

Design Challenges for Entity Linking

Xiao Ling, Sameer Singh, Daniel S. Weld

Computer Science & Engineering

UNIVERSITY *of* WASHINGTON



Entity Linking

Seattle beat Portland yesterday.

Entity Linking

Seattle beat Portland yesterday.

Entity Linking

Seattle beat Portland yesterday.



Seattle
(city)



Seattle
Sounders



Sea-Tac
(airport)



Entity Linking

Seattle beat Portland yesterday.



Seattle
(city)



Seattle
Sounders



Sea-Tac
(airport)



WIKIPEDIA
The Free Encyclopedia

~3-4 M
entries



Applications

- Relation Extraction
(e.g. Koch et al. 2014)
- Coreference Resolution
(e.g. Hajishirzi et al. 2013, Durrett & Klein 2014)
- Question Answering
(e.g. Sun et al. 2015)
- Web Search
(e.g. Knowledge Graph)
- many others...
(see Shen et al. 2014; Roth et al. 2014)

Ambiguity

- **Seattle** beat Portland yesterday.
- **Seattle** scores high in the latest report of startup hubs.

Ambiguity

Seattle Sounders



- **Seattle** beat Portland yesterday.
- **Seattle** scores high in the latest report of startup hubs.



Ambiguity

Seattle Sounders



- **Seattle** beat Portland yesterday.

- **Seattle** scores high in the latest report of startup hubs.

Seattle (city)

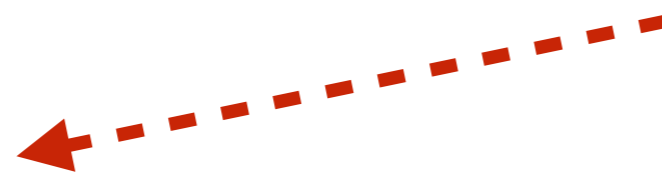
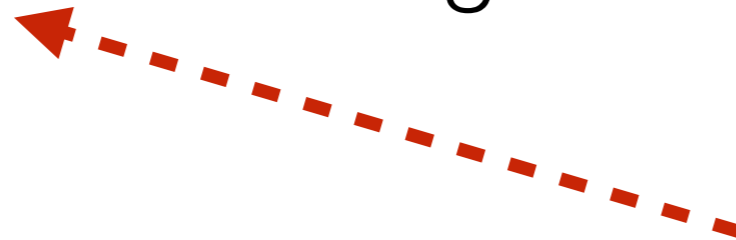


Variability

- **Seattle** scores high in the latest report of startup hubs.
- **The Emerald City** Council To Make Decision on Antibiotic Resolution

Variability

- **Seattle** scores high in the latest report of startup hubs.



Seattle (city)



- **The Emerald City** Council To Make Decision on Antibiotic Resolution

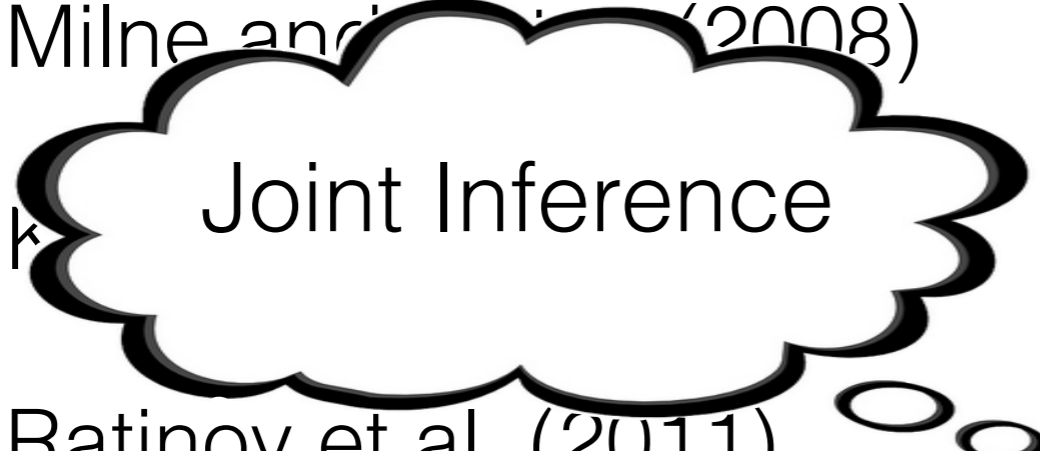

Related Work

- Cucerzan (2007)
- Milne and Witten (2008)
- Kulkarni et al. (2009)
- Ratinov et al. (2011)
- Hoffart et al. (2011)
- Han and Sun (2012)
- He et al. (2013a)
- He et al. (2013b)
- Cheng and Roth (2013)
- Sil and Yates (2013)
- Li et al. (2013)
- Cornolti et al. (2013)
- ... many others

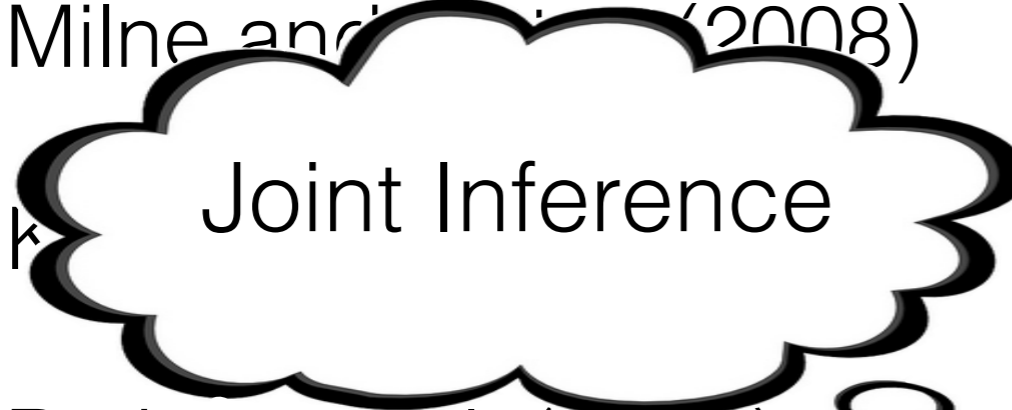

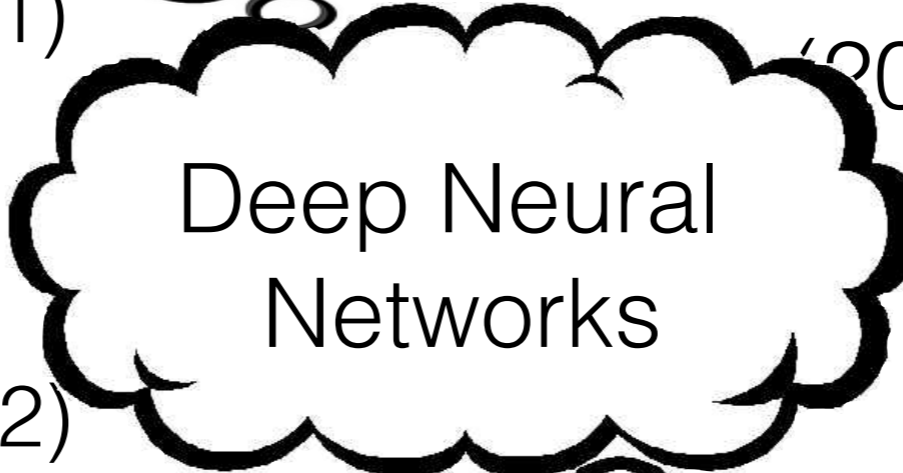
Related Work

- Cucerzan (2007)
- Milne and ... (2008)
- Joint Inference
- Ratinov et al. (2011)
- Hoffart et al. (2011)
- Han and Sun (2012)
- He et al. (2013a)
- He et al. (2013b)
- Cheng and Roth (2013)
- Sil and Yates (2013)
- Li et al. (2013)
- Cornolti et al. (2013)
- ... many others

Related Work

- Cucerzan (2007)
- Milne and ... (2008)
-  Joint Inference
- Ratinov et al. (2011)
- Hoffart et al. (2011)
- Han and Sun (2012)
- He et al. (2013a)
- He et al. (2013b)
-  Learning to rank (2013)
- Chen (2013)
- Sil and Yates (2013)
- Li et al. (2013)
- Cornolti et al. (2013)
- ... many others

Related Work

- Cucerzan (2007)
- Milne and... (2008)
-  Joint Inference
- Ratinov et al. (2011)
- Hoffart et al. (2011)
- Han and Sun (2012)
- He et al. (2013a)
- He et al. (2013b)
-  Learning to rank (2013)
- Chen (2013)
- Sil and Yates (2013)
- ... (2013)
-  Deep Neural Networks al. (2013)
- ... many others

Popular Data Sets

	Datase	# of Mentions	Knowledge Base
UIUC	ACE	244	Wikipedia
	MSNBC	654	Wikipedia
AIDA (Hoffart et al. 2011)	AIDA-D	5917	Yago
	AIDA-T	5616	Yago
TAC KBP	TAC09	3904	Wikipedia 2008
	TAC10	2250	Wikipedia 2008
	TAC10T	1500	Wikipedia 2008
	TAC11	2250	Wikipedia 2008
	TAC12	2226	Wikipedia 2008

Unfortunately...

	ACE	MSNBC	AIDA-D	AIDA-T	KBP09	KBP10	KBP10T	KBP11	KBP12
Cucerzan (2007)		✓							
Milne & Witten (2008)									
Kulkarni et al. (2009)		✓							
Ratinov et al. (2011)	✓	✓							
Hoffart et al. (2011)				✓					
Han & Sun (2012)					✓				
He et al. (2013a)				✓		✓			
He et al. (2013b)	✓	✓							
Cheng & Roth (2013)	✓	✓						✓	
Sil & Yates (2013)	✓	✓		✓					
Li et al. (2013)				✓	✓				
Cornolti et al. (2013)		✓		✓					
TAC-KBP participants					✓	✓	✓	✓	✓

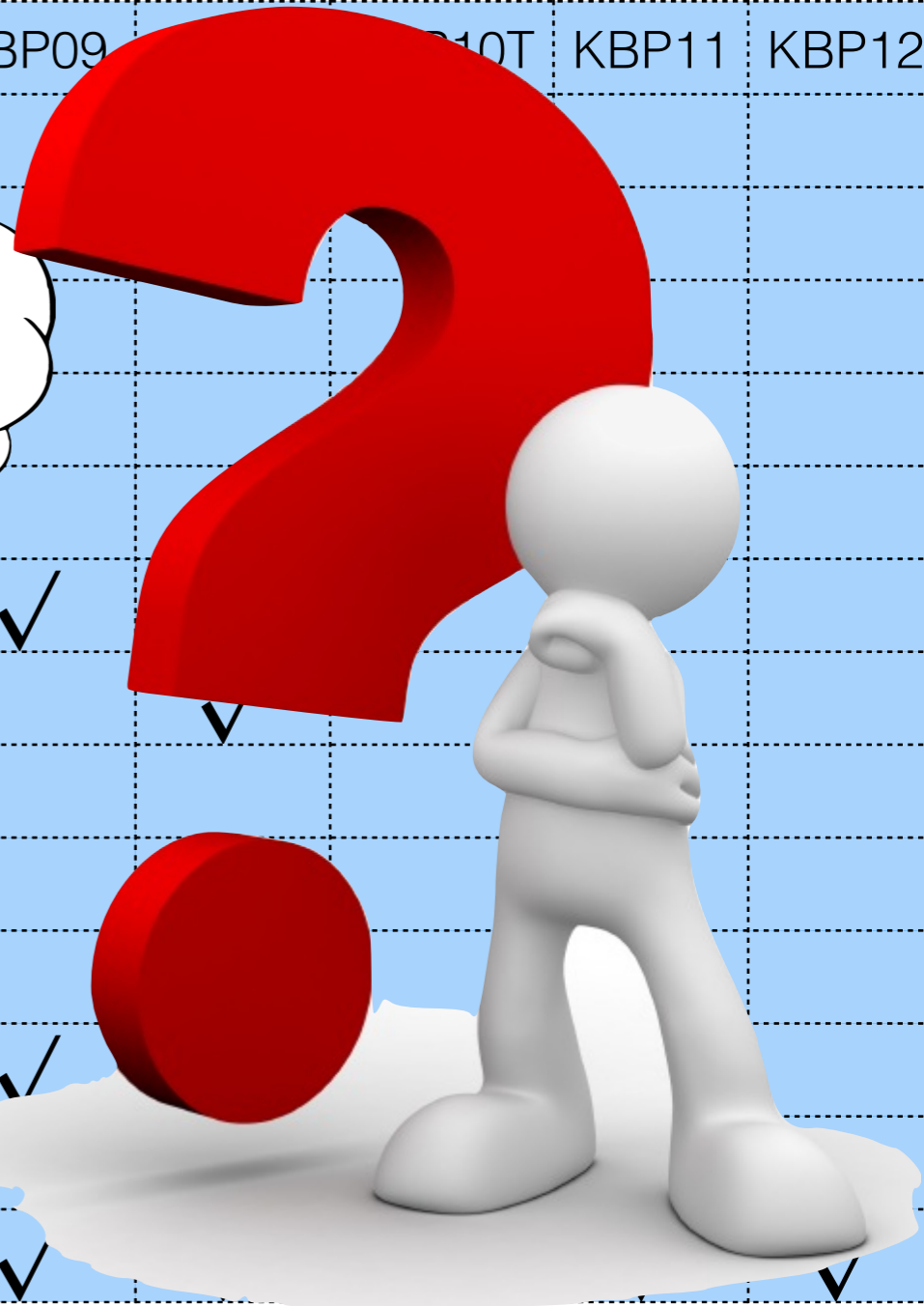
Unfortunately ...

	ACE	MSNBC	AIDA-D	AIDA-T	KBP09	KBP10T	KBP11	KBP12
Cucerzan (2007)		✓						
Milne & Witter								
Kulkarni								
Han & Sun (2012)					✓			
He et al. (2013a)				✓				
He et al. (2013b)								
Cheng & Roth (2013)								
Sil & Yates (2013)				✓				
Li et al. (2013)				✓				
Cornolti et al. (2013)		✓		✓				
TAC-KBP participants					✓			✓

Joint Inference

Learning to rank

Deep Neural Networks



Metonymy

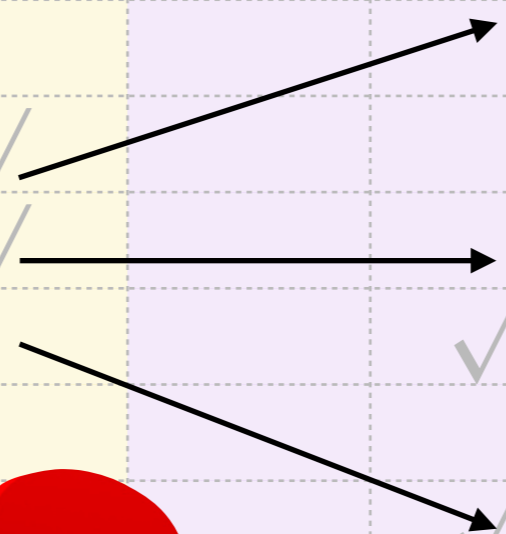
	ACE	MSNBC	AIDA-D	AIDA-T	KBP09	KBP10	KBP10T	KBP11	KBP12
Cucerzan (2007)		✓							
Milne & Witten (2008)									
Kulkarni et al. (2009)		✓							
Roth et al. (2010)		✓							
He et al. (2013a)				✓					
He et al. (2013b)	✓	✓							
Cheng & Roth (2013)	✓	✓						✓	
Sil & Yates (2013)	✓	✓		✓					
Li et al. (2013)				✓	✓				
Cornolti et al. (2013)		✓		✓					
TAC-KBP participants					✓	✓	✓	✓	✓

... Moscow 's as yet undisclosed proposals ...

