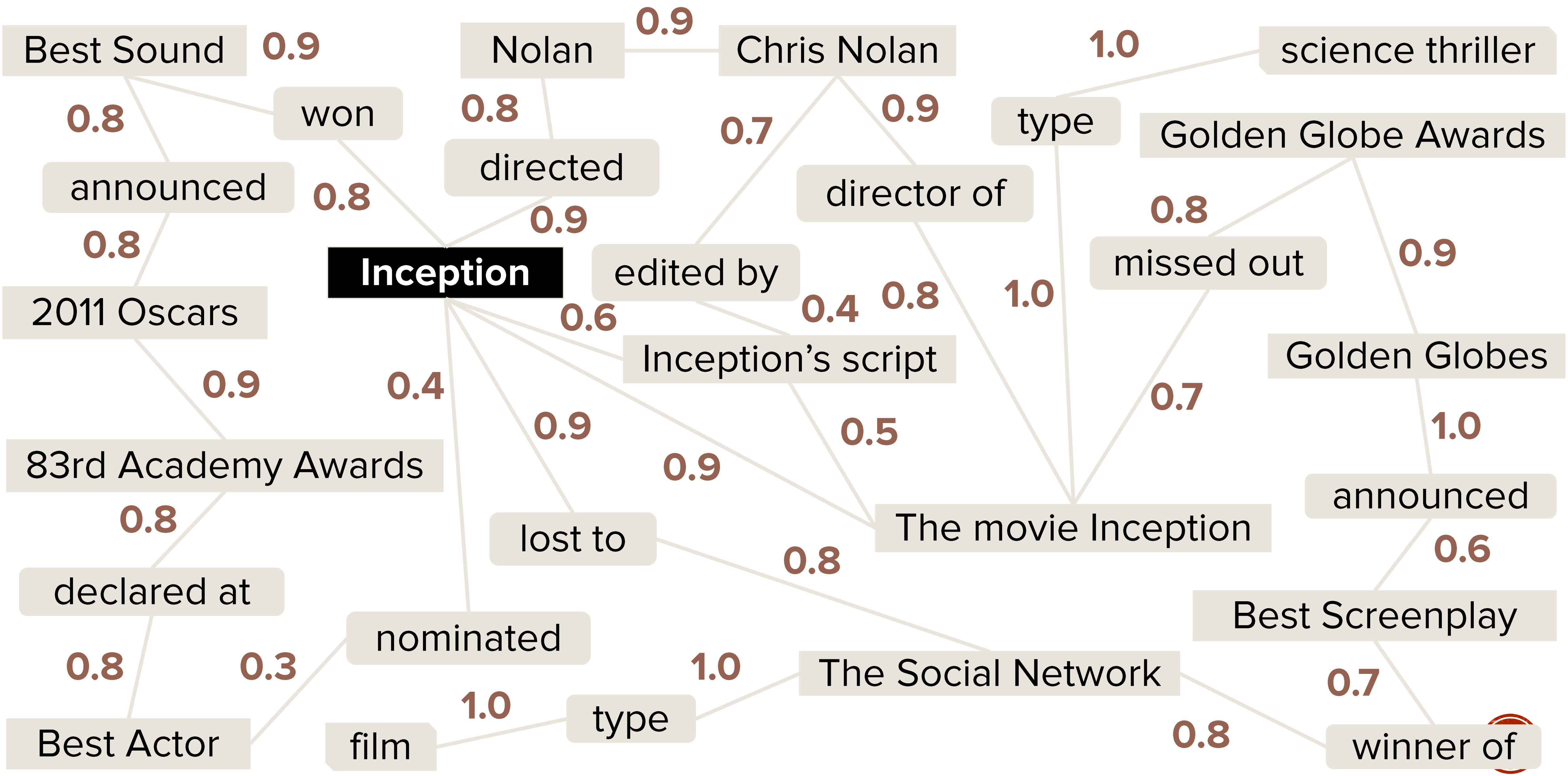


Question: Which Nolan films won an Oscar but missed a Golden Globe?

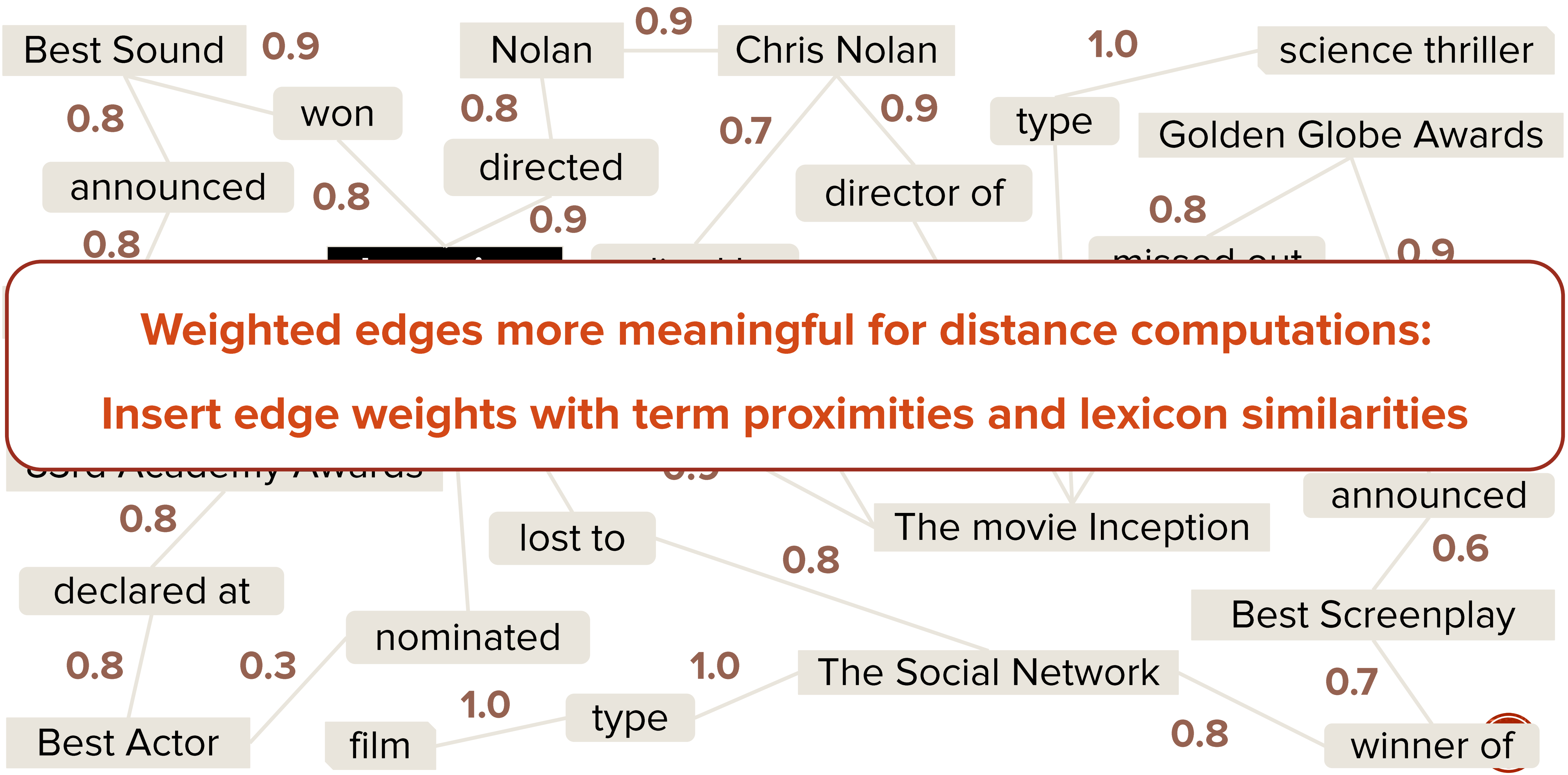


Shorter paths cleaner for finding answer candidates

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



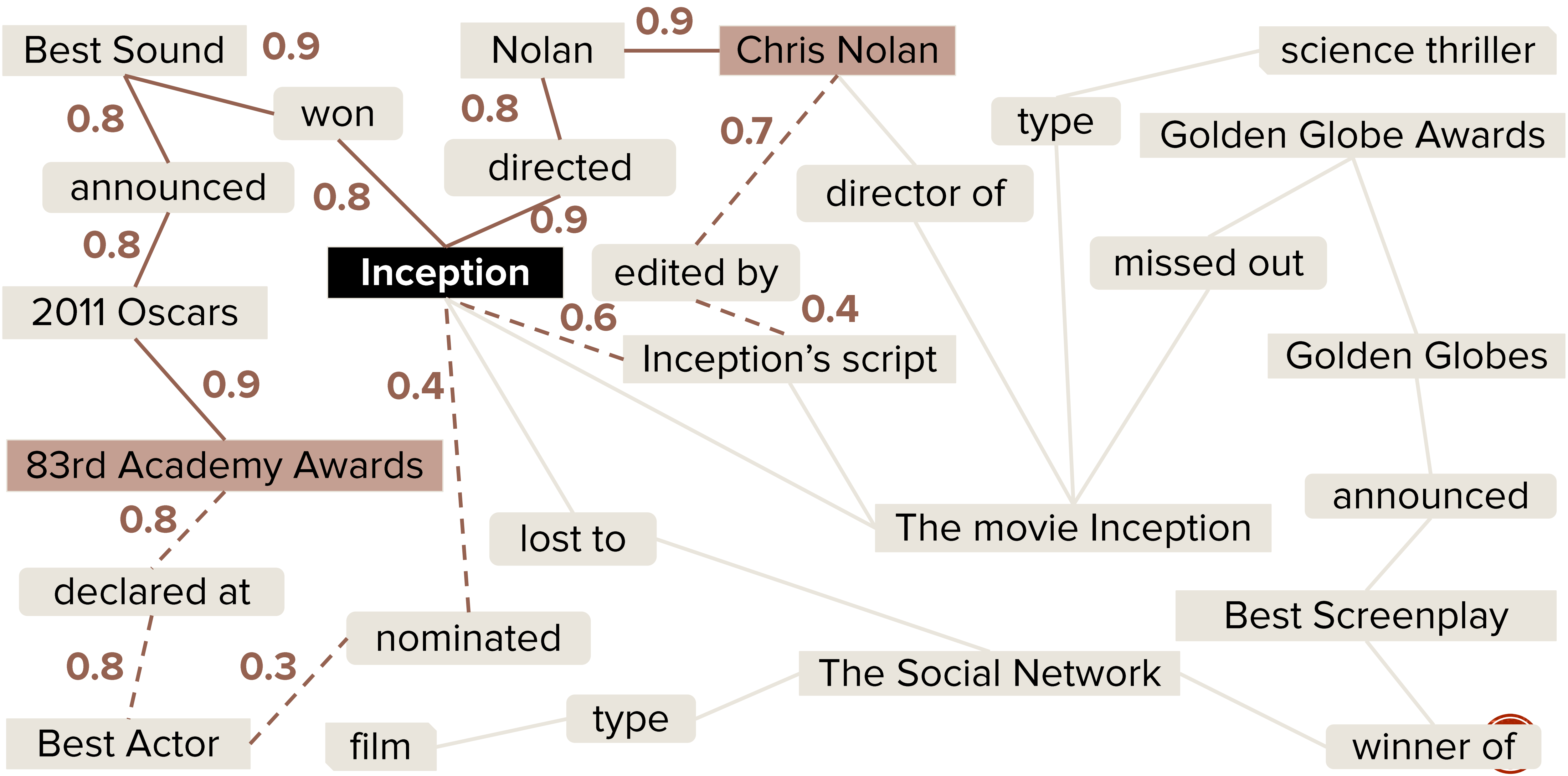
Question: Which Nolan films won an Oscar but missed a Golden Globe?



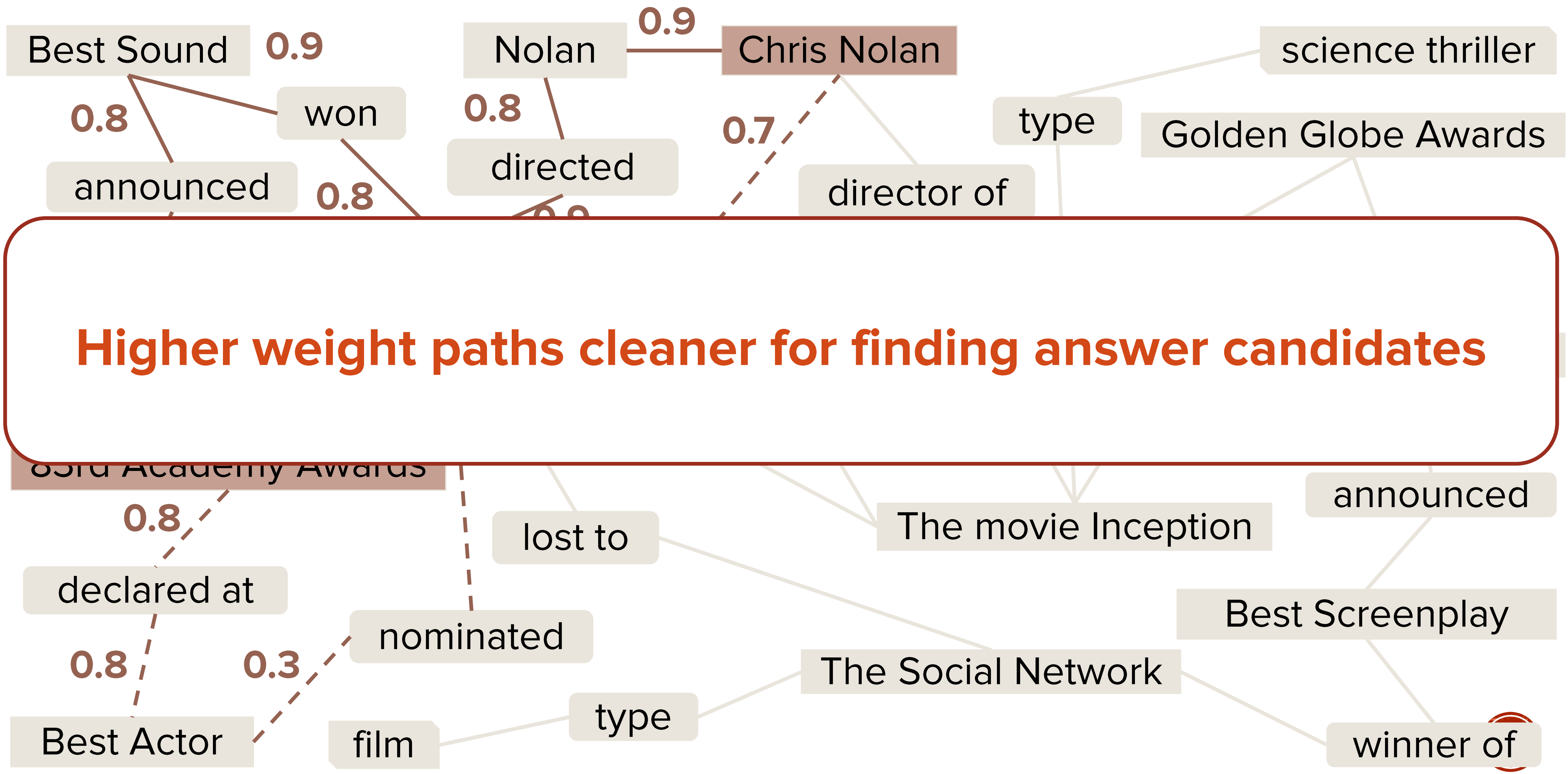
Weighted edges more meaningful for distance computations:

Insert edge weights with term proximities and lexicon similarities

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

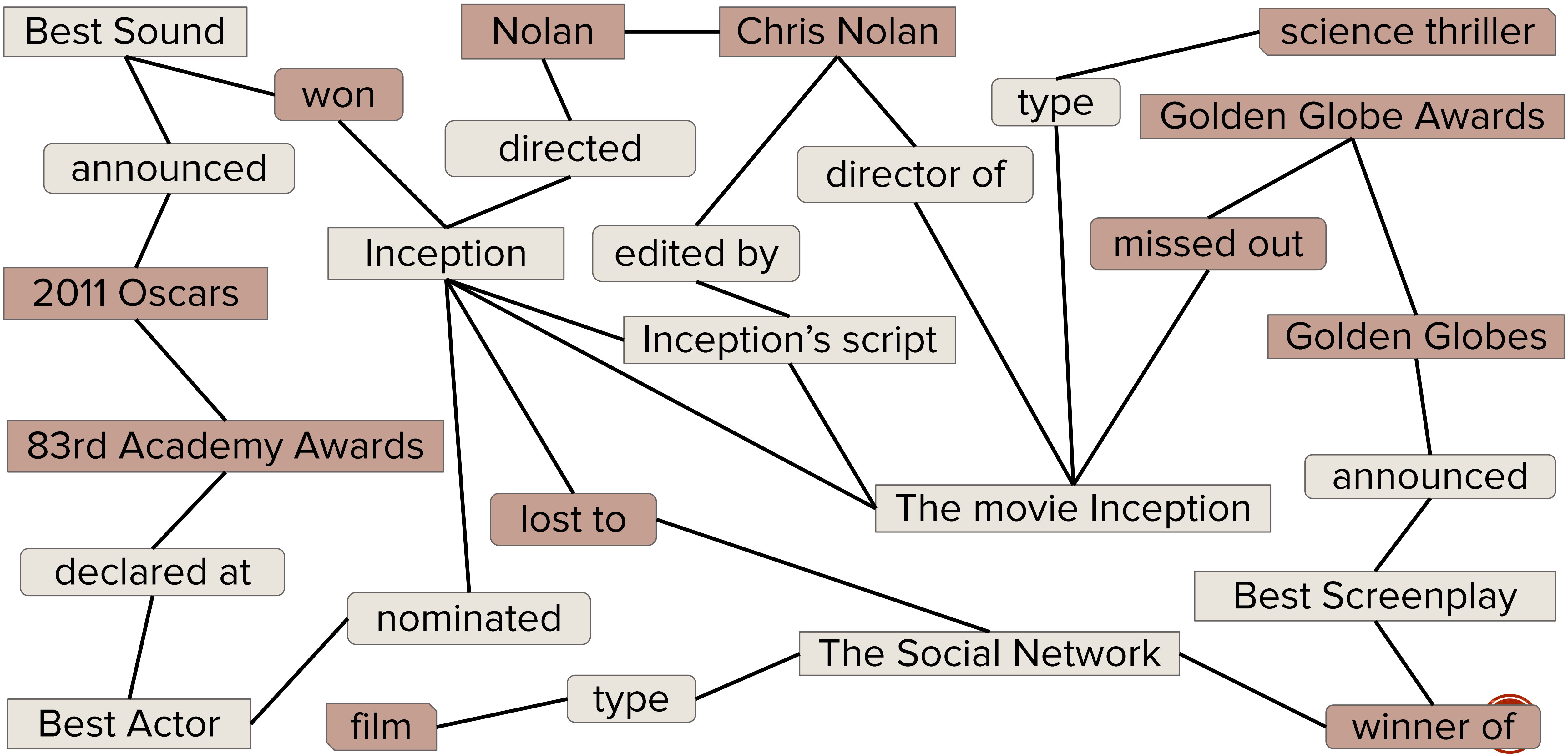


Question: Which Nolan films won an Oscar but missed a Golden Globe?

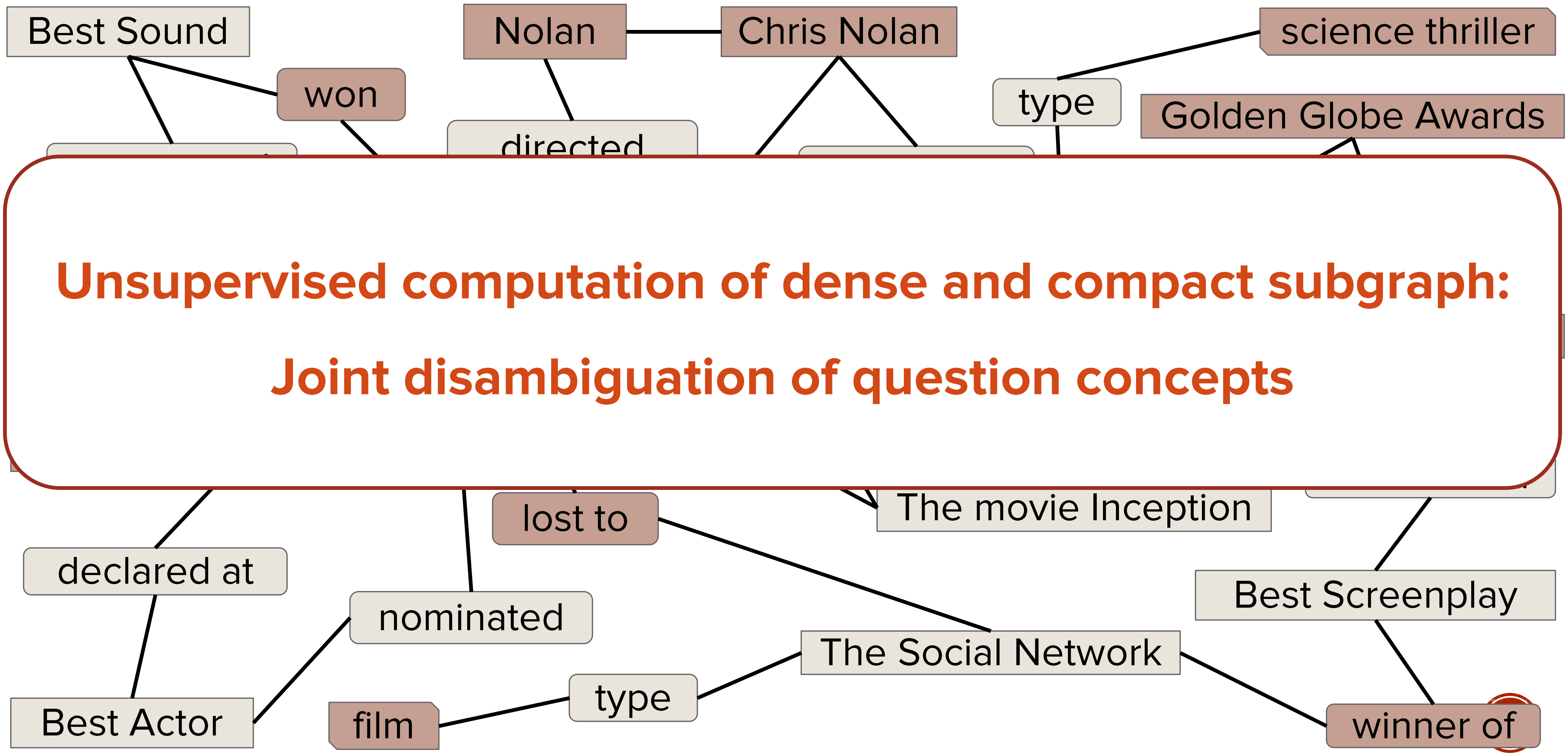


Higher weight paths cleaner for finding answer candidates

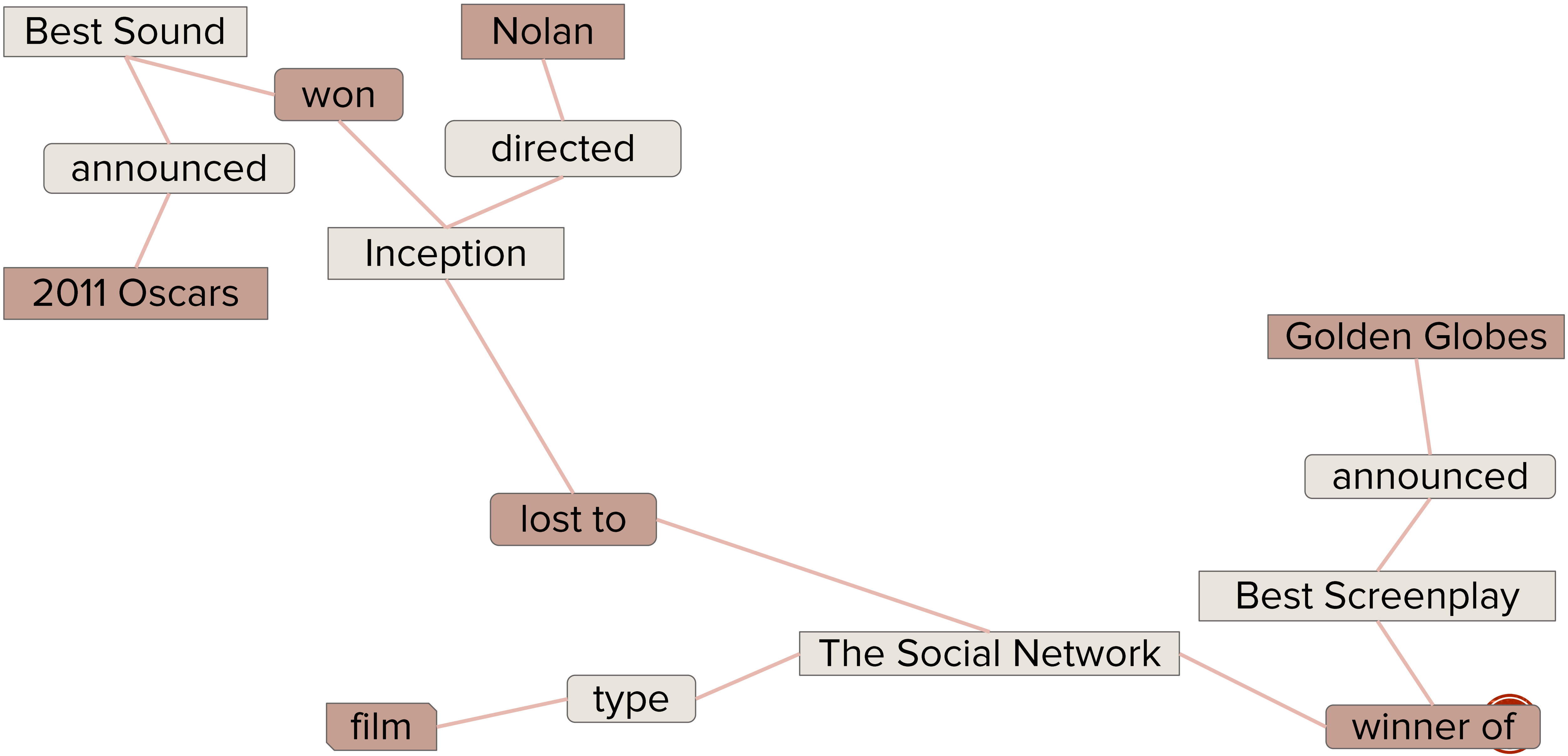
Question: Which Nolan films won an Oscar but missed a Golden Globe?



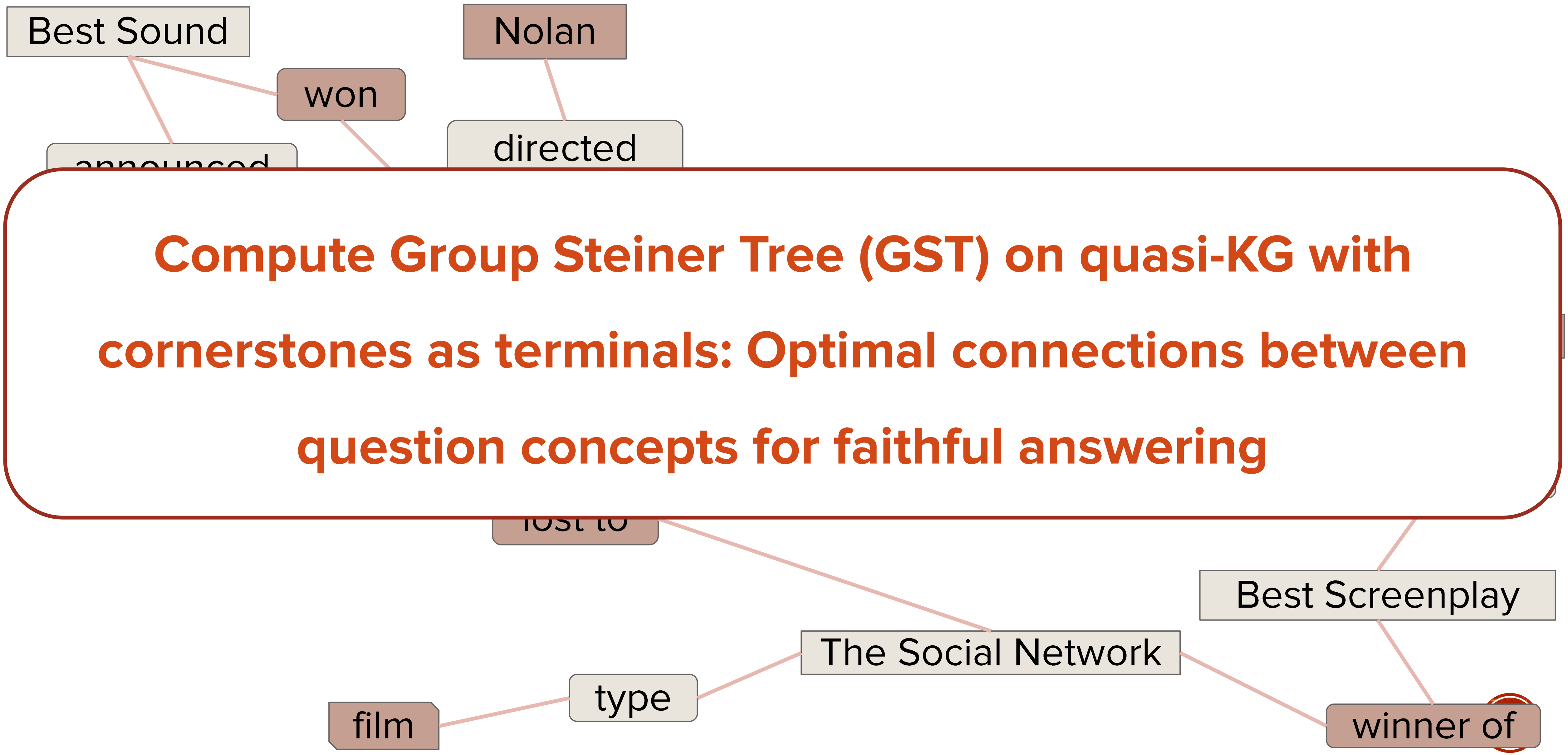
Question: Which Nolan films won an Oscar but missed a Golden Globe?



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



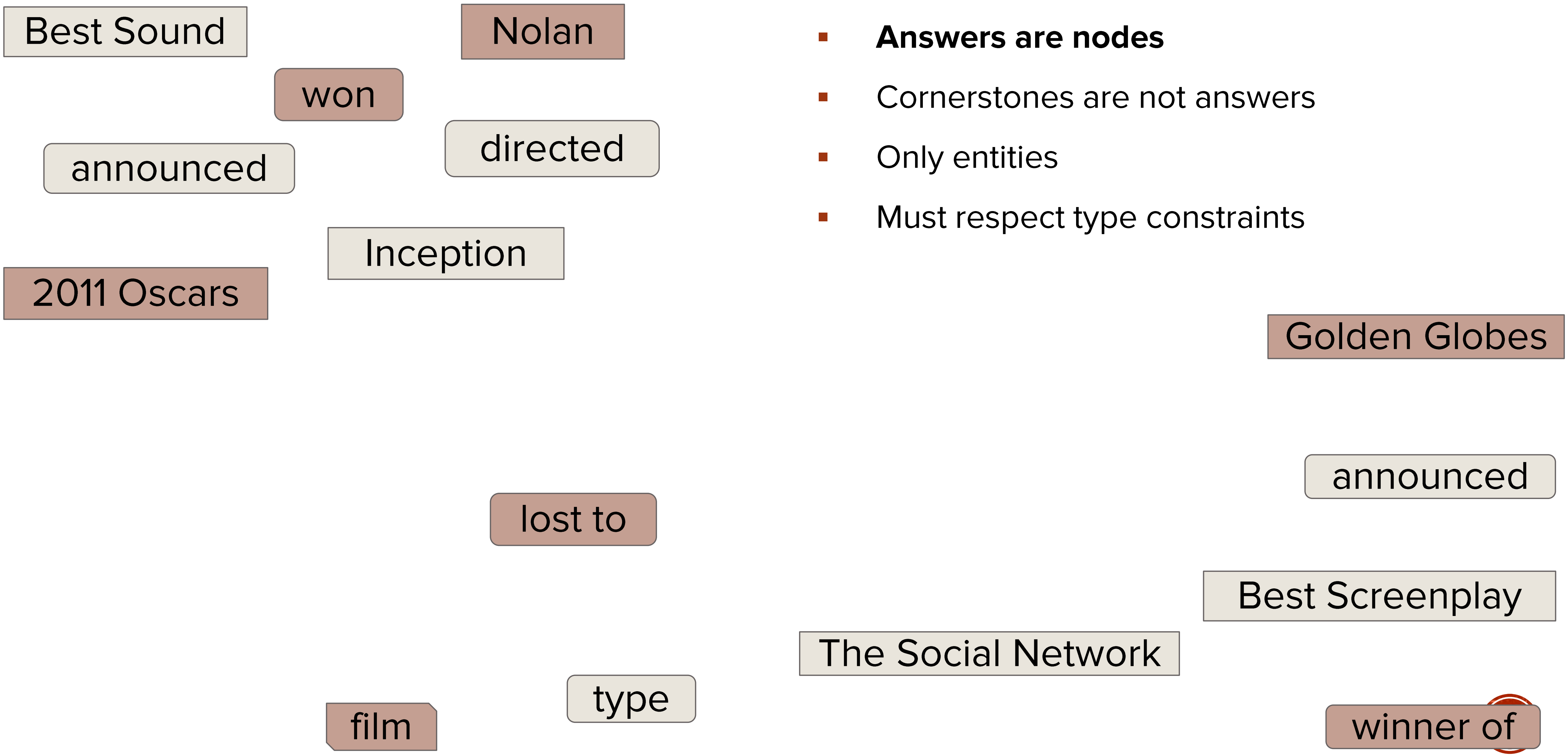
Question: Which Nolan films won an Oscar but missed a Golden Globe?



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?

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



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

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

Best Sound

won

directed

announced

Inception

lost to

type

announced

Best Screenplay

The Social Network

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?

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

Inception

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

Lu et al., Answering Complex Questions by Joining Multi-Document Evidence with Quasi Knowledge Graphs, SIGIR 2019.

Complex QA: Graph-based belief propagation

- The **QAmp** system (Vakulenko et al. 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?

Vakulenko et al., Message Passing for Complex Question Answering over Knowledge Graphs, CIKM 2019.

