

# Lecture 12: Information Extraction

Alan Ritter

(many slides from Greg Durrett)

# This Lecture

---

- ▶ How do we represent information for information extraction?
- ▶ Semantic role labeling / abstract meaning representation
- ▶ Relation extraction
- ▶ Slot filling
- ▶ Open Information Extraction

# Representing Information

# Semantic Representations

---

- ▶ “World” is a set of entities and predicates

person
Brutus
Caesar
Obama
Bush
...

president
Obama
Bush
...

stab
Brutus Caesar
...

# Semantic Representations

---

- ▶ “World” is a set of entities and predicates

person
Brutus
Caesar
Obama
Bush
...

president
Obama
Bush
...

stab
Brutus Caesar
...

- ▶ Statements are logical expressions that evaluate to true or false

# Semantic Representations

---

- ▶ “World” is a set of entities and predicates

person
Brutus
Caesar
Obama
Bush
...

president
Obama
Bush
...

stab
Brutus    Caesar
...

- ▶ Statements are logical expressions that evaluate to true or false

*Brutus stabs Caesar*

# Semantic Representations

---

- ▶ “World” is a set of entities and predicates

person
Brutus
Caesar
Obama
Bush
...

president
Obama
Bush
...

stab
Brutus    Caesar
...

- ▶ Statements are logical expressions that evaluate to true or false

*Brutus stabs Caesar*

`stab(Brutus, Caesar) => true`

# Semantic Representations

---

- ▶ “World” is a set of entities and predicates

person
Brutus
Caesar
Obama
Bush
...

president
Obama
Bush
...

stab
Brutus    Caesar
...

- ▶ Statements are logical expressions that evaluate to true or false

*Brutus stabs Caesar*

`stab(Brutus, Caesar) => true`

*Caesar was stabbed*



# Semantic Representations

---

- ▶ “World” is a set of entities and predicates

person	president	stab
Brutus	Obama	Brutus Caesar
Caesar	Bush	...
Obama	...	
Bush		
...		

- ▶ Statements are logical expressions that evaluate to true or false

*Brutus stabs Caesar*

$\text{stab}(\text{Brutus}, \text{Caesar}) \Rightarrow \text{true}$

*Caesar was stabbed*

$\exists x \text{stab}(x, \text{Caesar}) \Rightarrow \text{true}$

# Neo-Davidsonian Events

---

*Brutus stabbed Caesar with a knife at the theater on the Ides of March*

# Neo-Davidsonian Events

---

*Brutus stabbed Caesar with a knife at the theater on the Ides of March*

$\exists e \text{ stabs}(e, \text{Brutus}, \text{Caesar})$

# Neo-Davidsonian Events

---

*Brutus stabbed Caesar with a knife at the theater on the Ides of March*

$\exists e \text{ stabs}(e, \text{Brutus}, \text{Caesar}) \wedge \text{with}(e, \text{knife})$

# Neo-Davidsonian Events

---

*Brutus stabbed Caesar with a knife at the theater on the Ides of March*

$\exists e \text{ stabs}(e, \text{Brutus}, \text{Caesar}) \wedge \text{with}(e, \text{knife}) \wedge \text{location}(e, \text{theater})$

# Neo-Davidsonian Events

---

*Brutus stabbed Caesar with a knife at the theater on the Ides of March*

$\exists e \text{ stabs}(e, \text{Brutus}, \text{Caesar}) \wedge \text{with}(e, \text{knife}) \wedge \text{location}(e, \text{theater})$   
 $\wedge \text{time}(e, \text{Ides of March})$

# Neo-Davidsonian Events

---

*Brutus stabbed Caesar with a knife at the theater on the Ides of March*

$\exists e \text{ stabs}(e, \text{Brutus}, \text{Caesar}) \wedge \text{with}(e, \text{knife}) \wedge \text{location}(e, \text{theater})$   
 $\wedge \text{time}(e, \text{Ides of March})$

- ▶ Lets us describe events as having properties

# Neo-Davidsonian Events

---

*Brutus stabbed Caesar with a knife at the theater on the Ides of March*

$\exists e \text{ stabs}(e, \text{Brutus}, \text{Caesar}) \wedge \text{with}(e, \text{knife}) \wedge \text{location}(e, \text{theater})$   
 $\wedge \text{time}(e, \text{Ides of March})$

- ▶ Lets us describe events as having properties
- ▶ Unified representation of events and entities:



# Neo-Davidsonian Events

---

*Brutus stabbed Caesar with a knife at the theater on the Ides of March*

$\exists e \text{ stabs}(e, \text{Brutus}, \text{Caesar}) \wedge \text{with}(e, \text{knife}) \wedge \text{location}(e, \text{theater})$   
 $\wedge \text{time}(e, \text{Ides of March})$

- ▶ Lets us describe events as having properties
- ▶ Unified representation of events and entities:

*some clever driver in America*

# Neo-Davidsonian Events

---

*Brutus stabbed Caesar with a knife at the theater on the Ides of March*

$\exists e \text{ stabs}(e, \text{Brutus}, \text{Caesar}) \wedge \text{with}(e, \text{knife}) \wedge \text{location}(e, \text{theater})$   
 $\wedge \text{time}(e, \text{Ides of March})$

- ▶ Lets us describe events as having properties
- ▶ Unified representation of events and entities:

*some clever driver in America*

$\exists x \text{ driver}(x) \wedge \text{clever}(x) \wedge \text{location}(x, \text{America})$

# Real Text

---

*Barack Obama signed the Affordable Care act on Tuesday. He gave a speech later that afternoon on how the act would help the American people. Several prominent Republicans were quick to denounce the new law.*

# Real Text

---

*Barack Obama signed the Affordable Care act on Tuesday. He gave a speech later that afternoon on how the act would help the American people. Several prominent Republicans were quick to denounce the new law.*

$\exists e \text{ sign}(e, \text{Barack Obama}) \wedge \text{patient}(e, \text{ACA}) \wedge \text{time}(e, \text{Tuesday})$

# Real Text

---

*Barack Obama signed the Affordable Care act on Tuesday. He gave a speech later that afternoon on how the act would help the American people. Several prominent Republicans were quick to denounce the new law.*

which Tuesday?

$\exists e \text{ sign}(e, \text{Barack Obama}) \wedge \text{patient}(e, \text{ACA}) \wedge \text{time}(e, \text{Tuesday})$

# Real Text

---

who?

*Barack Obama signed the Affordable Care act on Tuesday. He gave a speech later that afternoon on how the act would help the American people. Several prominent Republicans were quick to denounce the new law.*

which Tuesday?

$\exists e \text{ sign}(e, \text{Barack Obama}) \wedge \text{patient}(e, \text{ACA}) \wedge \text{time}(e, \text{Tuesday})$

# Real Text

---

which afternoon?

who?

*Barack Obama signed the Affordable Care act on Tuesday. He gave a speech later that afternoon on how the act would help the American people. Several prominent Republicans were quick to denounce the new law.*

which Tuesday?

$\exists e \text{ sign}(e, \text{Barack Obama}) \wedge \text{patient}(e, \text{ACA}) \wedge \text{time}(e, \text{Tuesday})$

# Real Text

---

which afternoon?

who?

*Barack Obama signed the Affordable Care act on Tuesday. He gave a speech later that afternoon on how the act would help the American people. Several prominent Republicans were quick to denounce the new law.*

???

which Tuesday?

$\exists e \text{ sign}(e, \text{Barack Obama}) \wedge \text{patient}(e, \text{ACA}) \wedge \text{time}(e, \text{Tuesday})$



# Real Text

which afternoon?

who?

*Barack Obama signed the Affordable Care act on Tuesday. He gave a speech later that afternoon on how the act would help the American people. Several prominent Republicans were quick to denounce the new law.*

???

which Tuesday?

$\exists e \text{ sign}(e, \text{Barack Obama}) \wedge \text{patient}(e, \text{ACA}) \wedge \text{time}(e, \text{Tuesday})$

- Need to impute missing information, resolve coreference, etc.

# Real Text

which afternoon?

who?

*Barack Obama signed the Affordable Care act on Tuesday. He gave a speech later that afternoon on how the act would help the American people. Several prominent Republicans were quick to denounce the new law.*

???

which Tuesday?

$\exists e \text{ sign}(e, \text{Barack Obama}) \wedge \text{patient}(e, \text{ACA}) \wedge \text{time}(e, \text{Tuesday})$

- ▶ Need to impute missing information, resolve coreference, etc.
- ▶ Still unclear how to represent some things precisely or how that information could be leveraged (several prominent Republicans)

# Other Challenges

---

*Bob and Alice were friends until he moved away to attend college*

# Other Challenges

---

*Bob and Alice were friends until he moved away to attend college*

$\exists e1 \exists e2 \text{ friends}(e1, \text{Bob}, \text{Alice}) \wedge \text{moved}(e2, \text{Bob}) \wedge \text{end\_of}(e1, e2)$

# Other Challenges

---

*Bob and Alice were friends until he moved away to attend college*

$\exists e1 \exists e2 \text{ friends}(e1, \text{Bob}, \text{Alice}) \wedge \text{moved}(e2, \text{Bob}) \wedge \text{end\_of}(e1, e2)$

- ▶ How to represent temporal information?

# Other Challenges

---

*Bob and Alice were friends until he moved away to attend college*

$\exists e1 \exists e2 \text{ friends}(e1, \text{Bob}, \text{Alice}) \wedge \text{moved}(e2, \text{Bob}) \wedge \text{end\_of}(e1, e2)$

- ▶ How to represent temporal information?

*Bob and Alice were friends until **around the time** he moved away to attend college*

# Other Challenges

---

*Bob and Alice were friends until he moved away to attend college*

$\exists e1 \exists e2 \text{ friends}(e1, \text{Bob}, \text{Alice}) \wedge \text{moved}(e2, \text{Bob}) \wedge \text{end\_of}(e1, e2)$

- ▶ How to represent temporal information?

*Bob and Alice were friends until **around the time** he moved away to attend college*

- ▶ Representing truly open-domain information is very complicated! We don't have a formal representation that can capture everything

# (At least) Three Solutions

---



# (At least) Three Solutions

---

- ▶ Crafted annotations to capture some subset of phenomena: predicate-argument structures (semantic role labeling), time (temporal relations), ...

# (At least) Three Solutions

---

- ▶ Crafted annotations to capture some subset of phenomena: predicate-argument structures (semantic role labeling), time (temporal relations), ...
- ▶ Slot filling: specific ontology, populate information in a predefined way

# (At least) Three Solutions

---

- ▶ Crafted annotations to capture some subset of phenomena: predicate-argument structures (semantic role labeling), time (temporal relations), ...
- ▶ Slot filling: specific ontology, populate information in a predefined way

(Earthquake: magnitude=8.0, epicenter=central Italy, ...)

# (At least) Three Solutions

---

- ▶ Crafted annotations to capture some subset of phenomena: predicate-argument structures (semantic role labeling), time (temporal relations), ...
- ▶ Slot filling: specific ontology, populate information in a predefined way  
(Earthquake: magnitude=8.0, epicenter=central Italy, ...)
- ▶ Entity-relation-entity triples: focus on entities and their relations (note that prominent events can still be entities)

# (At least) Three Solutions

---

- ▶ Crafted annotations to capture some subset of phenomena: predicate-argument structures (semantic role labeling), time (temporal relations), ...
- ▶ Slot filling: specific ontology, populate information in a predefined way

(Earthquake: magnitude=8.0, epicenter=central Italy, ...)

- ▶ Entity-relation-entity triples: focus on entities and their relations (note that prominent events can still be entities)

(Lady Gaga, singerOf, Bad Romance)

# Open IE

---

- ▶ Entity-relation-entity triples aren't necessarily grounded in an ontology
- ▶ Extract strings and let a downstream system figure it out

# Open IE

---

- ▶ Entity-relation-entity triples aren't necessarily grounded in an ontology
- ▶ Extract strings and let a downstream system figure it out

*Barack Obama signed the Affordable Care act on Tuesday. He gave a speech later that afternoon on how the act would help the American people. Several prominent Republicans were quick to denounce the new law.*

# Open IE

---

- ▶ Entity-relation-entity triples aren't necessarily grounded in an ontology
- ▶ Extract strings and let a downstream system figure it out

*Barack Obama signed the Affordable Care act on Tuesday. He gave a speech later that afternoon on how the act would help the American people. Several prominent Republicans were quick to denounce the new law.*

(Barack Obama, signed, the Affordable Care act)



# Open IE

---

- ▶ Entity-relation-entity triples aren't necessarily grounded in an ontology
- ▶ Extract strings and let a downstream system figure it out

*Barack Obama signed the Affordable Care act on Tuesday. He gave a speech later that afternoon on how the act would help the American people. Several prominent Republicans were quick to denounce the new law.*

(Barack Obama, signed, the Affordable Care act)

(Several prominent Republicans, denounce, the new law)

# IE: The Big Picture

---

- ▶ How do we represent information? What do we extract?
  - ▶ Semantic roles
  - ▶ Abstract meaning representation
  - ▶ Slot fillers
  - ▶ Entity-relation-entity triples (fixed ontology or open)

# Semantic Role Labeling/ Abstract Meaning Representation

# Semantic Role Labeling

---

# Semantic Role Labeling

## Gold

# ARG1



# ARG2

# ARG3

Housing starts are expected to quicken a bit from August's pace

Figure from He et al. (2017)

# Semantic Role Labeling

---

- Identify predicate, disambiguate it, identify that predicate's arguments

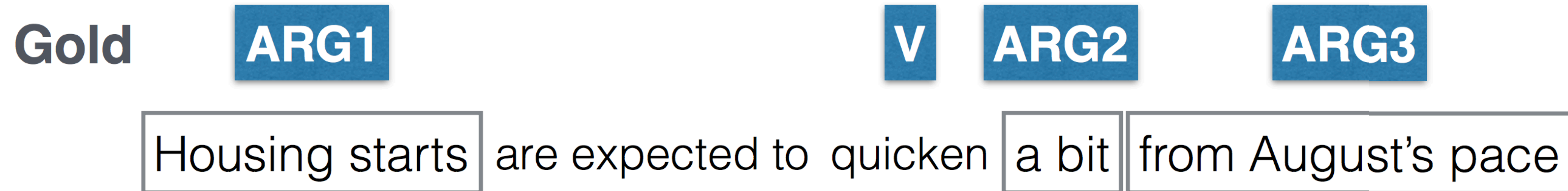


Figure from He et al. (2017)

# Semantic Role Labeling

---

- ▶ Identify predicate, disambiguate it, identify that predicate's arguments
- ▶ Verb roles from Propbank (Palmer et al., 2005)

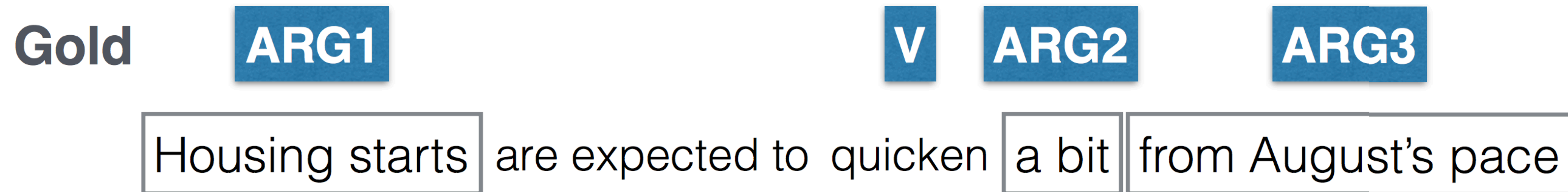


Figure from He et al. (2017)

# Semantic Role Labeling

- ▶ Identify predicate, disambiguate it, identify that predicate's arguments
- ▶ Verb roles from Propbank (Palmer et al., 2005)

