基于传统符号表示的知识库问答

冯岩松

北京大学

October 19, 2016

知识库问答

任务

利用知识库回答自然语言问题

知识库问答

任务

利用知识库回答自然语言问题

- **输入：** 自然语言问句
- **资源：** 结构化知识库
- **输出：** 答案

知识库问答

任务

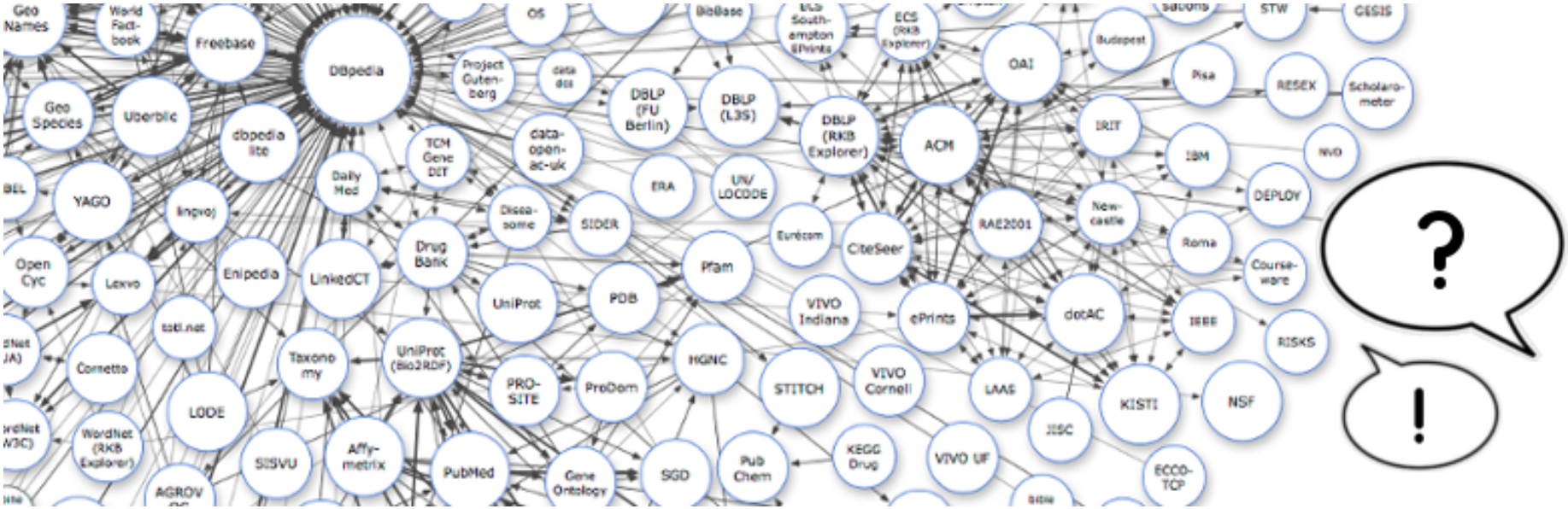
利用知识库回答自然语言问题

- **输入：** 自然语言问句
- **资源：** 结构化知识库，表格，结构化/半结构化记录，……
- **输出：** 答案

知识库问答

任务

利用知识库回答自然语言问题



任务框架

Interstellar

Theatrical release poster

Directed by	Christopher Nolan
Produced by	Emma Thomas
	Christopher Nolan
	Lynda Obst
Written by	Jonathan Nolan
	Christopher Nolan
Starring	Matthew McConaughey
	Anne Hathaway
	Jessica Chastain
	Bill Irwin
	Ellen Burstyn
	Michael Caine

What else did the director of the movie Interstellar direct ?

select ?y

fb:m.0fkf28	fb:object.type	fb:film.film
fb:m.0fkf28	fb:film.film.directed_by	?x
?x	fb:film.director.fim	?y
?y	fb:object.type	fb:film.film

Freebase

DBpedia

Inception,
The Dark Knight Rises
The Dark Knight
Batman Begins
.....

任务框架

What else did the director of the movie Interstellar direct ?

Convert to a Query

select ?y

fb:m.0fkf28	fb:object.type	fb:film.film
fb:m.0fkf28	fb:film.film.directed_by	?x
?x	fb:film.director.fim	?y
?y	fb:object.type	fb:film.film

Query over KBs

Inception,
The Dark Knight Rises
The Dark Knight
Batman Begins
.....

早期工作

- **Baseball**：有关棒球比赛的问答 [Green et al., 1961]
  - *How many games did the Yankees play in July?*
- **LUNAR**：有关科研数据的问答 [Woods, 1993]
  - *How many samples contain Titanium?*
- 订机票、地理知识，找工作……

早期工作

- **Baseball**：有关棒球比赛的问答 [Green et al., 1961]
  - *How many games did the Yankees play in July?*
- **LUNAR**：有关科研数据的问答 [Woods, 1993]
  - *How many samples contain Titanium?*
- 订机票、地理知识，找工作……

一些问题

- 面向特定领域
- 解答所需的知识库结构较为简单：数百个实体、关系
- 知识库规模有限：数千个三元组
- 大多可以采用人工设计的模板或规则来解析问题
- 难以应付开放域问题：更多的实体、关系；更多的三元组

大规模知识库

Key

更多实体、更多关系

- Freebase
- DBpedia

大规模知识库

Key

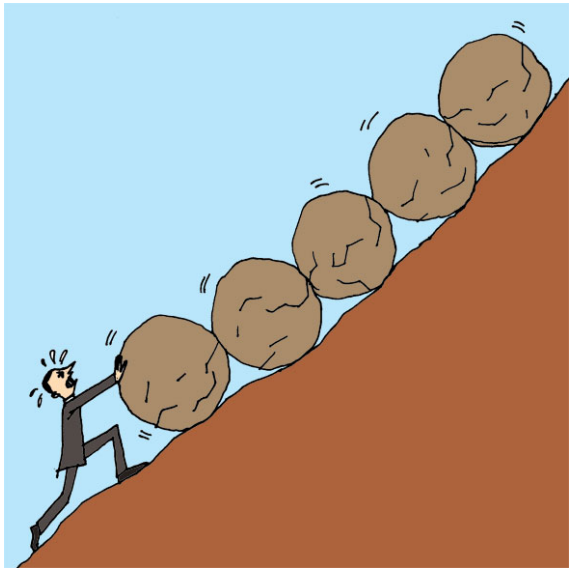
更多实体、更多关系

- Freebase → Free917, WebQuestion
- DBpedia → QALD

数据集

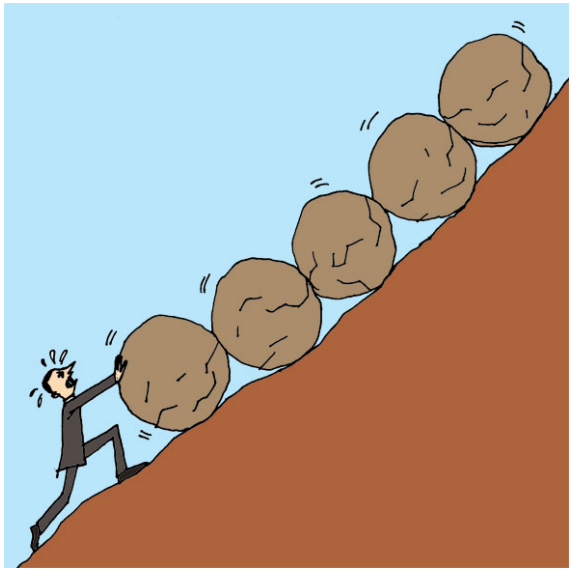
- Free917 [Cai and Yates, 2013]
  - 使用Freebase
  - 917个问题，标注了逻辑表达式
- WebQuestions [Berant et al., 2013]
  - 使用Freebase
  - 5,810个问题，通过Google Suggest API爬取
  - 利用 Amazon Mechanical Turk 服务得到答案 (存在错误)
  - 一个问题可能存在多个答案
  - 利用 Average F1 评价
  - WebQuestionsSP: WebQuestions with SPARQL annotations [Yih et al., 2016]
- QALD
  - 知识库问答评测，QALD：Question Answering over Linked Data
  - 使用 DBpedia
  - 每年100个问题左右
  - Hybrid Track：DBpedia不足以完整回答问题，必须依靠文本信息
- Simple Questions [Bordes et al., 2015]
  - 108,442个简单问题，附带一条Freebase 三元组作为答案

技术挑战



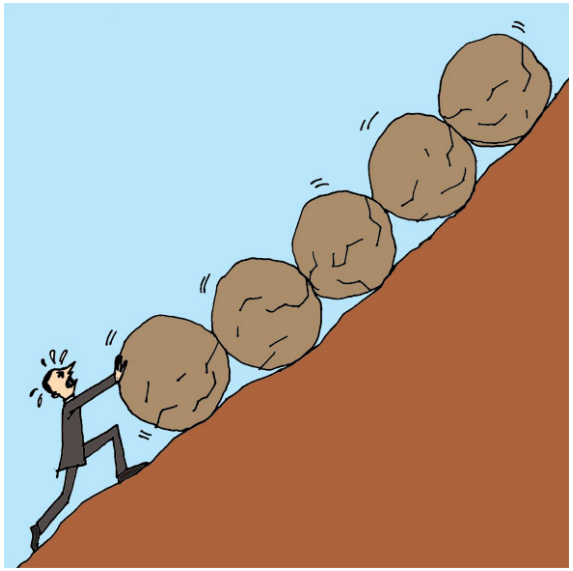
1. 如何恰当地表示问题的语义
2. 如何将问题的语义表示与知识库进行关联

技术挑战



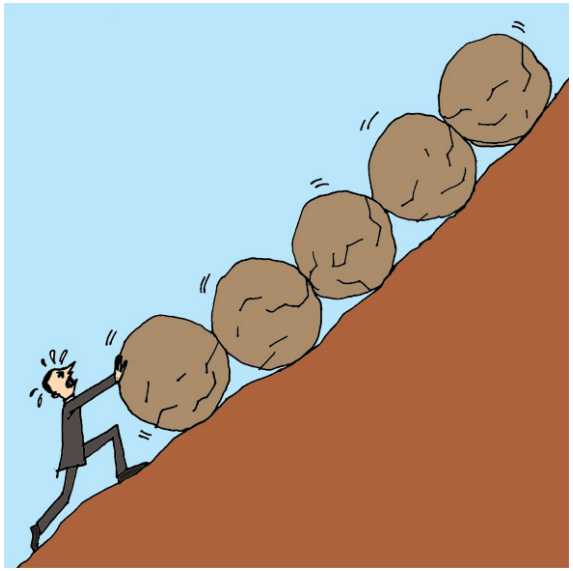
1. 如何恰当地表示问题的语义
   - 丰富的提问方式，复杂的提问意图
2. 如何将问题的语义表示与知识库进行关联

技术挑战



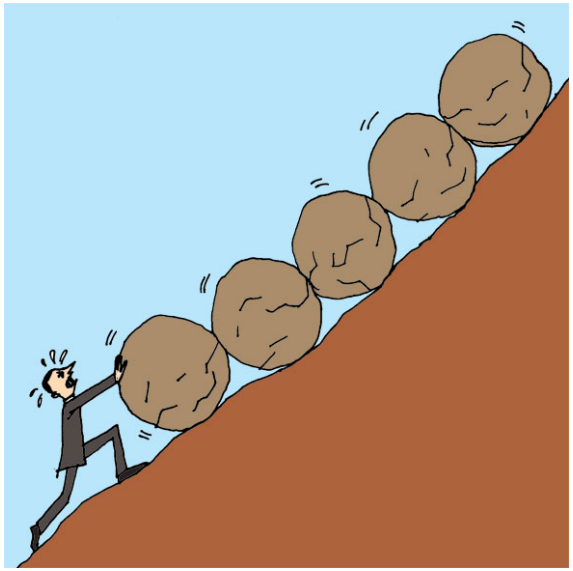
1. 如何恰当地表示问题的语义
   - 丰富的提问方式，复杂的提问意图
2. 如何将问题的语义表示与知识库进行关联
   - 大规模开放域知识库

技术挑战



1. 如何恰当地表示问题的语义
   - 丰富的提问方式，复杂的提问意图 $\rightarrow$ 语义分析
2. 如何将问题的语义表示与知识库进行关联
   - 大规模开放域知识库 $\rightarrow$ 知识库映射

技术挑战



1. 如何恰当地表示问题的语义
   - 丰富的提问方式，复杂的提问意图 → 语义分析
2. 如何将问题的语义表示与知识库进行关联
   - 大规模开放域知识库 → 实体链接、关系抽取