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

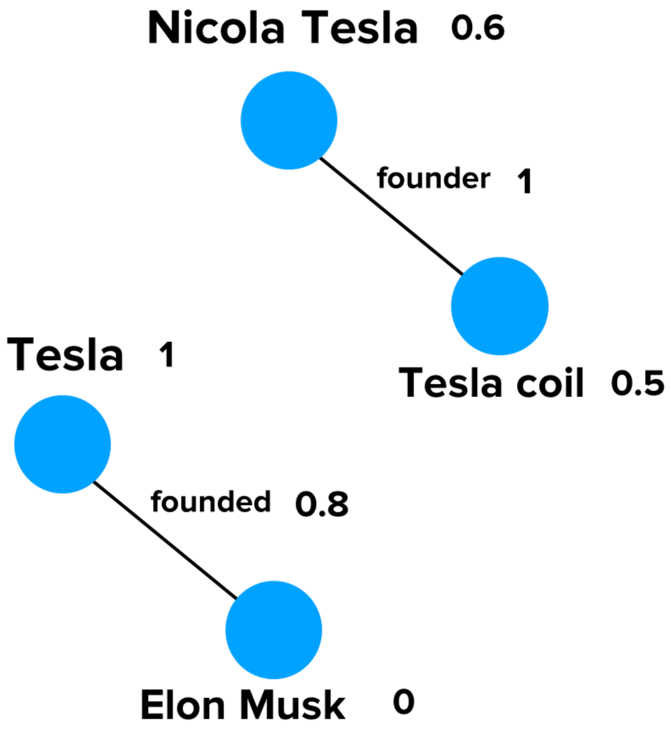
E1	
Tesla	1
Nicola Tesla	0.6
Tesla coil	0.5

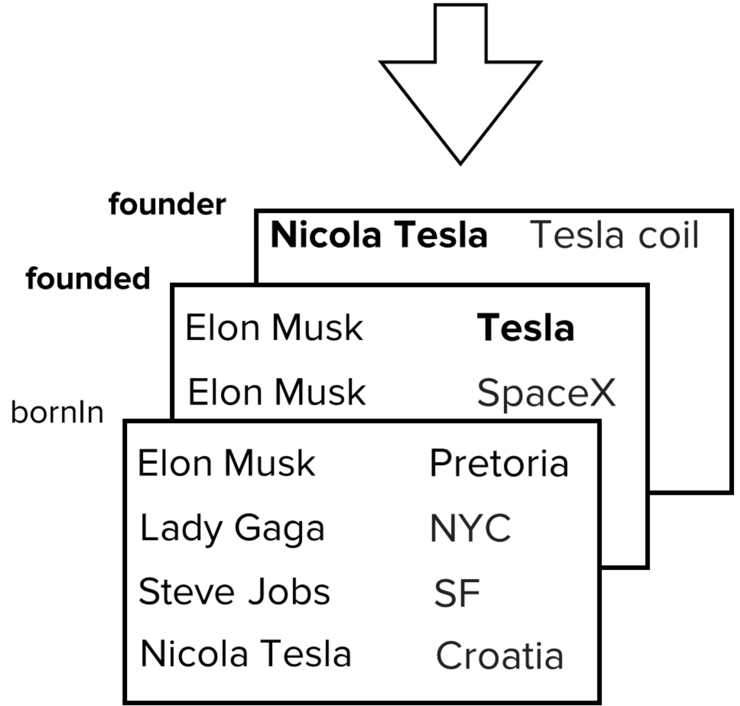
Reasoning: Message passing Hop 1

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



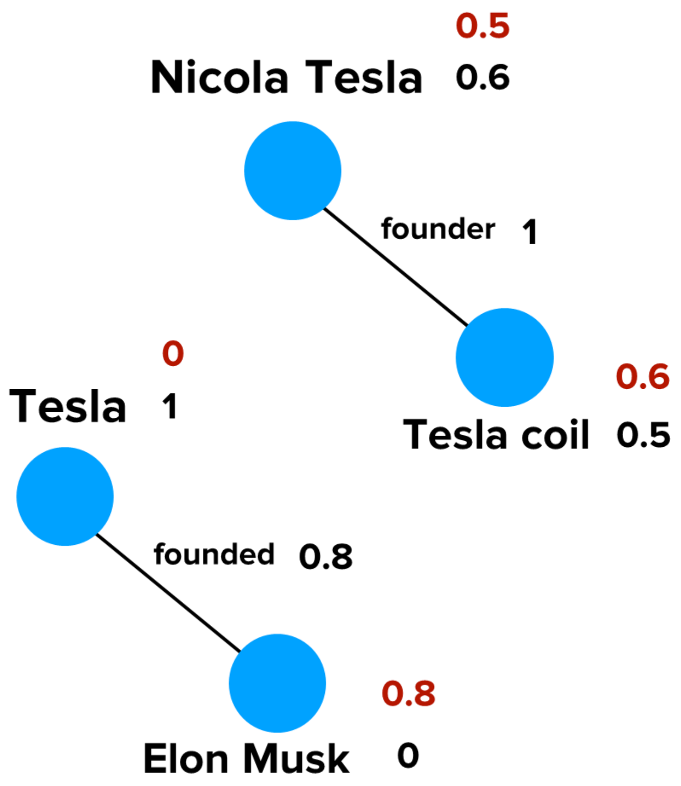


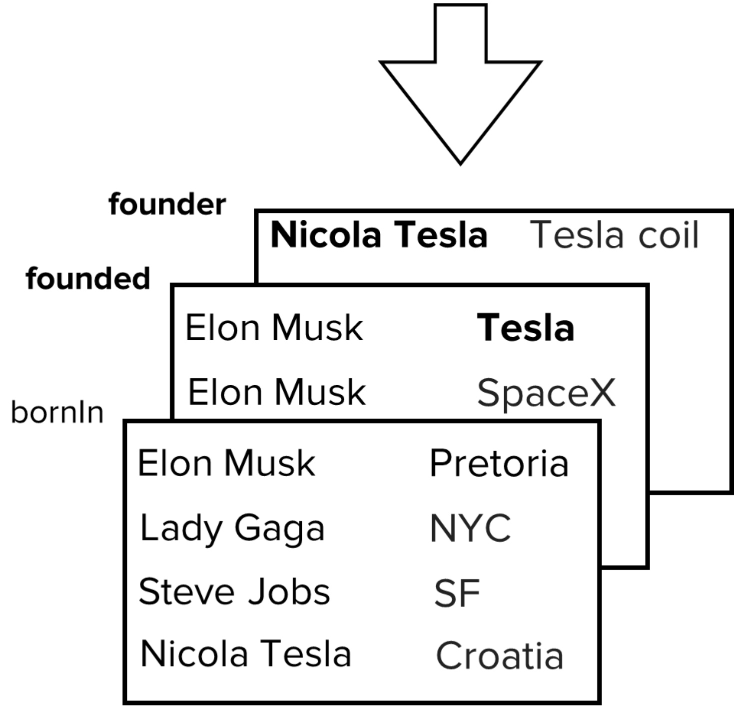
Reasoning: Message passing Hop 1

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

P1 E1 P2

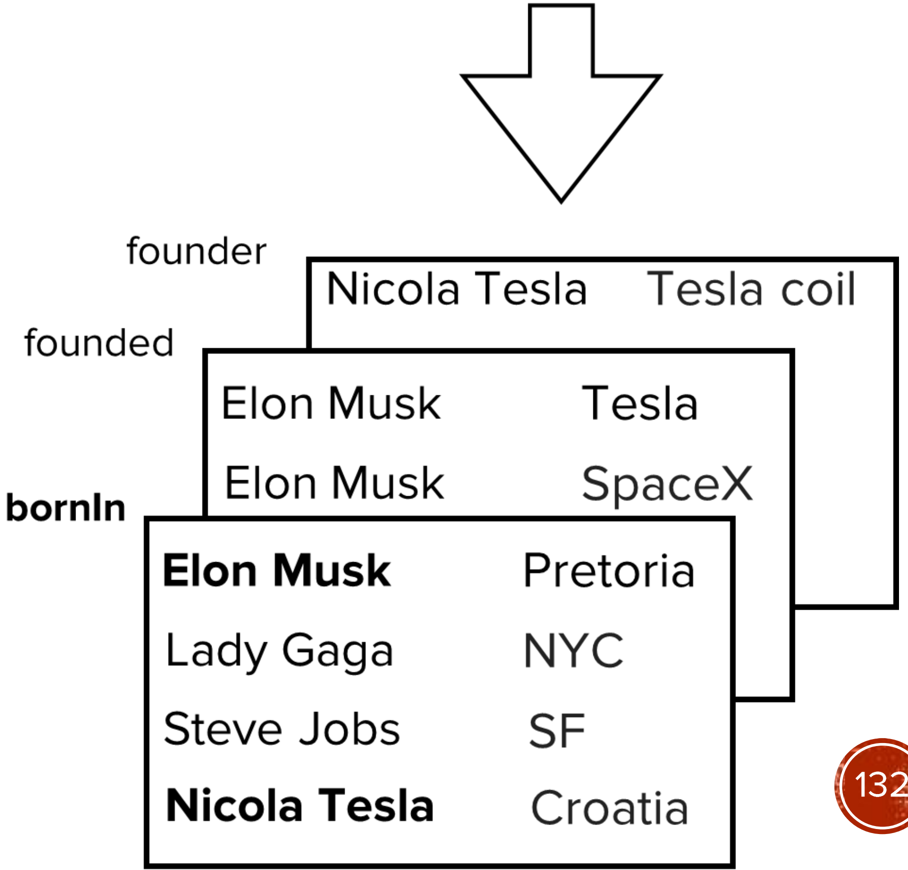
P1		E1	
founder	1	Tesla	1
founded	0.8	Nicola Tesla	0.6
		Tesla coil	0.5





Reasoning: Message passing Hop 2

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



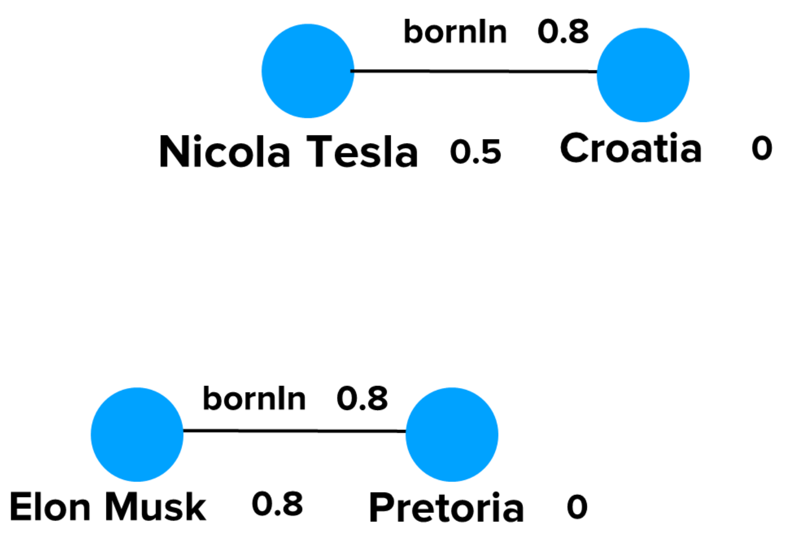
Reasoning: Message passing Hop 2

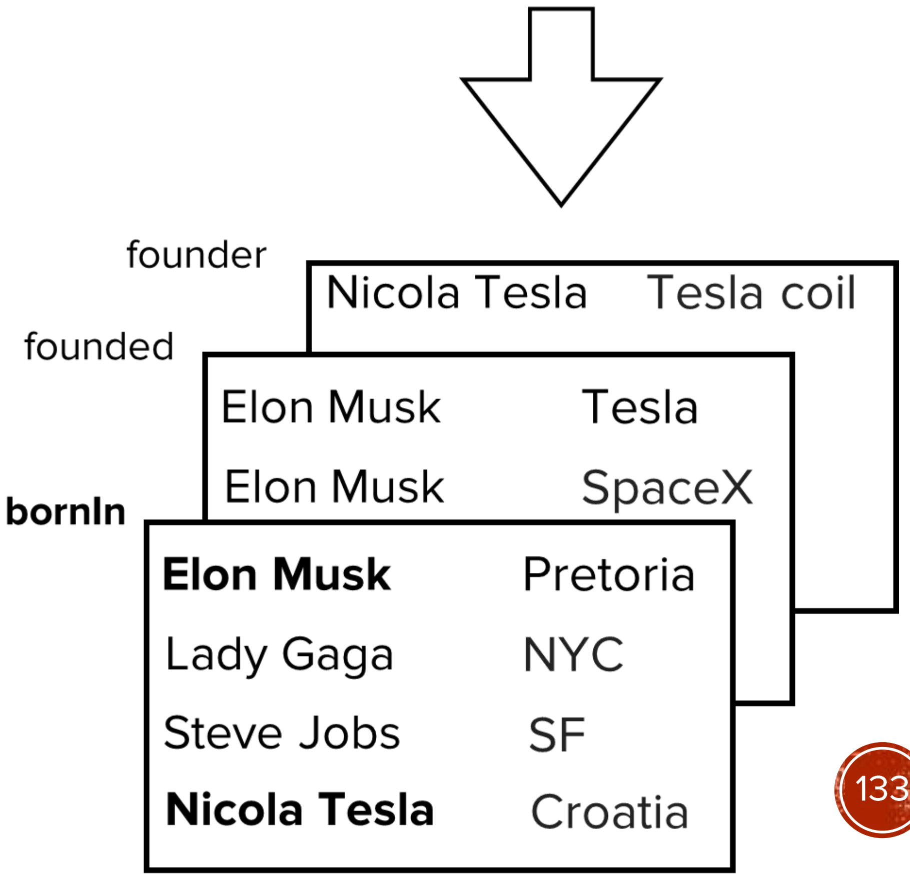
P2

bornIn	0.8

E2

Elon Musk	0.8
Tesla coil	0.6
Nicola Tesla	0.5



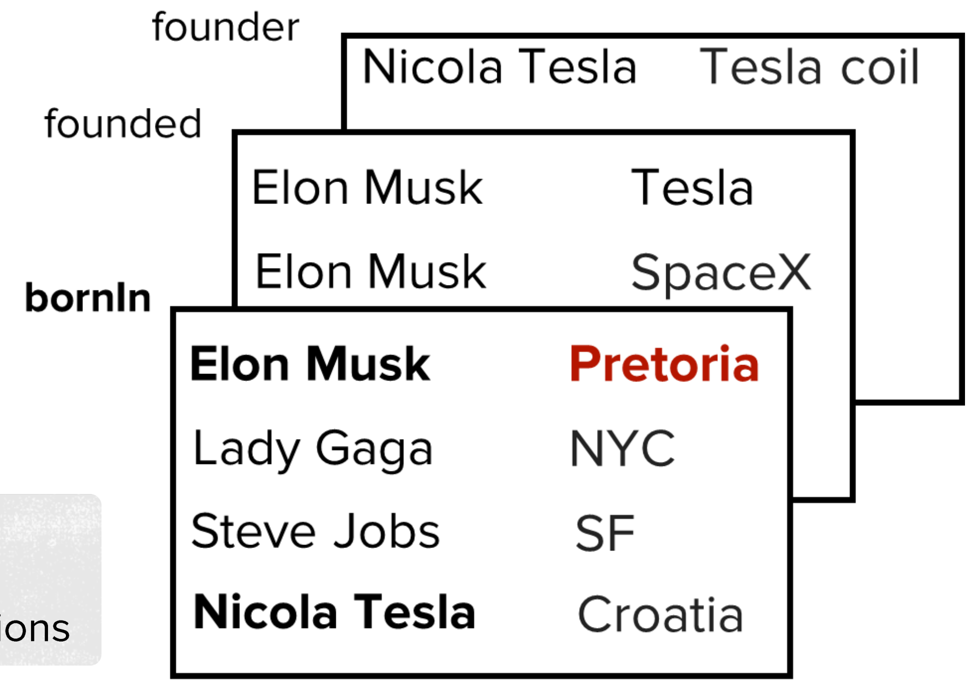


Reasoning: Message passing Hop 2

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

P2	
bornIn	0.8

E2	
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, a lot of 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

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), AQQUCN (Sawant et al. 2019), **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

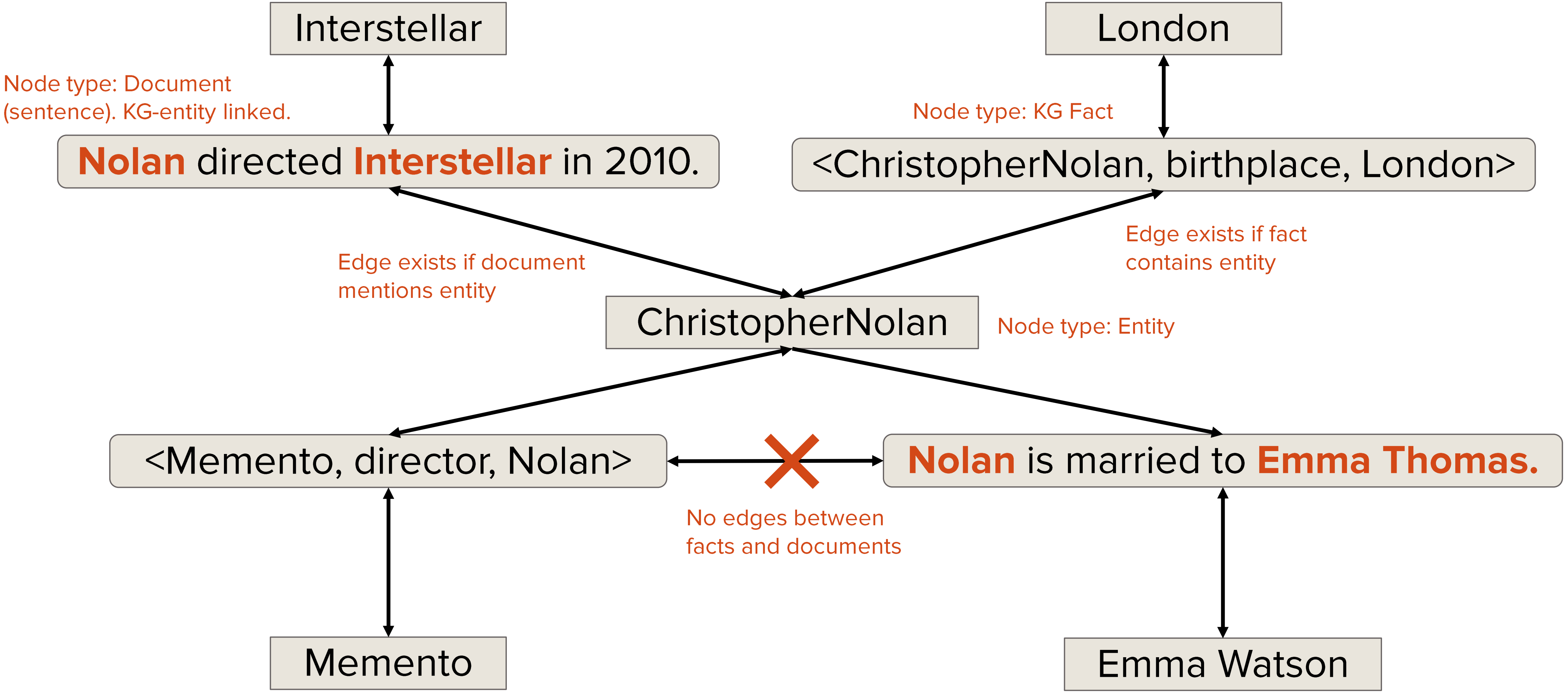
Sun et al., PullNet: Open Domain Question Answering with Iterative Retrieval on Knowledge Bases and Text, EMNLP 2019.

PullNet: Handling heterogeneity

KG facts	Entity-linked sentences
<ChristopherNolan, birthplace, London>	**Nolan** is married to **Emma Thomas.**
<Memento, director, Nolan>	**Nolan** directed **Interstellar** in 2010.
<Interstellar, castMember, AnneHathaway>	**Guy Pearce** was in **Memento** and **Flynn.**

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

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.

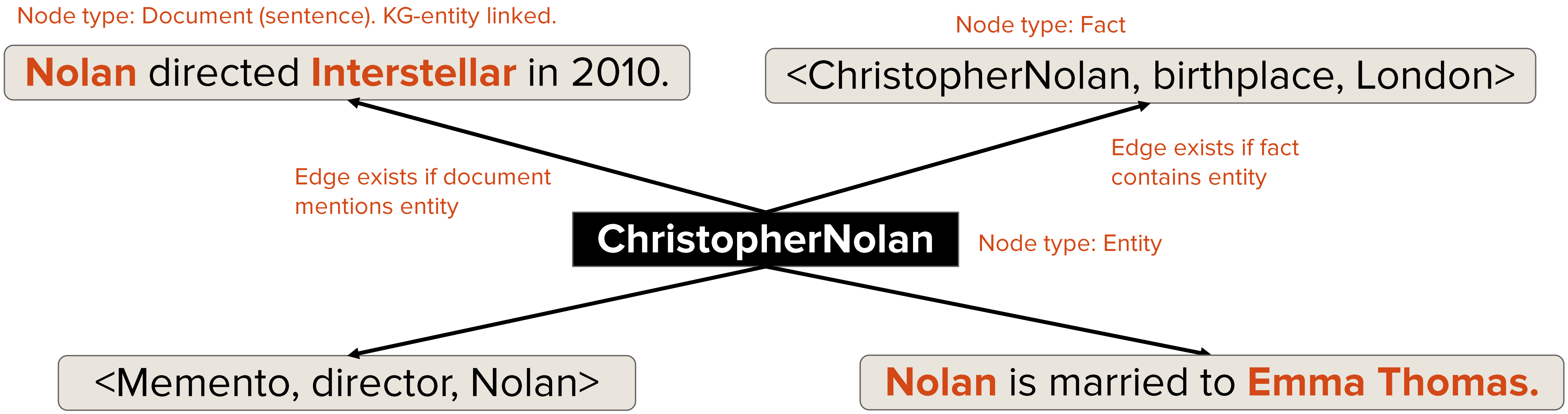
Question: Who are the actors in movies directed by Nolan?

NERD system

ChristopherNolan

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

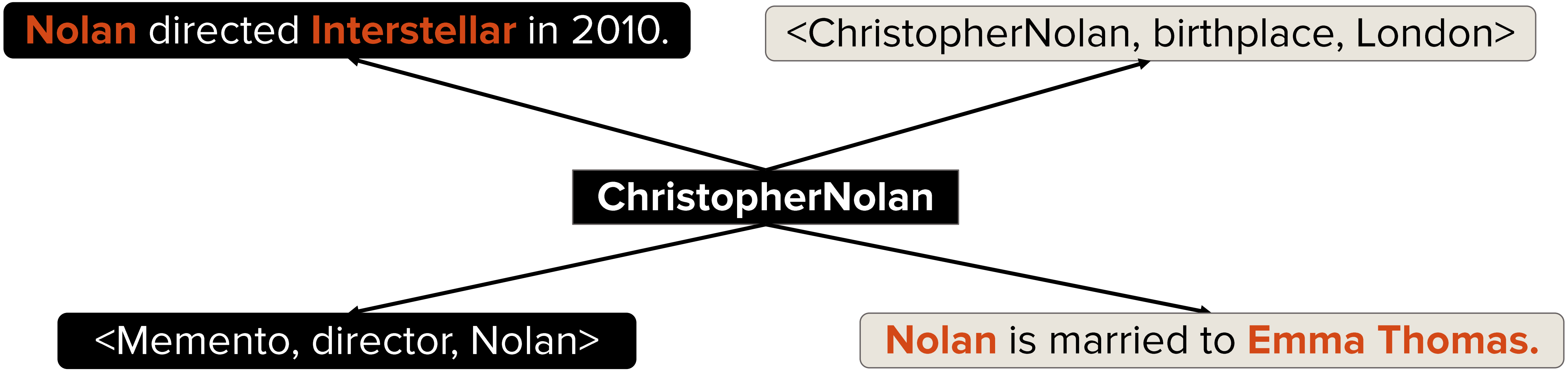
Question: Who are the actors in movies directed by Nolan?



Early fusion

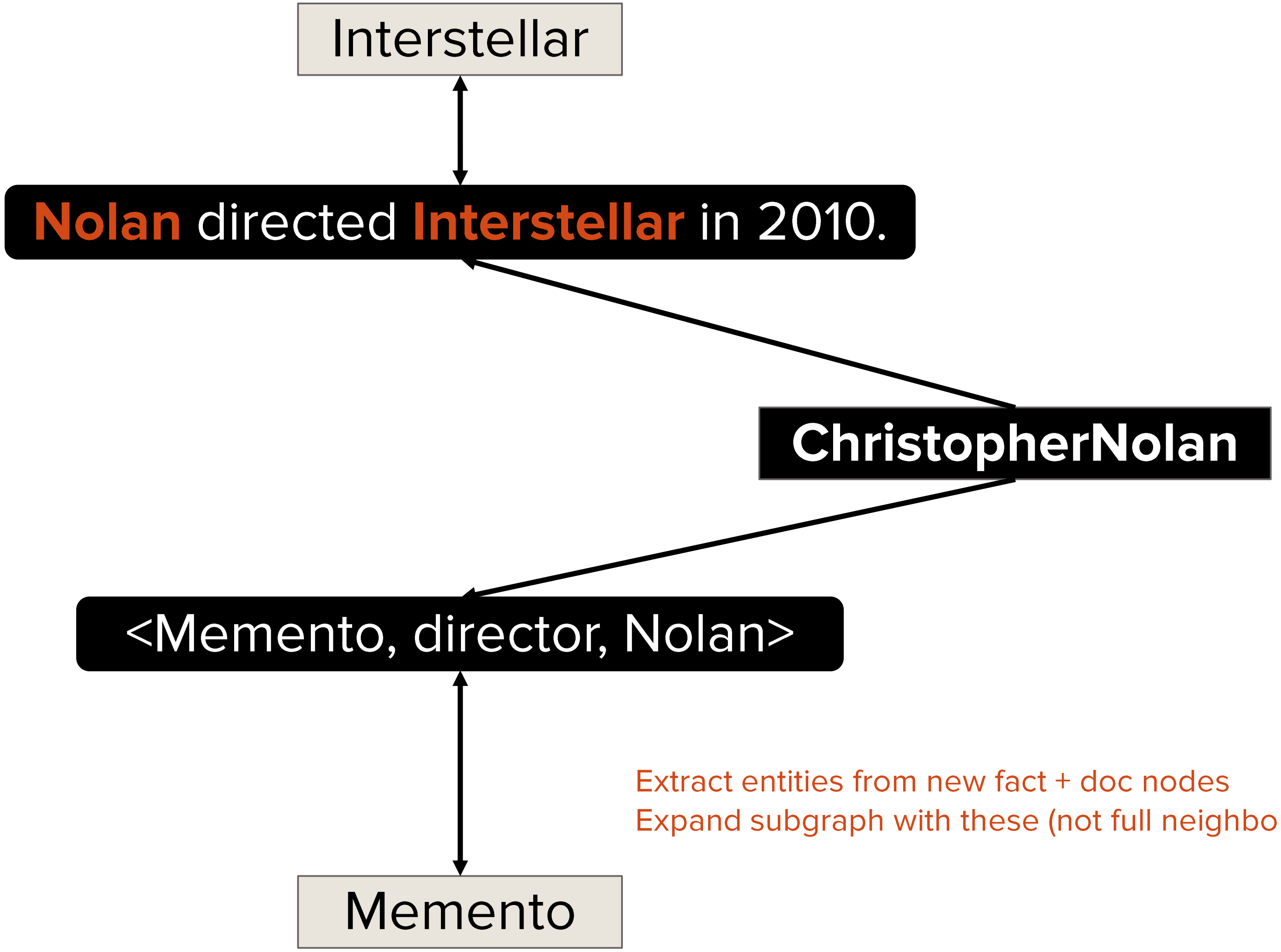
1. Pull sentences with linked entity from corpus (Lucene)
2. Pull facts of entity from the KG (using predicate similarity learned via LSTMs)

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



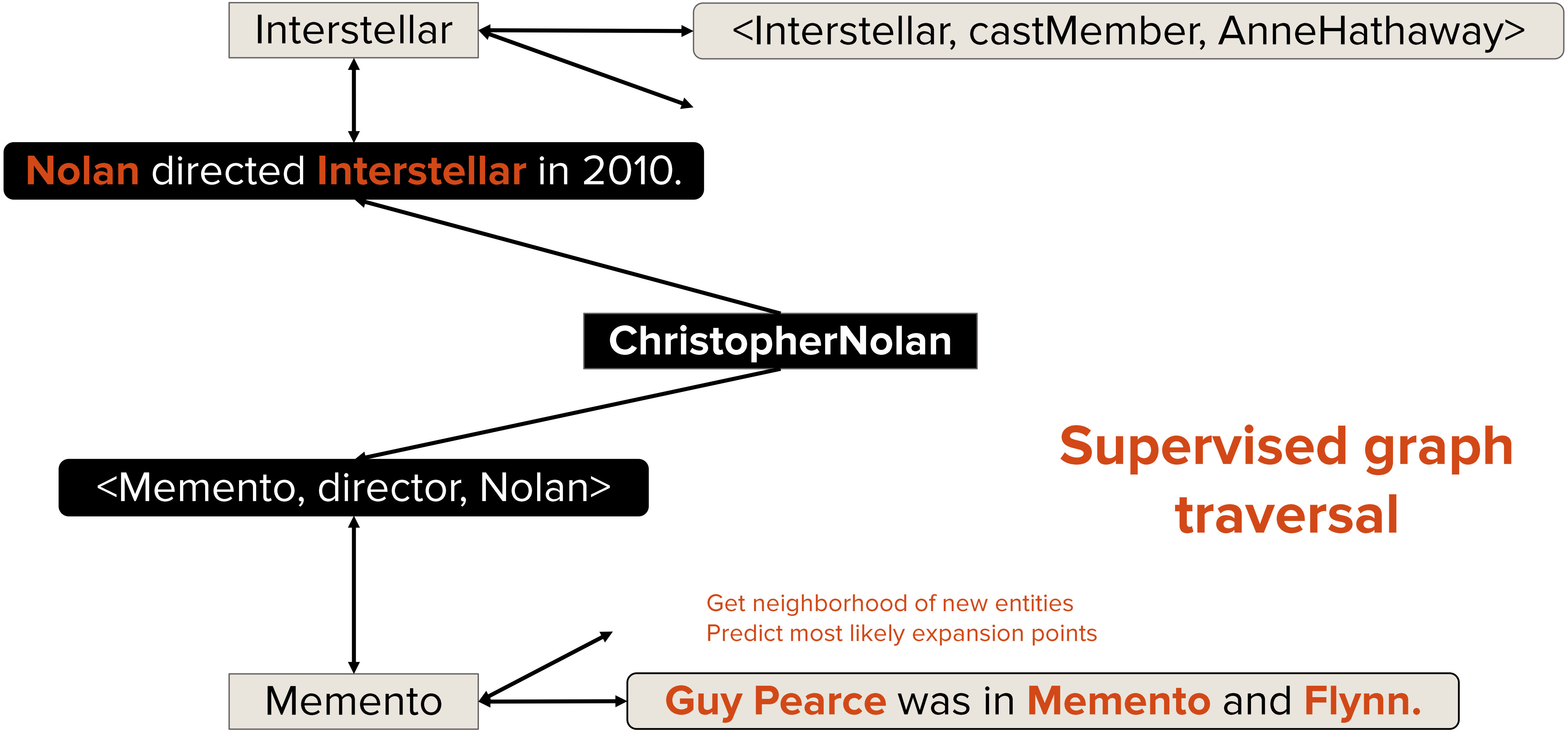
Predict most likely expansion points for next hop

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



Extract entities from new fact + doc nodes
Expand subgraph with these (not full neighborhood)

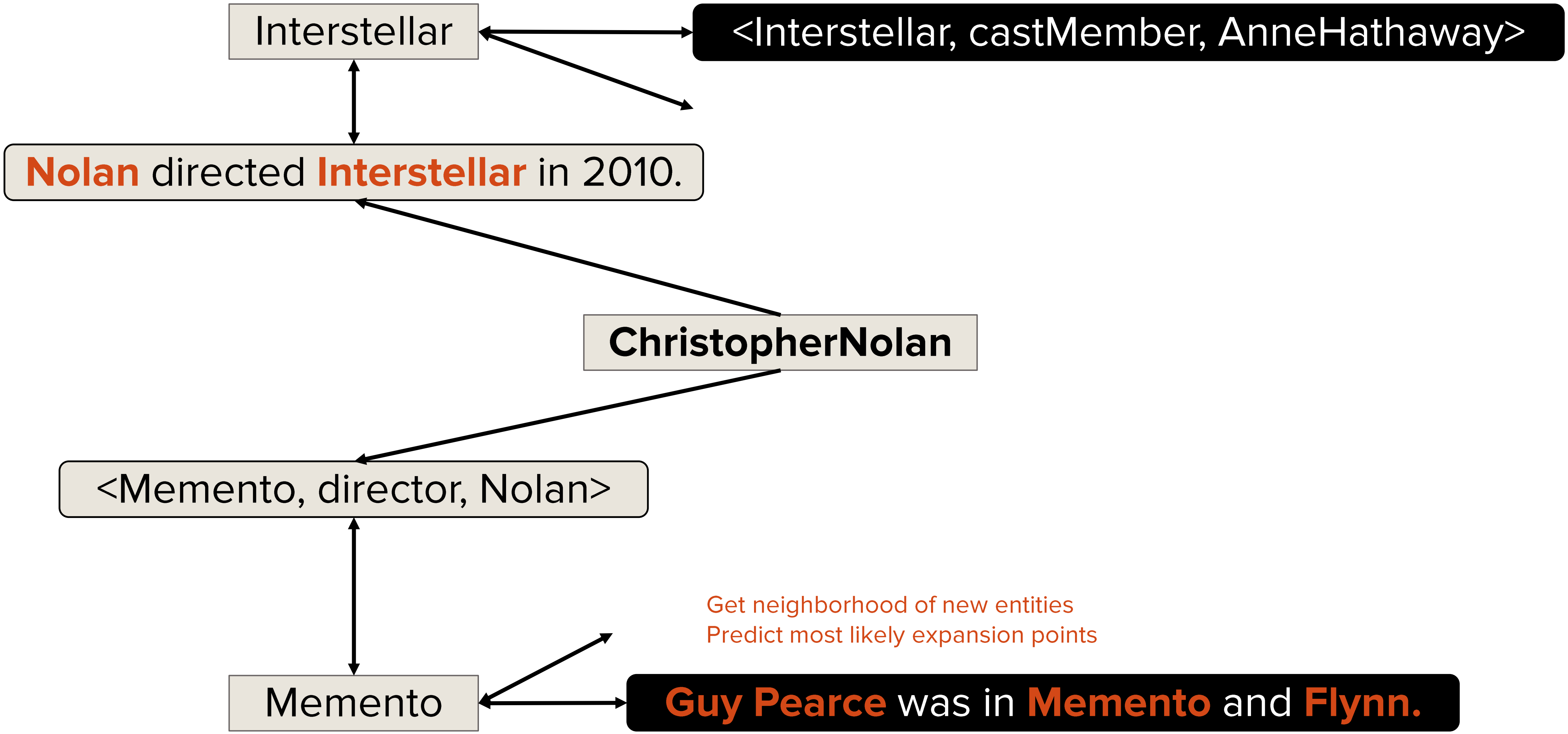
Question: Who are the actors in movies directed by Nolan?



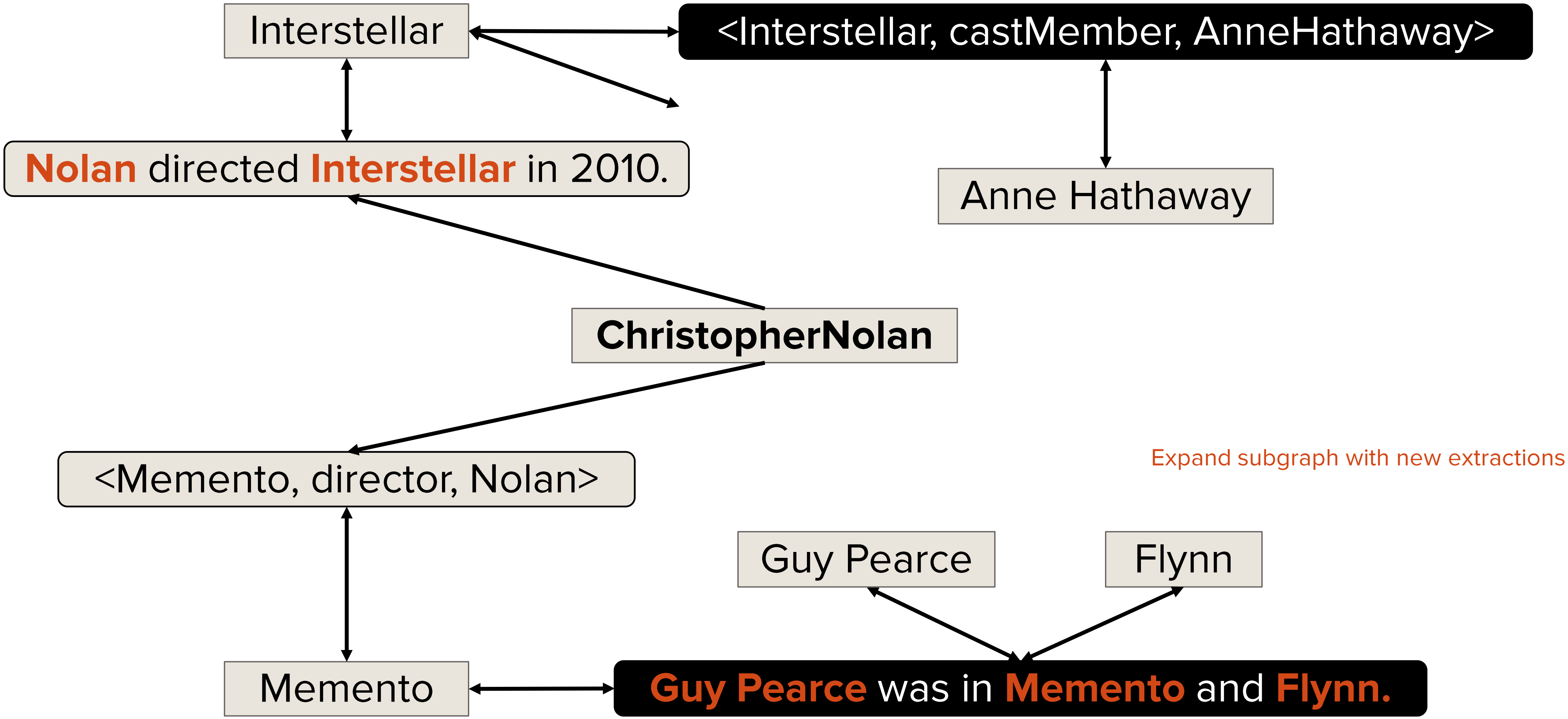
Supervised graph traversal

Get neighborhood of new entities
Predict most likely expansion points

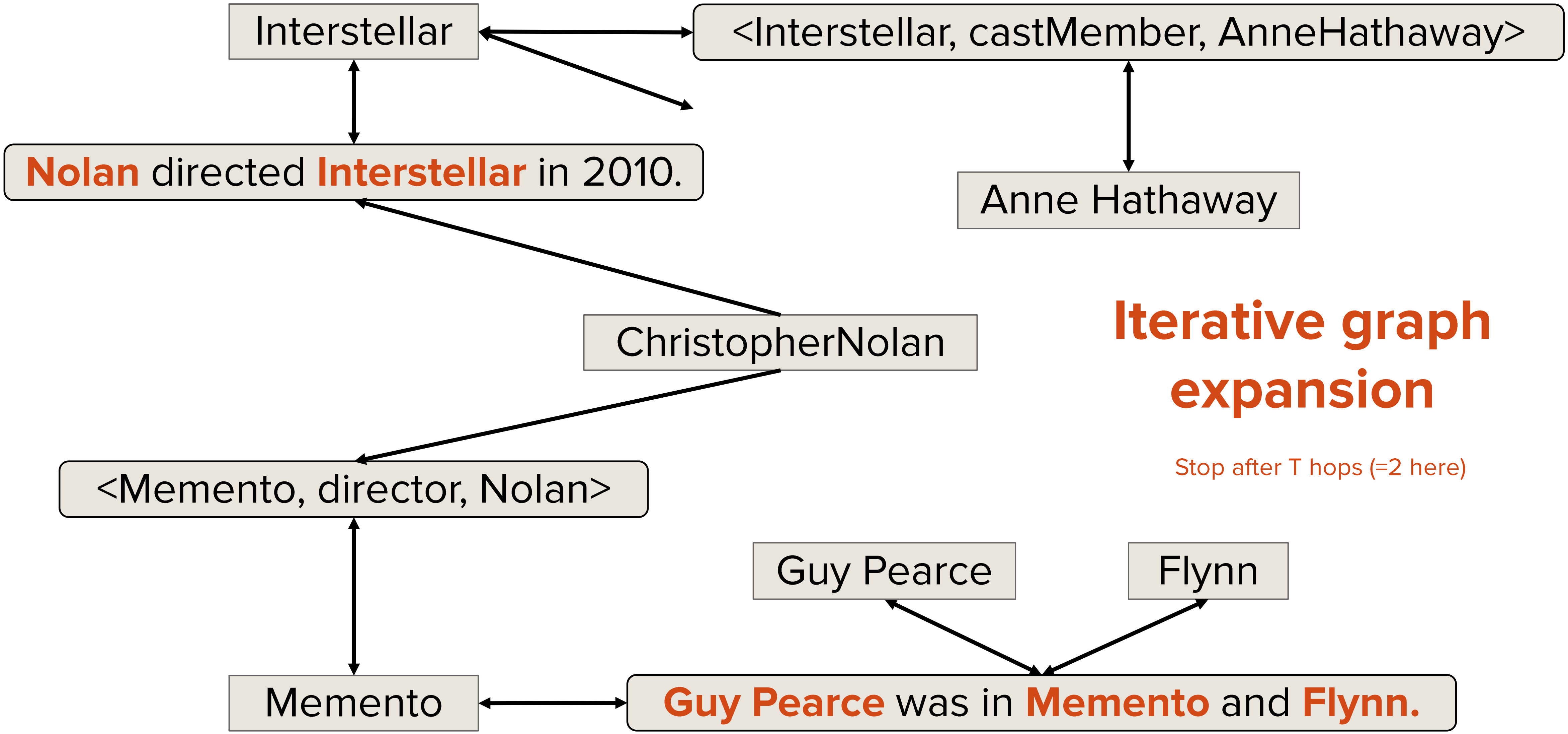
Question: Who are the actors in movies directed by Nolan?