Moscow (city)

Russia (country)

Government of Russia



Nested Entities

	ACE	MSNBC	AIDA-D	AIDA-T	KBP09	KBP10	KBP10T	KBP11	KBP12
Cucerzan (2007)		✓							
Milne & Witten (2008)									
Kulkarni et al. (2009)		✓							
Ratinov et al. (2011)	✓	✓							
Hoffart et al. (2011)	✓	✓							
Han & Sun (2012)					✓				
He et al. (2013a)				✓		✓			
He et al. (2013b)	✓	✓							
Cheng & Roth (2013)	✓	✓							✓
Sil & Yates (2013)	✓	✓							
Li et al. (2013)				✓	✓				
Cornolti et al. (2013)				✓					
TAC-KBP participants					✓	✓	✓	✓	✓

Green Party of the US

Florida Green Party

Green Party of Florida

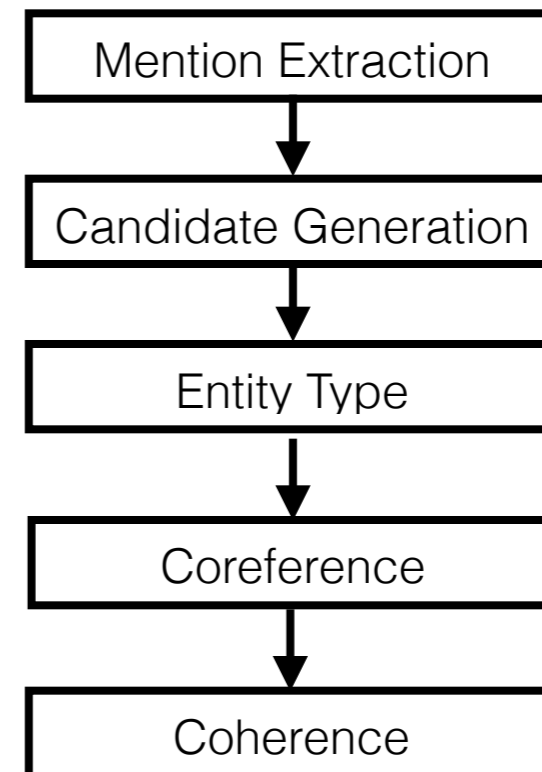


Contributions

- **Vinculum**: a simple, deterministic, modular EL sys.
- comprehensive evaluation over nine data sets
 - **candidate conditional prob.** can work quite well
 - **entity types** are important to the final performance
 - comparable results with two **state-of-the-art** sys.

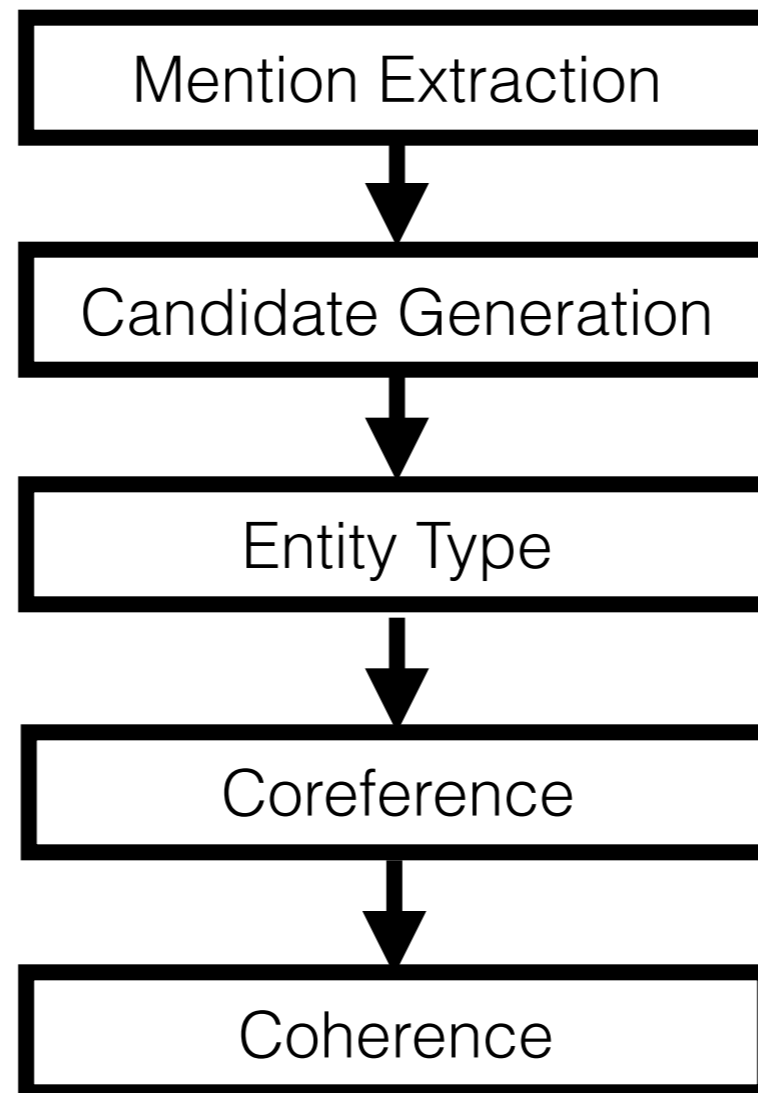
Agenda

- Introduction
- Vinculum
- Experiments
- Conclusion



Vinculum Architecture

Input: Seattle beat Portland yesterday.



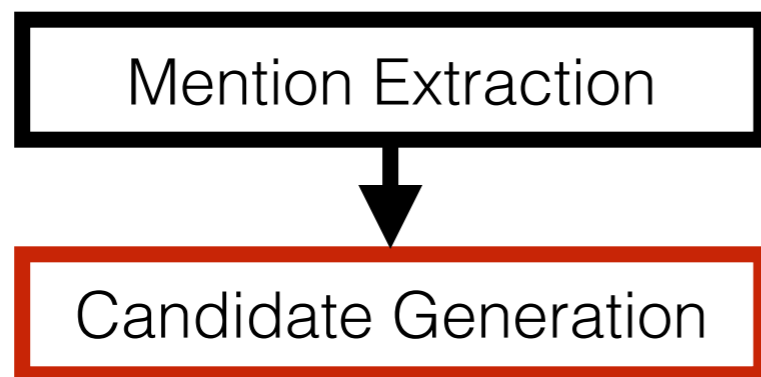
Mention Extraction

Seattle beat Portland yesterday.

Mention Extraction

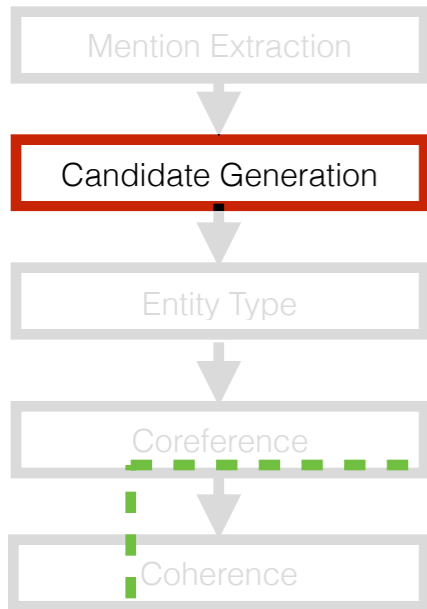
Candidate Generation

Seattle beat Portland yesterday.



Candidate Entities

- Seattle (city)
- Seattle Sounders
- Seattle-Tacoma (airport)



Conditional probability

... capital of the state of Washington .



In 1990, Washington starred as Bleek Gilliam ...

Washington refused to run for a third term ...

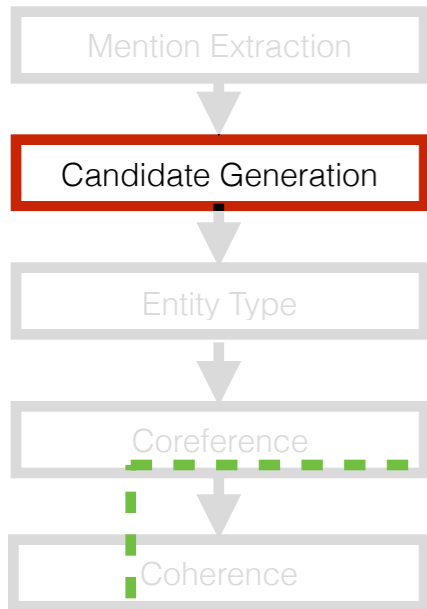


... Washington ...



WIKIPEDIA
The Free Encyclopedia

$$p(e | m) = \frac{\# [m \rightarrow e]}{\# m}$$



Conditional probability

... capital of the state of Washington .



In 1990, Washington starred as Bleek Gilliam ...

Washington refused to run for a third term ...



... Washington ...

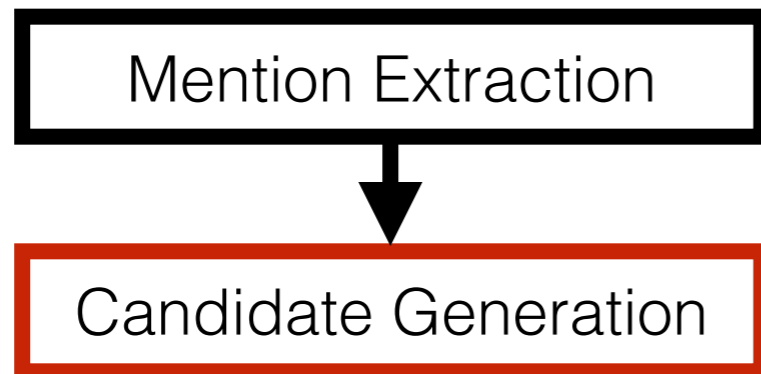


WIKIPEDIA
The Free Encyclopedia

$$p(\text{Washington Seal} \mid \text{"Washington"}) = \frac{\# \text{"W"} \rightarrow \text{Washington Seal}}{\# \text{"W"}}$$

Candidate Generation

Seattle beat Portland yesterday.

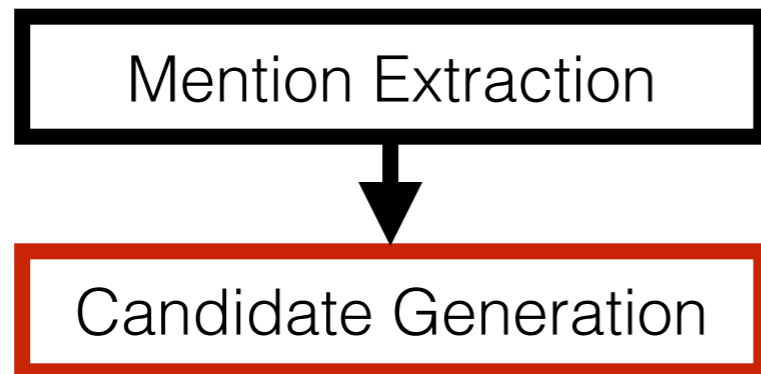


Candidate Entities

- Seattle (city)
- Seattle Sounders
- Seattle-Tacoma (airport)

Candidate Generation

Seattle beat Portland yesterday.

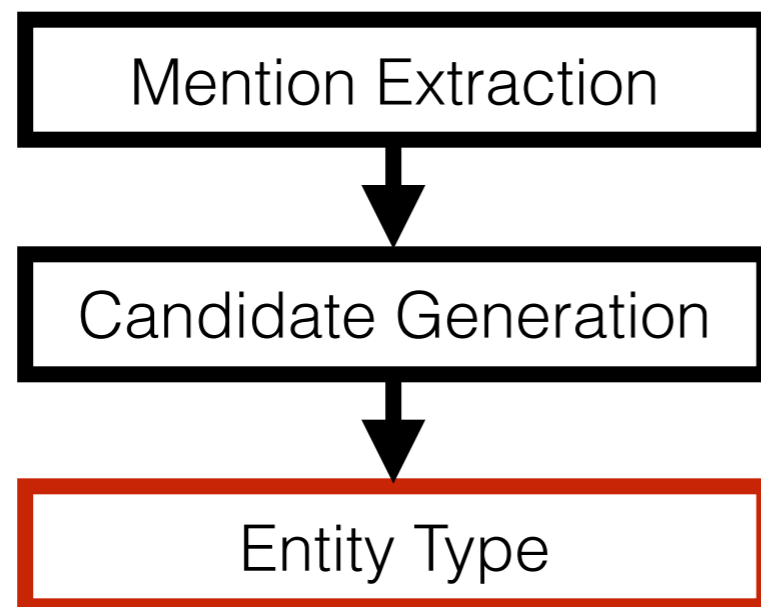


Candidate Entities

- Seattle (city) 0.6
- Seattle Sounders 0.2
- Seattle-Tacoma (airport) 0.1

Entity Types

Seattle beat Portland yesterday.

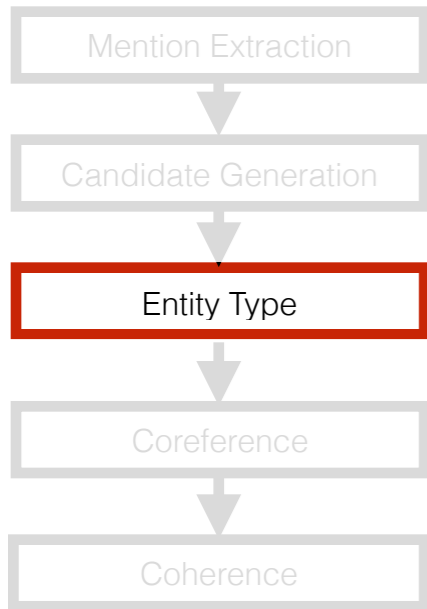


Candidate Entities

- Seattle (city) 0.6
- Seattle Sounders 0.2
- Seattle-Tacoma (airport) 0.1

Entity Type Prediction

- city 0.1
- sports_team 0.4
- facility/airport 0.1



Entity Types

Seattle beat Portland yesterday.

