Wikification and Beyond:
The Challenges of Entity and Concept Grounding

Heng Ji (RPI), Dan Roth (UIUC),

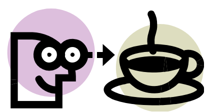
Ming-Wei Chang (MSR), Taylor Cassidy (ARL&IBM),

Xiaoman Pan (RPI), Hongzhao Huang (RPI)

http://nlp.cs.rpi.edu/paper/wikificationtutorial2.pdf **[pptx]**

Outline

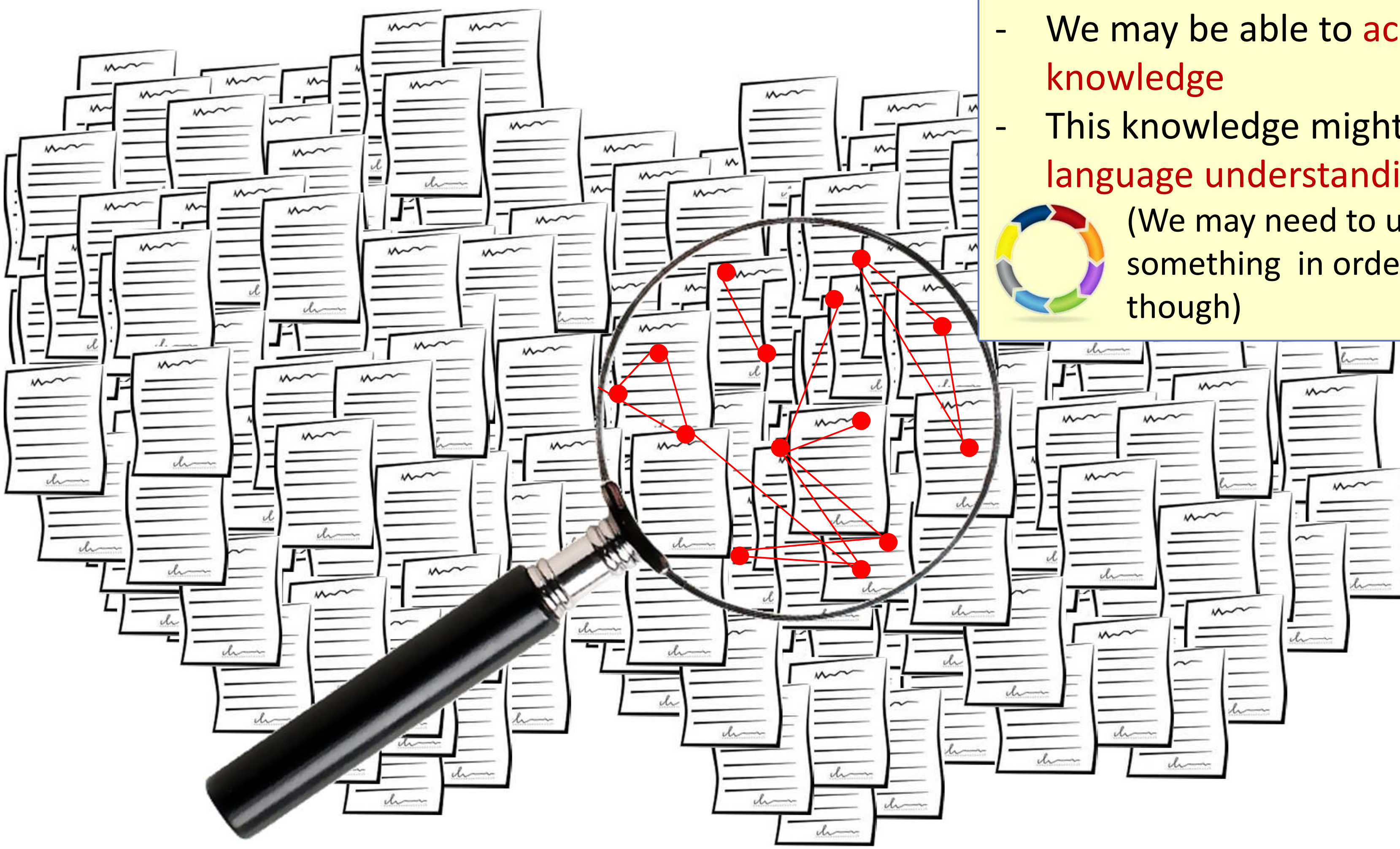
- Motivation and Definition)
- A Skeletal View of a Wikification System
  - High Level Algorithmic Approach
- Key Challenges



Coffee Break

- Recent Advances
- New Tasks, Trends and Applications
- What’s Next?
- Resources, Shared Tasks and Demos

Information overload



Downside:

- We need to deal with a lot of information

Upside:

- We may be able to acquire knowledge
- This knowledge might support language understanding

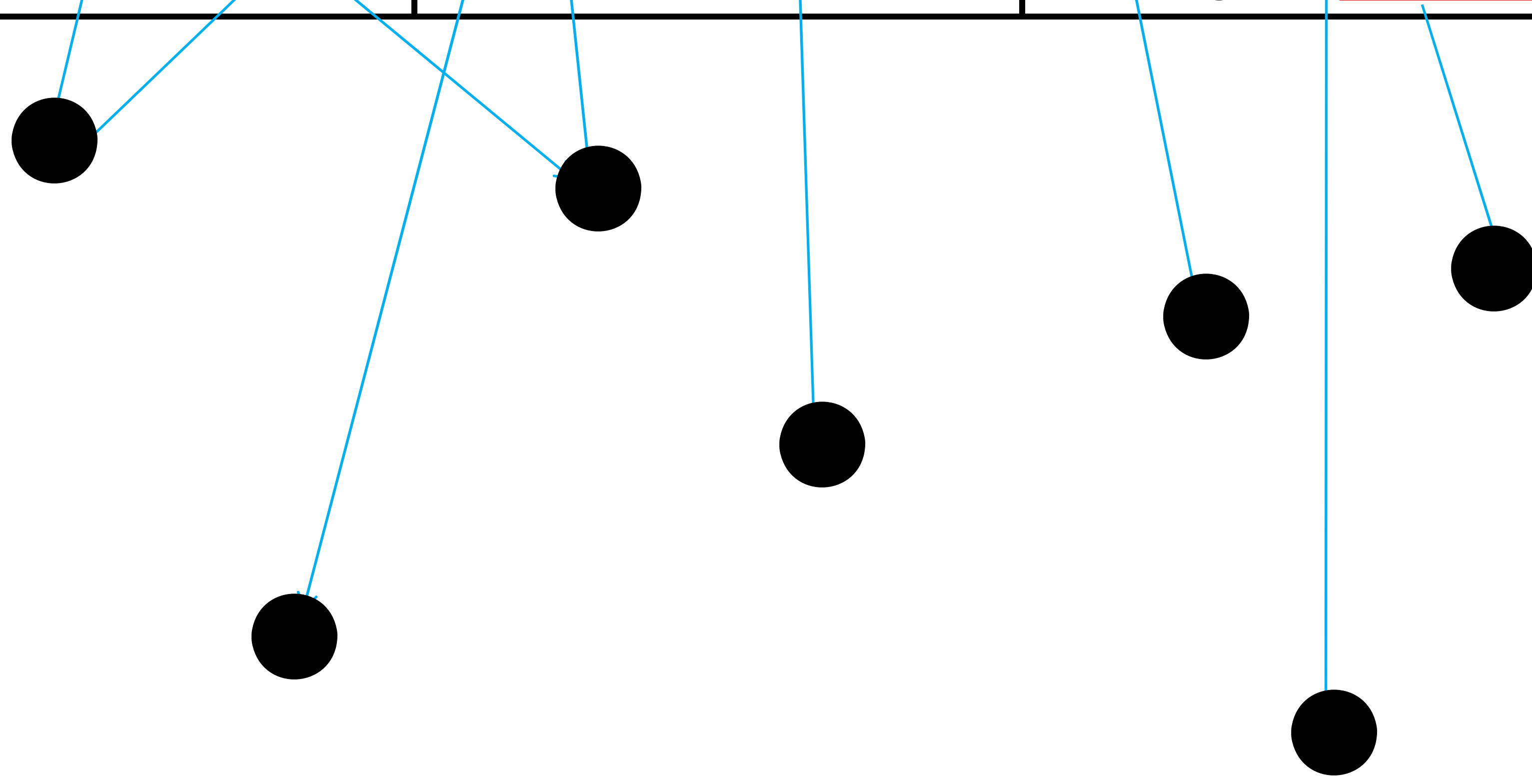
(We may need to understand something in order to do it, though)

Organizing knowledge

It's a version of ***Chicago*** – the standard classic Macintosh menu font, with that distinctive thick diagonal in the "N".	***Chicago*** was used by default for Mac menus through MacOS 7.6, and OS 8 was released mid-1997..	***Chicago VIII*** was one of the early 70s-era ***Chicago*** albums to catch my ear, along with ***Chicago II***.

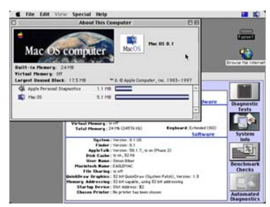
Cross-document co-reference resolution

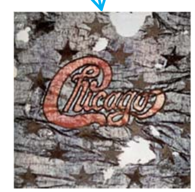
It’s a version of ***Chicago*** – the standard classic ***Macintosh*** menu font, with that distinctive thick diagonal in the ”N”.	***Chicago*** was used by default for ***Mac*** menus through ***MacOS 7.6***, and ***OS 8*** was released mid-1997..	***Chicago VIII*** was one of the early 70s-era ***Chicago*** albums to catch my ear, along with ***Chicago II***.

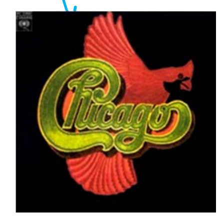


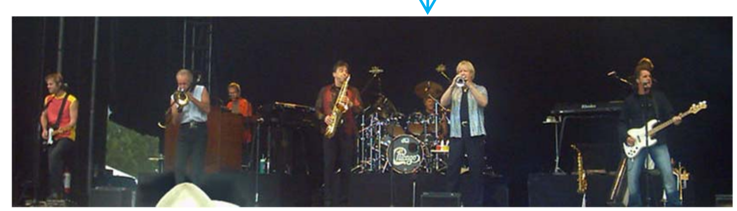
Reference resolution: (disambiguation to Wikipedia)

It's a version of ***Chicago*** – the standard classic ***Macintosh*** menu font, with that distinctive thick diagonal in the "N".	***Chicago*** was used by default for ***Mac*** menus through ***MacOS 7.6***, and ***OS 8*** was released mid-1997..	***Chicago VIII*** was one of the early 70s-era ***Chicago*** albums to catch my ear, along with ***Chicago II***.



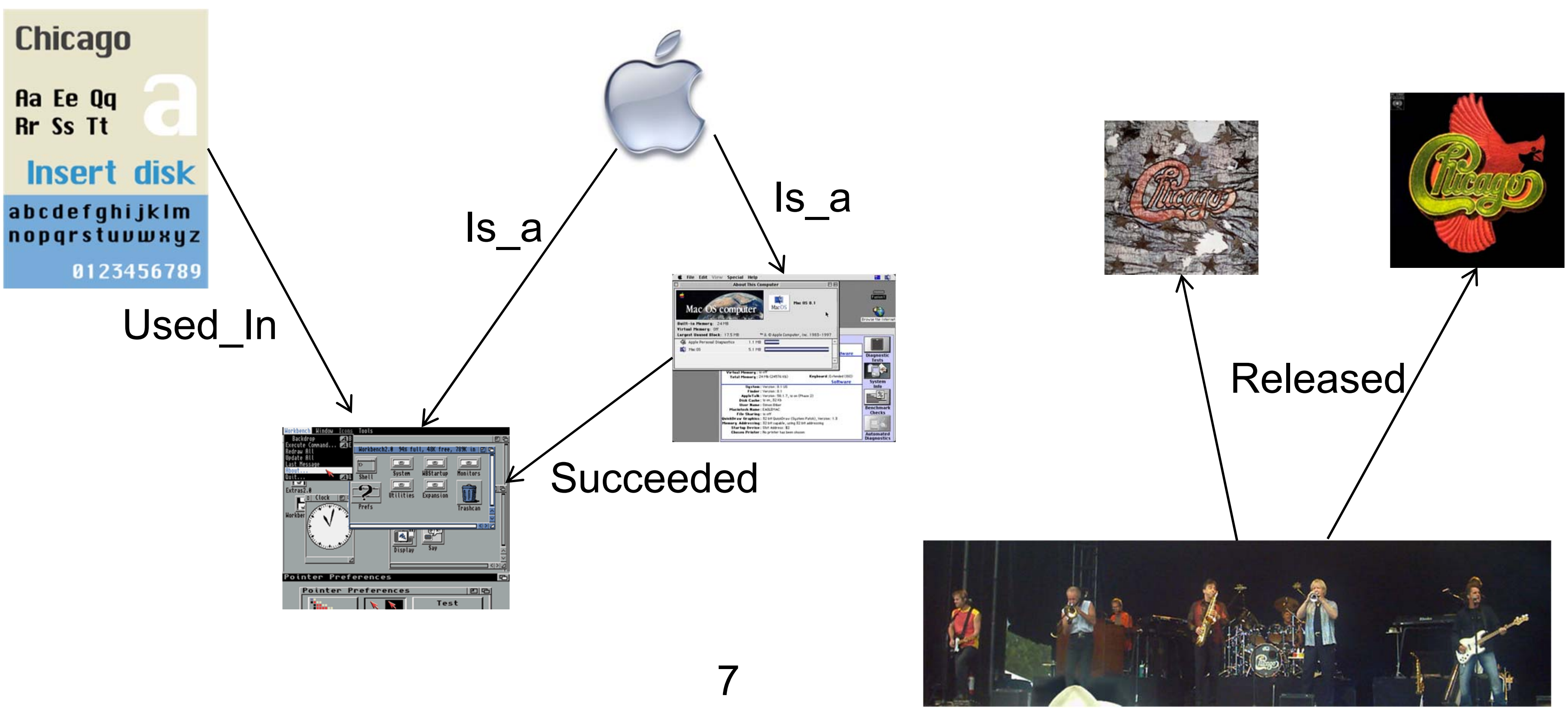






The “Reference” Collection has Structure

It’s a version of ***Chicago*** – the standard classic ***Macintosh*** menu font, with that distinctive thick diagonal in the ”N”.	***Chicago*** was used by default for ***Mac*** menus through ***MacOS 7.6***, and ***OS 8*** was released mid-1997..	***Chicago VIII*** was one of the early 70s-era ***Chicago*** albums to catch my ear, along with ***Chicago II***.

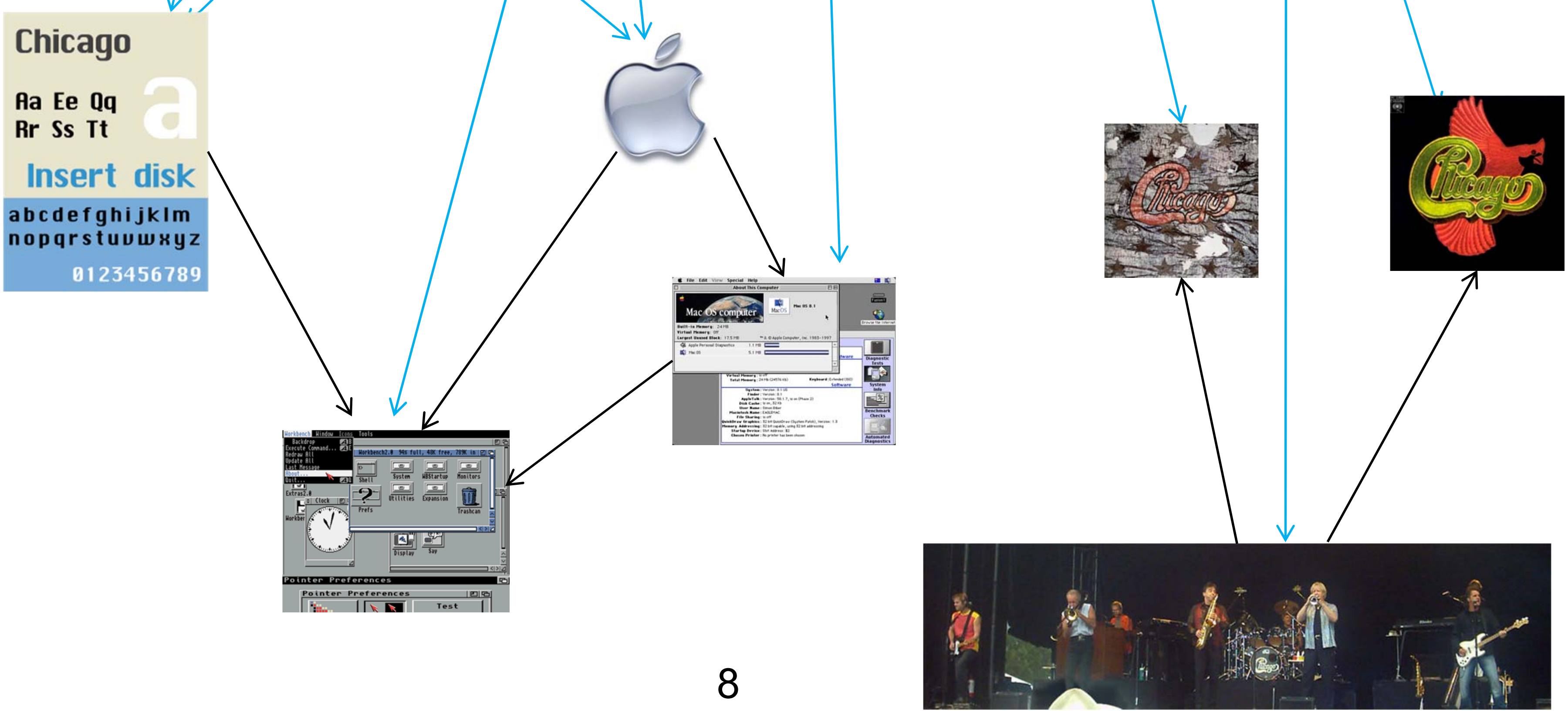


Analysis of Information Networks

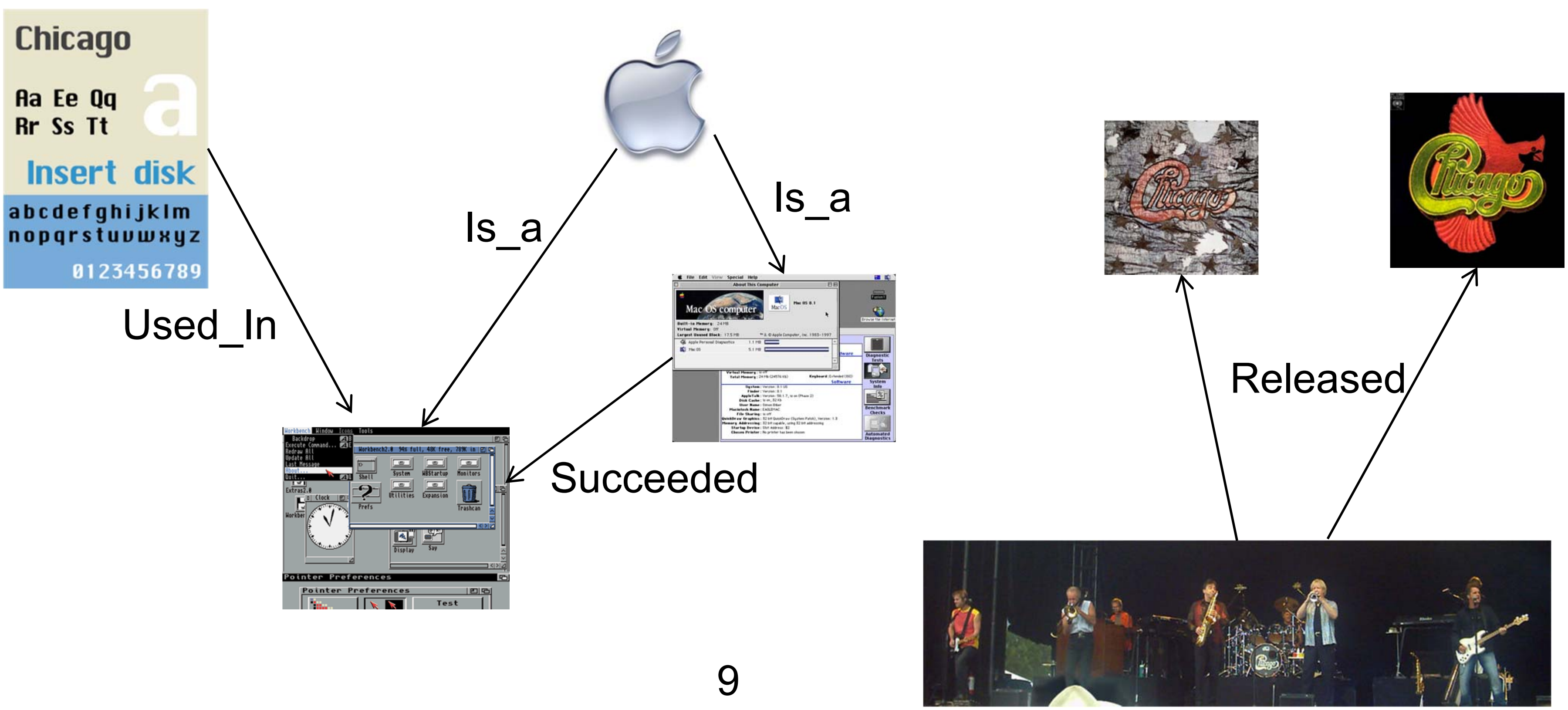
It's a version of *Chicago* – the standard classic *Macintosh* menu font, with that distinctive thick diagonal in the "N".

Chicago was used by default for *Mac* menus through *MacOS 7.6*, and *OS 8* was released mid-1997..

Chicago VIII was one of the early 70s-era *Chicago* albums to catch my ear, along with *Chicago II*.