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



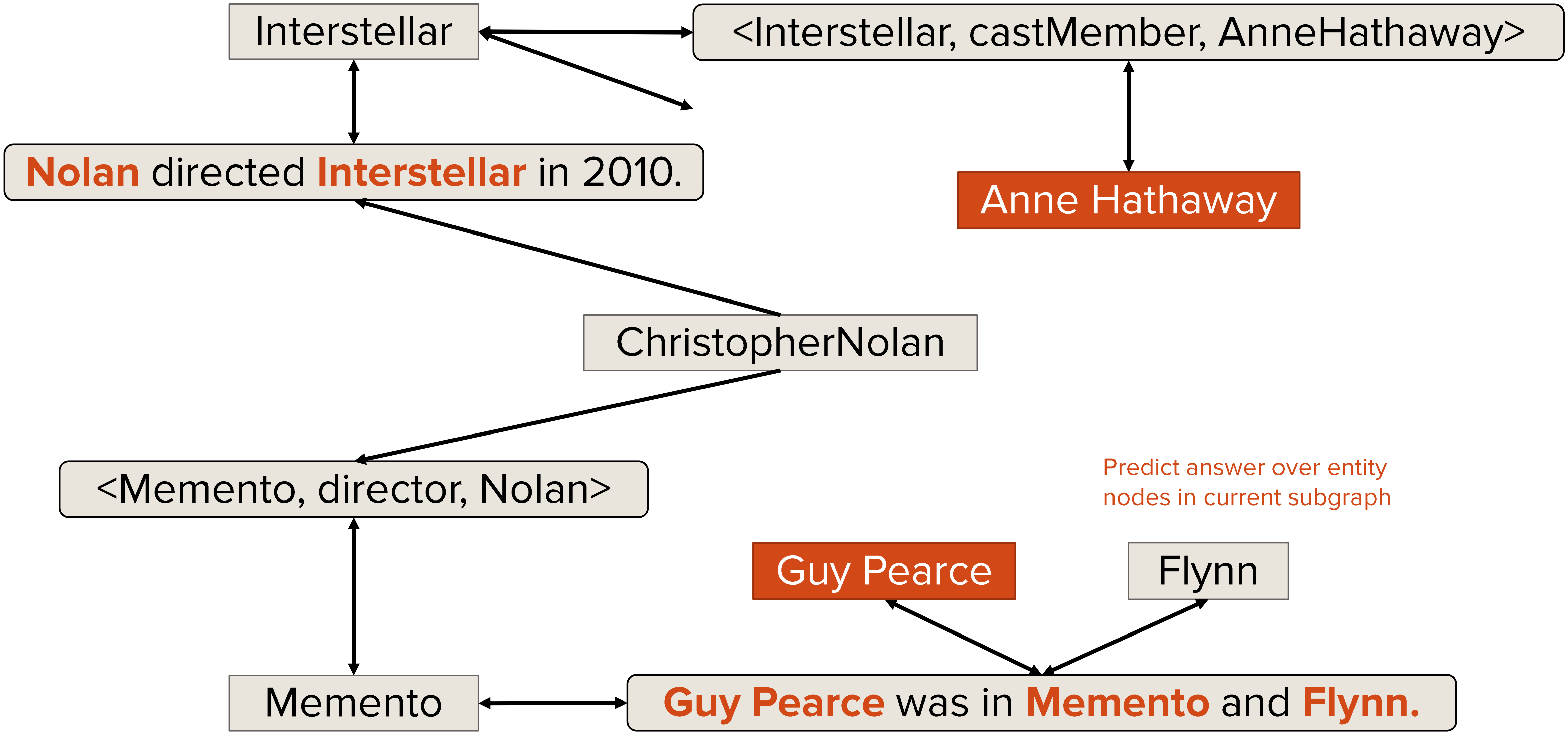
Question: Who are the actors in movies directed by Nolan?



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?

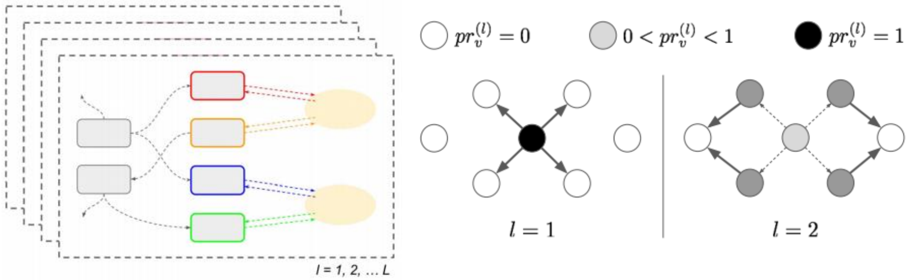
Interstellar

<Interstellar, castMember, AnneHathaway>

Nolan directed **Interstellar** in 2010.

Anne Hathaway

ChristopherNolan



Graph CNN-based classifier (GRAFT-Net) for both expansion point and answer prediction (Sun et al. 2018)

<Memento, director, Nolan>

Guy Pearce

Flynn

Memento

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

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

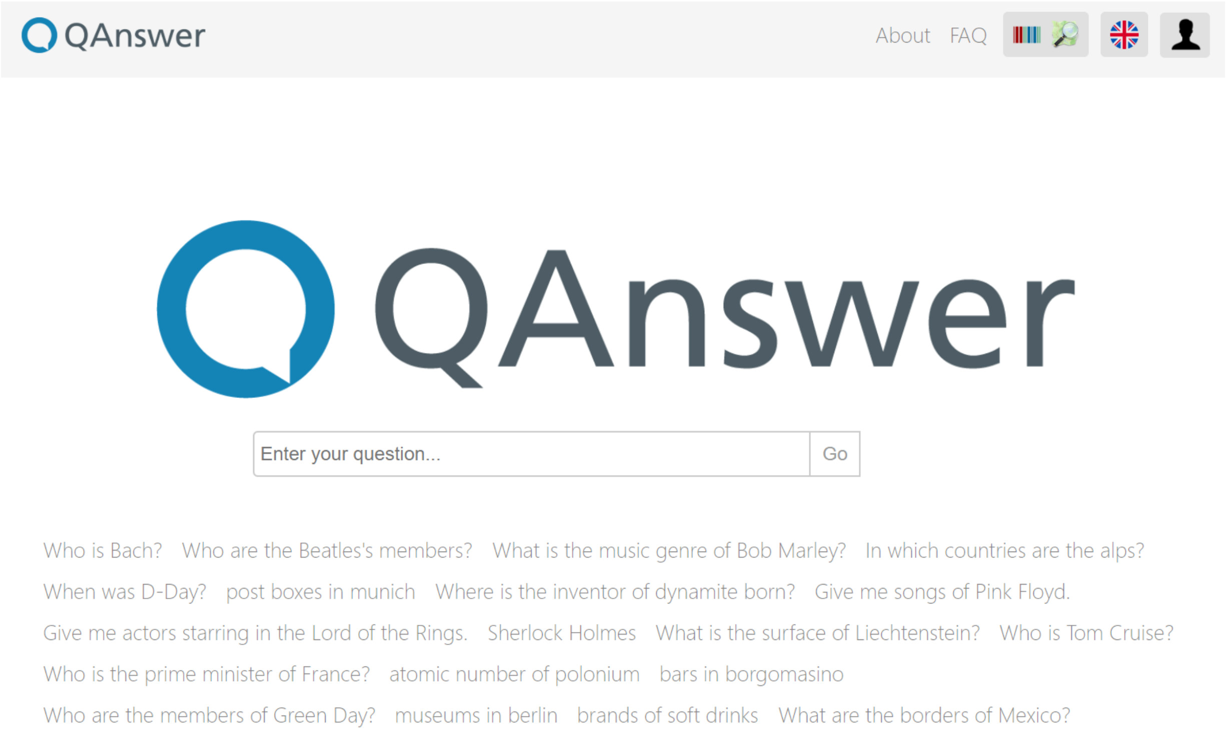
Sun et al., PullNet: Open Domain Question Answering with Iterative Retrieval on Knowledge Bases and Text, EMNLP 2019.

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)
- Multiple KGs as unified triple store
- KG-agnostic approach for QA

Give me actors born in Berlin.

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)
- 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}

Lucene-based lookup

Question

Expansion

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

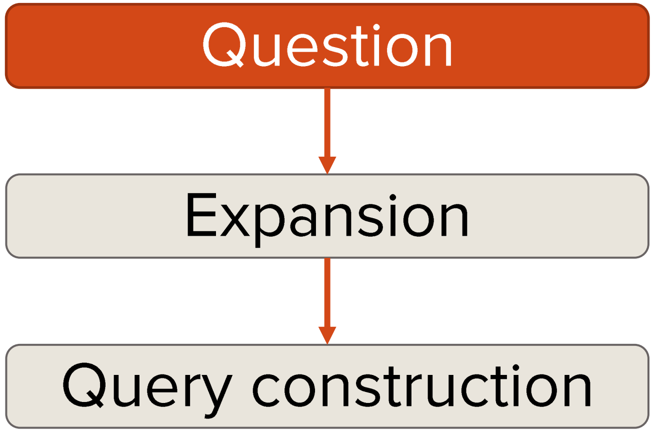
Efficient construction of SPARQL queries using a BFS of depth 2 on the KG **(exhaustive but valid)**

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

Give me actors born in Berlin.