# Gold

# ARG1

V

# ARG2

# ARG3

Housing starts are expected to quicken a bit from August's pace

# quicken:

**Arg0-PAG:** *causer of speed-up*

**Arg1-PPT:** *thing becoming faster* (vnrole: 45.4-patient)

**Arg2-EXT: EXT**

### Arg3-DIR: *old speed*

## Arg4-PRD: *new speed*

Figure from He et al. (2017)



# Semantic Role Labeling

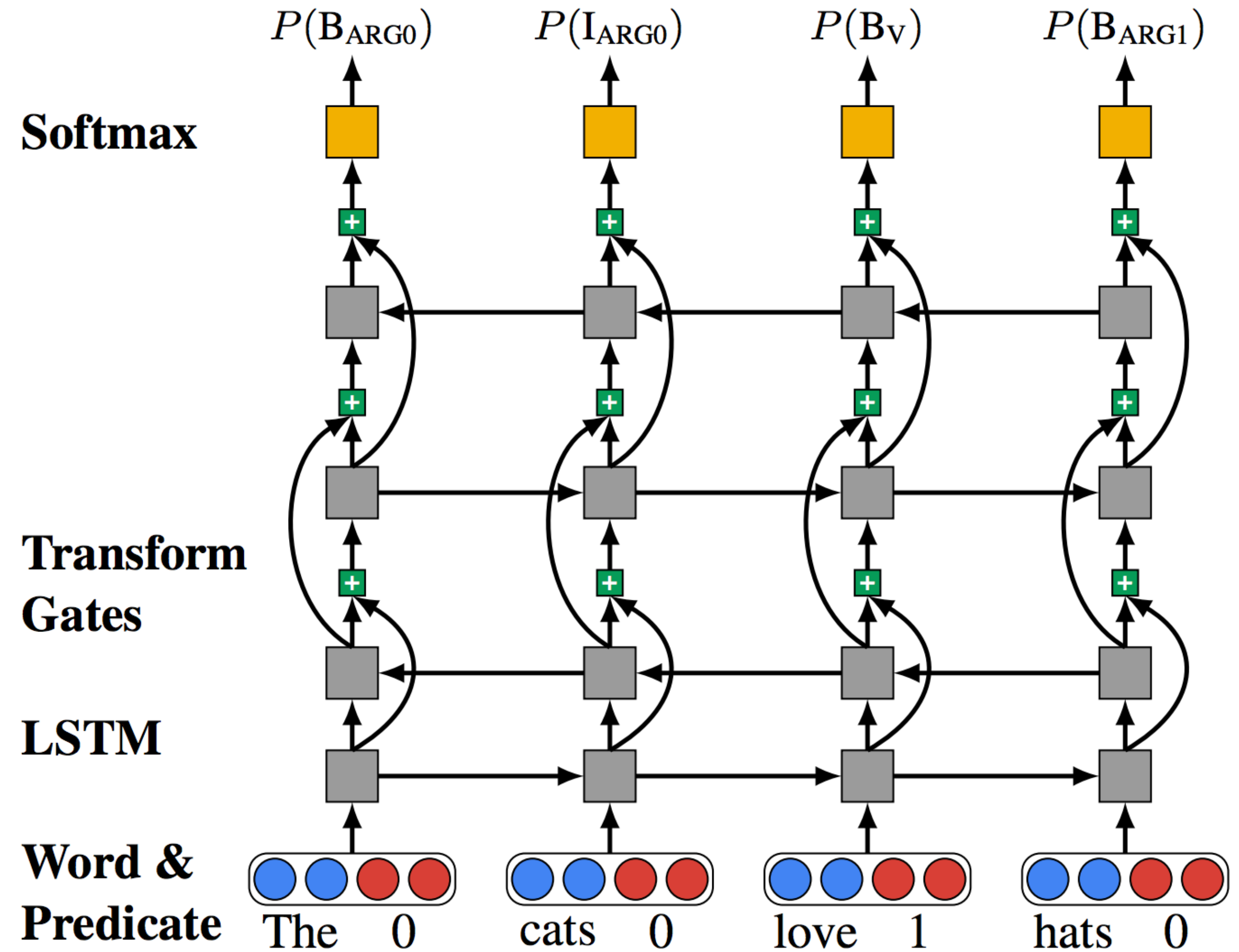


Figure from He et al. (2017)

# Semantic Role Labeling

- Identify predicates (*love*) using a classifier (not shown)

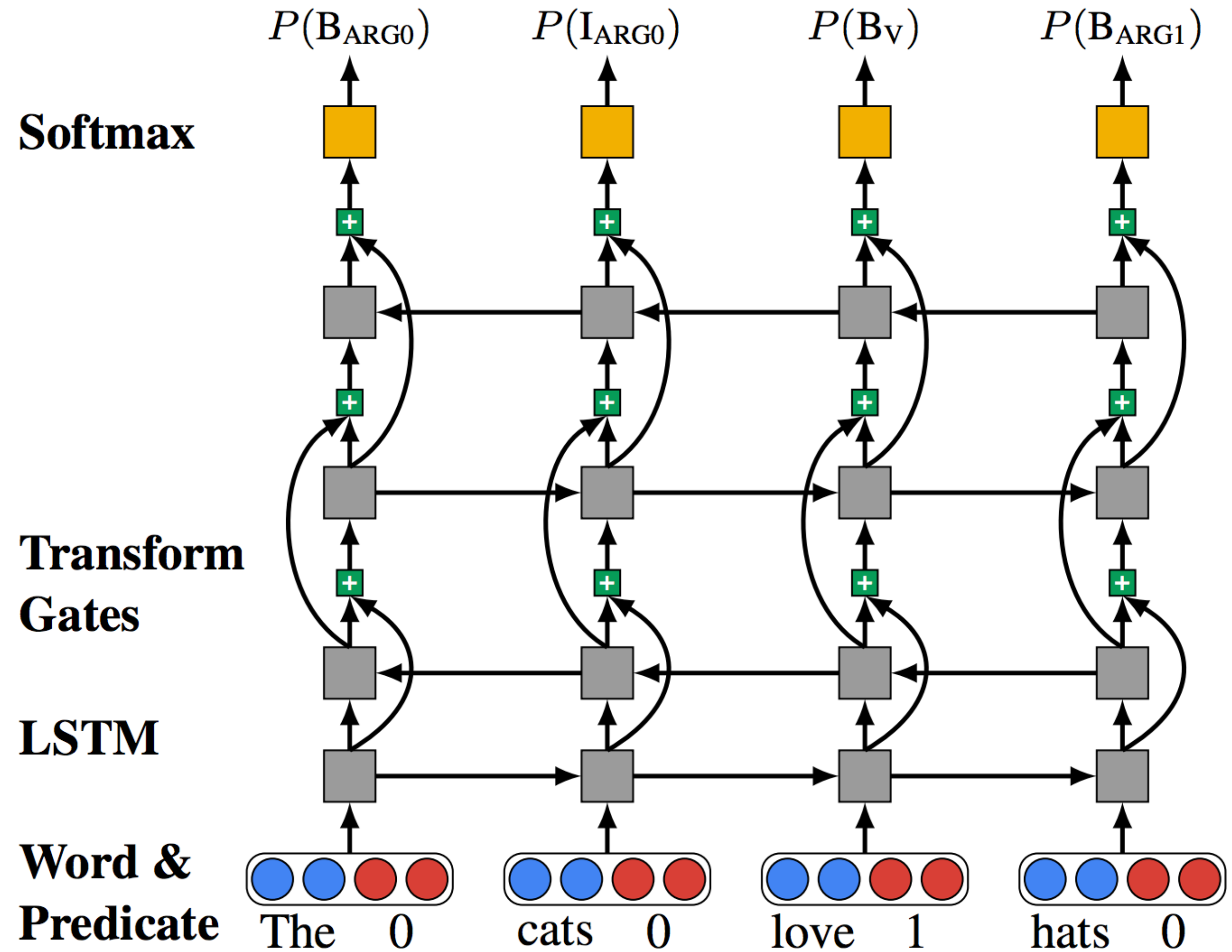


Figure from He et al. (2017)

# Semantic Role Labeling

- Identify predicates (*love*) using a classifier (not shown)
- Identify ARG0, ARG1, etc. as a tagging task with a BiLSTM conditioned on *love*

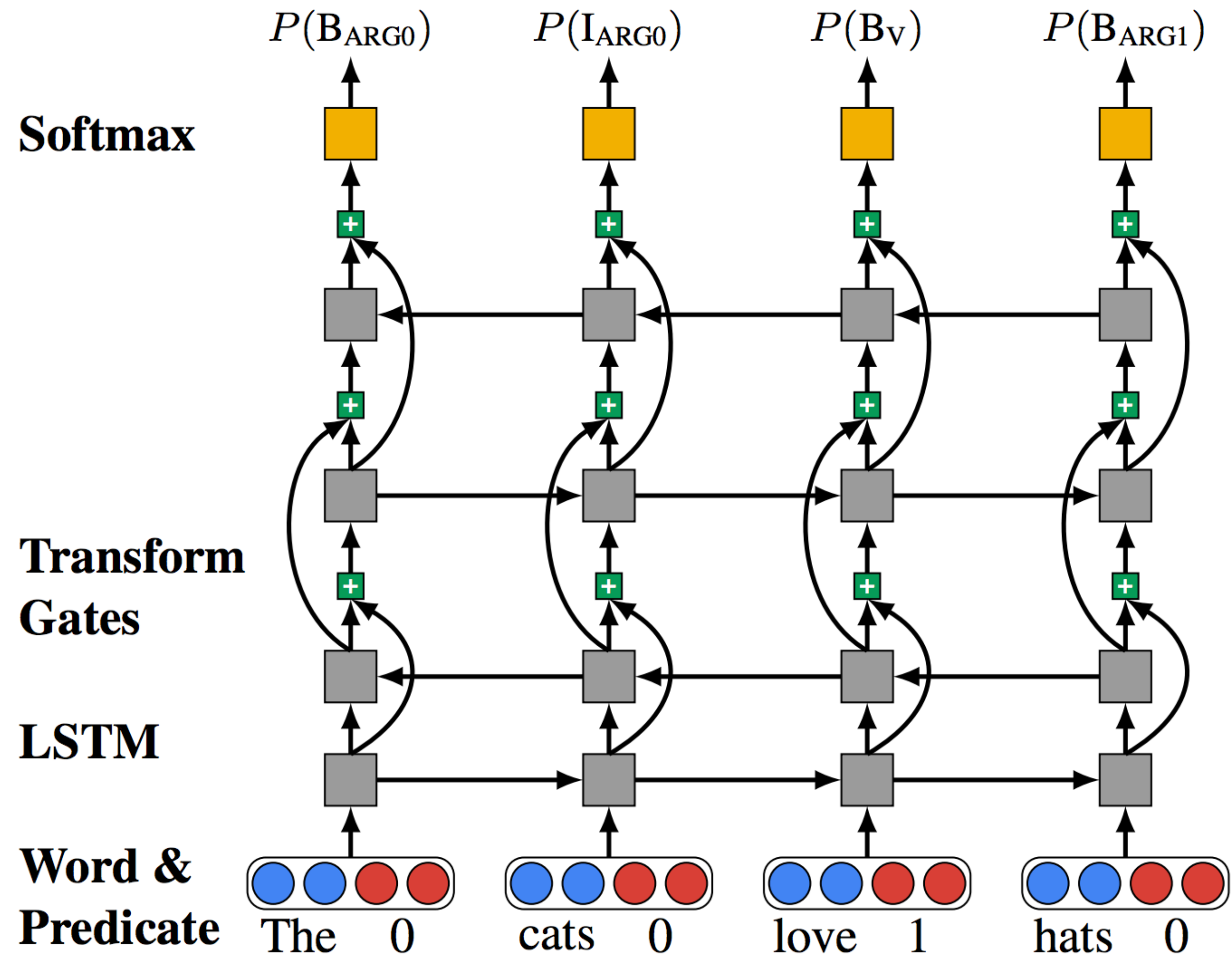


Figure from He et al. (2017)

# Semantic Role Labeling

- ▶ Identify predicates (*love*) using a classifier (not shown)
- ▶ Identify ARG0, ARG1, etc. as a tagging task with a BiLSTM conditioned on *love*
- ▶ Other systems incorporate syntax, joint predicate-argument finding

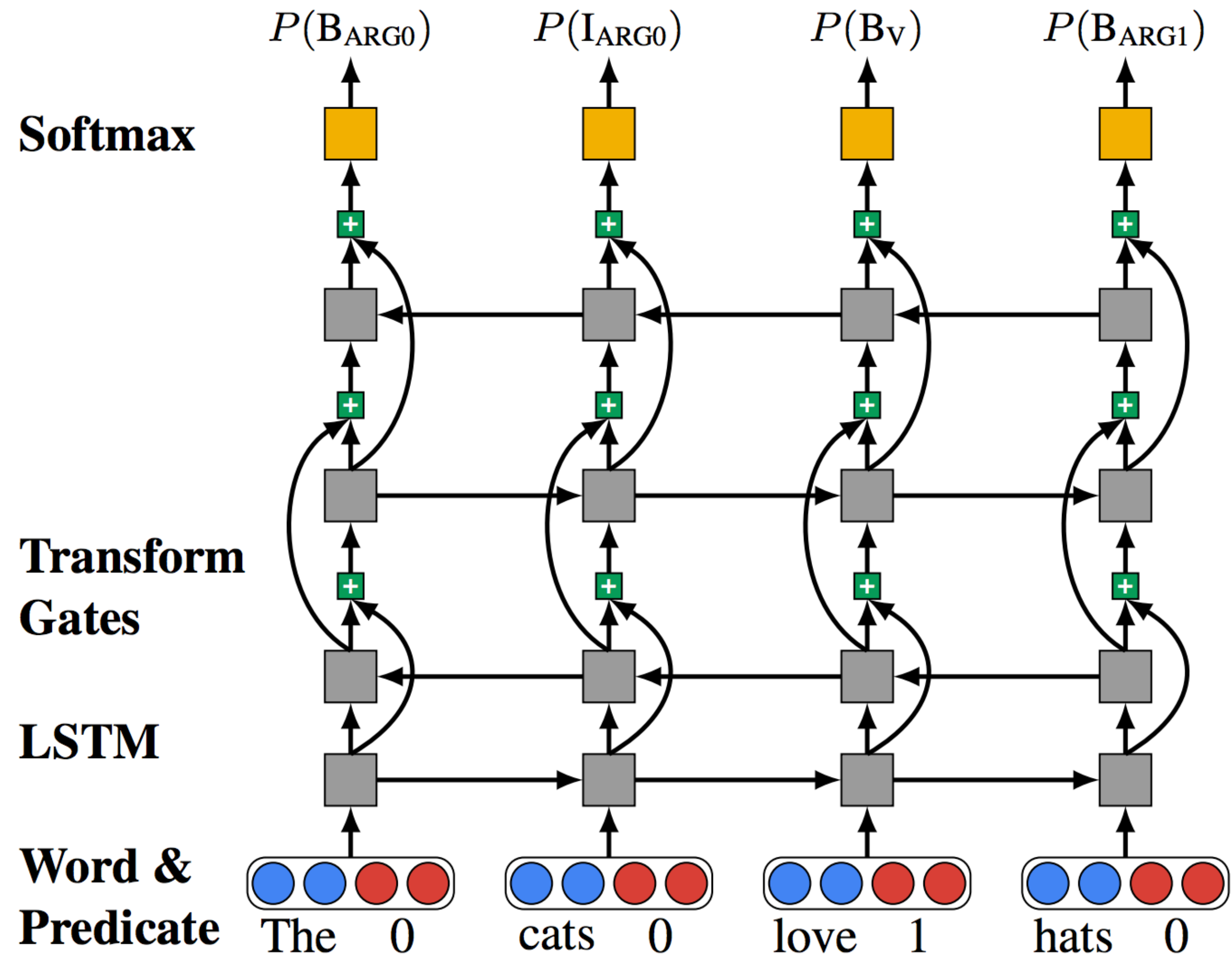


Figure from He et al. (2017)