Here – Wikipedia as a knowledge resource
…. but we can use other resources



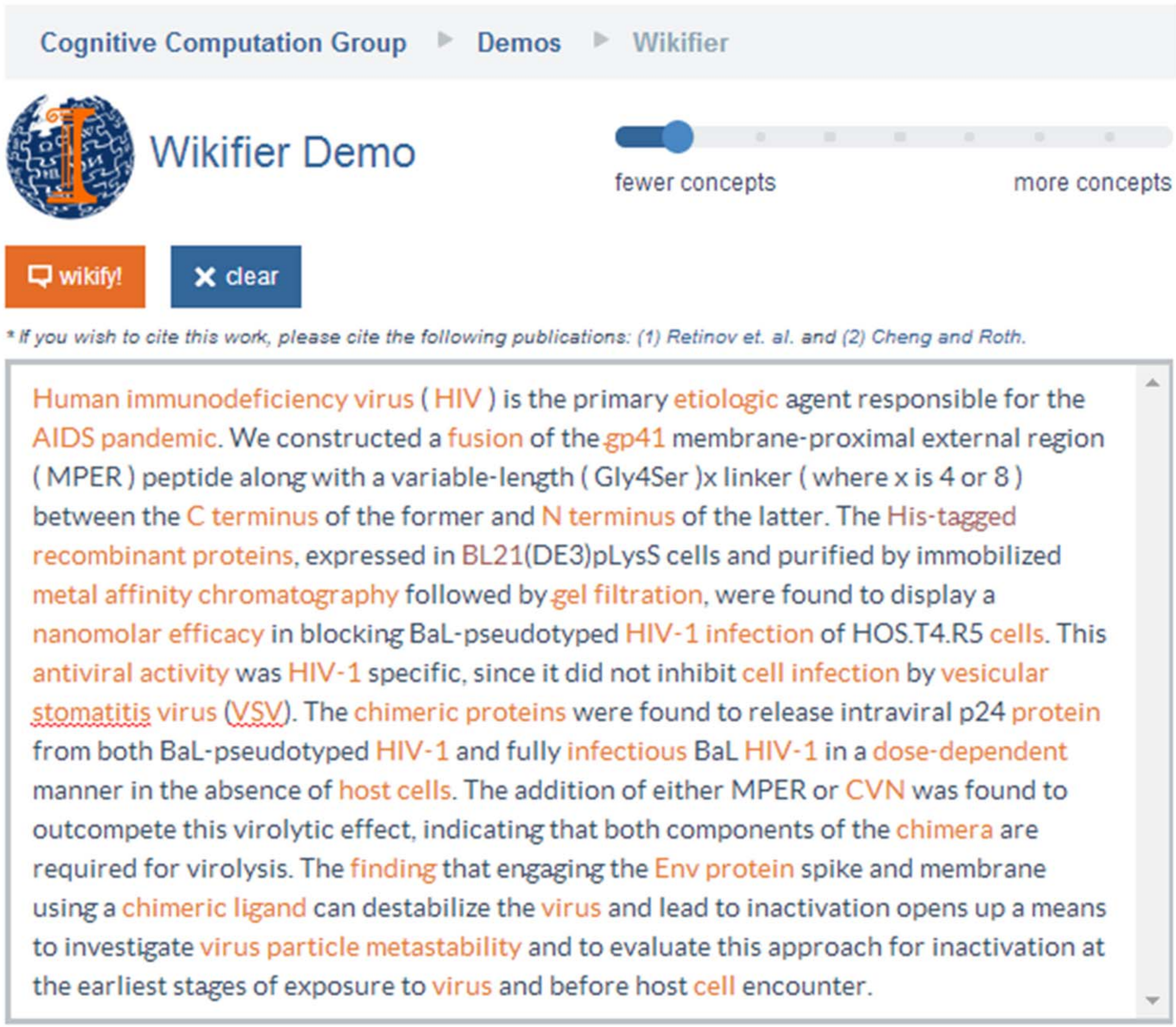
Motivation

- Dealing with Ambiguity of Natural Language
  - Mentions of entities and concepts could have multiple meanings
- Dealing with Variability of Natural Language
  - A given concept could be expressed in many ways

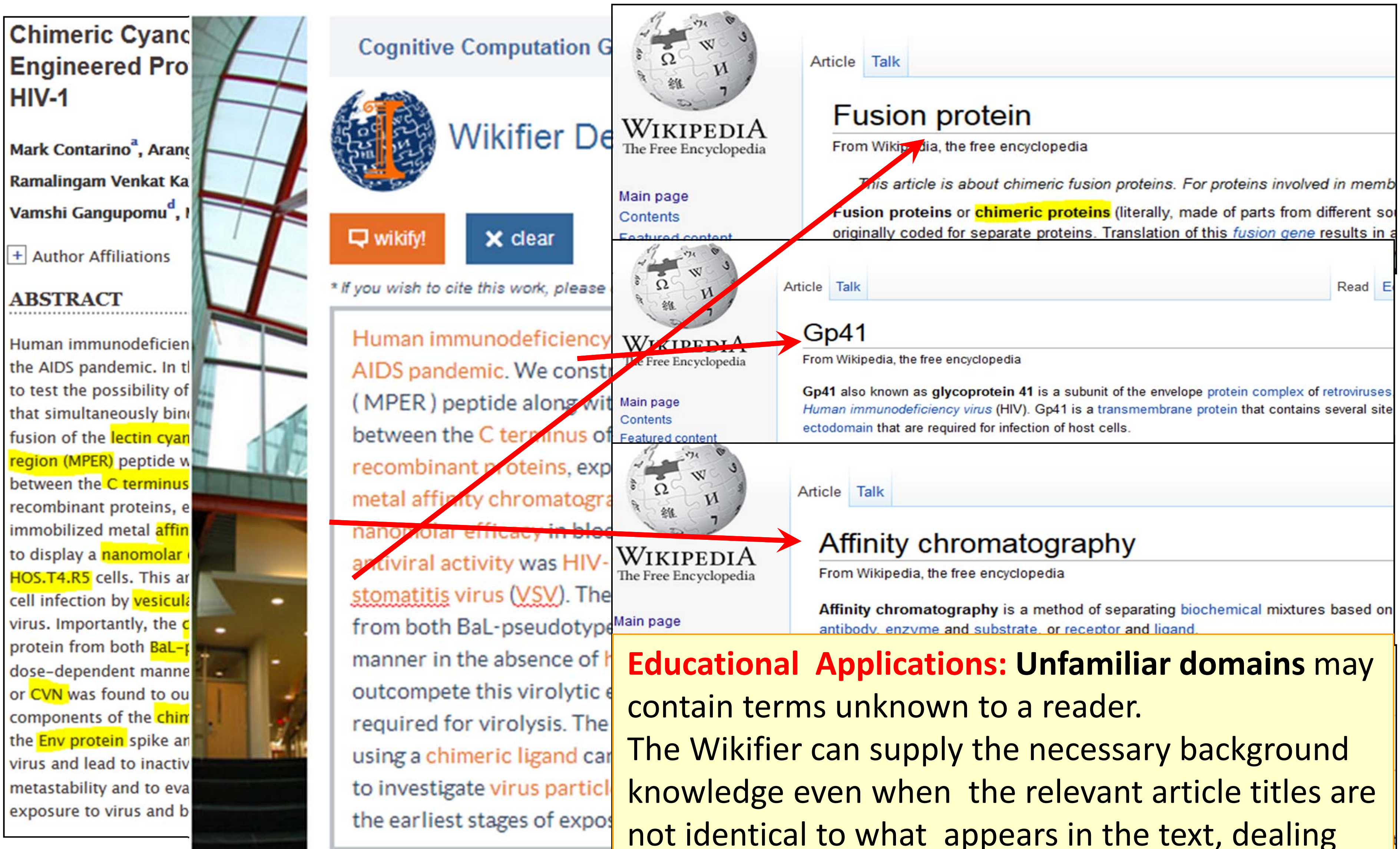
- Wikification addresses these two issues in a specific way:

- The Reference Problem
  - What is meant by this concept? (WSD + Grounding)
  - More than just co-reference (within and across documents)

Navigating Unfamiliar Domains



Navigating Unfamiliar Domains



Educational Applications: Unfamiliar domains may contain terms unknown to a reader.
The Wikifier can supply the necessary background knowledge even when the relevant article titles are not identical to what appears in the text, dealing with both **ambiguity and variability.**

Applications of Wikification

- Knowledge Acquisition (via grounding)
  - Still remains open: how to organize the knowledge in a useful way?
- Co-reference resolution (Ratinov & Roth, 2012)
  - "After the vessel suffered a catastrophic torpedo detonation, Kursk sank in the waters of Barents Sea…"
  - Knowing Kursk → Russian submarine K-141 Kursk helps system to co-ref "Kursk" and "vessel"
- Document classification
  - Tweets labeled World, US, Science & Technology, Sports, Business, Health, Entertainment (Vitale et. al., 2012)
  - Datalesss classification (ESA-based representations; Song & Roth' 14)
  - Document and concepts are represented via Wikipedia titles
- Visualization: Geo- visualization of News (Gao et. al. CHI'14)

Task Definition

- A formal definition of the task consists of:

1. A definition of the **mentions** (concepts, entities) to highlight

2. Determining the target encyclopedic resource (**KB**)

3. Defining what to point to in the KB (**title**)

1. Mentions

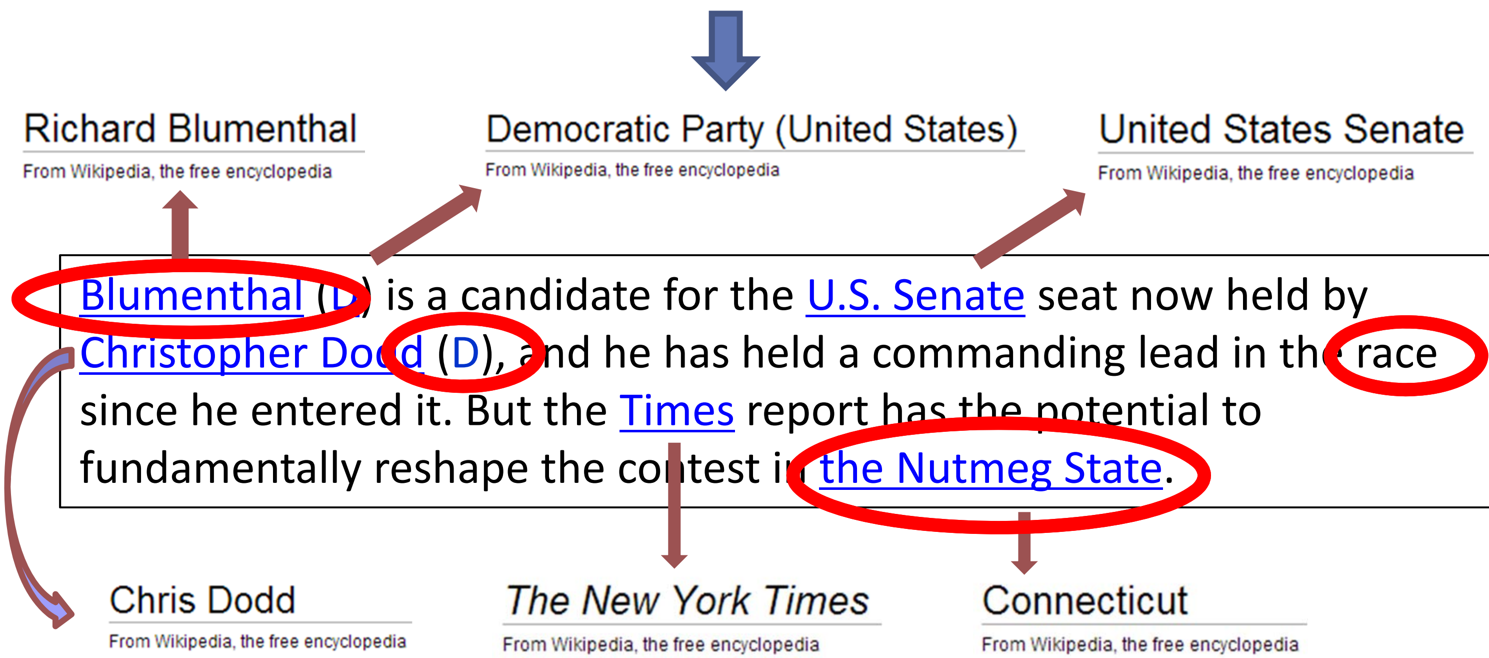
- A mention: a phrase used to refer to something in the world
  - Named entity (person, organization), object, substance, event, philosophy, mental state, rule …

- Task definitions vary across the definition of mentions
  - All N-grams (up to a certain size); Dictionary-based selection; Data-driven controlled vocabulary (e.g., all Wikipedia titles); **only named entities.**

- Ideally, one would like to have a mention definition that adapts to the application/user

Examples of Mentions (1)

Blumenthal (D) is a candidate for the U.S. Senate seat now held by Christopher Dodd (D), and he has held a commanding lead in the race since he entered it. But the Times report has the potential to fundamentally reshape the contest in the Nutmeg State.

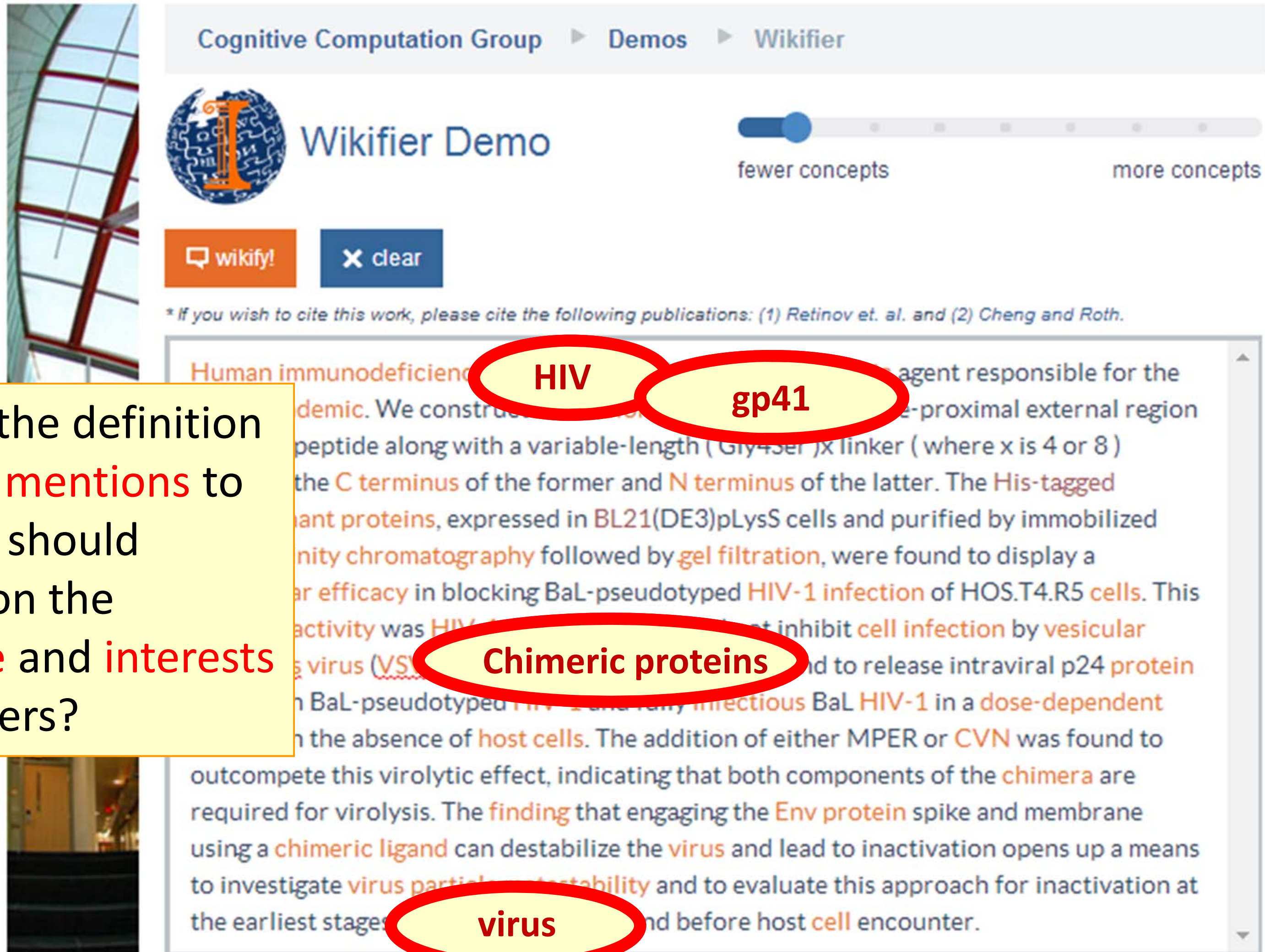


Examples of Mentions (2)

Some task definitions insist on dealing only with mentions that are named entities

How about: *Hosni Mubarak's wife?*
Both entities have a Wikipedia page

Examples of Mentions (3)



HIV

gp41

Chimeric proteins

virus

Perhaps the definition of which mentions to highlight should depend on the expertise and interests of the users?

2. Concept Inventory (KB)

- Multiple KBs can be used, in principle, as the target KB.
- Wikipedia has the advantage of a broad coverage, regularly maintair KB, with significant amount of tex associated with each title.
  - o All type of pages?
    - Content pages
    - Disambiguation pages
    - List pages
- What should happened to mentions that do not have entries in the target KB?

3. Null Links

- Often, there are multiple sensible links.

Dorothy Byrne, a state coordinator for the Florida Green Party,…

- How to capture the fact that Dorothy Byrne does not refer to any concept in Wikipedia?

- Wikification: Simply map Dorothy Byrne → Null
- Entity Linking: If multiple mentions in the given document(s) correspond to the same concept, which is outside KB
  - First cluster relevant mentions as representing a single concept
  - Map the cluster to Null

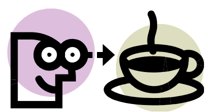
Naming Convention

- Wikification:
  - Map Mentions to KB Titles
  - Map Mentions that are not in the KB to NIL

- Entity Linking:
  - Map Mentions to KB Titles
  - If multiple mentions in correspond to the same Title, which is outside KB:
    - First cluster relevant mentions as representing a single Title
    - Map the cluster to Null
- If the set of target mentions only consists of named entities we call the task: Named Entity [Wikification, Linking]

Outline

- Motivation and Definition
- A Skeletal View of a Wikification System
    - o High Level Algorithmic Approach
- Key Challenges



Coffee Break

- Recent Advances
- New Tasks, Trends and Applications
- What's Next?
- Resources, Shared Tasks and Demos

Wikification: Subtasks

- Wikification and Entity Linking requires addressing several sub-tasks:
  - Identifying Target Mentions
    - Mentions in the input text that should be Wikified
  - Identifying Candidate Titles
    - Candidate Wikipedia titles that could correspond to each mention
  - Candidate Title Ranking
    - Rank the candidate titles for a given mention
  - NIL Detection and Clustering
    - Identify mentions that do not correspond to a Wikipedia title
    - Entity Linking: cluster NIL mentions that represent the same entity.

High-level Algorithmic Approach.

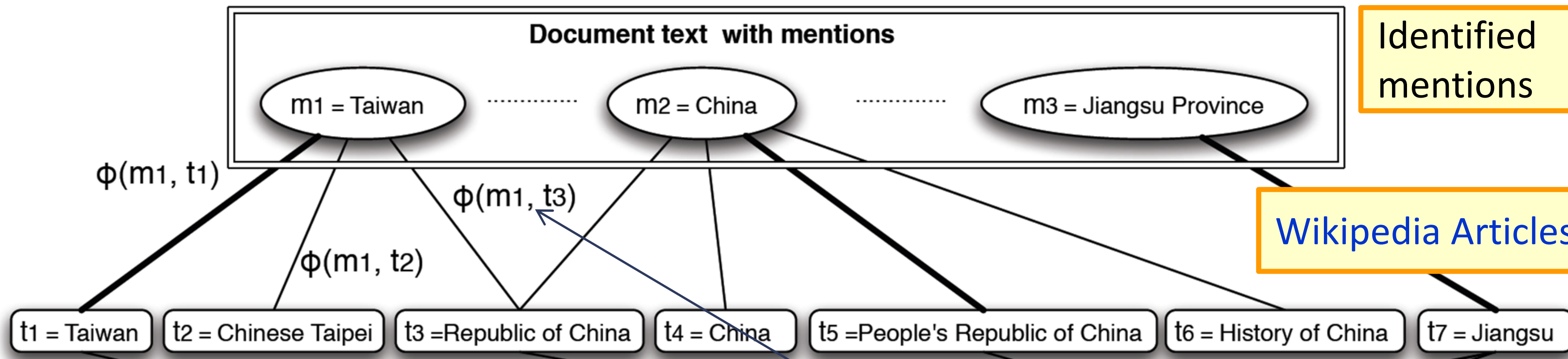
- Input: A text document d; Output: a set of pairs (m_i, t_i)
  - m_i are mentions in d; $t_j(m_i)$ are corresponding Wikipedia titles, or NIL.
- (1) Identify mentions m_i in d
- (2) Local Inference
  - For each m_i in d:
    - Identify a set of relevant titles $T(m_i)$
    - Rank titles $t_i \in T(m_i)$

    [E.g., consider local statistics of edges [(m_i, t_i), $(m_i, *)$, and $(*, t_i)$] occurrences in the Wikipedia graph]
- (3) Global Inference
  - For each document d:
    - Consider all $m_i \in d$; and all $t_i \in T(m_i)$
    - Re-rank titles $t_i \in T(m_i)$

    [E.g., if m, m' are related by virtue of being in d, their corresponding titles t, t' may also be related]

Local approach

A text Document



Identified mentions

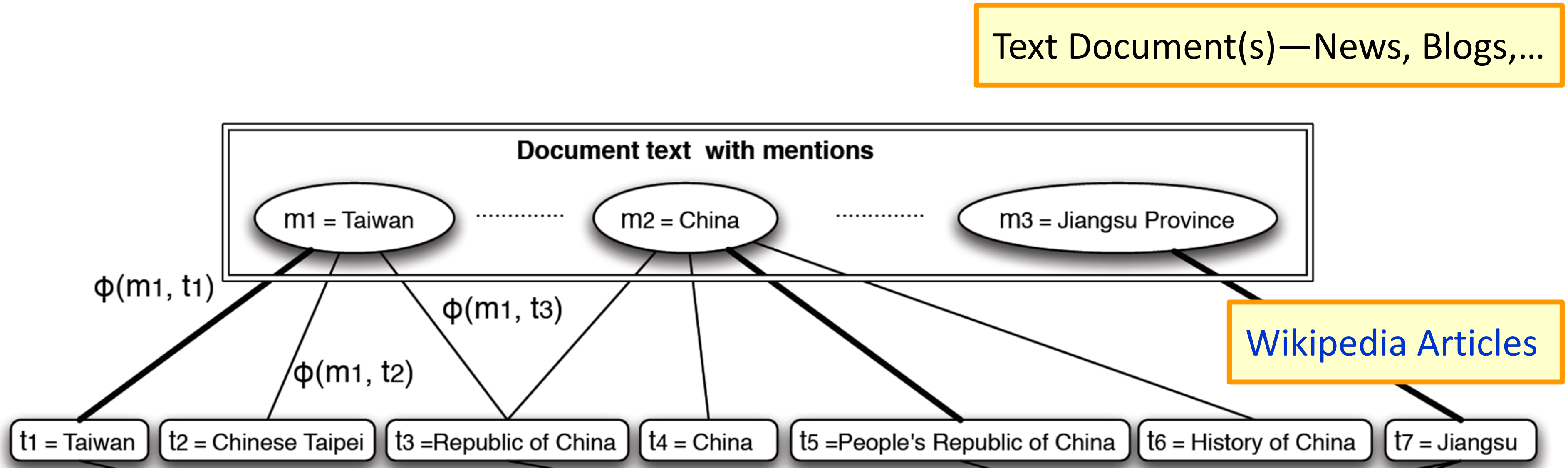
Wikipedia Articles

Local score of matching the mention to the title (decomposed by m_i)

- Γ is a solution to the problem
  - A set of pairs (m,t)
- m: a mention in the document
- t: the matched Wikipedia Title

$$\Gamma^*_{\text{local}} = \arg\max_{\Gamma} \sum_{i=1}^{N} \phi(m_i, t_i) \qquad (1)$$

Global Approach: Using Additional Structure



$$\Gamma^* \approx \arg\max_{\Gamma} \sum_{i=1}^{N} [\phi(m_i, t_i) + \sum_{t_i \in \Gamma, t_j \in \Gamma'} \psi(t_i, t_j)]$$

Adding a "global" term to evaluate how good the structure of the solution is.