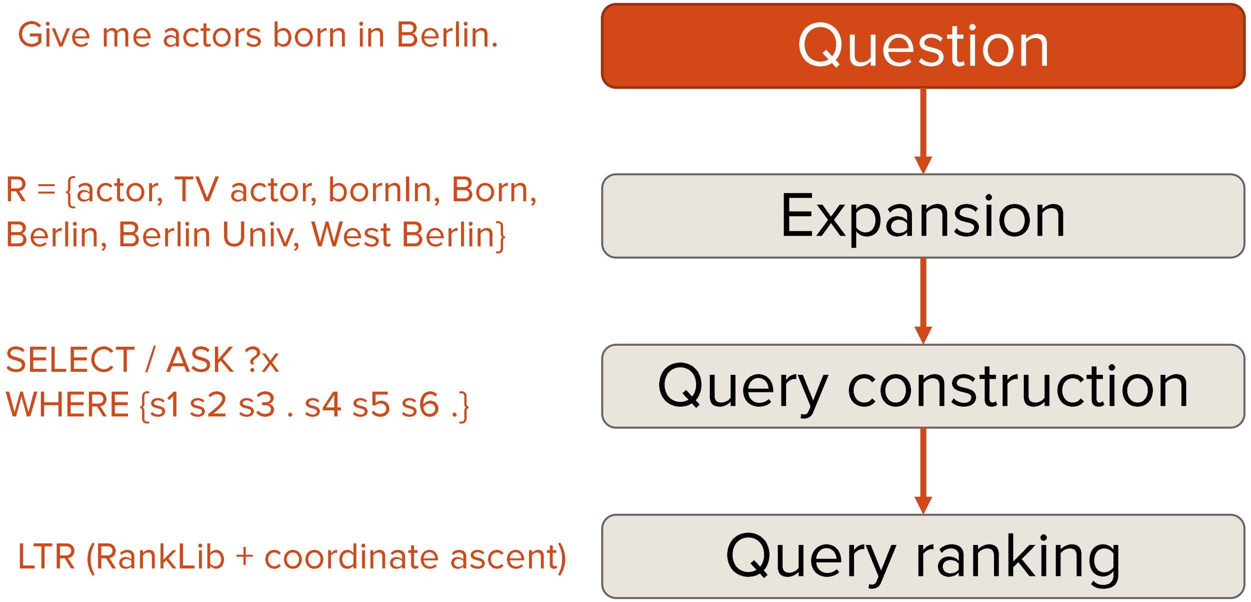
R = {actor, TVActor, bornIn, Born, Berlin, BerlinUniv, WestBerlin}

```
SELECT / ASK ?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



Structured querying over multiple KGs

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

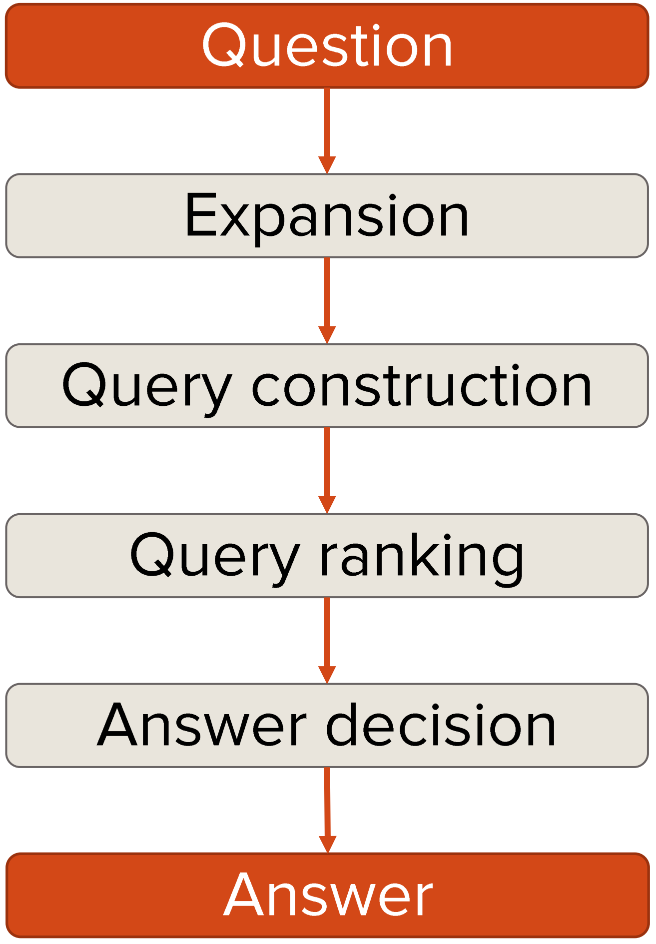
Give me actors born in Berlin.

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 ☹

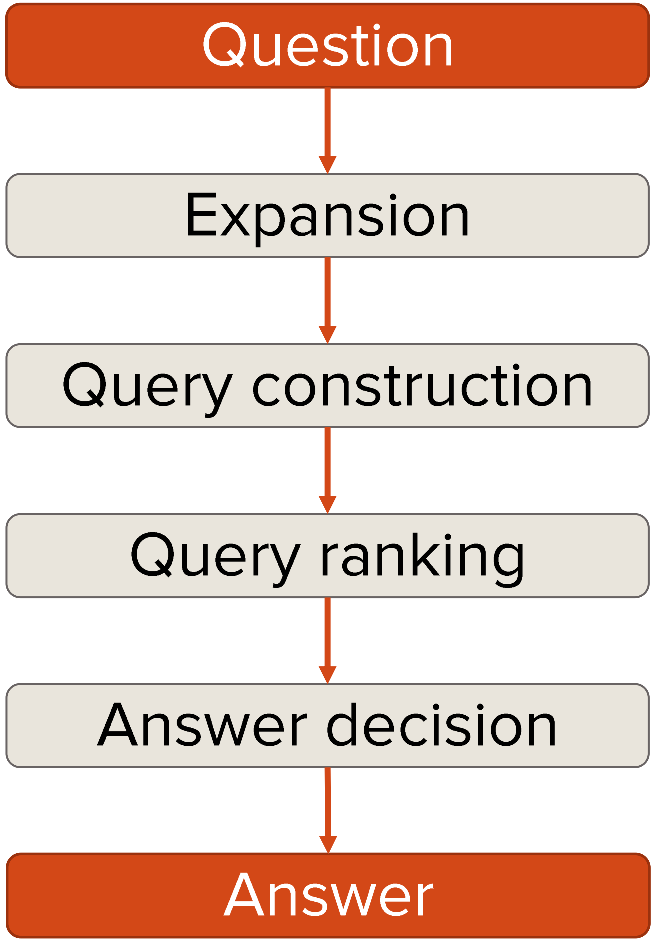
Give me actors born in Berlin.

R = {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

Which actor voiced the character Unicorn in The Last Unicorn?

Mia Farrow

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

Schmendrick

Who performed the songs in the movie The Last Unicorn?

America

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

Folk rock

Who was the director of the movie The Last Unicorn?

Jules Bass

Conversational KG-QA

Which actor voiced the character Unicorn in The Last Unicorn?

Mia Farrow

And Alan Arkin was behind …?

Schmendrick

The songs were by…?

America

Genre of this band?

Folk rock

By the way, who directed the movie?

Jules Bass

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

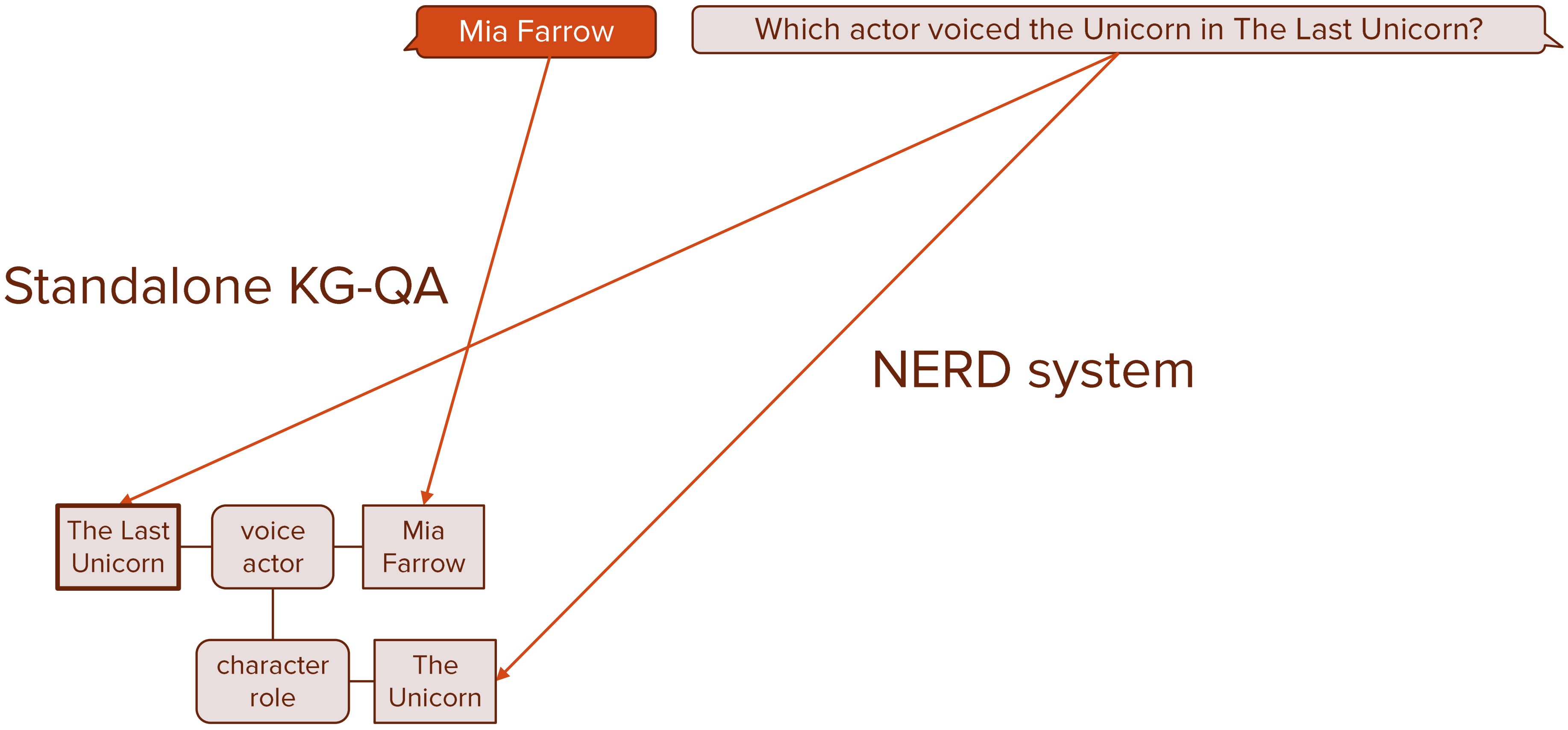
- **Key challenges** in conversational (KG-)QA
  - Infer implicit context
  - Handle ad hoc formulations
- Initially explored over **small tables** as sequential QA (Iyyer et al. 2017)
- Key direction for KG-QA now (Saha et al. 2018, Guo et al. 2018, **Christmann et al. 2019**, **Shen et al. 2019**)

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

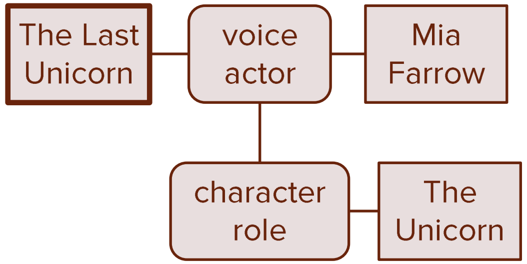


Initial context

Mia Farrow

Which actor voiced the Unicorn in The Last Unicorn?

And Alan Arkin was behind . . .?



How to expand the context?

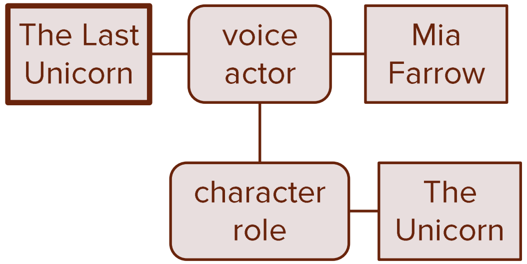

Neighborhood of
Mia Farrow
Neighborhood of
The Last Uni
orhood of
Unicorn

Judicious context expansion

Mia Farrow

Which actor voiced the Unicorn in The Last Unicorn?

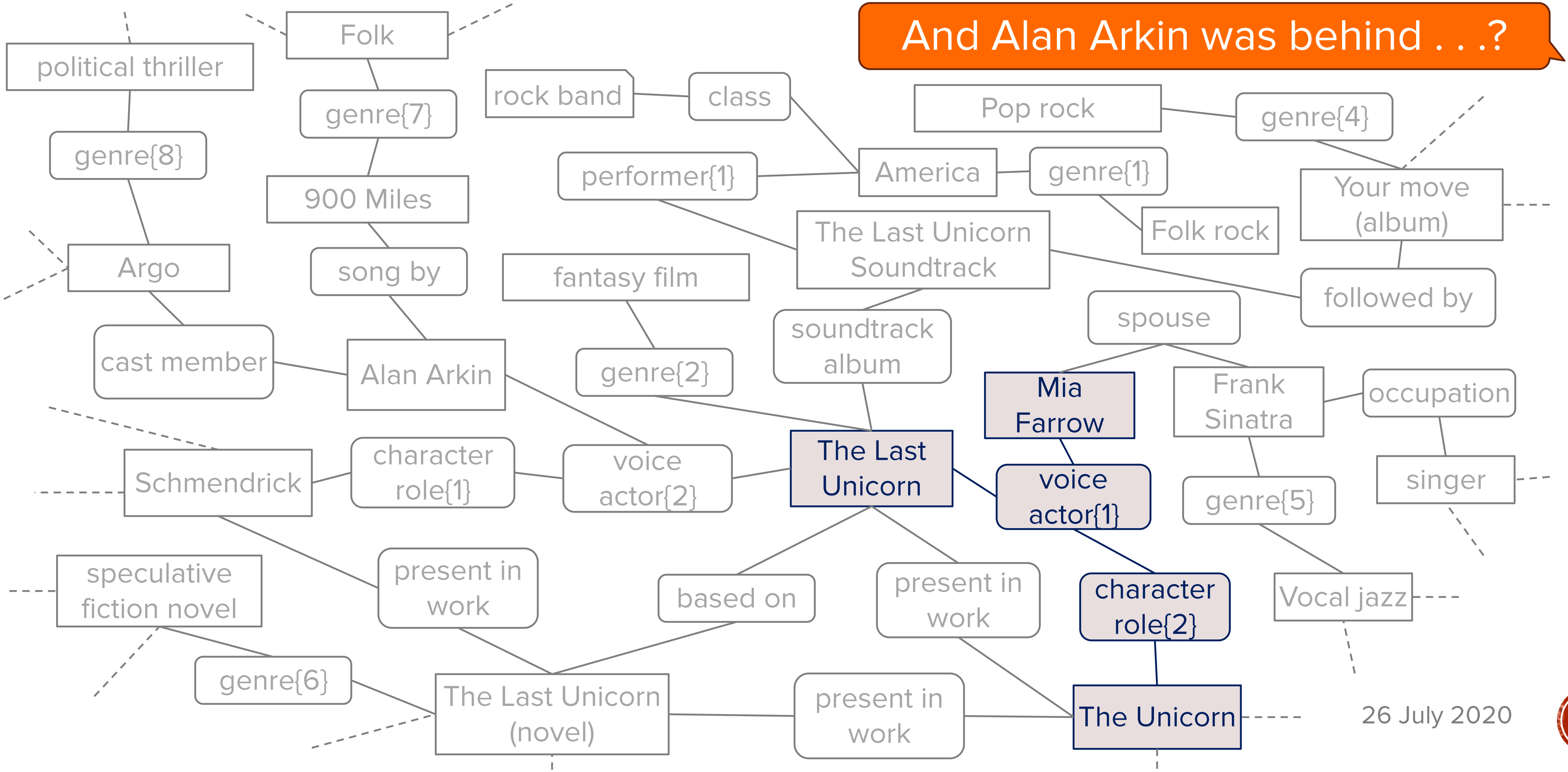
And Alan Arkin was behind . . .?



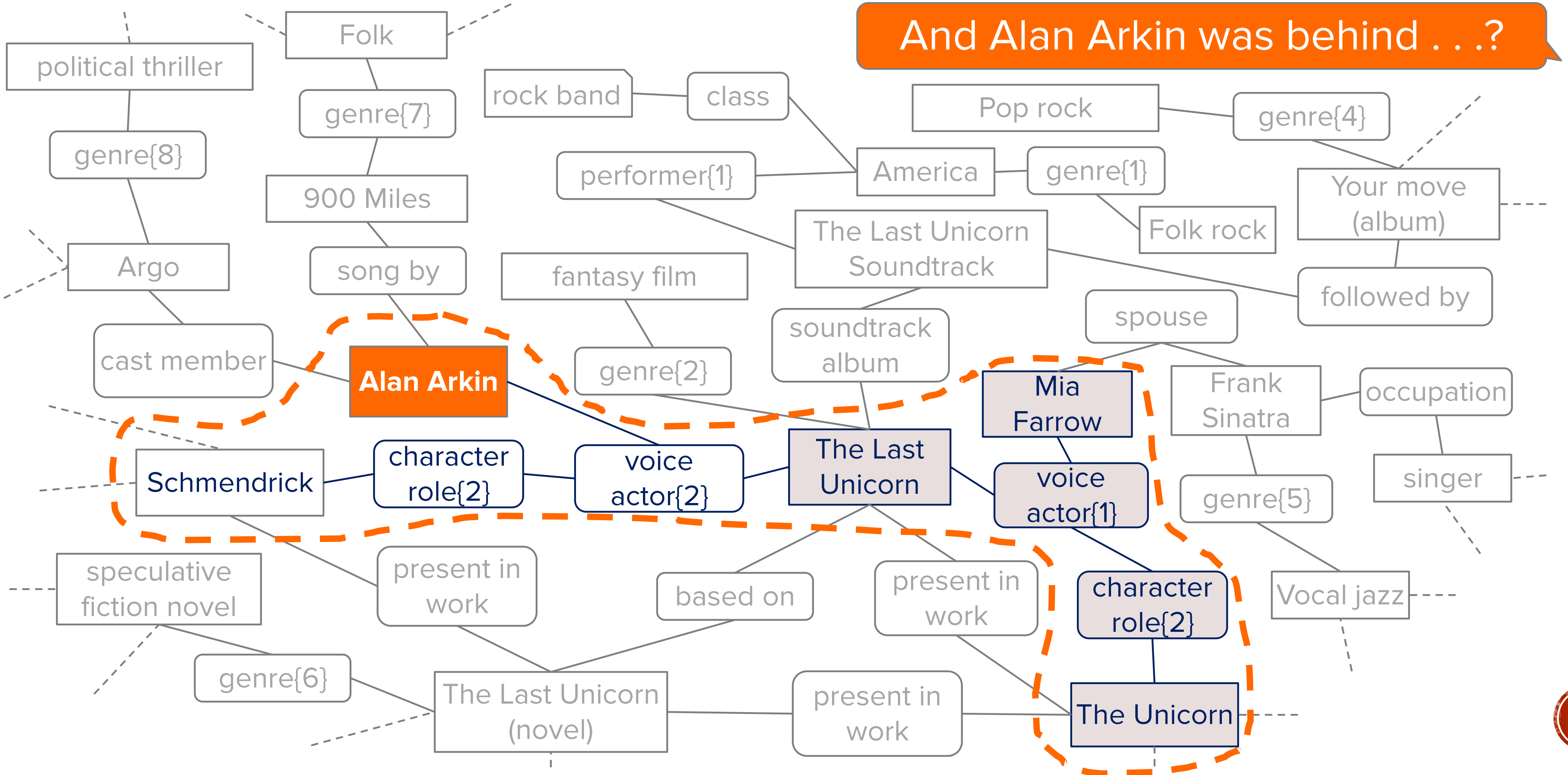
Do not expand with the complete neighborhood!

Exploring context neighborhood

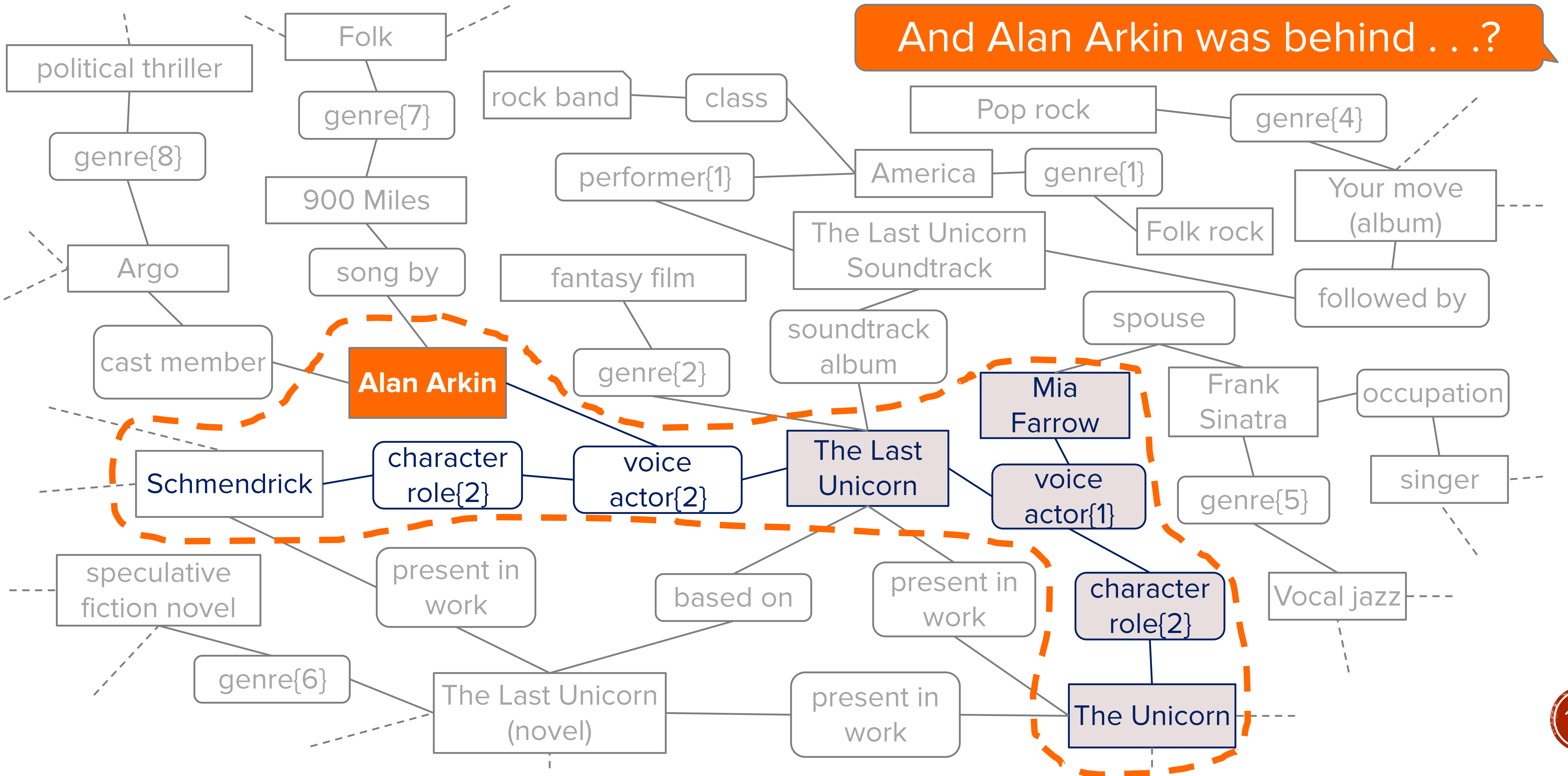
And Alan Arkin was behind . . .?



Exploring context neighborhood



Find frontier nodes to define expansion border

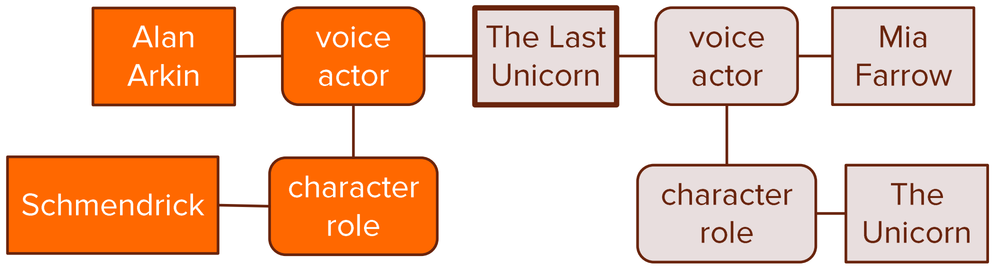


Context graph

Mia Farrow

Which actor voiced the Unicorn in The Last Unicorn?

And Alan Arkin was behind . . .?



Graph expanded with relevant facts only

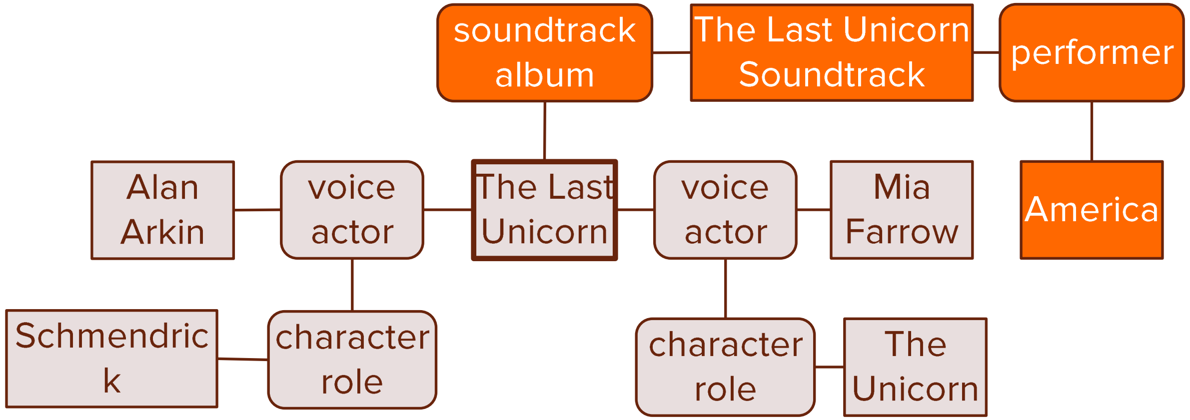
Context graph

Mia Farrow

Which actor voiced the Unicorn in The Last Unicorn?

And Alan Arkin was behind . . .?

So who performed the songs?



Graph expanded with relevant facts only

Context graph

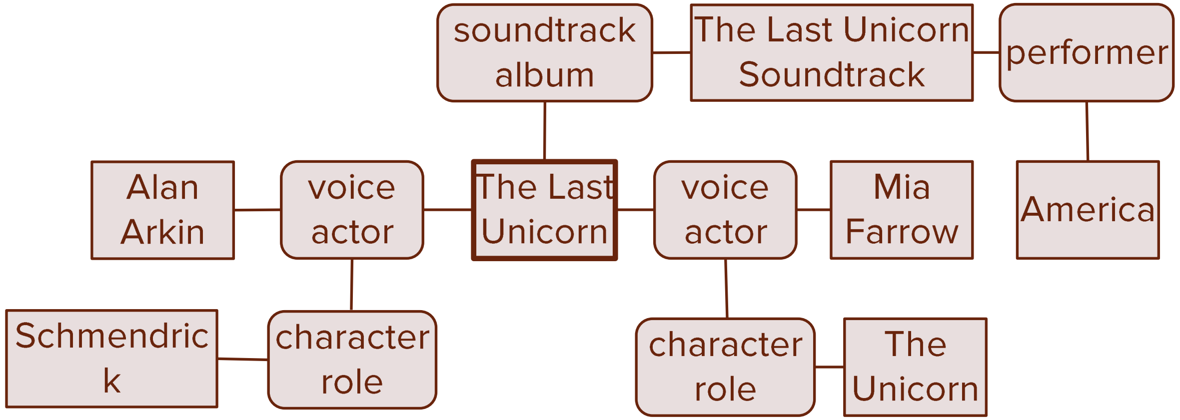
Which actor voiced the Unicorn in The Last Unicorn?

Mia Farrow

And Alan Arkin was behind . . .?

So who performed the songs?

Genre of this band?



How to determine Frontier nodes?

Frontier score

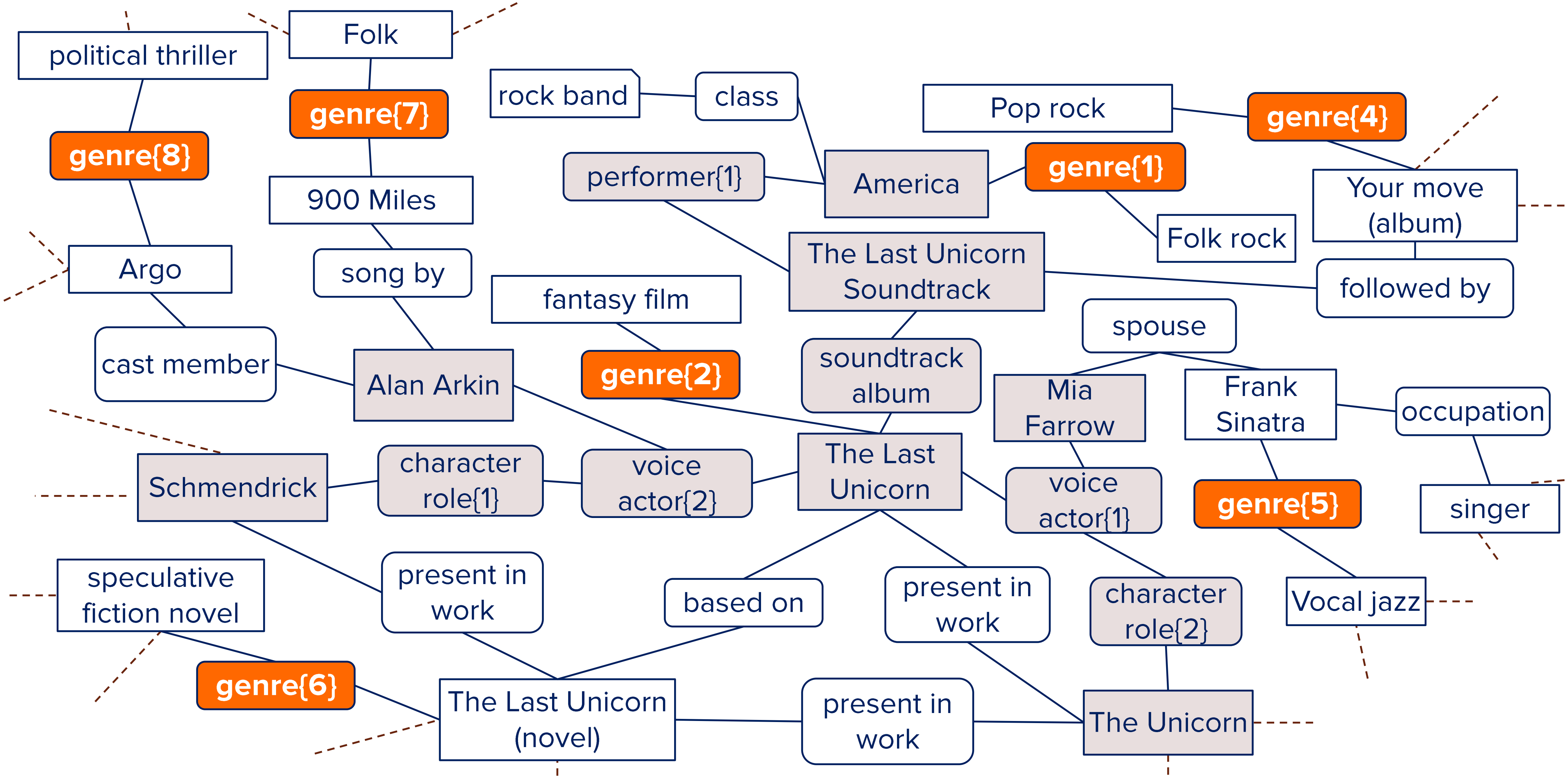
Matching similarity	**Context relevance**	**KG priors**
match (candidate c)	*prox (candidate c)*	*prior (candidate c)*

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

With hyperparameters h_1, h_2, h_3

The great disambiguation

Genre of this band?



Frontier nodes

Matching similarity

Candidate	*Match*
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

Candidate	*Prox*
genre{1}	0.91
folk rock band	0.86
RSH-Gold for Cult Band	0.86
…	…
genre{2}	0.34
fantasy film	0.36
…	…

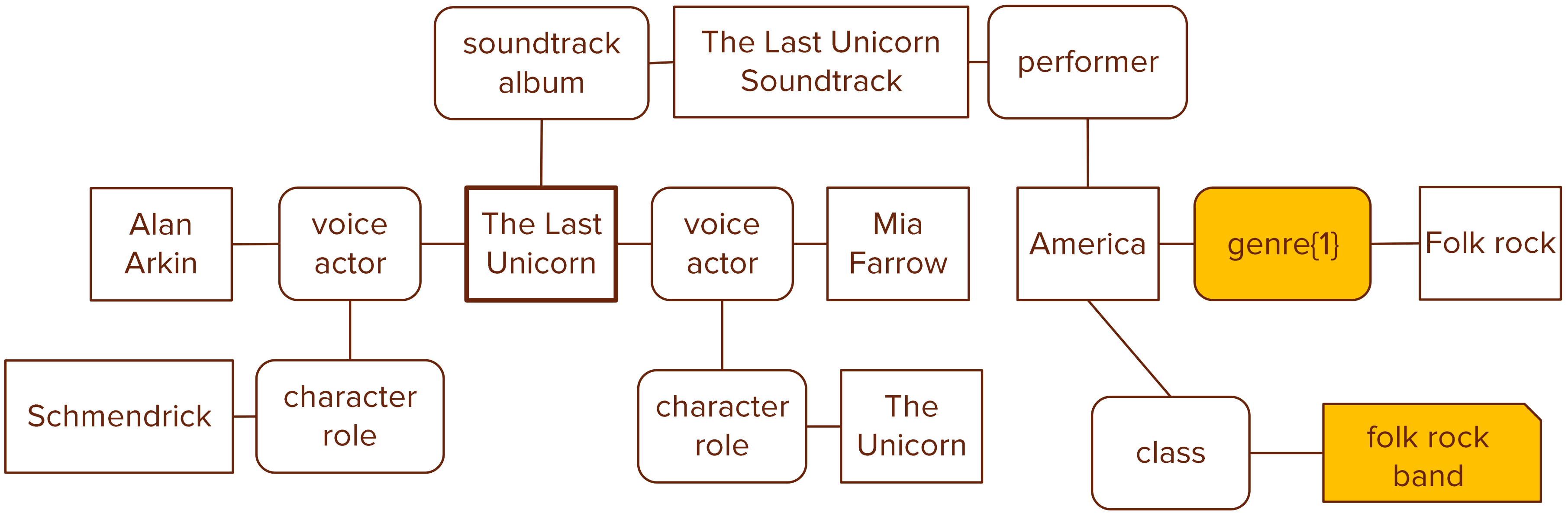
KG priors

Candidate	*KG priors*
…	…
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

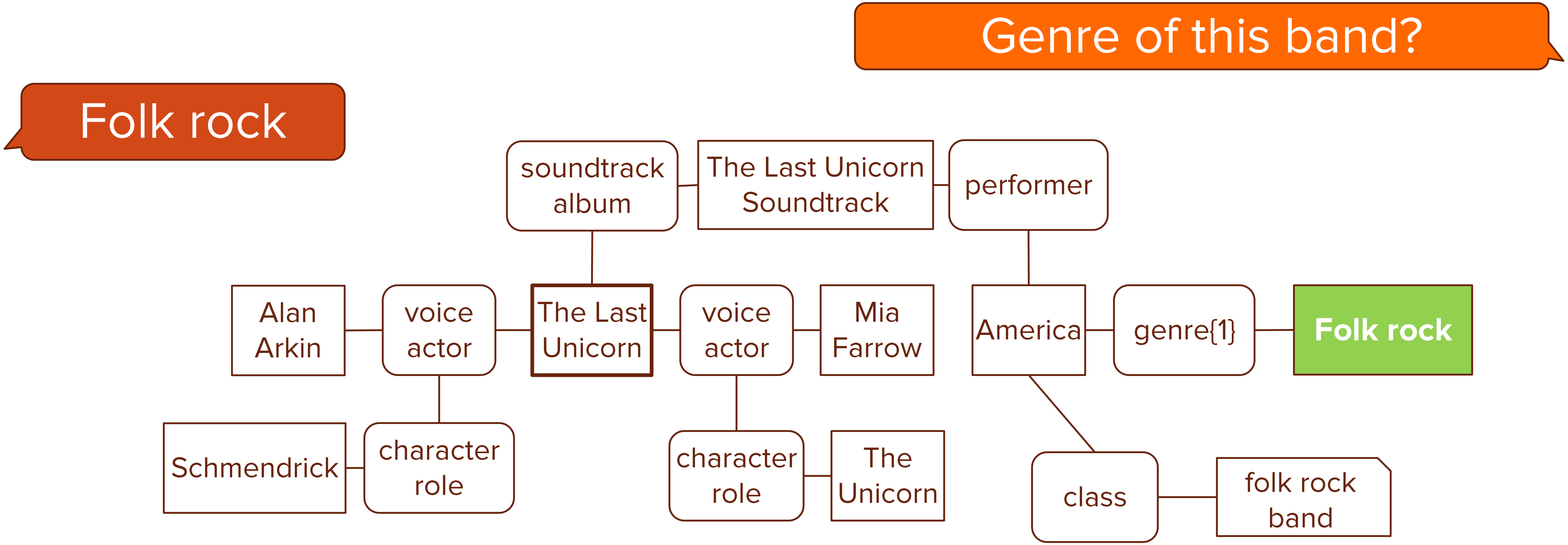
- Distance to all nodes in context graph X
  - Weighted by the turn they occurred in
  - distance_X

=> Implicit part

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

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

Answer detection



Top-ranked node according to *answer_score*

Conversational QA: Sequence-to-sequence modeling

- The **MaSP** model (Shen et al. 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

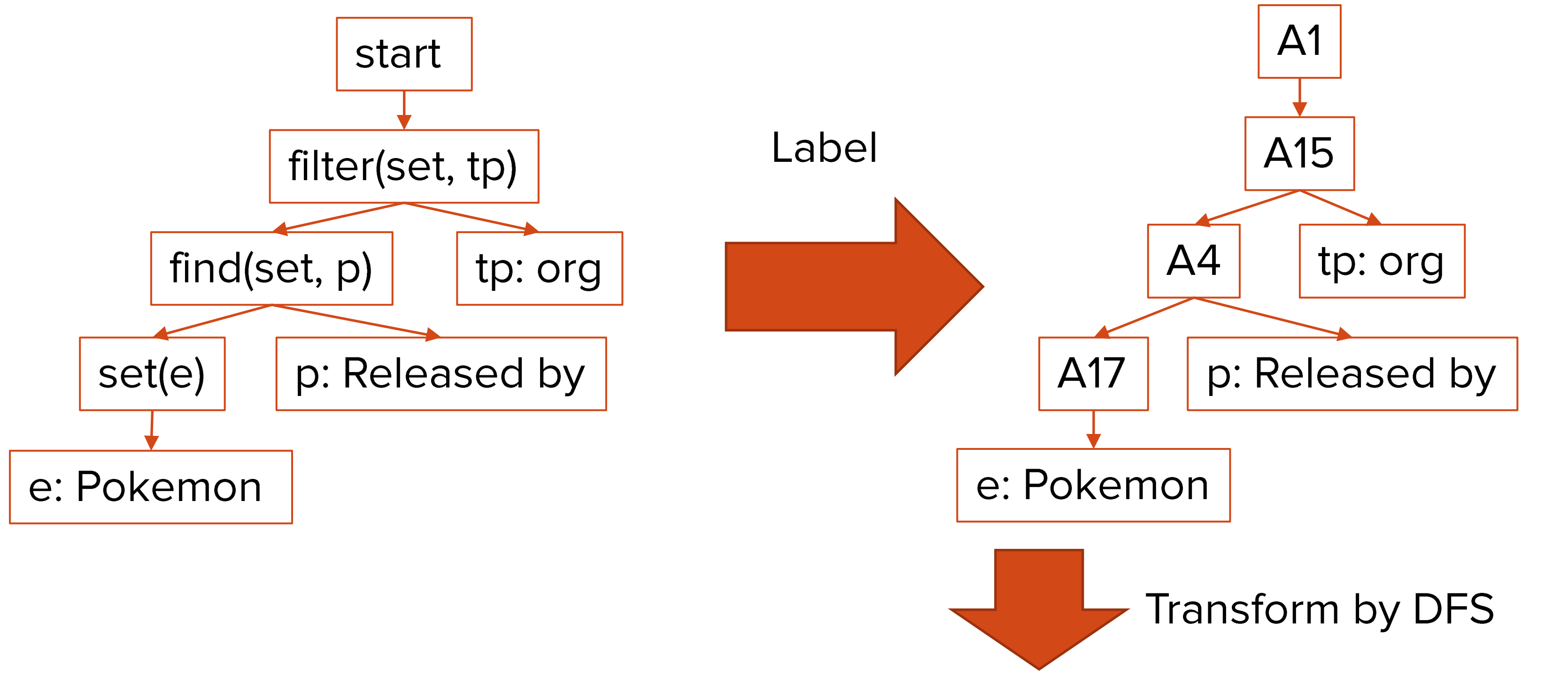
Shen et al., Multi-Task Learning for Conversational Question Answering over a Large-Scale Knowledge Base, EMNLP 2019.

Sequence-to-sequence model

NL question as sequence

Question: Who released Pokemon?

Tree-structured KB-query



Query as sequence

A1 A15 A4 A17 e:Pokemon p:ReleasedBy tp:org

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

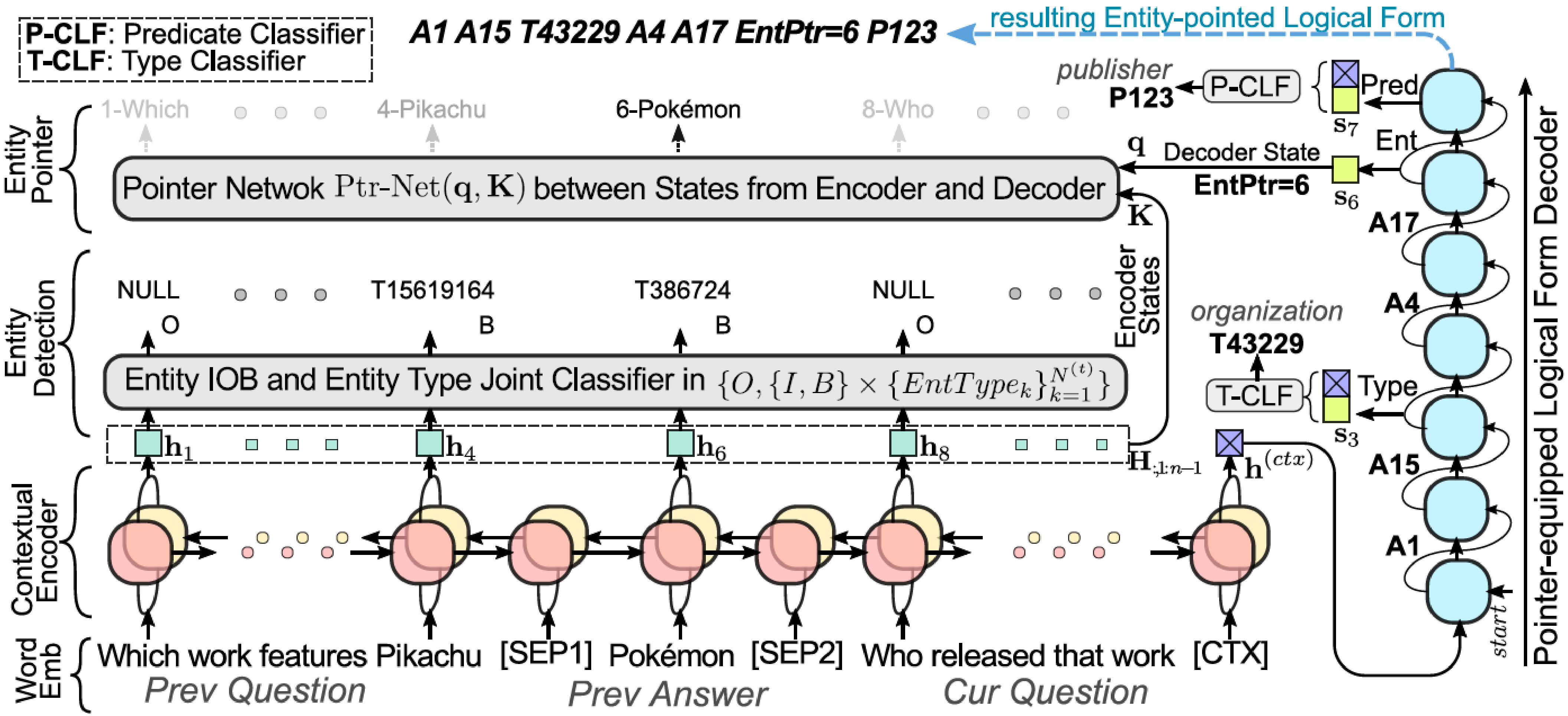
Sequence-to-sequence model

NL question as sequence

Question: [CONTEXT] Who released that work?

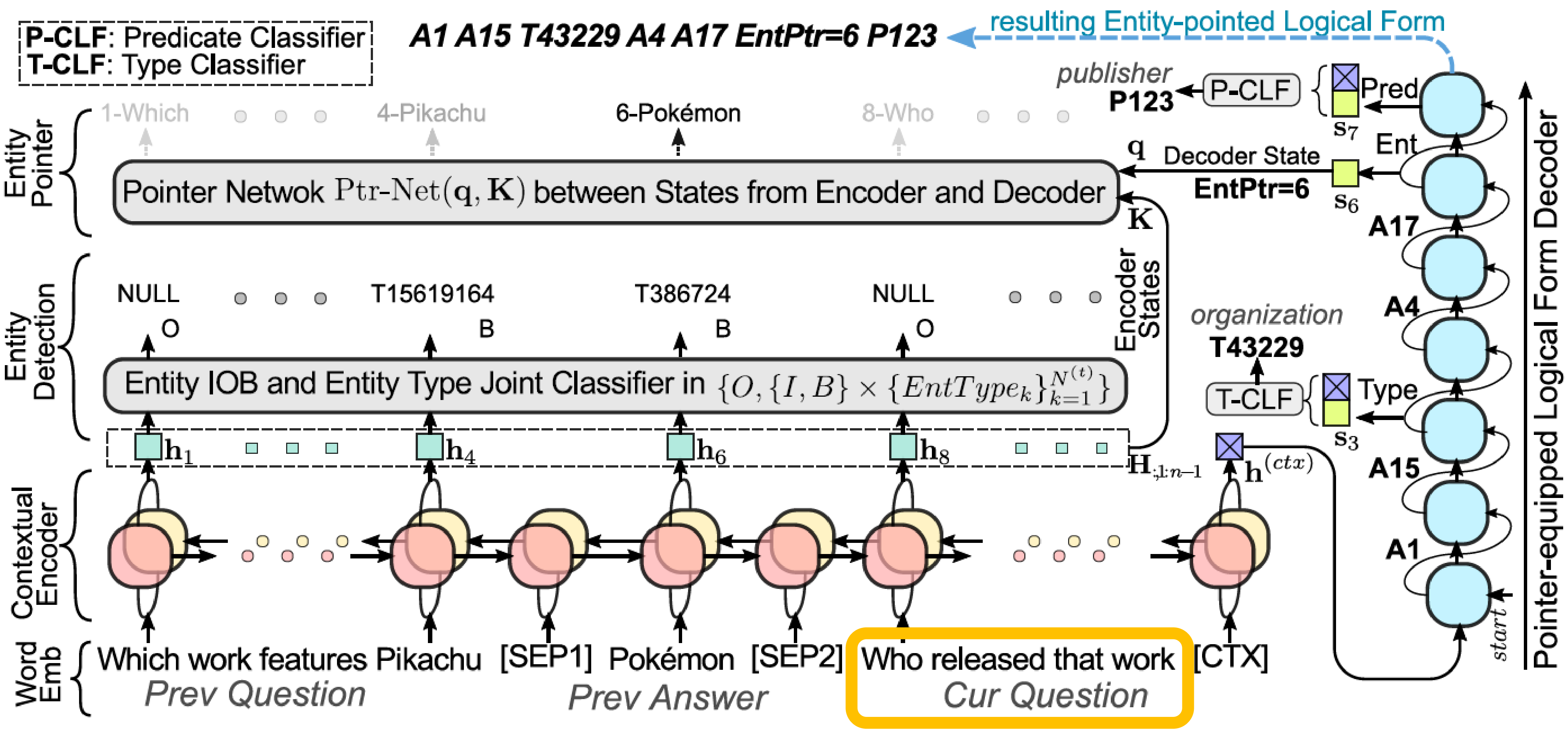
Shen et al., Multi-Task Learning for Conversational Question Answering over a Large-Scale Knowledge Base, EMNLP 2019.

Seq-to-seq: The MaSP model



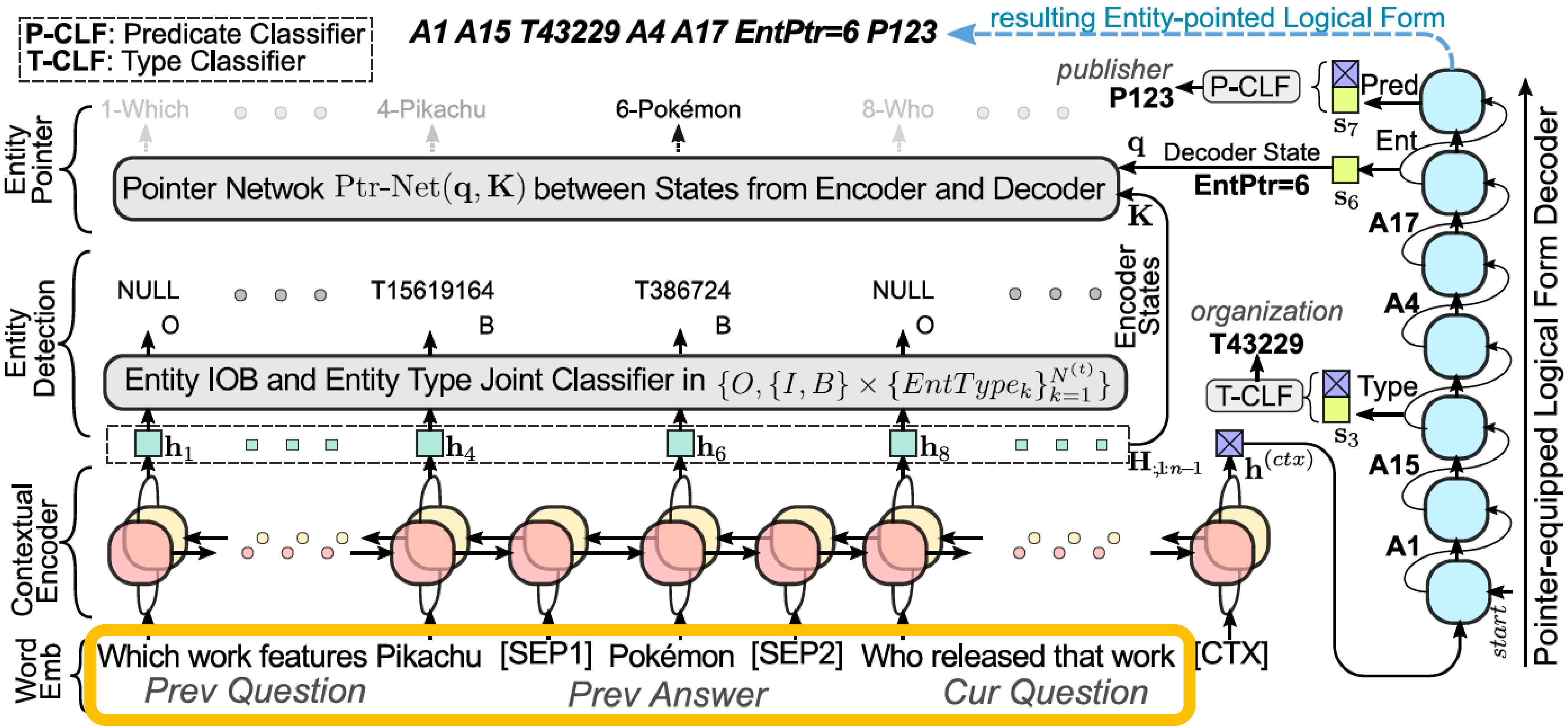
Shen et al., Multi-Task Learning for Conversational Question Answering over a Large-Scale Knowledge Base, EMNLP 2019.

MaSP: Step-by-step



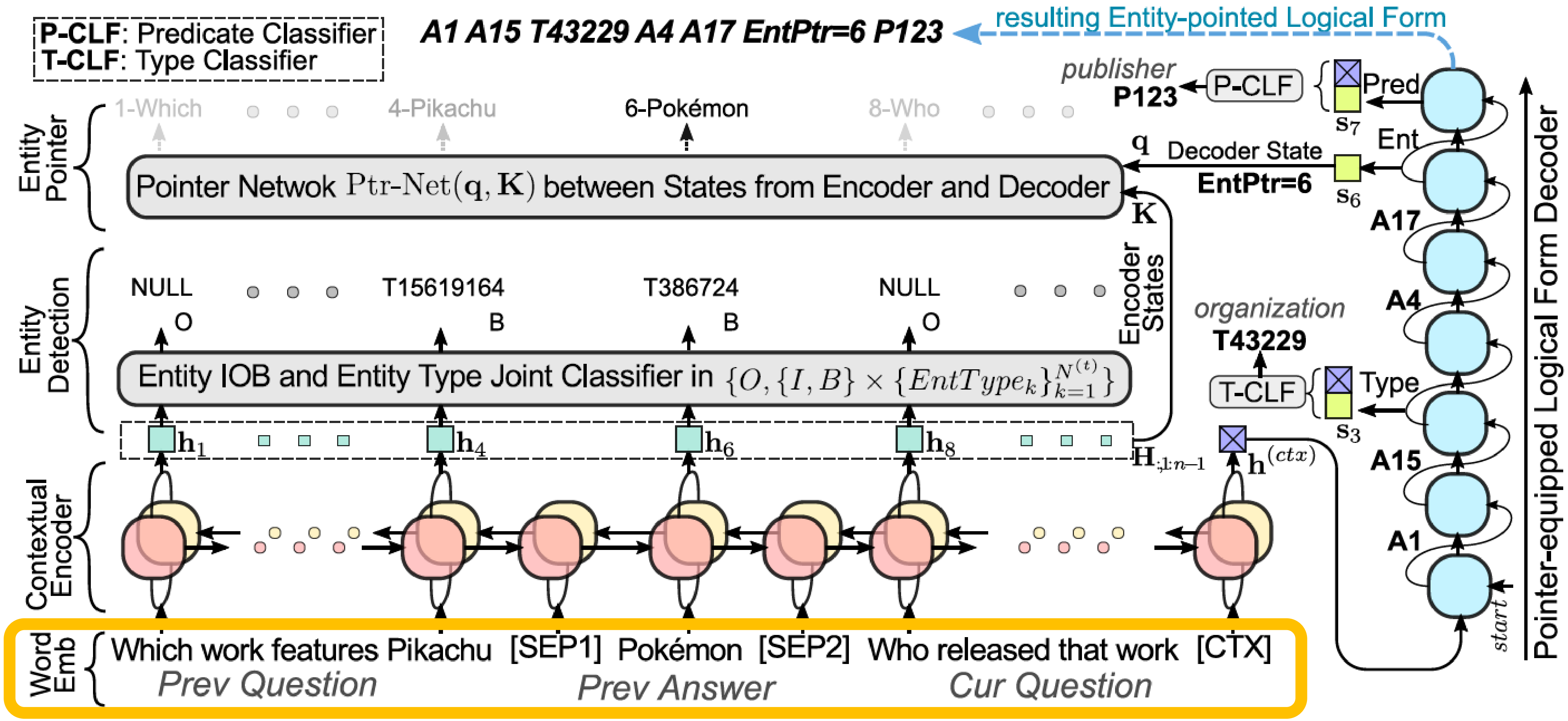
Shen et al., Multi-Task Learning for Conversational Question Answering over a Large-Scale Knowledge Base, EMNLP 2019.

MaSP: Step-by-step