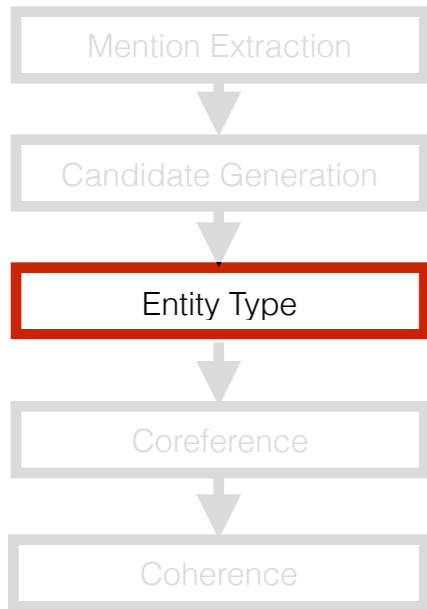
$$\begin{aligned}
 p(e \mid m) &= \sum_t p(e, t \mid m) \\
 &= \sum_t p(e \mid t, m) p(t \mid m)
 \end{aligned}$$

Candidate Entities

- Seattle (city) 0.6
- Seattle Sounders 0.2
- Seattle-Tacoma (airport) 0.1

Entity Type Prediction

- city 0.1
- sports_team 0.4
- facility/airport 0.1



Entity Types

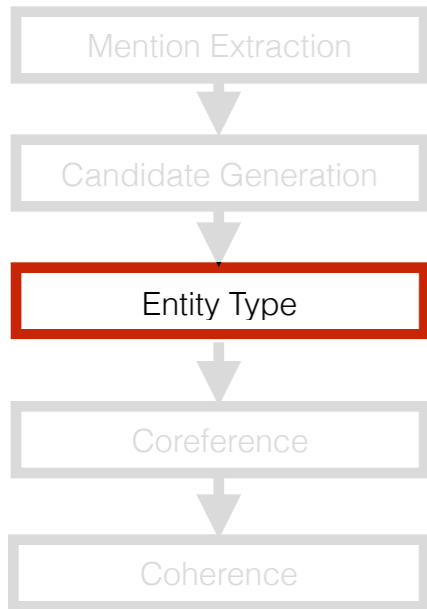
Seattle beat Portland yesterday.

Candidate Entities

- Seattle (city) 0.6
- Seattle Sounders 0.2
- Seattle-Tacoma (airport) 0.1

$$\begin{aligned}
 p(e \mid m) &= \sum_t p(e, t \mid m) \\
 &= \sum_t \mathbf{p(e \mid t, m)} p(t \mid m)
 \end{aligned}$$

$p(e \mid t, m)$: re-normalization of cond. prob.



Entity Types

Seattle beat Portland yesterday.

Candidate Entities

- Seattle (city) 0.6
- Seattle Sounders 0.2
- Seattle-Tacoma (airport) 0.1

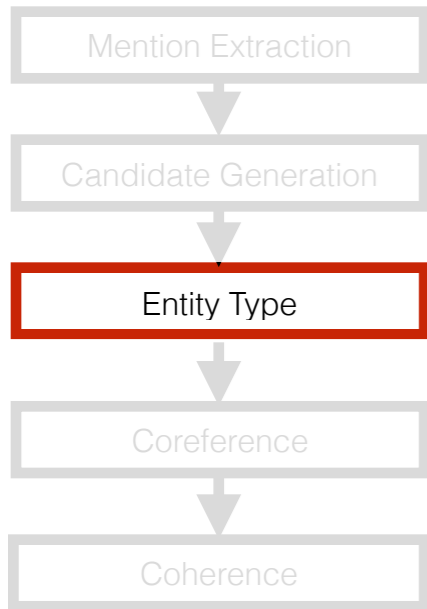
$$\begin{aligned}
 p(e \mid m) &= \sum_t p(e, t \mid m) \\
 &= \sum_t \mathbf{p(e \mid t, m)} p(t \mid m)
 \end{aligned}$$

$p(e \mid t, m)$: re-normalization of cond. prob.

e.g. $t = \mathbf{LOC}$

$$p(\text{Seattle-city} \mid \mathbf{LOC}, \text{"Seattle"}) = 0.6 / 0.7$$

$$p(\text{Sea-Tac} \mid \mathbf{LOC}, \text{"Seattle"}) = 0.1 / 0.7$$



Entity Types

Seattle beat Portland yesterday.

$$\begin{aligned}
 p(e \mid m) &= \sum_t p(e, t \mid m) \\
 &= \sum_t p(e, t \mid m) \mathbf{p(t \mid m)}
 \end{aligned}$$

Candidate Entities

- Seattle (city) 0.6
- Seattle Sounders 0.2
- Seattle-Tacoma (airport) 0.1

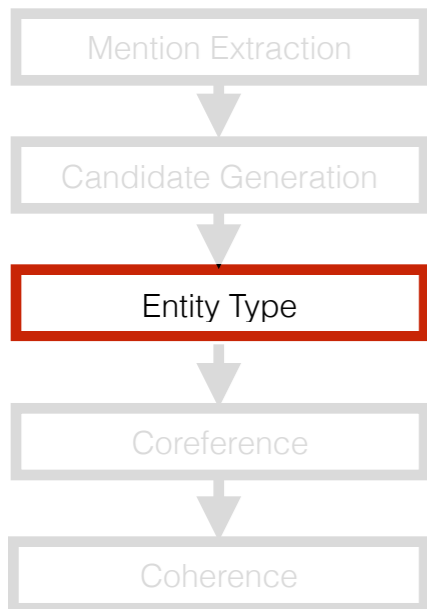
$$\mathbf{p(t \mid m)}$$

Entity Type Prediction

- city 0.1
- sports_team 0.4
- facility/airport 0.1

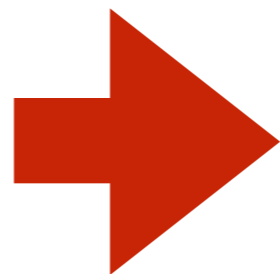
Entity Types

Seattle beat Portland yesterday.



Candidate Generation

- Seattle (city) 0.6
- **Seattle Sounders** 0.2
- Seattle-Tacoma (airport) 0.1



Entity Type

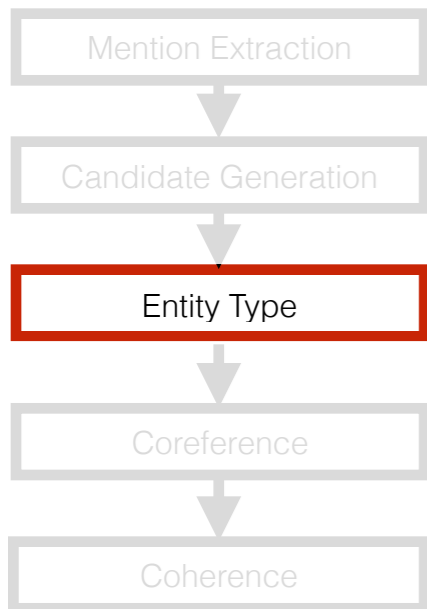
- Seattle (city) 0.2
- **Seattle Sounders** 0.4
- Seattle-Tacoma (airport) 0.1

Entity Type Prediction

- city 0.1
- **sports_team** 0.4
- facility/airport 0.1

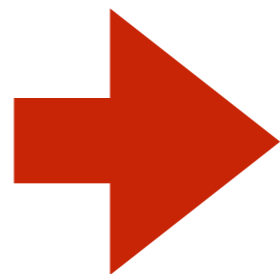
Entity Types

Seattle beat Portland yesterday.



Candidate Generation

- Seattle (city) 0.6
- **Seattle Sounders** 0.2
- Seattle-Tacoma (airport) 0.1



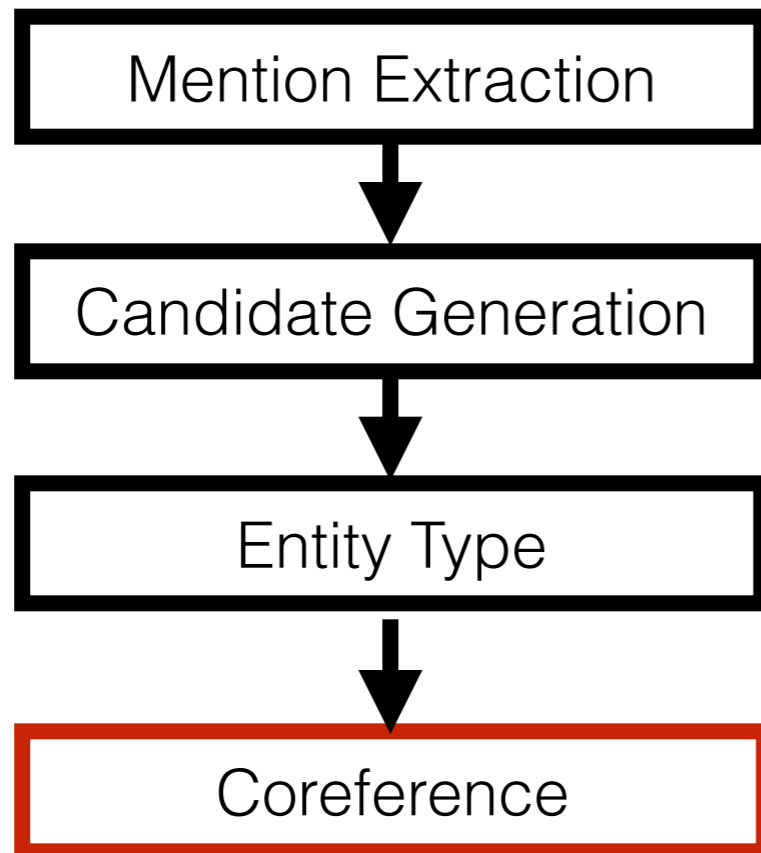
Entity Type

- **Seattle Sounders** 0.4
- Seattle (city) 0.2
- Seattle-Tacoma (airport) 0.1

Entity Type Prediction

- city 0.1
- **sports_team** 0.4
- facility/airport 0.1

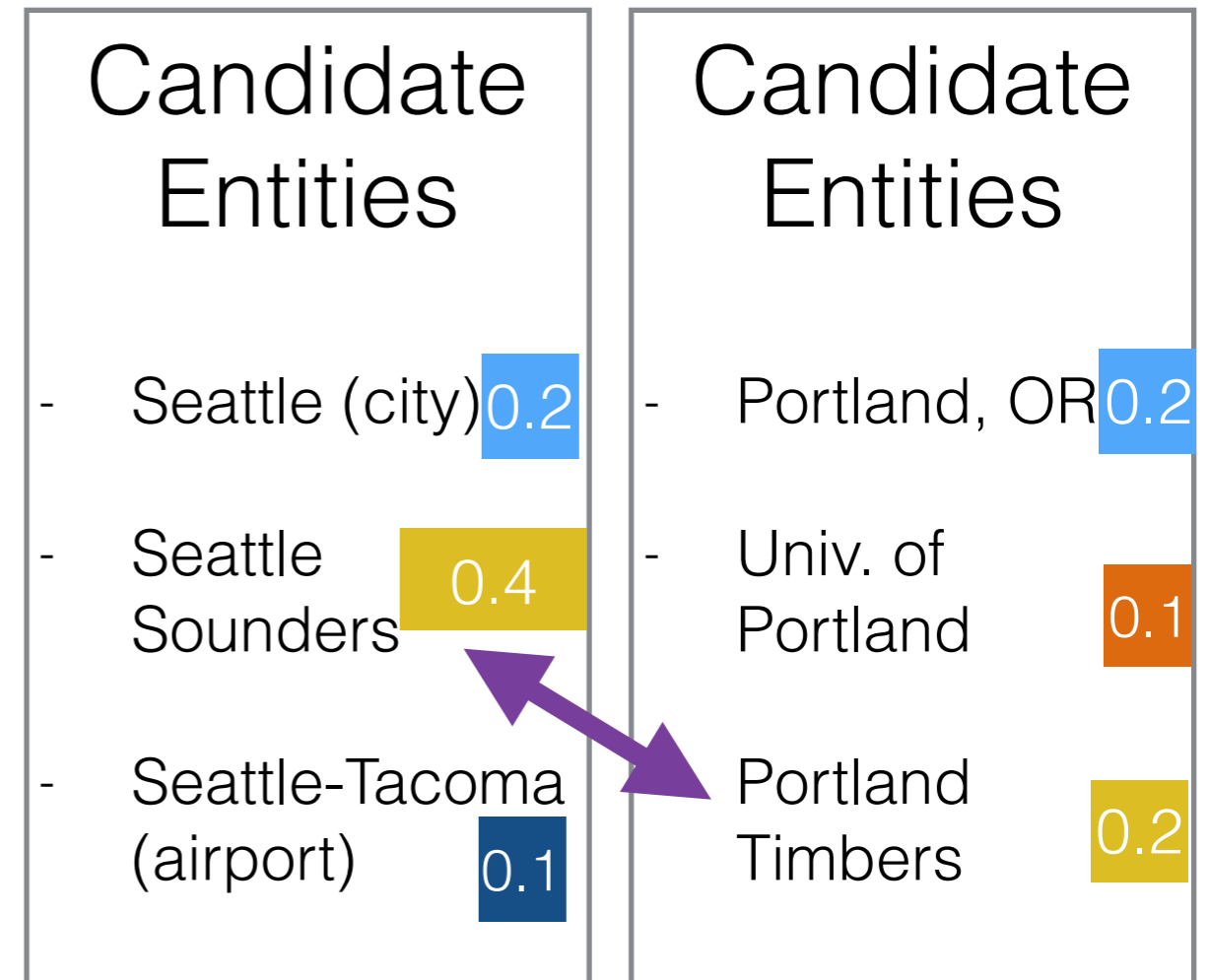
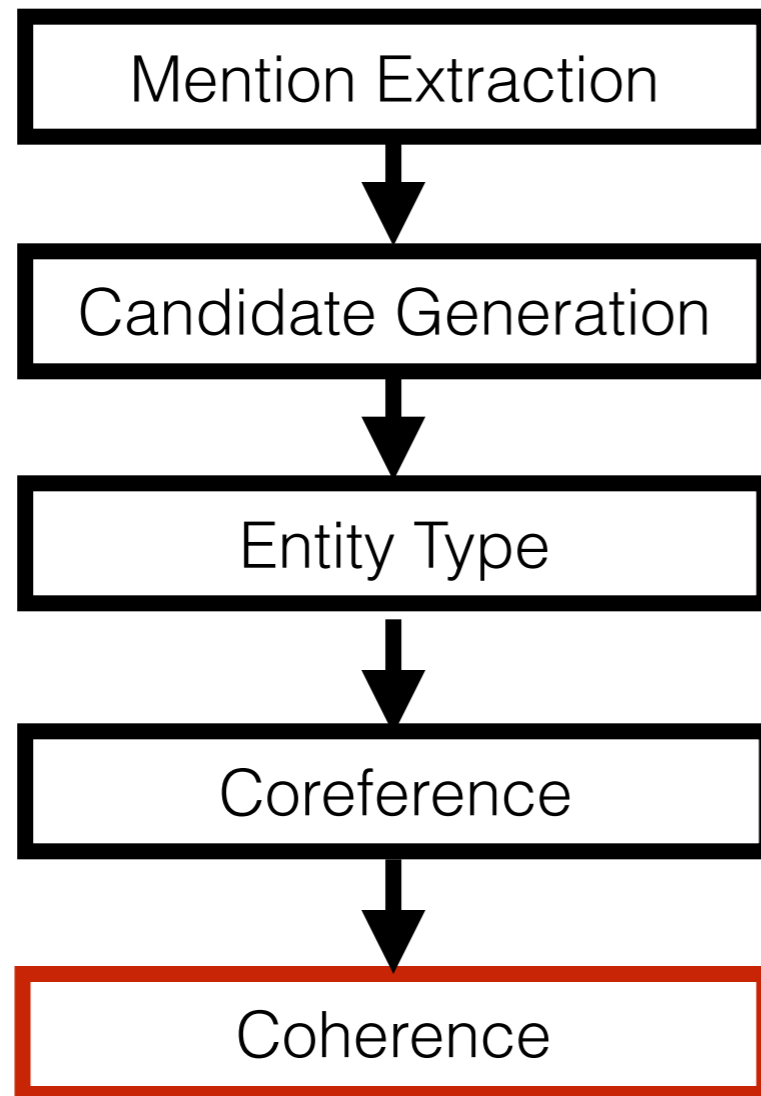
Coreference



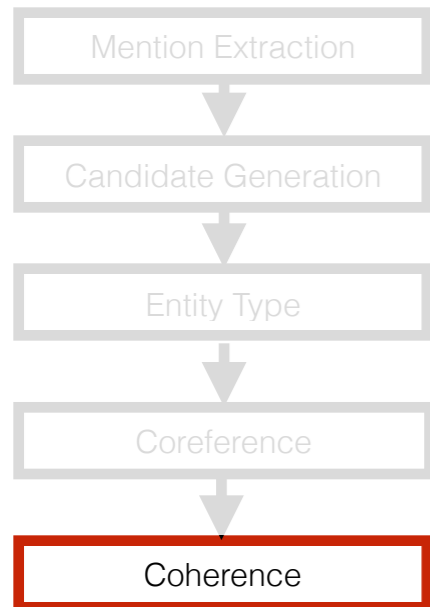
Seattle Sounders head coach Sigi Schmid has some ideas ...
Seattle beat Portland yesterday.