语义分析

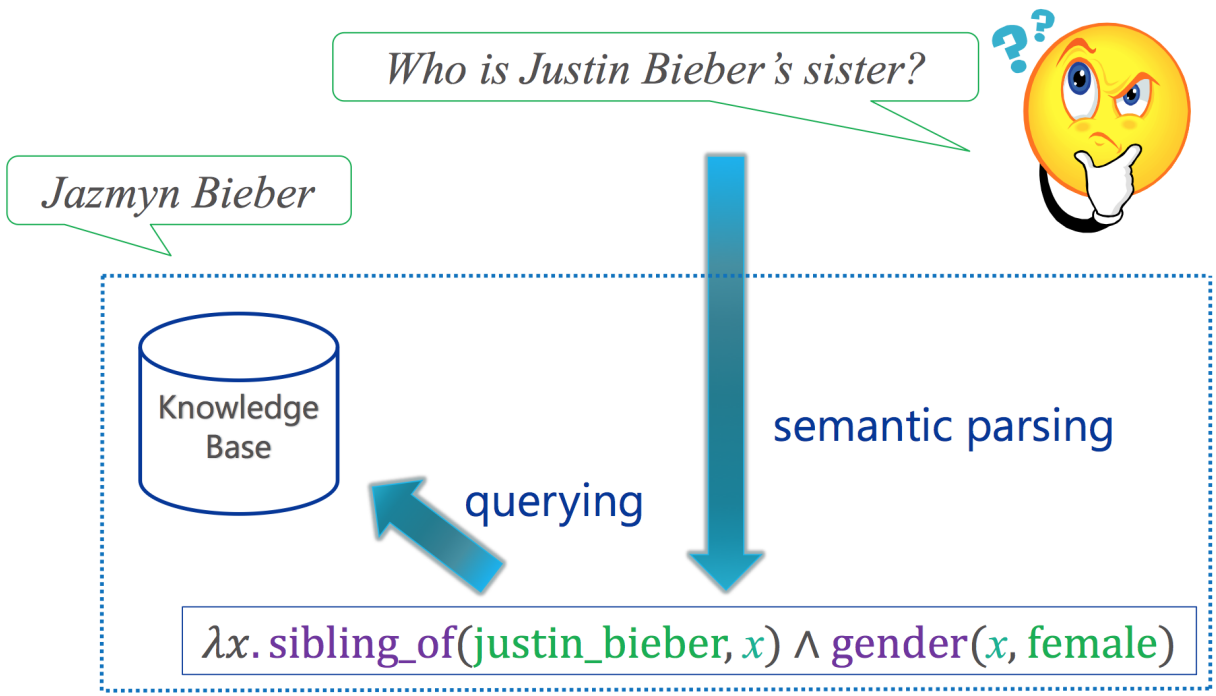
利用形式化方法表示问题语义

- 一步到位
- 两步方法

语义分析

利用形式化方法表示问题语义

- 一步到位
  - 直接获得与具体知识库相关的语义表示

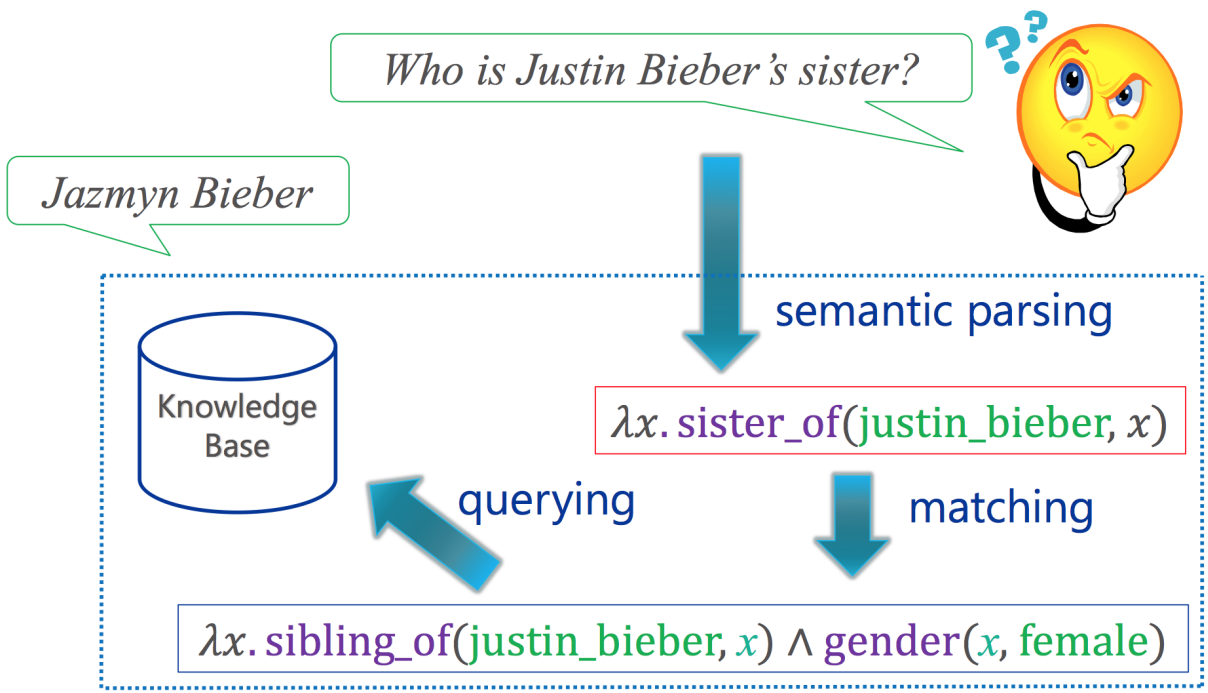


[Kwiatkowski et al., 2013, Yih and Ma, 2016]

语义分析

利用形式化方法表示问题语义

- 一步到位
- 两步方法
  - 先通用语义表示
  - 与再实现知识库映射



语义分析

利用形式化方法表示问题语义

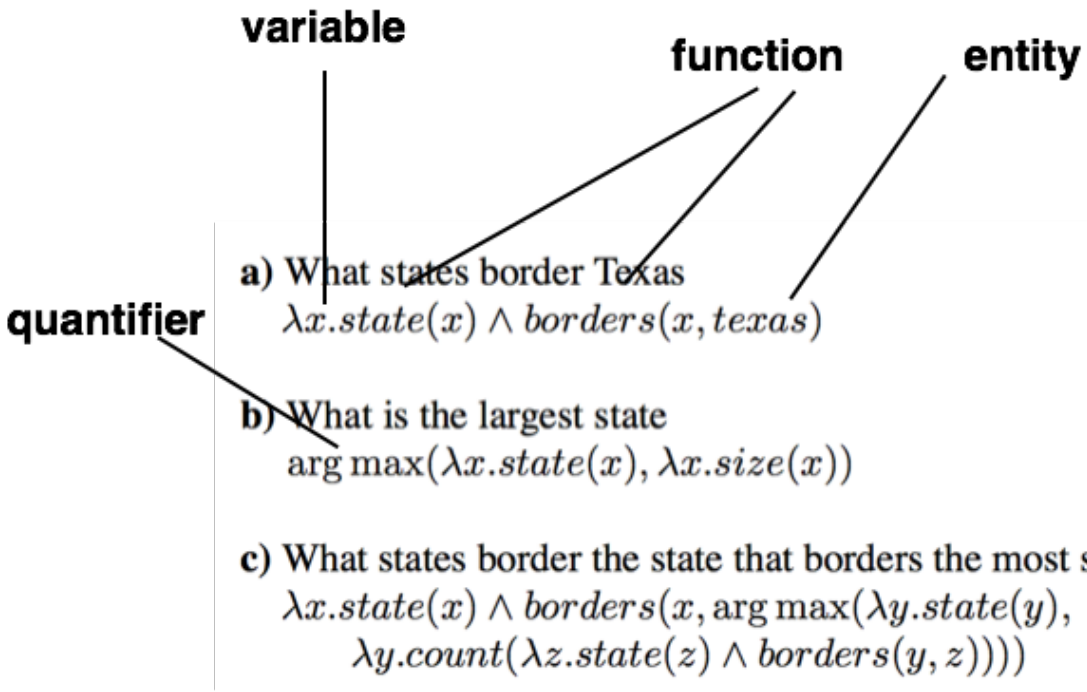
- 一步到位
- 两步方法

语义表示

- λ Calculus: $\lambda x.sibling\_of(justin\_bieber, x) \wedge gender(x, female)$
- Lambda Dependency-based Compositional Semantics (λ-DCS) $SisterOf.Justin\_Bieber$
- 借助于现有句法、语义分析技术，PCFG、CCG、Phrase Dependency Graph
- Query Graph
- ...

语义表示：λ Calculus

- constants
  - entities, numbers, functions
- logical connectors
  - $\vee, \wedge, \neg, \rightarrow$
- quantification
  - $\exists, \forall$
- additional quantifiers
  - argmax, argmin, ...

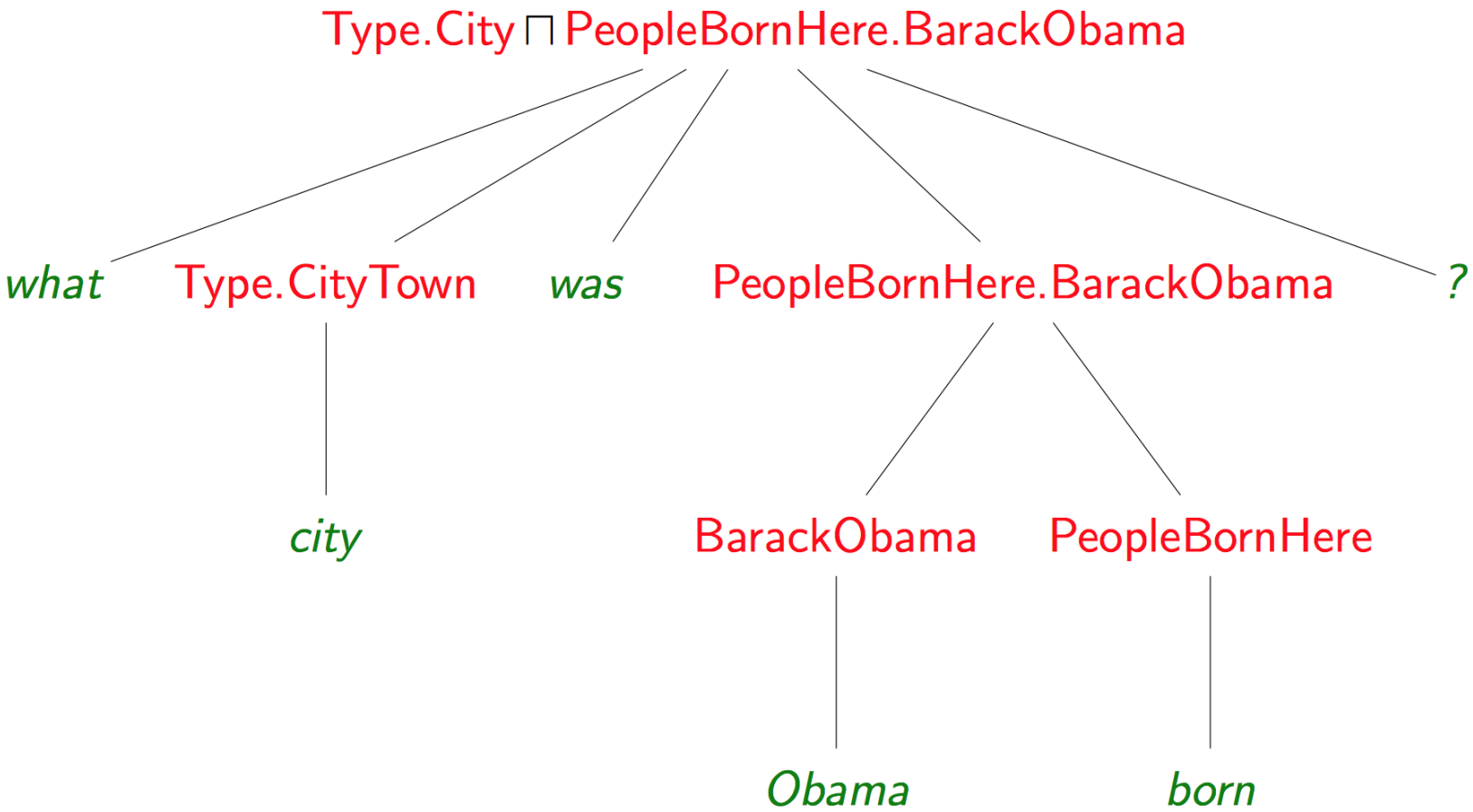


[Liu, 2015]

语义表示：λ-DCS

Lambda Dependency-based Compositional Semantics [Liang, 2011]