- Use the local solutions Γ' (each mention considered independently.
- Evaluate the structure based on pair-wise coherence scores $\Psi(t_i,t_j)$
- Choose those that satisfy document coherence conditions.

High-level Algorithmic Approach

- Input: A text document d; Output: a set of pairs (m_i ,t_i)
  - o m_i are mentions in d; t_i are corresponding Wikipedia titles, or NIL.
- (1) Identify mentions m_i in d
- (2) Local Inference
  - o For each m_i in d:
    - Identify a set of relevant titles T(m_i)
    - Rank titles $t_i \in$ T(m_i)

    [E.g., consider local statistics of edges (m_i ,t_i) , (m_i *), and (*, t_i) occurrences of in the Wikipedia graph]
- (3) Global Inference
  - o For each document d:
    - Consider all $m_i \in$ d; and all t_i 2 T(m_i)
    - Re-rank titles $t_i \in$ T(m_i)

    [E.g., if m, m' are related by virtue of being in d, their corresponding titles t, t' should also be related]

Mention Identification

- Highest recall: Each n-gram is a potential concept mention
  - Intractable for larger documents
- Surface form based filtering
  - Shallow parsing (especially NP chunks), NP's augmented with surrounding tokens, capitalized words
  - Remove: single characters, "stop words", punctuation, etc.
- Classification and statistics based filtering
  - Name tagging (Finkel et al., 2005; Ratinov and Roth, 2009; Li et al., 2012)
  - Mention extraction (Florian et al., 2006, Li and Ji, 2014)
  - Key phrase extraction, independence tests (Mihalcea and Csomai, 2007), common word removal (Mendes et al., 2012;)

Mention Identification (Cont')

- Wikipedia Lexicon Construction based on prior link knowledge
  - Only n-grams linked in training data (prior anchor text) (Ratinov et al., 2011; Davis et al., 2012; Sil et al., 2012; Demartini et al., 2012; Wang et al., 2012; Han and Sun, 2011; Han et al., 2011; Mihalcea and Csomai, 2007; Cucerzan, 2007; Milne and Witten, 2008; Ferragina and Scaiella, 2010)
    - E.g. all n-grams used as anchor text within Wikipedia
  - Only terms that exceed link probability threshold (Bunescu, 2006; Cucerzan, 2007; Fernandez et al., 2010; Chang et al., 2010; Chen et al., 2010; Meij et al., 2012; Bysani et al., 2010; Hachey et al., 2013; Huang et al., 2014)
  - Dictionary-based chunking
  - String matching (n-gram with canonical concept name list)
- Mis-spelling correction and normalization (Yu et al., 2013; Charton et al., 2013)

Mention Identification (Cont')

- Multiple input sources are being used
  - Some build on the given text only, some use external resources.
- Methods used by some popular systems
  - Illinois Wikifier (Ratinov et al., 2011; Cheng and Roth, 2013)
    - NP chunks and substrings, NER (+nesting), prior anchor text
  - TAGME (Ferragina and Scaiella, 2010)
    - Prior anchor text
  - DBPedia Spotlight (Mendes et al., 2011)
    - Dictionary-based chunking with string matching (via DBpedia lexicalization dataset)
  - AIDA (Finkel et al., 2005; Hoffart et al., 2011)
    - Name Tagging
  - RPI Wikifier (Chen and Ji, 2011; Cassidy et al., 2012; Huang et al., 2014)
    - Mention Extraction (Li and Ji, 2014)

Need Mention Expansion

- Medical Domain: 33% of abbreviations are ambiguous (Liu et al., 2001), major source of errors in medical NLP (Friedman et al., 2001)

RA	“rheumatoid arthritis”, “tenal artery”, “right atrium”, “right atrial”, “refractory anemia”, “radioactive”, “right arm”, “rheumatic arthritis”, …
PN	“Penicillin”; “Pneumonia”; “Polyarteritis”; “Nodosa”; “Peripheral neuropathy”; “Peripheral nerve”; “Polyneuropathy”; “Pyelonephritis”; “Polyneuritis”; “Parenteral nutrition”; “Positional Nystagmus”; “Periarteritis nodosa”, …

- Military Domain
  - *“GA ADT 1, USDA, USAID, ADP, Turkish PRT, and the DAIL staff met to create the Wardak Agricultural Steering Committee. “*
  - *“DST”* = *“District Stability Team”* or *“District Sanitation Technician”*?
  - *“ADP” = “Adrian Peterson” (Person) or “Arab Democratic Party” (Organization) or “American Democracy Project” (Initiative)?*

Mention Expansion

- Co-reference resolution
  - Each mention in a co-referential cluster should link to the same concept
  - Canonical names are often less ambiguous
  - Correct types: “*Detroit*” = “*Red Wings*”; “*Newport*” = “*Newport-Gwent Dragons*”
- Known Aliases
  - KB link mining (e.g., Wikipedia “re-direct”) (Nemeskey et al., 2010)
  - Patterns for Nicknames/ Acronym mining (Zhang et al., 2011; Tamang et al., 2012)

  “full-name” (acronym) or “acronym (full-name)”, “city, state/country”
- Statistical models such as weighted finite state transducer (Friburger and Maurel, 2004)
  - CCP = “Communist Party of China”; “MINDEF” = “Ministry of Defence”
- Ambiguity drops from 46.3% to 11.2% (Chen and Ji, 2011; Tamang et al., 2012).

High-level Algorithmic Approach

- Input: A text document d; Output: a set of pairs (m_i, t_i)
  - o m_i are mentions in d; t_i are corresponding Wikipedia titles, or NIL.
- (1) Identify mentions m_i in d
- (2) Local Inference
  - o For each m_i in d:
    - Identify a set of relevant titles $T(m_i)$
    - Rank titles $t_i \in T(m_i)$

    [E.g., consider local statistics of edges (m_i, t_i), $(m_i, *)$, and $(*, t_i)$ occurrences of in the Wikipedia graph]
- (3) Global Inference
  - o For each document d:
    - Consider all $m_i \in d$; and all $t_i \in T(m_i)$
    - Re-rank titles $t_i \in T(m_i)$

    [E.g., if m, m' are related by virtue of being in d, their corresponding titles t, t' should also be related]

Generating Candidate Titles

- 1. Based on canonical names (e.g. Wikipedia page title)
  - Titles that are a super or substring of the mention
    - Michael Jordan is a candidate for "Jordan"
  - Titles that overlap with the mention
    - "William Jefferson Clinton" →Bill Clinton;
    - "non-alcoholic drink"→Soft Drink
- 2. Based on previously attested references
  - All Titles ever referred to by a given string in training data
    - Using, e.g., Wikipedia-internal hyperlink index
    - More Comprehensive Cross-lingual resource (Spitkovsky & Chang, 2012)

Initial Ranking of Candidate Titles

- Initially rank titles according to…
  - Wikipedia article length
  - Incoming Wikipedia Links (from other titles)
  - Number of inhabitants or the largest area (for geo-location titles)
- More sophisticated measures of prominance
  - Prior link probability
  - Graph based methods

P(t|m): "Commonness"

$$Commonness(m \Rightarrow t) = \frac{count(m \to t)}{\sum_{t' \in W} count(m \to t')}$$

Typography

By default, a font called Charcoal is used to replace the similar Chicago typefa
additional system fonts are also provided including Capitals, Gadget, Sand, Te
operating system need to be provided, such as the Command key symbol, ⌘. I

Airlines and destinations

Although the population of Iceland is only about 300,000, there are scheduled flights to and from seven locations in the United States (Boston, Chicago, Minneapolis, New York, Orlando, Seattle, and Washington), three in Canada (Halifax, Toronto and Winnipeg) and 30 cities across Europe. The largest carriers at Keflavík are Icelandair and Iceland Express.

P(Title|"Chicago")

The Greatest Show on Earth were a British rock band, who recorded two albums for Harvest Records in 1970.

The band had been conceived by Harvest Records in an attempt to create a horn-based rock combo, such as Blood Sweat & Tears or Chicago.[1]

P(t|m): "Commonness"

Rank	t	P(t|"Chicago")
1	Chicago	.76
2	Chicago (band)	.041
3	Chicago (2002_film)	.022
20	Chicago Maroons Football	.00186
100	1985 Chicago Whitesox Season	.00023448
505	Chicago Cougars	.0000528
999	Kimbell Art Museum	.00000586

- First used by Medelyan et al. (2008)
- Most popular method for initial candidate ranking

Note on Domain Dependence

- "Commonness" Not robust across domains

Formal Genre

Corpus	Recall
ACE	86.85%
MSNBC	88.67%
AQUAINT	97.83%
Wiki	98.59%

Ratinov et al. (2011)

Tweets

Metric	Score
P1	60.21%
R-Prec	52.71%
Recall	77.75%
MRR	70.80%
MAP	58.53%

Meij et al. (2012)

Graph Based Initial Ranking

- Centrality (Hachey et al., 2011; Hakimov et al., 2012)

$$Centrality(a) = \frac{\partial_a}{\sum_{b \in W} s(a,b)} * in\_links(a) * out\_links(a) * k$$

  - ∂_a : the number of all reachable nodes from a
  - $s(a,b)$: the distance between a and b

- Importance of the title with respect to Wikipedia - Similar to PageRank (Brin & Page, 1998)
  - Hachey et al. (2011) showed tha centrality works slightly better than PageRank

Basic Ranking Methods

- Local: Mention-Concept Context Similarity
  - Use **similarity measure** to compare the context of the mention with the text associated with a candidate title (the text in the corresponding page)
- Global: Document-wide Conceptual Coherence
  - Use topical/semantic **coherence** measures between the set of referent concepts for all mentions in a document

Context Similarity Measures

Determine assignment that maximizes pairwise similarity

$$\Gamma^* = \underset{\Gamma}{\operatorname{argmax}} \sum_i \varphi(m_i, t_i)$$

Γ

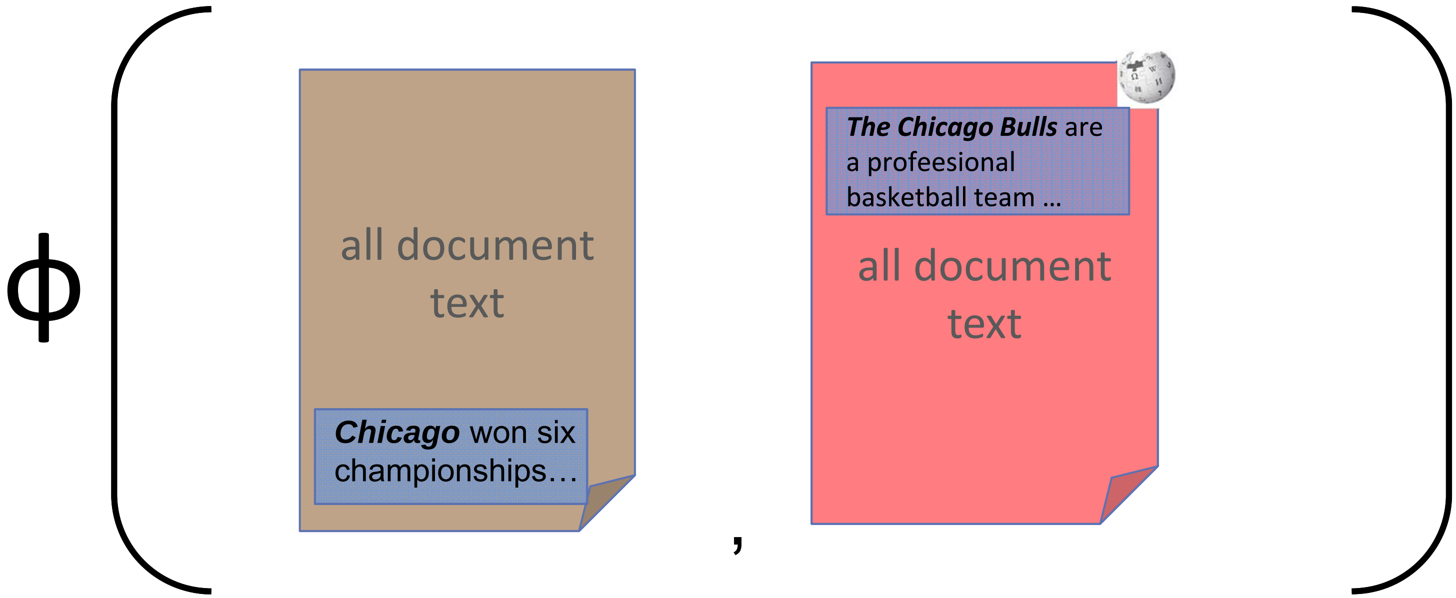
Mention-concept assignment

Mapping from mentions to titles

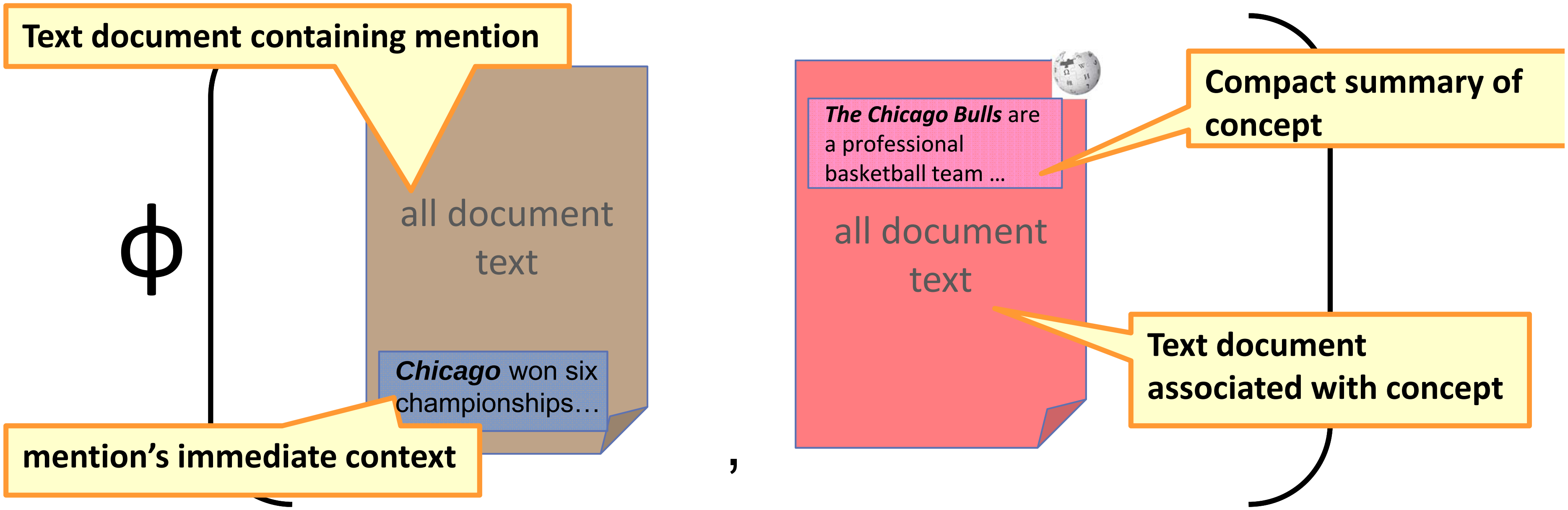
Feature vector to capture degree of contextual similarity

φ (Mention, Title)

Context Similarity Measures: *Context Source*

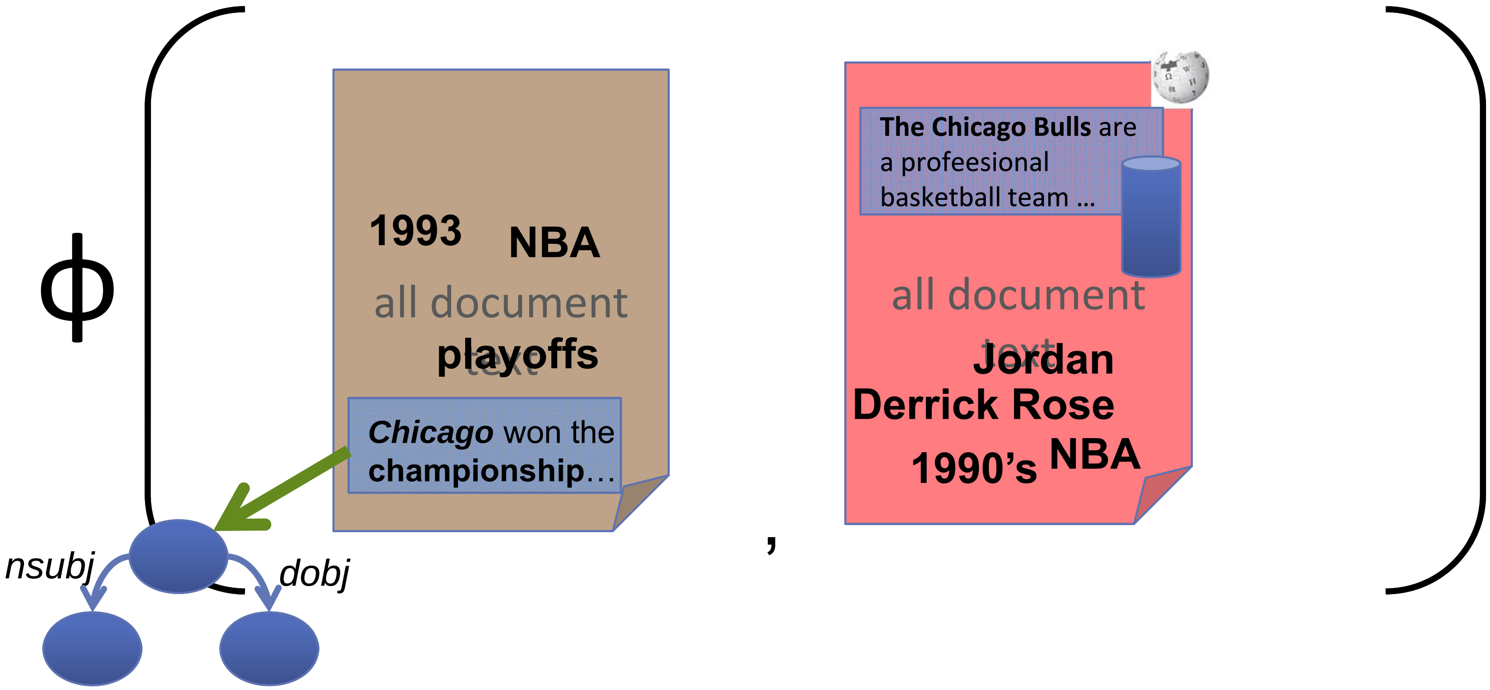


Context Similarity Measures: *Context Source*



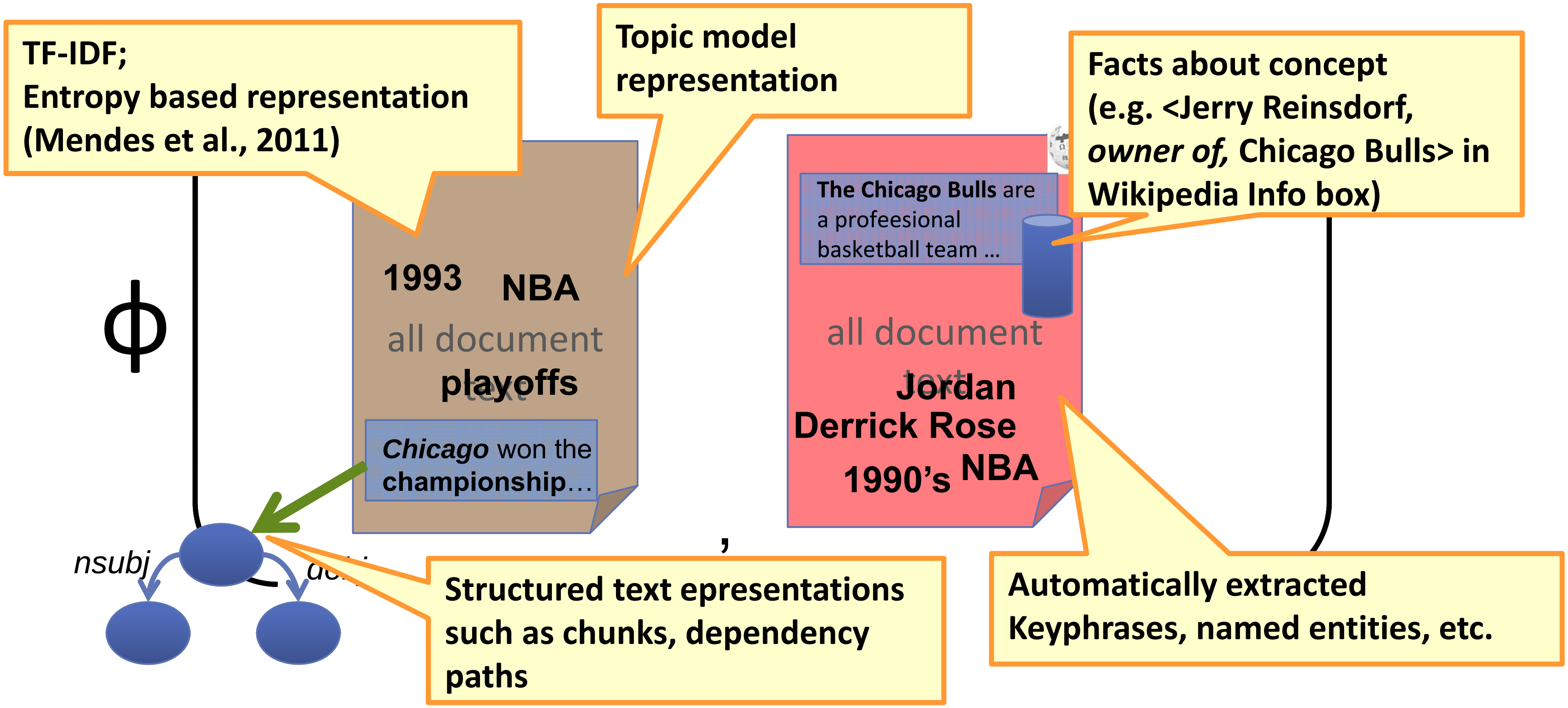
- Varying notion of distance between mention and context tokens
  - Token-level, discourse-level
- Varying granularity of concept description
  - Synopsis, entire document