# SRL for QA

---

- ▶ Question and several answer candidates

*Q: Who discovered prions?*

*AC1: In 1997, Stanley B. Prusiner, a scientist in the United States, discovered prions...*

*AC2: Prions were researched by...*

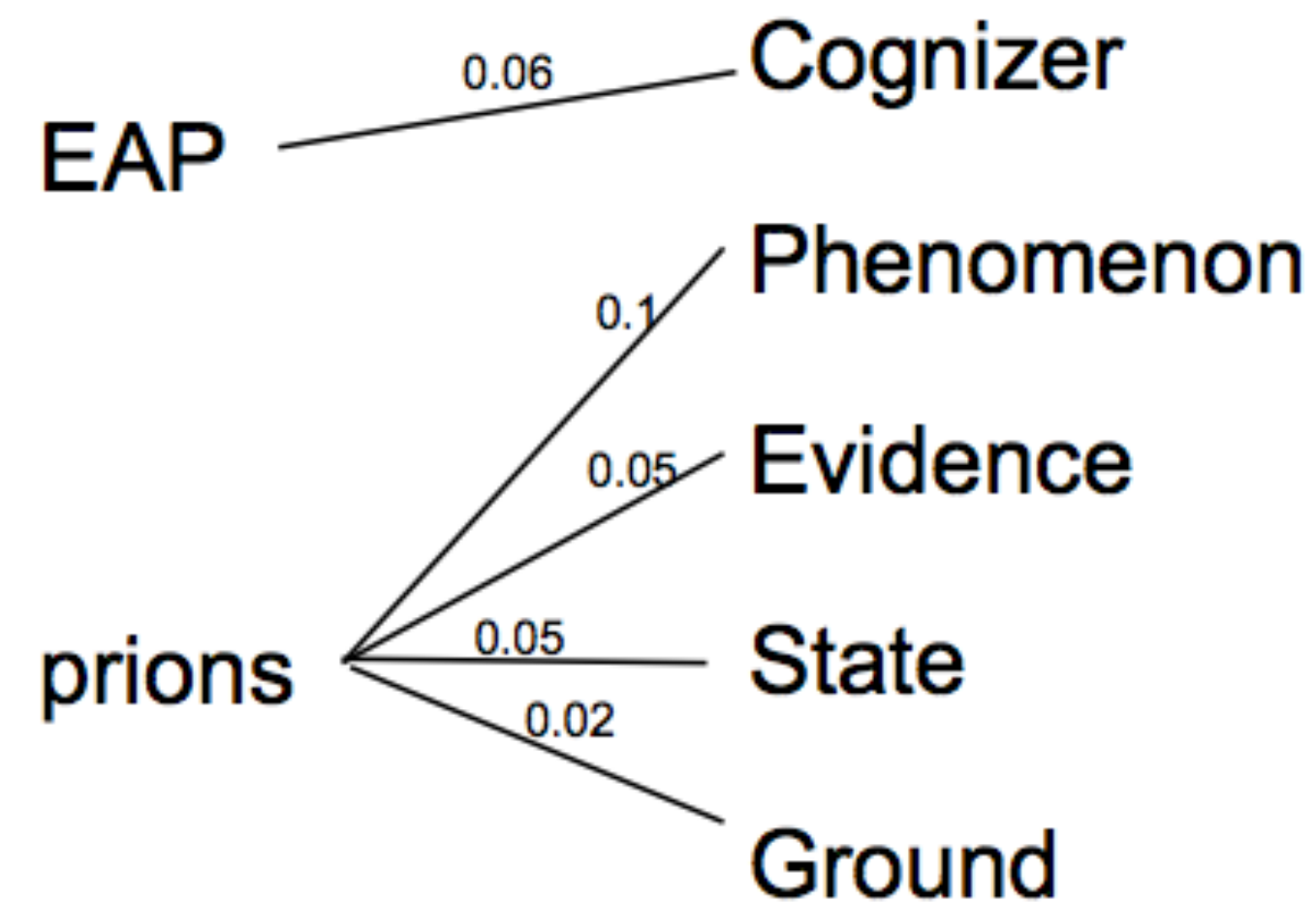
# SRL for QA

- Question and several answer candidates

Q: *Who discovered prions?* →

AC1: *In 1997, Stanley B. Prusiner, a scientist in the United States, discovered prions...*

AC2: *Prions were researched by...*



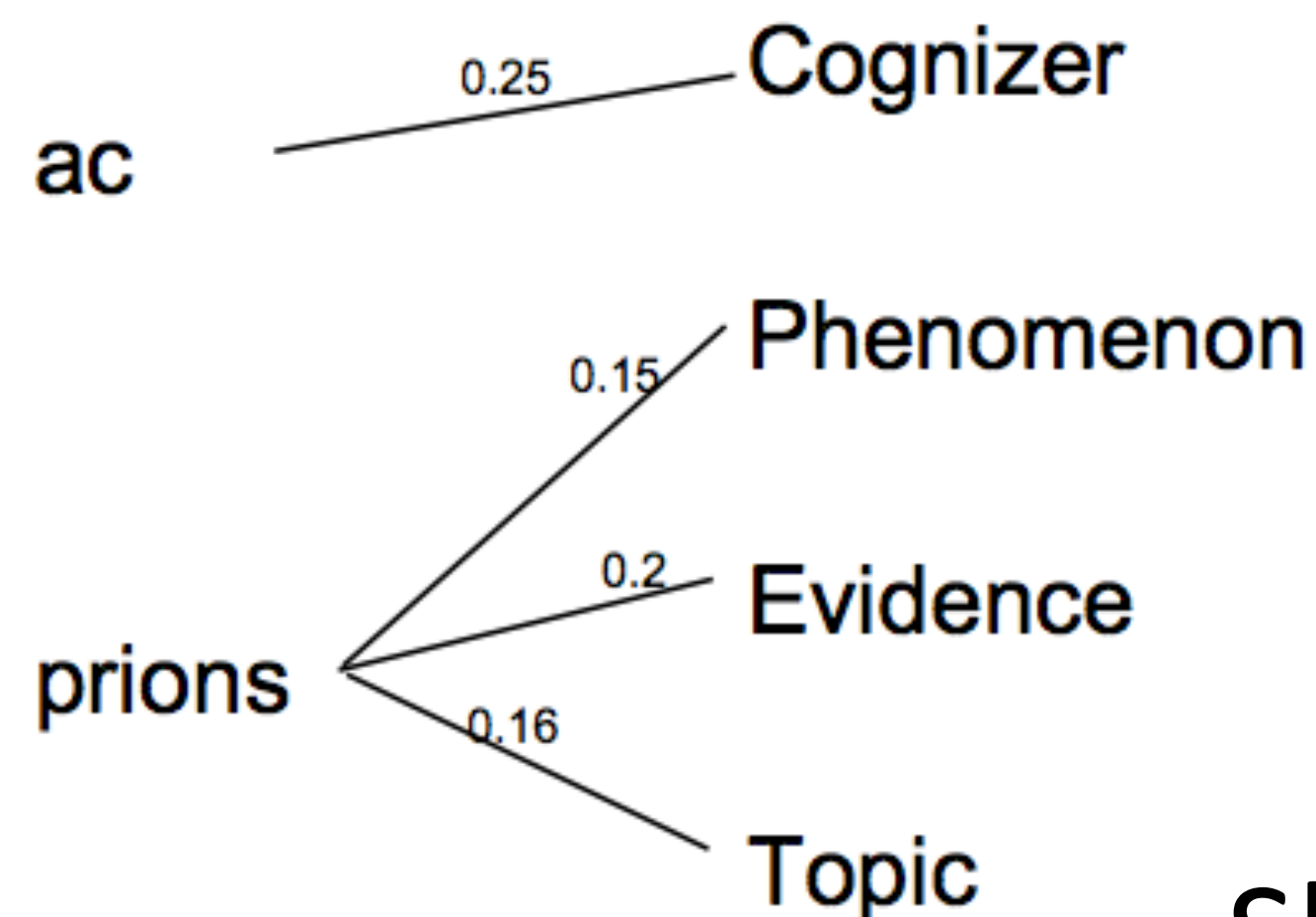
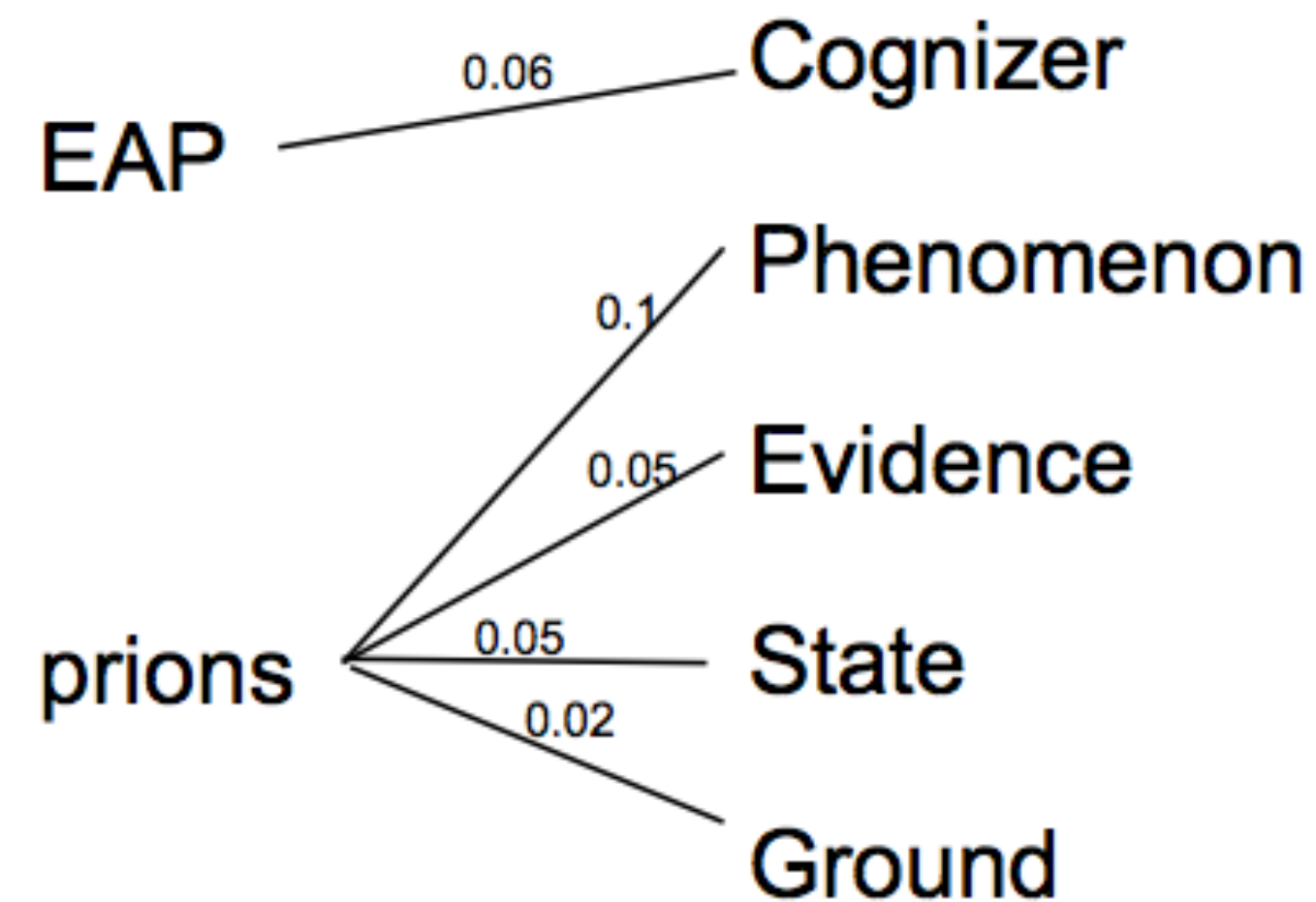
# SRL for QA

- Question and several answer candidates

Q: *Who discovered prions?* →

AC1: *In 1997, Stanley B. Prusiner, a scientist in the United States, discovered prions...*

AC2: *Prions were researched by...*



Shen and Lapata (2007)