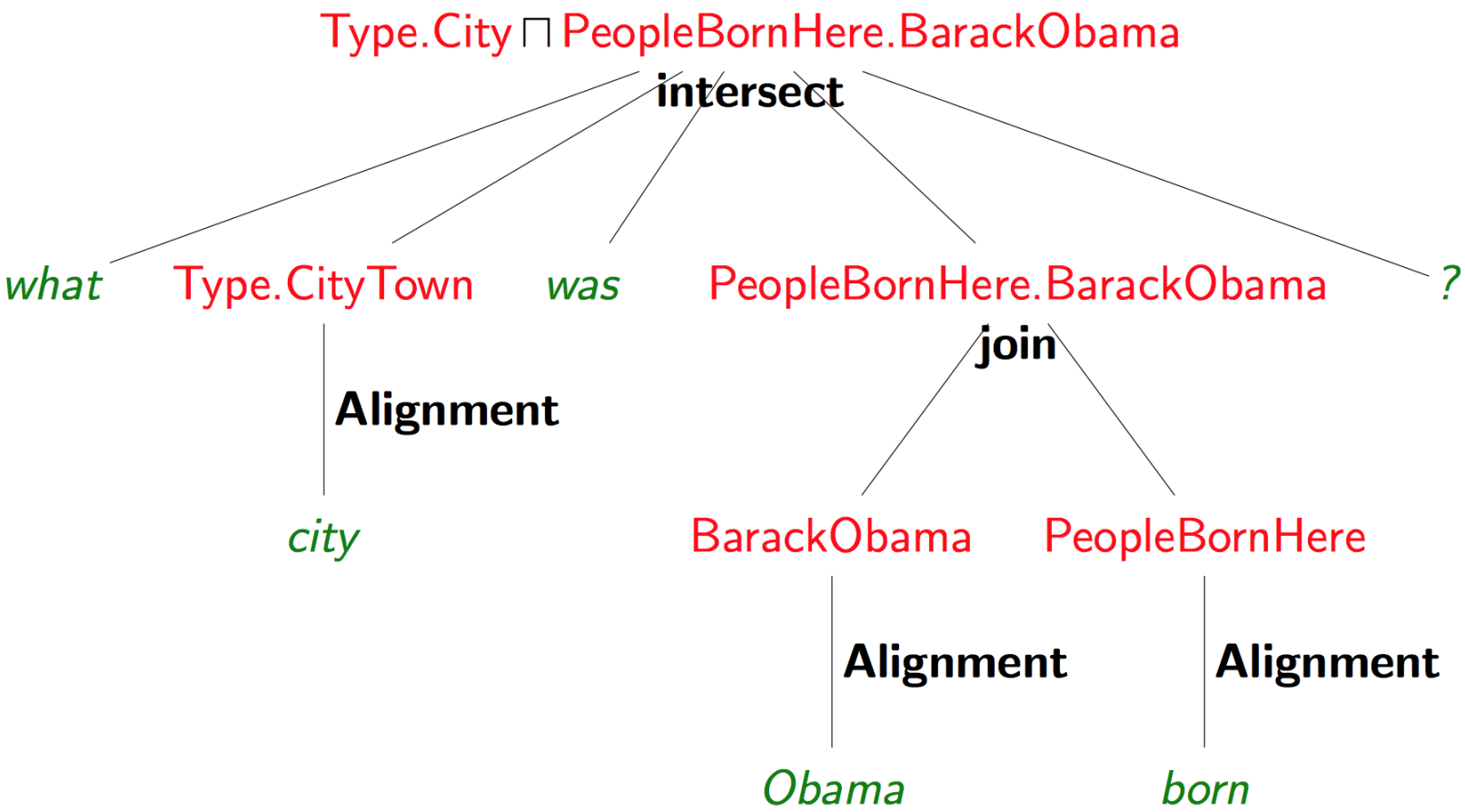
- 更简洁、易用
- 组成：实体、关系、Join/Intersection 操作



语义表示：λ-DCS

Lambda Dependency-based Compositional Semantics [Liang, 2011]

- 更简洁、易用
- 组成：实体、关系、Join/Intersection 操作



语义表示：CCG

- 组合范畴语法：Combinatory Categorial Grammar

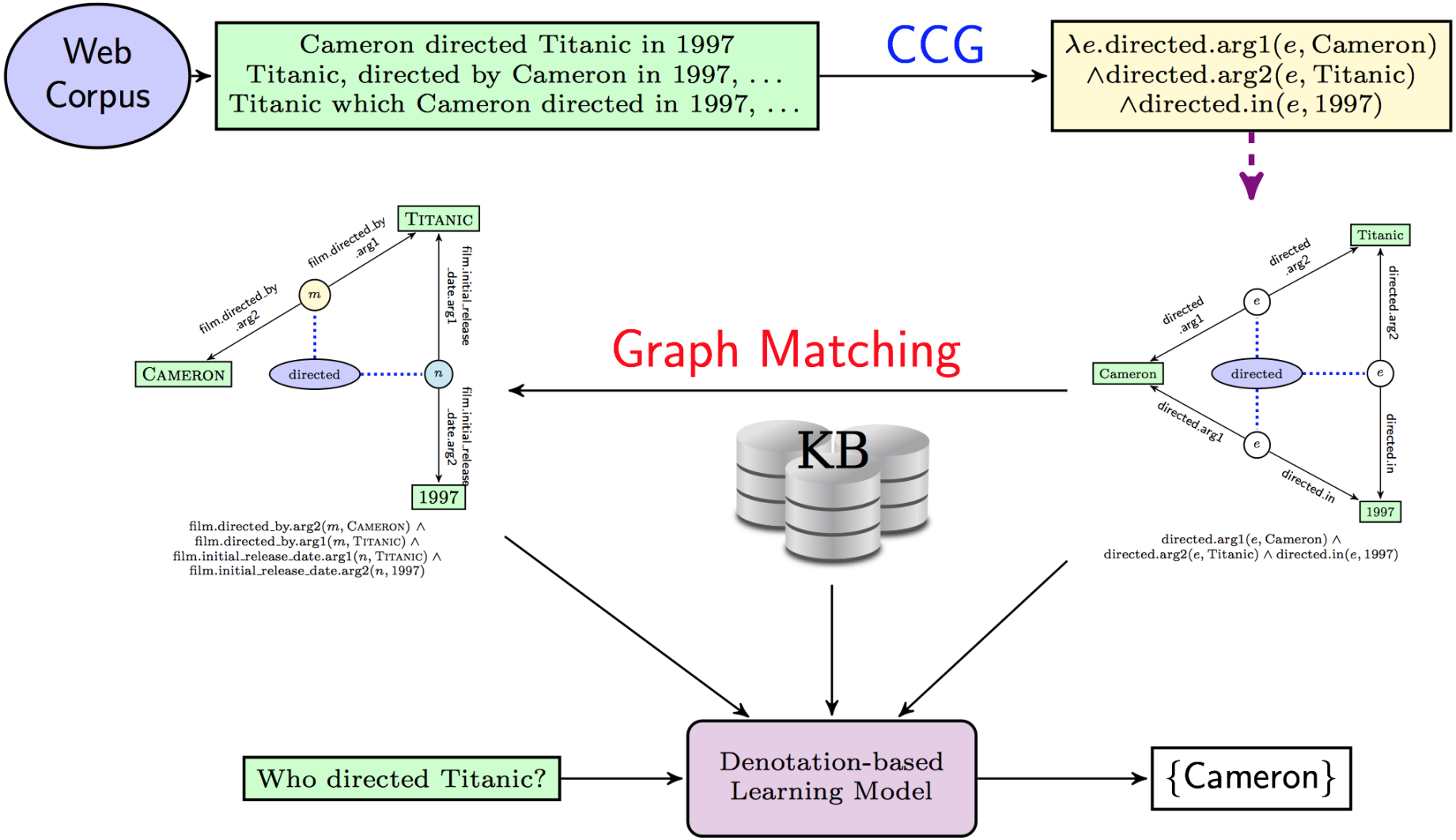
Cameron directed Titanic in 1997

Cameron	directed	Titanic	in	1997
NP	$((S\backslash NP)/PP[in])/NP$	NP	PP/NP	NP
Cameron	$\lambda z\lambda y\lambda x\lambda e.$ directed.arg1(e,x) $\wedge$ directed.arg2(e,z) $\wedge$ directed.in(e,y)	Titanic	$\lambda x.x$	1997
	$(S\backslash NP)/PP$ (>)		PP (>)	
	$\lambda y\lambda x\lambda e.$ directed.arg1(e,x) $\wedge$ directed.arg2$(e,$ Titanic$)$ $\wedge$ directed.in(e,y)		1997	
	$S\backslash NP$ (>) $\lambda x\lambda e.$ directed.arg1$(e,x)\wedge$ directed.arg2$(e,$ Titanic$)$ $\wedge$ directed.in$(e,$ 1997$)$			
S (<) $\lambda e.$ directed.arg1$(e,$ Cameron$)\wedge$ directed.arg2$(e,$ Titanic$)\wedge$ directed.in$(e,$ 1997$)$				

[Reddy et al., 2015]

语义表示：CCG

- 组合范畴语法：Combinatory Categorial Grammar

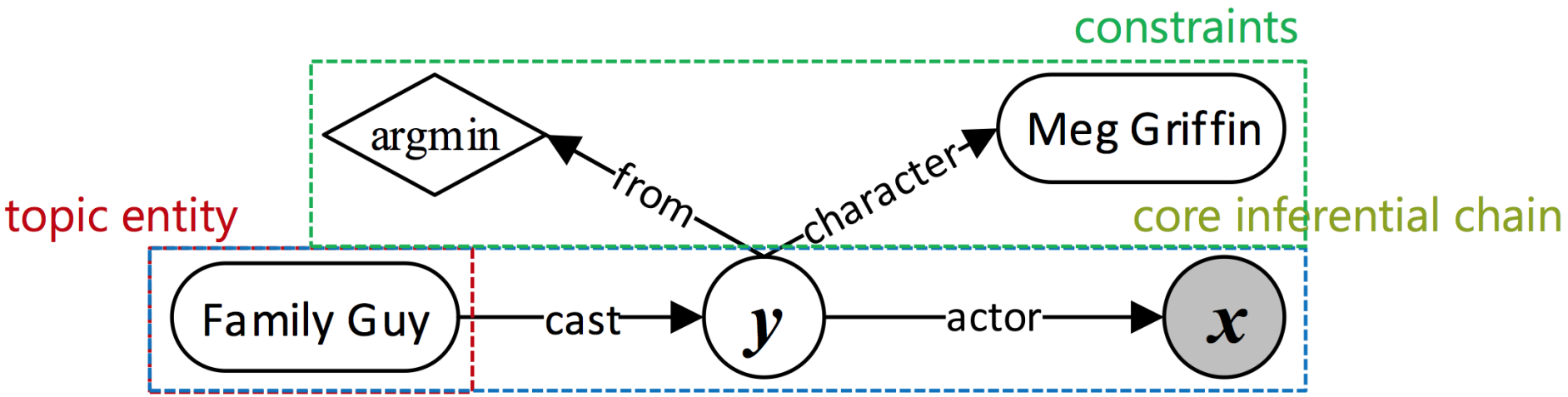


语义表示：Query Graph

- 借助于知识图谱中的片段，以及实体链接/关系抽取

Who first voiced Meg on Family Guy?

$\Longrightarrow \lambda x.\exists y.cast(FamilyGuy, y) \wedge actor(y, x) \wedge character(y, MegGriffin)$

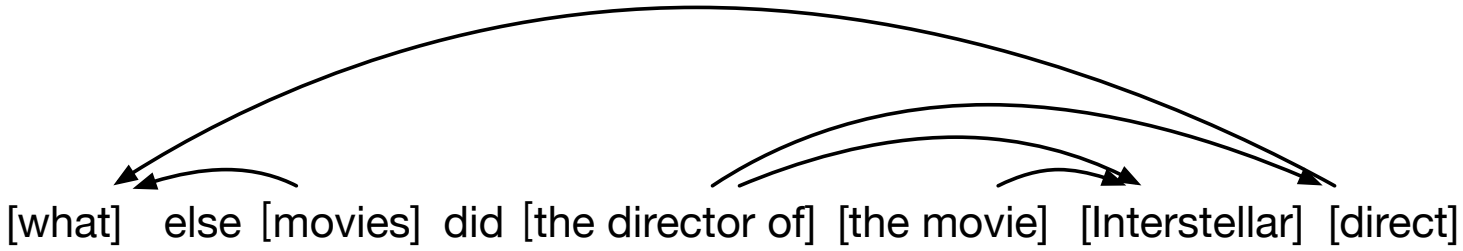


- 直接有效: 疑问词、实体、关系、及一些限制节点组成的图

[Yih et al., 2015]

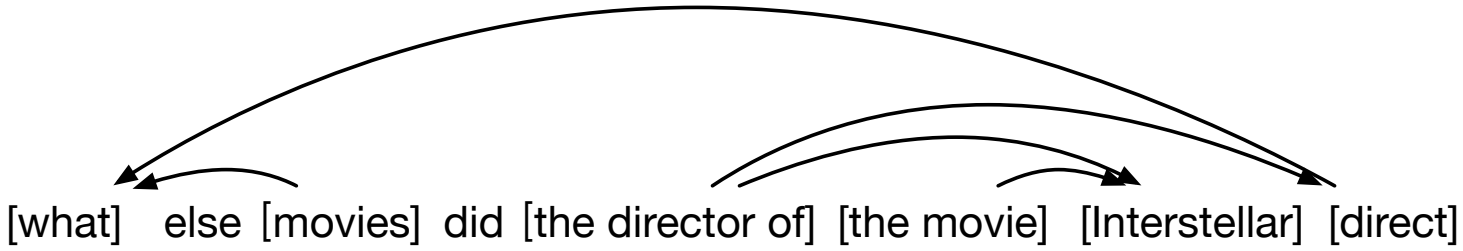
语义表示：Phrase Dependency Graph

1. 语义表示需独立于某一具体知识库



语义表示：Phrase Dependency Graph

1. 语义表示需独立于某一具体知识库



2. 与知识库的映射可以更灵活

Freebase

select ?y

fb:m.0fkf28	fb:object.type	fb:film.film
fb:m.0fkf28	fb:film.film.directed_by	?x
?x	fb:film.director.fim	?y
?y	fb:object.type	fb:film.film