Context Similarity Measures: *Context Analysis*



- Context is processed and represented in a variety of ways

Context Similarity Measures: *Context Analysis*



- Context is processed and represented in a variety of ways

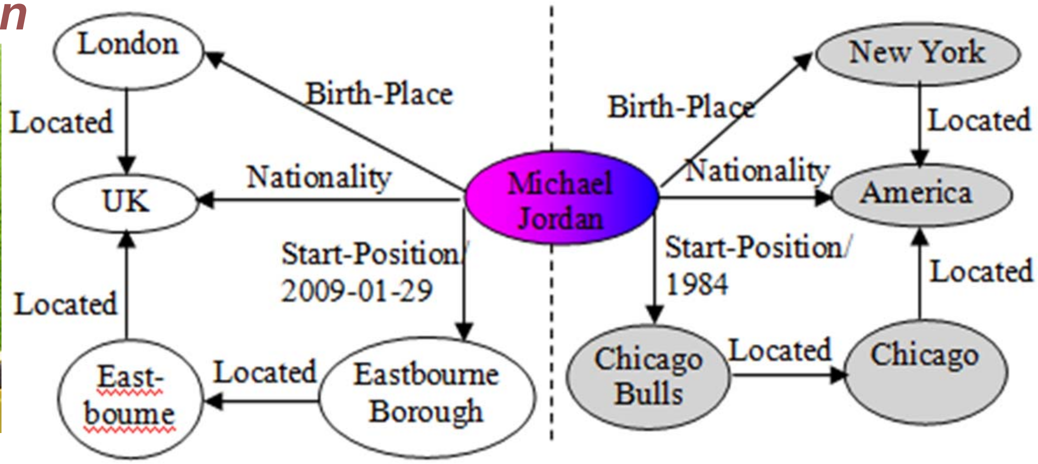
Typical Features for Ranking

Mention/Concept Attribute			Description
Name	Spelling match		Exact string match, acronym match, alias match, string matching…
	KB link mining		Name pairs mined from KB text redirect and disambiguation pages
	Name Gazetteer		Organization and geo-political entity abbreviation gazetteers
Document surface	Lexical		Words in KB facts, KB text, mention name, mention text.
			Tf.idf of words and ngrams
	Position		Mention name appears early in KB text
	Genre		Genre of the mention text (newswire, blog, …)
	Local Context		Lexical and part-of-speech tags of context words
Entity Context	Type		Mention concept type, subtype
	Relation/Event		Concepts co-occurred, attributes/relations/events with mention
	Coreference		Co-reference links between the source document and the KB text
Profiling			Slot fills of the mention, concept attributes stored in KB infobox
Concept			Ontology extracted from KB text
Topic			Topics (identity and lexical similarity) for the mention text and KB text
KB Link Mining			Attributes extracted from hyperlink graphs of the KB text
Popularity		Web	Top KB text ranked by search engine and its length
		Frequency	Frequency in KB texts

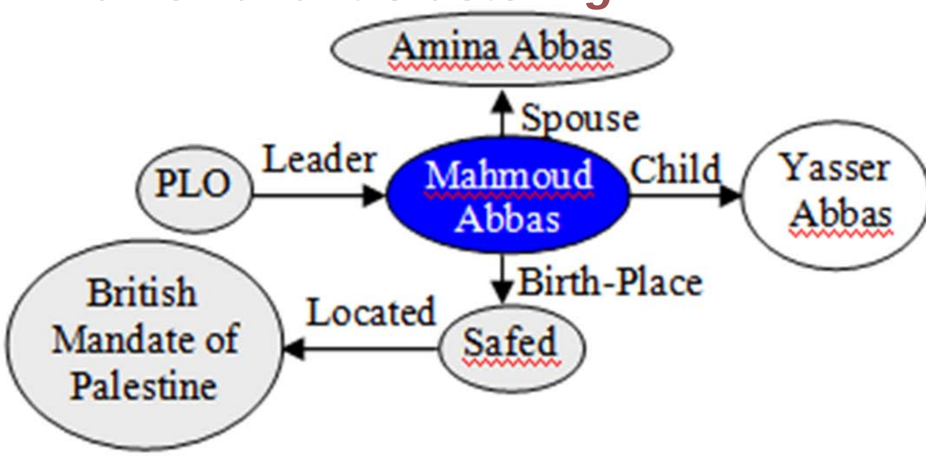
- (Ji et al., 2011; Zheng et al., 2010; Dredze et al., 2010; Anastacio et al., 2011; Zhou et al., 2014)

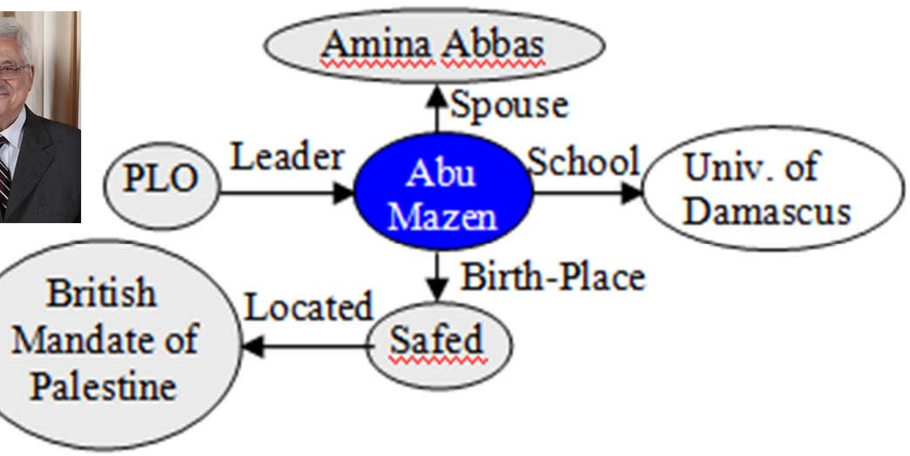
Entity Profiling Feature Examples

Disambiguation



Name Variant Clustering





- Deep semantic context exploration and indicative context selection (Gao et al., 2010; Chen et al., 2010; Chen and Ji, 2011; Cassidy et al., 2012)
- Exploit name tagging, Wikipedia infoboxes, synonyms, variants and abbreviations, slot filling results and semantic categories

Topic Feature Example

Li Na

player

tennis

Russia

single

final

gain

half

female

Li Na

Pakistan

relation

express

vice president

country

Prime minister

Topical features or topic based document clustering for context expansion (Milne and Witten, 2008; Syed et al., 2008; Srinivasan et al., 2009; Kozareva and Ravi, 2011; Zhang et al., 2011; Anastacio et al., 2011; Cassidy et al., 2011; Pink et al., 2013)

Putting it All Together

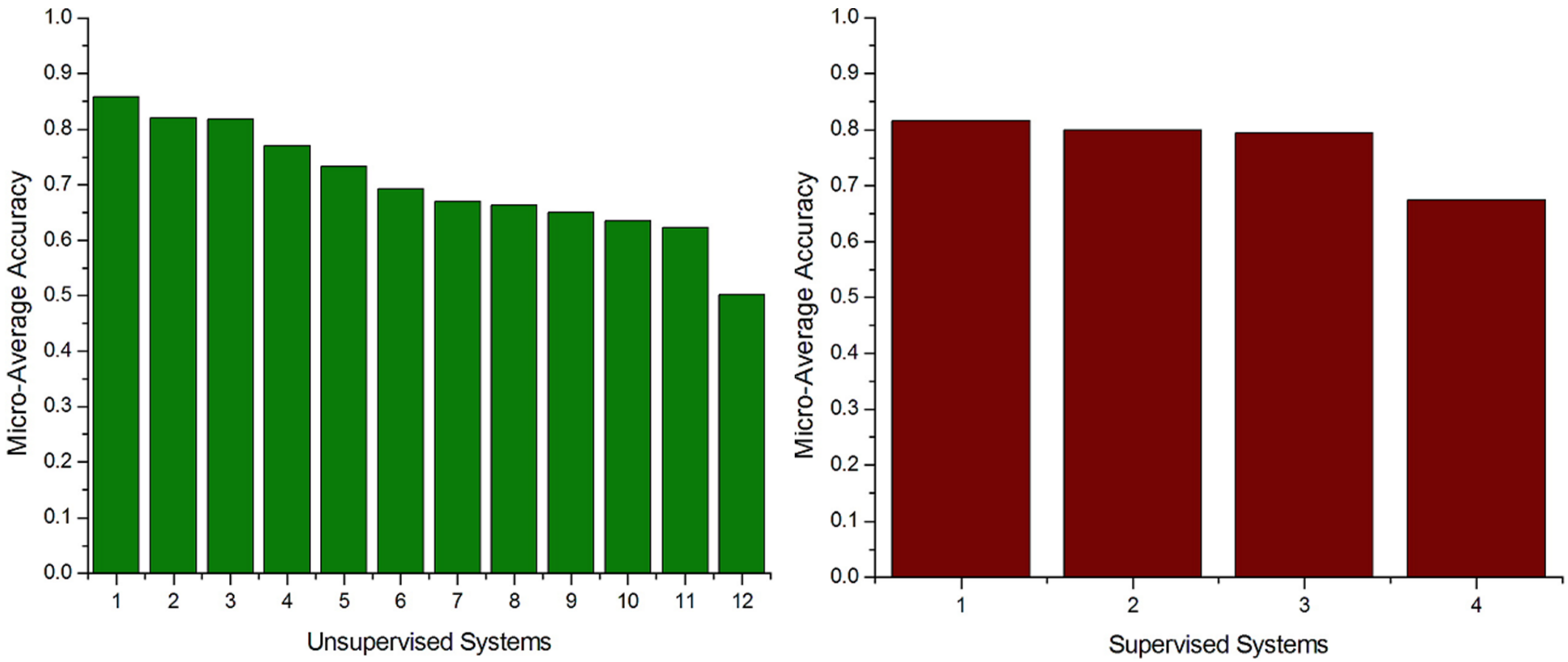
	Score Baseline	Score Context	Score Text
Chicago_city	0.99	0.01	0.03
Chicago_font	0.0001	0.2	0.01
Chicago_band	0.001	0.001	0.02

- Learning to Rank [Ratinov et. al. 2011]
  - Consider all pairs of title candidates
    - Supervision is provided by Wikipedia
  - Train a ranker on the pairs (learn to prefer the correct solution)
  - A Collaborative Ranking approach: outperforms many other learning approaches (Chen and Ji, 2011)

Ranking Approach Comparison

- Unsupervised or weakly-supervised learning (Ferragina and Scaiella, 2010)
  - Annotated data is minimally used to tune thresholds and parameters
  - The similarity measure is largely based on the unlabeled contexts
- Supervised learning (Bunescu and Pasca, 2006; Mihalcea and Csomai, 2007; Milne and Witten, 2008, Lehmann et al., 2010; McNamee, 2010; Chang et al., 2010; Zhang et al., 2010; Pablo-Sanchez et al., 2010, Han and Sun, 2011, Chen and Ji, 2011; Meij et al., 2012)
  - Each <mention, title> pair is a classification instance
  - Learn from annotated training data based on a variety of features
  - ListNet performs the best using the same feature set (Chen and Ji, 2011)
- Graph-based ranking (Gonzalez et al., 2012)
  - context entities are taken into account in order to reach a global optimized solution together with the query entity
- IR approach (Nemeskey et al., 2010)
  - the entire source document is considered as a single query to retrieve the most relevant Wikipedia article

Unsupervised vs. Supervised Ranking



- KBP2010 Entity Linking Systems (Ji et al., 2010)

High-level Algorithmic Approach

- Input: A text document d; Output: a set of pairs (m_i ,t_i)
  - o m_i are mentions in d; t_i are corresponding Wikipedia titles, or NIL.
- (1) Identify mentions m_i in d
- (2) Local Inference
  - o For each m_i in d:
    - Identify a set of relevant titles T(m_i)
    - Rank titles t_i ∈ T(m_i)

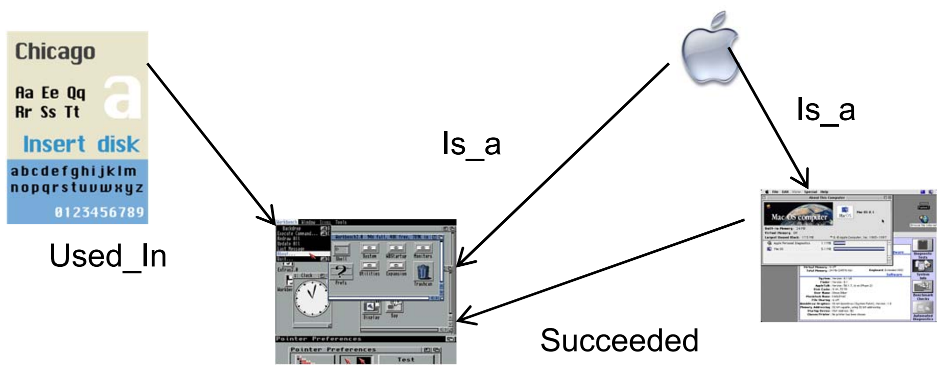
    [E.g., consider local statistics of edges (m_i ,t_i) , (m_i *), and (*, t_i) occurrences of in the Wikipedia graph]
- (3) Global Inference
  - o For each document d:
    - Consider all m_i ∈ d; and all t_i ∈ T(m_i)
    - Re-rank titles t_i ∈ T(m_i)

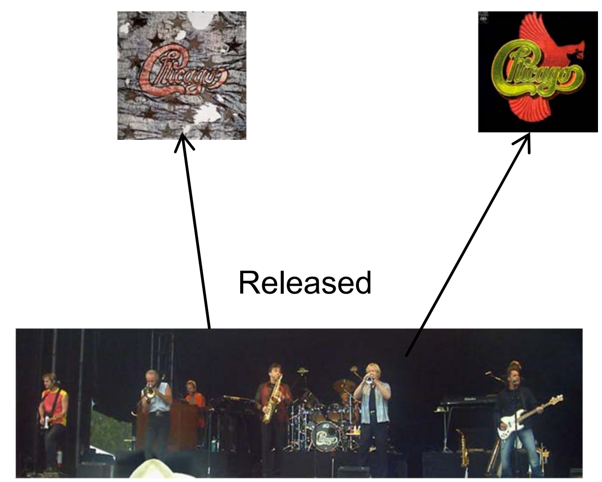
    [E.g., if m, m' are related by virtue of being in d, their corresponding titles t, t' should also be related]

Conceptual Coherence

- Recall: The reference collection (might) have structure.

It's a version of ***Chicago*** – the standard classic ***Macintosh*** menu font, with that distinctive thick diagonal in the "N".	***Chicago*** was used by default for ***Mac*** menus through ***MacOS 7.6***, and ***OS 8*** was released mid-1997..	***Chicago VIII*** was one of the early 70s-era ***Chicago*** albums to catch my ear, along with ***Chicago II***.





- Hypothesis:
  - Textual co-occurrence of concepts is reflected in the KB (Wikipedia)
- Incite:
  - Preferred disambiguation Γ contains structurally coherent concepts

Co-occurrence (Title 1, Title 2)

Typography

By default, a font called Charcoal is used to replace the similar Chicago typefa
additional system fonts are also provided including Capitals, Gadget, Sand, Te
operating system need to be provided, such as the Command key symbol, ⌘. I

Airlines and destinations

Although the population of Iceland is only about 300,000, there are scheduled flights to and from seven locations in the United States (Boston, Chicago, Minneapolis, New York, Orlando, Seattle, and Washington), three in Canada (Halifax, Toronto and Winnipeg) and 30 cities across Europe. The largest carriers at Keflavík are Icelandair and Iceland Express.

The city senses of Boston and Chicago appear together often.

The Greatest Show on Earth were a British rock band, who recorded two albums for Harvest Records in 1970.

The band had been conceived by Harvest Records in an attempt to create a horn-based rock combo, such as Blood Sweat & Tears or Chicago.[1]

Co-occurrence(Title1, Title2)

Typography

By default, a font called Charcoal is used to replace the similar Chicago typefa
additional system fonts are also provided including Capitals, Gadget, Sand, Te
operating system need to be provided, such as the Command key symbol, ⌘. I

Airlines and destinations

Although the population of Iceland is only about 300,000, there are scheduled flights to and from seven locations in the United States (Boston, Chicago, Minneapolis, New York, Orlando, Seattle, and Washington), three in Canada (Halifax, Toronto and Winnipeg) and 30 cities across Europe. The largest carriers at Keflavík are Icelandair and Iceland Express.

Rock music and albums appear together often

The Greatest Show on Earth were a British rock band, who recorded two albums for Harvest Records in 1970.

The band had been conceived by Harvest Records in an attempt to create a horn-based rock combo, such as Blood Sweat & Tears or Chicago.[1]

Global Ranking

$$\Gamma^* \approx \arg\max_{\Gamma} \sum_{i=1}^{N} [\phi(m_i, t_i) + \sum_{t_i \in \Gamma, t_j \in \Gamma'} \psi(t_i, t_j)]$$

- How to approximate the “global semantic context” in the document”?
  - It is possible to only use non-ambiguous mentions as a way to approximate it.
- How to define relatedness between two titles? (What is Ψ?)

Title Coherence & Relatedness

- Let c, d be a pair of titles …
- Let C and D be their sets of incoming (or outgoing) links
  - Unlabeled, directed link structure

Introduced by Milne &Witten (2008) Used by Kulkarni et al. (2009), Ratinov et al (2011), Hoffart et al (2011),

$$relatedness(c,d) = \frac{\log\left(\max\left(|C|,|D|\right)\right) - \log\left(|C \cap D|\right)}{\log(W) - \log\left(\min\left(|C|,|D|\right)\right)}$$

See García et al. (JAIR2014) for variational details

$$PMI(c,d) = \frac{|C \cap D| / |W|}{\left(|C| / |W|\right) * \left(D / |W|\right)}$$

Relatedness Outperforms Pointwise Mutual Information (Ratinov et al., 2011)

- Let C and D $\in \{0,1\}^K$, where K is the set of all categories

$$relatedness(c,d) = \langle C, D \rangle$$

Category based similarity introduced by Cucerzan (2007)

More Relatedness Measures (Ceccarelli et al., 2013)

Singleton Features	
P(a)	probability of a mention to entity a: $P(a) = \|\|in(a)\|/\|W\|$.
H(a)	entropy of a: $H(a) = -P(a)\log(P(a)) - (1-P(a))\log(1-P(a))$.

Asymmetric Features	
P(a|b)	conditional probability of the entity a given b: $P(a\|b) = \|in(a) \cap in(b)\| \,/\, \|in(b)\|$.
Link$(a \to b)$	equals 1 if a links to b, and 0 otherwise.
$P(a \to b)$	probability that a links to b: equals $1/\|out(a)\|$ if a links to b, and 0 otherwise.
Friend(a, b)	equals 1 if a links to b, and $\|out(a) \cap in(b)\|/\|out(a)\|$ otherwise.
$KL(a\|\|b)$	Kullback-Leibler divergence: $KL(a\|\|b) = \log\frac{P(a)}{P(b)}P(a) + \log\frac{1-P(a)}{1-P(b)}(1 - P(a))$.

More Relatedness Measures (Ceccarelli et al., 2013)

Symmetric Features	
$\rho^{\mathsf{MW}}(a,b)$	co-citatation based similarity [19].
$J(a,b)$	Jaccard similarity: $J(a,b) = \frac{in(a)\cap in(b)}{in(a)\cup in(b)}$.
$P(a,b)$	joint probability of entities a and b: $P(a,b) = P(a\|b)\cdot P(b) = P(b\|a)\cdot P(a)$.
$\mathsf{Link}(a \leftrightarrow b)$	equals 1 if a links to b and vice versa, 0 otherwise.
$\mathsf{AvgFr}(a,b)$	average friendship: $(\mathsf{Friend}(a,b) + \mathsf{Friend}(b,a))/2$.
$\rho^{\mathsf{MW}}_{\mathsf{out}}(a,b)$	ρ^{MW} considering outgoing links.
$\rho^{\mathsf{MW}}_{\mathsf{in\text{-}out}}(a,b)$	ρ^{MW} considering the union of the incoming and outgoing links.
$J_{\mathsf{out}}(a,b)$	Jaccard similarity considering the outgoing links.
$J_{\mathsf{in\text{-}out}}(a,b)$	Jaccard similarity considering the union of the incoming and outgoing links.
$\chi^2(a,b)$	χ^2 statistic: $\chi^2(a,b) = (\|in(b)\cap in(a)\|\cdot(\|W\| - \|in(b)\cup in(a)\|) + -\|in(b)\setminus in(a)\|\cdot\|in(a)\setminus in(b)\|)^2 \cdot \frac{\|W\|}{\|in(a)\|\cdot\|in(b)\|(\|W\|-\|in(a)\|)(\|W\|-\|in(b)\|)}$
$\chi^2_{\mathsf{out}}(a,b)$	χ^2 statistic considering the outgoing links.
$\chi^2_{\mathsf{in\text{-}out}}(a,b)$	χ^2 statistic considering the union of the incoming and outgoing links.
$\mathsf{PMI}(a,b)$	point-wise mutual information: $\log\frac{P(b\|a)}{P(b)} = \log\frac{P(a\|b)}{P(a)} = \log\frac{\|in(b)\cap in(a)\|\|W\|}{\|in(b)\|\|in(a)\|}$

More Relatedness Measures (Ceccarelli et al., 2013)

Features	Rank	NDCG@5	NDCG@10	P@5	P@10	MRR
P(c|e)	**1**	**0.68**	**0.72**	**0.47**	**0.33**	**0.80**
J(e, c)	**2**	0.62	0.66	0.44	0.31	0.75
Friend(e,c)	24	0.59	0.64	0.42	0.31	0.71
$\rho^{\mathsf{MW}}(e,c)$	19	0.59	0.63	0.42	0.31	0.72
$J_{\mathsf{in-out}}(e,c)$	26	0.60	0.63	0.42	0.30	0.74
AvgFr(e,c)	**3**	0.57	0.62	0.40	0.30	0.69
P(e,c)	27	0.56	0.60	0.39	0.28	0.70
$\rho^{\mathsf{MW}}_{\mathsf{in-out}}(a,b)$	9	0.56	0.60	0.40	0.29	0.71
$J_{\mathsf{in-out}}(e,c)$	**4**	0.54	0.58	0.39	0.28	0.67
$\rho^{\mathsf{MW}}_{\mathsf{out}}(a,b)$	17	0.52	0.55	0.37	0.27	0.65
$\chi^2(e,c)$	25	0.51	0.55	0.37	0.27	0.64
P(e|c)	22	0.48	0.54	0.36	0.28	0.60
H(c)	**5**	0.48	0.51	0.30	0.20	0.68
$\chi^2_{\mathsf{out}}(e,c)$	16	0.47	0.50	0.34	0.24	0.61
AvgFr(c,e)	21	0.44	0.49	0.33	0.25	0.56
P(c)	13	0.47	0.49	0.29	0.19	0.66
PMI(e,c)	23	0.42	0.48	0.32	0.25	0.53
$\chi^2_{\mathsf{in-out}}(e,c)$	11	0.44	0.46	0.33	0.23	0.58
$P(e \rightarrow c)$	18	0.37	0.38	0.24	0.15	0.55
Link$(e \rightarrow c)$	20	0.37	0.38	0.24	0.15	0.55
$P(c \rightarrow e)$	12	0.35	0.36	0.22	0.14	0.52
Link$(c \rightarrow e)$	15	0.31	0.33	0.21	0.14	0.46
$KL(c\|\|e)$	10	0.32	0.32	0.19	0.12	0.51
Link$(c \leftrightarrow e)$	14	0.28	0.29	0.17	0.11	0.45
$KL(e\|\|c)$	8	0.26	0.28	0.17	0.11	0.44
P(e)	6	0.08	0.11	0.06	0.06	0.17
H(e)	7	0.08	0.11	0.06	0.06	0.17

Densest Subgraph Heuristic (Moro et al., TACL2014)

- Target KB: **Babelnet** (Navigli & Ponzetto, *AI* 2012)
  - A **semantic network** of concepts (including named entities), with typed edges for semantic relations, in multiple languages.
- Babelfy System
  - 1. Assign weights to and remove labels from edges using *directed triangles*
    - Inspired by (Watts & Strogatz 1998)
  - 2. Create *semantic signature* via Random Walk with Restart (RWR) (Tong et al., 2006) using edge weights for probability
    - SemSign_c – set of concepts most related to *c* based on RWR
  - 3. Graph - V: (m, c) candidates E: based on SemSign
  - 4. Reduce ambiguity by approximating Densest Subgraph
    - Hypothesis: The best concept for a mention comes from the densest portion of the graph

NIL Detection and Clustering

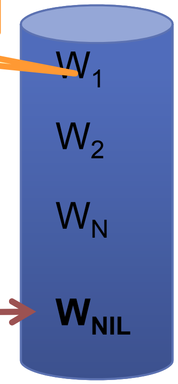
- The key difference between Wikification and Entity Linking is the way NIL are treated.
- In Wikification:
  - Local Processing
  - Each mention m_i that does not correspond to title t_i is mapped to NIL.
- In Entity Linking:
  - Global Processing
  - Cluster all mentions m_i that represent the same concept
  - If this cluster does not correspond to a title t_i, map it to NIL.
- Mapping to NIL is challenging in both cases

NIL Detection

1. Augment KB with NIL entry and <u>treat it like any other entry</u>
2. Include general NIL-indicating features

Is it in the KB?