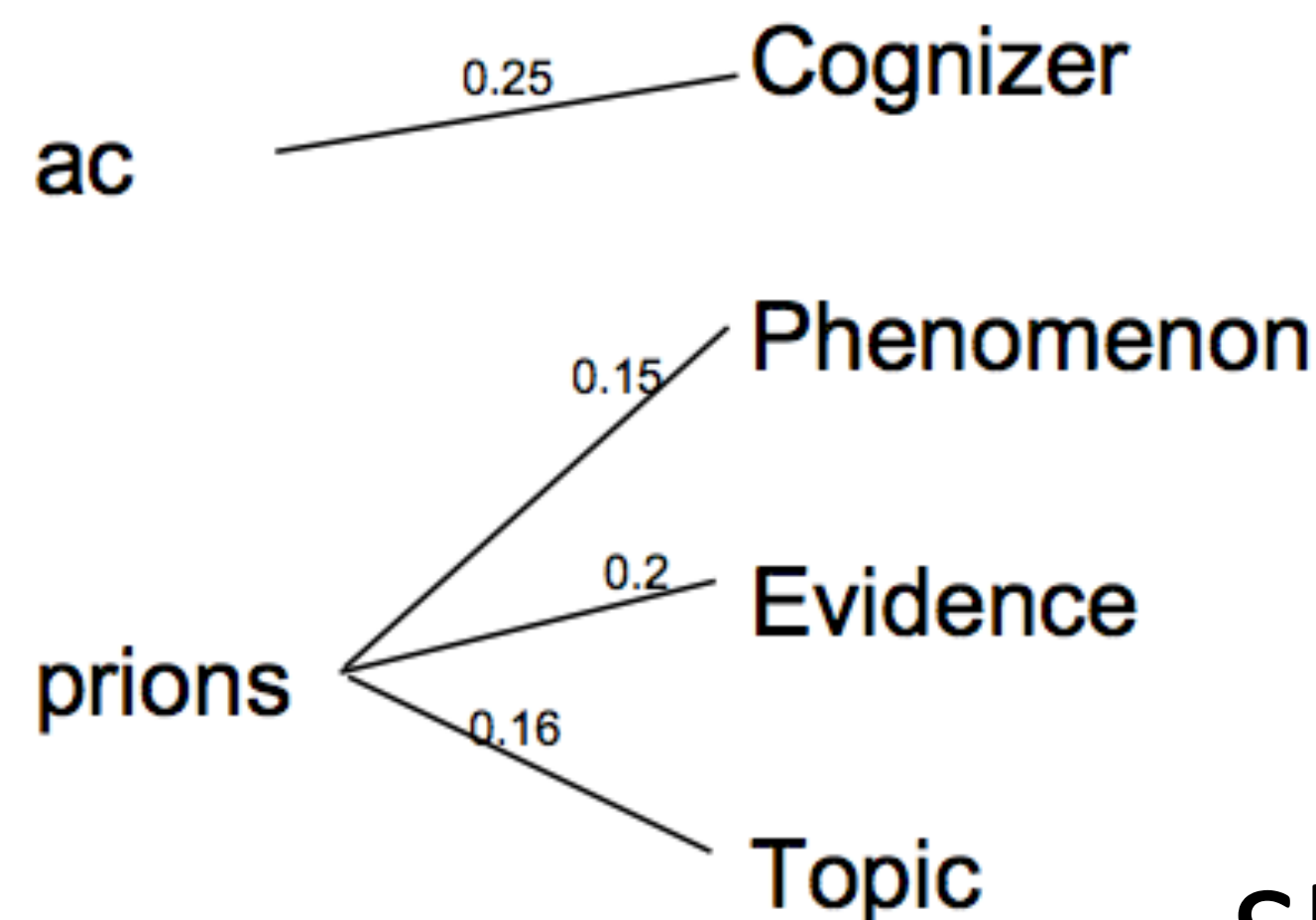
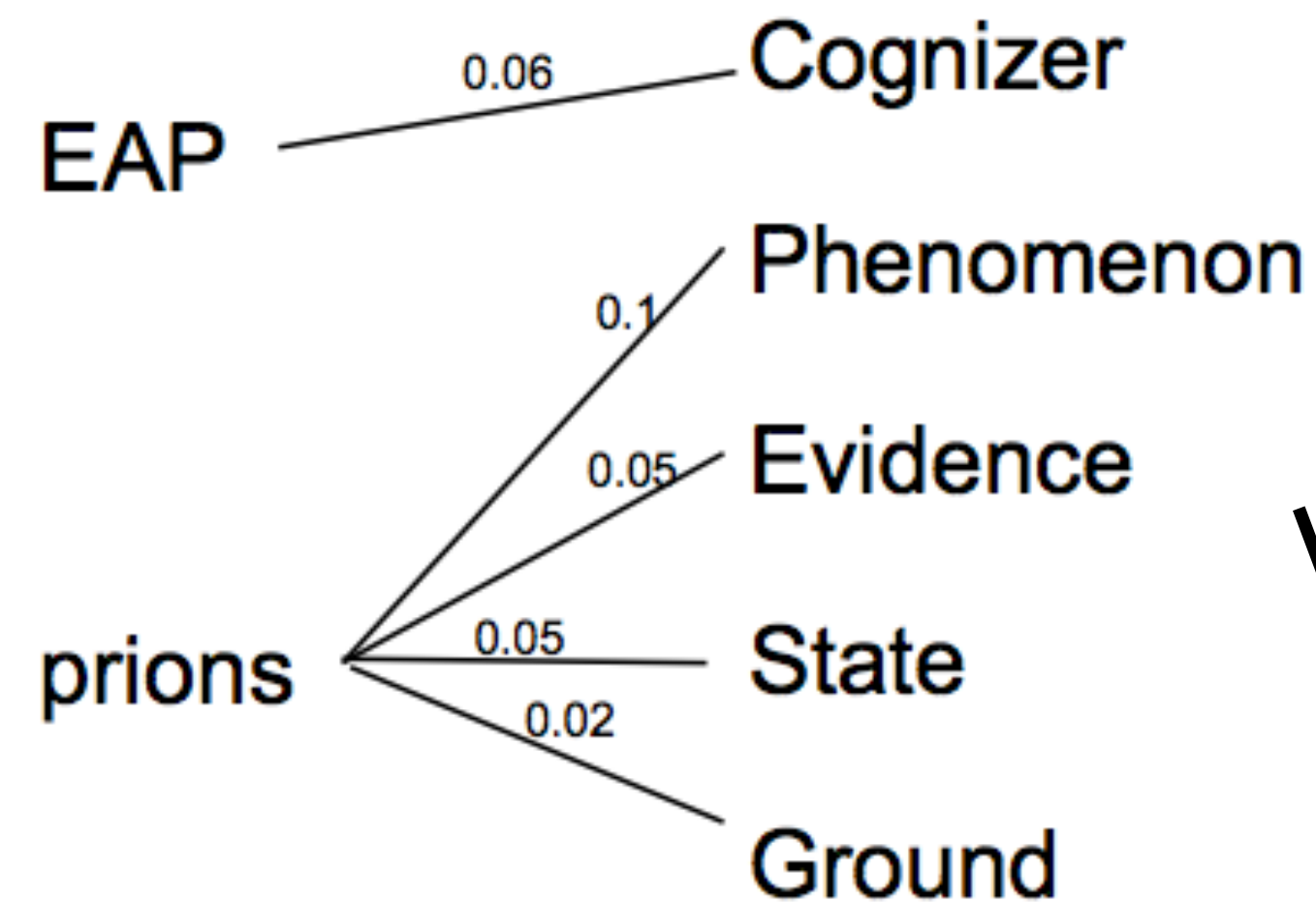
# SRL for QA

- Question and several answer candidates

Q: *Who discovered prions?*

AC1: *In 1997, Stanley B. Prusiner, a scientist in the United States, discovered prions...*

AC2: *Prions were researched by...*



Score by matching expected answer phrase (EAP) against answer candidate (AC)

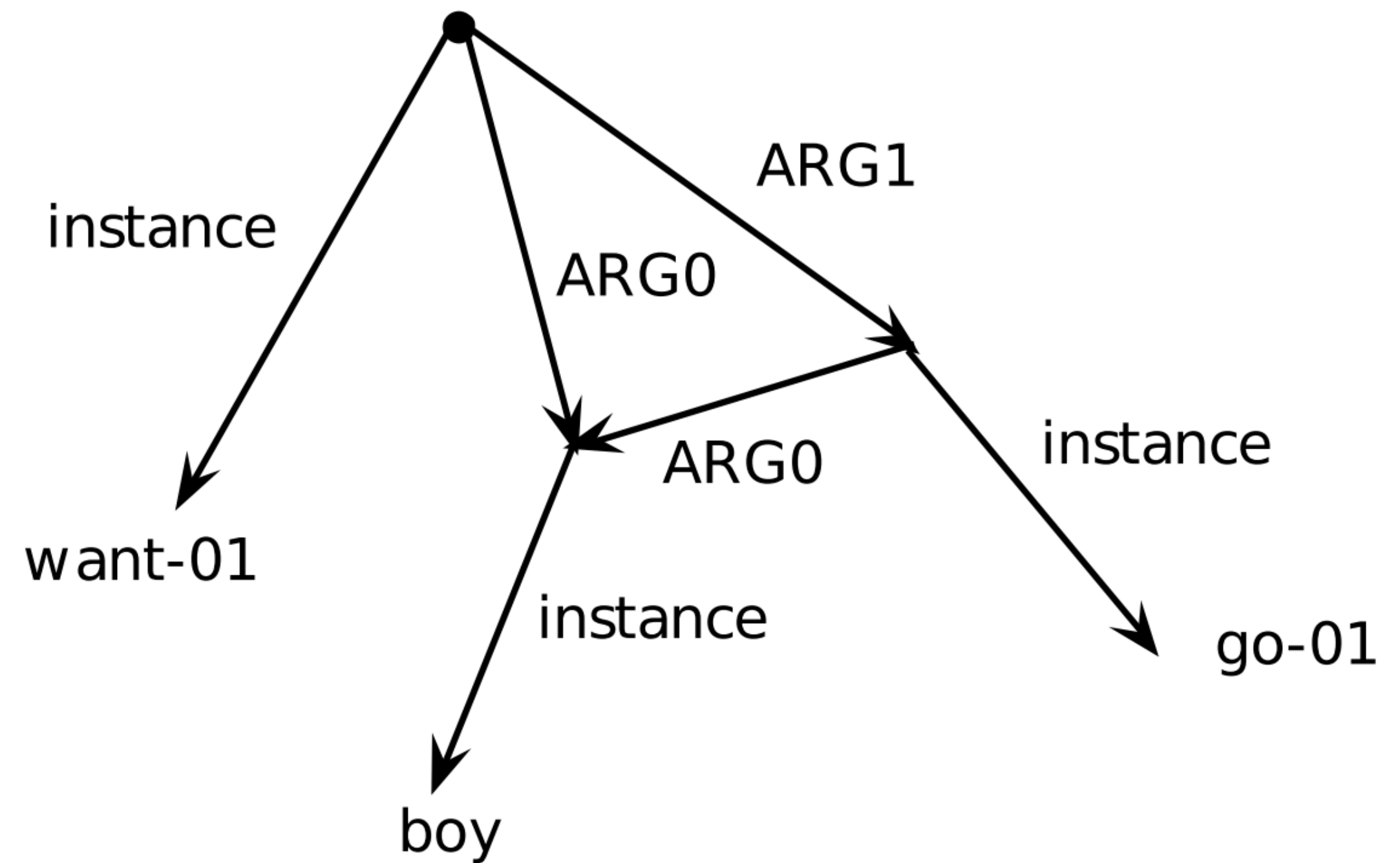
Shen and Lapata (2007)



# Abstract Meaning Representation

Banarescu et al. (2014)

- ▶ Graph-structured annotation

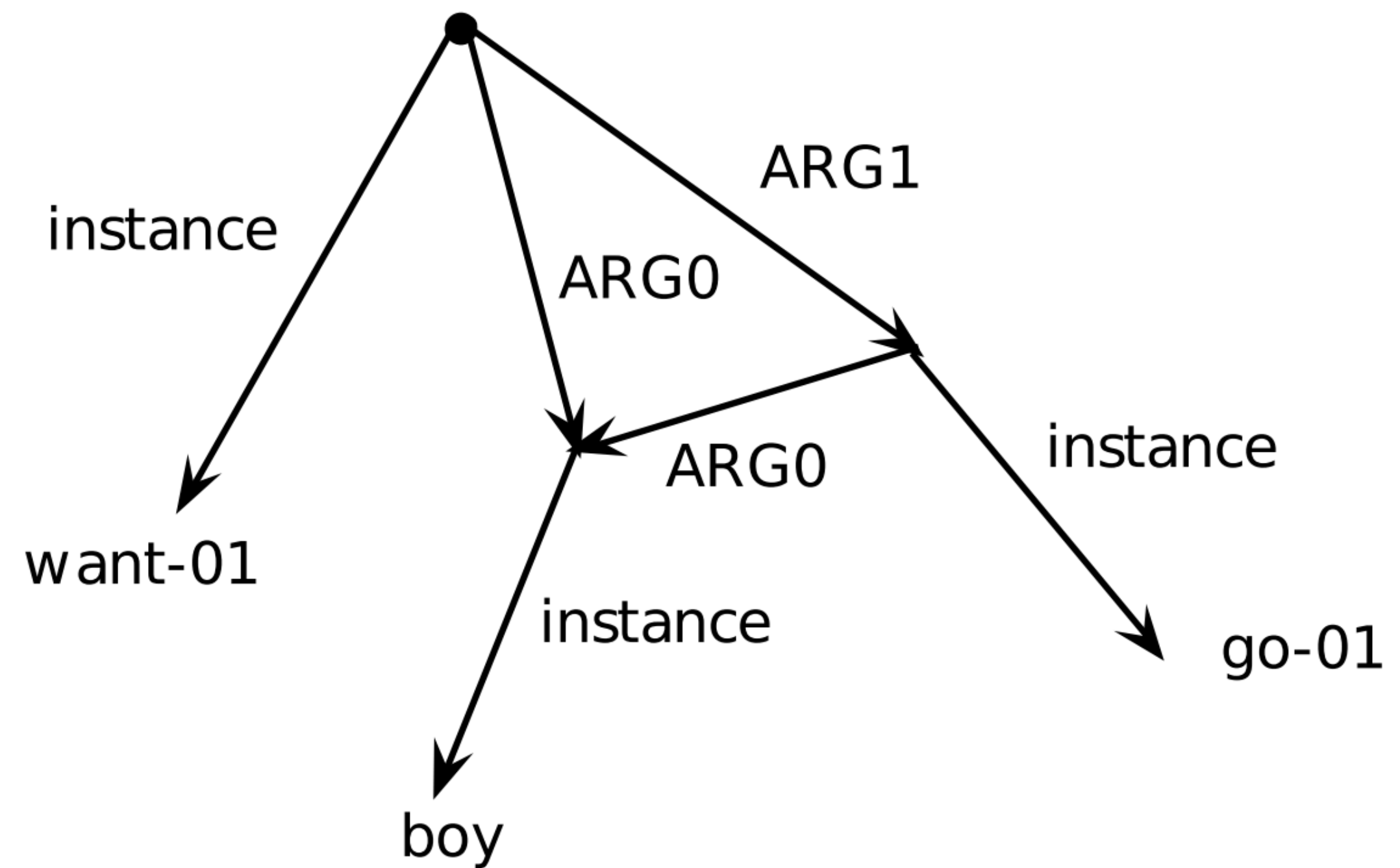


The boy wants to go

# Abstract Meaning Representation

Banarescu et al. (2014)

- ▶ Graph-structured annotation
- ▶ Superset of SRL: full sentence analyses, contains coreference and multi-word expressions as well

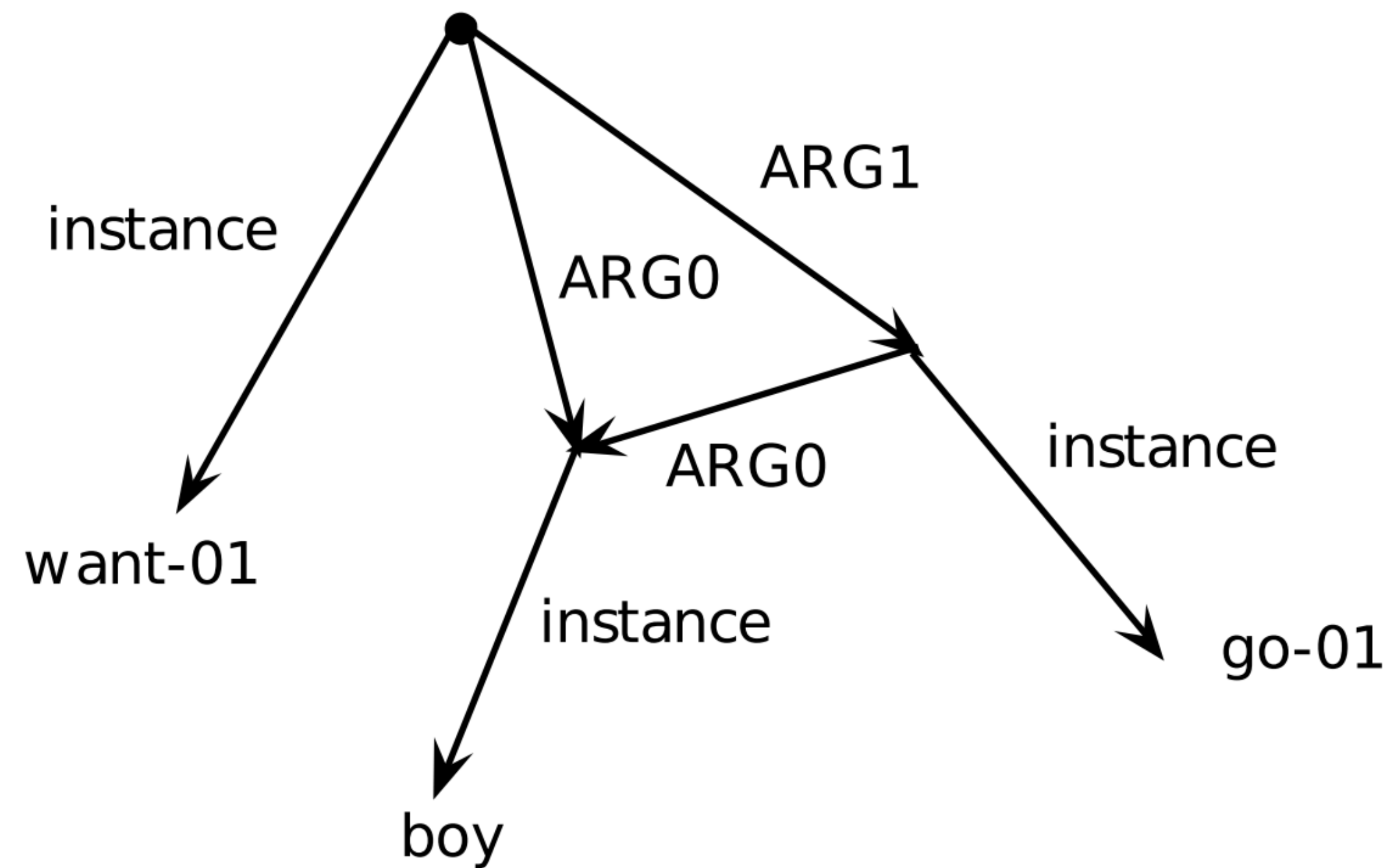


The boy wants to go

# Abstract Meaning Representation

Banarescu et al. (2014)

- ▶ Graph-structured annotation
- ▶ Superset of SRL: full sentence analyses, contains coreference and multi-word expressions as well
- ▶ F1 scores in the 60s: hard!