DBpedia

select ?y

ns:Interstellar	dbo:type	dbo:film
ns:Interstellar	dbp:director	?x
?y	dbp:director	?x
?y	dbo:type	dbo:film

[Xu et al., 2014]

Framework

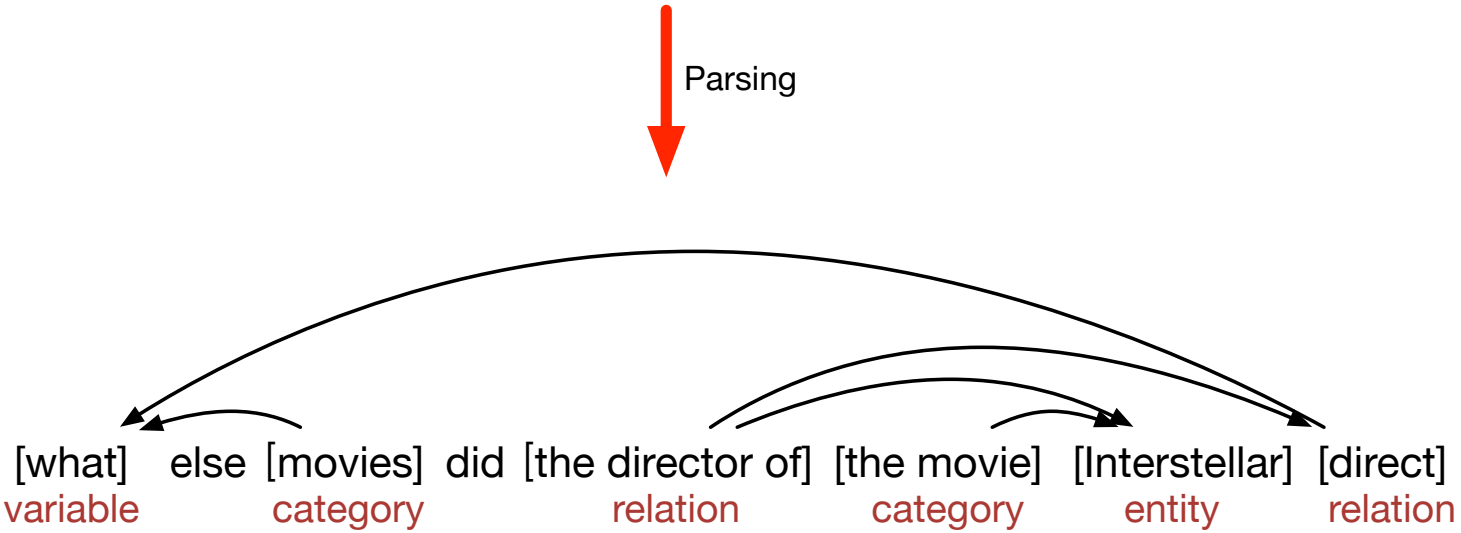
what else movies did the director of the movie Interstellar direct

Framework

what else movies did the director of the movie Interstellar direct

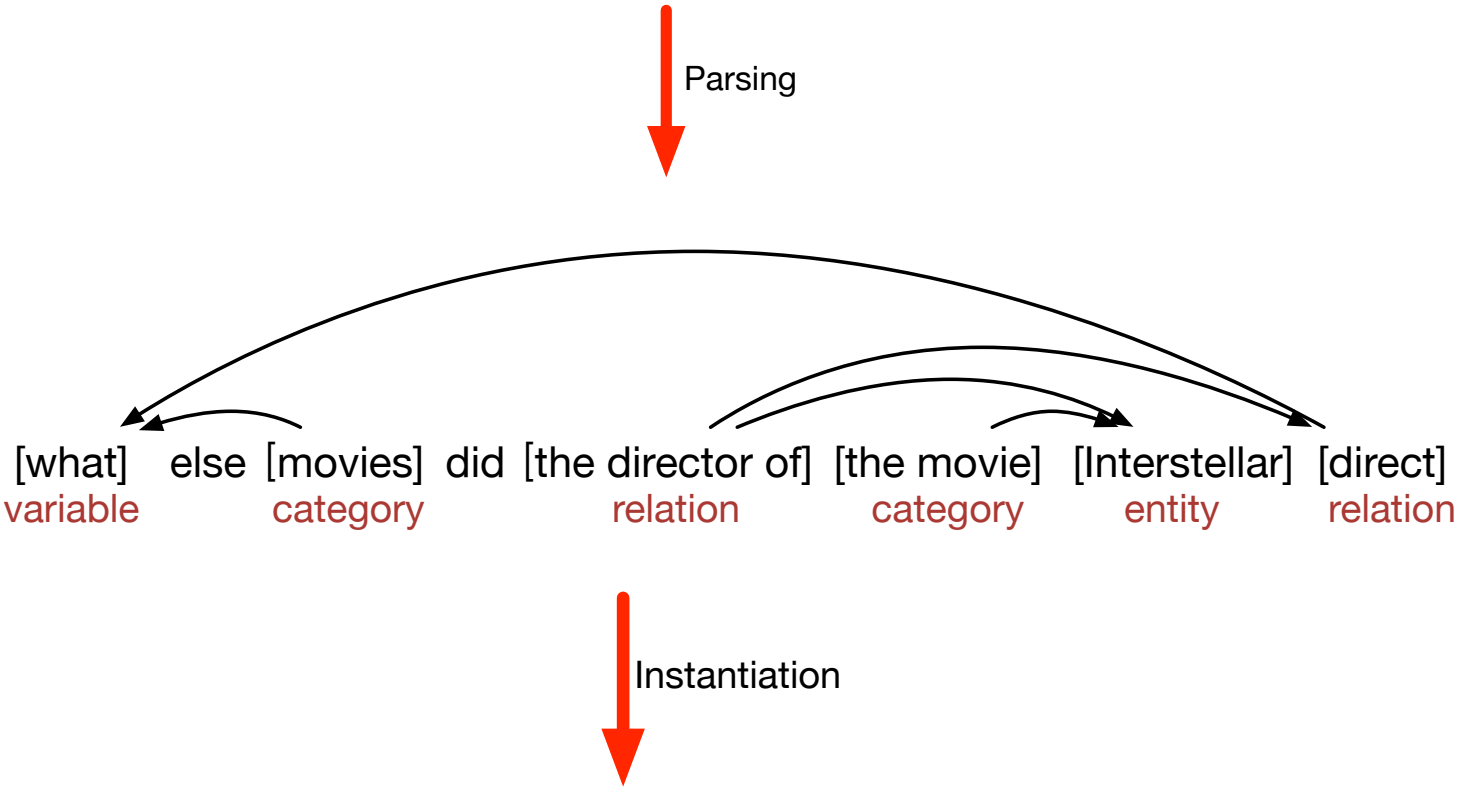
Framework

what else movies did the director of the movie Interstellar direct



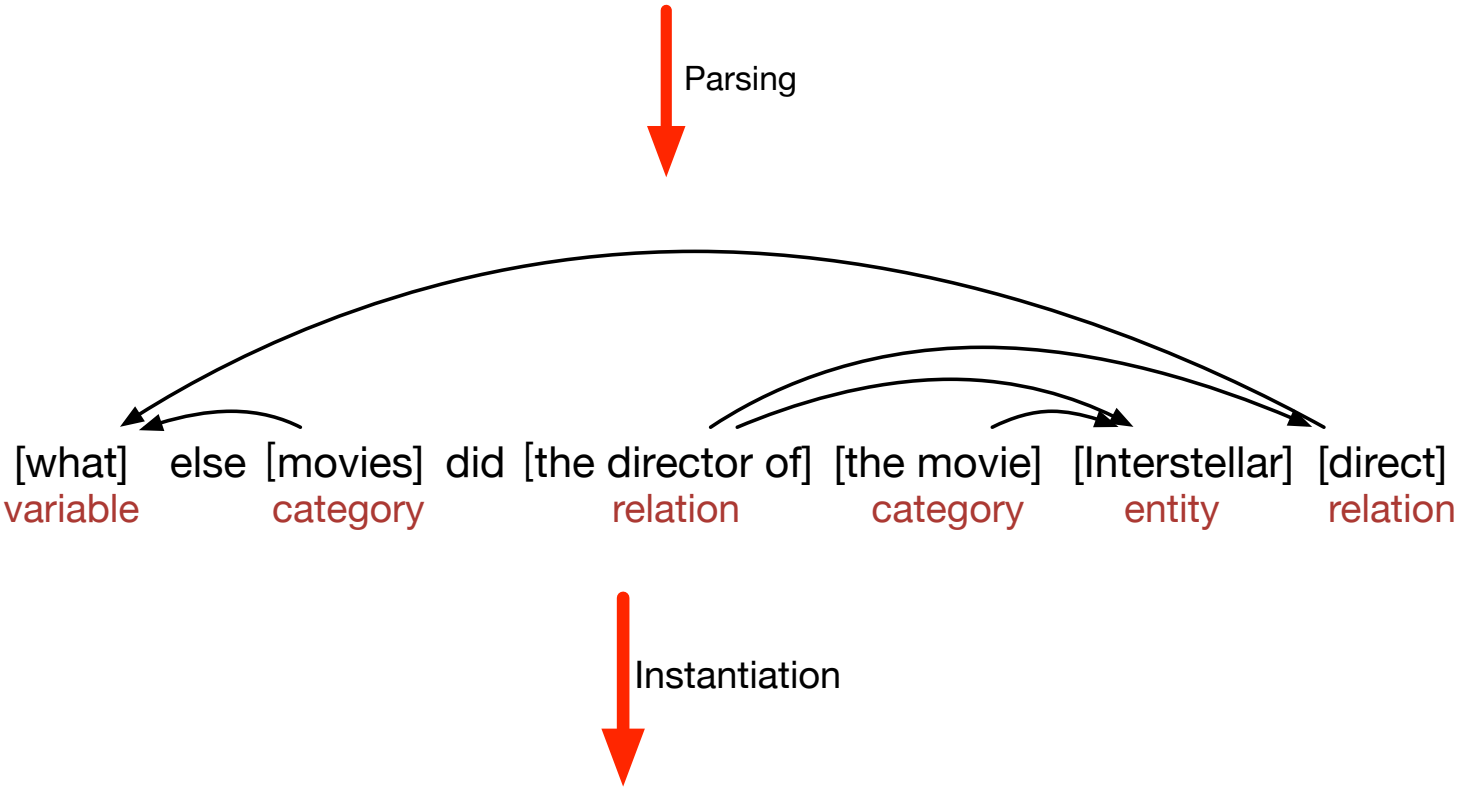
Framework

what else movies did the director of the movie Interstellar direct

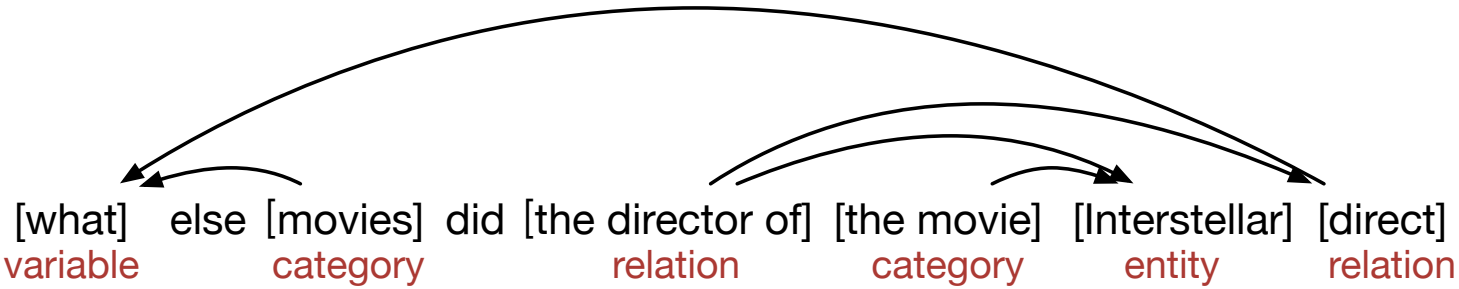


Framework

what else movies did the director of the movie Interstellar direct



Phrase Dependency Graph



Node

每个短语都具有语义标签 $l \in$ {entity, category, variable, relation}

Edge

短语之间的 谓词–论元 结构

unary predicate

binary predicate

结构预测问题

输入: 自然语言问题

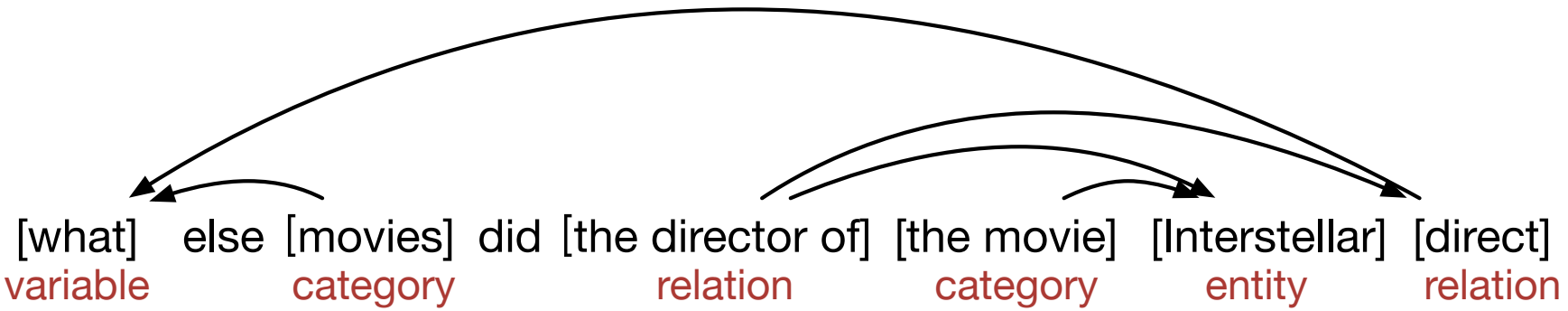
输出: Phrase Dependency Graph

流水线框架

1. Phrase Detection

what	else	movies	did	the director of	the movie	Interstellar	direct
variable		category		relation	category	entity	relation