{ , *NIL* }



KB

Jordan accepted a basketball scholarship to North Carolina, …

In the 1980's **Jordan** began developing recurrent neural networks.

Local man **Michael Jordan** was appointed county coroner …

1. Binary classification (Within KB vs. NIL)
2. Select NIL cutoff by tuning confidence threshold

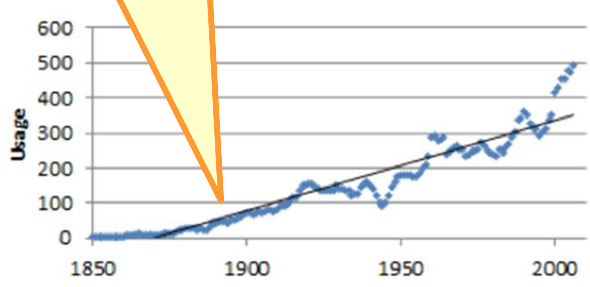
Is it an *entity*?

- Concept Mention Identification (above)
- Not all NP's are linkable

No spike: Not an entity

Sudden Google Books frequency spike: Entity

"Prices Quoted"



"Soluble Fiber"

NIL Detection: Main Challenges

- Wikipedia's hyperlinks offer a wealth of disambiguated mentions that can be leveraged to train a Wikification system.
- However, relative to mentions from general text, Wikipedia mentions are disproportionately likely to have corresponding Wikipedia pages
- Accounting for this bias from statistical models requires more than simply training a Wikification system on a moderate number of examples from non-Wikipedia text
- Applying distinct semi-supervised and active learning approaches to the task is a primary area of future work
- More advanced selectional preference methods should be applied to solve the cases when the correct referent is ranked very low by statistical models, and combine multi-dimensional clues

NIL Clustering

“All in one”

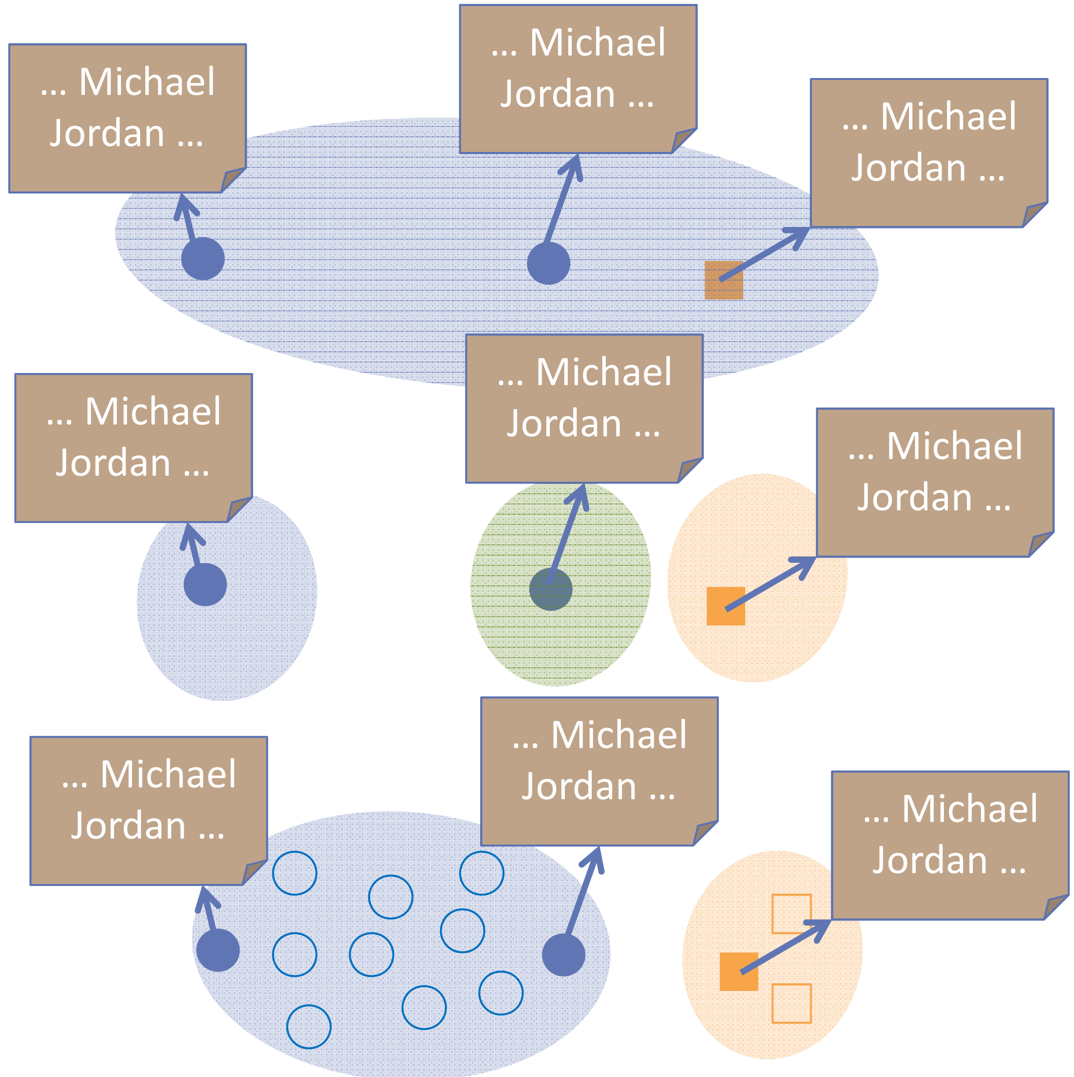
Simple string matching

“One in one”

Often difficult to beat!

Collaborative Clustering

Most effective when ambiguity is high



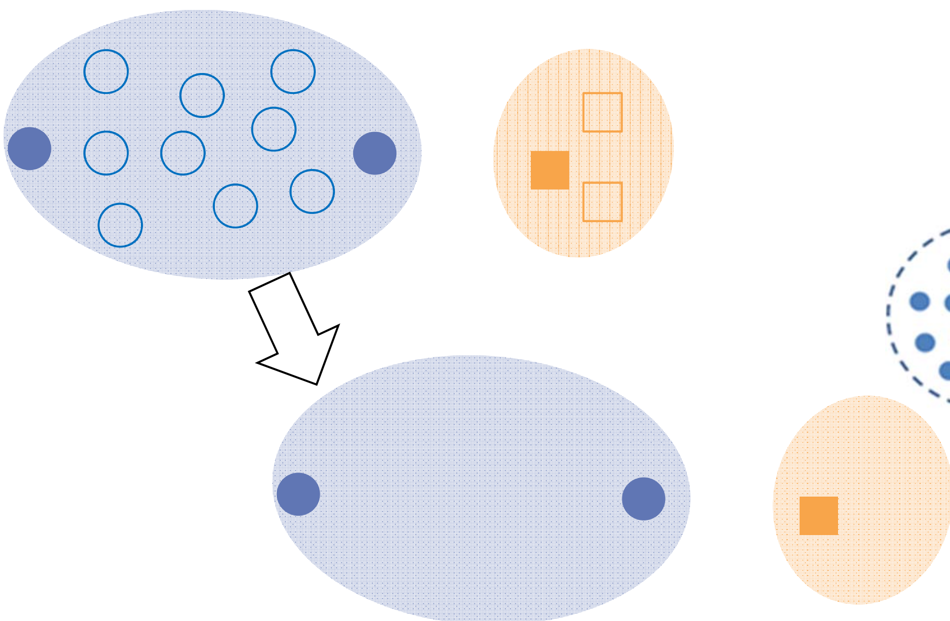
NIL Clustering Methods Comparison (Chen and Ji, 2011; Tamang et al., 2012)

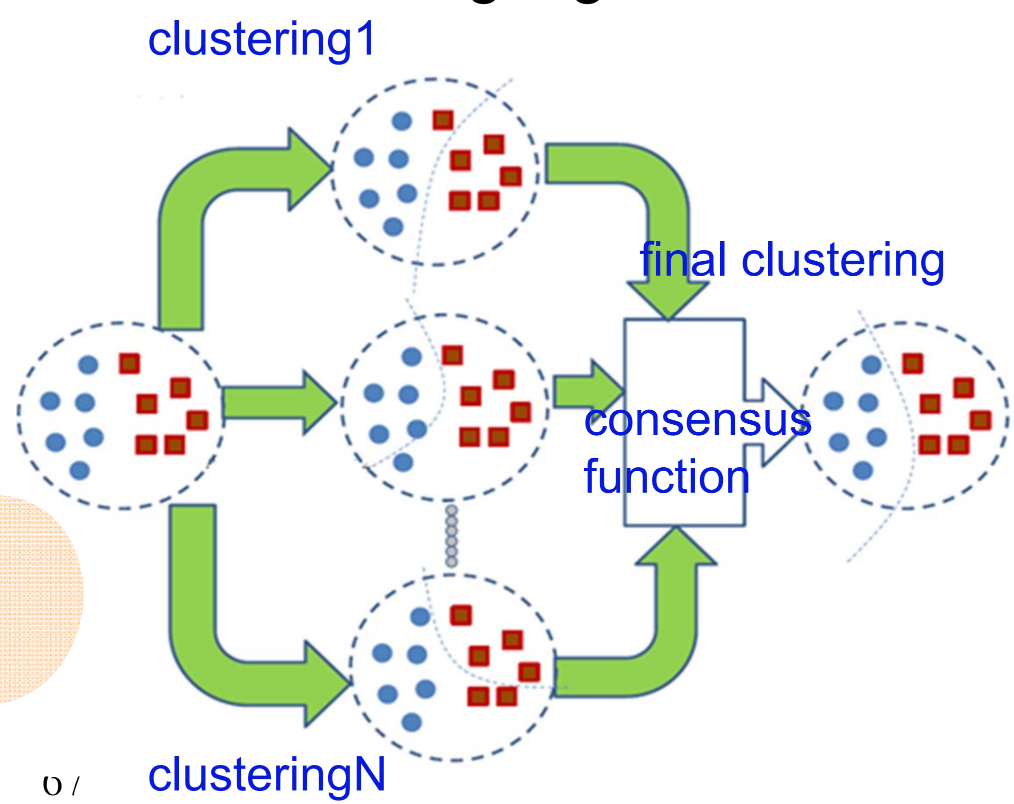
Algorithms		B-cubed+ F-Measure	Complexity
Agglomerative clustering	3 linkage based algorithms (single linkage, complete linkage, average linkage) (Manning et al., 2008)	85.4%-85.8%	$O(n^2)$ $O(n^2 \log n)$ n: the number of mentions
	6 algorithms optimizing internal measures cohesion and separation	85.6%-86.6%	$O(n^2 \log n)$ $O(n^3)$
Partitioning Clustering	6 repeated bisection algorithms optimizing internal measures	85.4%-86.1%	$O(NNZ \times k + m \times k)$ NNZ: the number of non-zeroes in the input matrix M: dimension of feature vector for each mention k: the number of clusters
	6 direct k-way algorithms optimizing internal measures (Zhao and Karypis, 2002)	85.5%-86.9%	$O(NNZ \times \log k)$

- **Co-reference methods** were also used to address NIL Clustering (E.g., Cheng et. al 2013): L^{3}M Latent Left Linking jointly learn metric and clusters mentions

Collaborative Clustering (Chen and Ji, 2011; Tamang et al., 2012)

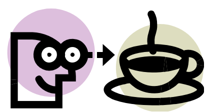
- Consensus functions
  - Co-association matrix (Fred and Jain,2002)
  - Graph formulations (Strehl and Ghosh, 2002; Fern and Brodley, 2004): instance-based; cluster-based; hybrid bipartite
- 12% gain over the best individual clustering algorithm





Outline

- Motivation and Definition
- A Skeletal View of a Wikification System
    - o High Level Algorithmic Approach
- Key Challenges



Coffee Break

- Recent Advances
- New Tasks, Trends and Applications
- What's Next?
- Resources, Shared Tasks and Demos

General Challenges

Blumenthal (D) is a candidate for the U.S. Senate seat now held by Christopher Dodd (D), and he has held a commanding lead in the race since he entered it. But the Times report has the potential to fundamentally reshape the contest in the Nutmeg State.

- Ambiguity

Times → The New York Times
Times → The Times

⋮

- Variability

CT → Connecticut
The Nutmeg State → Connecticut

⋮

- Concepts outside of Wikipedia (NIL)
    - Blumenthal ?

- Scale
    - Millions of labels

General Challenges

- A few researchers focused on efficiency of Wikification (e.g. stacking (He et al., 2013) and distributional hierarchical clustering (Bekkerman et al., 2005; Singh et al. 2011)); most others focus on improving quality

> Blumenthal (D) is a candidate for the U.S. Senate seat now held by Christopher Dodd (D), and he has held a commanding lead in the race since he entered it. But the Times report has the potential to fundamentally reshape the contest in the Nutmeg State.

- State-of-the-art systems (Ratinov et al. 2011) can achieve the above with local and global statistical features
  - Reaches bottleneck around 70%~ 85% F1 on non-wiki datasets
  - What is missing?

Challenges

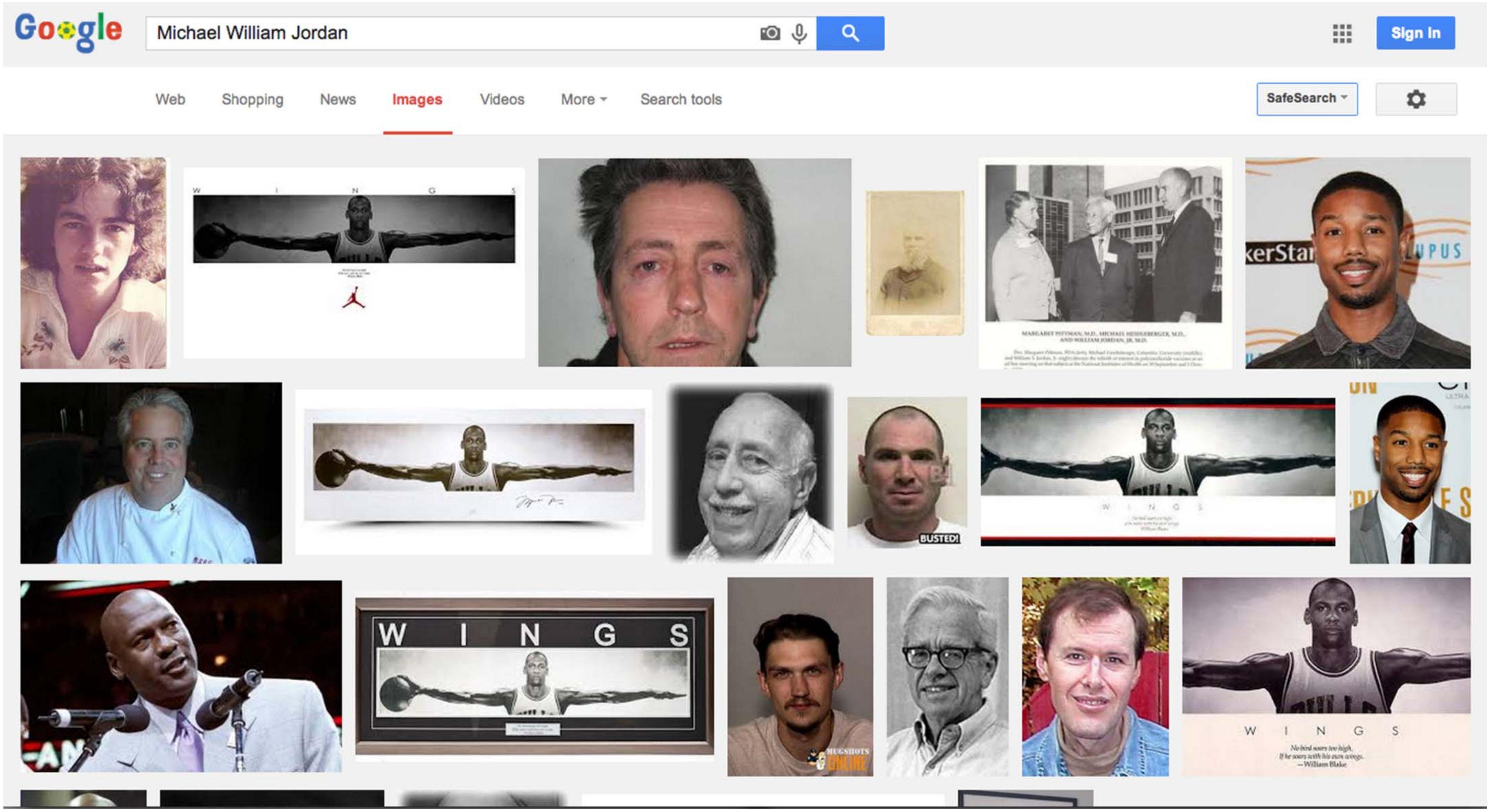
- Dealing with Popularity Bias

- Exploiting Semantic Knowledge to Improve Wikification
  - Relational Information in the text

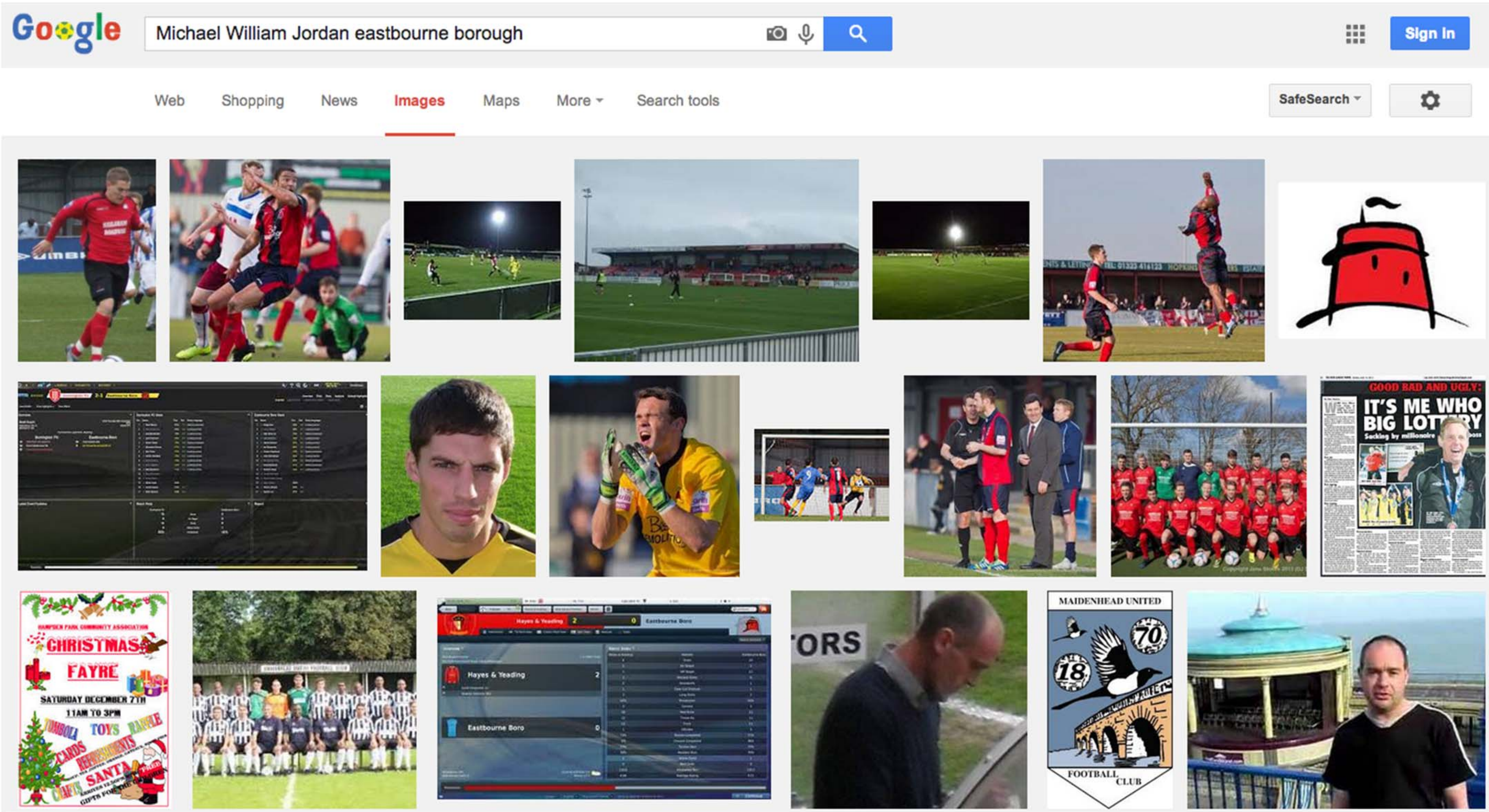
- Recovering from gaps in background knowledge
  - Mostly when dealing with short texts and social media

- Exploiting common sense knowledge

Popularity Bias: If you are called Michael Jordan…



A Little Better…



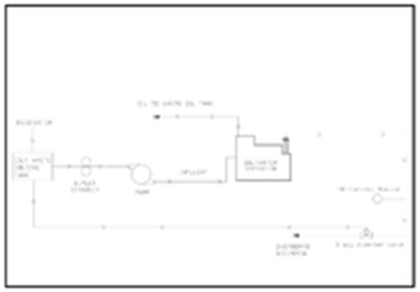
Deep Semantic Knowledge

Local **OWS** activists were part of this **protest**.

Order of World Scouts

Occupy Wall Street

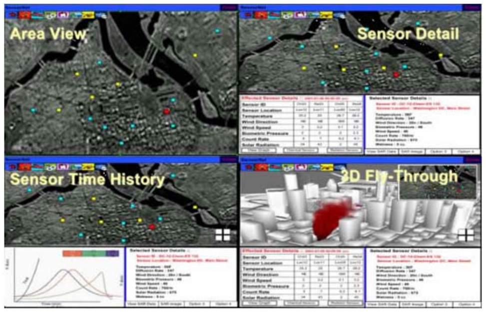
Oily Water Separator



Overhead Weapon Station

Open Window School

Open Geospatial Consortium



Deep Semantic Knowledge

- An Australian jury found that an Uzbekistan Olympic boxing official was defamed in a book about alleged Olympic corruption in which he was depicted as a major international heroin dealer and racketeer.
- Rakhimov was also said to have bribed members of the International Boxing Federation in the vote for the Federation Presidency.