Coherence

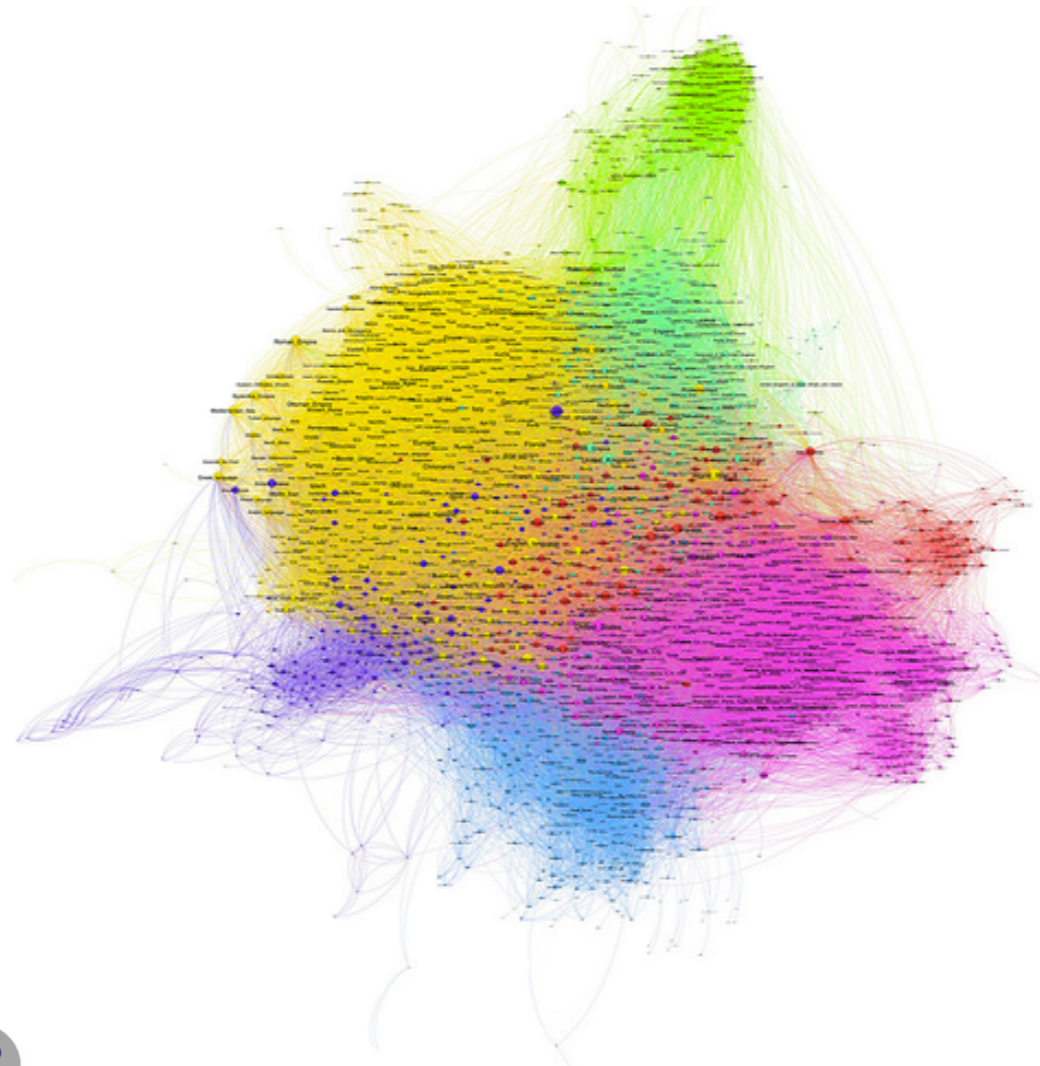
Seattle beat **Portland** yesterday.



Normalized Google Distance (NGD) (Milne & Witten, 2008)

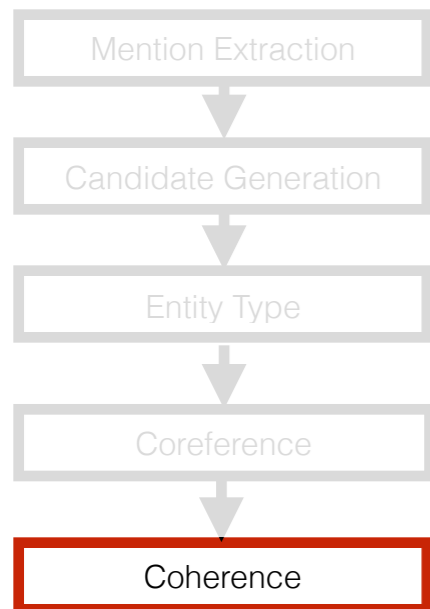


$$\phi_{NGD}(c_i, c_j) = 1 - \frac{\log(\max(|L_i|, |L_j|)) - \log(|L_i \cap L_j|)}{\log(W) - \log(\min(|L_i|, |L_j|))}$$

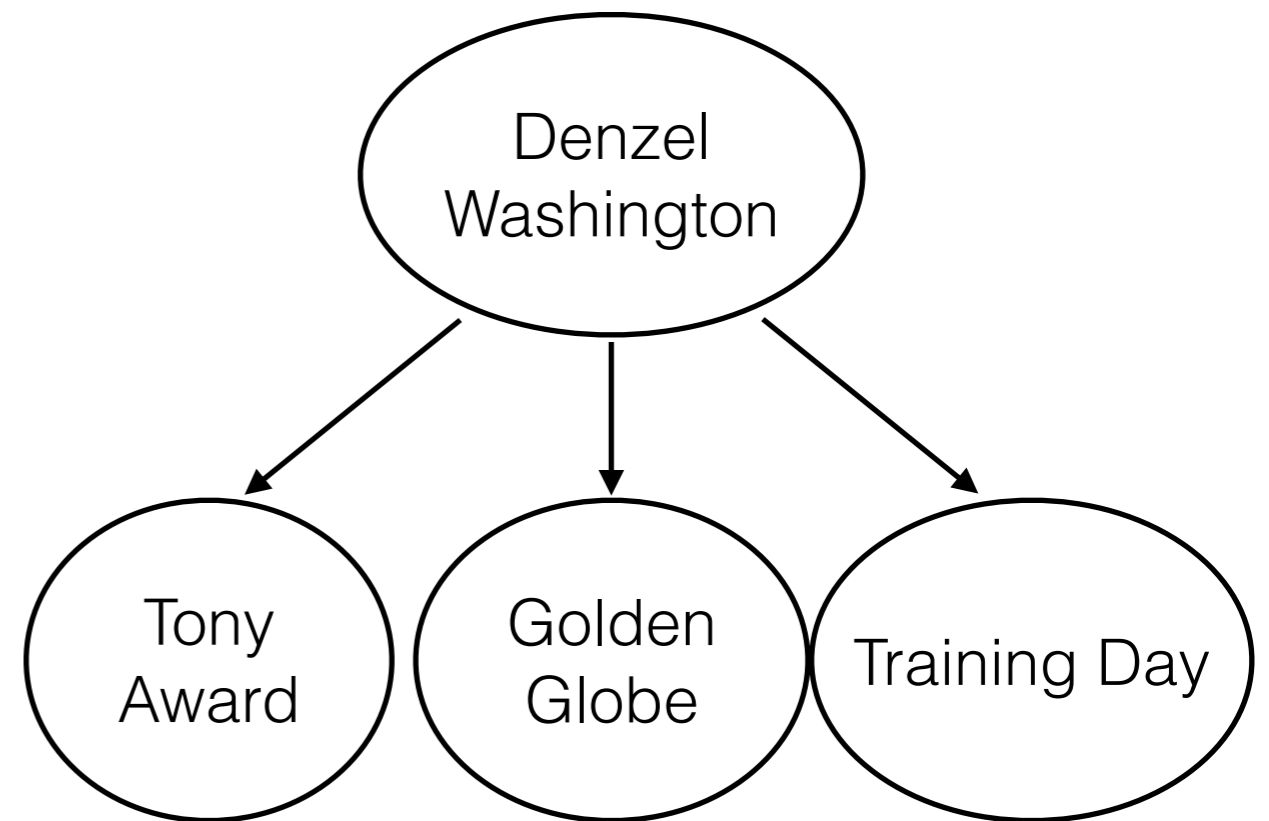
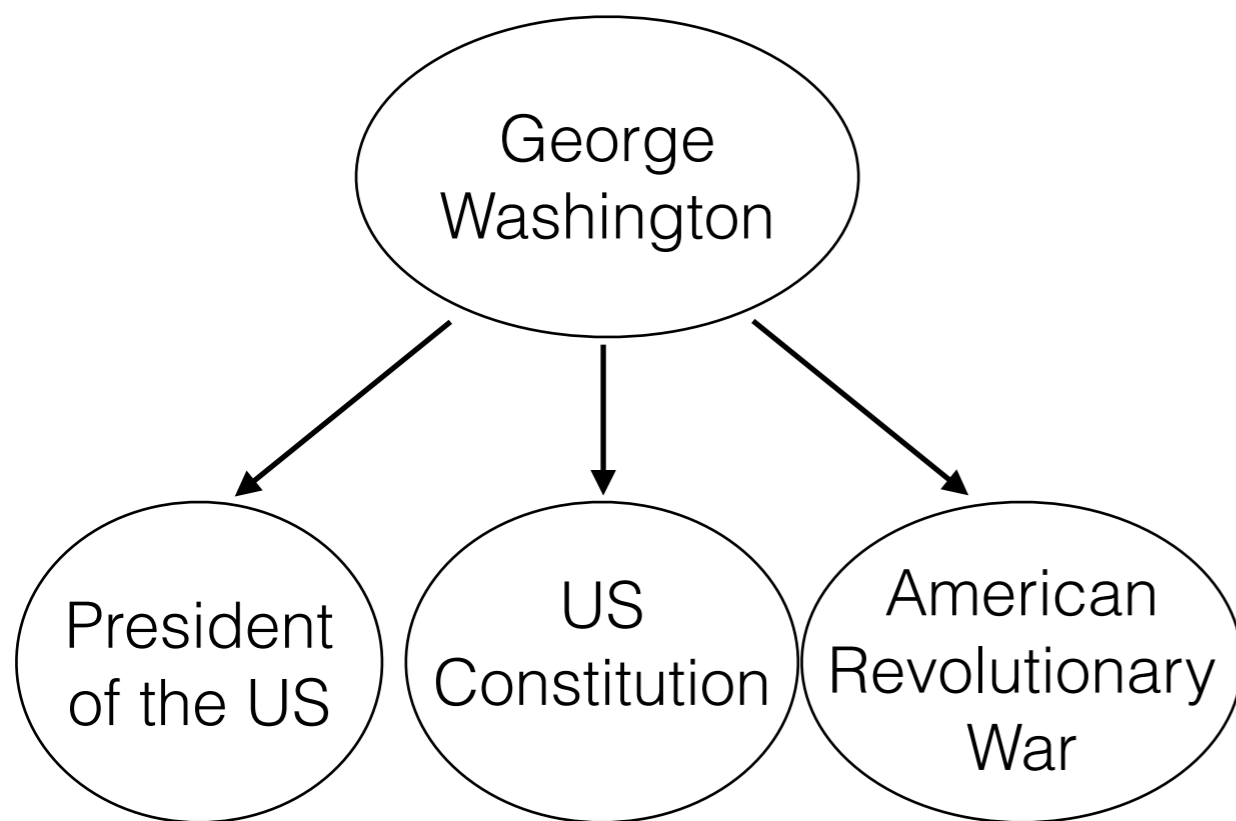


Normalized Google Distance (NGD)

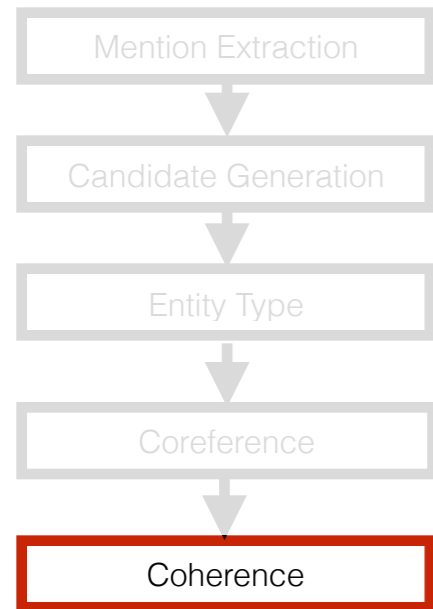
(Milne & Witten, 2008)