2. Phrase Dependency Parsing

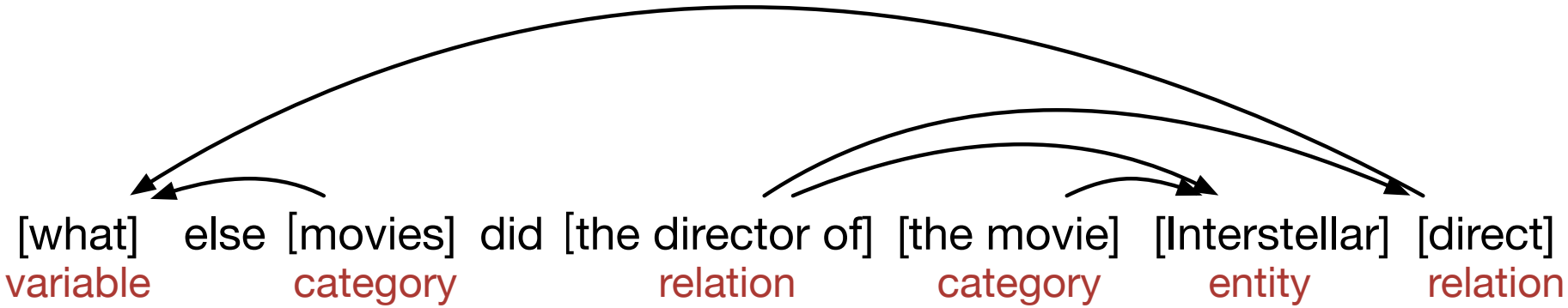


Phrase Detection

what	else	movies	did	the	director	of	the	movie	Interstellar	direct
V-B	O	C-B	O	R-B	R-I	R-I	C-B	C-I	E-B	R-B

- Sequence labeling problem
- Structured perceptron with lexical features

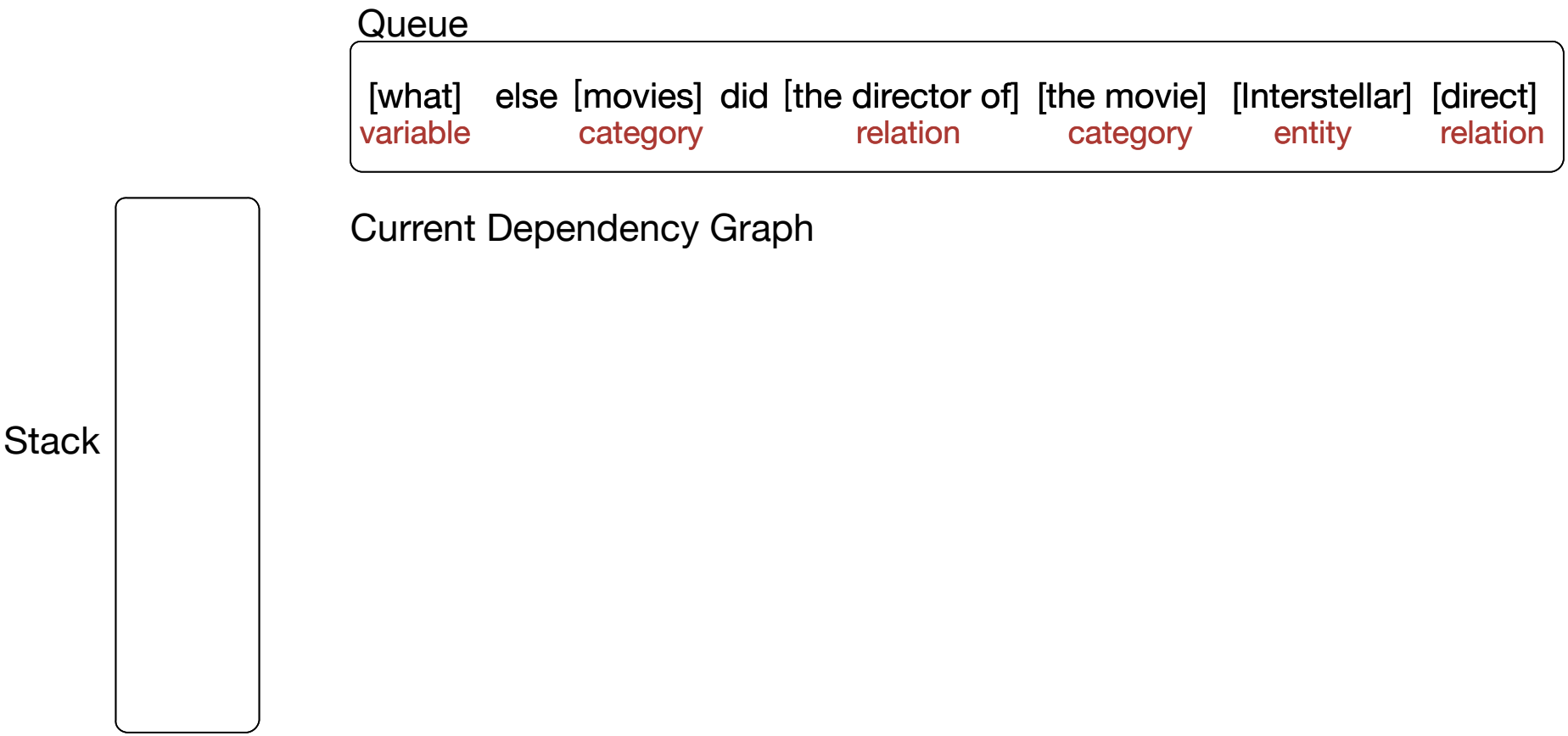
Phrase Dependency Parsing



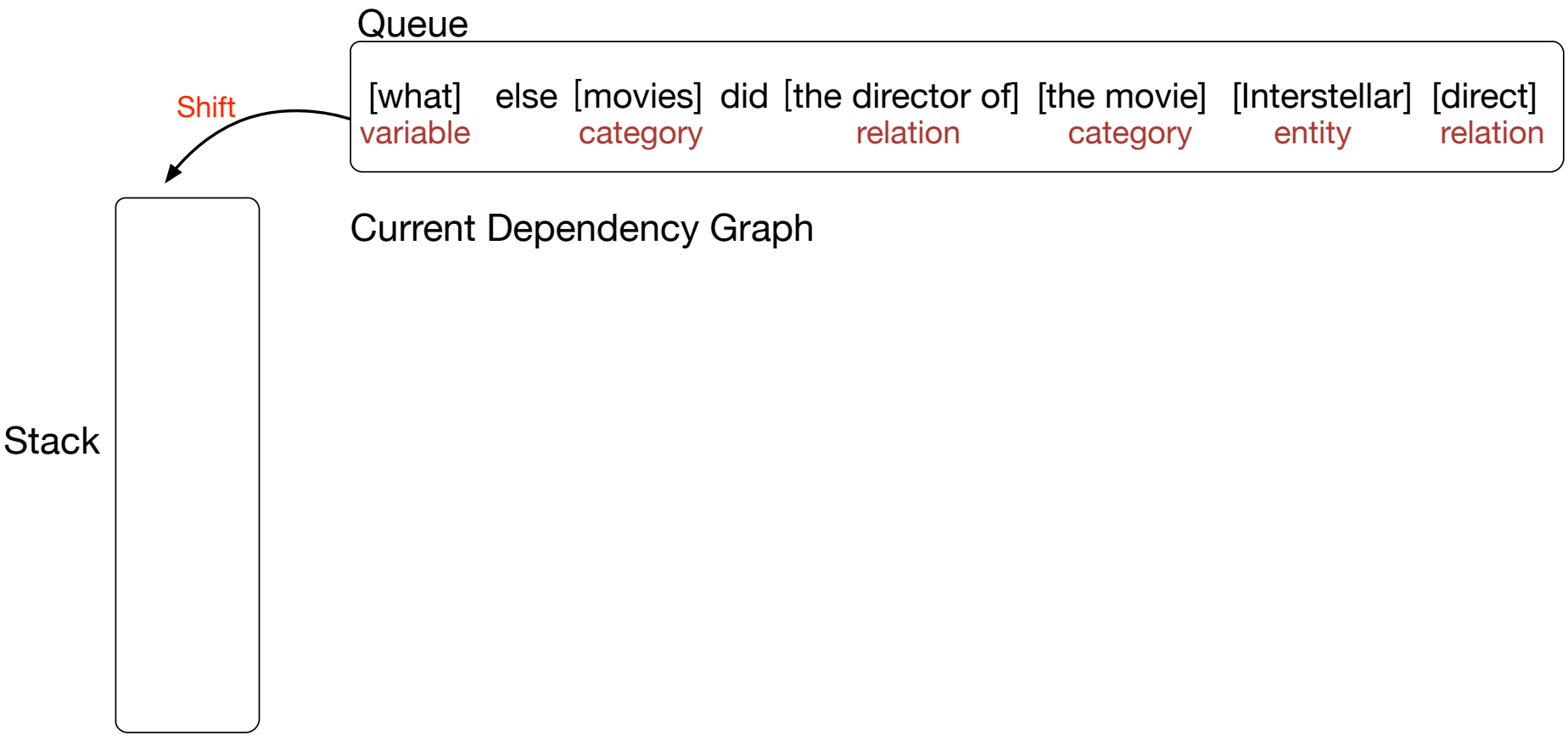
Transition-based parsing

- A queue of incoming phrases
- A stack of processed phrases
- Four actions (ArcLeft, ArcRight, Shift, Reduce)
- Multiple heads