International Boxing Association (amateur), olympic-style

International Boxing Association (professional body), organization that sanctions professional boxing

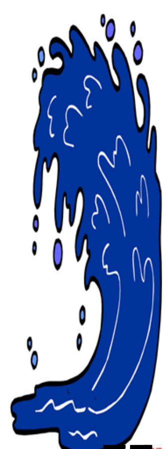
Deep Semantic Knowledge

- It was a pool **report** typo. Here is exact **Rhodes** quote: ”this is not gonna be a couple of weeks. It will be a period of days.”
- At a **WH briefing** here in Santiago, **NSA** spox **Rhodes** came with a litany of pushback on idea **WH** didn’t consult with **Congress**.
- **Rhodes singled out** a **Senate** resolution that passed on March 1st which denounced **Khaddafy’s** atrocities. **WH** says **UN** rez incorporates it

Ben Rhodes (Speech Writer)

Knowledge Gap between Source and KB

Source: breaking news/new information/rumor	KB: bio, summary, snapshot of life
According to **Darwin** it is the **Males** who do the vamping.	**Charles Robert Darwin**, was an English **naturalist** and **geologist** best known for his contributions to **evolutionary theory**.
I had no idea the **victim** in the **Jackson cases** was publicized.	In the summer of 1993, **Jackson** was accused of **child sexual abuse** by a 13-year-old boy named Jordan Chandler and his father, Dr. Evan Chandler, a dentist.
I went to youtube and checked out the **Gulf oil crisis**: all of the posts are one month old, or older…	On April 20, 2010, the Deepwarter Horizon oil platform, located in the Mississippi Canyon about 40 miles (64 km) off the Louisiana coast, suffered **a catastrophic explosion**; it sank a day-and-a-half later



Fill in the Gap with Background Knowledge

Source: breaking news/new information/rumors	KB: bio, summary, snapshot of life
Christies denial of marriage privledges to gays will alienate independents and his "I wanted to have the people vote on it" will ring hollow.	Christie has said that he favoured New Jersey's law allowing same-sex couples to form civil unions, but would veto any bill legalizing same-sex marriage in New Jersey
Translation out of hype-speak: some kook made threatening noises at Brownback and go arrested	Samuel Dale "Sam" Brownback (born September 12, 1956) is an American politician, the 46th and current Governor of Kansas.

Connect/Sort Background Knowledge

Man Accused Of Making Threatening Phone Call To Kansas Gov. Sam Brownback May Face Felony Charge

Background Knowledge

Making **pesto**! I had to soak my **nuts** for 3 hours

Background Knowledge

Awesome post from wolfblitzercnn: Behind the scenes on **Clinton**'s Mideast trip - URL - #cnn

Background Knowledge

Iran and **Russia** will be the next war along with **Chavez** if we do not create a successful **democracy** in **Iraq**.

Portrait of Carlos Chávez by Carl van Vechten (1937)

Cesar Chavez

Hugo Chávez

- Chavez's opposition to Zionism and close relations with **Iran**, have led to accusations of antisemitism
- Soon after this speech, in August Chávez announced that his government would nationalize Venezuela's gold industry,… to banks in Venezuela's political allies like **Russia**, China and Brazil.
- The CD and other opponents of Chávez's Bolivarian government accused it of trying to turn Venezuela from a **democracy** into a dictatorship by…

Background Knowledge

2005-06-05
Taiwan (TW)
International; weapons
Taiwan successfully fired its first cruise missile.
This will enable Taiwan to hit major military targets in southeast China.
The China Times reported that Taiwan has successfully test fired the Hsiung Feng its first cruise missile enabling Taiwan to hit major military targets in southeast China.

Hsiung Feng IIE

Hsiung Feng

From Wikipedia, the free encyclopedia

Hsiung Feng can refer to:

- Hsiung Feng I
- Hsiung Feng II
- Hsiung Feng IIE
- Hsiung Feng III

Background Knowledge

1995-12-18
Germany (DE)
International; weapons; war and conflict
There are huge obstacles to achieving peace and cooperation among combatants in the former Yugoslavia.
German Foreign Minister Klaus Kinkel said in opening remarks at the one-day meeting that there can be no peace in the former Yugoslavia if some parties to the conflict remain heavily armed and others try to catch up.

1918-1929: *Kingdom* of *Serbs*, *Croats and Slovenes*
1929-1941: Kingdom of Yugoslavia
1945-1946: Yugoslavia Democratic Union
1946-1963: Federal People's Republic of Yugoslavia
1963-1992: Socialist Federal Republic of Yugoslavia
1992-2003: Federal Republic of Yugoslavia
2003-2006: Serbia and Montenegro
…

Commonsense Knowledge

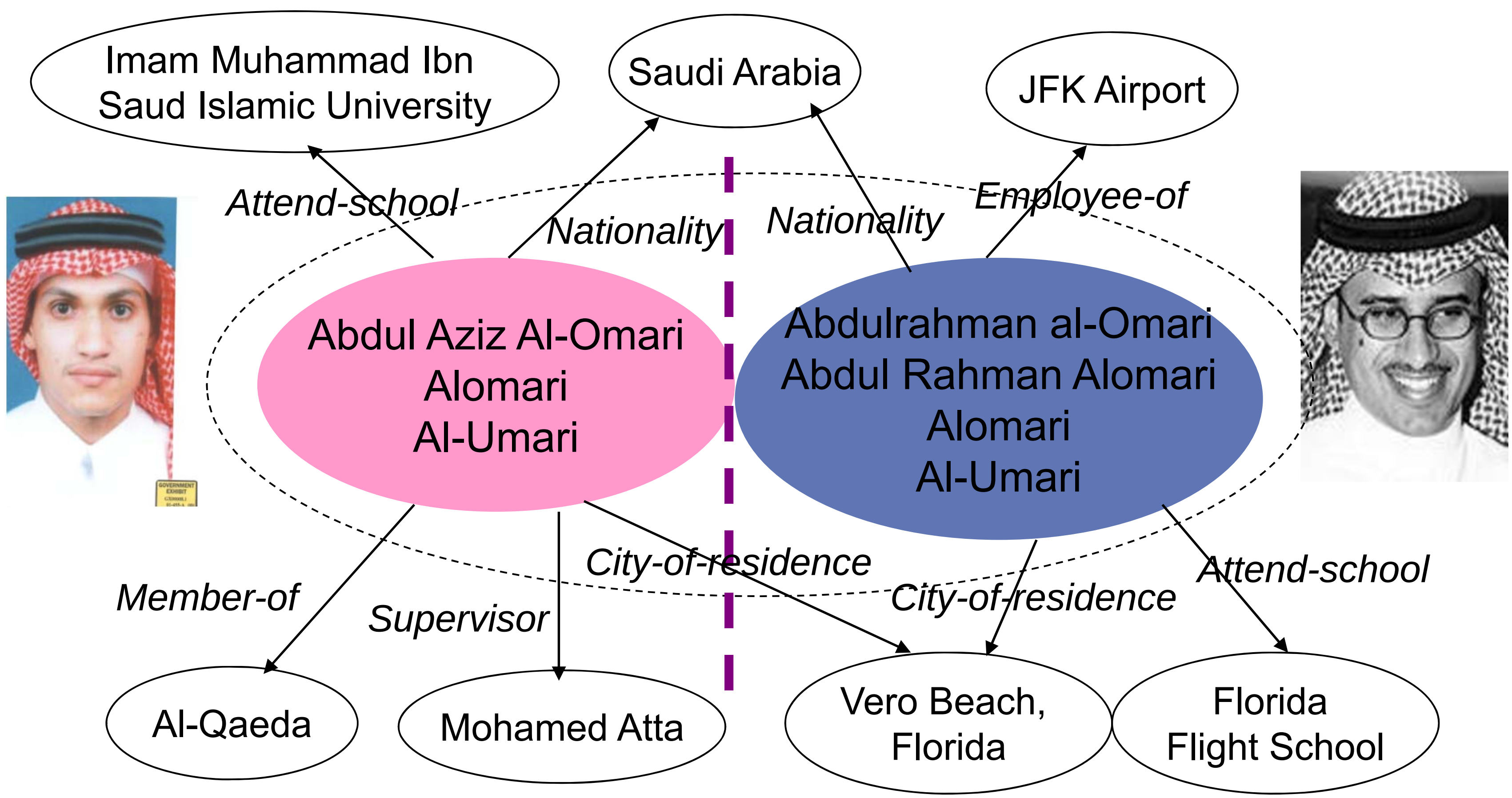
2008-07-26

During talks in Geneva attended by William J. Burns Iran refused to respond to Solana's offers.

William_J._Burns (1861-1932)

William_Joseph_Burns (1956-)

Rich Context (Information Networks)



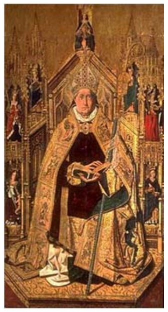
Rich Context

- “Supreme Court” (in Japan, China, U.S., Macedonia, etc.)
- “LDP (Liberty and Democracy Party)” (in Australia, Japan, etc.)
- “Newcastle University” can be located in UK or Australia
- Many person entities share the same common names such as “Albert”, “Pasha”, etc.
- “Ji county” can be located in “Shanxi” or “Tianjin”

Rich Context: Coreferential Mentions

Brazilian government and **Abbott Laboratories** agree on lower price for AIDS drug Kaletra in response to Brazilian threat to violate the patent.

According to WHO studies the price of the drug was exorbitant and the Brazilian government demanded that **Abbot** lower the price.



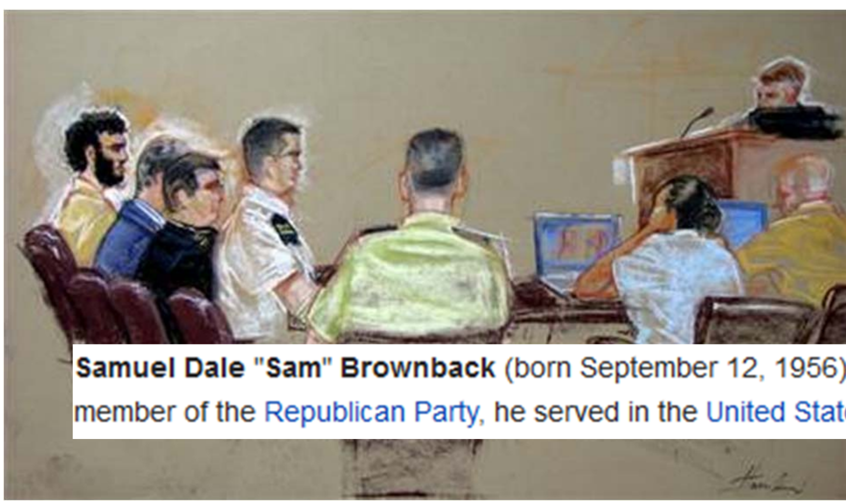
- Finding collaborators based cross-document entity clustering (Chen and Ji, 2011)

Rich Context: Related Employer

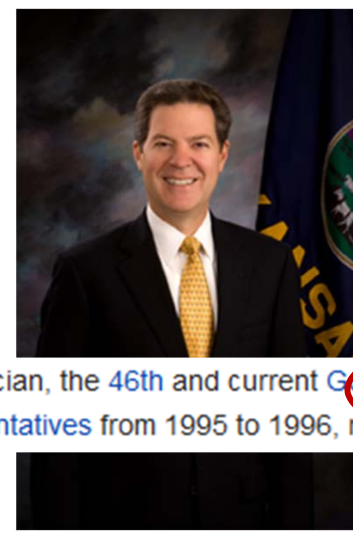
Hundreds of protesters from various groups converged on the state capitol in Topeka, Kansas today…

Second, I have a really hard time believing that there were any ACTUAL "explosives" since the news story they link to talks about one guy getting arrested for THREATENING Governor Brownback.

Peter Brownback



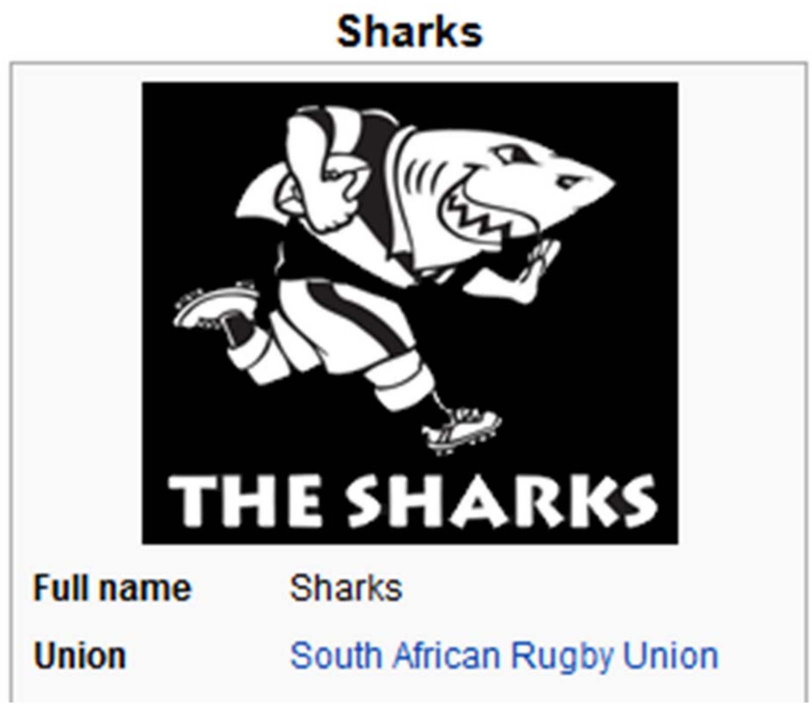
Sam Brownback



Samuel Dale "Sam" Brownback (born September 12, 1956) is an American politician, the 46th and current Governor of Kansas. A member of the Republican Party, he served in the United States House of Representatives from 1995 to 1996, representing Kansas's

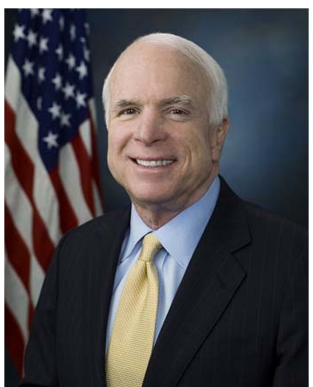
Rich Context: Related Employees

*I check the numbers, and am shocked to learn the Sharks have over 10 million in available cap room. **Antropov** would fit in great with the **Sharks**, while **McCabe** would be the big shot (you will remember last year how much they pursued **Souray**).*



Rich Context: Related Colleagues

Where would McCain be without Sarah?



Alaska Governor Sarah Palin was revealed as McCain's surprise choice for running mate on August 29, 2008.[234]

Sarah Louise Palin (/ˈpeɪlɪn/; née **Heath**; born February 11, 1964) is an American politician, commentator and author who served as the ninth Governor of Alaska, from 2006 to 2009. As the Republican Party nominee for Vice President in the 2008 presidential election alongside Arizona Senator John McCain, she was the first Alaskan on the national ticket of a major party and first Republican

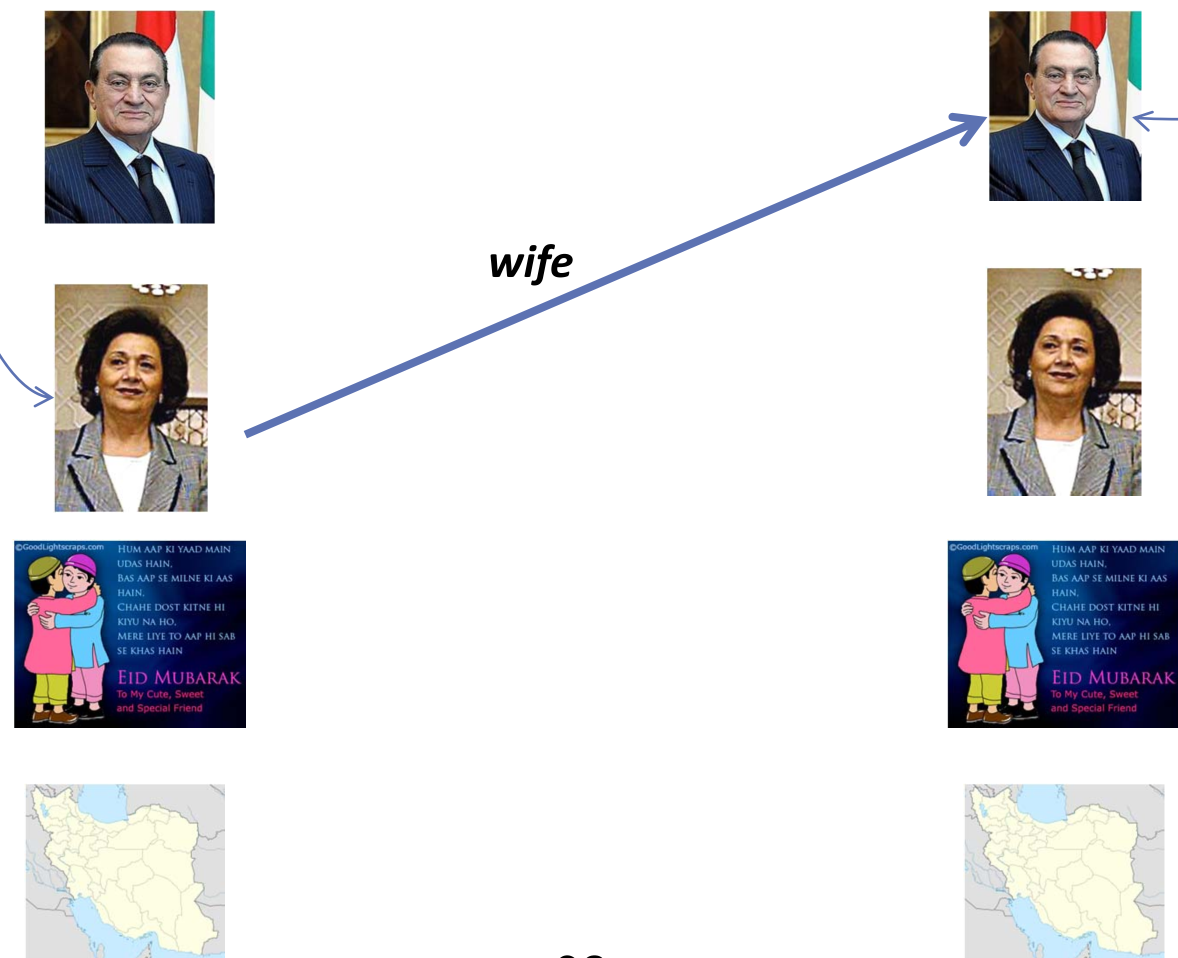
Rich Context: Related Colleagues

No matter what, he never should have given Michael Jackson that propofol. He seems to think a “proper” court would have let Murray go free.

The trial of Conrad Murray was the American criminal trial of Michael Jackson's personal physician, Conrad Murray.

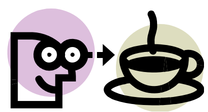
Rich Context: Related Family Members

Mubarak, the wife of deposed Egyptian President Hosni Mubarak, …



Outline

- Motivation and Definition
- A Skeletal View of a Wikification System
    - o High Level Algorithmic Approach
- Key Challenges



Coffee Break

Recent Advances

- New Tasks, Trends and Applications
- What’s Next?
- Resources, Shared Tasks and Demos

Recent Advances

Improving Wikification by

- Acquiring Rich Knowledge
  - Better Meaning Representation
  - Collaborative Title Collection

- Global Inference Using the Additional Knowledge
  - Joint Mention Extraction and Linking
  - Collective Inference

Semantic Relations and Event Extraction

- Co-occurrence
  - Two pilots had their wedding in **Spain** on 15th, and so they became the first homosexual couple who got married in Spanish troops. The wedding was held in **Sevilla** city hall.
  - The assistant of **Bosnia** Premier Taqik said …two **Democratic Progressive Party** members who held important duties in the central government…

- Part-whole Relation
  - Verizon coverage in **WV** is good along the interstates and in the major cities like Charleston, Clarksburg, **Fairmont**, Morgantown, Huntington, and Parkersburg.-
  - **Manchester** (**New Hampshire**)

Semantic Relations and Event Extraction (Cont')

- Employer/Title
  - **Milton**, the senior **representative** of **Brazil** government
  - **Milton**, the **Governor** of **Pichincha Province**, **Ecuador**

- Affiliation
  - **Bulgarian National Medicines Agency**

- Located Relation
  - **Fine Chemical Plant** in Wuhu City

- Event
  - The leader of **Chilean** Fencing Federation **Ertl** was **elected** as the new **chairman** of this country's **Olympic Committee** tonight.

Another Challenging Example

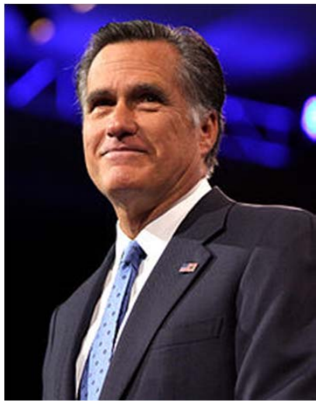
- I am cautiously anticipating the **GOP** nominee in 2012 not to be **Mitt Romney.**
- When **Romney** was the Governor of **Massachusetts**, he helped develop health care law. I appreciate his work.
- I think **Newt** is going to be the last of the "Not **Romneys**".
- **Romney** is the great-great-grandson of a **Mormon pioneer**, from being a Mormon to having taken so many positions in the past that annoy conservatives.
- I don't think **Republican candidates** like **Romney**, **Newt**, and **Johnson** have a real chance for the election.