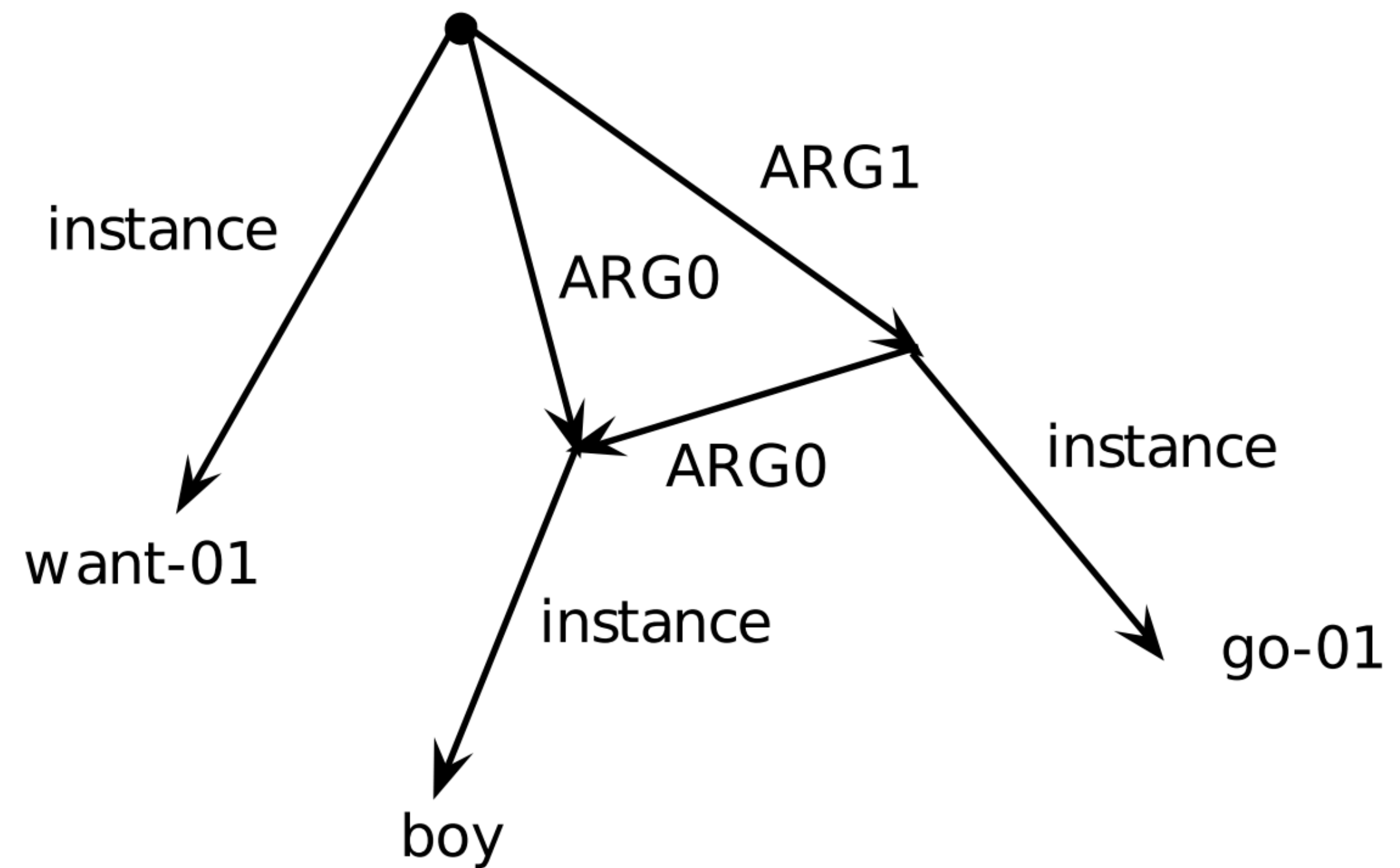


The boy wants to go

# Abstract Meaning Representation

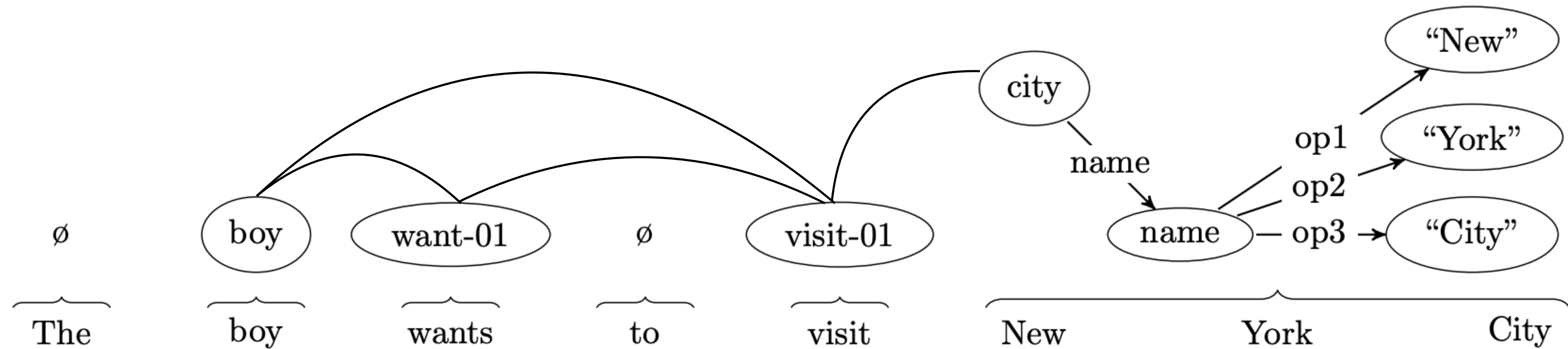
Banarescu et al. (2014)

- ▶ Graph-structured annotation
- ▶ Superset of SRL: full sentence analyses, contains coreference and multi-word expressions as well
- ▶ F1 scores in the 60s: hard!
- ▶ So comprehensive that it's hard to predict, but still doesn't handle tense or some other things...



The boy wants to go

# Abstract Meaning Representation

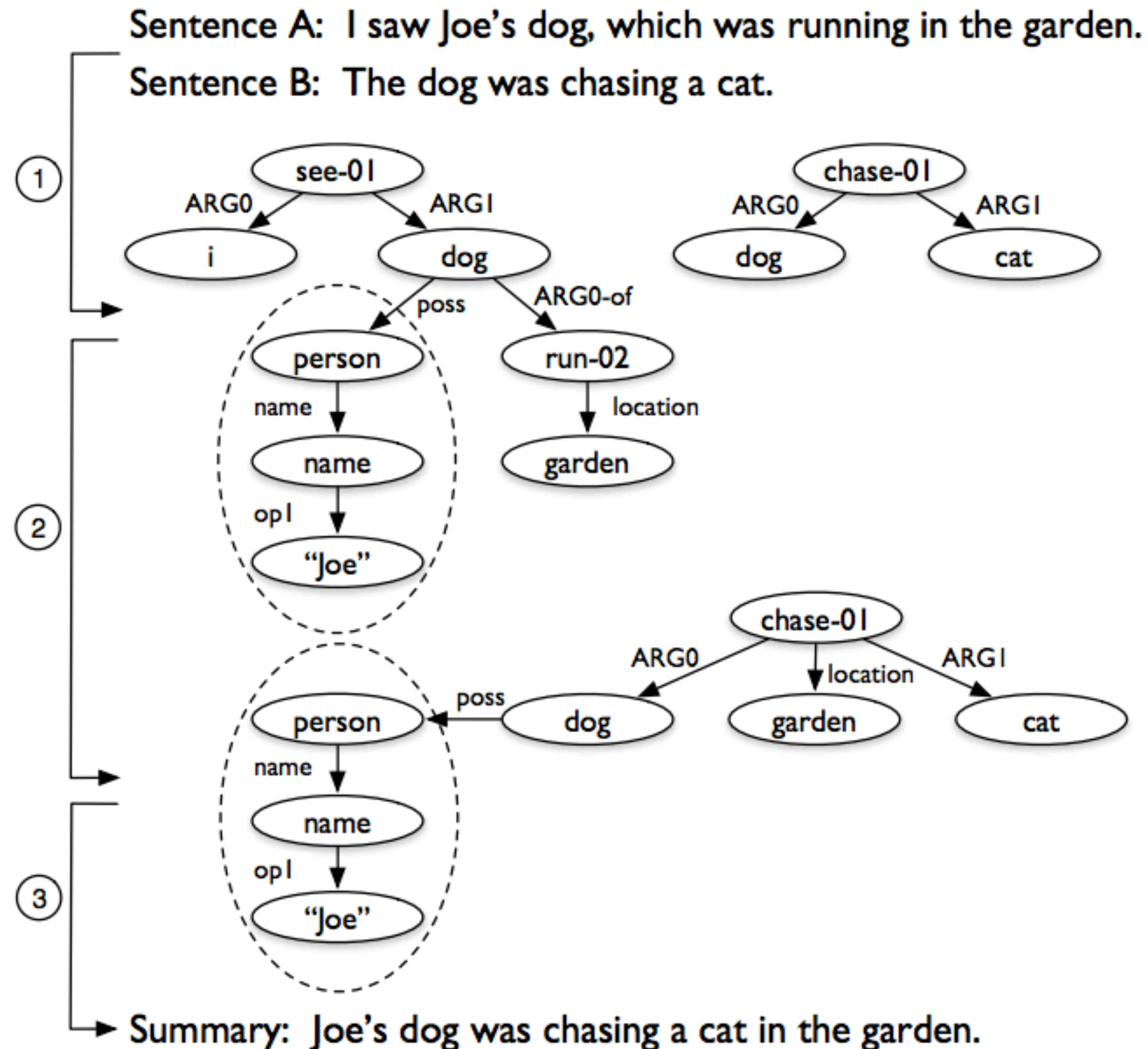


- ▶ First predict mapping from concepts to graph nodes (many-to-many)
- ▶ Then use an edge scoring module similar to dependency parsers to predict edges
- ▶ Predicting a coherent graph is *hard*, lots of constraints on it and no dynamic program

Flanigan et al. (2016), Lyu et al. (2018)