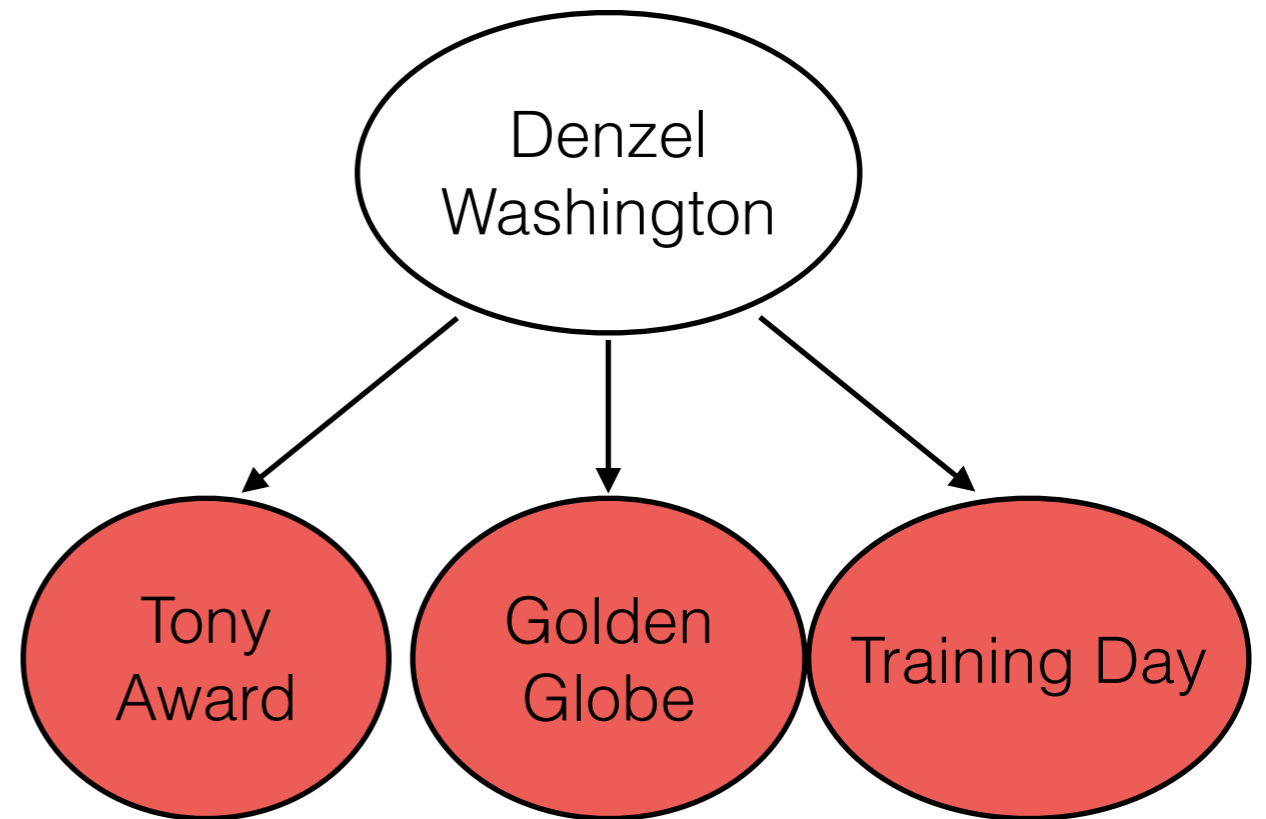
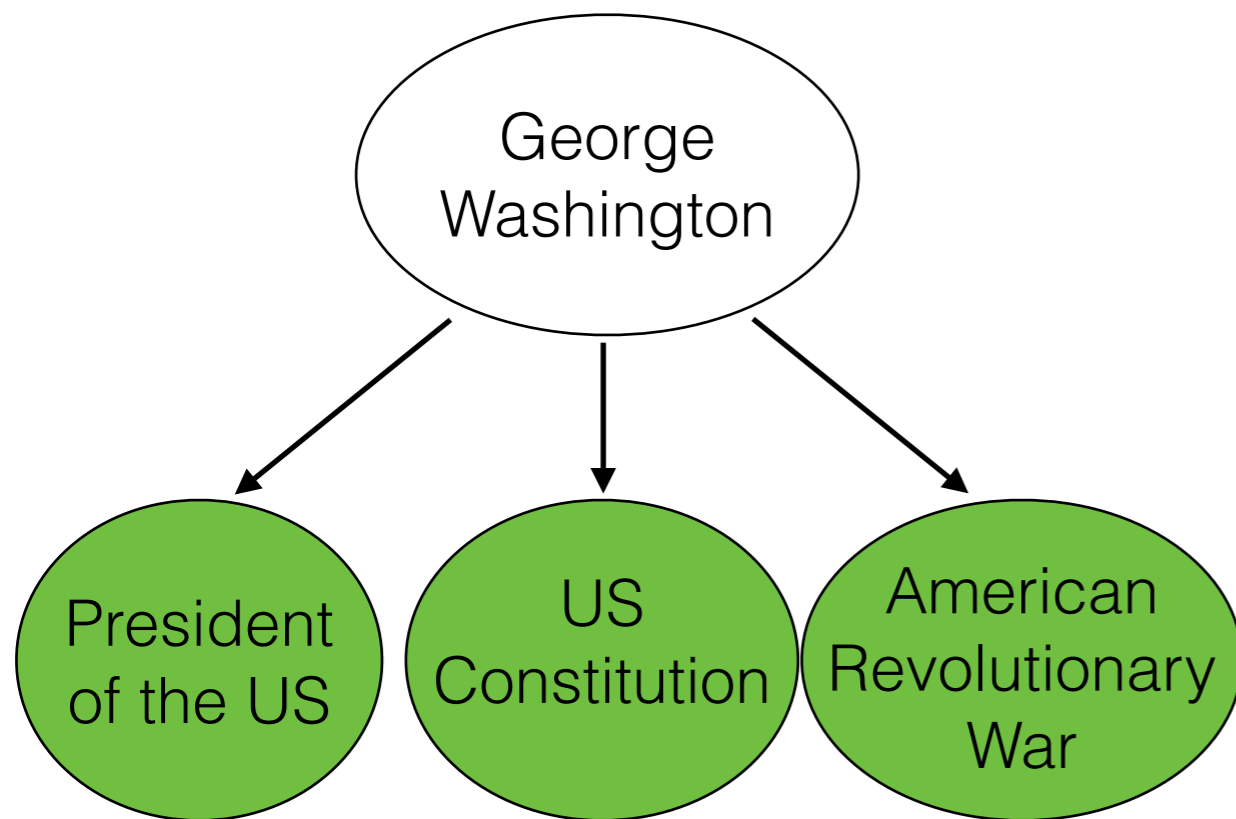
$$\phi_{NGD}(c_i, c_j) = 1 - \frac{\log(\max(|L_i|, |L_j|)) - \log(|L_i \cap L_j|)}{\log(W) - \log(\min(|L_i|, |L_j|))}$$



Normalized Google Distance (NGD) (Milne & Witten, 2008)

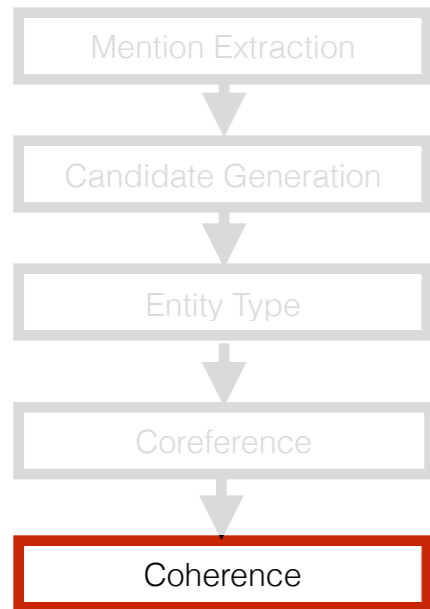


$$\phi_{NGD}(c_i, c_j) = 1 - \frac{\log(\max(|L_i|, |L_j|)) - \log(|L_i \cap L_j|)}{\log(W) - \log(\min(|L_i|, |L_j|))}$$

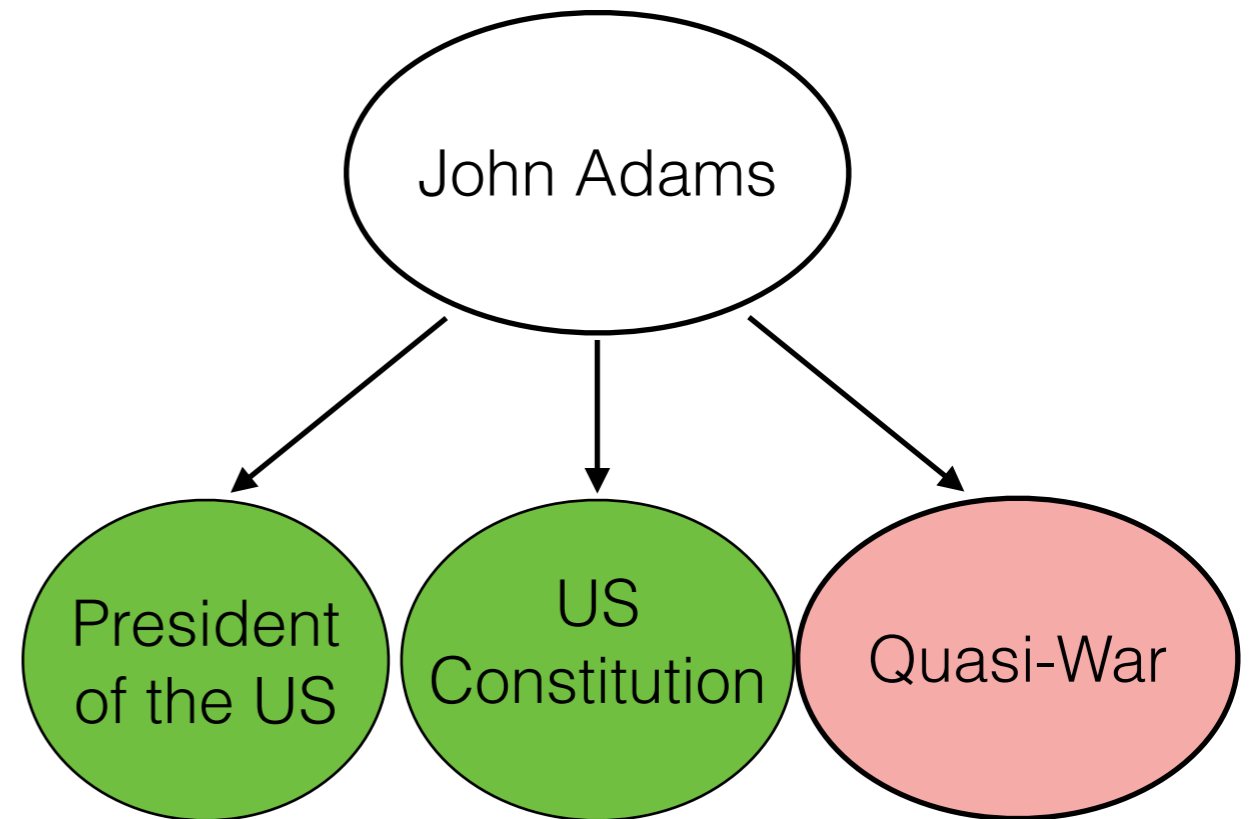
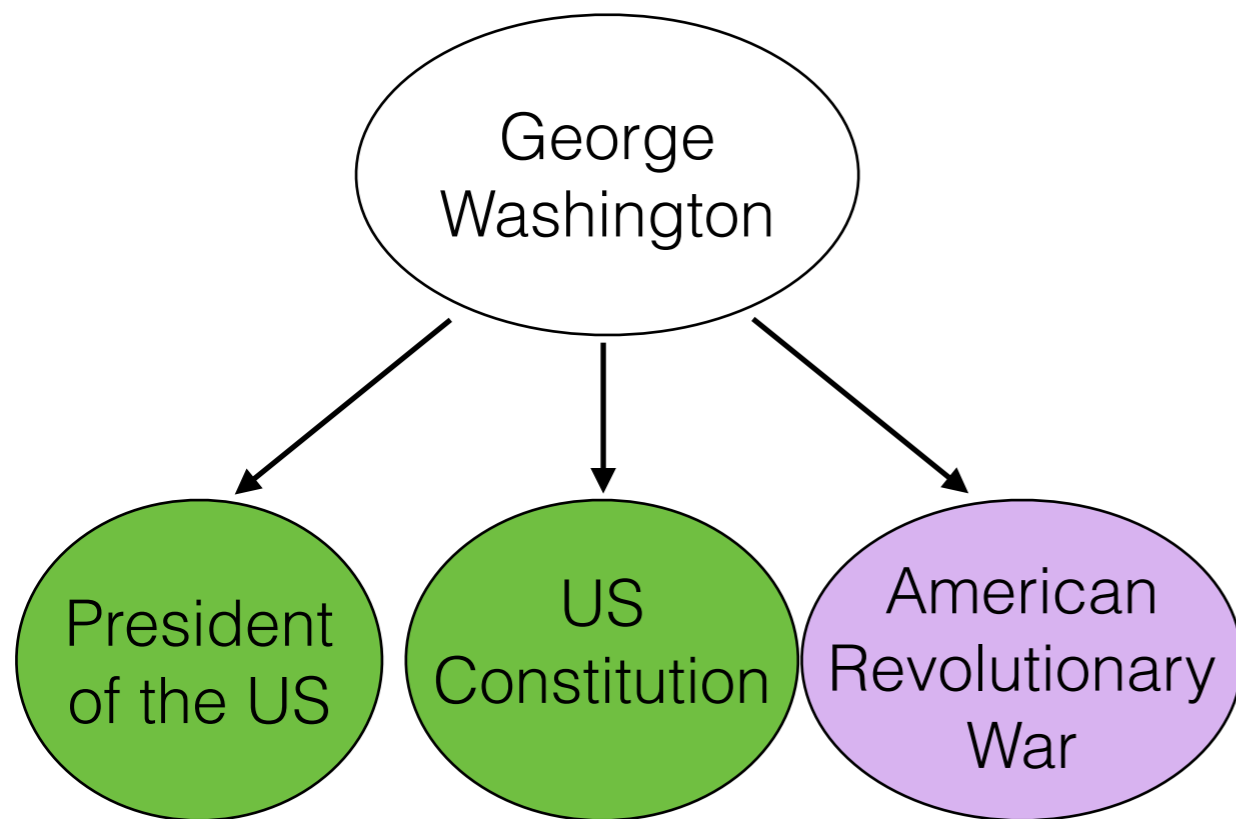


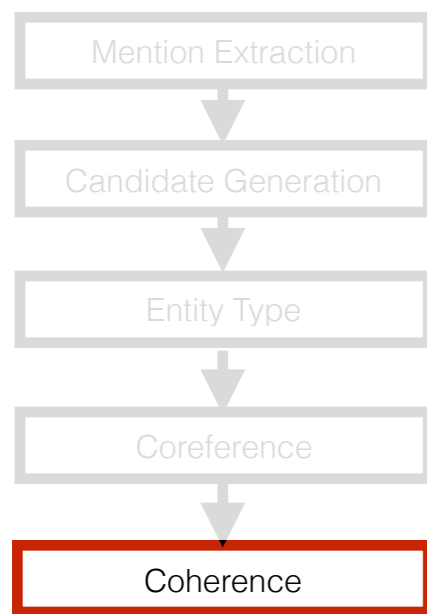
Normalized Google Distance (NGD)

(Milne & Witten, 2008)



$$\phi_{NGD}(c_i, c_j) = 1 - \frac{\log(\max(|L_i|, |L_j|)) - \log(|L_i \cap L_j|)}{\log(W) - \log(\min(|L_i|, |L_j|))}$$