Parsing Example



Parsing Example



Parsing Example

Queue

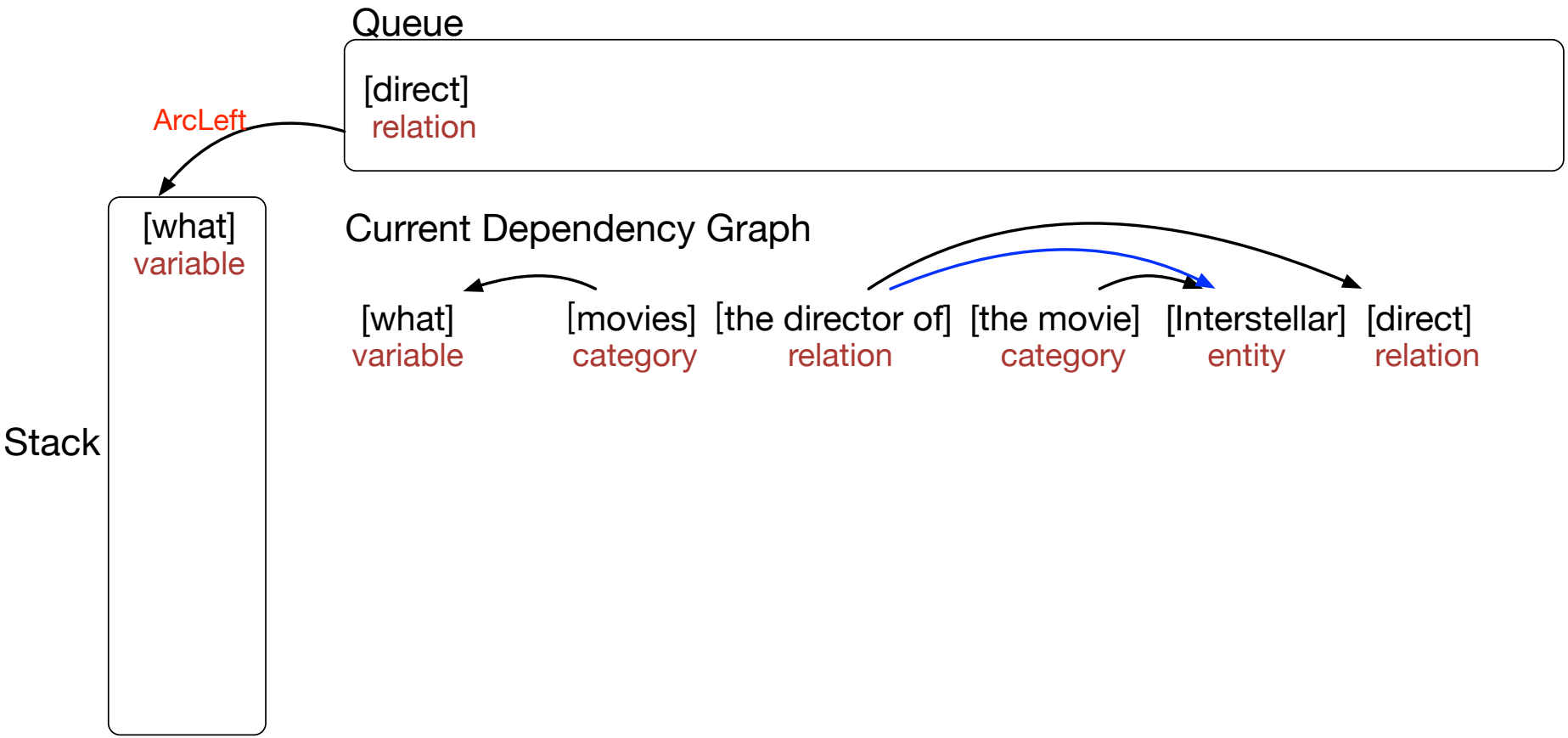
else	[movies]	did	[the director of]	[the movie]	[Interstellar]	[direct]
	category		relation	category	entity	relation

Stack

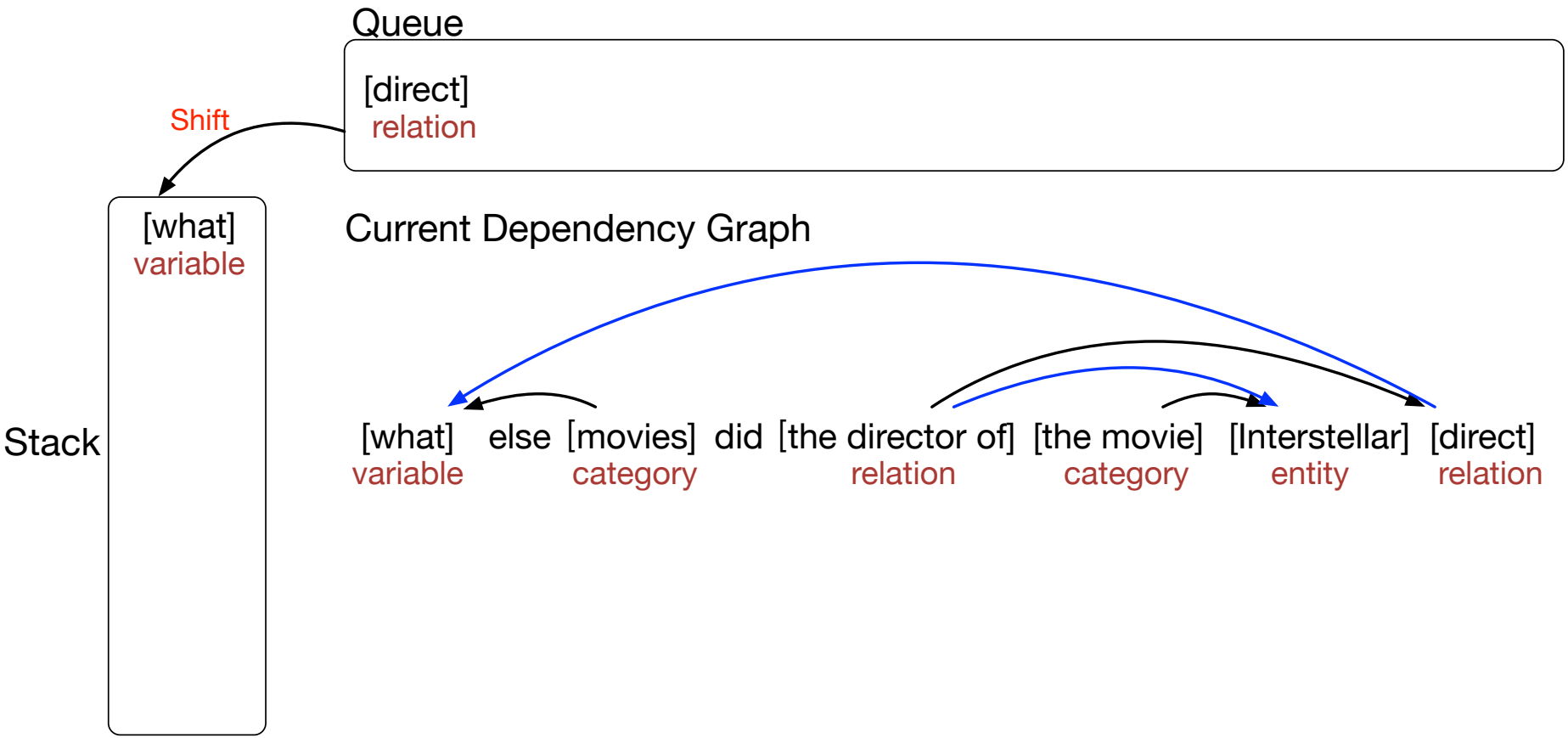
[what]
variable

Current Dependency Graph

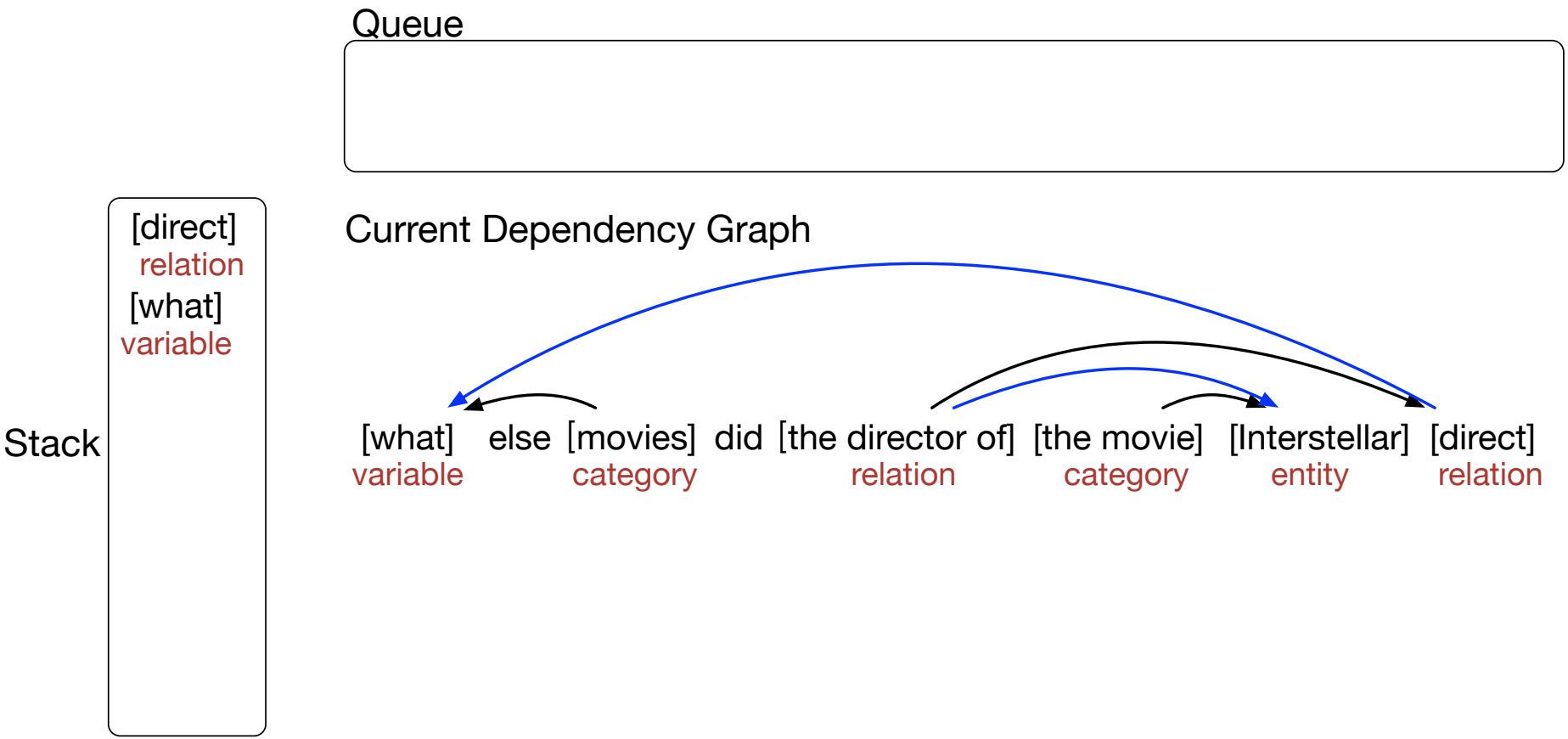
Parsing Example



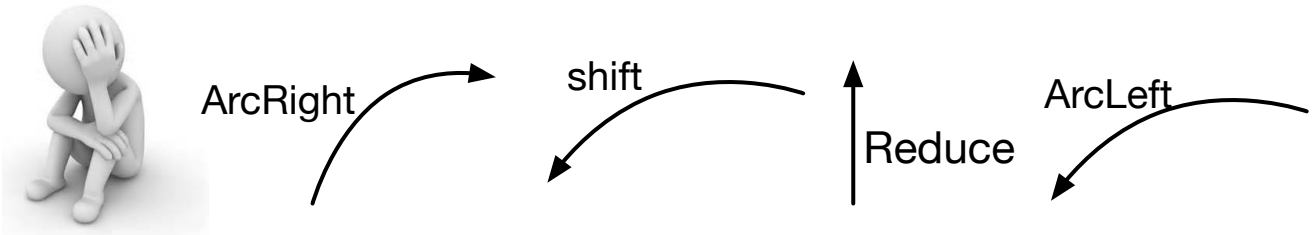
Parsing Example



Parsing Example



解码



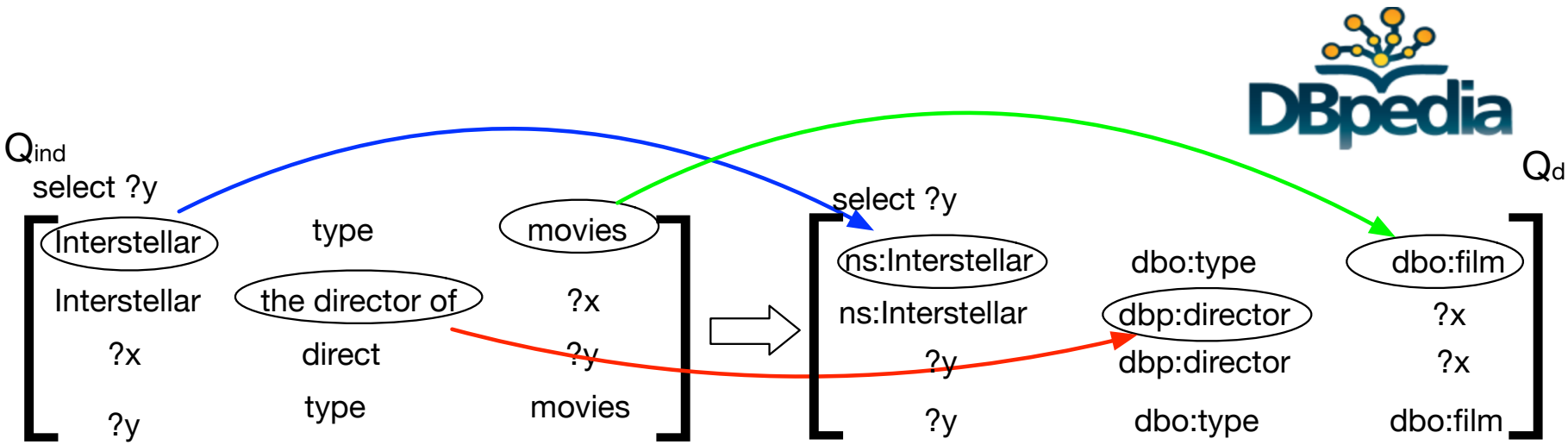
- incremental processing

$$z = \arg\max_{a \in A} w \cdot f(S, Q, a)$$

where A = {`Shift`, `Reduce`, `ArcLeft`, `ArcRight`}

- 特征
  - 词汇特征
  - 结构特征
  - 语义特征

与知识库的关联



- 与知识库中的实体链接
- 面向知识库的概念匹配
- 面向知识库的关系抽取

与知识库关联

关键

实体链接

实体链接

- 费城 $\Longrightarrow$ 费城 *(城市)*、费城 *(电影)*、费城 *(街道)*, ...
- 问题通常较短、缺乏足够的上下文、命名实体边界模糊, ...