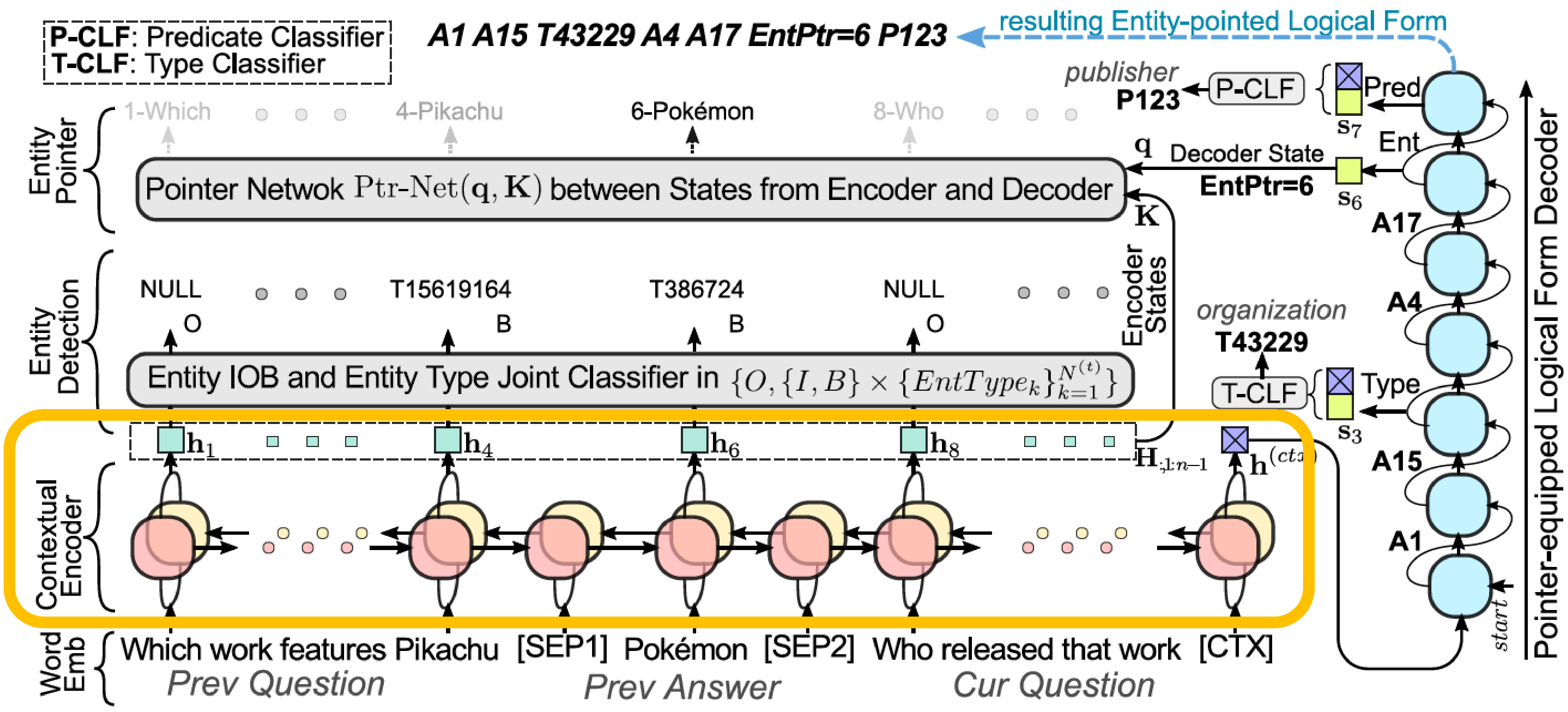
Shen et al., Multi-Task Learning for Conversational Question Answering over a Large-Scale Knowledge Base, EMNLP 2019.

MaSP: Step-by-step



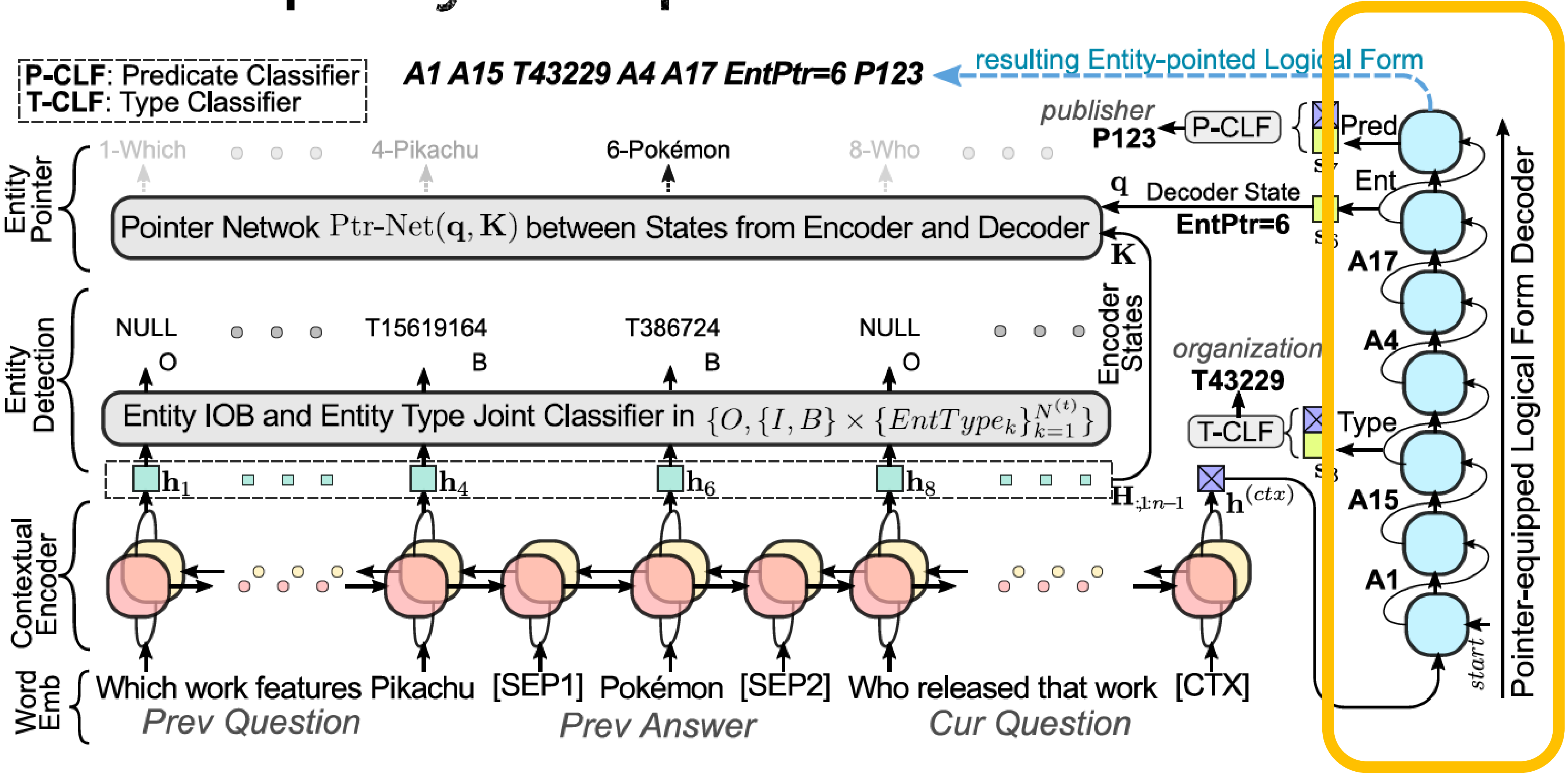
Shen et al., Multi-Task Learning for Conversational Question Answering over a Large-Scale Knowledge Base, EMNLP 2019.

MaSP: Step-by-step



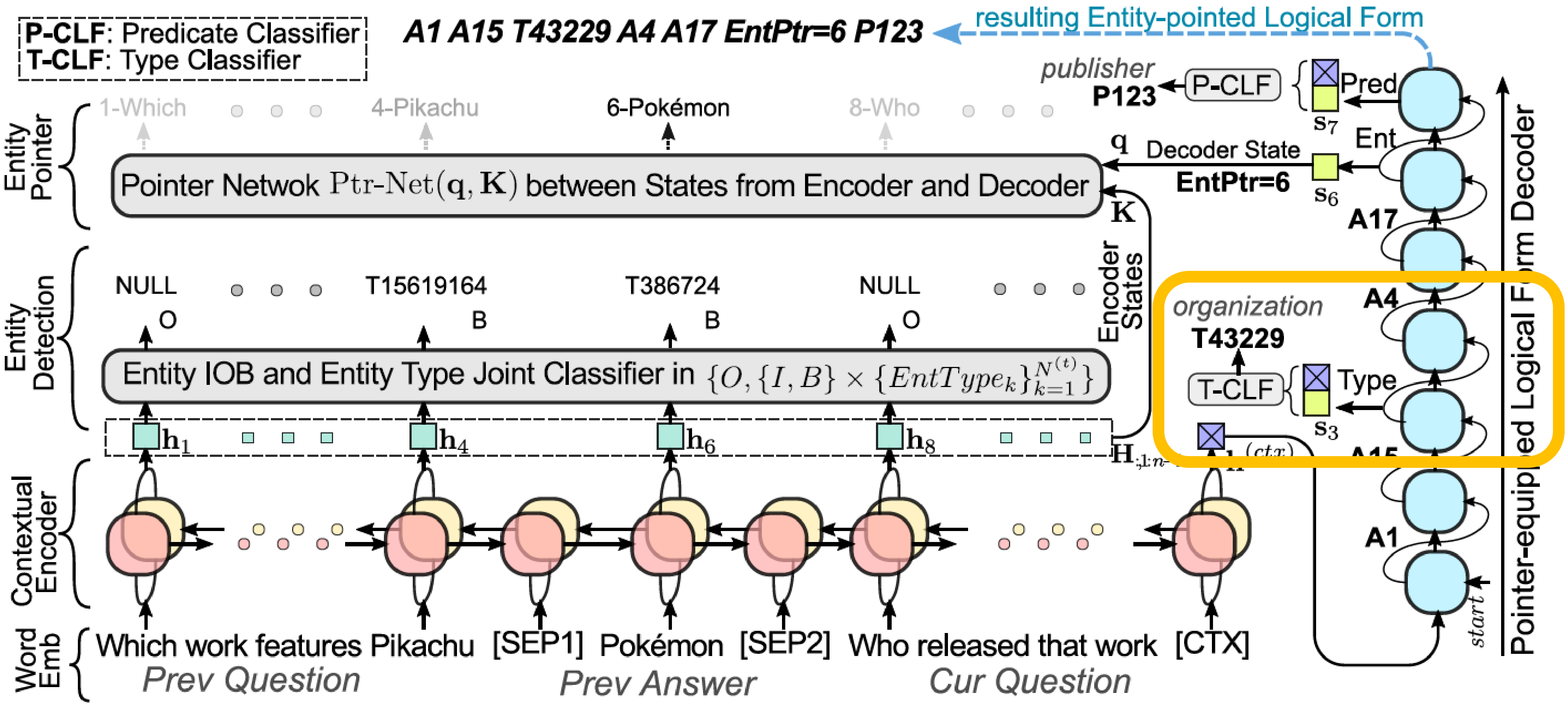
Shen et al., Multi-Task Learning for Conversational Question Answering over a Large-Scale Knowledge Base, EMNLP 2019.

MaSP: Step-by-step



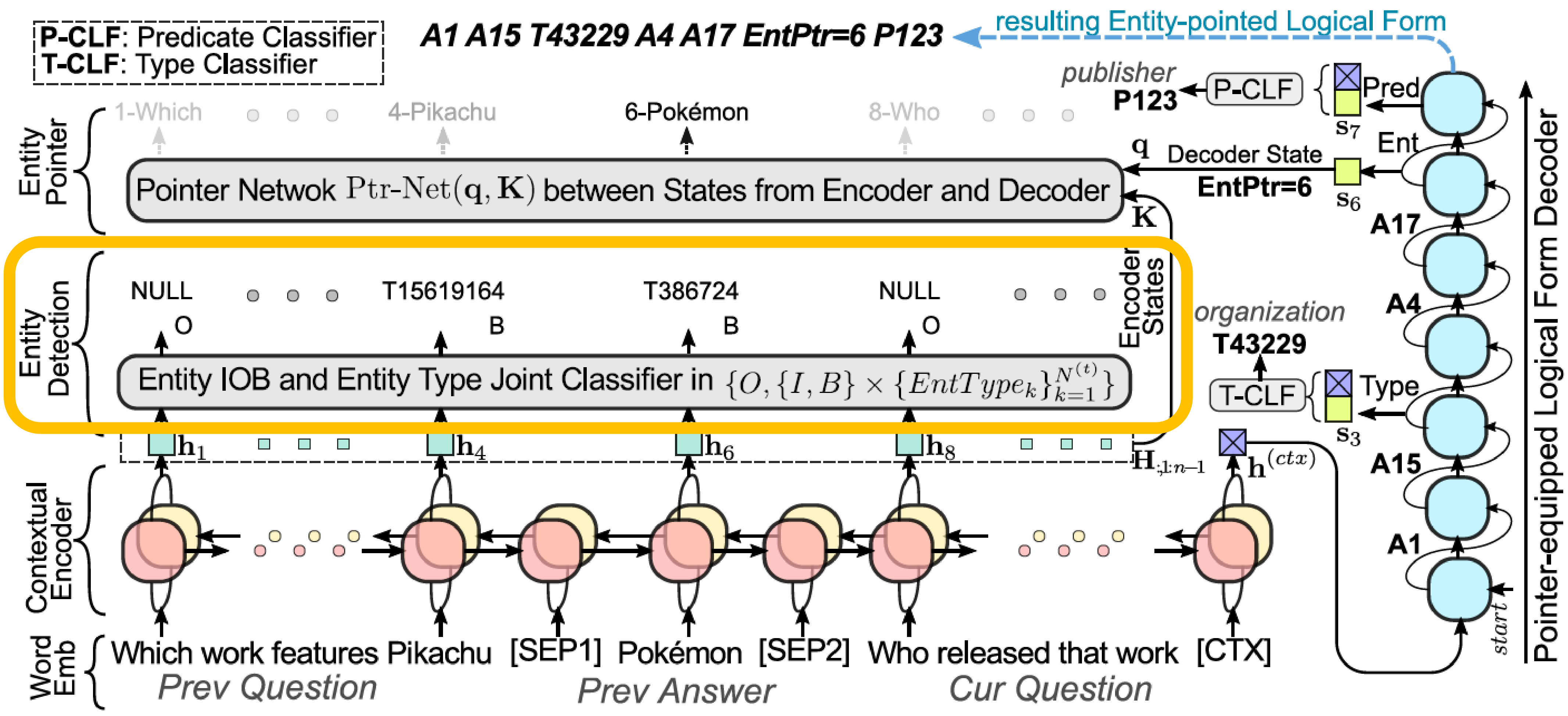
Shen et al., Multi-Task Learning for Conversational Question Answering over a Large-Scale Knowledge Base, EMNLP 2019.

MaSP: Step-by-step



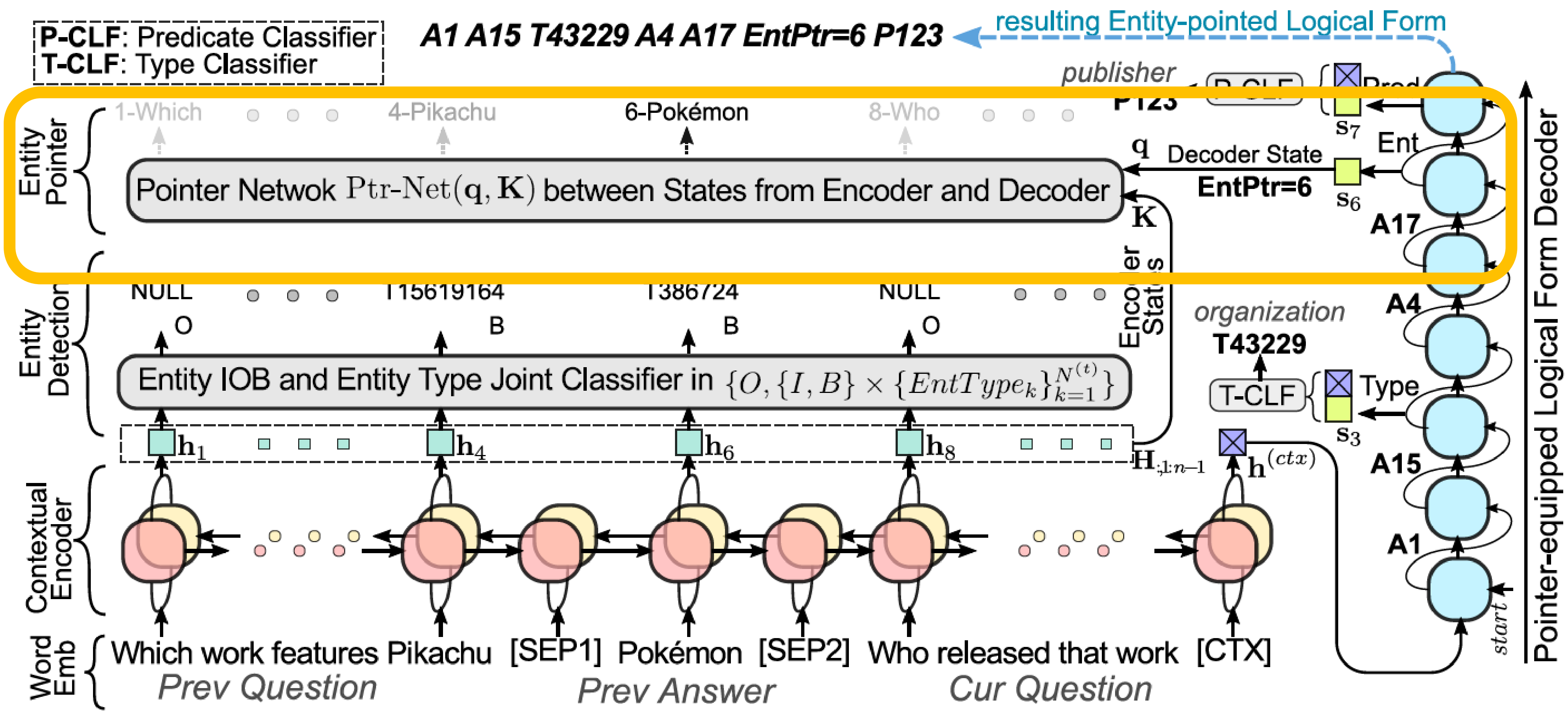
Shen et al., Multi-Task Learning for Conversational Question Answering over a Large-Scale Knowledge Base, EMNLP 2019.

MaSP: Step-by-step



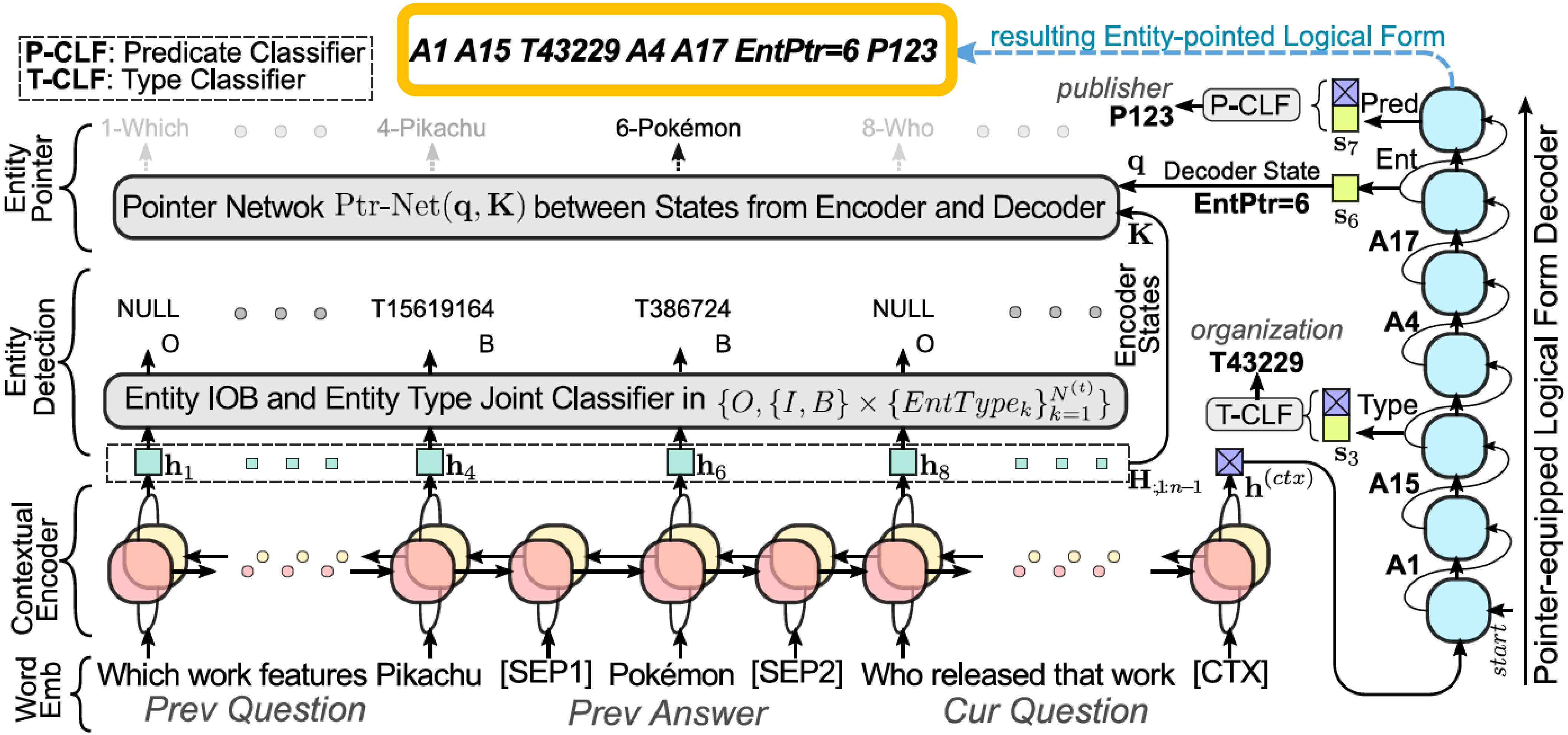
Shen et al., Multi-Task Learning for Conversational Question Answering over a Large-Scale Knowledge Base, EMNLP 2019.

MaSP: Step-by-step



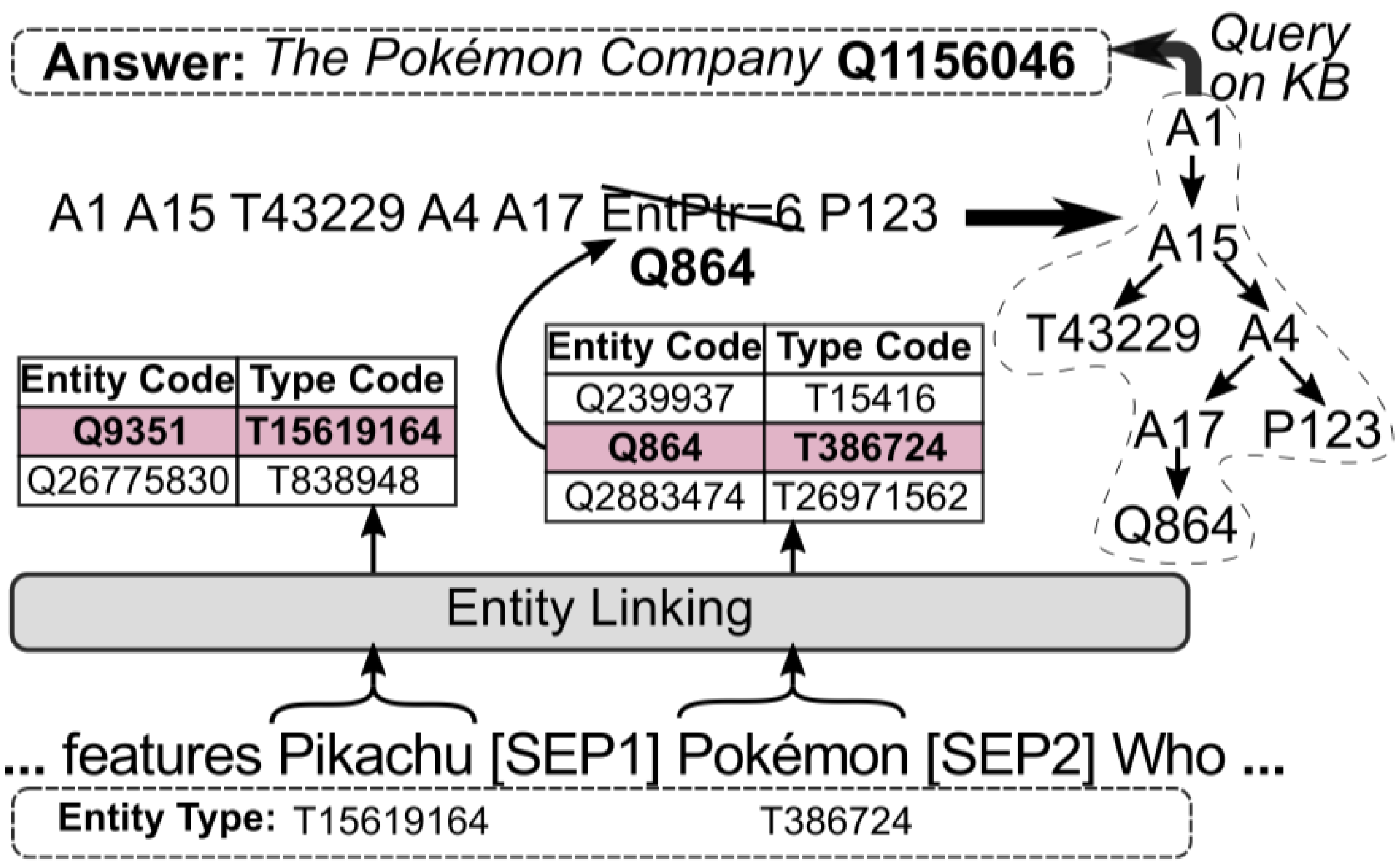
Shen et al., Multi-Task Learning for Conversational Question Answering over a Large-Scale Knowledge Base, EMNLP 2019.

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

Weakly/strongly supervised	MaSP, NEQA, TextRay	KEQA, PullNet
Unsupervised	QAnswer*	QAmp*, QUEST, CONVEX
	Explicit structured query	Without explicit query (approximate graph search)

Degree of supervision

Query generation

* Ranker/labeler supervised

Methodology quad chart

Degree of supervision ↑	Explicit structured query	Without explicit query (approximate graph search)
Weakly/strongly supervised	MaSP, TextRay, NEQA, D2A, Saha et al. 2018, SEMPRE, AQQU, STAGG, PARALEX, OQA, PARASEMPRE, Cai and Yates 2013	KEQA, PullNet, GRAFT-Net, GraphParser, Bordes et al. 2014
Unsupervised	QAnswer*, DEANNA, Unger et al. 2012	QAmp*, QUEST, CONVEX

Query generation →

* Ranker/labeler supervised

Methodology quad chart

	Explicit structured query	Without explicit query	Hybrid search methods?
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	There is some interplay in current systems but largely open area Unsupervised subgraph computations with small degree of supervised neural learning..?
Weakly supervised with paraphrases	PARASEMPRE, PARALEX, OQA	-	
Unsupervised	QAnswer*, DEANNA, Unger et al. 2012	QAmp*, QUEST, CONVEX	

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)

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

Open problems

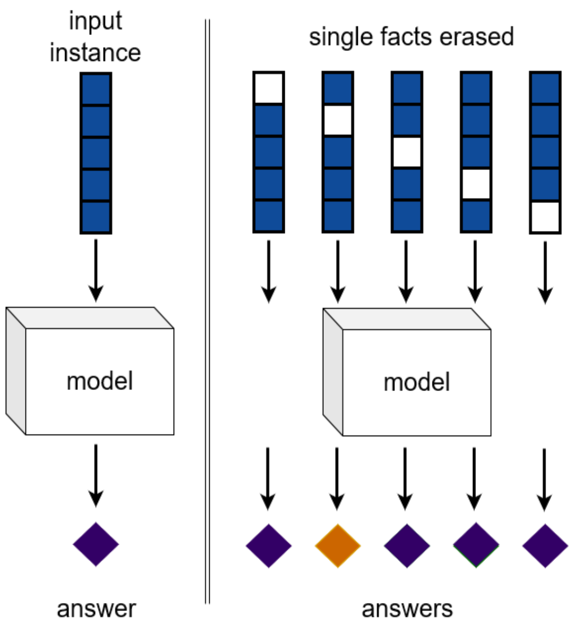
- Unanswerability
- Interpretability
- Interactivity
- Efficiency
- Robustness

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?

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

- **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 in case of any issues!

Acknowledgements

- Gerhard Weikum for valuable feedback on slides
- Authors of several papers for sharing additional content
- Members of D5@MPII for inputs
- Organizers @SIGIR2020

Question Answering over Curated and Open Web Sources

Rishiraj Saha Roy

Max Planck Institute for Informatics, Germany

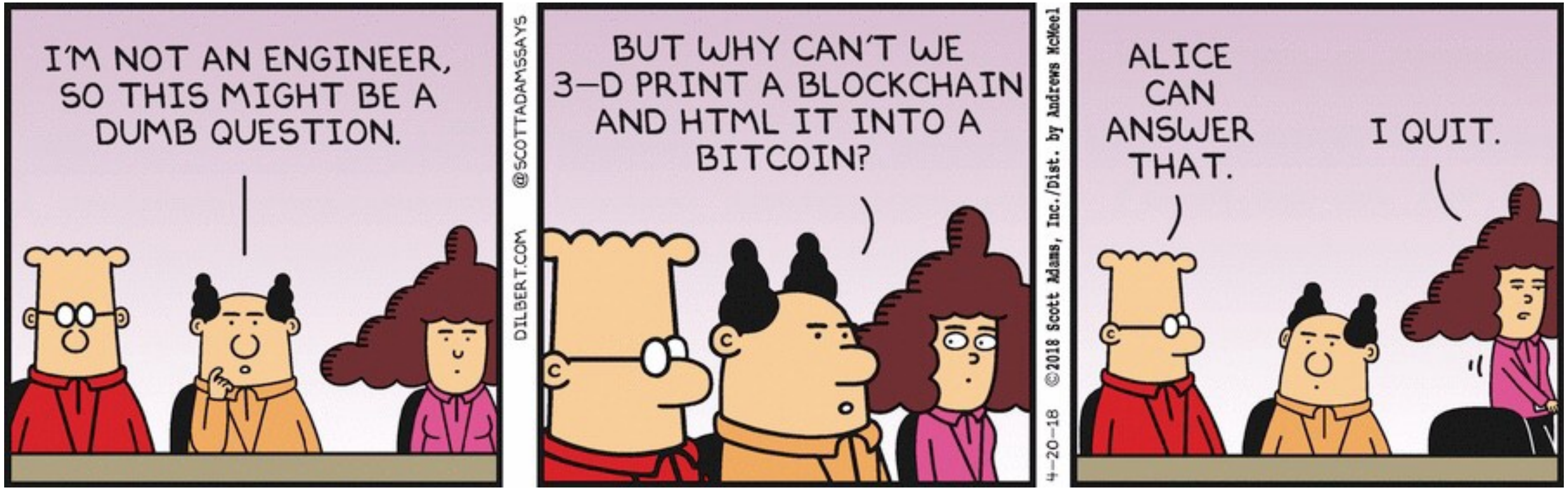
Avishek Anand

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

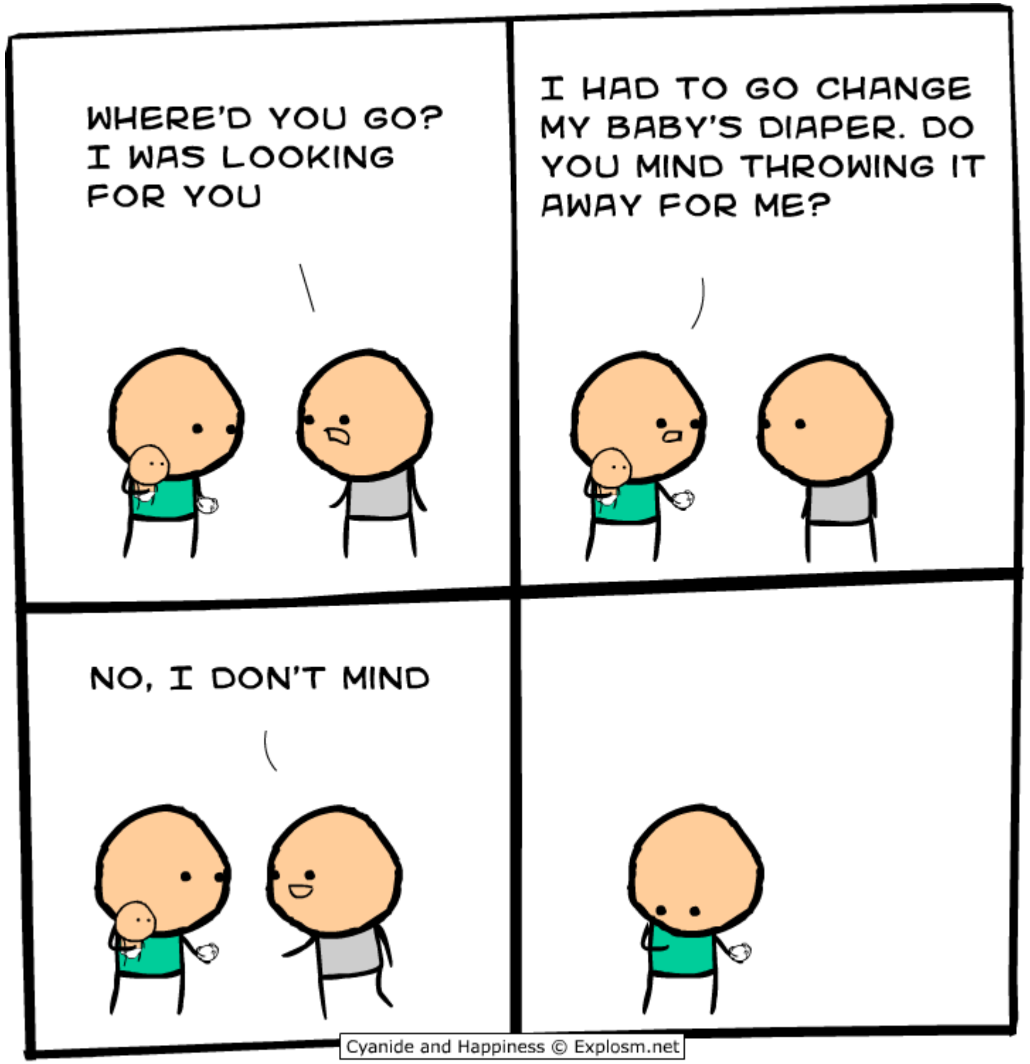
Passage: *Anita was stung by a bee and left the garden.*

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

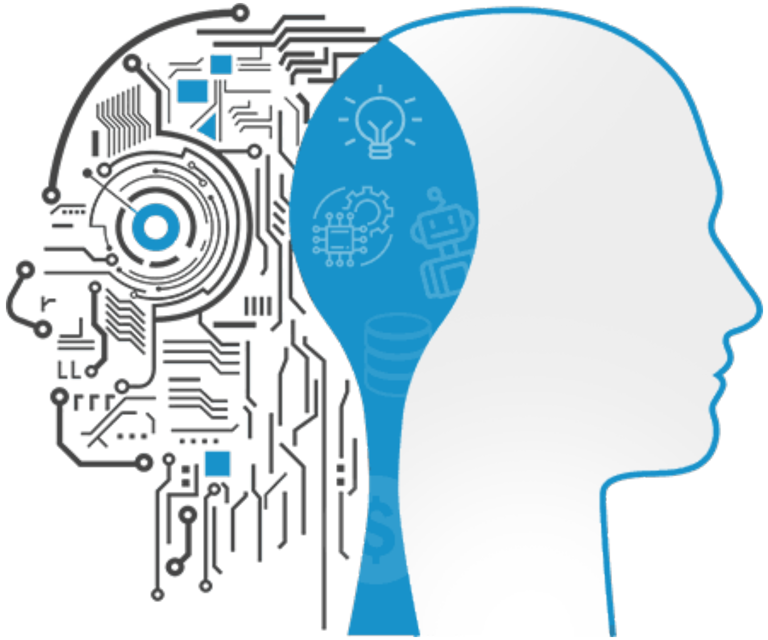
— *Eve, Nathan and Caleb*

Lots of Success

Lots of Applications

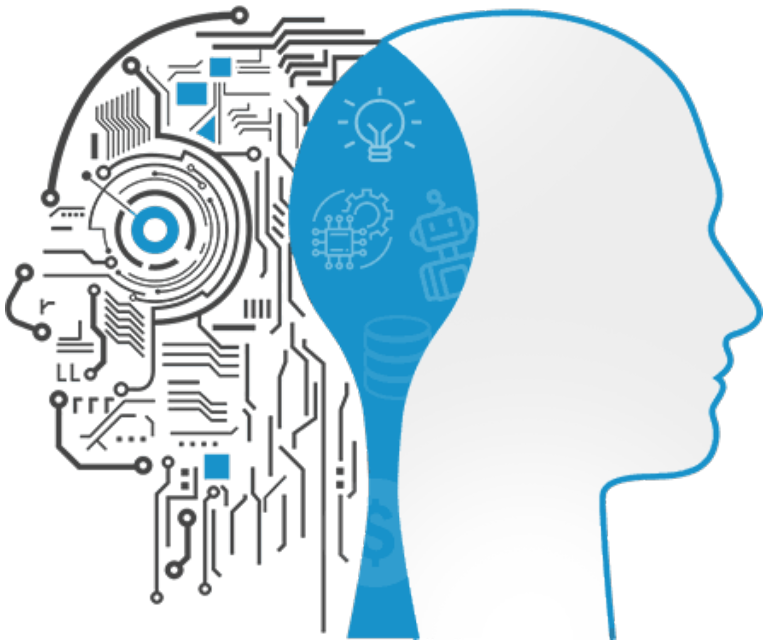
Problems in NLP, Dialog and Search can be formulated as QA

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

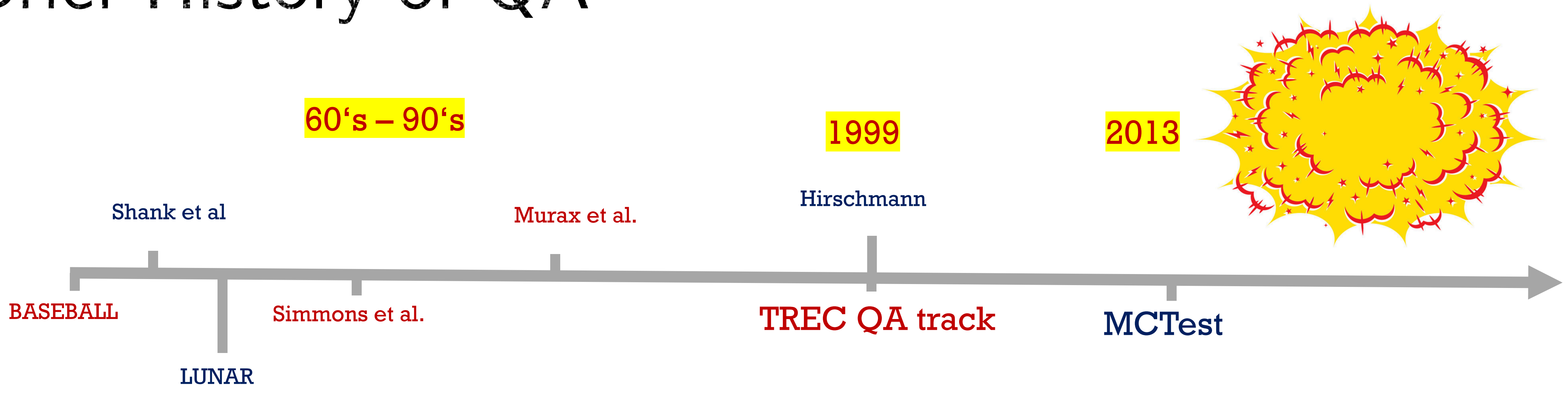
What we do not cover



- 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

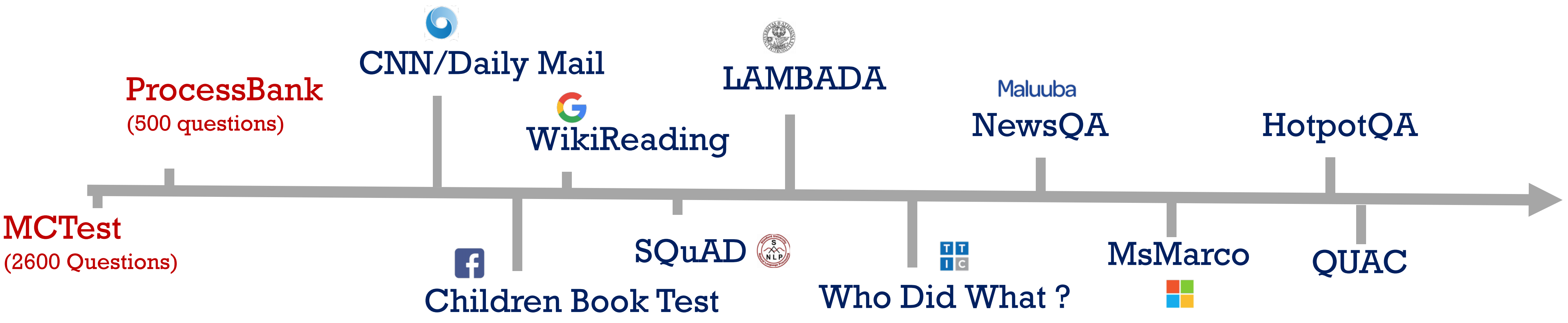
- Types of Questions:

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

context

question → Document or a Passage → answer

What is the capital of Australia ?

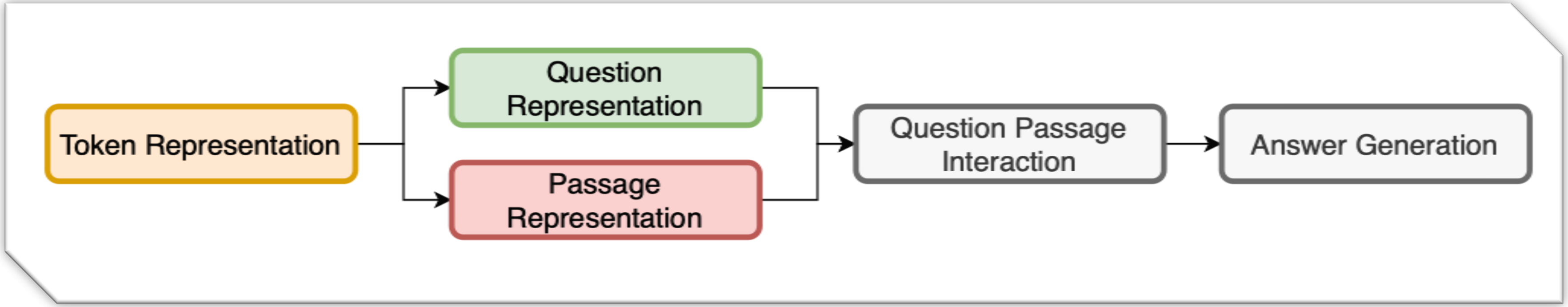
As the seat of the government of Australia, Canberra is home to many important institutions of the federal government, national monuments and museums. Canberra is also the capital of the country.

Canberra

- Extractive phrase
- Sentence

Canberra is also the capital of the country.

The MRC Pipeline



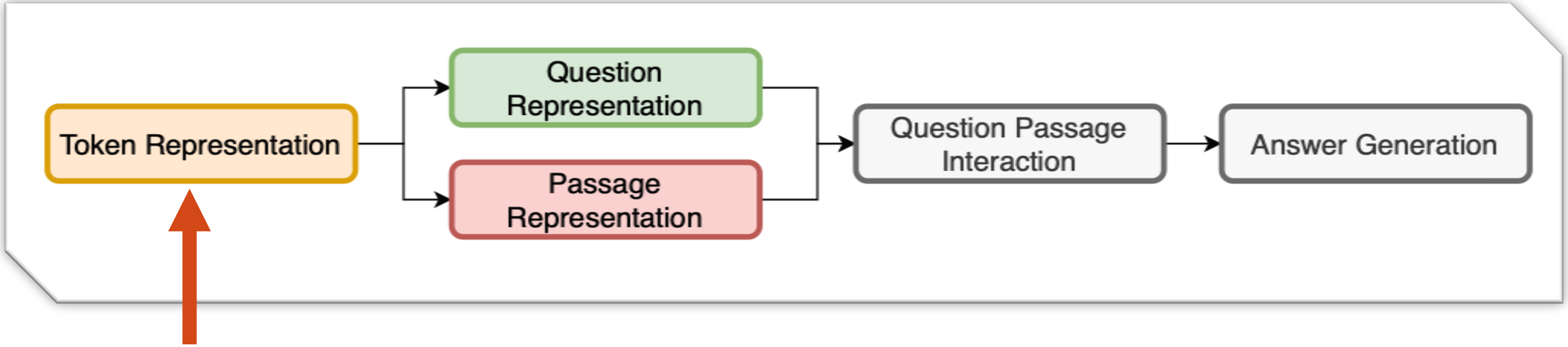
- Words, characters, sub-words embeddings
- Contextual Embeddings
- Other features – Matching, Alignment, Language structure

- Sequential representation
- Contextual representation
- Attentive reading

- Attentive reading
- Attention flows
- Multiple input passes inputs
- Re-representation of question and passages

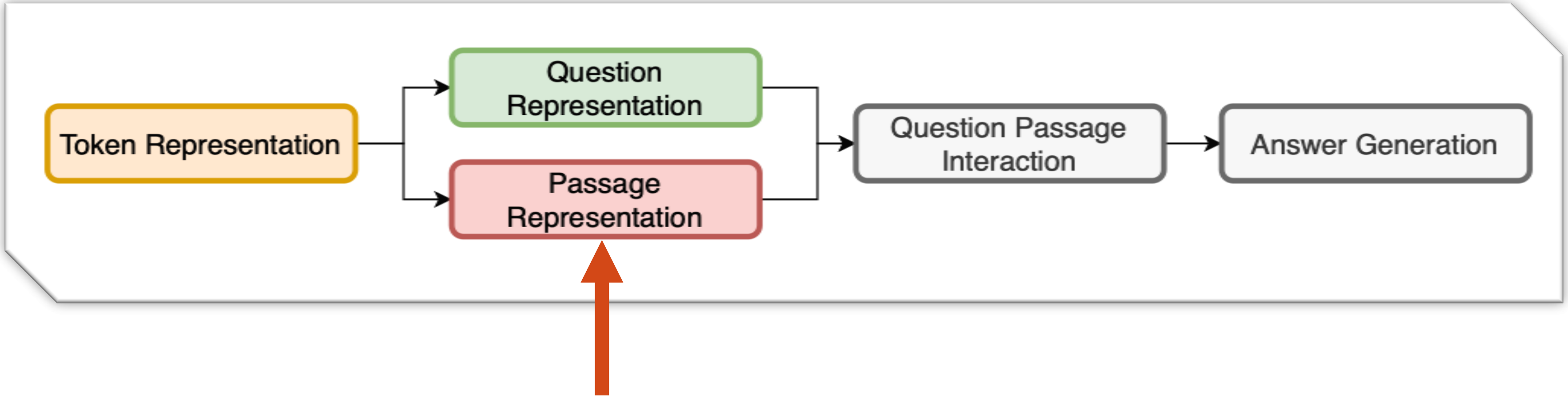
- Token prediction
- Span prediction
- Free-form generation

Token Representation



Conventional	Linguistic	Contextual
• One Hot	• POS	• Bi-LSTM
• Word Embeddings	• Named Entity	• BERT
• Sub-word, Character Embeddings	• Query Category	• ELMO

Question/Passage Representation



CNN

- CNN. for doc rep.
- Cross-attention
- …

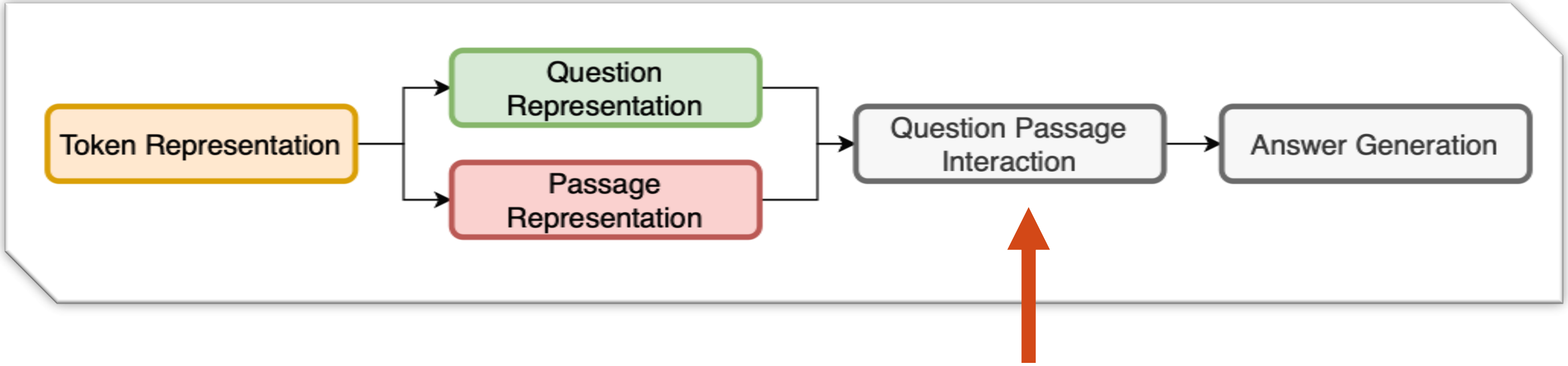
RNN

- LSTM
- Bi-LSTM
- Bi-GRU

Transformers

- GPT2
- BERT
- XLNET

Question And Passage Interaction