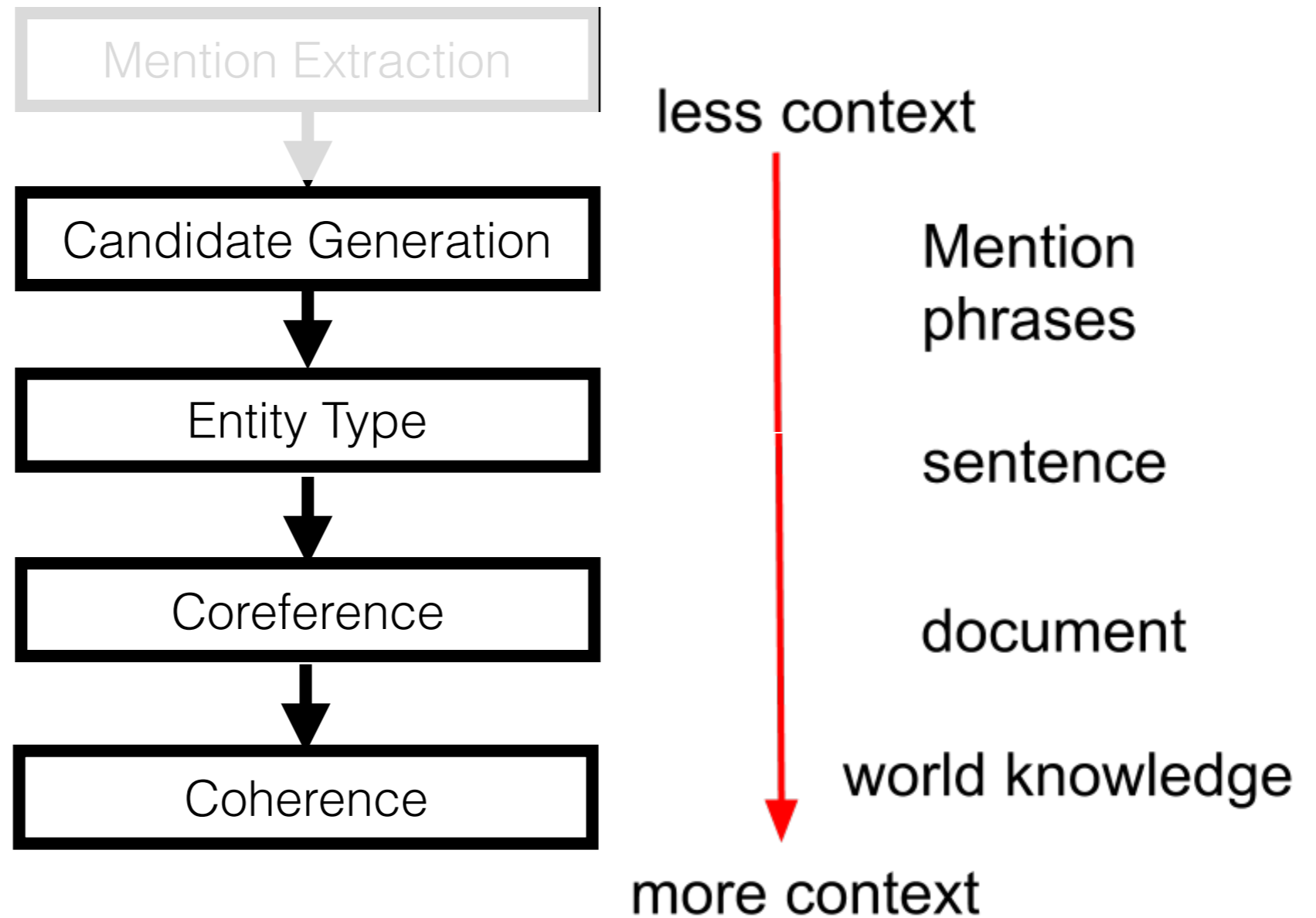


Relational Score

(Cheng & Roth, 2013)

- Relation triples from Freebase
- A binary score =
 - 1, if two entities appear in a triple
 - 0, otherwise
- E.g. (Barack Obama, birthplace, United States)
=> $r(\text{Barack Obama, United States}) = 1$

Context



Agenda

- Introduction
- Vinculum
- Experiments
- Conclusion

	ACE	MSNBC	AIDA-D	AIDA-T	KBP09	KBP1	KBP10	KBP1	KBP1
Cucerzan (2007)		✓							
Milne & Witten (2008)									
Kulkarni et al. (2009)		✓							
Ratinov et al. (2011)	✓	✓							
Hoffart et al. (2011)				✓					
Han & Sun (2012)					✓				
He et al. (2013a)				✓		✓			
He et al. (2013b)	✓	✓							
Cheng & Roth (2013)	✓	✓						✓	
Sil & Yates (2013)	✓	✓		✓					
Li et al. (2013)				✓	✓				
Cornolti et al. (2013)		✓		✓					
TAC-KBP participants					✓	✓	✓	✓	✓

Data Sets

Dataset	# of Mentions	Knowledge Base
ACE	244	Wikipedia
MSNBC	654	Wikipedia
AIDA-D	5917	Yago
AIDA-T	5616	Yago
TAC09	3904	Wikipedia 2008
TAC10	2250	Wikipedia 2008
TAC10T	1500	Wikipedia 2008
TAC11	2250	Wikipedia 2008
TAC12	2226	Wikipedia 2008

Mention
based F1

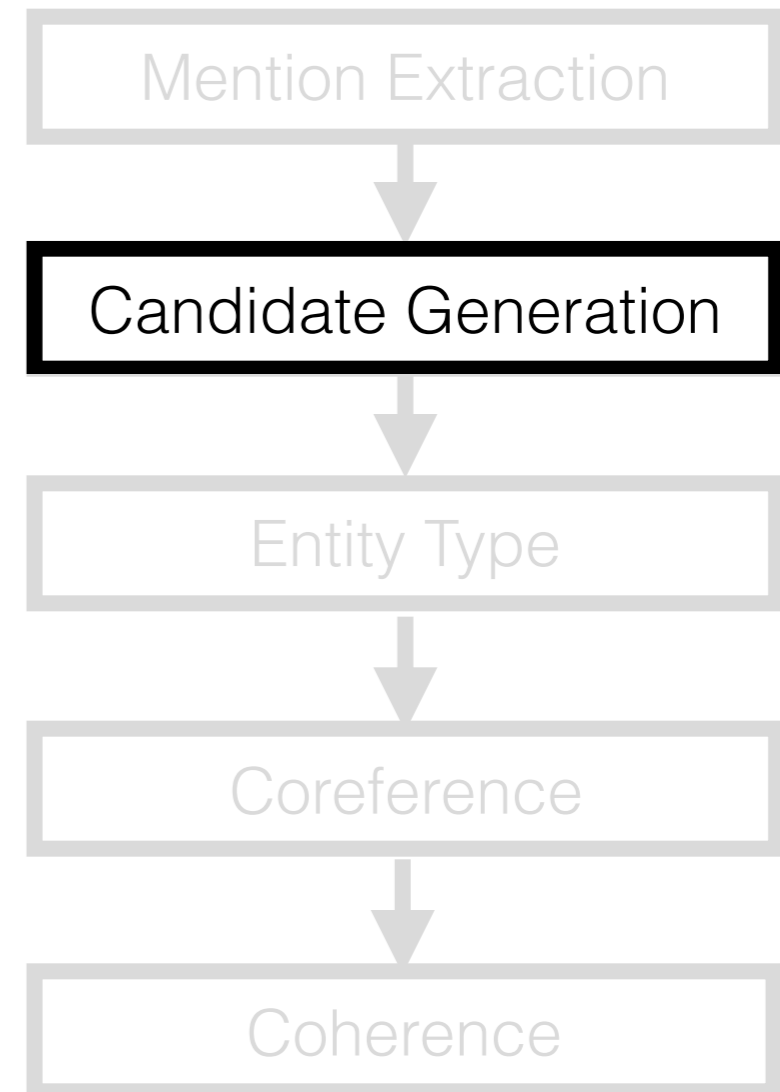
Official Eval.

Candidate Generation

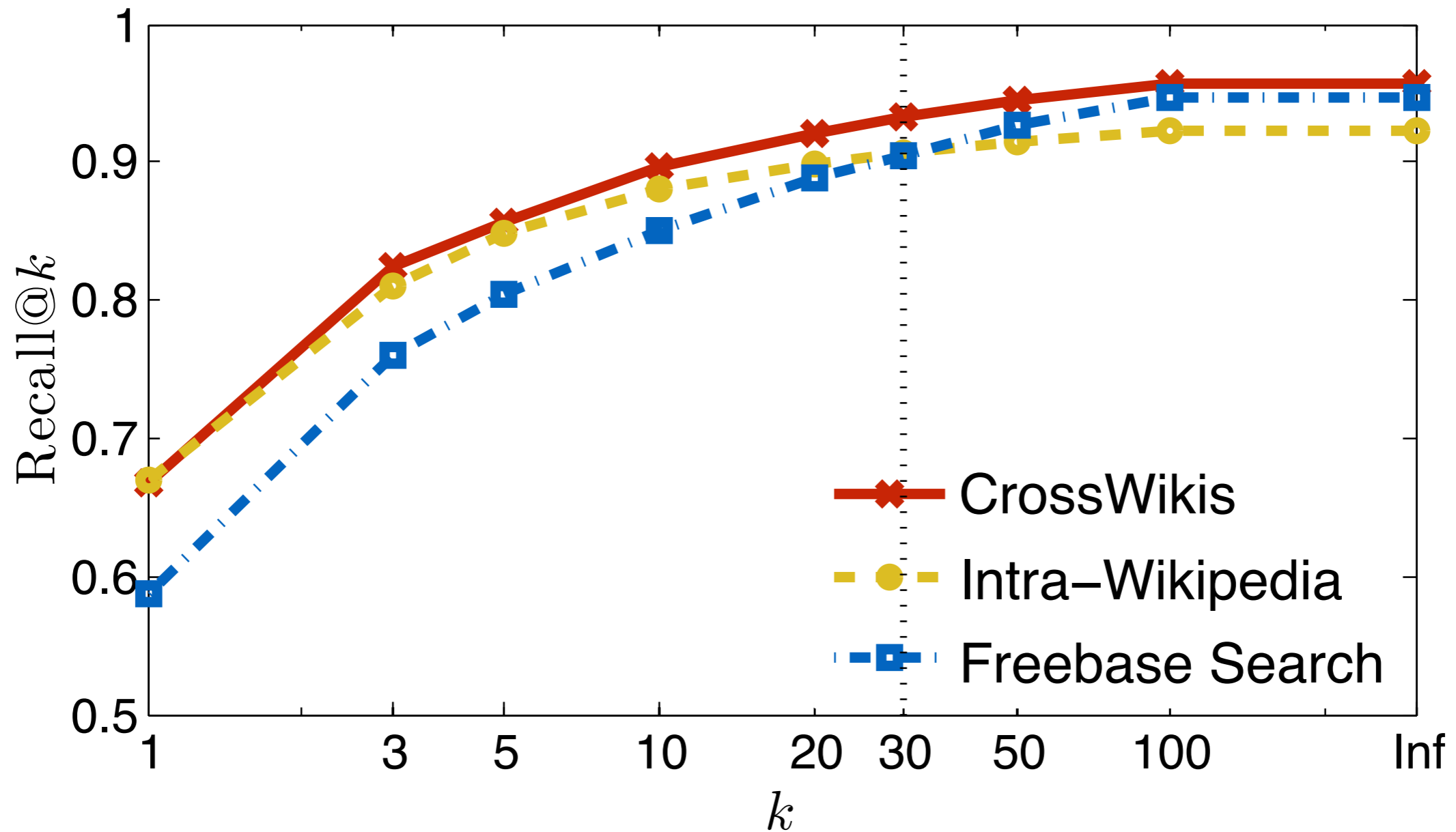
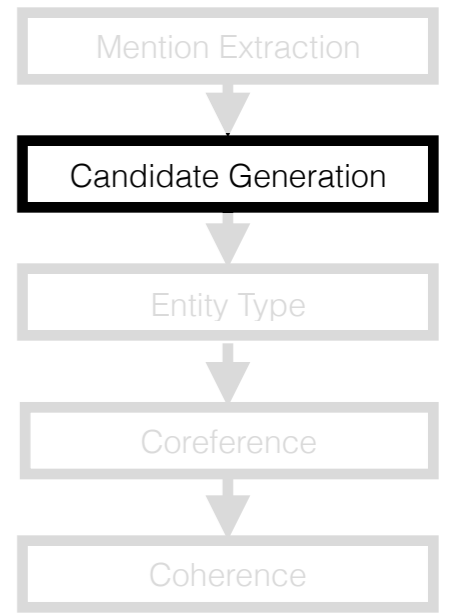
Conditional Probability $p(\mathbf{e} \mid \mathbf{m})$

e.g. $p(\text{🇺🇸} \mid \text{"Washington"})$

- intra-Wikipedia
- CrossWikis
(Spitkovsky & Chang, 2012)
- Freebase Search API



Aggregate Recall

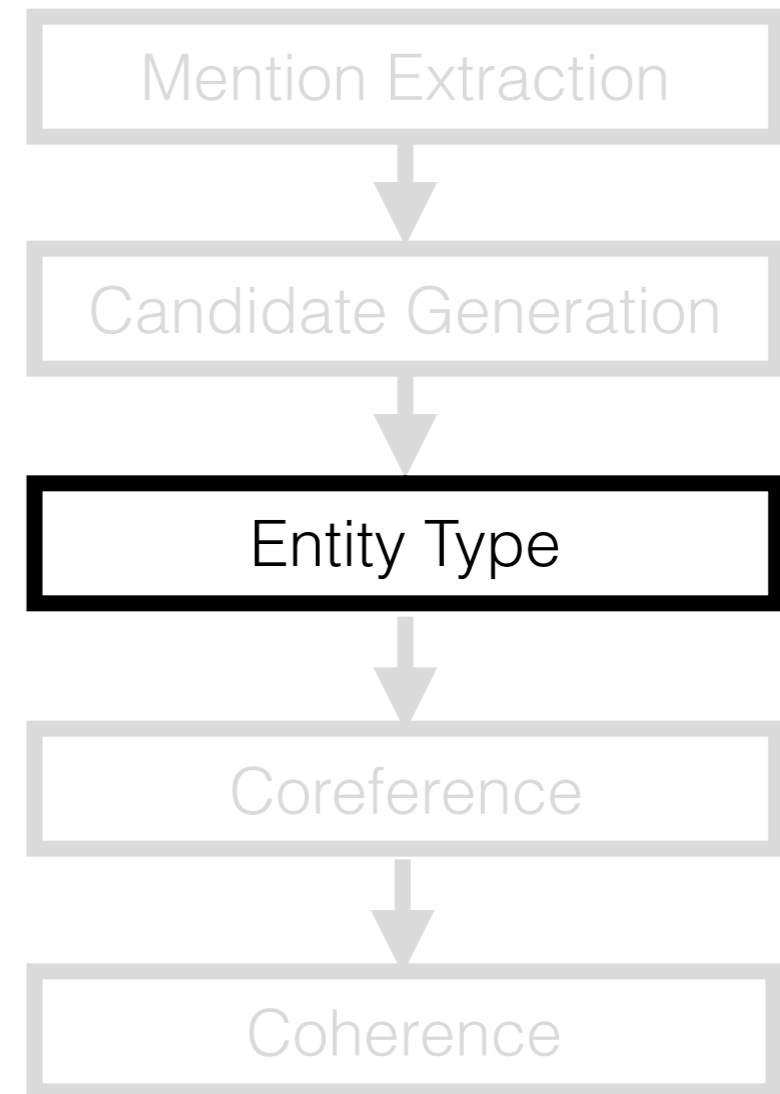


Effect of Entity Types

- Coarse-grained NER
(Stanford NER)
- Fine-grained Entity Types
(Ling & Weld, 2012)