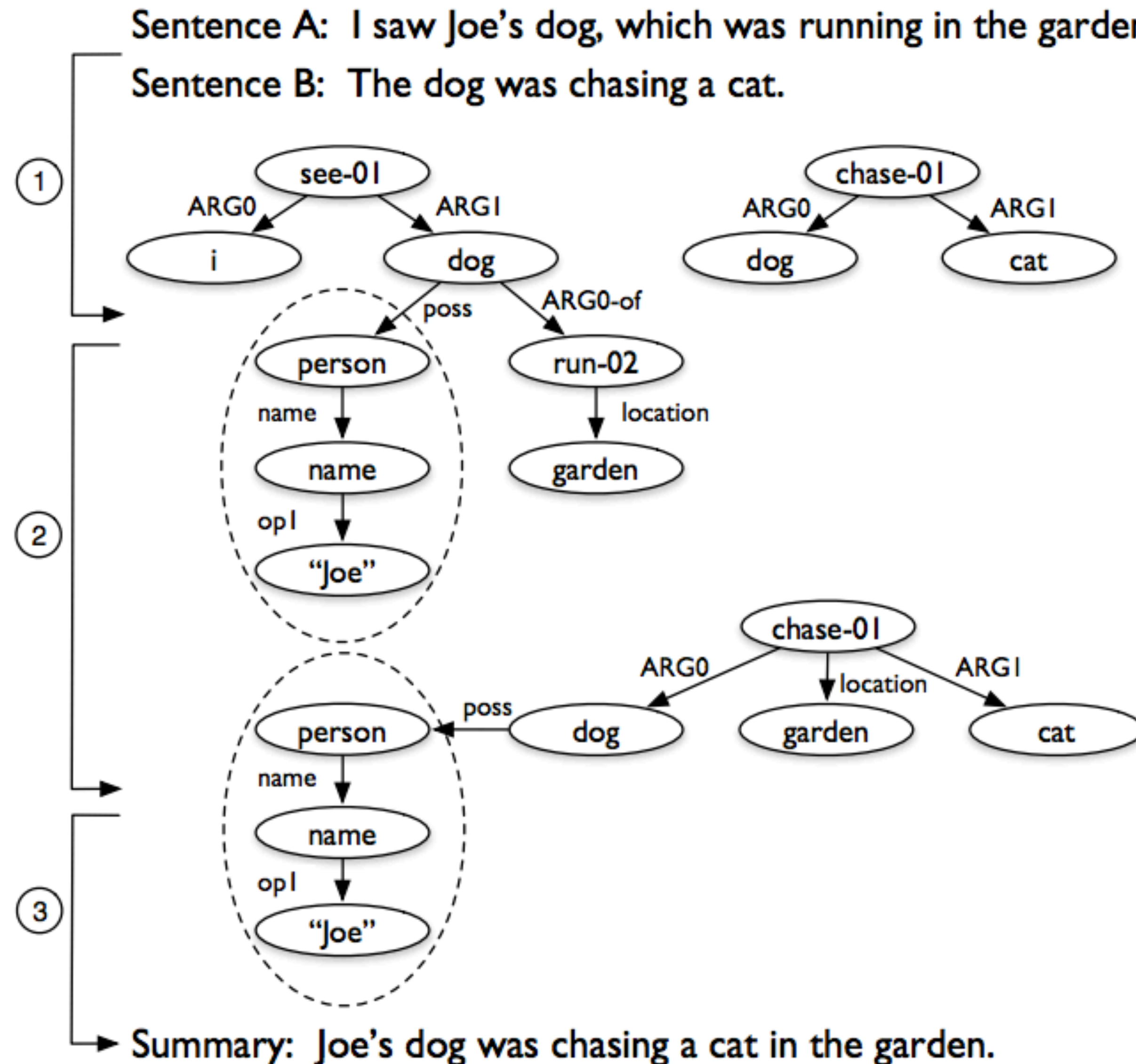


# Summarization with AMR

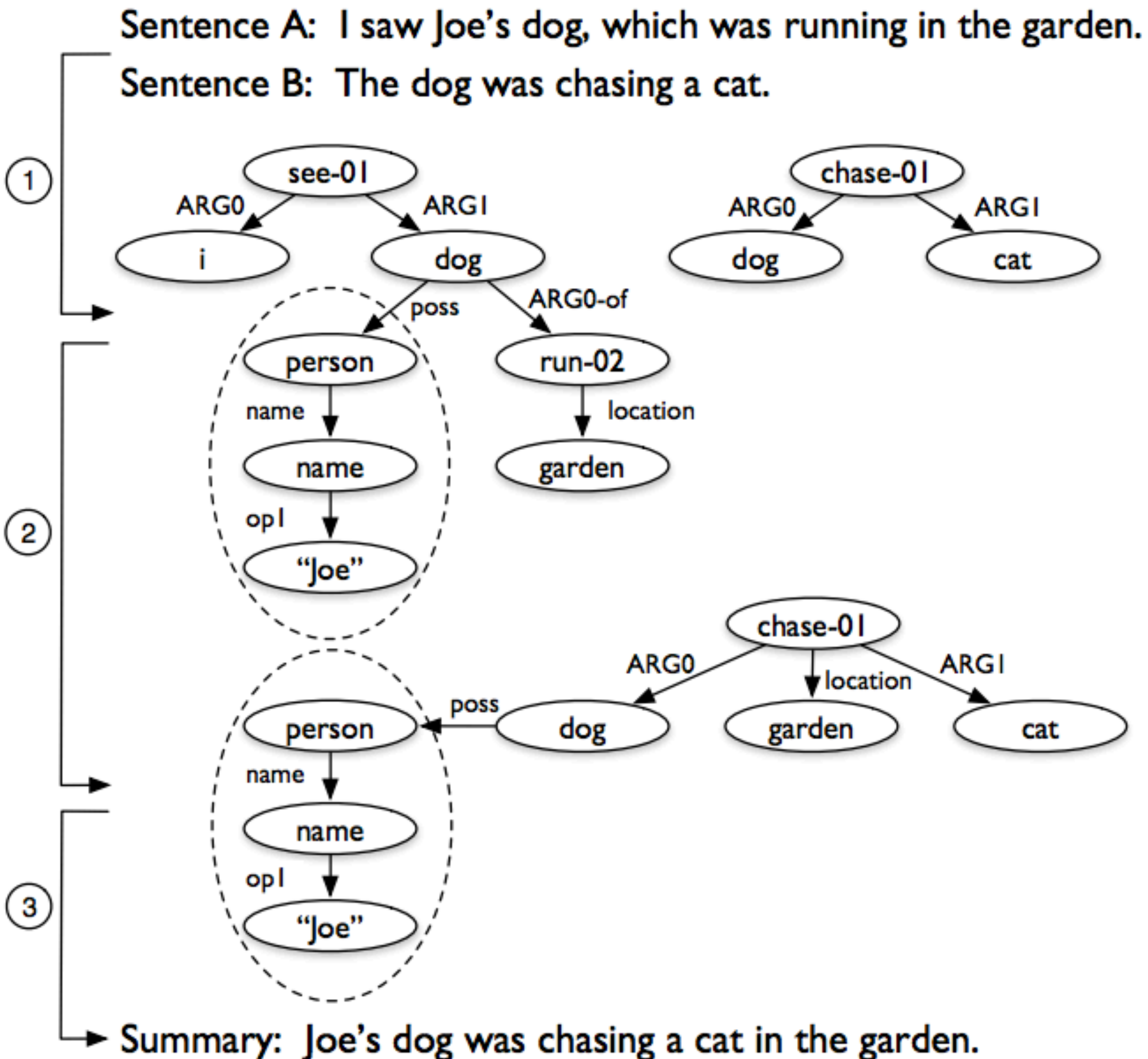


# Summarization with AMR

- Merge AMRs across multiple sentences



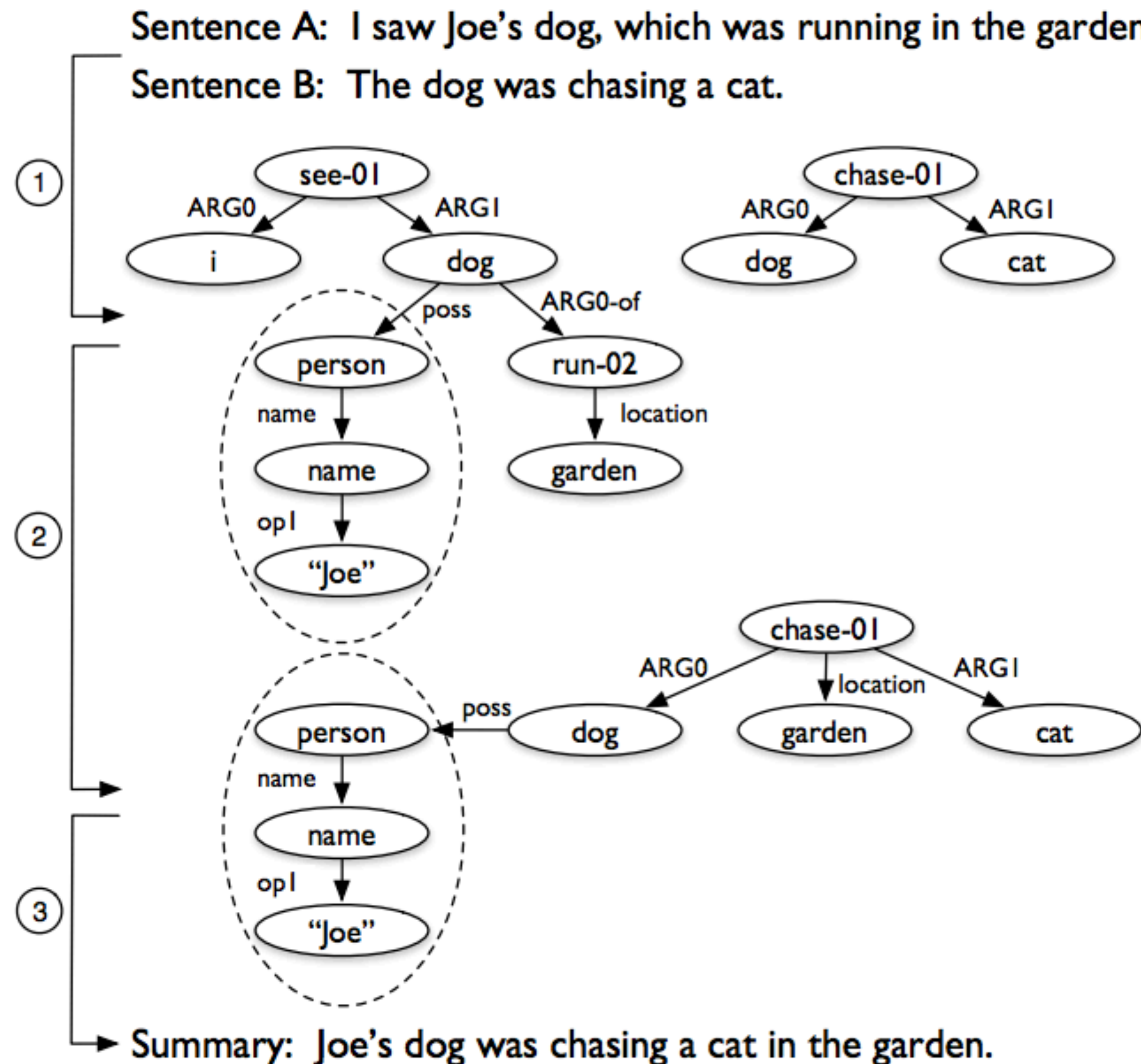
# Summarization with AMR



- ▶ Merge AMRs across multiple sentences
- ▶ Summarization = subgraph extraction



# Summarization with AMR



- ▶ Merge AMRs across multiple sentences
- ▶ Summarization = subgraph extraction
- ▶ No real systems actually work this way (more when we talk about summarization)

# Slot Filling

# Slot Filling

---

- ▶ Most conservative, narrow form of IE

# Slot Filling

---

- ▶ Most conservative, narrow form of IE

magnitude

time

*Indian Express — A massive earthquake of magnitude 7.3 struck Iraq on Sunday, 103 kms (64 miles) southeast of the city of As-Sulaymaniyah, the US Geological Survey said, reports Reuters. US Geological Survey initially said the quake was of a magnitude 7.2, before revising it to 7.3.*

epicenter

# Slot Filling

---

- ▶ Most conservative, narrow form of IE

magnitude

time

*Indian Express* — A massive earthquake of magnitude 7.3 struck Iraq on Sunday, 103 kms (64 miles) southeast of the city of As-Sulaymaniyah, the US Geological Survey said, reports Reuters. US Geological Survey initially said the quake was of a magnitude 7.2, before revising it to 7.3.

epicenter

Speaker: Alan Clark

speaker

“Gender Roles in the Holy Roman Empire”

title

Allagher Center Main Auditorium

location

This talk will discuss...

# Slot Filling

- ▶ Most conservative, narrow form of IE

magnitude

time

*Indian Express* — A massive earthquake of **magnitude 7.3** struck Iraq on **Sunday**, 103 kms (64 miles) southeast of the city of As-Sulaymaniyah, the US Geological Survey said, reports Reuters. US Geological Survey initially said the quake was of a magnitude 7.2, before revising it to 7.3.

epicenter

Speaker: **Alan Clark**

speaker

**"Gender Roles in the Holy Roman Empire"**

title

**Allagher Center Main Auditorium**

location

*This talk will discuss...*

- ▶ Old work: HMMs, later CRFs trained per role



# Slot Filling: MUC

---

## Template

(a)

SELLER	BUSINESS	ACQUIRED	PURCHASER
CSR Limited	Oil and Gas	Delhi Fund	Esso Inc.

## Document

(b) [S CSR] has said that [S it] has sold [S its] [B oil interests] held in [A Delhi Fund]. [P Esso Inc.] did not disclose how much [P they] paid for [A Dehli].

- ▶ Key aspect: need to combine information across multiple mentions of an entity using coreference

# Slot Filling: Forums

---

- ▶ Extract product occurrences in cybercrime forums, but not everything that looks like a product is a product

TITLE: [ buy ] Backconnect bot

BODY: Looking for a solid backconnect bot .

If you know of anyone who codes them please let me know

(a) File 0-initiator4856

TITLE: Exploit cleaning ?

BODY: Have some Exploits i need fud .

(b) File 0-initiator10815

Not a product in this context

Portnoff et al. (2017), Durrett et al. (2017)



# Relation Extraction

# Relation Extraction

---

- ▶ Extract entity-relation-entity triples from a fixed inventory

# Relation Extraction

---

- ▶ Extract entity-relation-entity triples from a fixed inventory

*During the war in Iraq, American journalists were sometimes caught in the line of fire*

# Relation Extraction

---

- ▶ Extract entity-relation-entity triples from a fixed inventory

*During the war in Iraq, American journalists were sometimes caught in the line of fire*

# Relation Extraction

---

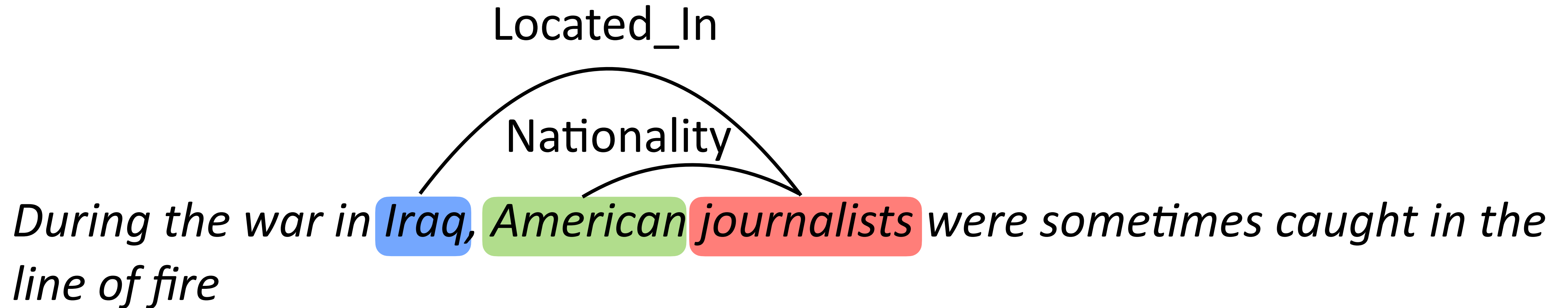
- ▶ Extract entity-relation-entity triples from a fixed inventory



# Relation Extraction

---

- ▶ Extract entity-relation-entity triples from a fixed inventory



- ▶ Use NER-like system to identify entity spans, classify relations between entity pairs with a classifier

# Relation Extraction

---

- ▶ Extract entity-relation-entity triples from a fixed inventory



- ▶ Use NER-like system to identify entity spans, classify relations between entity pairs with a classifier
- ▶ Systems can be feature-based or neural, look at surface words, syntactic features (dependency paths), semantic roles

# Relation Extraction

---

- ▶ Extract entity-relation-entity triples from a fixed inventory



- ▶ Use NER-like system to identify entity spans, classify relations between entity pairs with a classifier
  - ▶ Systems can be feature-based or neural, look at surface words, syntactic features (dependency paths), semantic roles
  - ▶ Problem: limited data for scaling to big ontologies
- ACE (2003-2005)



# Hearst Patterns

---

- ▶ Syntactic patterns especially for finding hypernym-hyponym pairs (“is a” relations)

# Hearst Patterns

---

- ▶ Syntactic patterns especially for finding hypernym-hyponym pairs (“is a” relations)

# Hearst Patterns

---

- ▶ Syntactic patterns especially for finding hypernym-hyponym pairs (“is a” relations)

*Y is a X*

*Berlin is a city*

# Hearst Patterns

---

- ▶ Syntactic patterns especially for finding hypernym-hyponym pairs (“is a” relations)

*Y is a X*

*Berlin is a city*

*X such as [list]*

*cities such as Berlin, Paris, and London.*

# Hearst Patterns

---

- ▶ Syntactic patterns especially for finding hypernym-hyponym pairs (“is a” relations)

*Y is a X*

*Berlin is a city*

*X such as [list]*

*cities such as Berlin, Paris, and London.*

*other X including Y*

*other cities including Berlin*

# Hearst Patterns

---

- ▶ Syntactic patterns especially for finding hypernym-hyponym pairs (“is a” relations)

*Y is a X*

*Berlin is a city*

*X such as [list]*

*cities such as Berlin, Paris, and London.*

*other X including Y*

*other cities including Berlin*

- ▶ Totally unsupervised way of harvesting world knowledge for tasks like parsing and coreference (Bansal and Klein, 2011-2012)

# Distant Supervision

---

# Distant Supervision

---

- ▶ Lots of relations in our knowledge base already (e.g., 23,000 film-director relations); use these to bootstrap more training data



# Distant Supervision

---

- ▶ Lots of relations in our knowledge base already (e.g., 23,000 film-director relations); use these to bootstrap more training data
- ▶ If two entities in a relation appear in the same sentence, assume the sentence expresses the relation

# Distant Supervision

---

- ▶ Lots of relations in our knowledge base already (e.g., 23,000 film-director relations); use these to bootstrap more training data
- ▶ If two entities in a relation appear in the same sentence, assume the sentence expresses the relation

Director

*[Steven Spielberg]'s film [Saving Private Ryan] is loosely based on the brothers' story*



# Distant Supervision

---

- ▶ Lots of relations in our knowledge base already (e.g., 23,000 film-director relations); use these to bootstrap more training data
- ▶ If two entities in a relation appear in the same sentence, assume the sentence expresses the relation

Director

*[Steven Spielberg]'s film [Saving Private Ryan] is loosely based on the brothers' story*

*Allison co-produced the Academy Award-winning [Saving Private Ryan], directed by [Steven Spielberg]*

Director



# Distant Supervision

- ▶ Learn decently accurate classifiers for ~100 Freebase relations
- ▶ Could be used to crawl the web and expand our knowledge base

Relation name	100 instances			1000 instances		
	Syn	Lex	Both	Syn	Lex	Both
/film/director/film	<b>0.49</b>	0.43	0.44	<b>0.49</b>	0.41	0.46
/film/writer/film	<b>0.70</b>	0.60	0.65	<b>0.71</b>	0.61	0.69
/geography/river/basin_countries	0.65	0.64	<b>0.67</b>	<b>0.73</b>	0.71	0.64
/location/country/administrative_divisions	0.68	0.59	<b>0.70</b>	<b>0.72</b>	0.68	<b>0.72</b>
/location/location/contains	0.81	<b>0.89</b>	0.84	<b>0.85</b>	0.83	0.84
/location/us_county/county_seat	0.51	0.51	<b>0.53</b>	0.47	<b>0.57</b>	0.42
/music/artist/origin	0.64	0.66	<b>0.71</b>	0.61	<b>0.63</b>	0.60
/people/deceased_person/place_of_death	0.80	0.79	<b>0.81</b>	0.80	<b>0.81</b>	0.78
/people/person/nationality	0.61	0.70	<b>0.72</b>	0.56	0.61	<b>0.63</b>
/people/person/place_of_birth	<b>0.78</b>	0.77	<b>0.78</b>	0.88	0.85	<b>0.91</b>
Average	0.67	0.66	<b>0.69</b>	<b>0.68</b>	0.67	0.67