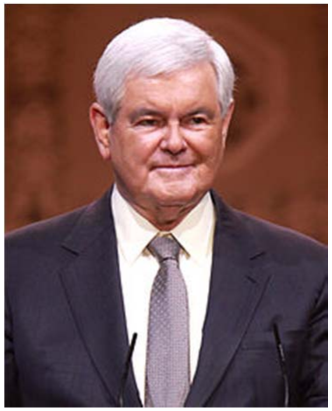
Candidates from Popularity Ranking

- **Mitt Romney**
- George W. Romney
- George Romney (painter)
- Marion G. Romney
- Ann Romney
- G. Scott Romney
- Lenore Romney
- George S. Romney
- Val Romney
- G. Ott Romney

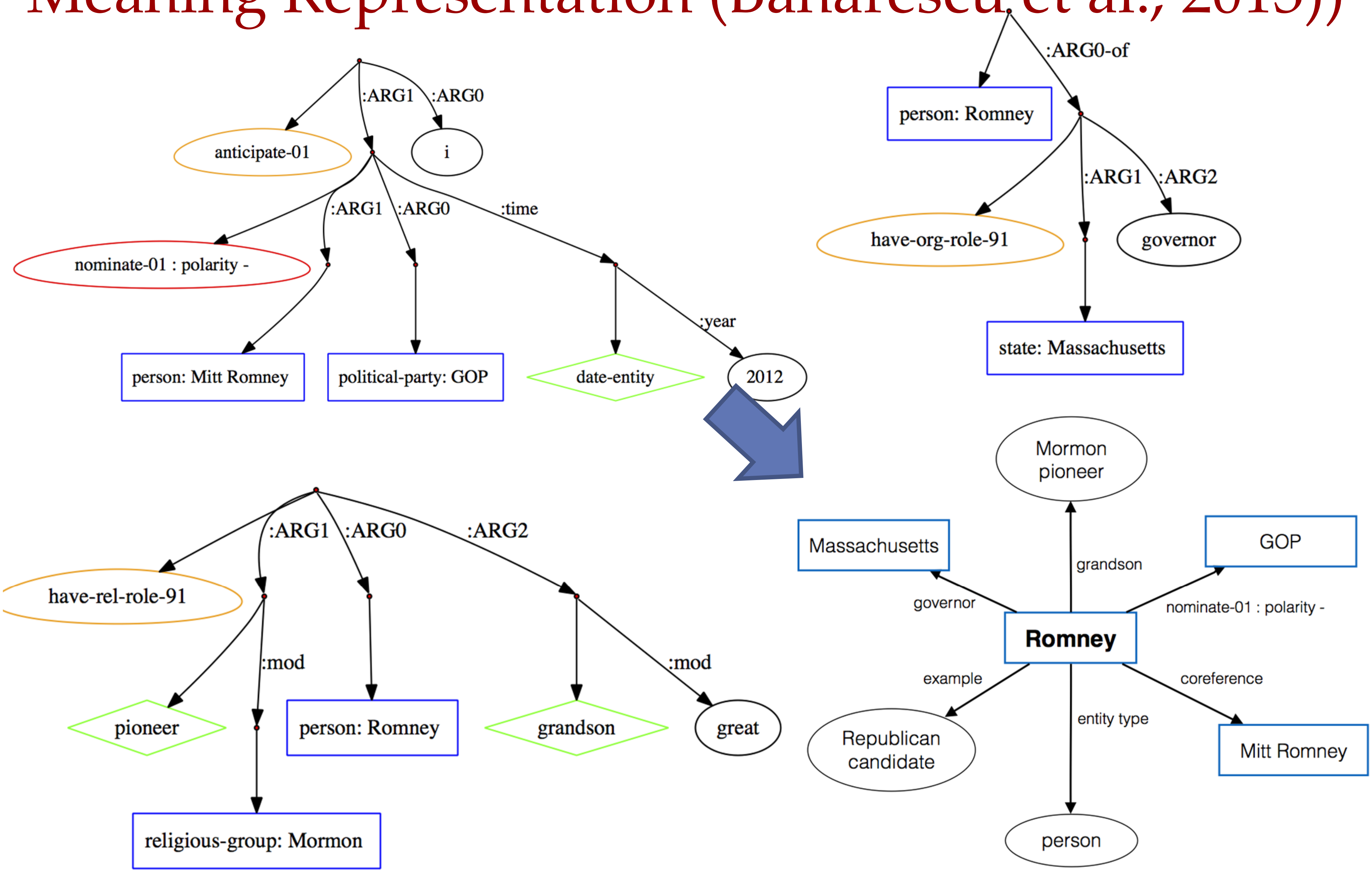
- Newt
- **Newt Gingrich**
- Newt (programming library)
- Eastern newt
- California newt
- Rough-skinned newt
- Alpine newt
- Iberian ribbed newt
- Kaiser's spotted newt
- Newt Allen

- Johnson & Johnson
- Johnson
- Lyndon B. Johnson
- Samuel Johnson
- Andrew Johnson
- B. S. Johnson
- Magic Johnson
- Robert Johnson
- **Gary Johnson**
- Randy Johnson





Better Semantic Representations (e.g., Abstract Meaning Representation (Banarescu et al., 2013))



Acquiring Rich Knowledge from KBs

- Wikipedia (Han and Zhao, 2009)
  - Wikipedia titles and their surface forms
  - Associative relation (internal page links), hierarchical relation and equivalence relation between concepts
  - Polysemy (disambiguation page) and synonymy (redirect page) between key terms
  - Templates (Zheng et al., 2014)
- DBPedia (Zheng et al., 2014)
  - o Rich relational structures and hierarchies, fine-grained types

Matching Rich Semantic Knowledge in KB

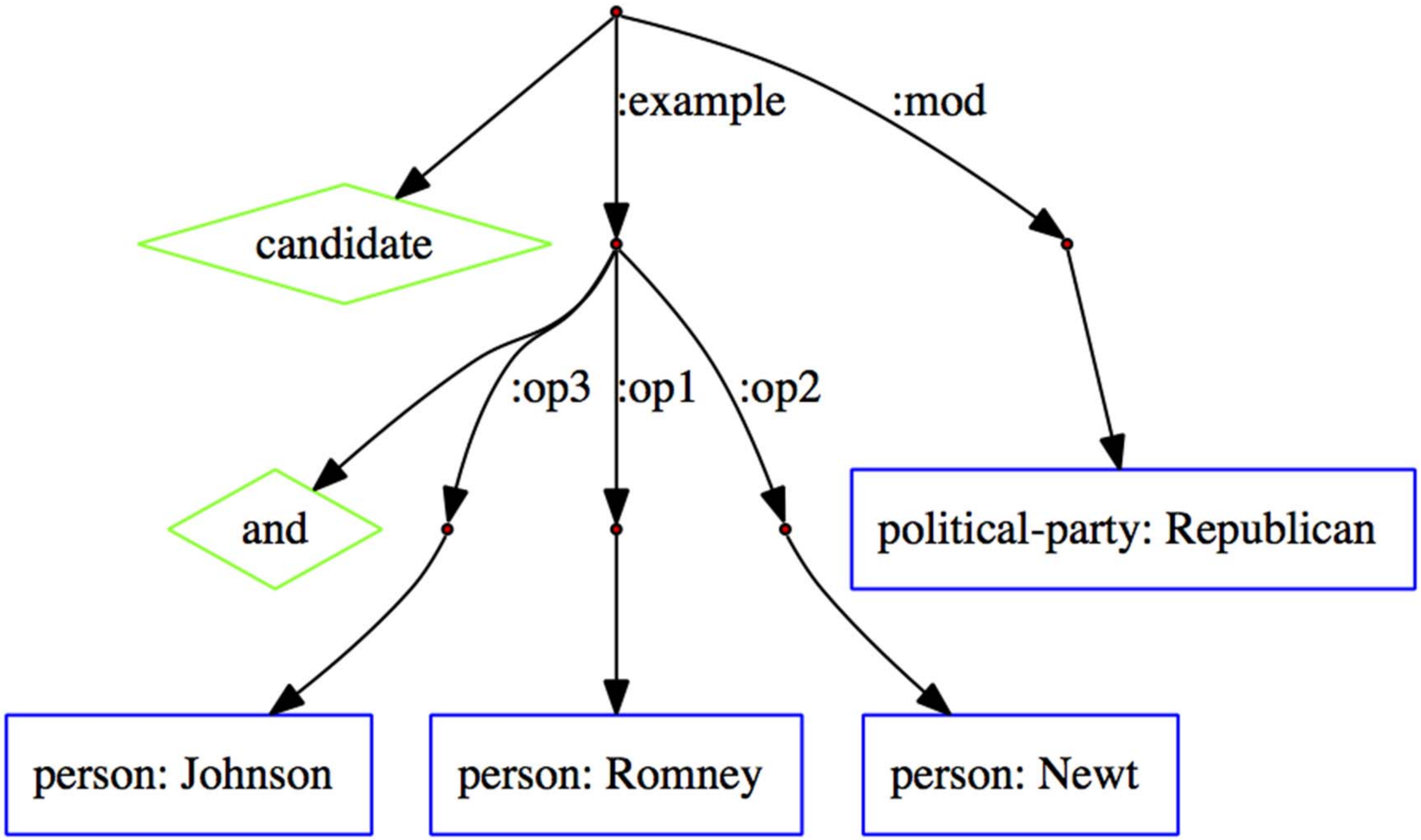
Willard Mitt Romney (born March 12, 1947) is an American businessman who was the ***Republican Party's*** nominee for President of the United States in the 2012 election. Before his presidential bid, he served as the ***70th Governor of Massachusetts*** from 2003 to 2007.

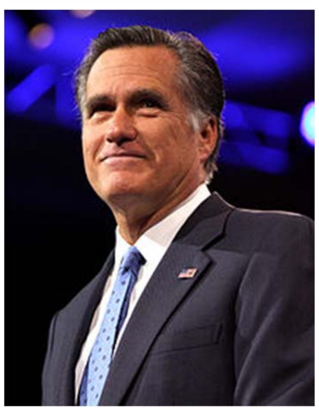
Religion	The Church of Jesus Christ of Latter-day Saints (Mormon)

Political party	Republican

WIKIPEDIA

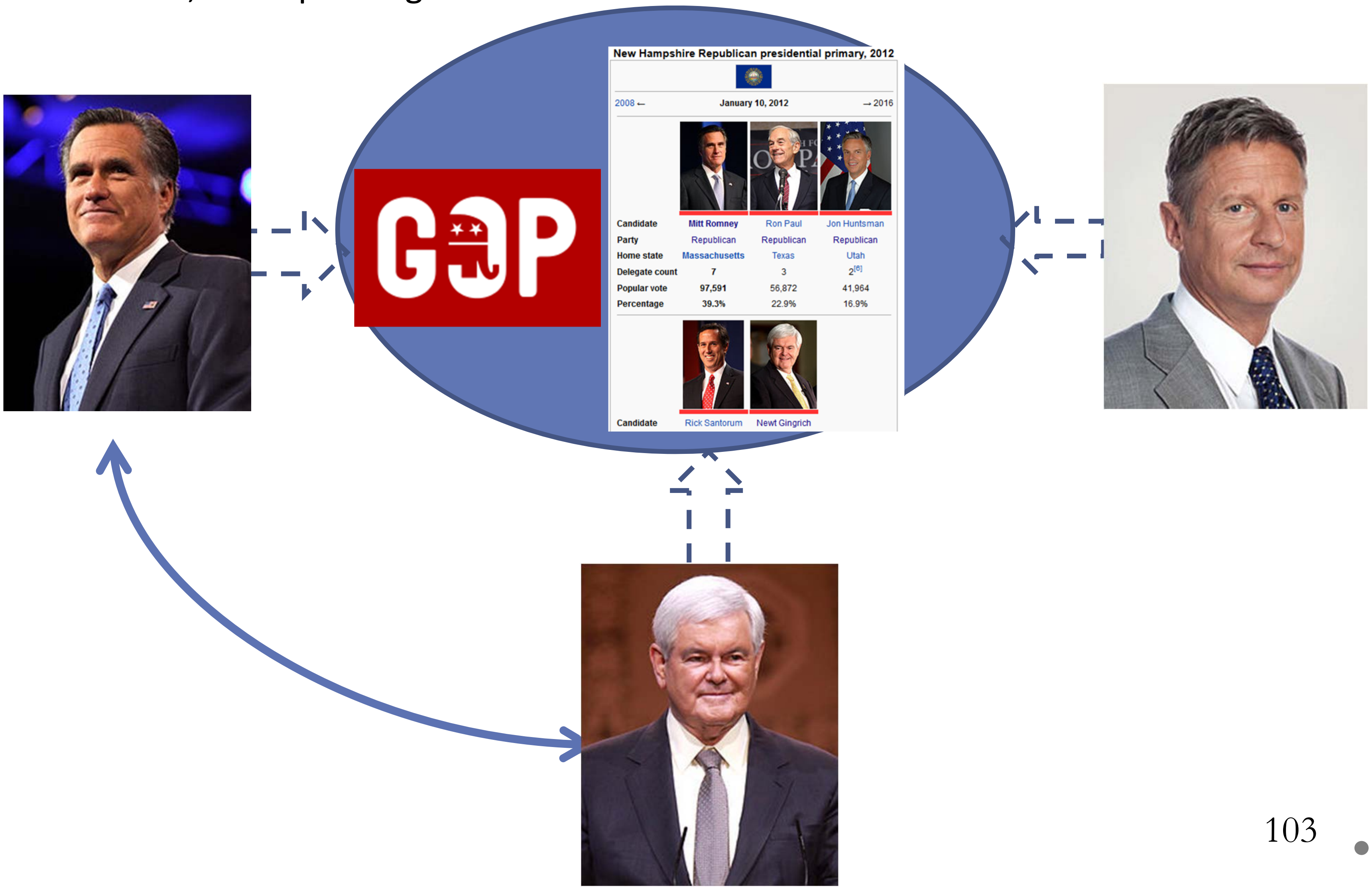
Collaborative Title Collection on Source





Collaborative Title Collection on KB

- Explicit inlinks / outlinks
- Implicit semantic relatedness: overlapped hyperlinks in Wikipedia titles, articles, infoboxes, concept categories and external links



Collaborative Title Collection on KB

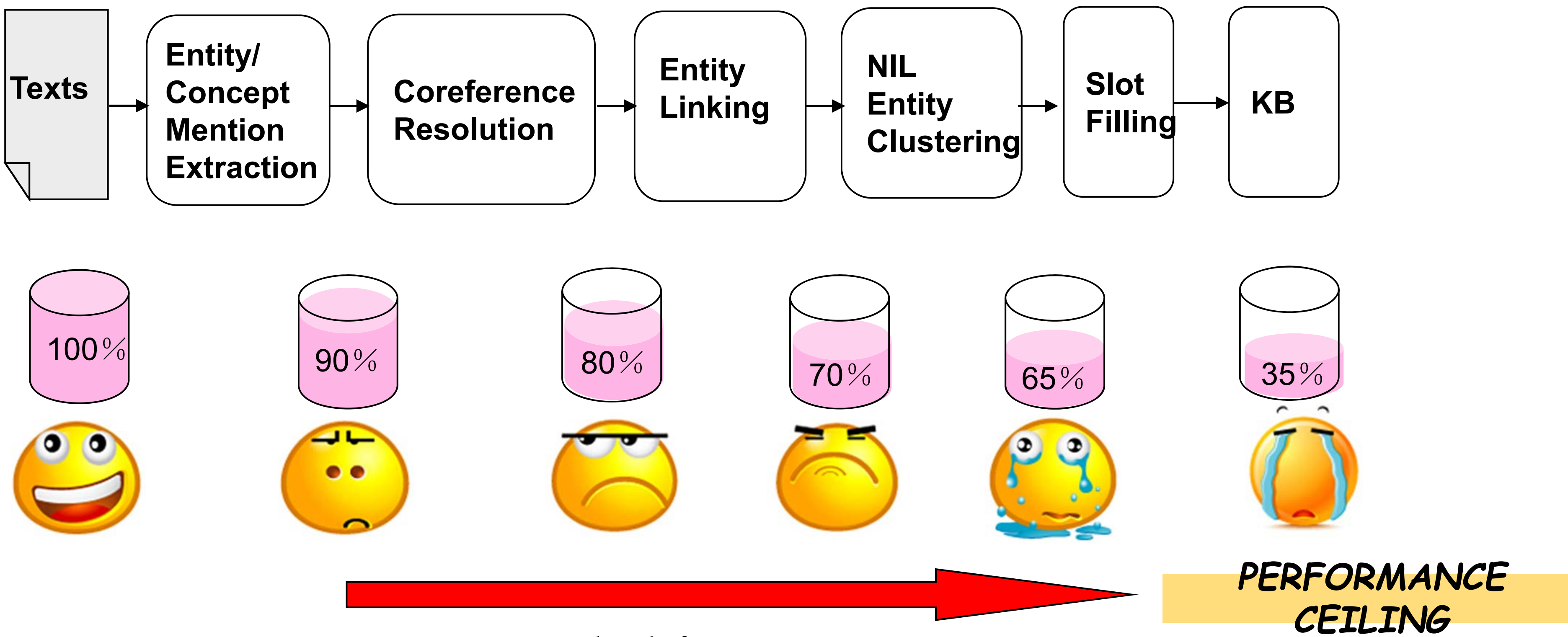
- Go beyond Wikipedia: Exploit rich structures in DBPedia, Freebase, YAGO, Ontologies
- Google Knowledge base: "people also search for"

Recent Advances

Improving Wikification by

- Acquiring Rich Knowledge
  - Better Meaning Representation
  - Collaborative Concept Collection
- Global Inference Using the Additional Knowledge
  - Joint Mention Extraction and Linking
  - Collective Inference

End-to-end Wikification: Traditional Pipeline Approach

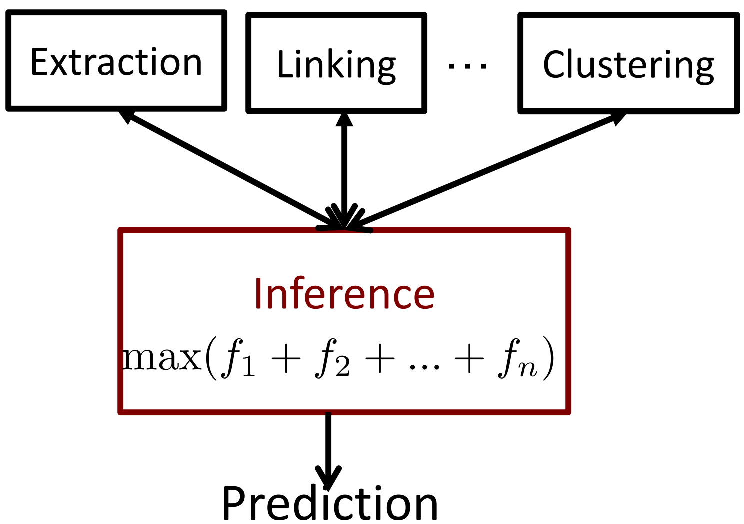


- Errors are compounded from stage to stage
- No interaction between individual predictions
- Incapable of dealing with global dependencies

Solution: Joint Extraction and Linking

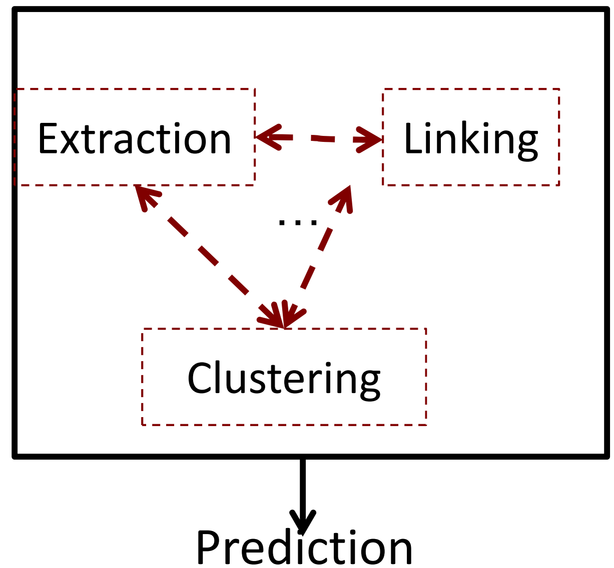
- [Blue Cross]$_{ORG}$ and [Blue Shield of Alabama]$_{ORG}$
- [Blue Cross and Blue Shield of Alabama]$_{ORG}$ → BCBS of Alabama

Joint Inference



- Constrained Conditional Models, ILP [Roth2004,Punyakanok2005,Roth2007, Chang2012, Yang2013]
- Re-ranking [**Sil2013**,Ji2005,McClosky2011]
- Dual decomposition [Rush2010]

Joint Modeling



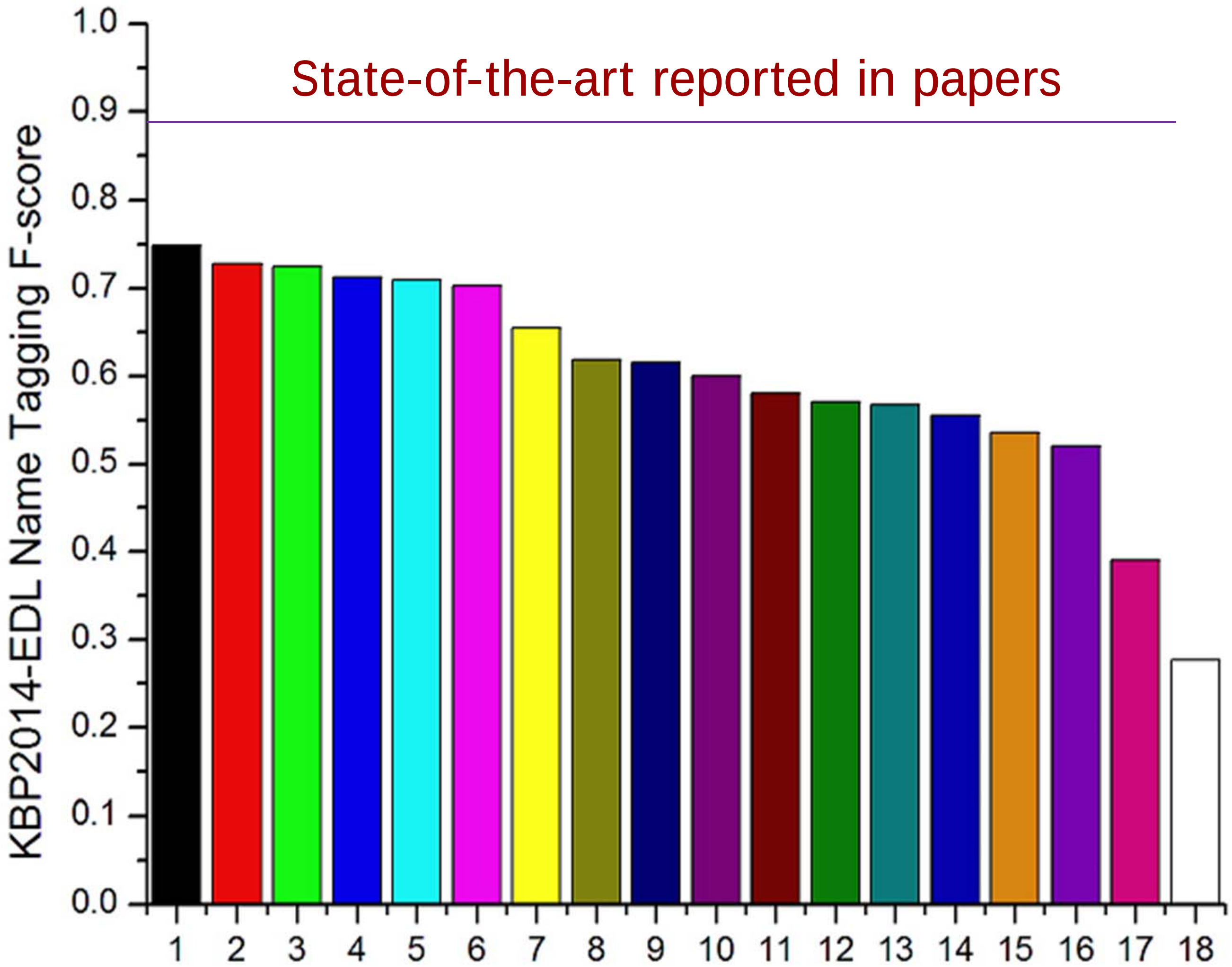
- Probabilistic Graphical Models [Sutton2004,Wick2012,**Wick2013**, Singh2013]
- Markov logic networks [Poon2007,Poon2010,Kiddon2012]
- Linking for extraction [**Meij2012,Guo2013,Fahrni2013,Huang2014**]