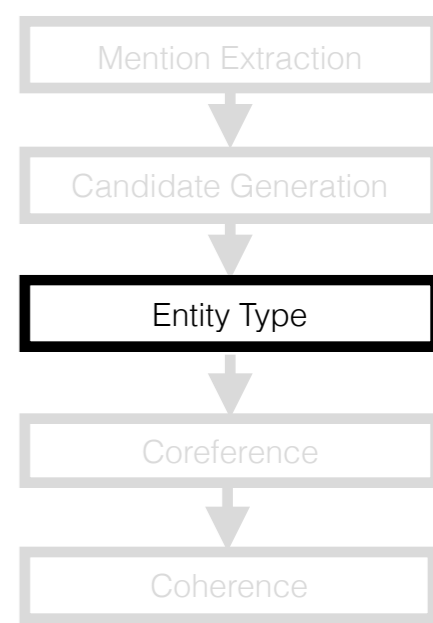
$$\begin{aligned} p(e | m) &= \sum_t p(e, t | m) \\ &= \sum_t p(e, t | m) \mathbf{p(t | m)} \end{aligned}$$

Entity Type Probability



FIGER

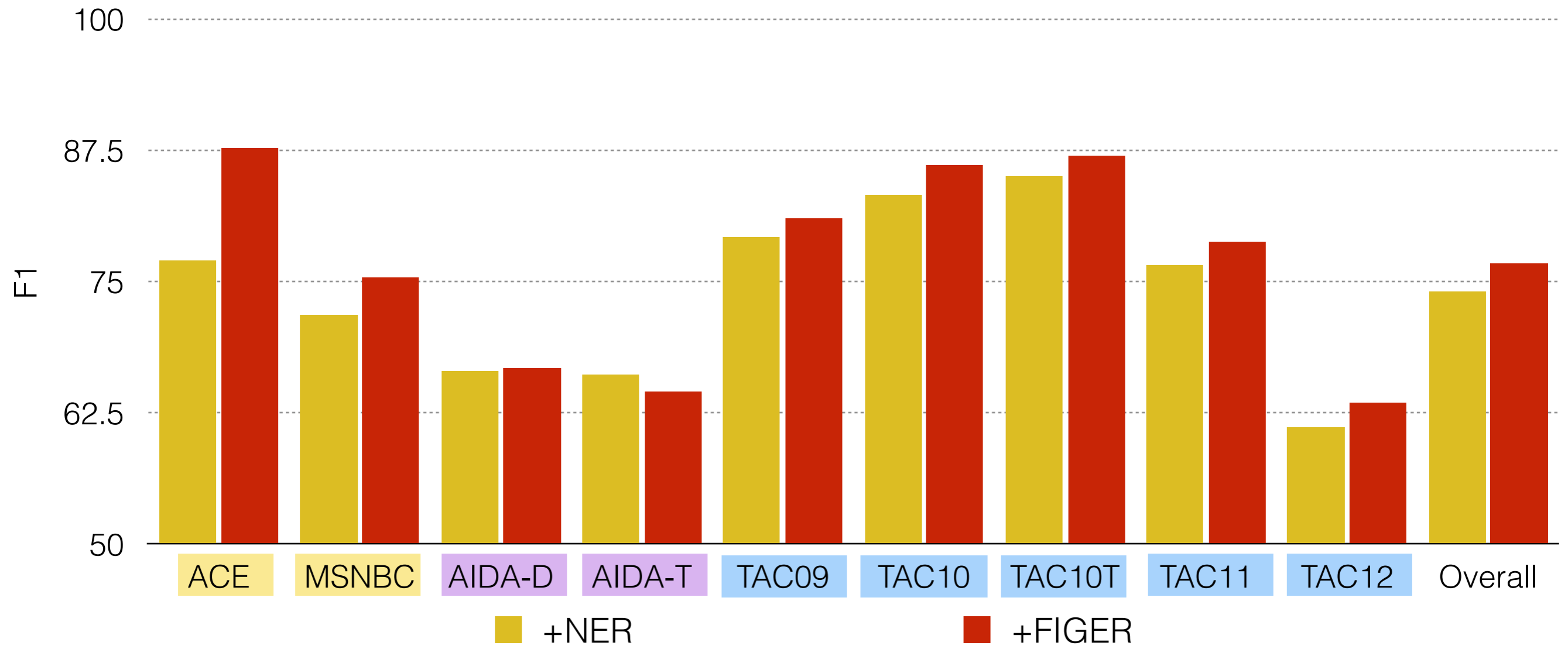
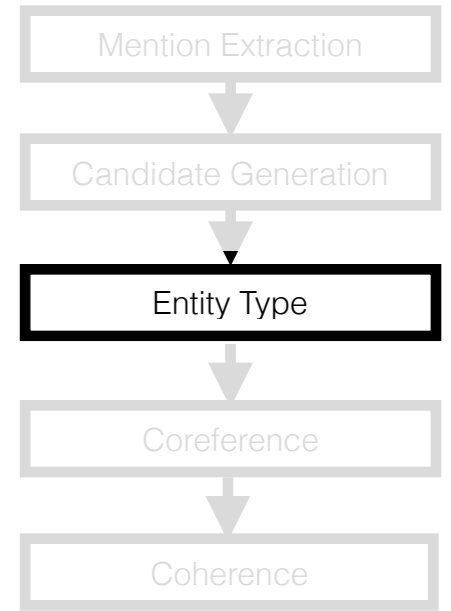
(Ling & Weld, 2012)



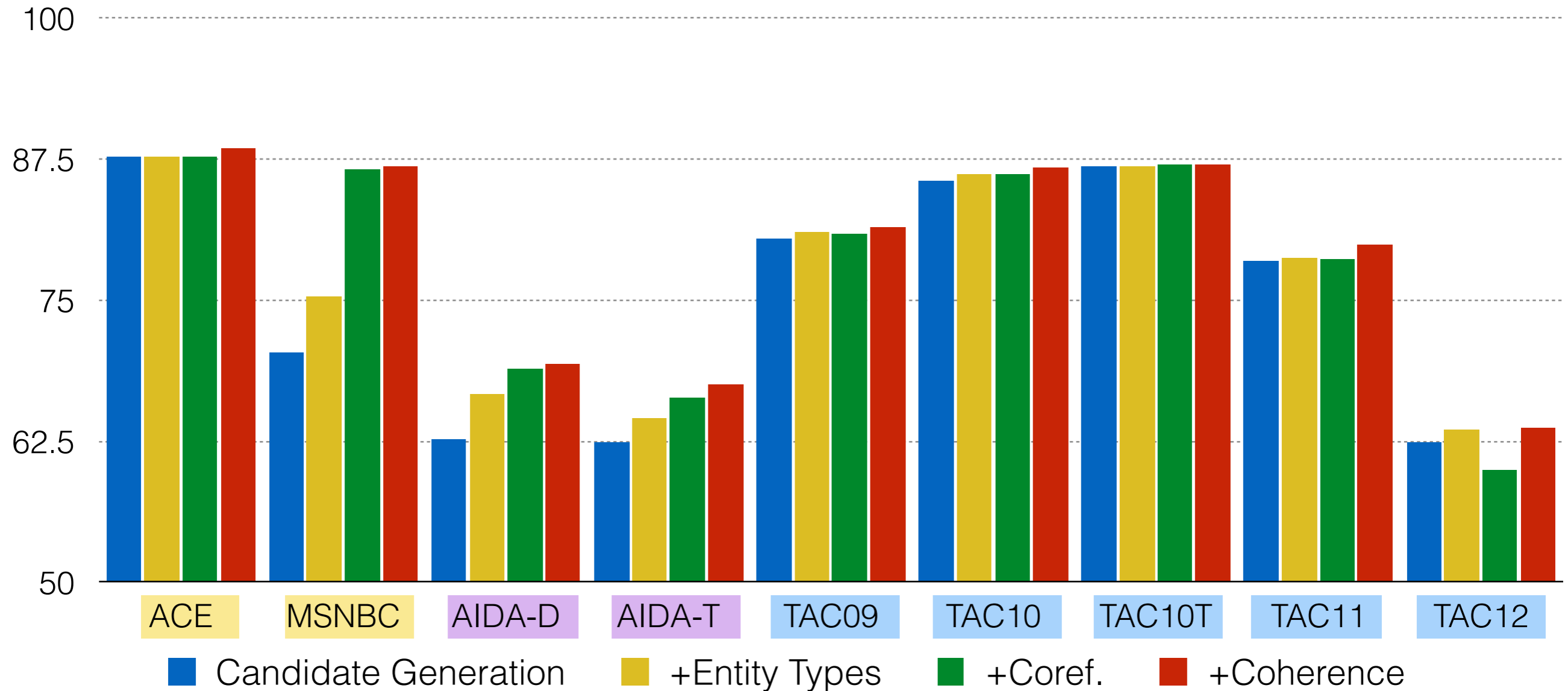
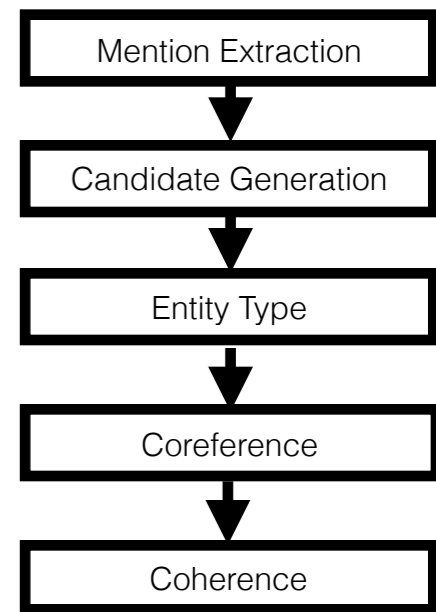
- 112 entity types
- multi-label multi-class

person actor architect artist athlete author coach director	doctor engineer monarch musician politician religious_leader soldier terrorist	organization airline company educational_institution fraternity_sorority sports_league sports_team	terrorist_organization government_agency government political_party educational_department military news_agency
location city country county province railway road bridge	body_of_water island mountain glacier astral_body cemetery park	product engine airplane car ship spacecraft train	camera mobile_phone computer software game instrument weapon
building airport dam hospital hotel library power_station restaurant sports_facility theater	time color award educational_degree title law ethnicity language religion god	chemical_thing biological_thing medical_treatment disease symptom drug body_part living_thing animal food	art film play written_work newspaper music event attack election protest military_conflict natural_disaster sports_event terrorist_attack
			website broadcast_network broadcast_program tv_channel currency stock_exchange algorithm programming_language transit_system transit_line

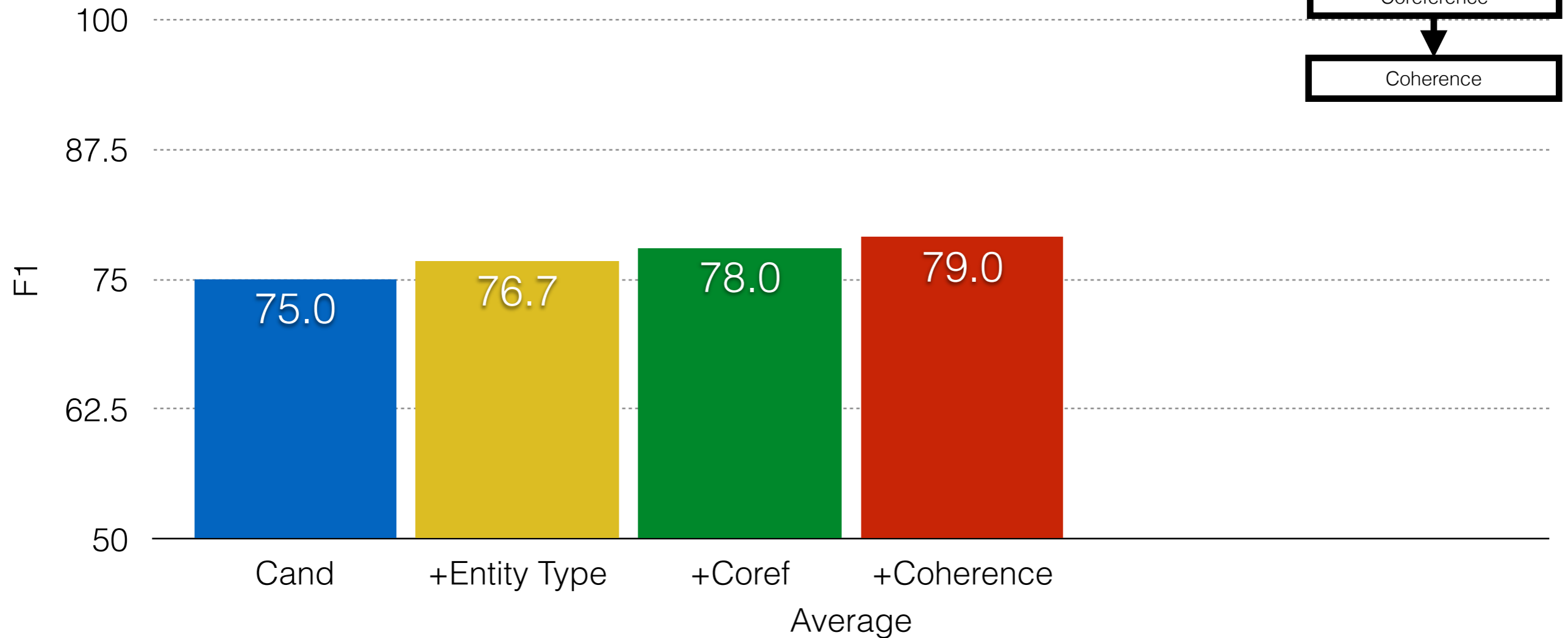
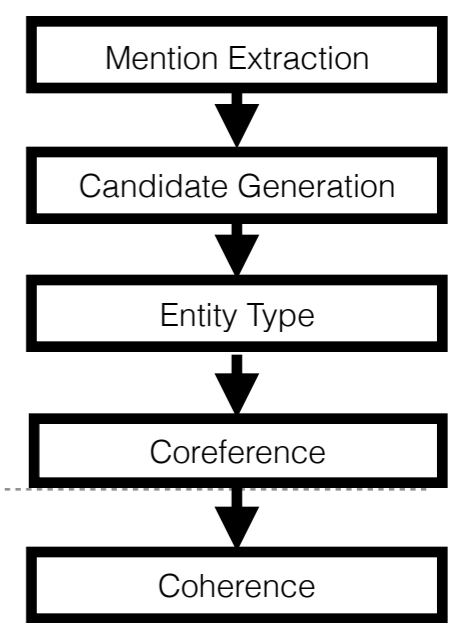
Predicted Entity Types