# FewRel

- ▶ Treats relation classification as a few-shot classification problem
- ▶ 100 classes x 700 instances, goal is to generalize to each class with just a few instances
- ▶ BERT can handle this fairly well (Soares et al., 2019)
- ▶ “FewRel 2.0”: new dataset with “none of the above” type, which makes things much harder

Supporting Set	
(A) capital_of	(1) <i>London</i> is the capital of <i>the U.K.</i> (2) <i>Washington</i> is the capital of <i>the U.S.A.</i>
(B) member_of	(1) <i>Newton</i> served as the president of <i>the Royal Society.</i> (2) <i>Leibniz</i> was a member of <i>the Prussian Academy of Sciences.</i>
(C) birth_name	(1) <i>Samuel Langhorne Clemens</i> , better known by his pen name <i>Mark Twain</i> , was an American writer. (2) <i>Alexei Maximovich Peshkov</i> , primarily known as <i>Maxim Gorky</i> , was a Russian and Soviet writer.
Test Instance	
(A) or (B) or (C)	<i>Euler</i> was elected a foreign member of <i>the Royal Swedish Academy of Sciences.</i>

Han et al. (2018), Gao et al. (2019)



# Entity Tracking / Procedural Text

# Entity Tracking

- ▶ Information extraction for “procedural text”: text describing some kind of process
- ▶ For a recipe: what ingredients are involved at each timestep?
- ▶ Involves global constraints and being able to model complex entity interactions

## Recipes Dataset

Seq. of Steps	sugar	eggs	flour
Combine sugar, oil, and vanilla	1	0	0
Add eggs one at a time	1	1	0
In a separate bowl, combine flour, soda, and salt.	0	0	1
Add to the <b>sugar mixture</b> alternately with milk	1	1	1
Stir <b>remaining ingredients</b> one at a time.	1	1	1

Tracking  
**Intermediate  
Compositions**

**Global Tracking  
without Explicit  
Entity Mentions**

0 → Ingredient Absent  
1 → Ingredient Present

Kiddon et al. (2016), Bosselut et al. (2018)

Slide credit: Aditya Gupta



# Entity Tracking

- ▶ Process paragraphs: predict when objects are created, moved, or destroyed in a scientific process
- ▶ Structured prediction problem, tied to the particular information conveyed in these paragraphs
- ▶ Use a neural CRF to make a coherent prediction for each entity

## ProPara Dataset

Seq. of Steps	water	mixture	sugar
Roots absorb water from soil.	M	O	O
The water flows to the leaf.	M	O	O
Light from the sun and CO <sub>2</sub> enter the leaf.	E	O	O
Light, water, and CO <sub>2</sub> combine into mixture.	D	C	O
Mixture forms sugar.	O	D	C

**Implicit Events**  
requiring Global  
Knowledge

**Structural  
Constraints**  
C → M → D

C → Creation  
E → Existence  
M → Movement  
D → Destruction  
O → Outside Process

Dalvi et al. (2018), Gupta and Durrett (2019)

Slide credit: Aditya Gupta

Open IE

# Open Information Extraction

---

- ▶ “Open”ness — want to be able to extract all kinds of information from open-domain text
- ▶ Acquire commonsense knowledge just from “reading” about it, but need to process lots of text (“machine reading”)
- ▶ Typically no fixed relation inventory

# TextRunner

---

- ▶ Extract positive examples of (e, r, e) triples via parsing and heuristics
- ▶ Train a Naive Bayes classifier to filter triples from raw text: uses features on POS tags, lexical features, stopwords, etc.

# TextRunner

---

- ▶ Extract positive examples of (e, r, e) triples via parsing and heuristics
- ▶ Train a Naive Bayes classifier to filter triples from raw text: uses features on POS tags, lexical features, stopwords, etc.

*Barack Obama, 44th president of the United States, was born on August 4, 1961 in Honolulu*

*=> Barack\_Obama, was born in, Honolulu*

# TextRunner

---

- ▶ Extract positive examples of (e, r, e) triples via parsing and heuristics
- ▶ Train a Naive Bayes classifier to filter triples from raw text: uses features on POS tags, lexical features, stopwords, etc.

*Barack Obama, 44th president of the United States, was born on August 4, 1961 in Honolulu*

*=> Barack\_Obama, was born in, Honolulu*

- ▶ 80x faster than running a parser (which was slow in 2007...)

# TextRunner

---

- ▶ Extract positive examples of (e, r, e) triples via parsing and heuristics
- ▶ Train a Naive Bayes classifier to filter triples from raw text: uses features on POS tags, lexical features, stopwords, etc.

*Barack Obama, 44th president of the United States, was born on August 4, 1961 in Honolulu*

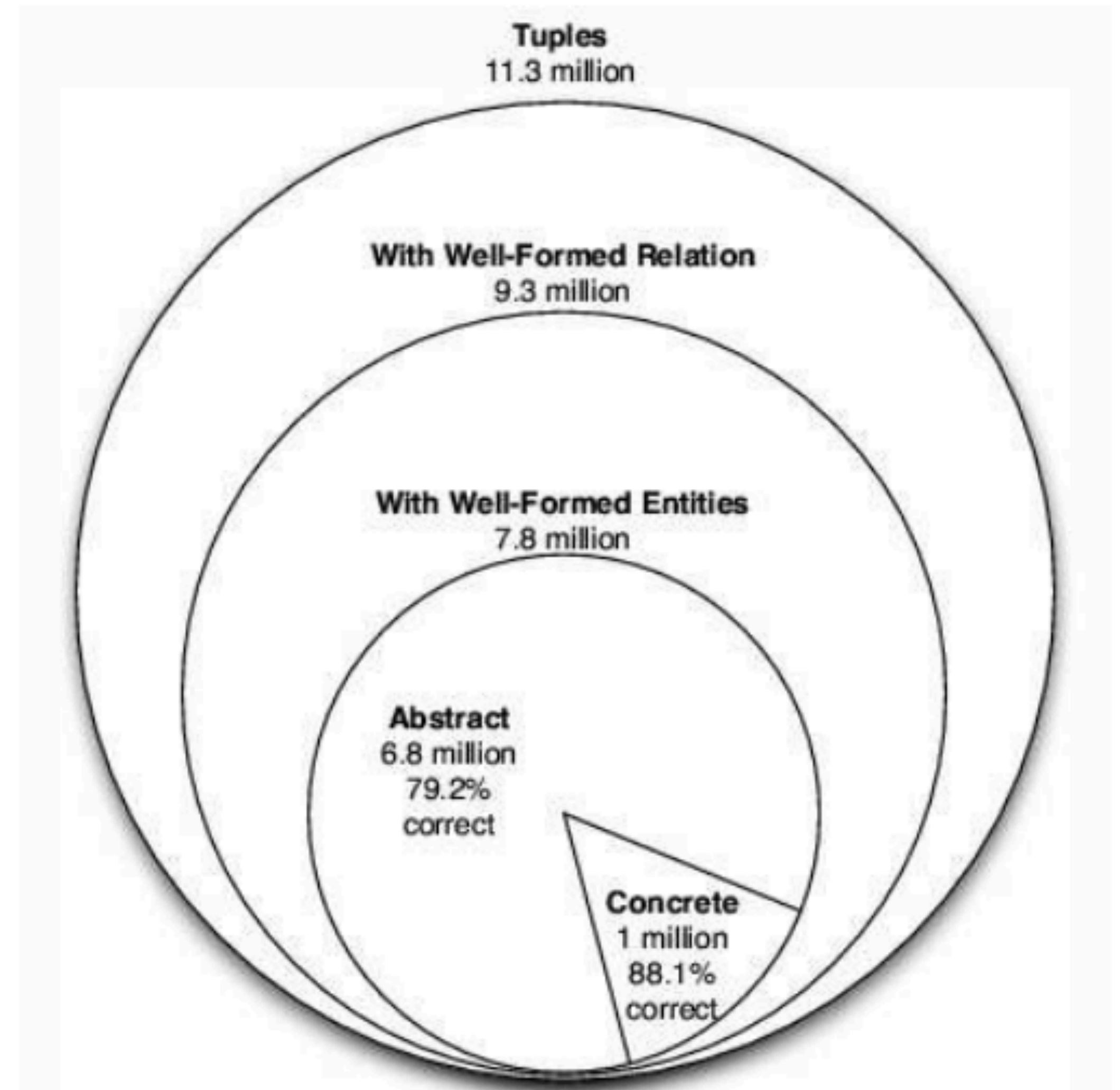
*=> Barack\_Obama, was born in, Honolulu*

- ▶ 80x faster than running a parser (which was slow in 2007...)
- ▶ Use multiple instances of extractions to assign probability to a relation



# Exploiting Redundancy

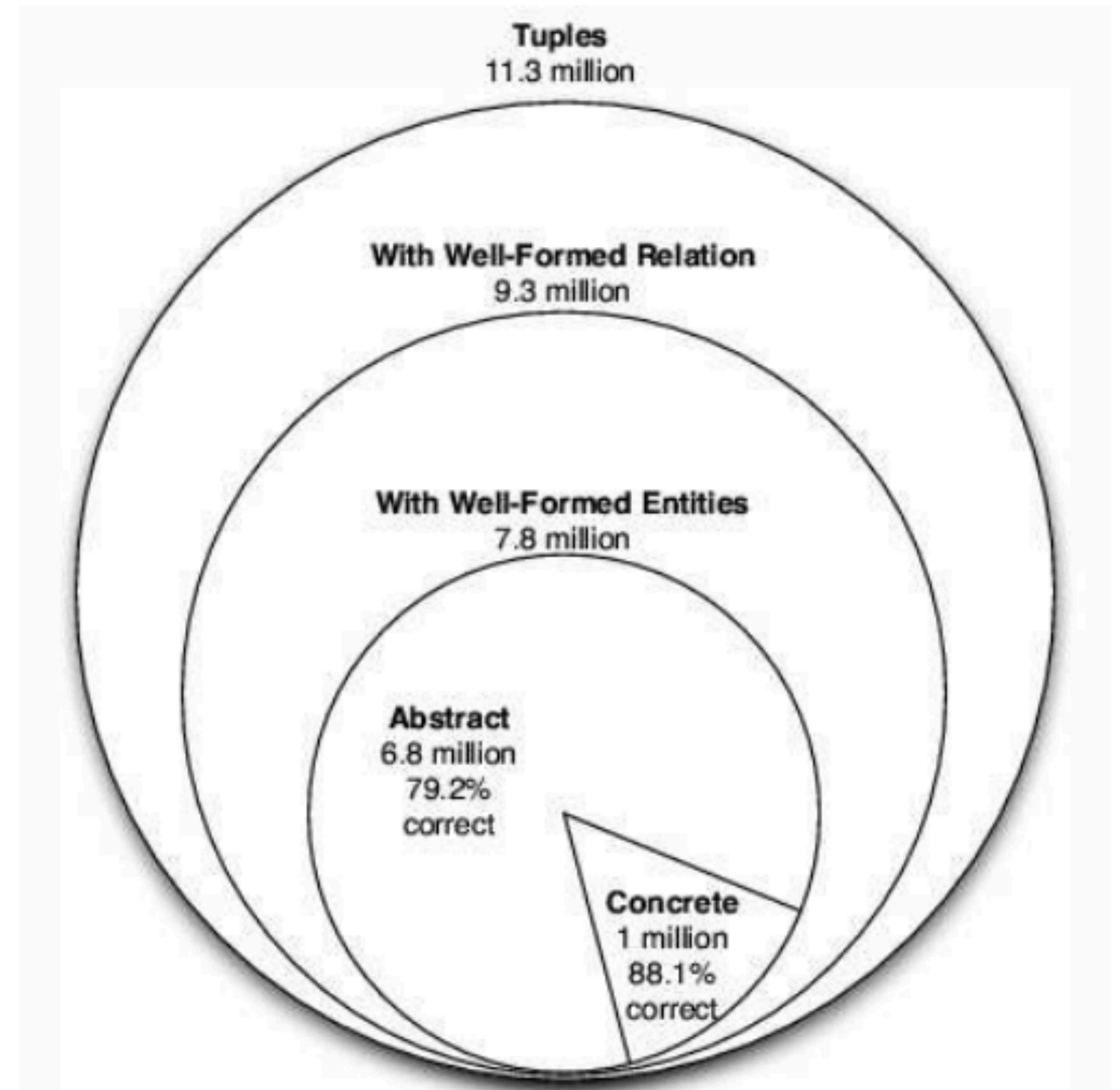
---



Banko et al. (2007)

# Exploiting Redundancy

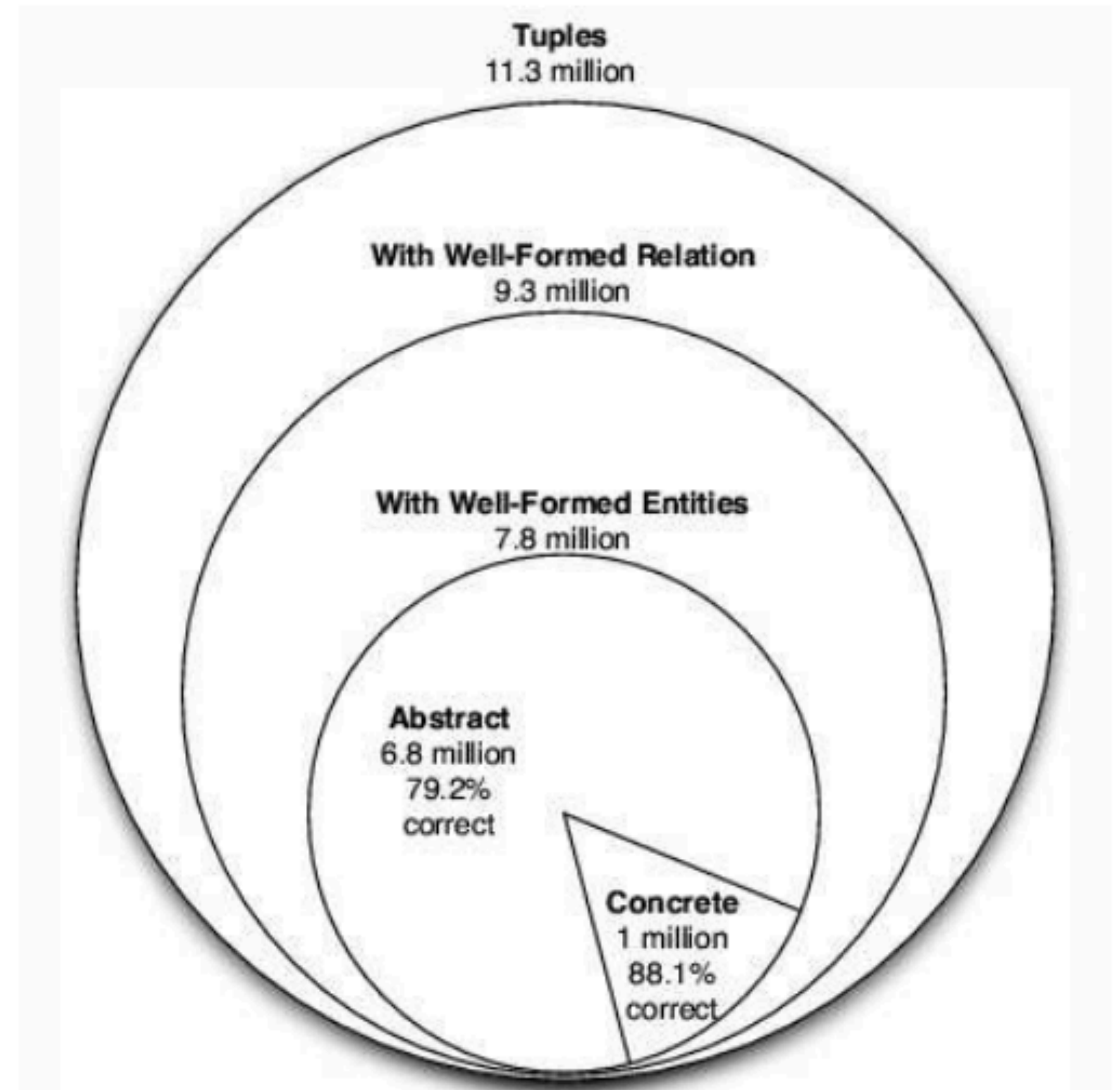
- ▶ 9M web pages / 133M sentences



Banko et al. (2007)