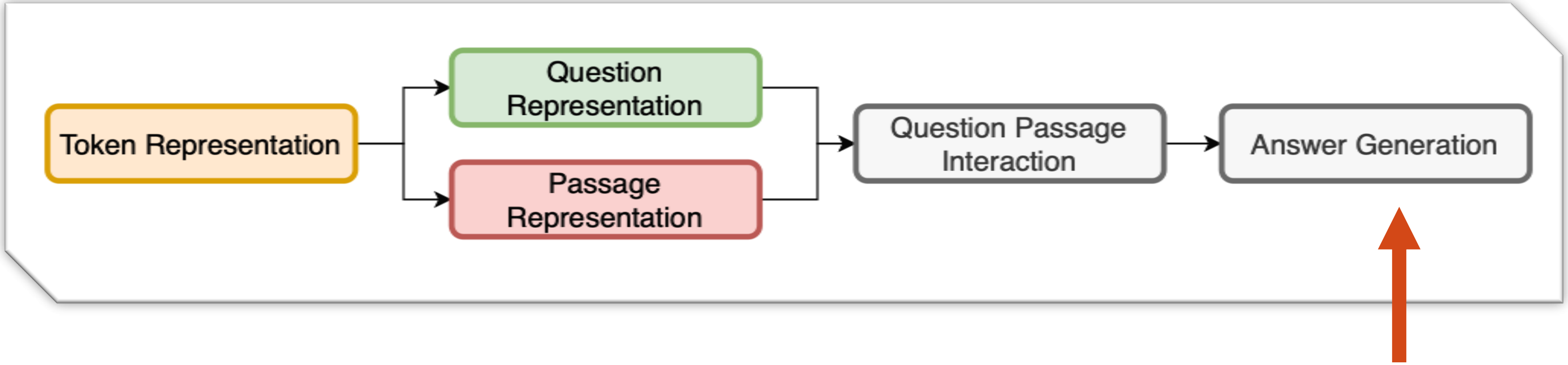
Attention

- Uni directional
- Bi-directional
- Cross-attention

Interaction Type

- Single Hop Interaction
- Multi hop Interaction

Answer Generation



Cloze	MCQ	Free Text	Span Pred.
• Word-level prediction (BC)	• Choosing one answer among many (MC)	• Generative models	• Predict begin and end of sequence

Attention Mechanism

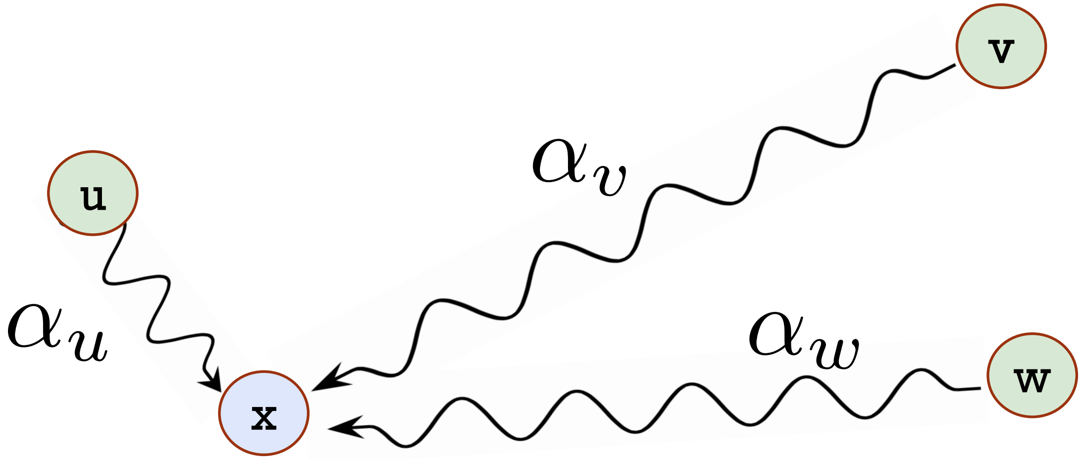
- Attention is used to represent tokens, question and passages
  - How do we re-represent otherwise independent token representations ?
  - How do we leverage contextualization ?

- Hard attention
- **Soft Attention**
- Co-attention
- Self-attention

Attention – Influence Point Of View

Attention encodes how much influence the context u has on x

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



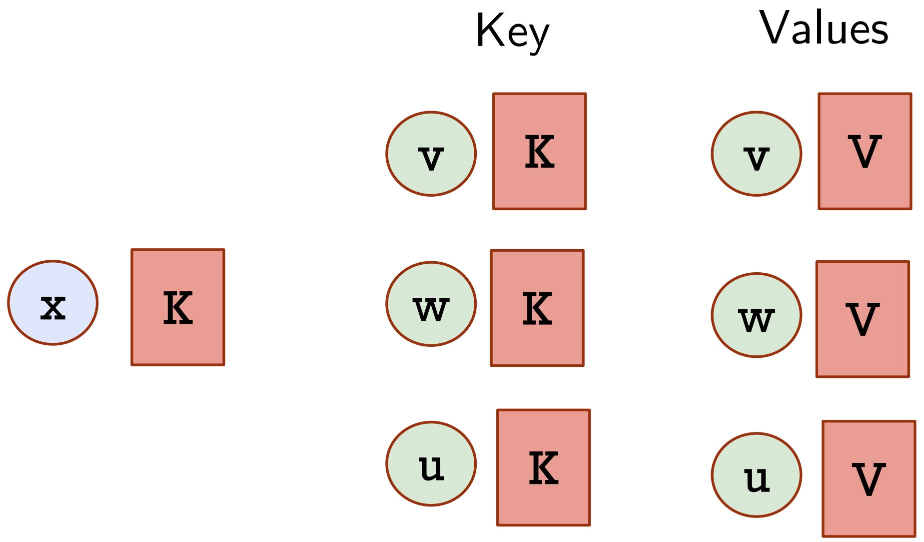
$$\mathbf{x}' = \alpha_u \mathbf{u} + \alpha_v \mathbf{v} + \alpha_w \mathbf{w}$$

- 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

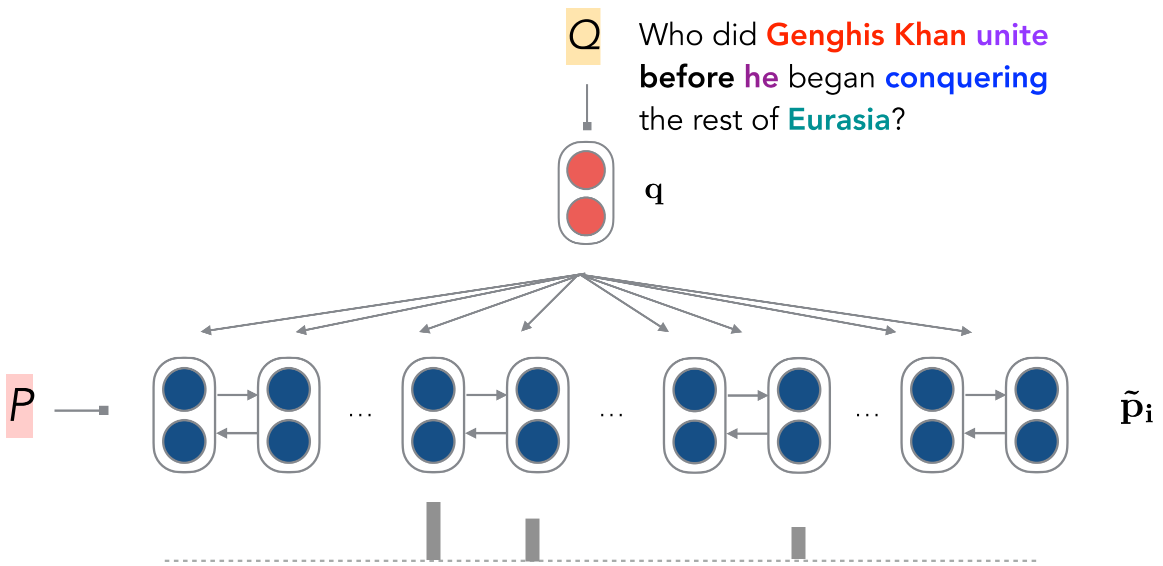
[Chen '16] Attentive reader

Tokens
- Word embeddings
- Exact matching features
- POS,NEF, TF features
- Alignment

Passage
- Bi-LSTM

Query-Passage Interaction
- Soft Attention

Answer Gen.
- Span Pred.

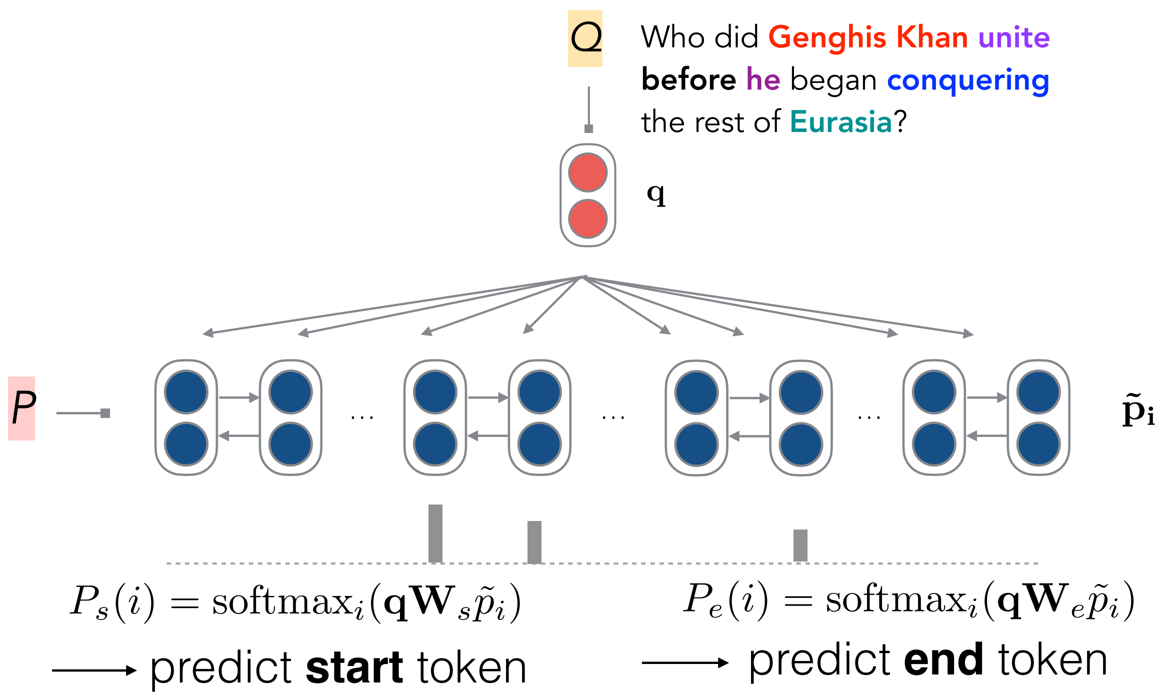


$$P_s(i) = \mathrm{softmax}_i(\mathbf{q}\mathbf{W}_s\tilde{p}_i)$$
→ predict **start** token

$$P_e(i) = \mathrm{softmax}_i(\mathbf{q}\mathbf{W}_e\tilde{p}_i)$$
→ predict **end** token

Attentive Reader

[Chen '16]

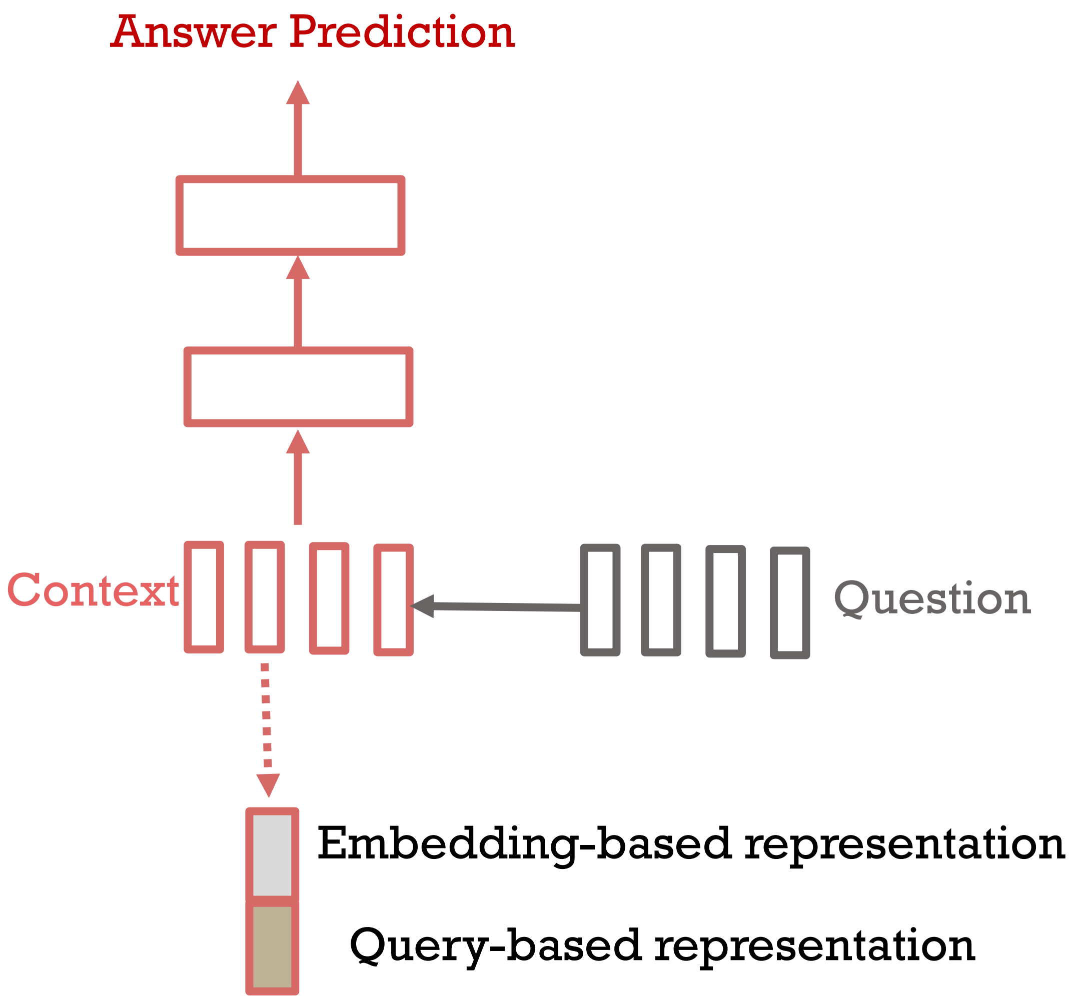


$$\Pr(a|q, p_i) = P_s(a_s)P_e(a_e)$$

Input Representation

DrQA [Chen '17]

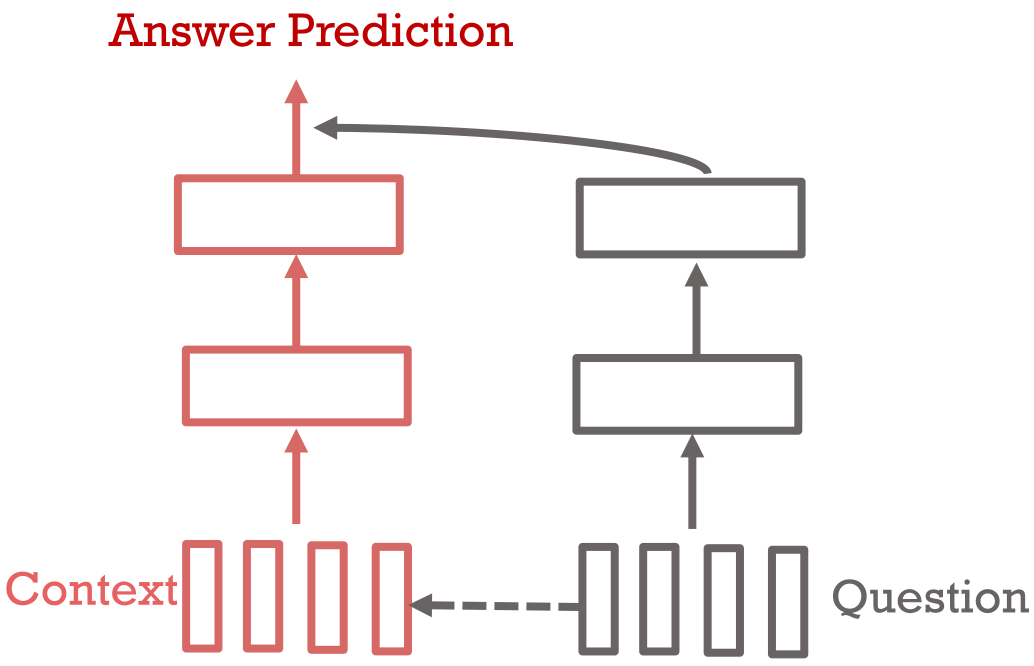
- Context words are represented based on similarity with the query
- Semantic similarity
  - Word embeddings
- Matching similarity
  - Direct word-level matching
  - Weighted matching
  - Attention mechanism



MatchLSTM [Wang & Jiang '16], DCN [Xiong '16], BiDAF [Seo '17]

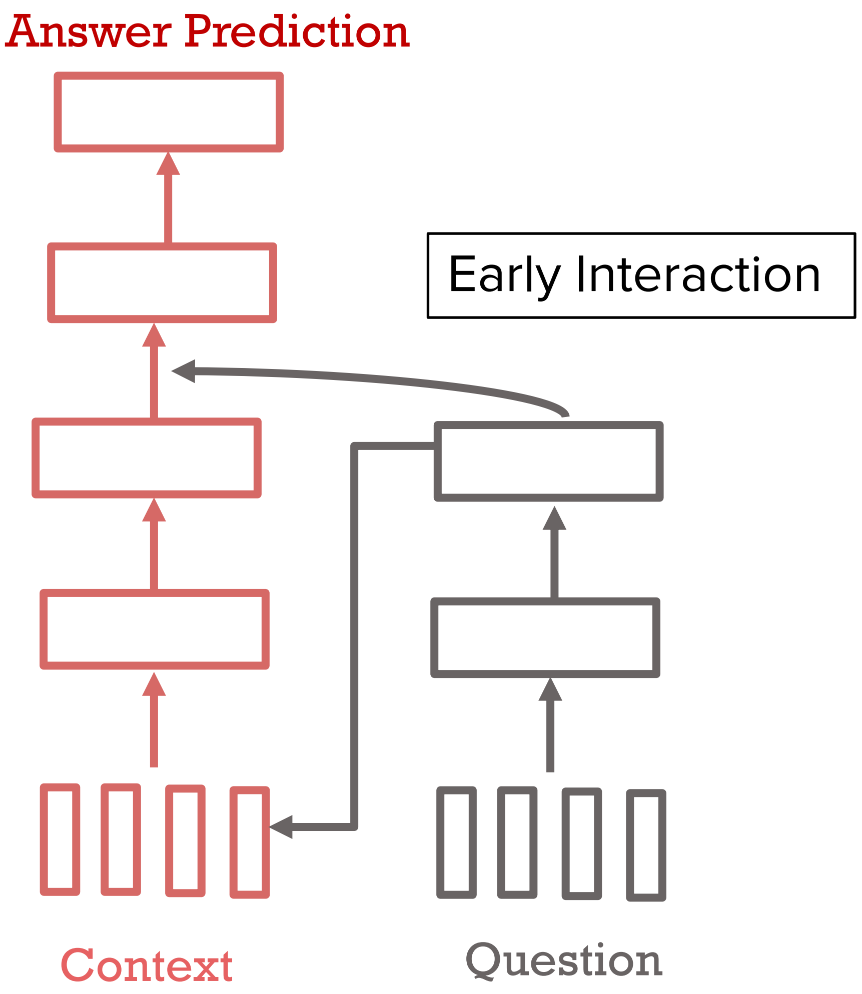
Late Interaction

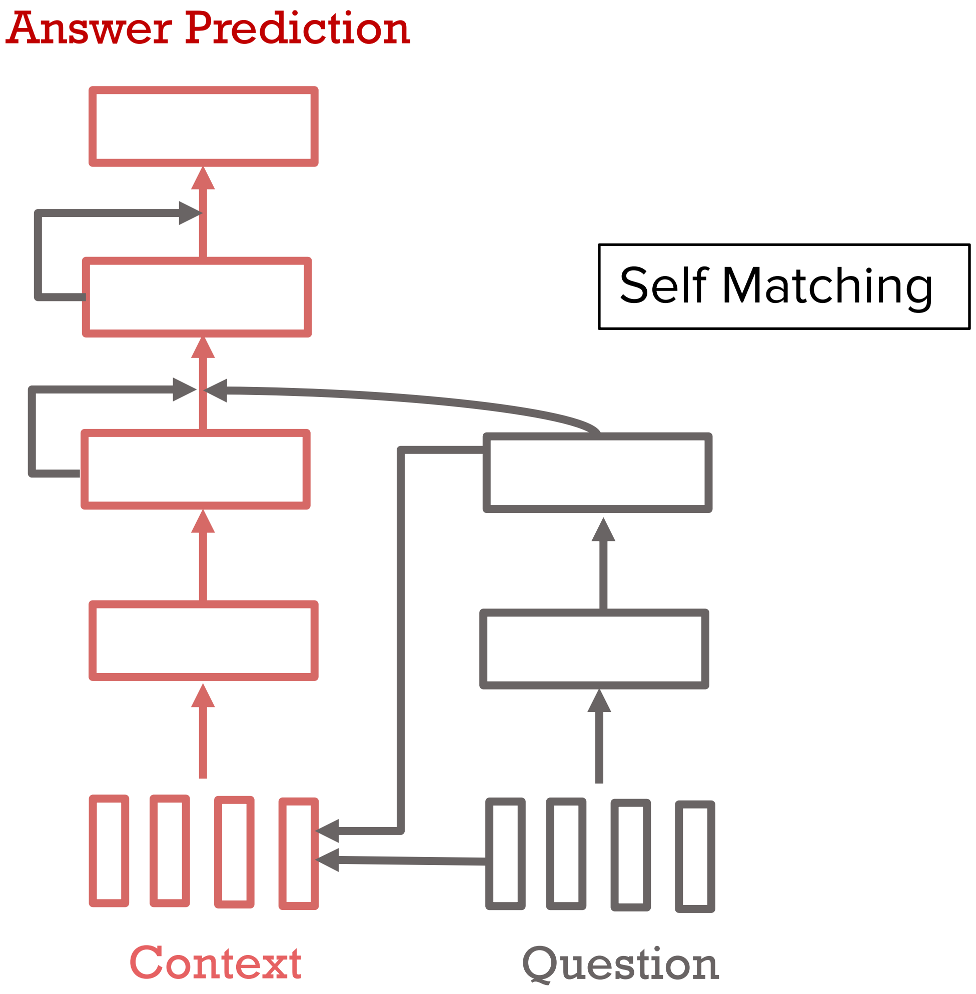
- First encode question and context sufficiently
- Choice of encoders
  - Bi-LSTMs
  - Conv Nets
- Most popular Model
  - Bi-directional attention flow [Seo '17]



Rasor [Lee '16], fastQA [Weissenborn'17]

Other Variants

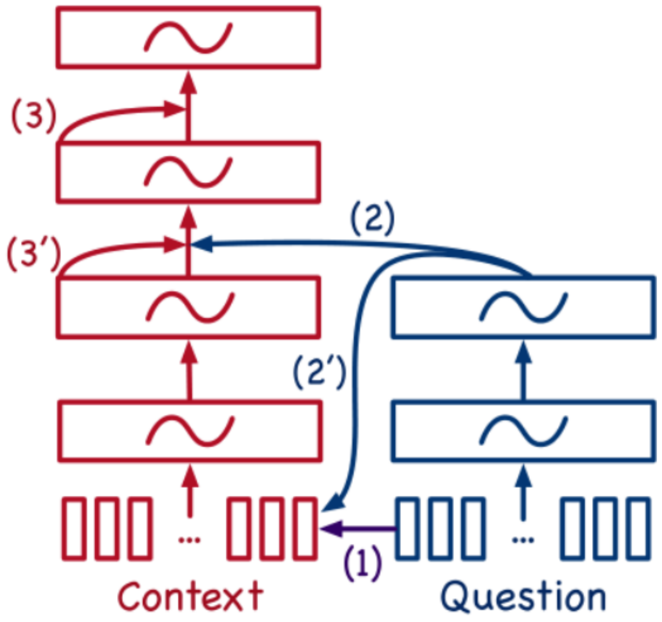




Attention Based Architectures

- 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)		✓		✓	

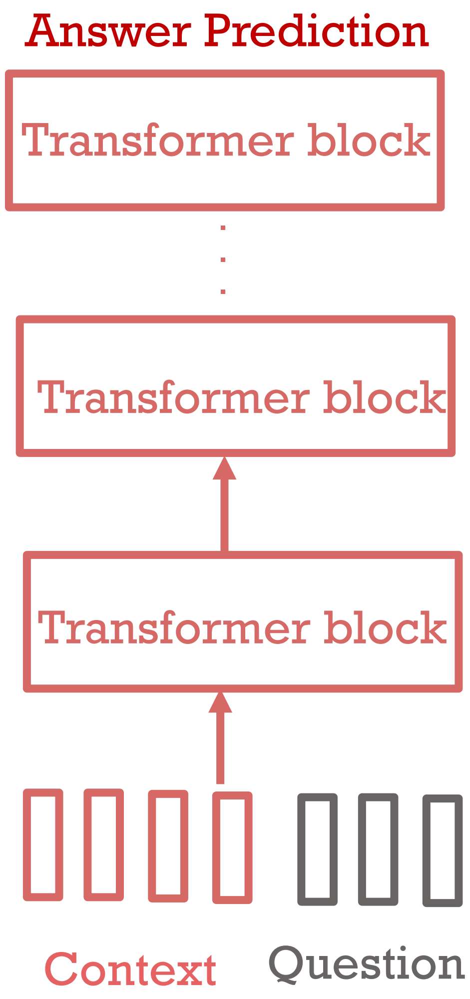


(1) Word-level fusion, (2) high-level fusion, (2') high-level fusion (alternative), (3) self-boosted fusion, and (3') self-boosted fusion (alternative).

[Devlin '18]

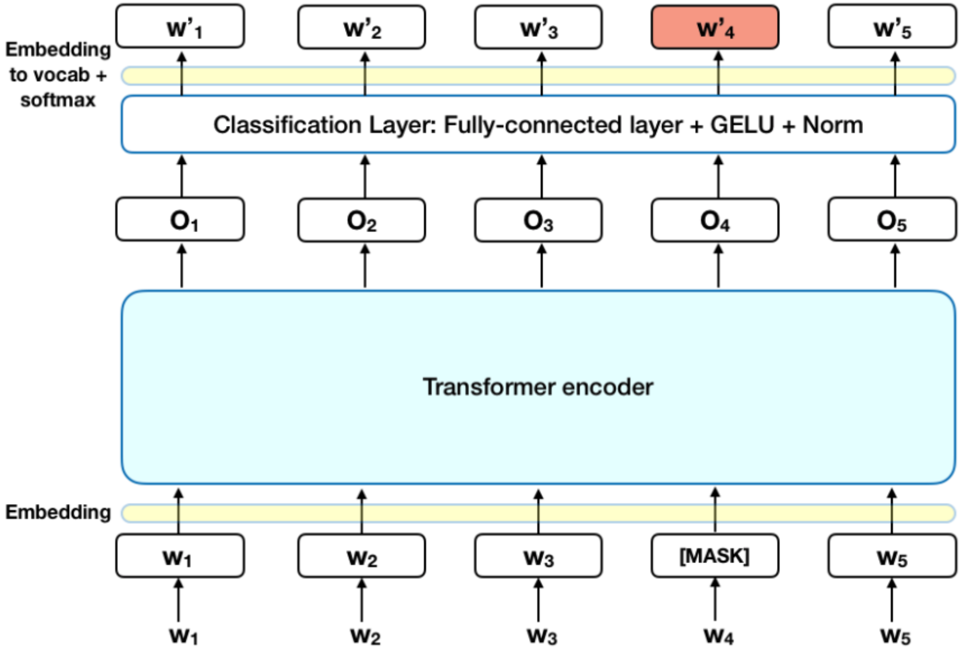
Contextual Language Models

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



BERT

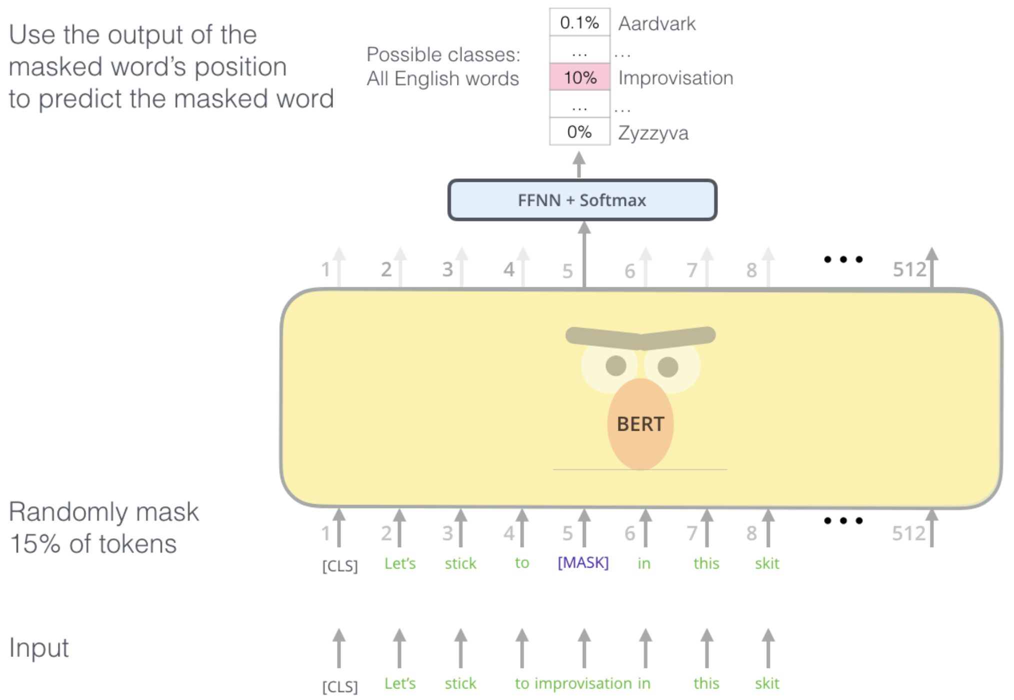
- Bi-directional : Transformer encoder reads the entire sequence of words at once.
  - Learns the context of a word based on all of its surroundings (left and right of the word).



BERT– Masked Language Model

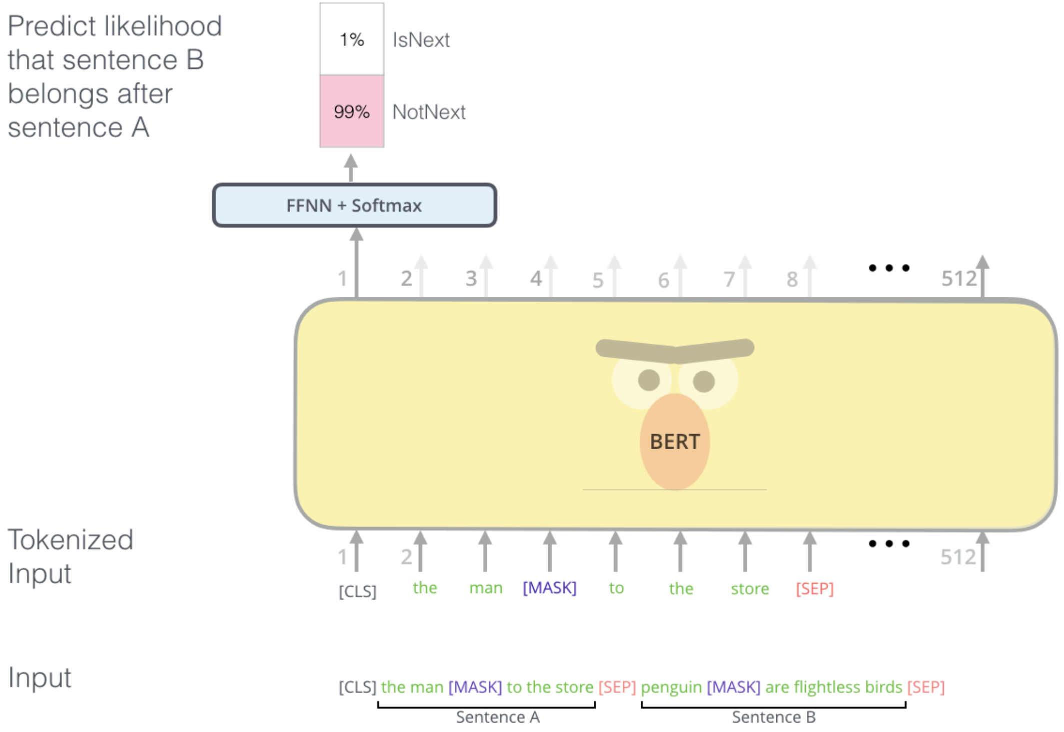
Masked word prediction

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



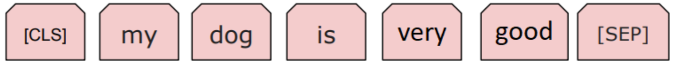
Next Sentence Prediction

- 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

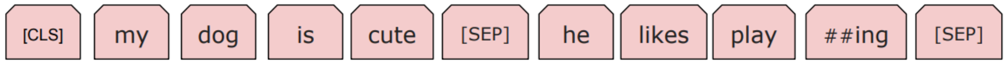


BERT Fine Tuning

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

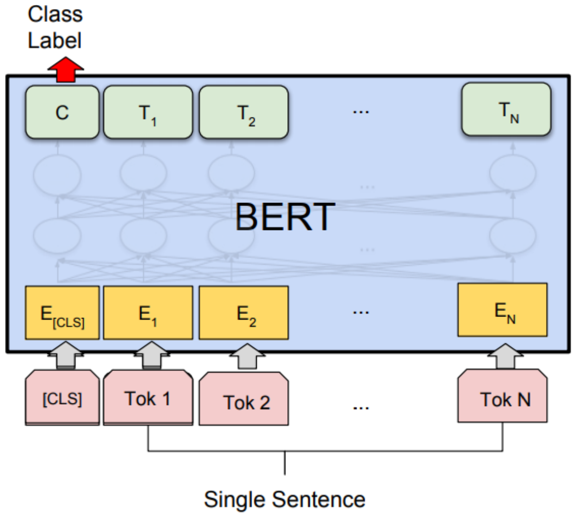


Single sentence input



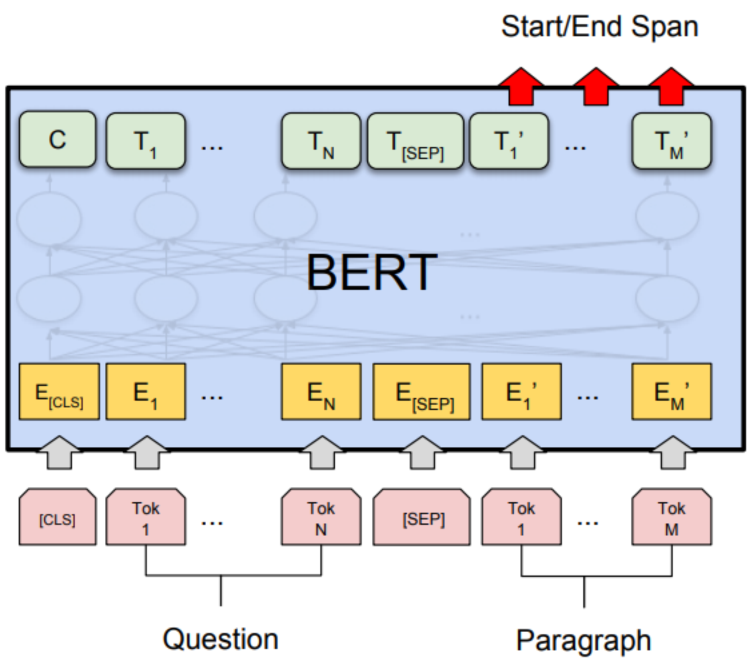
Single sentence input

- 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.



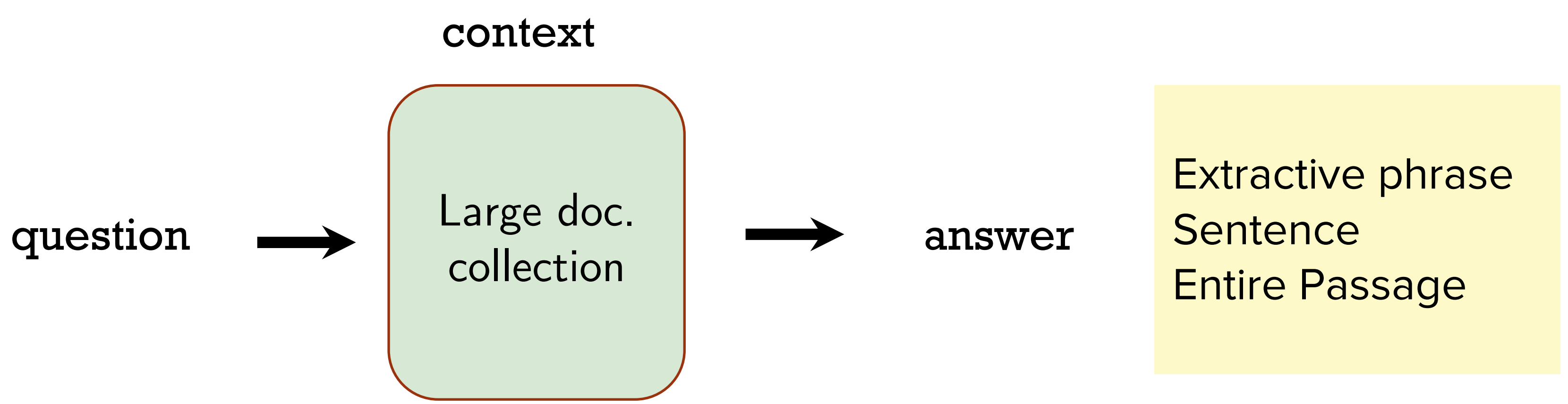
BERT Fine Tuning

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] — Trivia questions — Web pages from BING search
- SearchQA [Dunn et al., 2017] — Jeopardy — Google search snippets
- Quasar-T [Dhingra et al., 2017] — Reddit — ClueWeb09
- Natural Questions [Kwiatkowski et al., 2019] — Google queries — Wikipedia pages in results

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

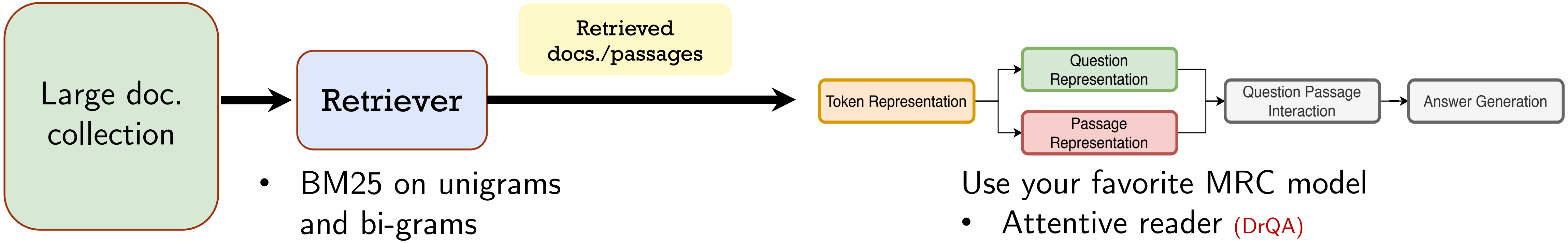
Repurposed for ODQA

- 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



- BM25 on unigrams and bi-grams
- Vector Index

Use your favorite MRC model

- Attentive reader (DrQA)
- BiDaf (Seo 17)
- BERT (Bertserini)

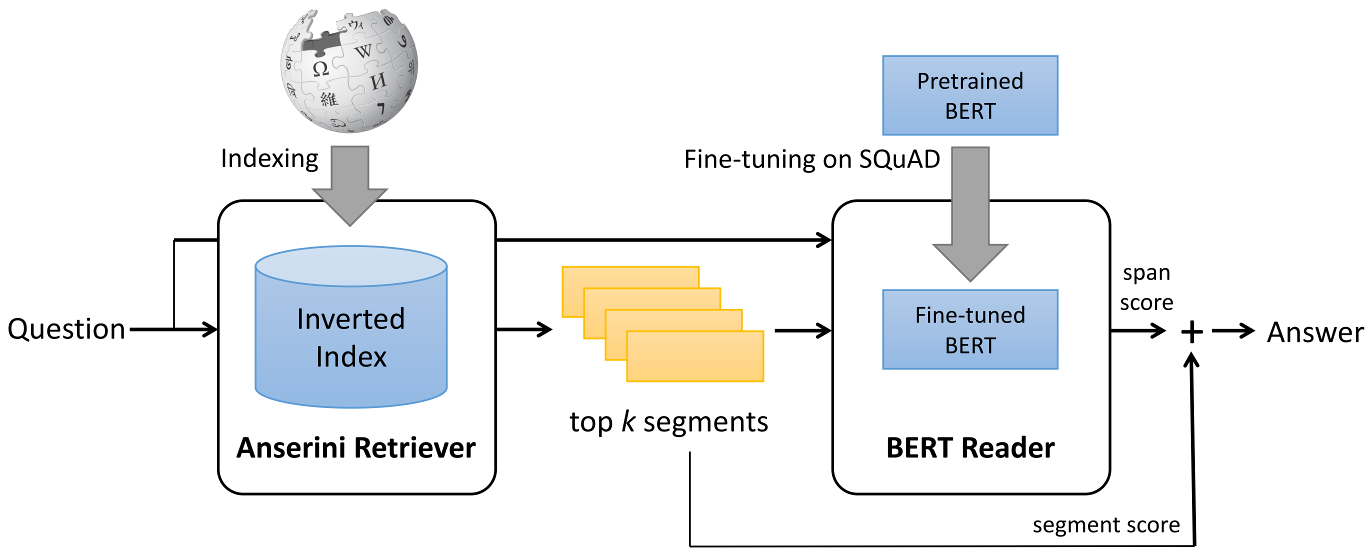
How is the reader model trained ?

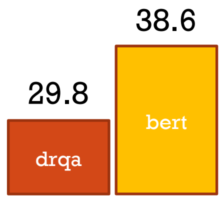
Using an existing QA dataset (e.g. SQUAD)

How does it answer questions ?

Independently find answers for tok-k passage and return the most "probable" span

BERTserini





Squad (EM)

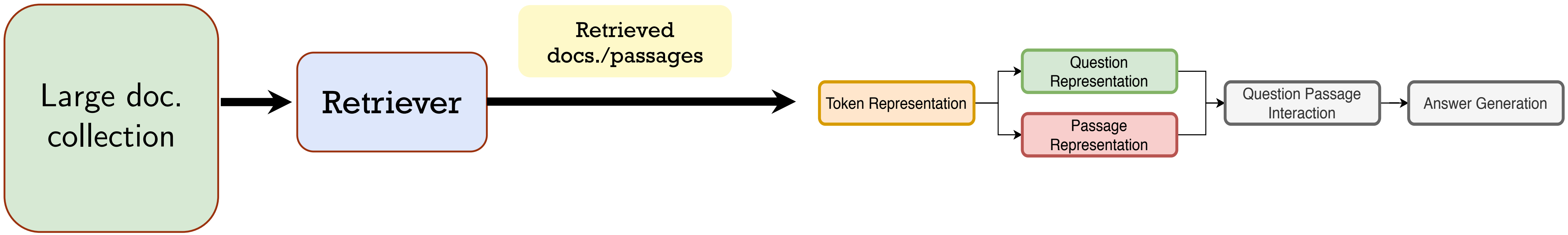
Retriever

- 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)

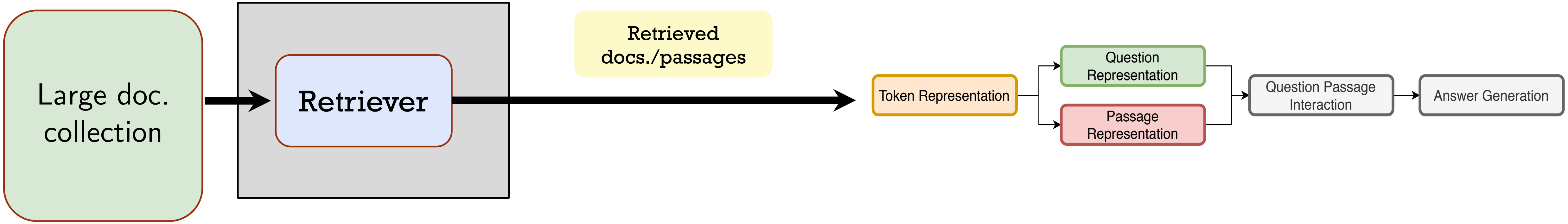
Design Questions



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

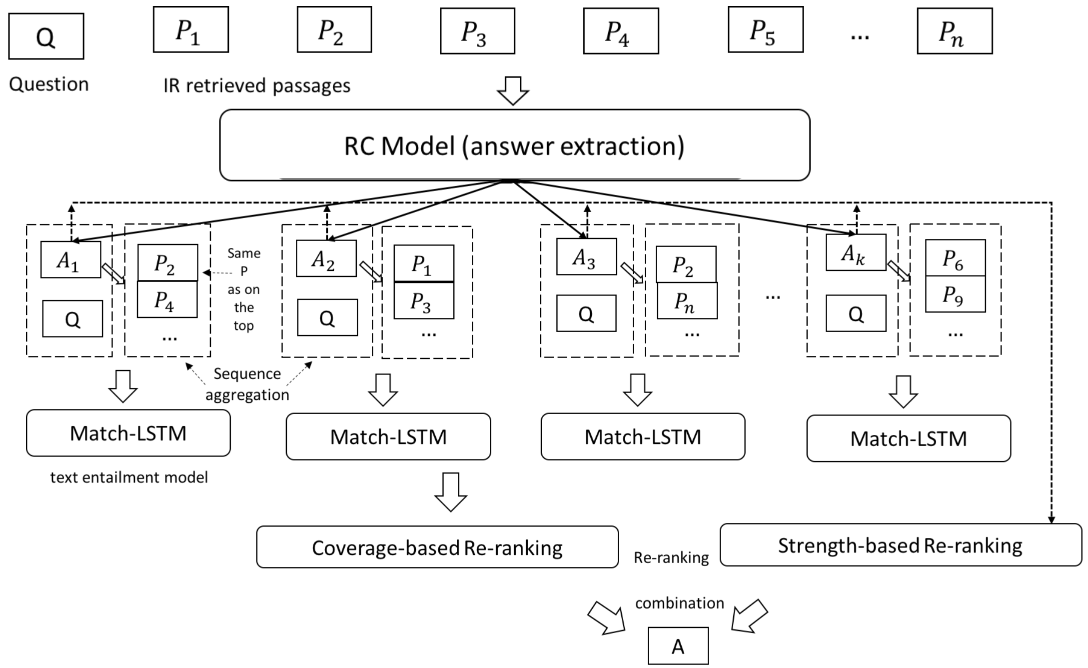
A1: Isaac Newton
P1: **Sir Isaac Newton** was an English physicist , mathematician , astronomer , natural philosopher , alchemist and theologian ...
P2: **Sir Isaac Newton** was an English mathematician, astronomer, and physicist who is widely recognized as one of the most influential scientists ...

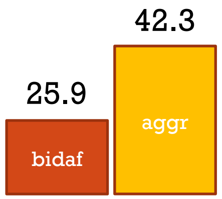
A2: Galileo Galilei
P1: **Galileo Galilei** was an Italian physicist , mathematician , astronomer , and philosopher who played a major role in the Scientific Revolution .