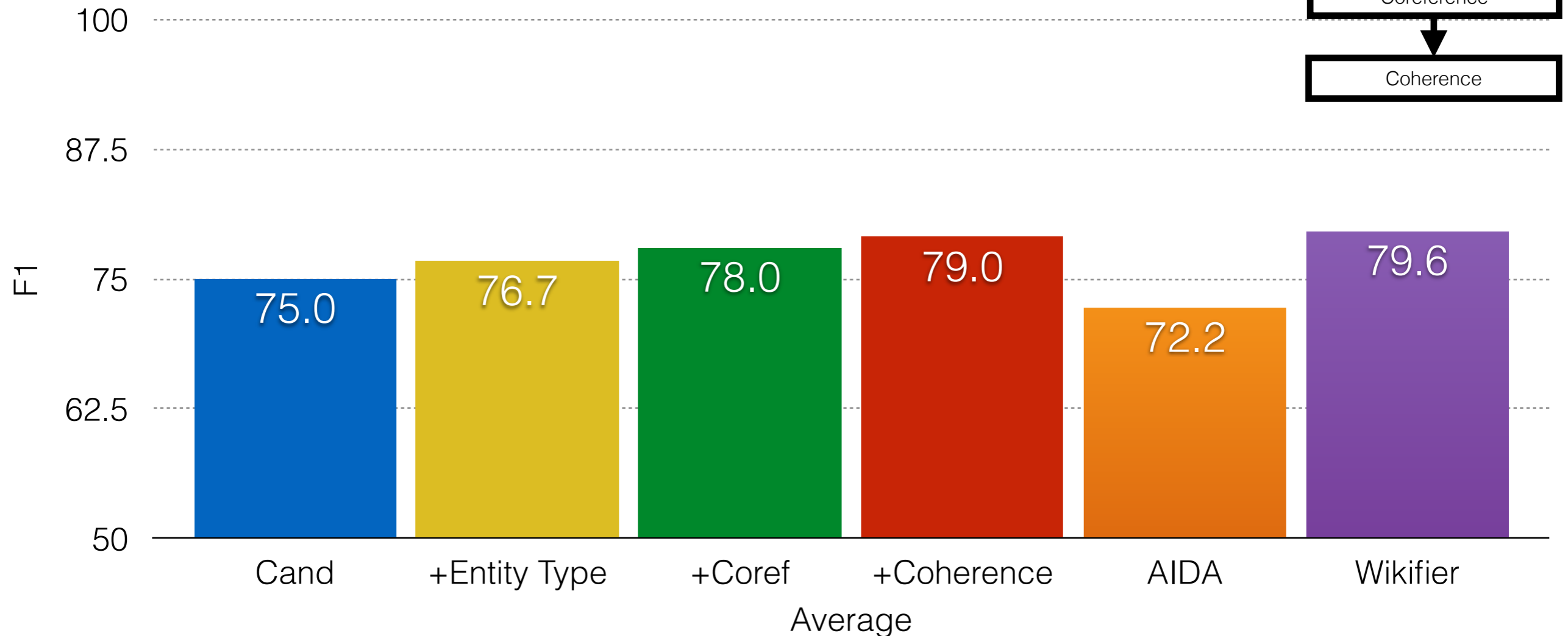
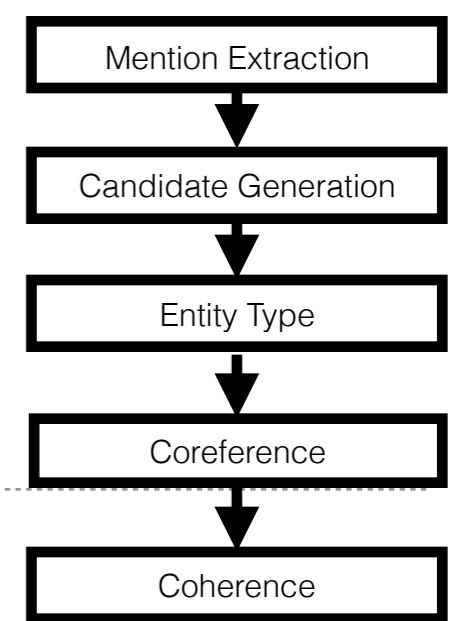
Overall Performance




Overall Performance



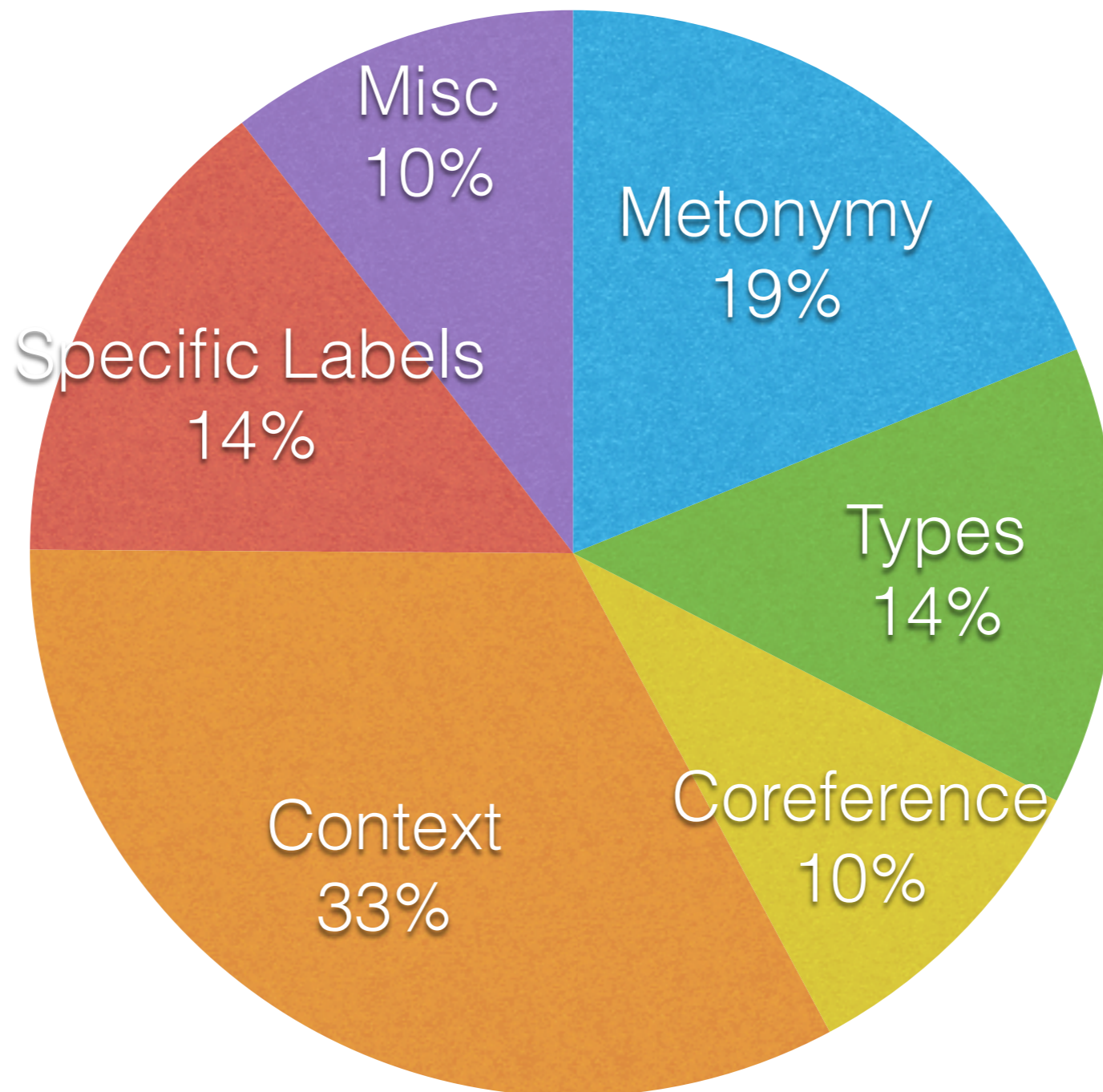
Overall Performance



 AIDA (Hoffart et al. 2011)

 Illinois Wikifier 2.0 (Cheng & Roth, 2013)

Error Analysis



Conclusion

Vinculum

- a modular deterministic system achieves good performance

Conclusion

Vinculum

- a modular deterministic system achieves good performance
- a comprehensive evaluation over nine data sets

Conclusion

Vinculum

- a modular deterministic system achieves good performance
- a comprehensive evaluation over nine data sets
 - CrossWikis provides better cond. prob.

Conclusion

Vinculum

- a modular deterministic system achieves good performance
- a comprehensive evaluation over nine data sets
 - CrossWikis provides better cond. prob.
 - Fine-grained entity types are very useful

Conclusion

Vinculum

- a modular deterministic system achieves good performance
- a comprehensive evaluation over nine data sets
 - CrossWikis provides better cond. prob.
 - Fine-grained entity types are very useful
 - Coreference and Coherence also improve the performance

Conclusion

Vinculum

- a modular deterministic system achieves good performance
- a comprehensive evaluation over nine data sets
 - CrossWikis provides better cond. prob.
 - Fine-grained entity types are very useful
 - Coreference and Coherence also improve the performance
- <http://github.com/xiaoling/vinculum>

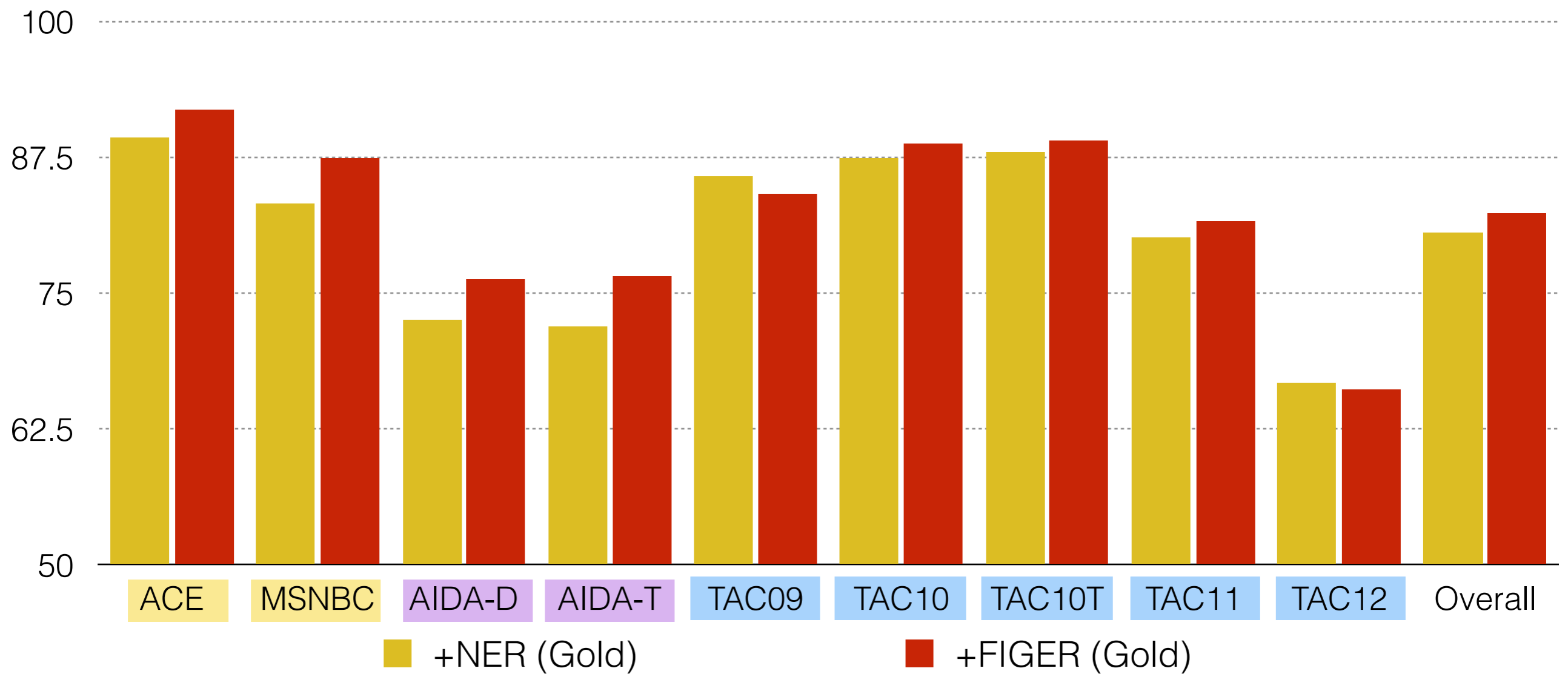
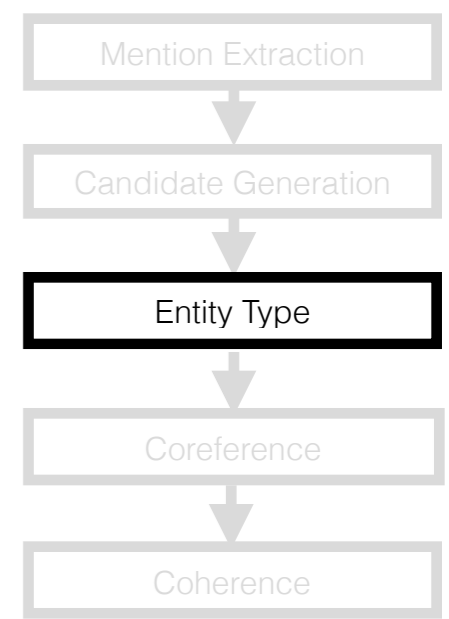
Thanks!
Questions?

Conclusion

Vinculum

- a modular deterministic system achieves good performance
- a comprehensive evaluation over nine data sets
 - CrossWikis provides better cond. prob.
 - Fine-grained entity types are very useful
 - Coreference and Coherence also improve the performance
- <http://github.com/xiaoling/vinculum>

Oracle Entity Types



Implementation Details

Component	Implementation
Mention Extraction	Stanford NER
Candidate Generation	CrossWikis
Entity Type Prediction	Fine-grained Entity Types
Coreference	Stanford Coreference
Coherence	NGD + relational triples

System Comparison

	VINCULUM	AIDA	WIKIFIER
Mention Extraction	NER	NER	NER, noun phrases
Candidate Generation	CrossWikis	intra-Wikipedia	intra-Wikipedia
Entity Types	FIGER	NER	NER
Coreference	representative mention	-	re-rank the candidates
Coherence	NGD, relational	NGD	NGD, relational
Learning	deterministic	trained on AIDA	trained on Wiki

Error Analysis: Metonymy

- South Africa managed to avoid a fifth successive defeat in 1996 at the hands of the All Blacks ...
- Prediction : **South Africa**
- Label : **South Africa national rugby union team**

Error Analysis: Entity Types

- Instead of Los Angeles International, for example, consider flying into Burbank or John Wayne Airport ...
- Prediction : **Burbank, California**
- Label : **Bob Hope Airport**

Error Analysis: Coreference

- It is about his mysterious father, Barack Hussein Obama, an imperious if alluring voice gone distant and then missing.
- Prediction : **Barack Obama**
- Label : **Barack Obama Sr.**

Error Analysis: Context

- Scott Walker removed himself from the race, but Green never really stirred the passions of former Walker supporters, nor did he garner outsized support “outstate”.
- Prediction : **Scott Walker (singer)**
- Label : **Scott Walker (politician)**