# Exploiting Redundancy

- ▶ 9M web pages / 133M sentences
- ▶ 2.2 tuples extracted per sentence, filter based on probabilities

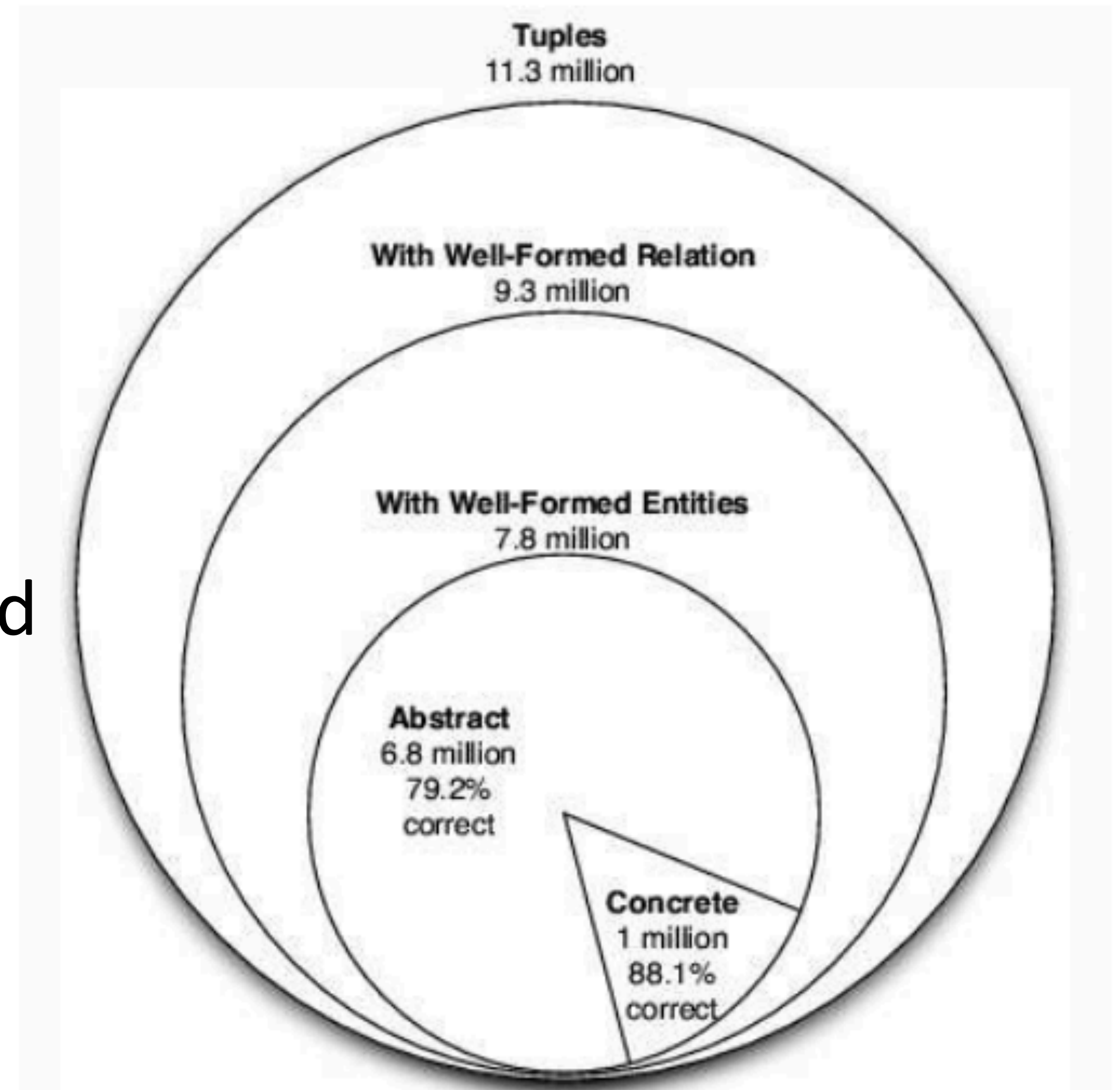


Banko et al. (2007)



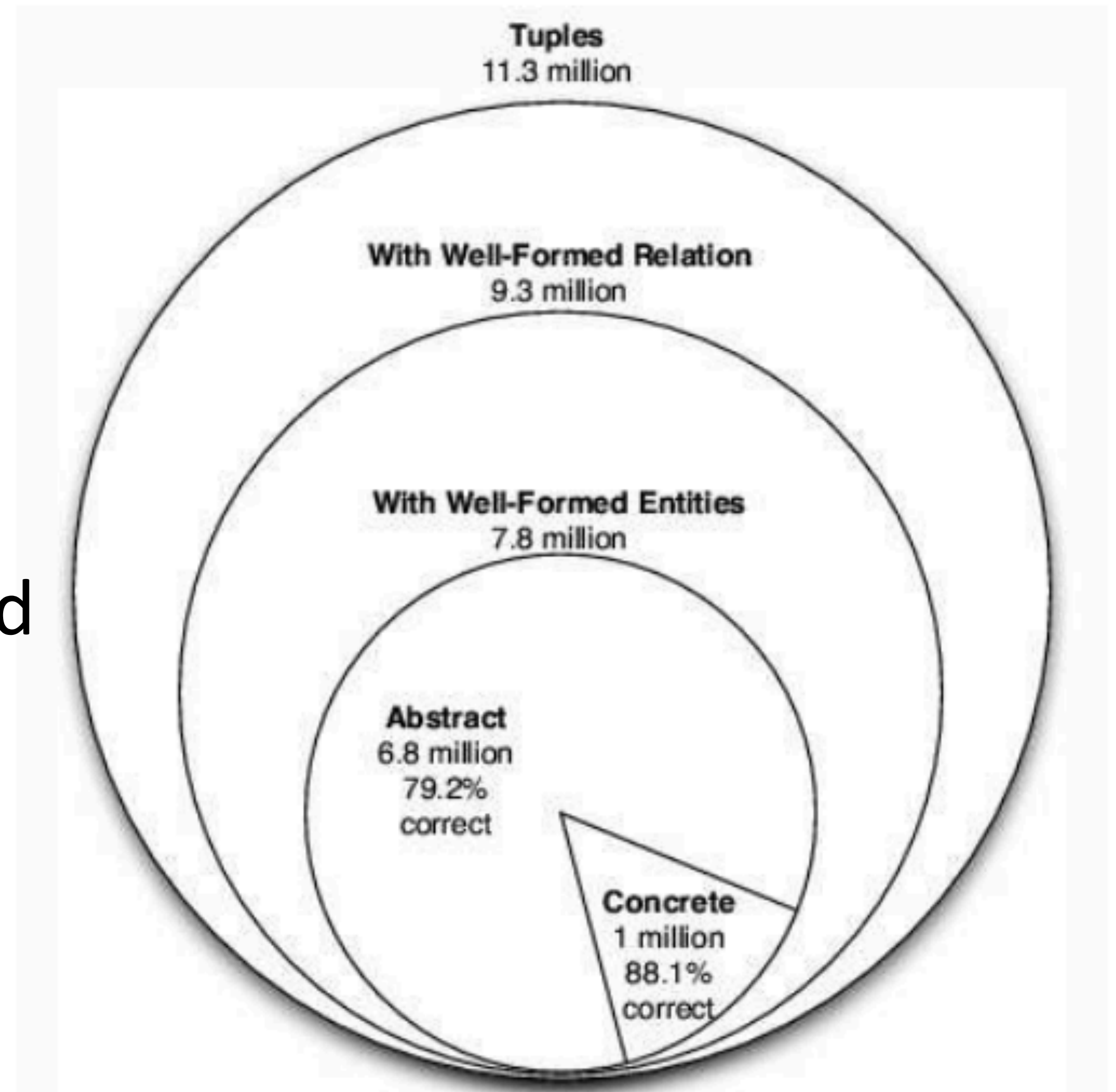
# Exploiting Redundancy

- ▶ 9M web pages / 133M sentences
- ▶ 2.2 tuples extracted per sentence, filter based on probabilities
- ▶ Concrete: definitely true  
Abstract: possibly true but underspecified



# Exploiting Redundancy

- ▶ 9M web pages / 133M sentences
- ▶ 2.2 tuples extracted per sentence, filter based on probabilities
- ▶ Concrete: definitely true  
Abstract: possibly true but underspecified
- ▶ Hard to evaluate: can assess precision of extracted facts, but how do we know recall?



# ReVerb

---

- ▶ More constraints: open relations have to begin with verb, end with preposition, be contiguous (e.g., *was born on*)



# ReVerb

---

- ▶ More constraints: open relations have to begin with verb, end with preposition, be contiguous (e.g., *was born on*)
- ▶ Extract more meaningful relations, particularly with light verbs

is	is an album by, is the author of, is a city in
has	has a population of, has a Ph.D. in, has a cameo in
made	made a deal with, made a promise to
took	took place in, took control over, took advantage of
gave	gave birth to, gave a talk at, gave new meaning to
got	got tickets to, got a deal on, got funding from



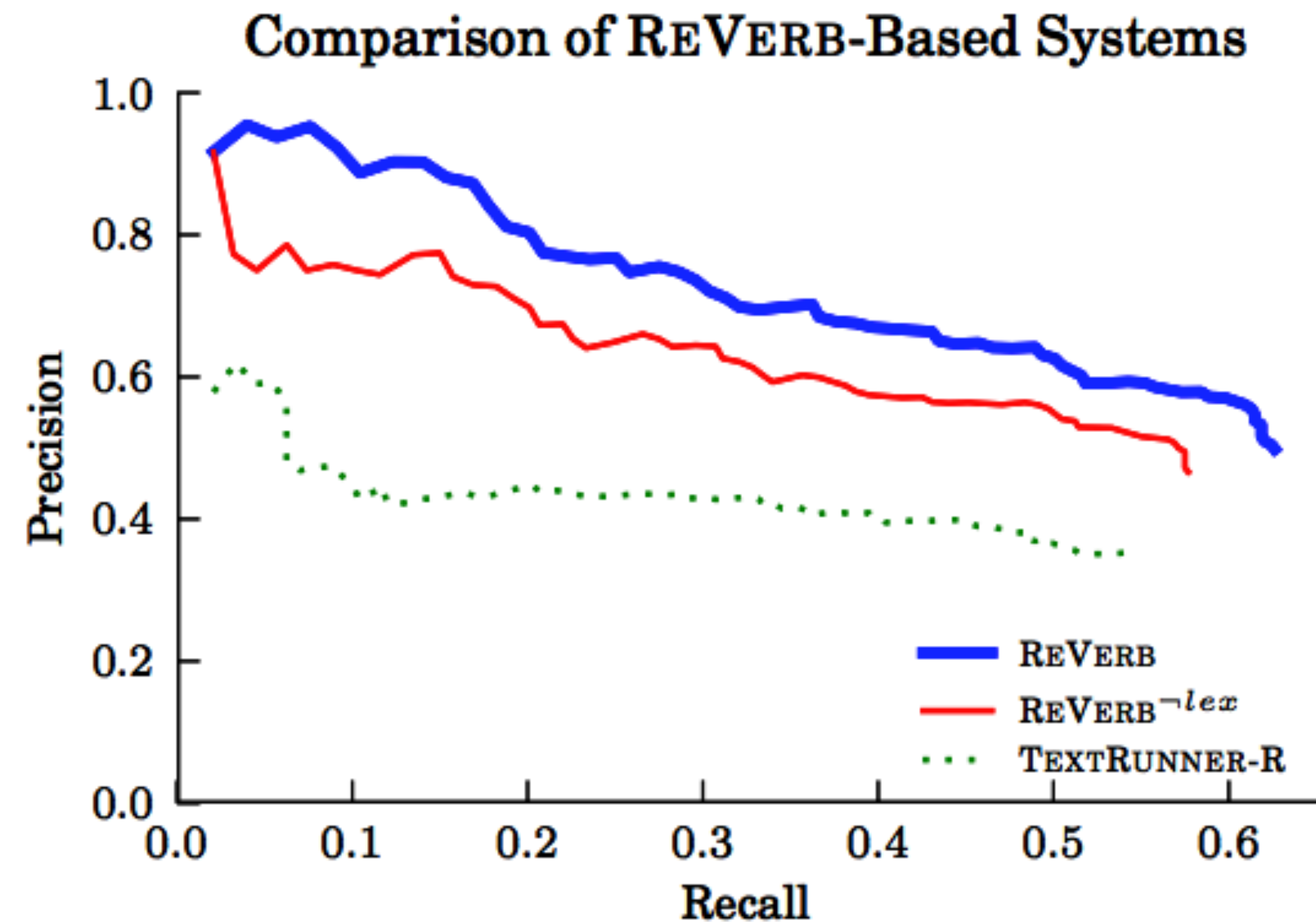
# ReVerb

---

- ▶ For each verb, identify the longest sequence of words following the verb that satisfy a POS regex ( $V \cdot^* P$ ) and which satisfy heuristic lexical constraints on specificity
- ▶ Find the nearest arguments on either side of the relation

# ReVerb

- ▶ For each verb, identify the longest sequence of words following the verb that satisfy a POS regex ( $V .^* P$ ) and which satisfy heuristic lexical constraints on specificity
- ▶ Find the nearest arguments on either side of the relation
- ▶ Annotators labeled relations in 500 documents to assess recall





# QA from Open IE

(a) **CCG parse** builds an underspecified semantic representation of the sentence.

Former	municipalities	in	Brandenburg
$N/N$	$N$	$N \setminus N/NP$	$NP$
$\lambda f \lambda x. f(x) \wedge former(x)$	$\lambda x. municipalities(x)$	$\lambda f \lambda x \lambda y. f(y) \wedge in(y, x)$	$Brandenburg$
$\xrightarrow{>}$		$\xrightarrow{>}$	
$N$		$N \setminus N$	
$\lambda x. former(x) \wedge municipalities(x)$		$\lambda f \lambda y. f(y) \wedge in(y, Brandenburg)$	
		$\xrightarrow{<}$	
$N$			
$l_0 = \lambda x. former(x) \wedge municipalities(x) \wedge in(x, Brandenburg)$			

(b) **Constant matches** replace underspecified constants with Freebase concepts

$$l_0 = \lambda x. former(x) \wedge municipalities(x) \wedge in(x, Brandenburg)$$

$$l_1 = \lambda x. former(x) \wedge municipalities(x) \wedge in(x, Brandenburg)$$

$$l_2 = \lambda x. former(x) \wedge municipalities(x) \wedge location.containedby(x, Brandenburg)$$

$$l_3 = \lambda x. former(x) \wedge OpenRel(x, Municipality) \wedge location.containedby(x, Brandenburg)$$

$$l_4 = \lambda x. OpenType(x) \wedge OpenRel(x, Municipality) \wedge location.containedby(x, Brandenburg)$$

# Takeaways

---

- ▶ SRL/AMR: handle a bunch of phenomena, but more or less like syntax++ in terms of what they represent
- ▶ Relation extraction: can collect data with distant supervision, use this to expand knowledge bases
- ▶ Slot filling: tied to a specific ontology, but gives fine-grained information
- ▶ Open IE: extracts lots of things, but hard to know how good or useful they are
  - ▶ Can combine with standard question answering
  - ▶ Add new facts to knowledge bases
- ▶ Many, many applications and techniques