P2: **Galileo Galilei** is credited with discovering the first four moons of Jupiter .

Support And Coverage

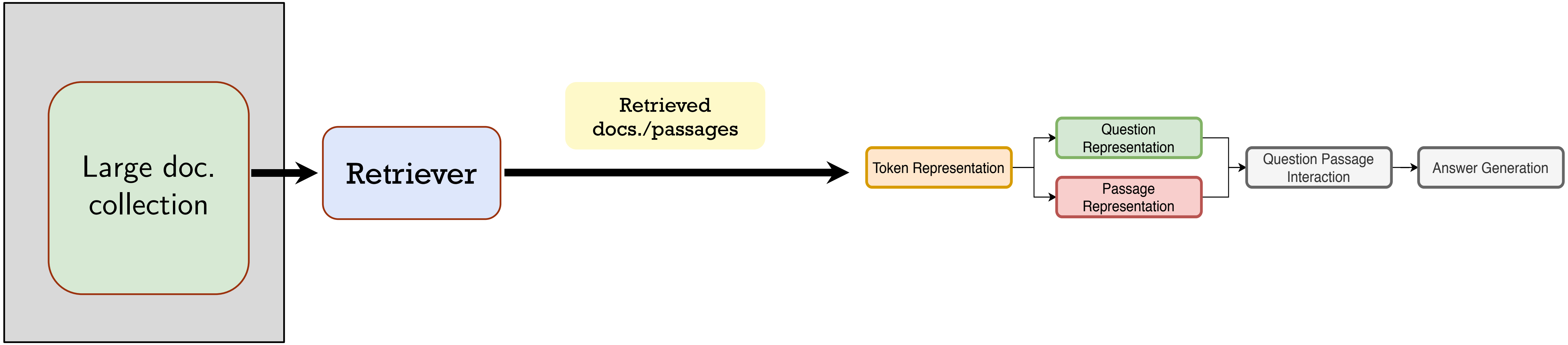
[Wang et al.' 18]

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





Quasar (EM)

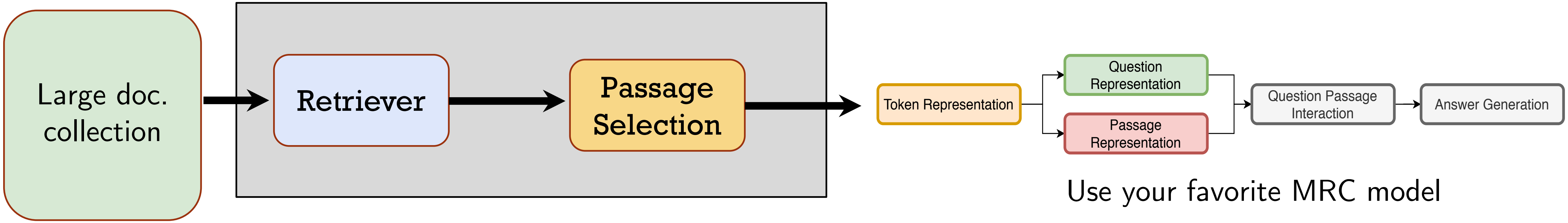


How do we Exploit Evidence from the collection ?

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



Use your favorite MRC model

BM25 on unigrams and bi-grams

- 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]

Likelihood of the passage containing the answer

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

Likelihood of the answer given a cand. passage

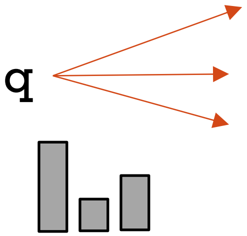
Selecting Passages

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

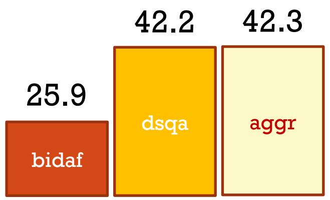
$$\Pr(a|q, P) = \sum_{p_i \in P} \boxed{\Pr(a|q, p_i)} \boxed{\Pr(p_i|q, P)}$$

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

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

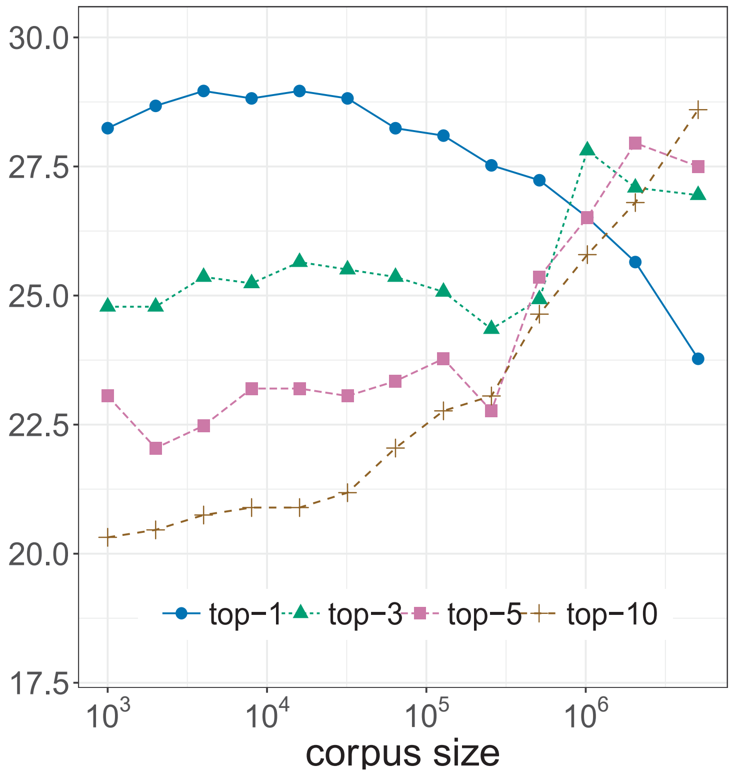
Answer Selection

1. Detect spans for each passage
2. Multiple answers possible in a passage
3. Use the same rep. space for passage sel. and answer sel.

Passage Selection

1. Compute representations for query and passage independently
2. Compute relevance of passage to the query
3. Relevance is used as weights later

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

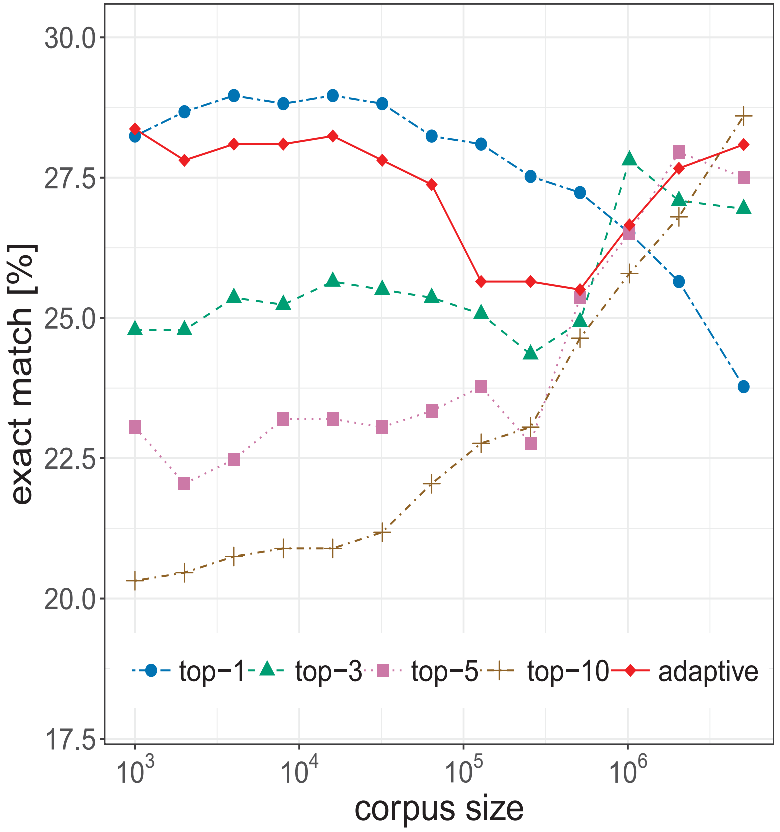
$$s_i = \left[s_i^{(1)}, \ldots, s_i^{(\tau)}\right]^T$$

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

Choose passages until surpassing a certain confidence threshold

- if document retrieval is certain → selects fewer docs/passages
- If uncertain → retrieval depth is higher

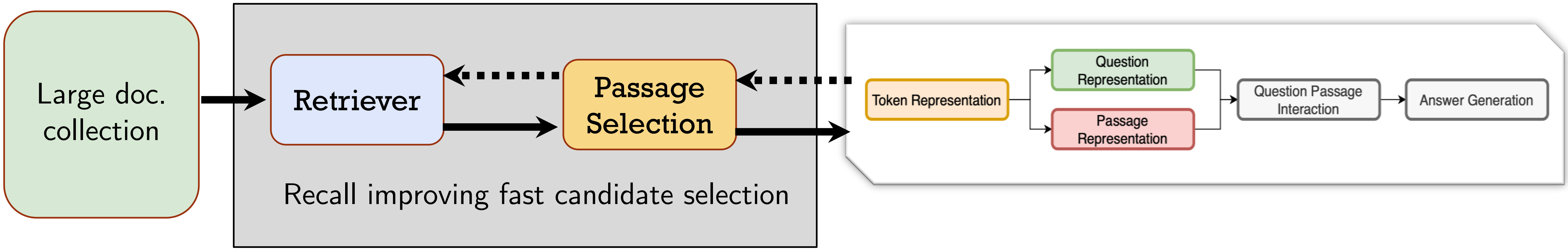
Retrieval Depth and Collection Size



Slightly more involved depth prediction

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

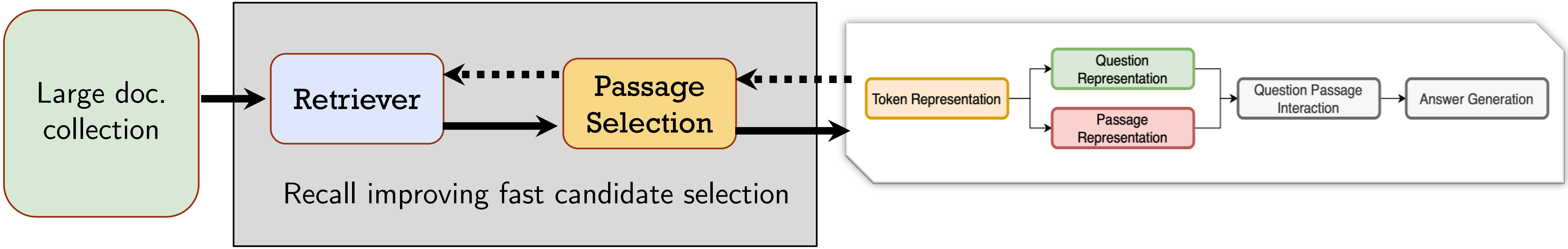
$$\underbrace{n_i}_{\text{Ret. depth}} = \left\lceil s_i^T \underbrace{\beta}_{\text{Learnable param.}} \right\rceil + \underbrace{b}_{\text{tolerance}}$$



How we exploit reader feedback for better retrieval ?

[Das et al. ICLR '19]

Retriever Reader Interaction

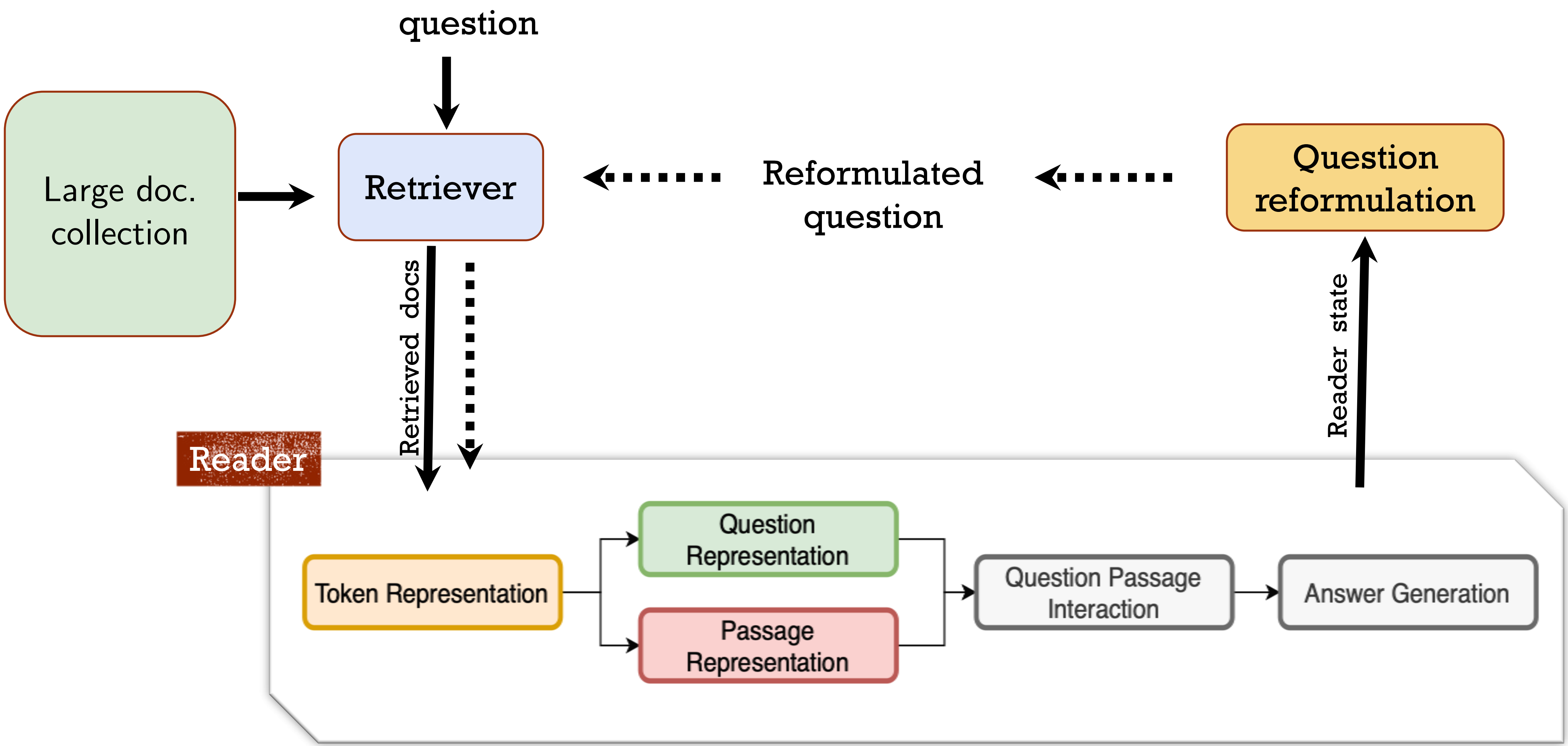


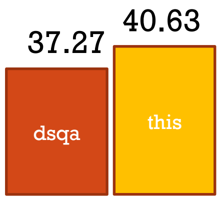
- Single retrieve and read step is limiting – vocabulary gap between question and corpus passages

- How can we enable multi-stage retriever-reader interaction ?
  - Akin to Neural Query Expansion
  - Take care about efficiency concerns

[Das et al. ICLR '19]

Retriever Reader Interaction

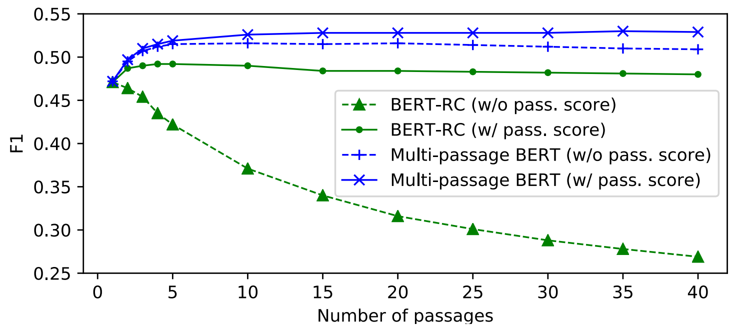




Quasar (EM)

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

- LIME
- SHAP
- Instance-wise feature selection

Model introspective

- Simple gradients
- LRP
- Integrated gradients
- Analyzing attention

By Design

Simple Models

- Linear models
- Templated models
- Matching models
- Analytical sol. possible

Neural Models

- Coarse-to-fine models
- Generate explanations and task prediction jointly

[Choi '17]

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.

[Choi '17]

Select Sentences as Explanations

What is the capital of Australia ?

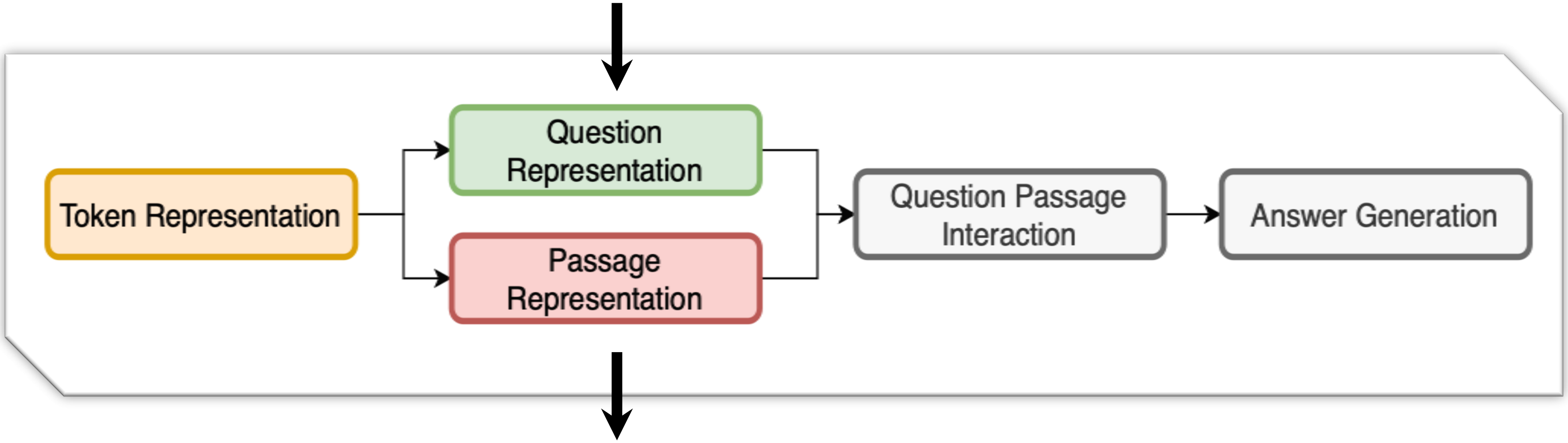
The country's other major metropolitan areas are Melbourne, Brisbane, Perth, and Adelaide. As the seat of the government of Australia, Canberra is home to many important institutions of the federal government, national monuments and museums. Canberra is also the capital of the country.

Input to Reader

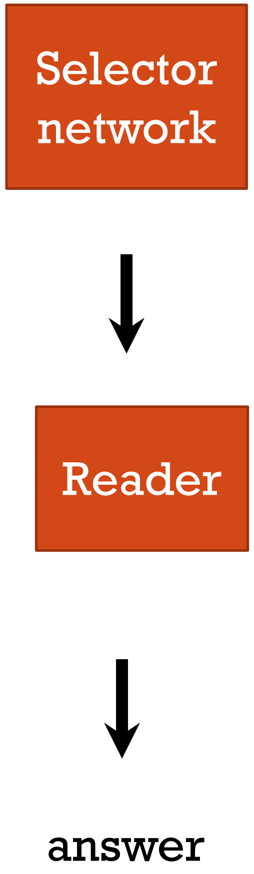
[Choi '17]

What is the capital of Australia ?

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



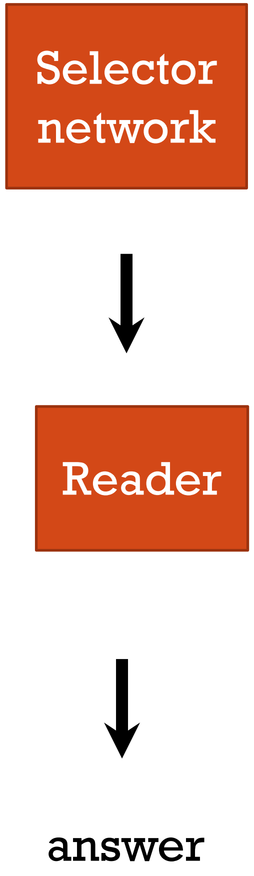
Canberra



[Choi '17]

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

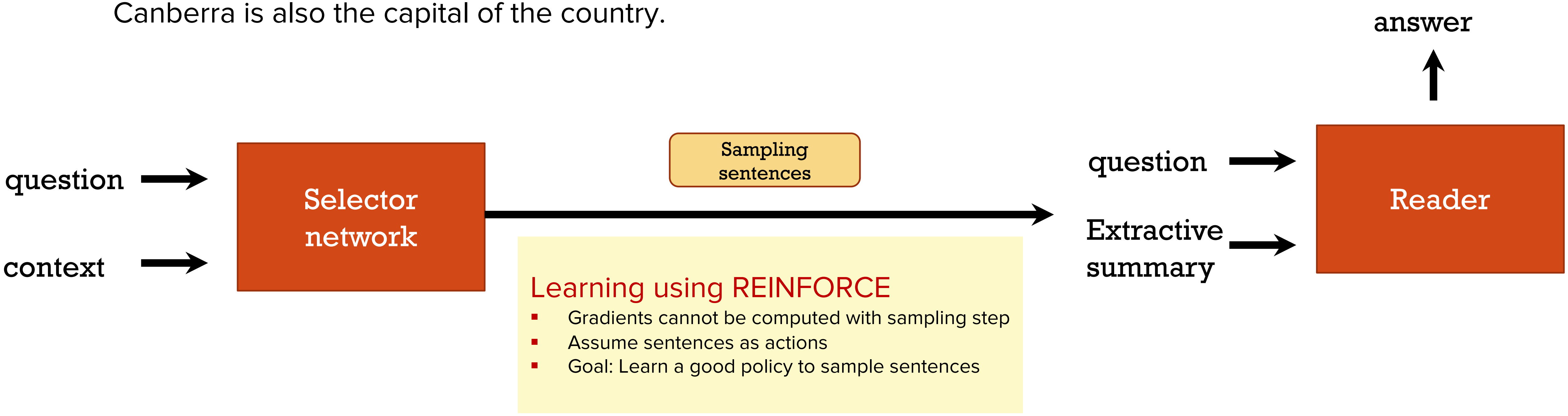


End-to-end Models

[Choi '17]

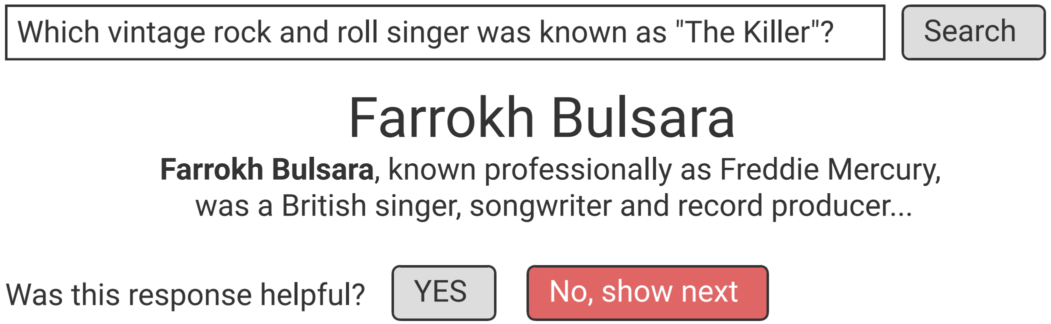
What is the capital of Australia ?

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.



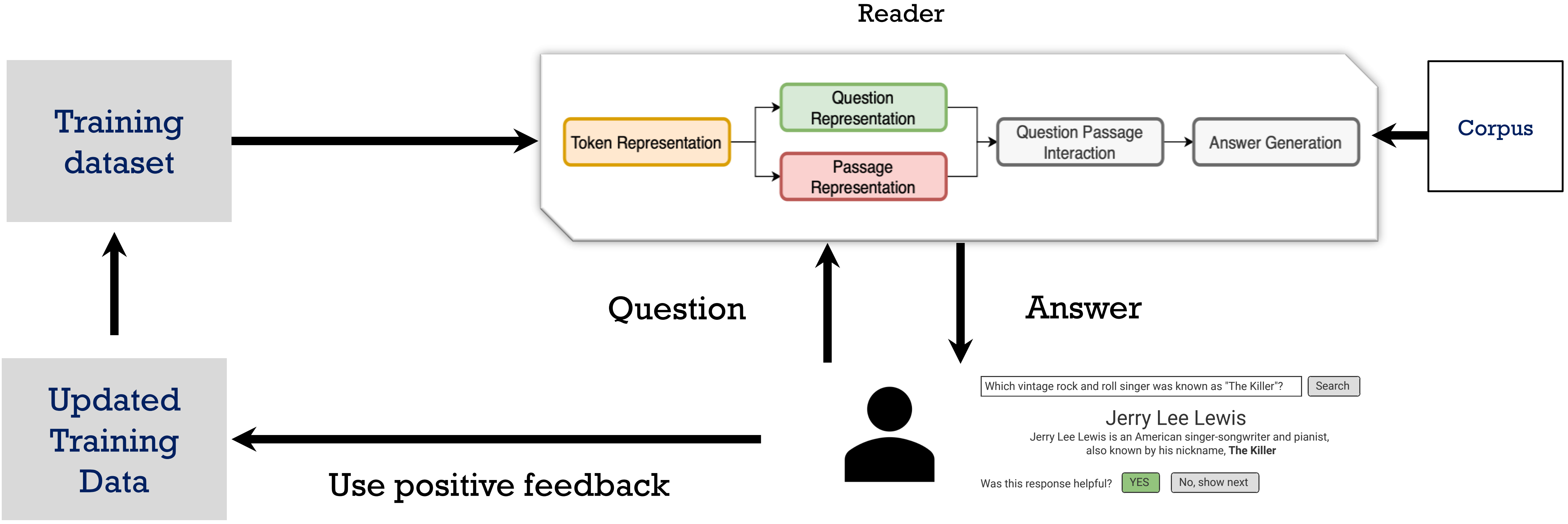
User Feedback

- 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 ?



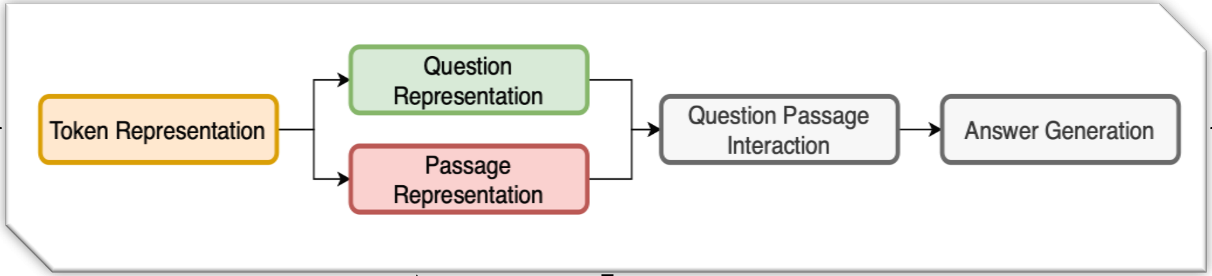
[Kratzwald & Feuerriegel, WebConf '19]

Updating Training Set



Credibility Validation

Reader



Training dataset

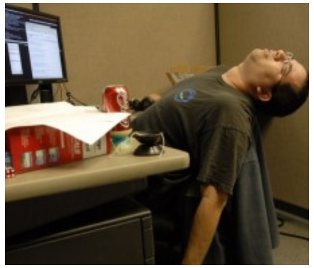
Corpus

Question

Answer

Updated Training Data

Evaluate 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

Credibility Validation

- Use question and answer using restrictive semantics to find support documents
- Supported by reader – top-k reader docs
- Length restrictions, NE restrictions

CONVERSATIONAL QA

[Reddy et al.'18]

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)

[Reddy et al. ‘18]

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

[Reddy '18]

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 moneyman, 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_1: What are the candidates running for?

Q_2: Where?

Q_3: Who is the democratic candidate?

Q_4: Who is his opponent?

Q_5: What party does he belong to?

Q_6: Which of them is winning?

A_1: Governor, R_1: The Virginia governor's race