与知识库关联

关键

实体链接

实体链接

- 费城 $\Longrightarrow$ 费城 *(城市)*、费城 *(电影)*、费城 *(街道)*, ...
- 问题通常较短、缺乏足够的上下文、命名实体边界模糊, ...
- 需充分利用现有实体链接工具资源 (Heng Ji's EDL tools collection)

与知识库关联

关键

实体链接，关系抽取

实体链接

- 费城 $\Longrightarrow$ 费城 *(城市)*、费城 *(电影)*、费城 *(街道)*, ...
- 问题通常较短、缺乏足够的上下文、命名实体边界模糊, ...
- 需充分利用现有实体链接工具资源 (Heng Ji's EDL tools collection)

关系抽取/匹配

- 缺乏足够的上下文、表述方式灵活多样、候选关系众多、与知识库表述不匹配、易受错误传递影响, ...
- 复合关系表示，如Freebase中的CVT(compound value type)节点
  - 球队-球员-球衣号-场上位置，夫妻婚姻存续期, ...
- 多管齐下：关系抽取、通过实体猜测、联合消解, ...

联合消解歧义

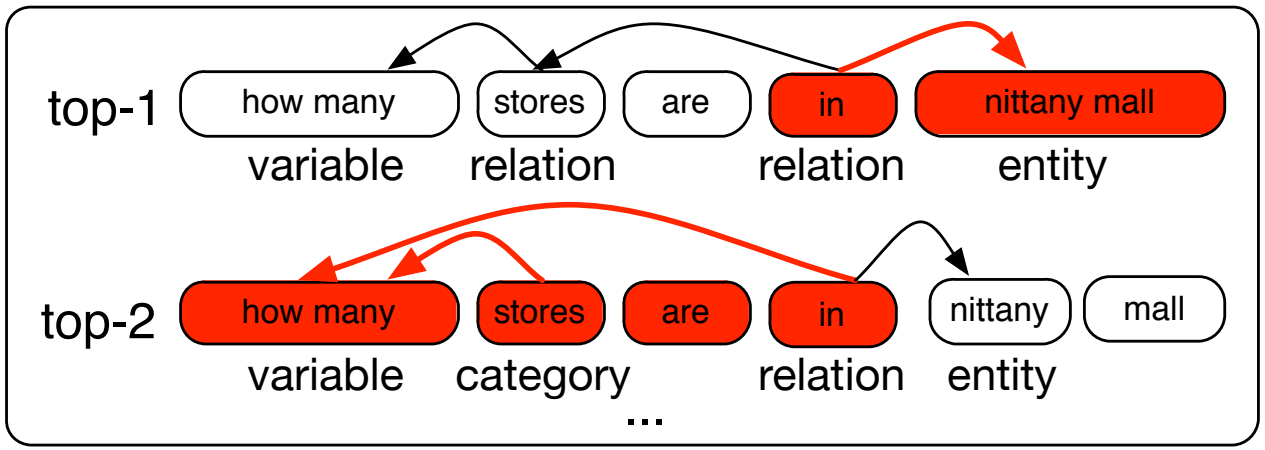
- 将子任务的中间结果暂存，多个子任务联合寻找最优结果
- 克服流水线框架中错误传递

[Xu et al., 2016]

联合消解歧义

- 将子任务的中间结果暂存，多个子任务联合寻找最优结果
- 克服流水线框架中错误传递

- 短语切分与短语依存结构抽取之间的联合消解

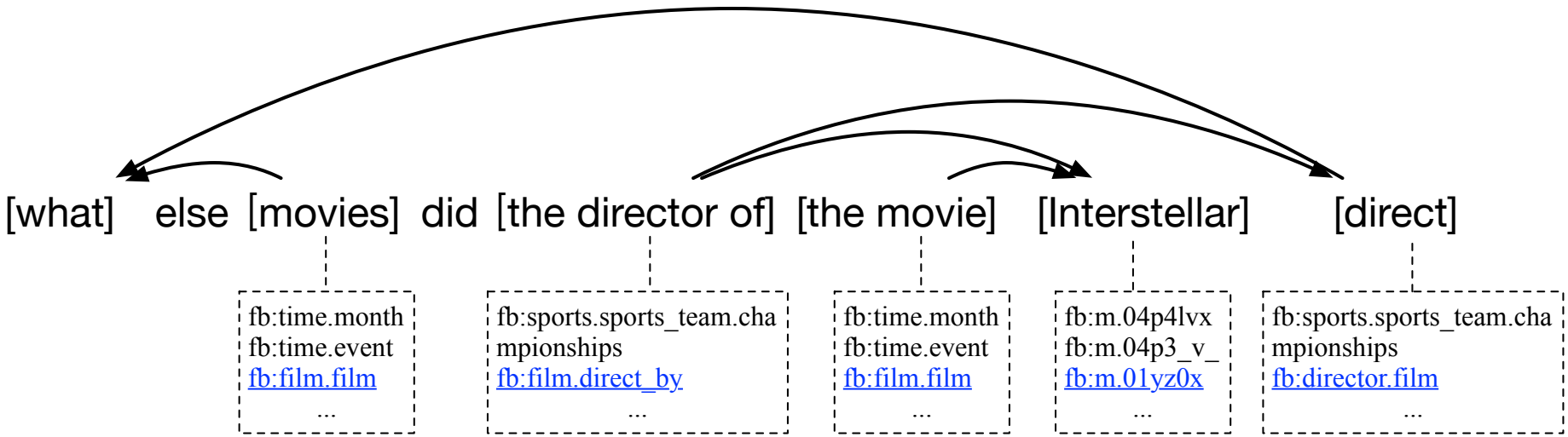


[Xu et al., 2016]

联合消解歧义

- 将子任务的中间结果暂存，多个子任务联合寻找最优结果
- 克服流水线框架中错误传递

- 实体链接与关系抽取之间的联合消解



[Xu et al., 2016]

结构化知识库之外：文本的作用

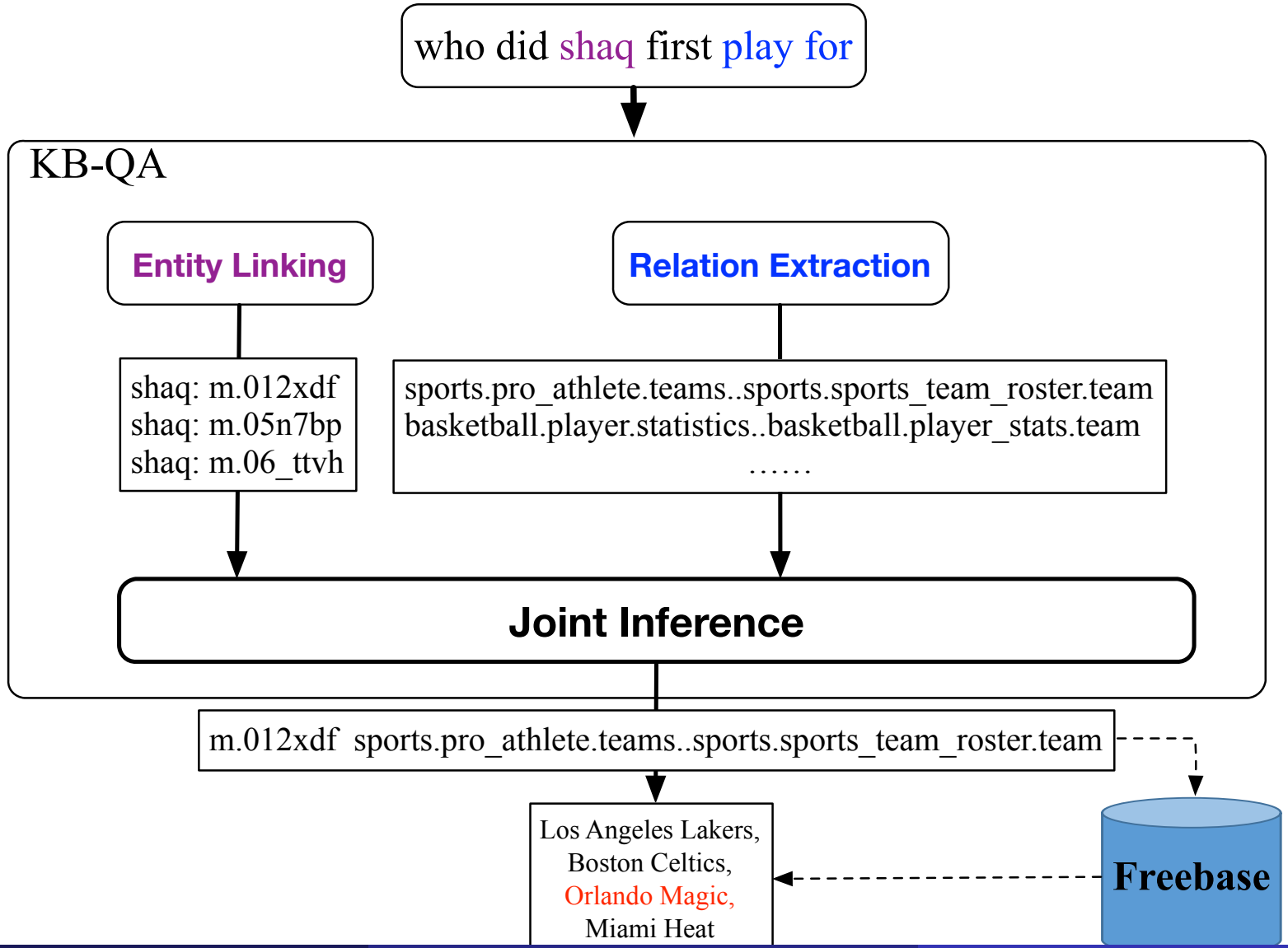
单独依靠知识库也许是不够的

- 难以显示的依靠知识库解析： *Who did Shaq first play for ?*
- 偏主观的表述方式： *What is the most popular movie in the past 5 years in UK ?*

[Zhang et al., 2015, Xu et al., 2016a, Xu et al., 2016b]