A_2: Virginia, R_2: The Virginia governor's race

A_3: Terry McAuliffe, R_3: Democrat Terry McAuliffe

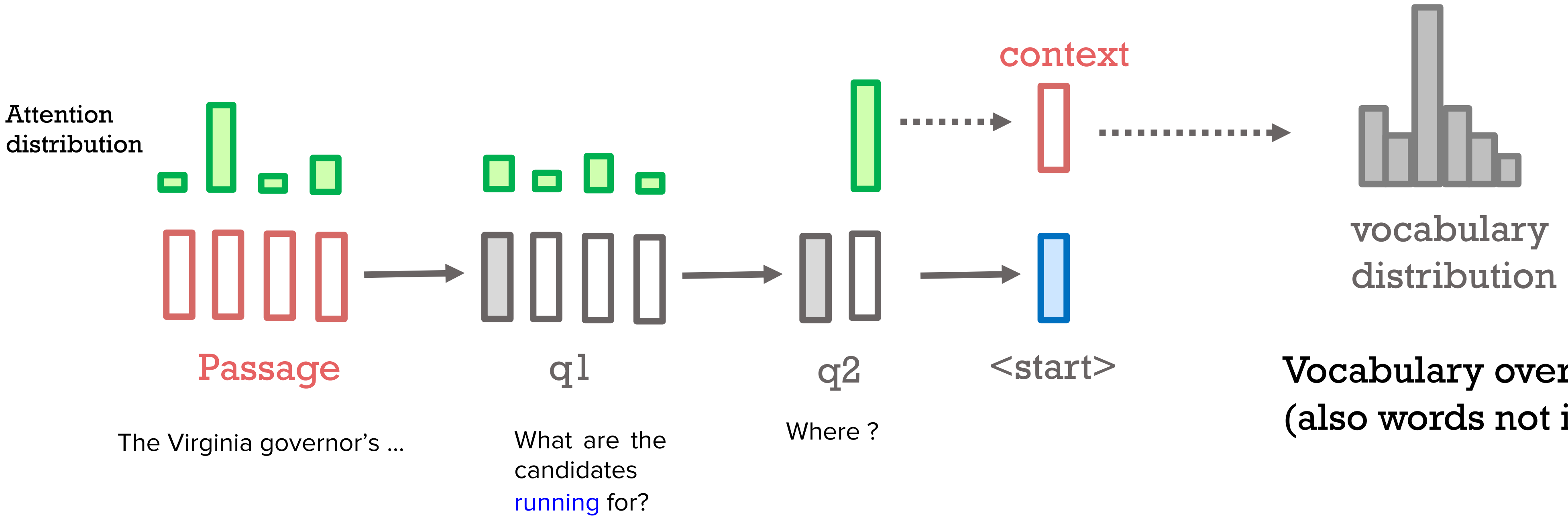
A_4: Ken Cuccinelli , R_4 Republican Ken Cuccinelli

A_5: Republican, R_5: Republican Ken Cuccinelli

A_6: Terry McAuliffe, R_6: Democrat Terry McAuliffe, the longtime political fixer and moneyman, hasn't trailed in a poll since May

[Reddy et al. '18]

Seq2Seq Abstractive Response Generation

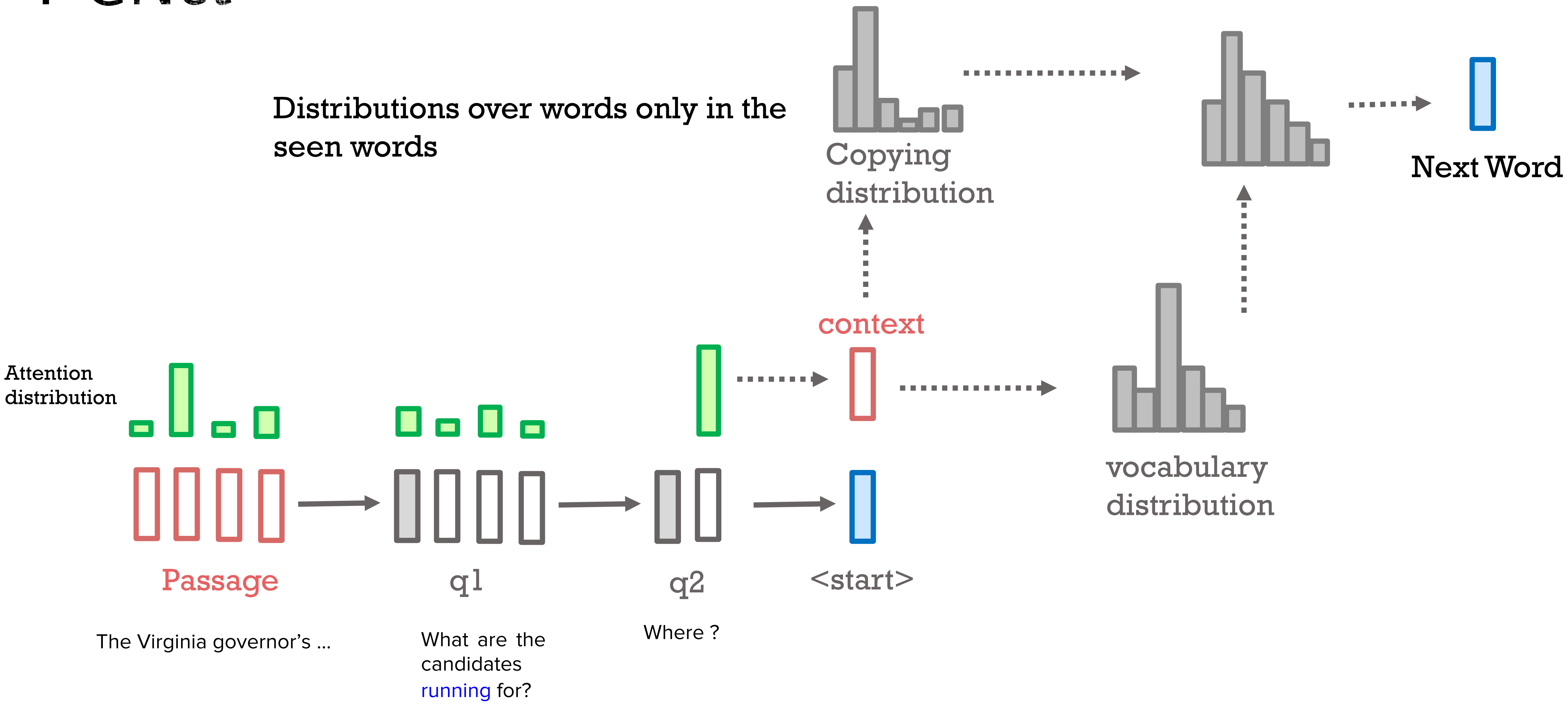


Vocabulary over a fixed set of words (also words not in the sentence)

PGNet

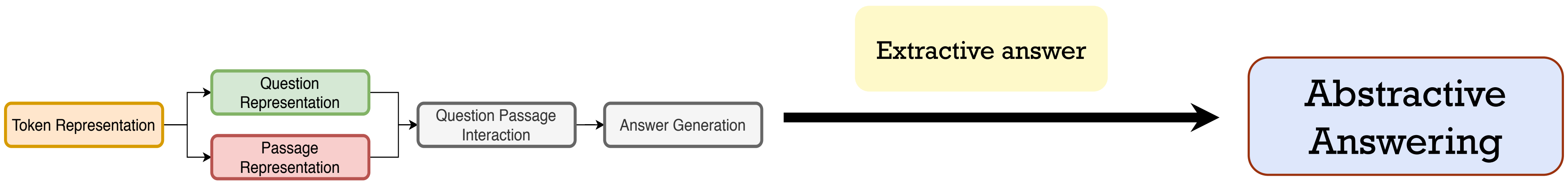
[Reddy et al.'18]

Distributions over words only in the seen words



[Reddy et al. '18]

Hybrid Model



Type	Seq2seq	PGNet	DrQA	Augmt. DrQA	DrQA+ PGNet	Human
			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

[Reddy et al. '18]

Leaderboard Sneakpeak

Rank	Model	In-domain	Out-of-domain	Overall
	Human Performance *Stanford University* (Reddy & Chen et al. TACL '19)	89.4	87.4	88.8
1 Sep 05, 2019	RoBERTa + AT + KD (ensemble) *Zhuiyi Technology* https://arxiv.org/abs/1909.10772	91.4	**89.2**	**90.7**
1 Apr 22, 2020	TR-MT (ensemble) *WeChatAI*	**91.5**	88.8	**90.7**
2 Sep 05, 2019	RoBERTa + AT + KD (single model) *Zhuiyi Technology* https://arxiv.org/abs/1909.10772	90.9	**89.2**	90.4
3 Jan 01, 2020	TR-MT (ensemble) *WeChatAI*	91.1	87.9	90.2
4 Mar 29, 2019	Google SQuAD 2.0 + MMFT (ensemble) *MSRA + SDRG*	89.9	88.0	89.4
5 Dec 18, 2019	TR-MT (single model) *WeChatAI*	90.4	86.8	89.3
6 Sep 13, 2019	XLNet + Augmentation (single model) *Xiaoming* https://github.com/stevezheng23/xlnet_extension_tf	89.9	86.9	89.0

39 Aug 21, 2018	DrQA + seq2seq with copy attention (single model) *Stanford University* https://arxiv.org/abs/1808.07042	67.0	60.4	65.1

[Reddy et al.‘18]

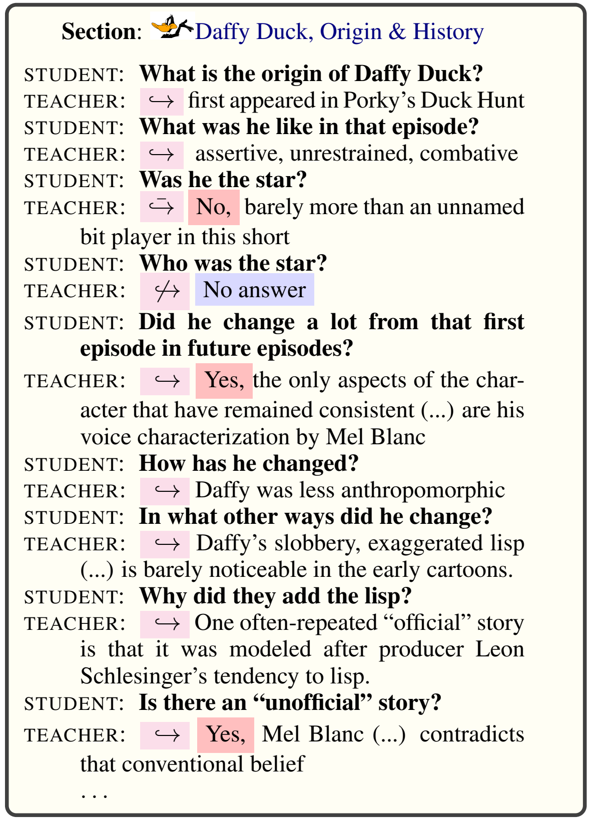
Other Conversational Datasets

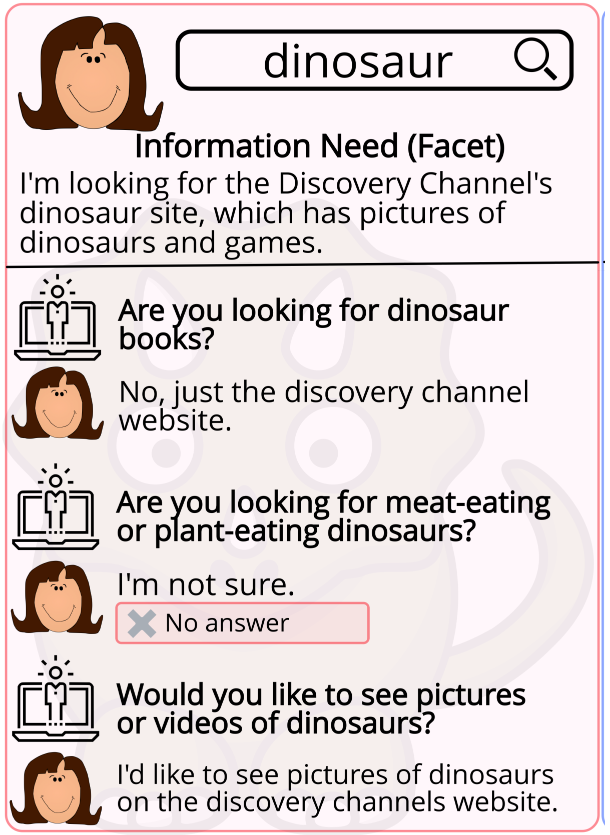
QuAC [Choi ’19]

- 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]





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,yuxche,gordonl,fdiaz,pauben,nickcr,sdumais}@microsoft.com

Open-Retrieval Conversational Question Answering

Chen Qu[1] Liu Yang[1] Cen Chen[2] Minghui Qiu[3] W. Bruce Croft[1] Mohit Iyyer[1]
[1] University of Massachusetts Amherst [2] Ant Financial [3] 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[1] Dan Li[1] Pengjie Ren[1] Evangelos Kanoulas[1] Maarten de Rijke[1, 2]
[1]University of Amsterdam, Amsterdam, The Netherlands [2]Ahold Delhaize, Zaandam, The Netherlands
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

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 Red eld, 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 Gard-ner. 2019. Quoref: A reading comprehension dataset with questions requiring coreferential reasoning. In EMNLP-IJCNLP.

[18] EunsolChoi,HeHe,MohitIyyer,MarkYatskar,Wen-tauYih,YejinChoi,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] PavelBraslavski,DenisSavenkov,EugeneAgichtein,andAlinaDubatovka.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.

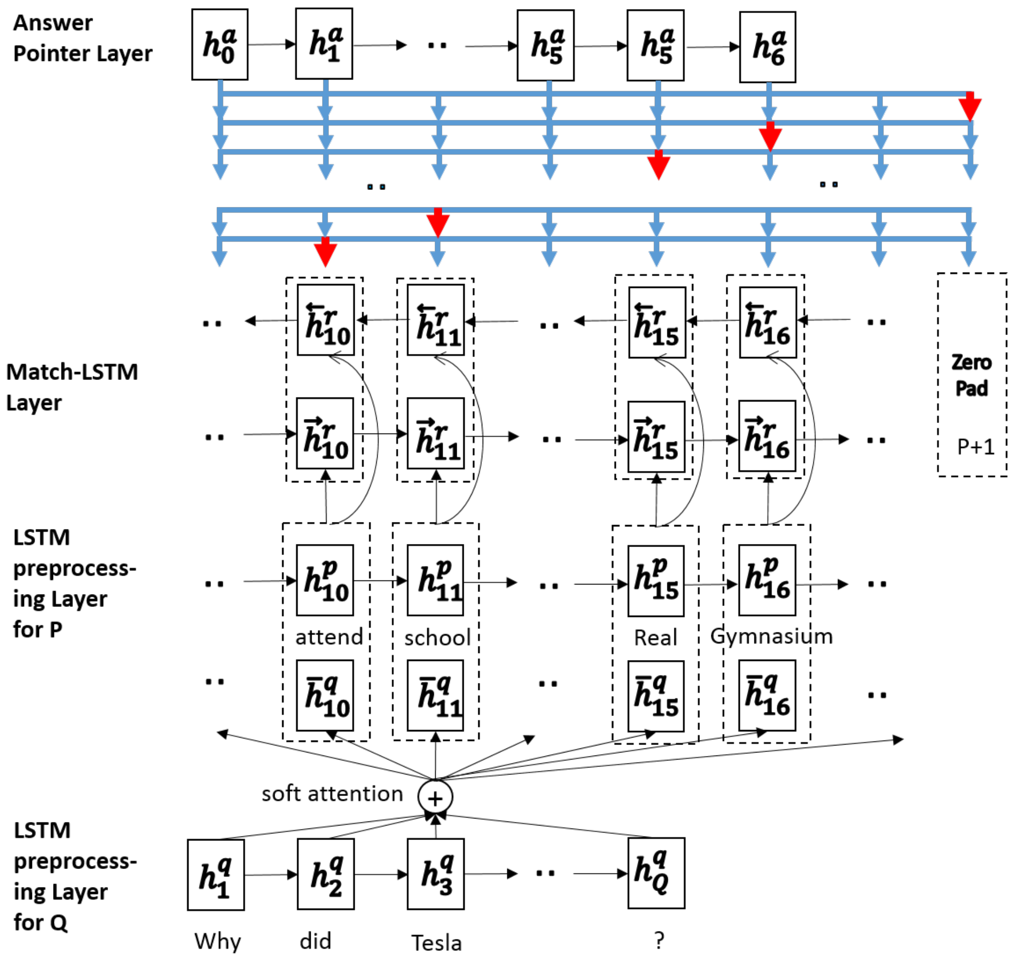
[67] Marzieh Saeidi, Max Bartolo, Patrick Lewis, Sameer Singh, Tim Rocktäschel, Mike Sheldon, Guillaume Bouchard, and Sebastian Riedel. 2018.
Interpretation of Natural Language Rules in Conversational Machine Reading. In EMNLP.

[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.

ARCHITECTURES [BACKUP..]

Match-LSTM And Pointer Nets



(a) Sequence Model

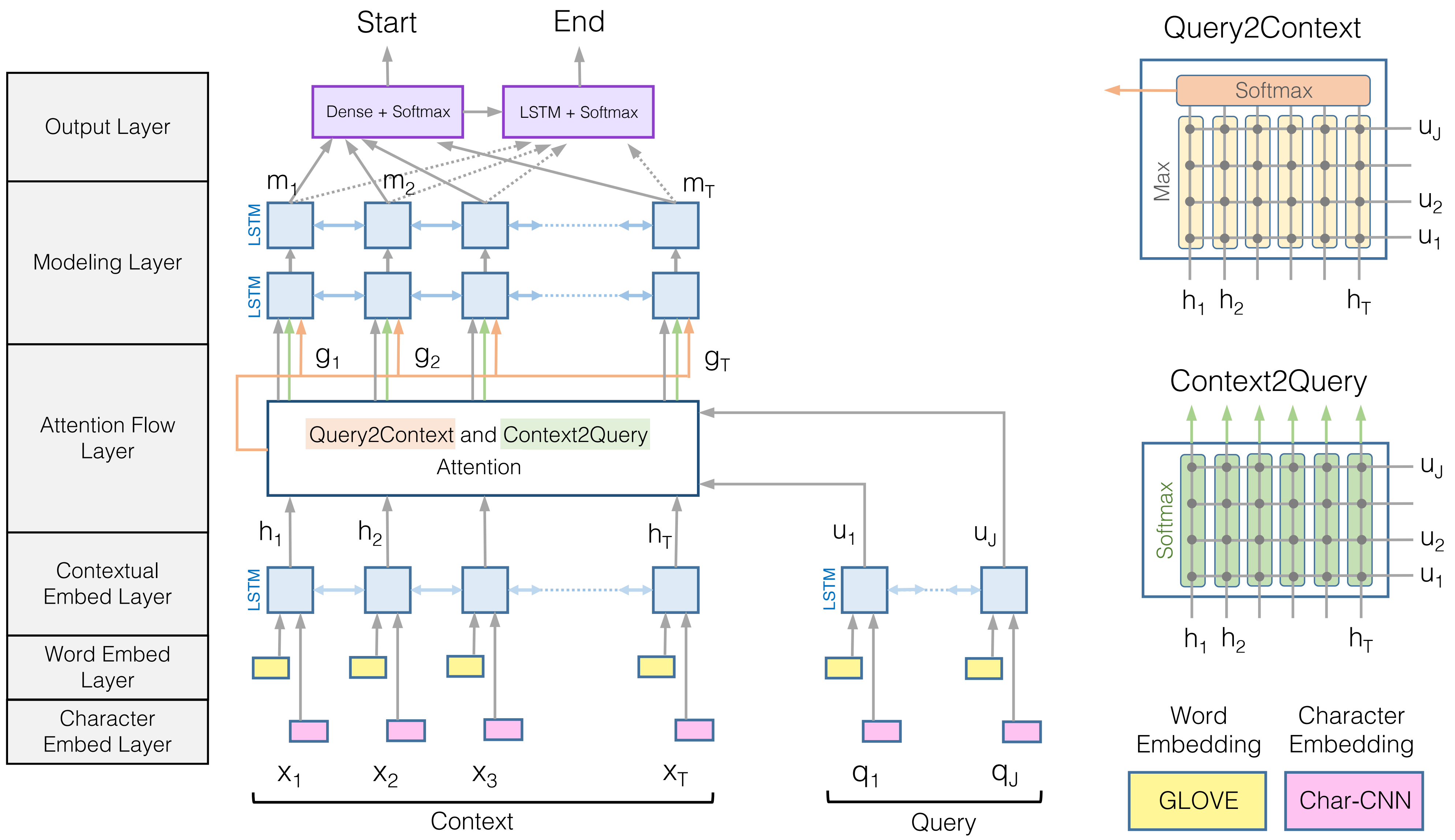
- Originally proposed for entailment

- Get a query representation - q
- Get a passage representation – p conditioned by a query representation (soft-attention)

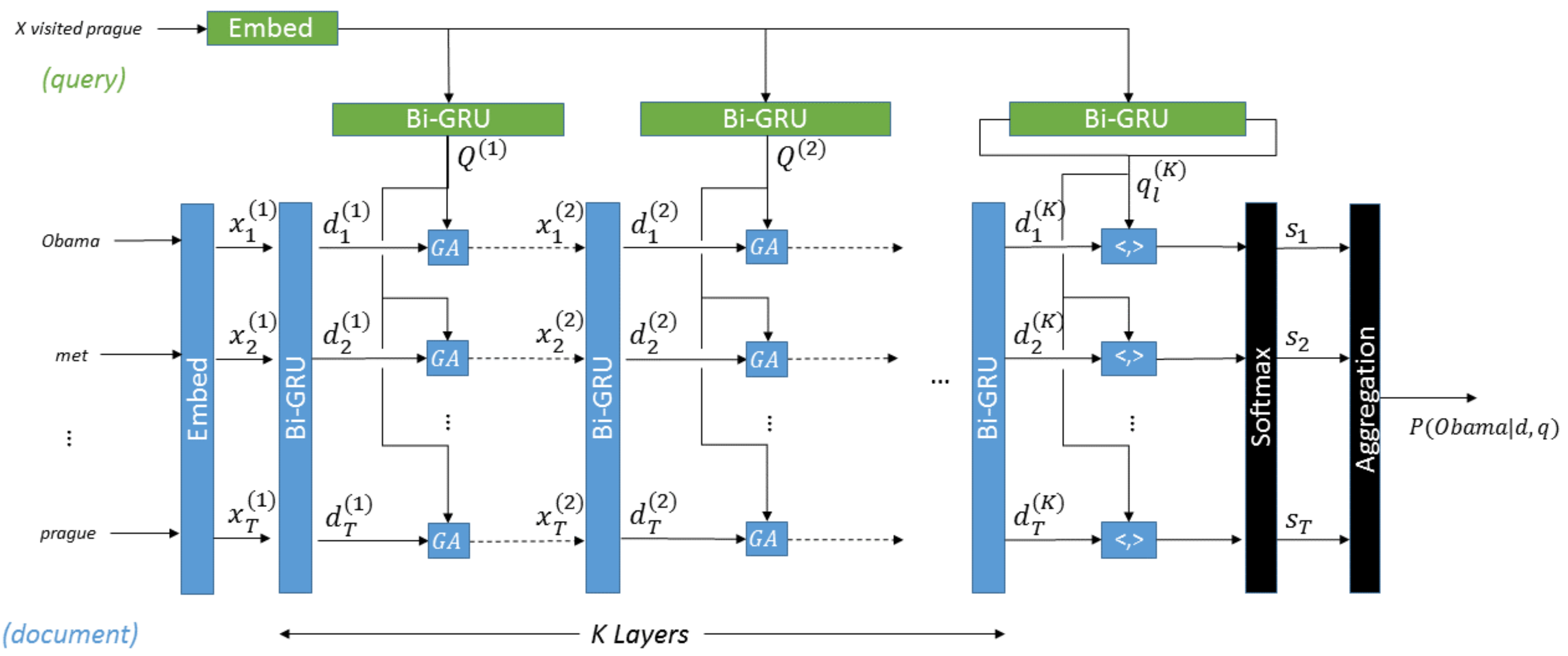
- Pointer Net: Select tokens from p
  - EOS is an explicit marker
  - Ptr Net gets the p(.) over the input sequence

- Boundary model – predicts begin and end of the answer seq. (assumes answer to be continuous)

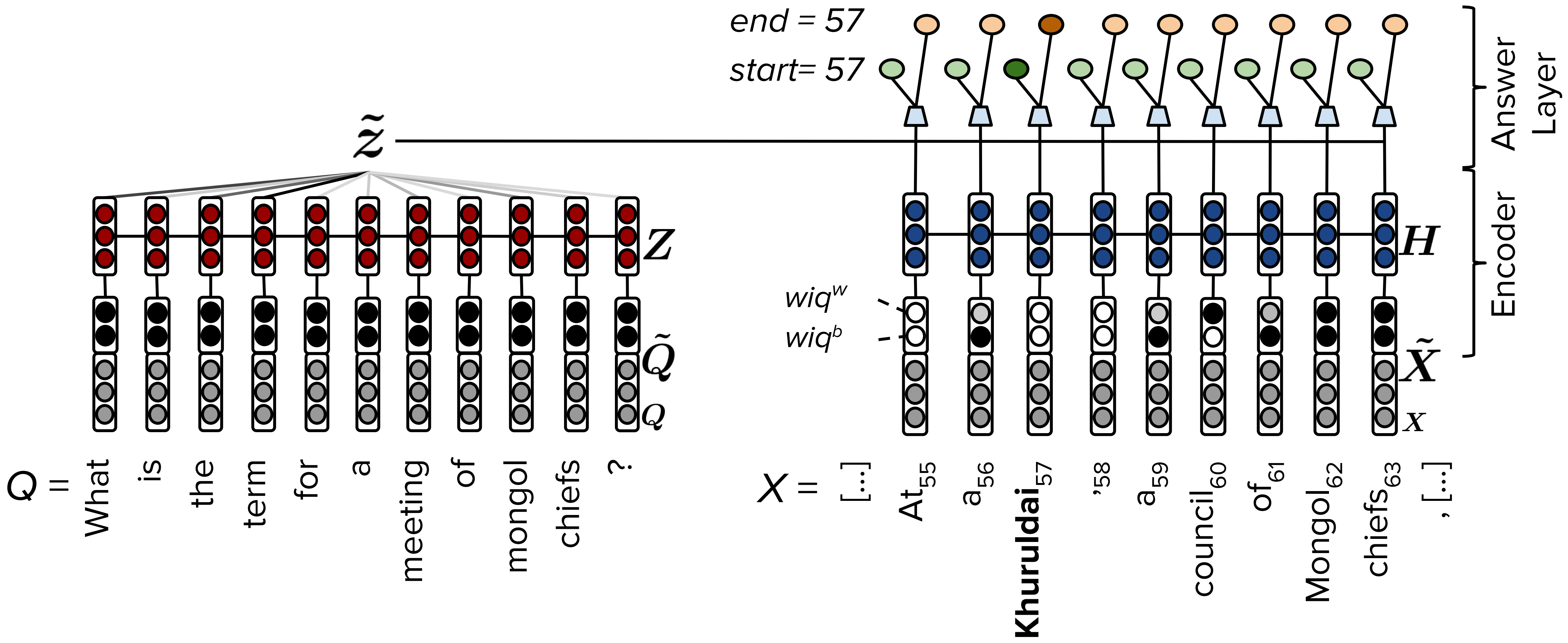
Bi-directional Attention Flow



Gated Attention

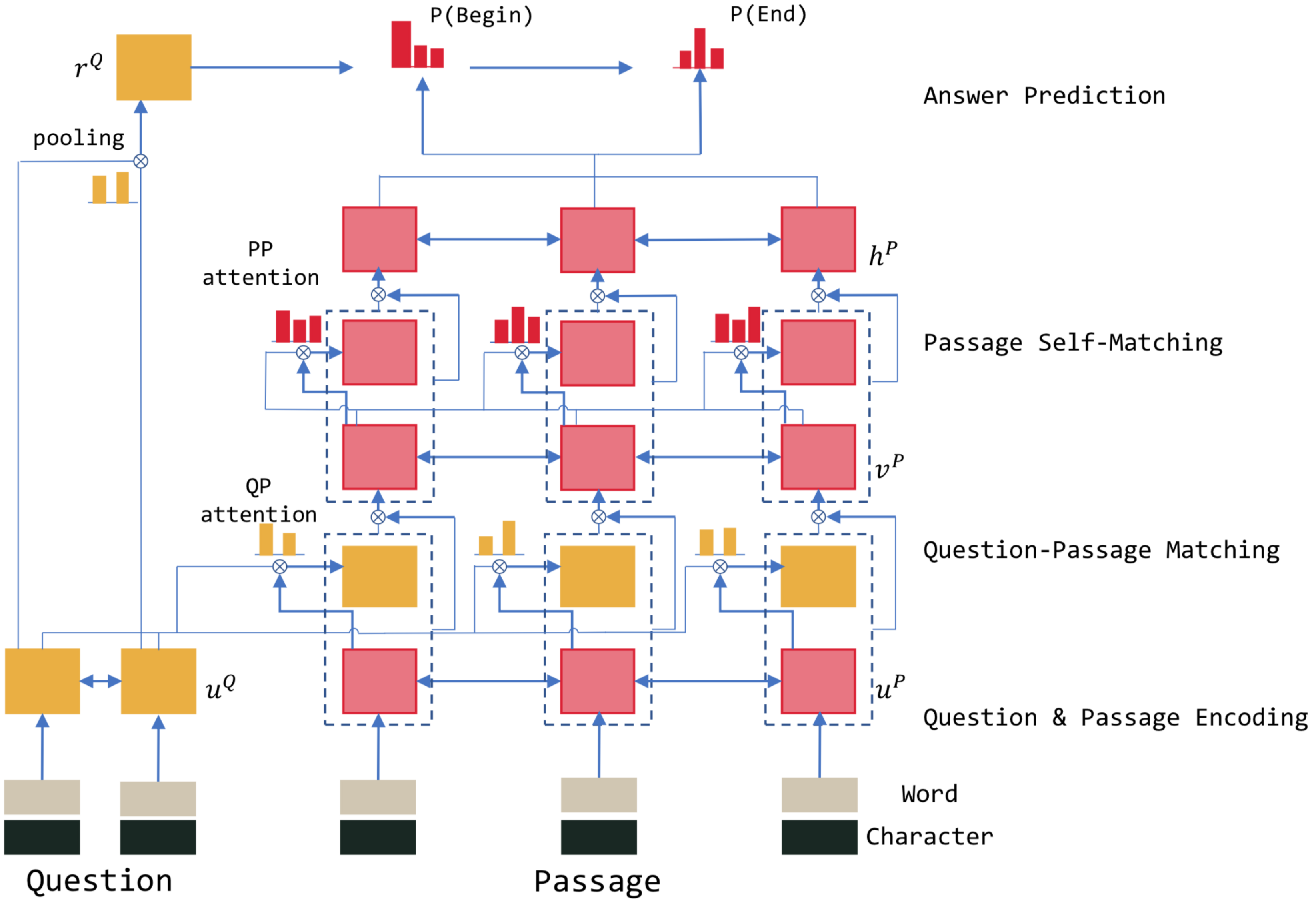


Fastnet

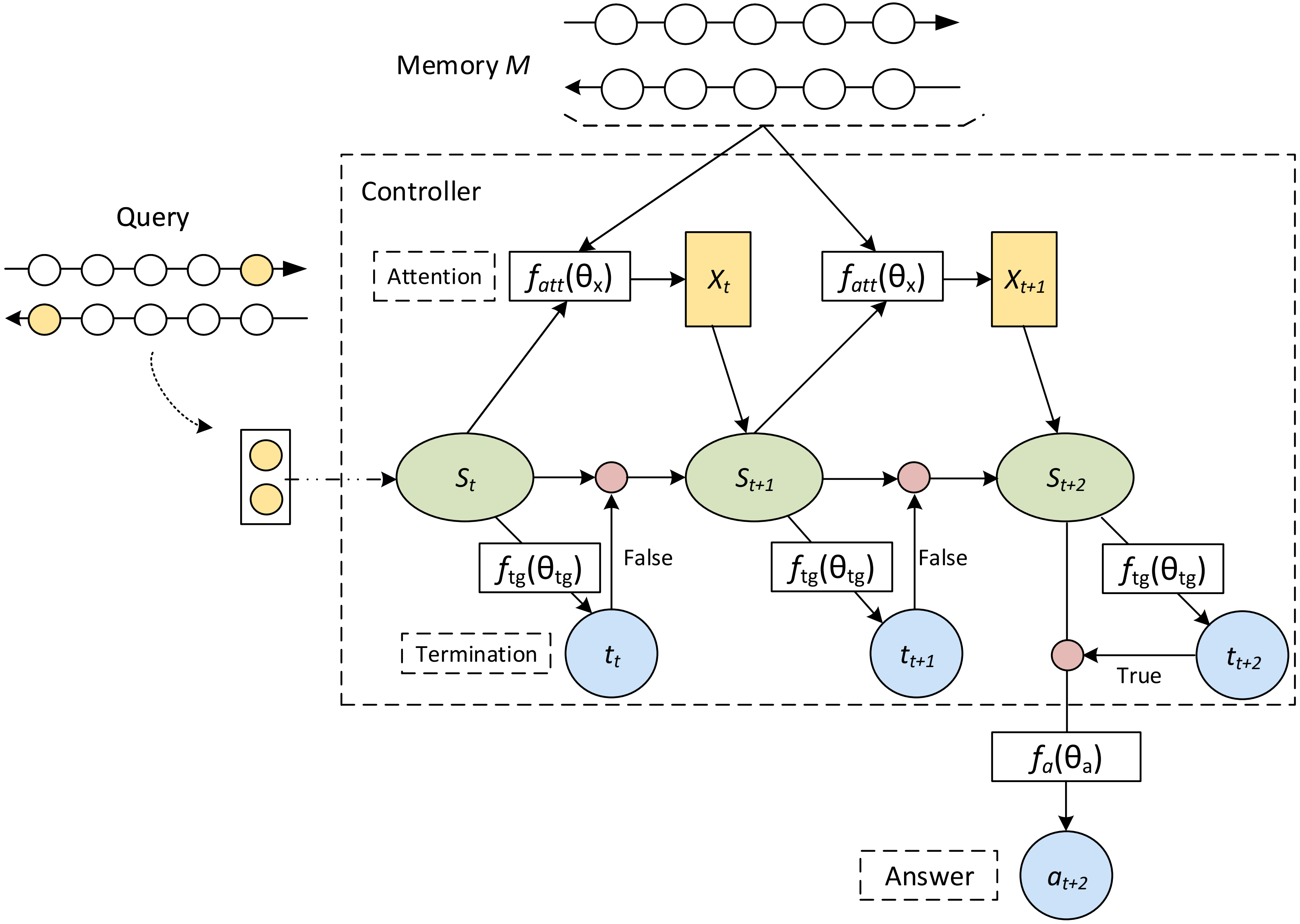


end = 57
start= 57
$\tilde{z}$
Answer Layer
Encoder
Z
H
wiq^w
wiq^b
$\tilde{Q}$
Q
$\tilde{X}$
x
Q = What is the term for a meeting of mongol chiefs ?
X = [...] At₅₅ a₅₆ **Khuruldai**₅₇ ,₅₈ a₅₉ council₆₀ of₆₁ Mongol₆₂ chiefs₆₃ , [...]

R-NET

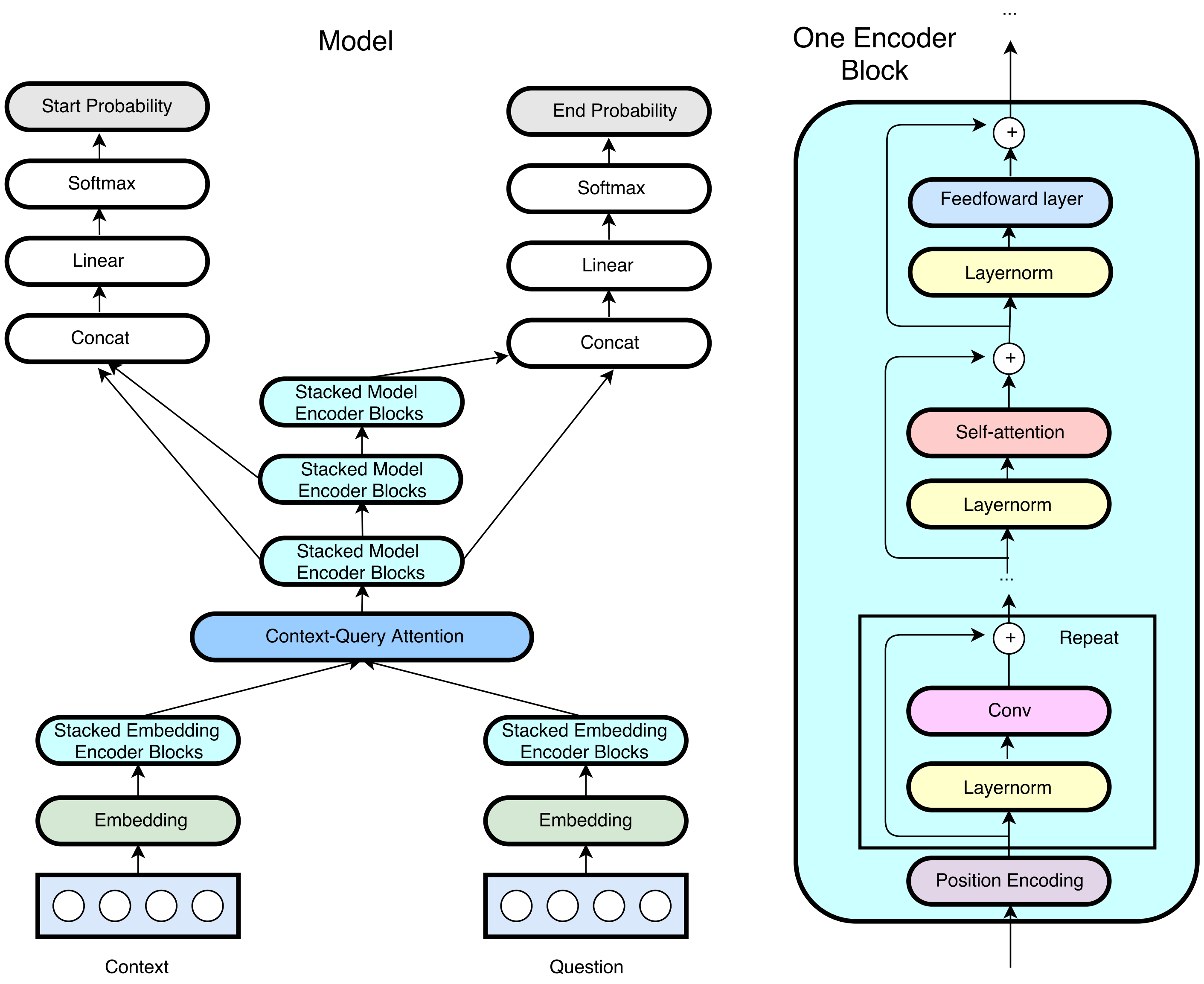

P(Begin)
P(End)
r^Q
pooling
Answer Prediction
PP
attention
h^P
Passage Self-Matching
v^P
QP
attention
Question-Passage Matching
u^Q
u^P
Question & Passage Encoding
Word
Character
Question
Passage

ReasonNet



Memory M
Controller
Query
Attention
$f_{att}(\theta_x)$
X_t
$f_{att}(\theta_x)$
X_{t+1}
S_t
S_{t+1}
S_{t+2}
$f_{tg}(\theta_{tg})$
False
$f_{tg}(\theta_{tg})$
False
$f_{tg}(\theta_{tg})$
Termination
t_t
t_{t+1}
t_{t+2}
True
$f_a(\theta_a)$
Answer
a_{t+2}

QANET


Model
Start Probability
Softmax
Linear
Concat
End Probability
Softmax
Linear
Concat
Stacked Model Encoder Blocks
Stacked Model Encoder Blocks
Stacked Model Encoder Blocks
Context-Query Attention
Stacked Embedding Encoder Blocks
Embedding
Stacked Embedding Encoder Blocks
Embedding
Context
Question
One Encoder Block
...
+
Feedfoward layer
Layernorm
+
Self-attention
Layernorm
...
+
Repeat
Conv
Layernorm
Position Encoding