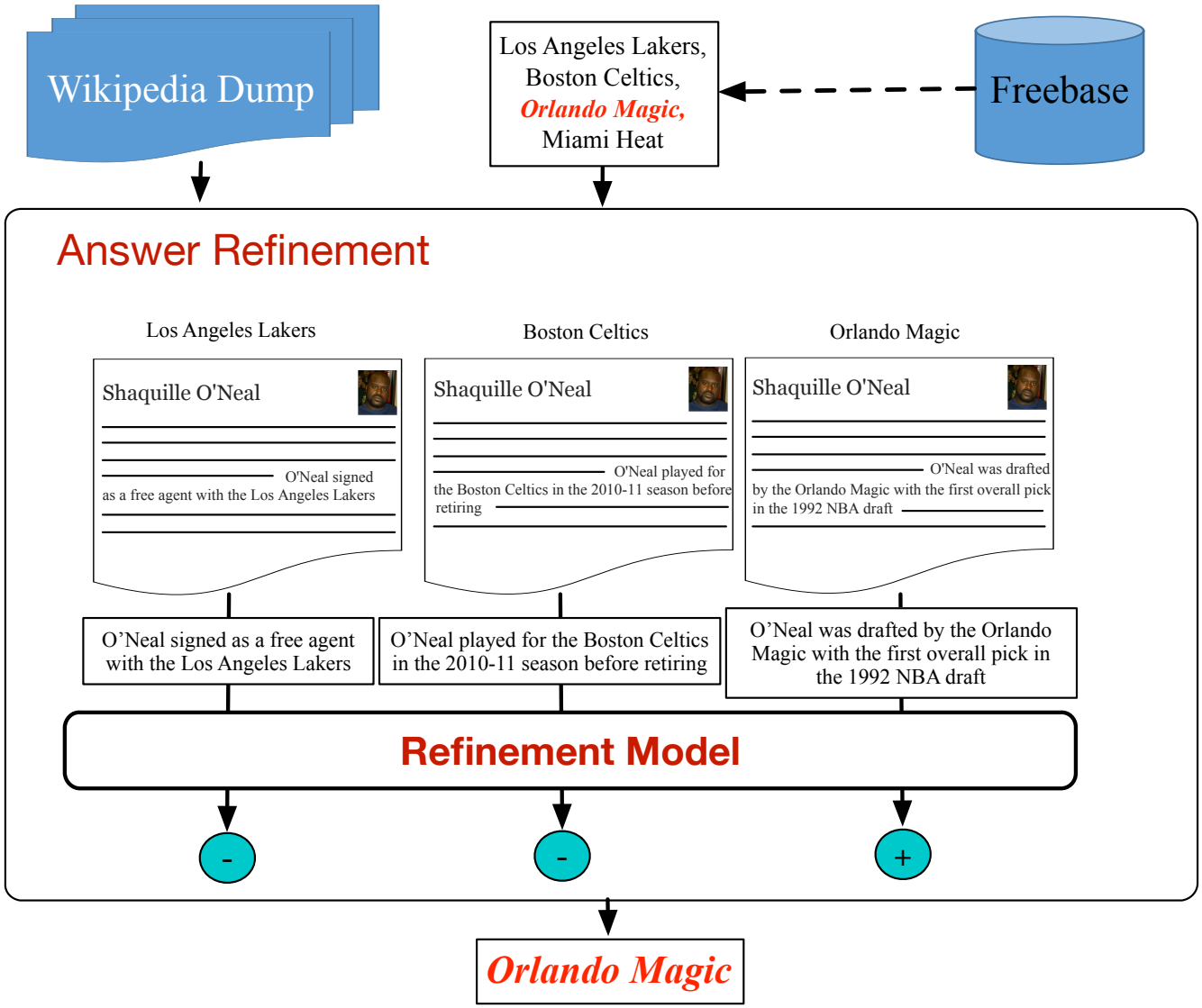
结构化知识库之外：文本的作用

单独依靠知识库也许是不够的$\Longrightarrow$ 实体与关系的联合消解



结构化知识库之外：文本的作用

单独依靠知识库也许是不够的 $\Longrightarrow$ 利用维基正文清洗候选答案



在WebQuestions上

Question & Answers
1. what is the largest nation in europe Before: Kazakhstan, Turkey, Russia, ... After: Russia
2. which country in europe has the largest land area Before: Georgia, France, Russia, ... After: Russian Empire, Russia
3. what year did ray allen join the nba Before: 2007, 2003, 1996, 1993, 2012 After: 1996
4. who is emma stone father Before: Jeff Stone, Krista Stone After: Jeff Stone
5. where did john steinbeck go to college Before: Salinas High School, Stanford University After: Stanford University

WebQuestions 上的结果

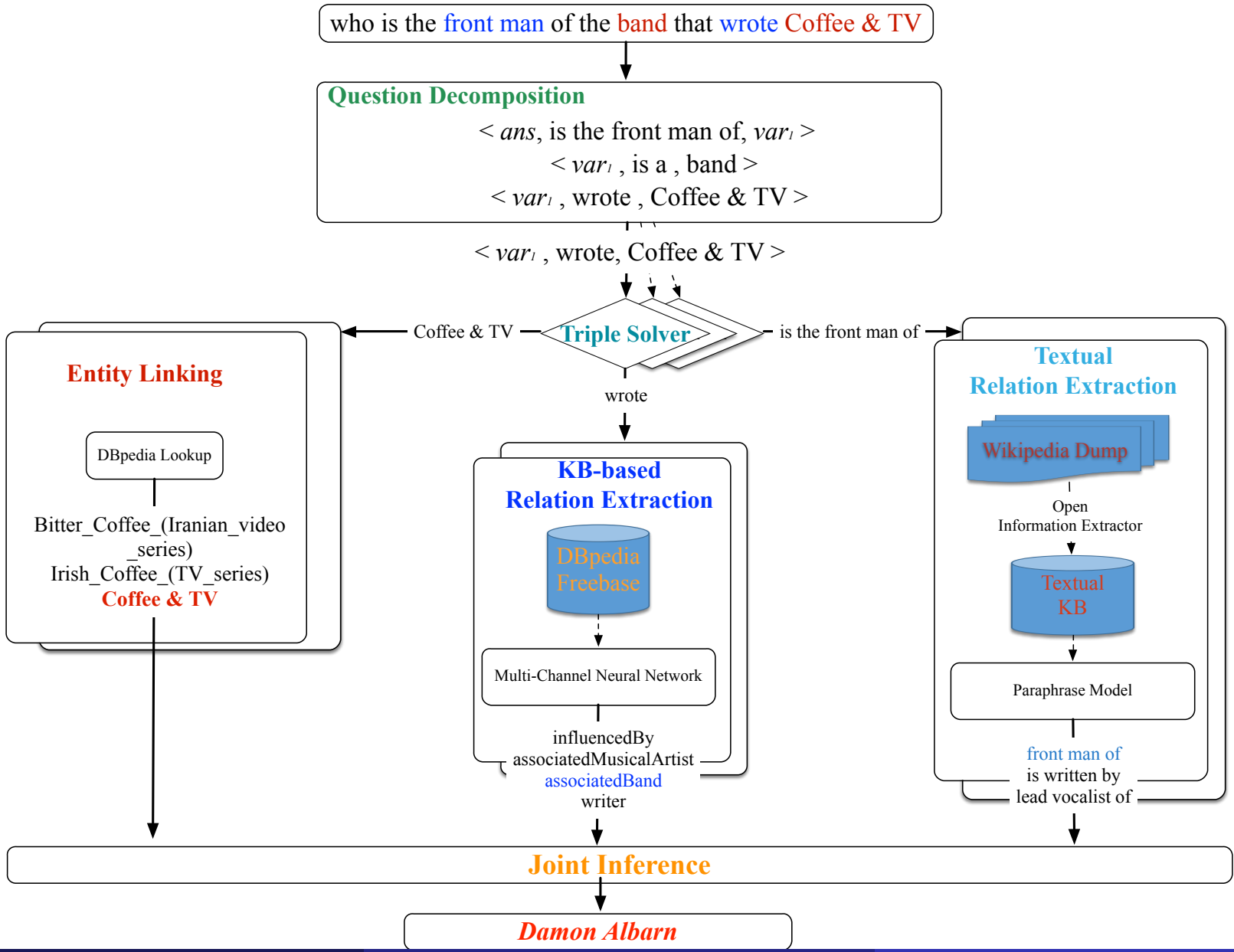
Method	average F_1
Berant et al. (2013)	35.7
Yao et al. (2014)	33.0
Xu et al. (2014)	39.1
Berant et al. (2014)	39.9
Bao et al. (2014)	37.5
Dong et al. (2015)	40.8
Yao et al. (2015)	44.3
Bast et al. (2015)	49.4
Berant et al. (2015)	49.7
Reddy et al. (2016)	50.3
Yih et al. (2015)	52.5
EL/RE (KB)	44.1
EL/RE + Joint	47.1
EL/RE + Refine	47.0
EL/RE + Joint + Refine	**53.3**

Hybrid-QA: 基于混合资源的知识库问答

- 采用 疑问词—实体—关系 表示问题语义
- 同时考虑结构化知识库与维基正文来抽取规范化谓词和开放式关系
- 实体链接
- 利用实体关系抽取技术抽取规范化的知识库谓词关系：X *associatedBand* Y
- 利用开放信息抽取技术抽取开放式的文本关系：X *is written by* Y
- 利用整数线性规划来选择最佳组合，构成知识库查询

[Xu et al., 2016b]

Hybrid-QA


who is the front man of the band that wrote Coffee & TV
Question Decomposition
< ans, is the front man of, var1 >
< var1 , is a , band >
< var1 , wrote , Coffee & TV >
< var1 , wrote, Coffee & TV >
Coffee & TV
Triple Solver
is the front man of
wrote
Entity Linking
DBpedia Lookup
Bitter_Coffee_(Iranian_video_series)
Irish_Coffee_(TV_series)
Coffee & TV
KB-based Relation Extraction
DBpedia Freebase
Multi-Channel Neural Network
influencedBy
associatedMusicalArtist
associatedBand
writer
Textual Relation Extraction
Wikipedia Dump
Open Information Extractor
Textual KB
Paraphrase Model
front man of
is written by
lead vocalist of
Joint Inference
Damon Albarn

WebQuestions 上的结果

Method	average F_1
Berant et al. (2013)	35.7
Yao et al. (2014)	33.0
Xu et al. (2014)	39.1
Berant et al. (2014)	39.9
Bao et al. (2014)	37.5
Dong et al. (2015)	40.8
Yao et al. (2015)	44.3
Bast et al. (2015)	49.4
Berant et al. (2015)	49.7
Reddy et al. (2016)	50.3
Yih et al. (2015)	52.5
EL/RE (KB)	44.1
EL/RE + Joint	47.1
EL/RE + Refine	47.0
EL/RE + Joint + Refine	**53.3**
Hybrid (KB + Text + Joint)	**53.8**

在QALD-6上的结果

Method	WebQ	QALD
EL/RE (KB)	44.1	10.1
KB + Joint	47.1	14.3
Text	40.3	28.7
Text + Joint	45.5	37.4
KB + Text + Joint (Hybrid)	**53.8**	**40.9**

QALD-6
What is the most common language in norway
What currency do they use in switzerland
What countries does queen elizabeth ii reign
What is the best sandals resort in st lucia
WebQuestions
What is the largest city in the county in which
Where was the Father of Singapore born
Who is the architect of the tallest building in Japan

小结

- 选择适当的语义表示形式
  - 是否需要与知识库分离?
  - 现有基础分析工具是否足够好?
- 利用更丰富的资源，例如，文本，多知识库等
- 但是
  - 仍然依靠很多人工介入、规则、模板
- 关键
  - 自然语言与知识库之间的映射 （实体、关系）
  - 基于知识库的浅层推断
  - 利用常识

参考文献-1

- John Zelle and Raymond Mooney. Learning to parse database queries using inductive logic programming. In AAAI 1996
- Luke S Zettlemoyer and Michael Collins. Learning to map sentences to logical form: Structured classification with probabilistic categorial grammars. In UAI 2005
- Yuk Wah Wong and Raymond J Mooney. Learning synchronous grammars for semantic pars- ing with lambda calculus. In ACL 2007
- Percy Liang, Michael I Jordan, and Dan Klein. Learning dependency-based compositional semantics. Computational Linguistics, 2013.
- Qingqing Cai and Alexander Yates. Large-scale semantic parsing via schema matching and lexicon extension. In ACL 2013
- Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. Semantic parsing on freebase from question-answer pairs. In EMNLP 2013
- Tom Kwiatkowski, Eunsol Choi, Yoav Artzi, and Luke S. Zettlemoyer. Scaling semantic parsers with on-the-fly ontology matching. In EMNLP 2013

参考文献-2

- Siva Reddy, Mirella Lapata, and Mark Steedman. Large-scale semantic parsing without question-answer pairs. In TACL 2014
- Xuchen Yao and Benjamin Van Durme. 2014. Information extraction over structured data: Question answering with freebase. In ACL 2014
- Jonathan Berant and Percy Liang. Semantic parsing via paraphrasing. In ACL 2014
- Junwei Bao, Nan Duan, Ming Zhou, and Tiejun Zhao. Knowledge-based question answering as ma- chine translation. In ACL 2014
- Kun Xu, Sheng Zhang, Yansong Feng, and Dongyan Zhao. Answering natural language questions via phrasal semantic parsing. In NLPCC 2014.
- Antoine Bordes, Sumit Chopra, and Jason Weston. Question answering with subgraph embed- dings. In EMNLP 2014
- Jonathan Berant and Percy Liang. Imitation learning of agenda-based semantic parsers. TACL 2015.

参考文献-3

- Antoine Bordes, Nicolas Usunier, Sumit Chopra, and Jason Weston. Large-scale simple question answering with memory networks. CoRR, abs/1506.02075.
- Wen-tau Yih, Ming-Wei Chang, Xiaodong He, and Jianfeng Gao. Semantic parsing via staged query graph generation: Question answering with knowledge base. In ACL-IJCNLP 2015.
- Li Dong, Furu Wei, Ming Zhou, and Ke Xu. Question answering over freebase with multi- column convolutional neural networks. In ACL-IJCNLP 2015
- Yi Yang and Ming-Wei Chang. S-MART: Novel tree-based structured learning algorithms applied to tweet entity linking. In ACL- IJCNLP 2015
- Hannah Bast and Elmar Haussmann. More accurate question answering on Freebase. In CIKM 2015
- Kun Xu, Siva Reddy, Yansong Feng, Songfang Huang, and Dongyan Zhao. Question Answering on Freebase via Relation Extraction and Textual Evidence. In ACL 2016

参考文献-4

- Wen-tau Yih, Matthew Richardson, Christopher Meek, Ming-Wei Chang, and Jina Suh. The Value of Semantic Parse Labeling for Knowledge Base Question Answering, In ACL 2016
- Kun Xu, Yansong Feng, Songfang Huang, and Dongyan Zhao. Hybrid Question Answering over Knowledge Base and Free Text. In COLING 2016
- Siva Reddy, Oscar Tackstrom, Michael Collins, Tom Kwiatkowski, Dipanjan Das, Mark Steedman, and Mirella Lapata. Transforming Dependency Structures to Logical Forms for Semantic Parsing. In TACL 2016
- Wen-tau Yih and Hao Ma, Question Answering with Knowledge Bases, Web and Beyond, In NAACL 2016, Tutorial