Name Tagging: "Old" Milestones

Year	Tasks & Resources	Methods	F-Measure	Example References
1966	-	First person name tagger with punch card 30+ decision tree type rules	-	(Borkowski et al., 1966)
1998	MUC-6	MaxEnt with diverse levels of linguistic features	97.12%	(Borthwick and Grishman, 1998)
2003	CONLL	System combination; Sequential labeling with Conditional Random Fields	89%	(Florian et al., 2003; McCallum et al., 2003; Finkel et al., 2005)
2006	ACE	Diverse levels of linguistic features, Re-ranking, joint inference	~89%	(Florian et al., 2006; Ji and Grishman, 2006)

- Our progress compared to 1966:
  - More data, a few more features and more fancy learning algorithms
- Not much active work after ACE because we tend to believe it's a solved problem…

The end of extreme happiness is sadness…



Cross-genre Name Tagging

- Experiments on ACE2005 data

Training Domain	Test Domain	Prec	Recall	F_1
nwire	nwire	89.6	89.1	89.3
	bn	92.5	40.3	56.2
	bc	87.3	84.4	85.8
	cts	71.4	85.6	77.8
	wl	77.2	73.2	75.2
	un	77.7	59.4	67.3
bn	nwire	84.8	77.7	81.2
	bn	89.6	85.1	87.3
	bc	87.1	81.6	84.3
	cts	85.2	83.1	84.2
	wl	75.3	69.8	72.5
	un	75.5	57.8	65.5
bc	bc	88.4	82.5	85.4
cts	cts	91.9	77.9	84.3
wl	wl	75.4	67.2	71.1
un	un	84.1	51.72	64.0

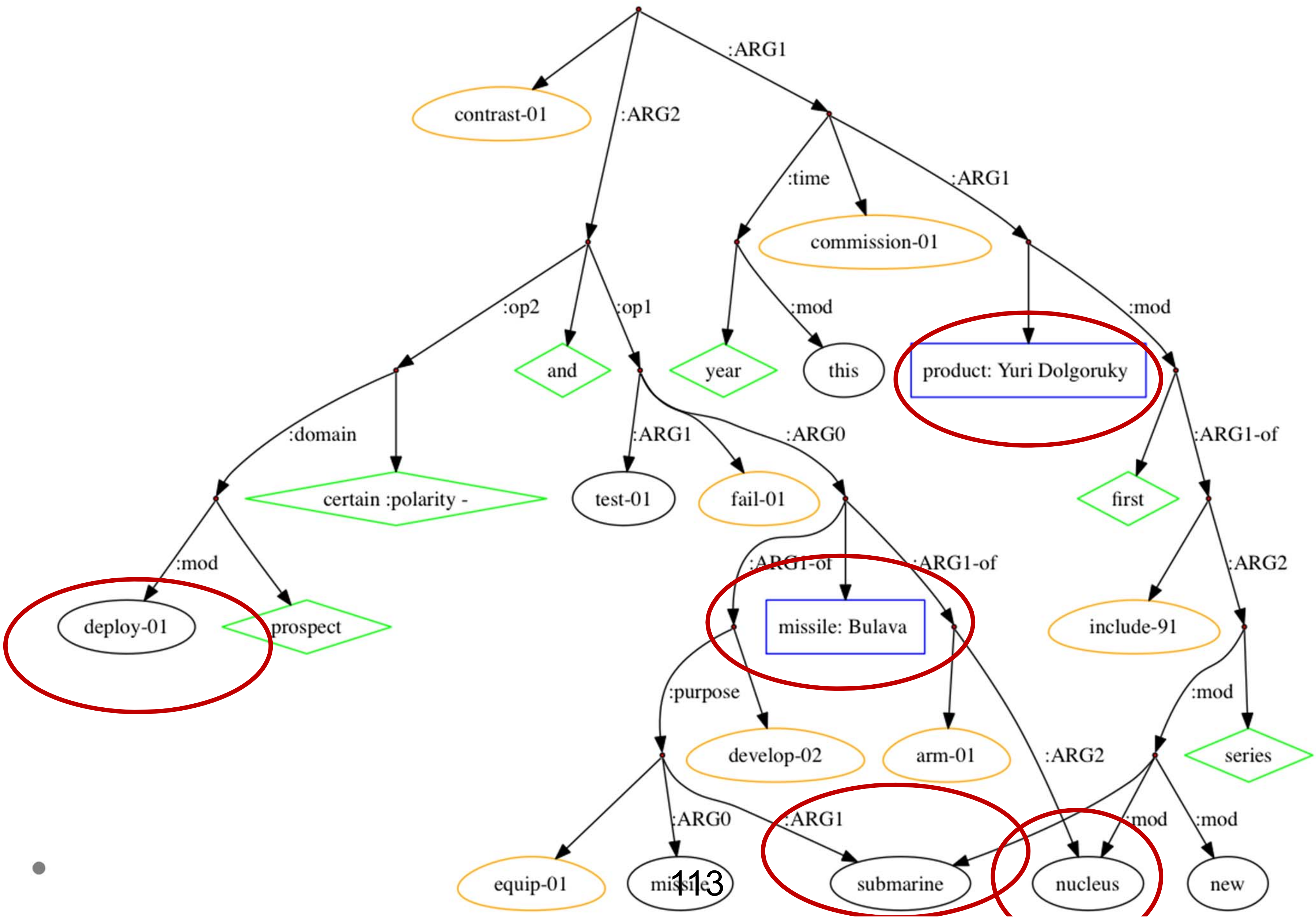
What's Wrong?

- Name taggers are getting old (trained from 2003 news & test on 2012 news)
- Genre adaptation (informal contexts, posters)
- Revisit the definition of name mention – extraction for linking
- Old unsolved problems
  - Identification: "Asian Pulp and Paper Joint Stock Company , Lt. of Singapore"
  - Classification: "**FAW** has also utilized the capital market to directly finance,…" (*FAW = First Automotive Works*)
- Potential Solutions for Quality
  - Word clustering, Lexical Knowledge Discovery (Brown, 1992; Ratinov and Roth, 2009; Ji and Lin, 2010)
  - Feedback from Linking, Relation, Event (Sil and Yates, 2013; Li and Ji, 2014)

Joint Extraction and Linking

The Yuri dolgoruky is the first in a series of new nuclear submarines to be commissioned this year but the bulava nuclear-armed missile developed to equip the submarine has failed tests and the deployment prospects are uncertain.

Joint Extraction and Linking



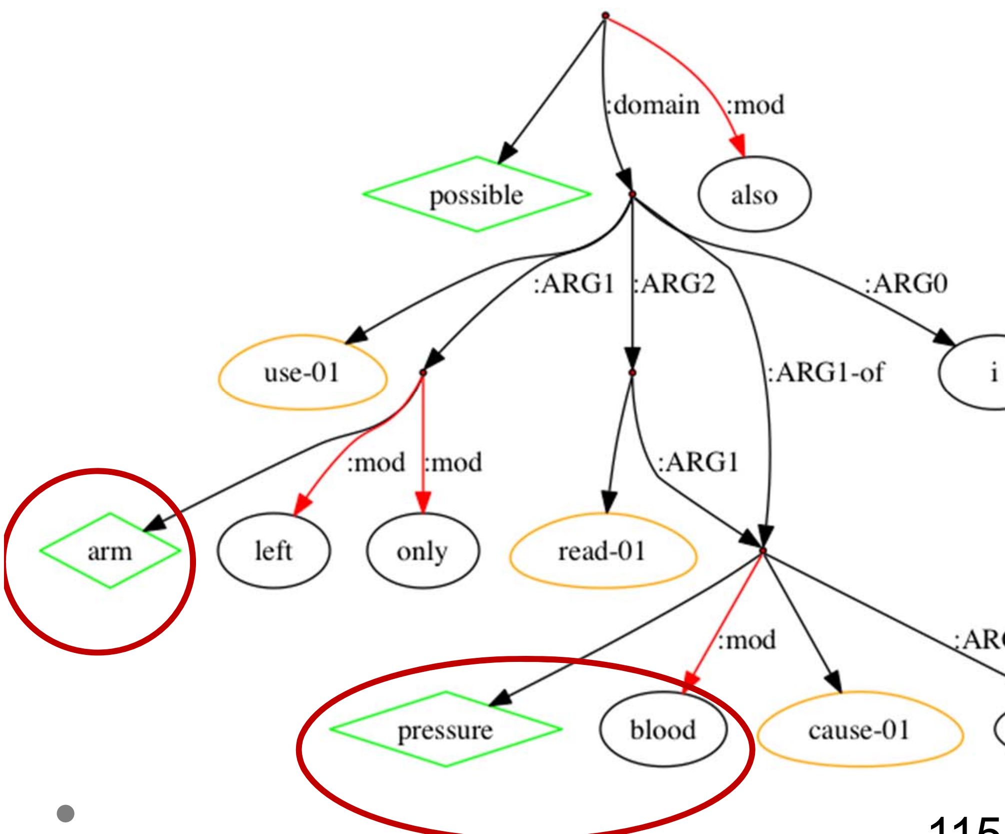
Joint Extraction and Linking

The Yuri dolgoruky is the first in a series of new nuclear submarines to be commissioned this year but the bulava nuclear-armed missile developed to equip the submarine has failed tests and the deployment prospects are uncertain.

Acquiring Rich Knowledge

I had it done three years ago at **Johns Hopkins**.

Also, because of a lumpectomy I can only use my left **arm** for **B. P.** readings.



Global Interaction Feature: Distinct-Links-Per-Mention (Sil and Yates, 2013)

- **Objective:** Penalize over-segmented phrases
  - **Example:**

[Home] [Depot]

Home_DepotHome_Depot

Indicates over-segmentation

#distinct Entities = 1
#mentions = 2
=> Feature Value= **0.5**

[Home Depot]

Home_Depot

Indicates **correct segmentation**

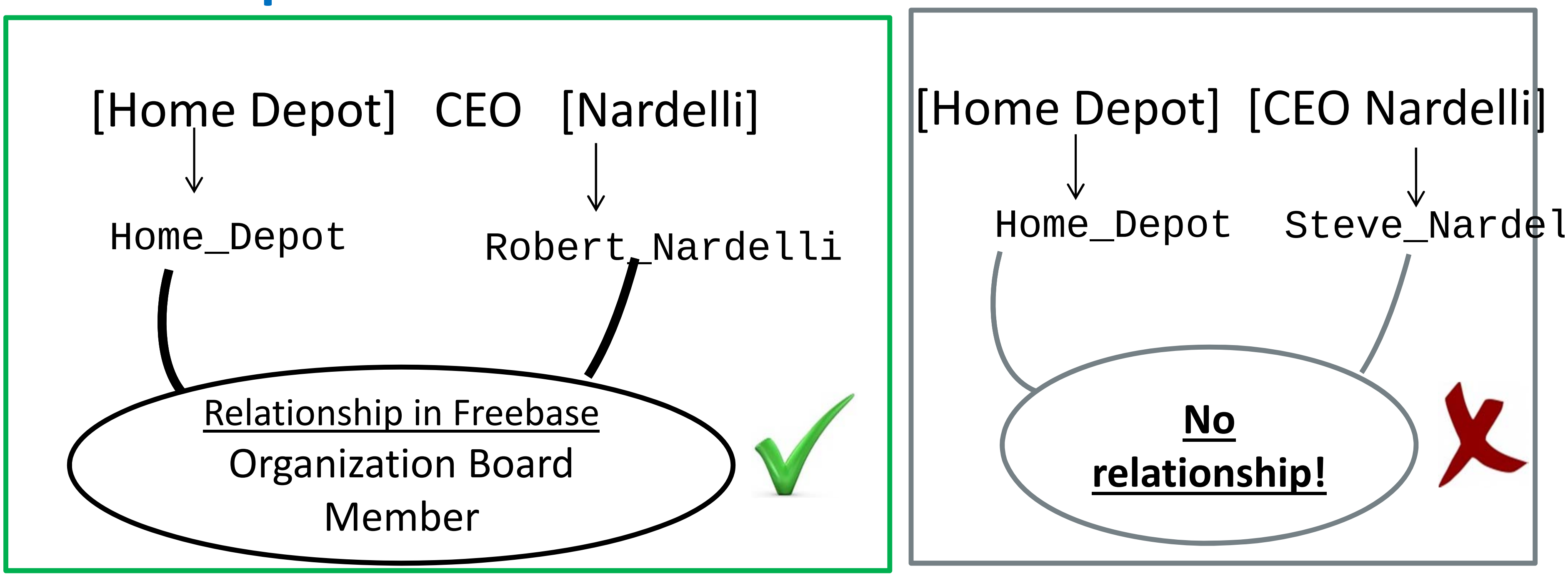
#distinct Entities = 1
#mentions = 1

=> Feature Value= 1

Global Interaction Feature: Binary Relation Count (Sil and Yates, 2013)

- Use Binary Relations between entities in Freebase

- **Example:**



Indicates: Under-segmentation

Global Interaction Feature: Entity Type PMI (Sil and Yates, 2013)

- Find patterns of entities appearing close to each other

$$PMI(T(e_1), T(e_2)) = \frac{\sum_{(e,e')\in T} \mathbf{1}[T(e_1) = T(e) \wedge T(e_2) = T(e')]}{\sum_{e\in T} \mathbf{1}[T(e_1) = T(e)] \times \sum_{e\in T} \mathbf{1}[T(e_2) = T(e)]}$$

- **Example:**

[Home Depot] CEO [Nardelli]

Home_Depot — **Type:** Organization

Robert_Nardelli — **Type:** Org_Leader

[Home Depot] [CEO Nardelli]

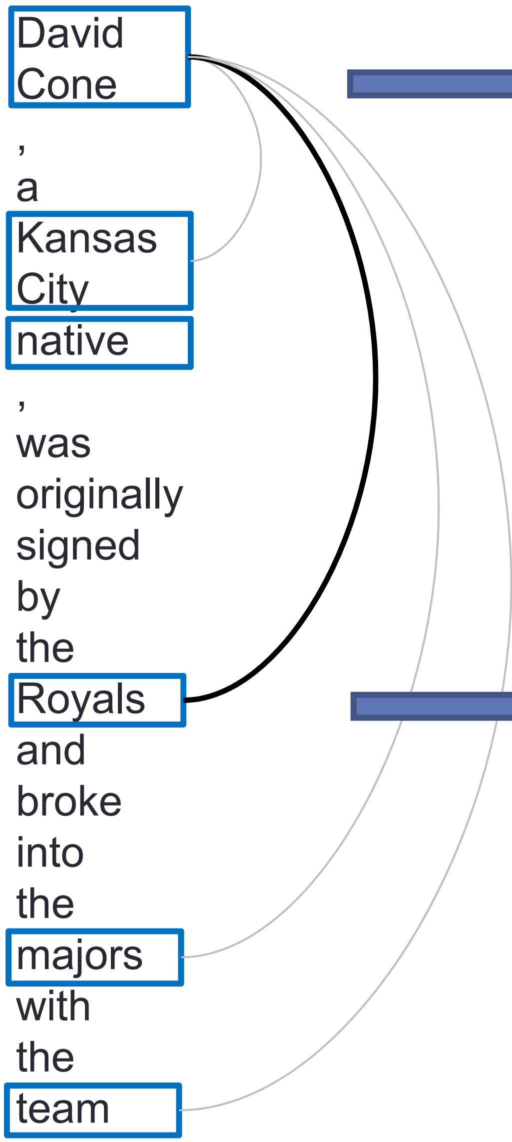
Home_Depot — **Type:** Organization

Steve_Nardel — **Type:** Music_Artist

Joint Extraction and Linking

- Some recent work (Sil and Yates, 2013; Meij et al., 2012; Guo et al., 2013; Huang et al., 2014b) proved extraction and linking can mutually enhance each other
  - **Bosch** will provide the rear axle. → Robert Bosch Tool Corporation → ORG
  - Parker was 15 for 21 from the field, putting up a season high while scoring nine of **San Antonio**'s final 10 points in regulation → San Antonio Spurs → ORG
- IBM (Sil and Florian, 2014), MSIIPL THU (Zhao et al., 2014), SemLinker (Meurs et al., 2014), UBC (Barrena et al., 2014) and RPI (Hong et al., 2014) used the properties in external KBs such as DBPedia as feedback to refine the identification and classification of name mentions.
  - RPI system successfully corrected 11.26% wrong mentions
- HITS team (Judea et al., 2014) proposed a joint approach that simultaneously solves extraction, linking and clustering using Markov Logic Networks
- Document Linking → Event Extraction (Ji and Grishman, 2008)
- Entity Linking → Relation Extraction (Chan and Roth, 2010)
- Toward more interactions and joint inferences between tasks → Marry EDL and SF in KBP2015

Entity Linking to Improve Relation Extraction (Chan and Roth, 2010)



David Brian Cone (born January 2, 1963) is a former Major League Baseball pitcher. He compiled an 8–3 postseason record over 21 postseason starts and was a part of five World Series championship teams (1992 with the Toronto Blue Jays and 1996, 1998, 1999 & 2000 with the New York Yankees). He had a career postseason ERA of 3.80. He is the subject of the book *A Pitcher's Story: Innings With David Cone* by Roger Angell. Fans of David are known as "Cone-Heads."
Cone lives in Stamford, Connecticut, and is formerly a color commentator for the Yankees on the YES Network.[1]

Contents

[hide]

WIKIPEDIA
The Free Encyclopedia

Partly because of the resulting lack of leadership, after the 1994 season the Royals decided to reduce payroll by trading pitcher David Cone and outfielder Brian McRae, then continued their salary dump in the 1995 season. In fact, the team payroll, which was always among the league's highest, was sliced in half from $40.5 million in 1994 (fourth-highest in the major leagues) to $18.5 million in 1996 (second-lowest in the major leagues)

WIKIPEDIA
The Free Encyclopedia

120

Recent Advances

Improving Wikification by

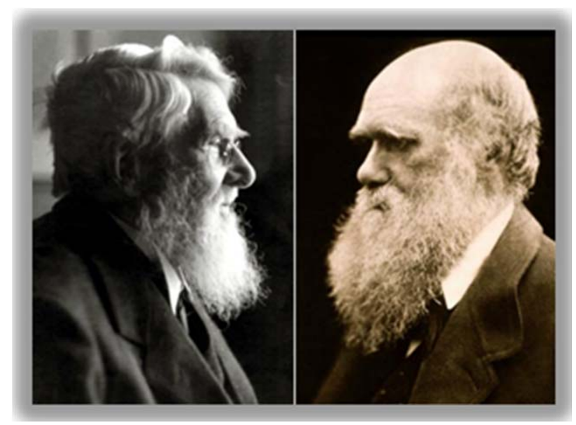
- Acquiring Rich Knowledge
  - Better Meaning Representation
  - Collaborative Concept Collection
- Global Inference Using the Additional Knowledge
  - Joint Concept Extraction and Linking
  - Collective Inference

From Non-collective to Collective

- Intuition: Promote semantically coherent pairs of titles
  - (1) Enrich text with external KB knowledge
    - Use graph metrics (as described earlier)
  - (2) Enrich text with (some) gold titles
    - Use graph propagation algorithms
  - (3) Enrich Text with (local) relational information
    - Use global inference methods

Collaborative Learning

Collective Animal Behavior



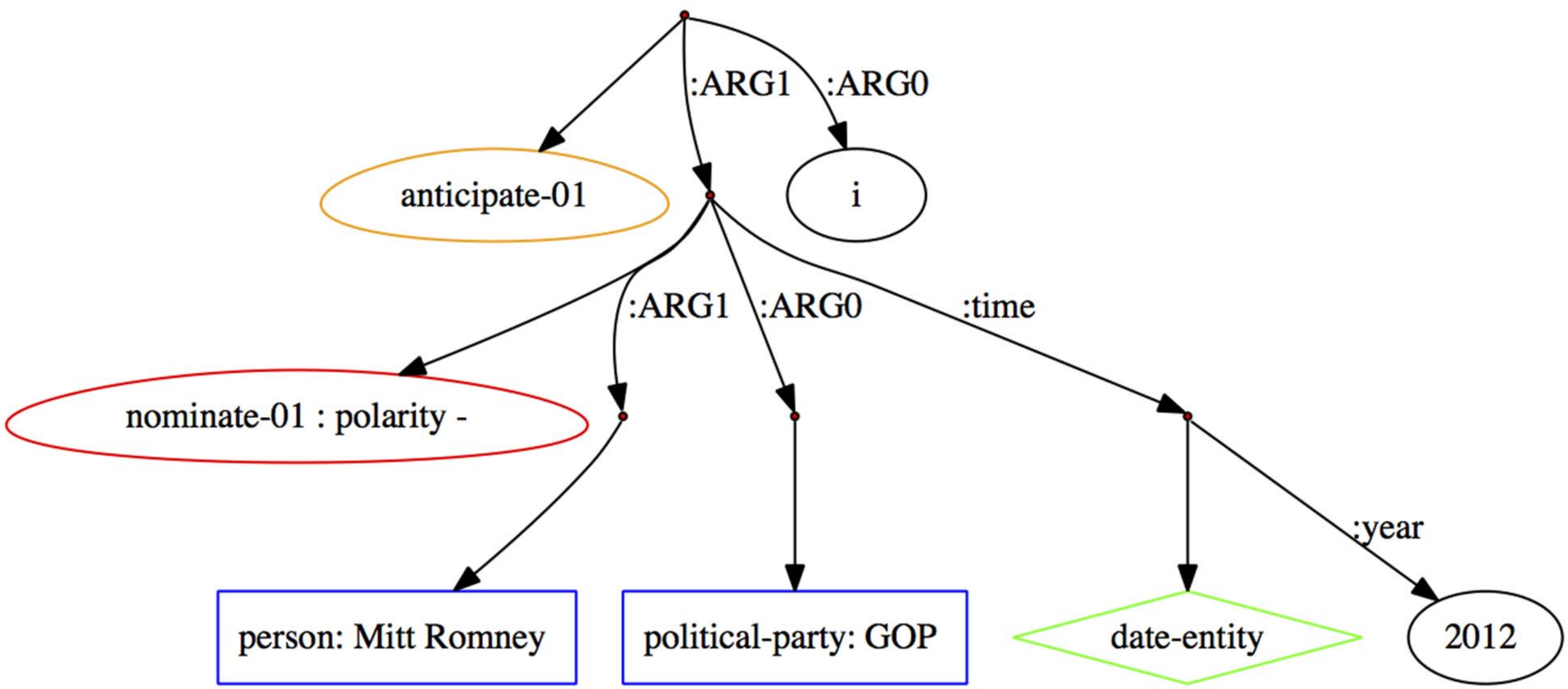
Great Minds Think Alike

(1) Collective Inference: Basic Idea

- Construct a Knowledge Graph from Source
- Construct a Knowledge Graph from KBs
- Each Knowledge Graph contains a thematically homogeneous coherent story/context
- Semantic Matches of Knowledge Graphs using Graph based Measures to Match Three Criteria:
  - Ideally we want to align two graphs directly (Yan and Han, 2002), current simplified solutions→
  - **Similarity:** The mention and the concept should have high similarity
  - **Salience:** The concept should be salient and popular in KB
  - **Coherence:** The concept and its collaborators decided by the mention's collaborators should be strongly connected in KB

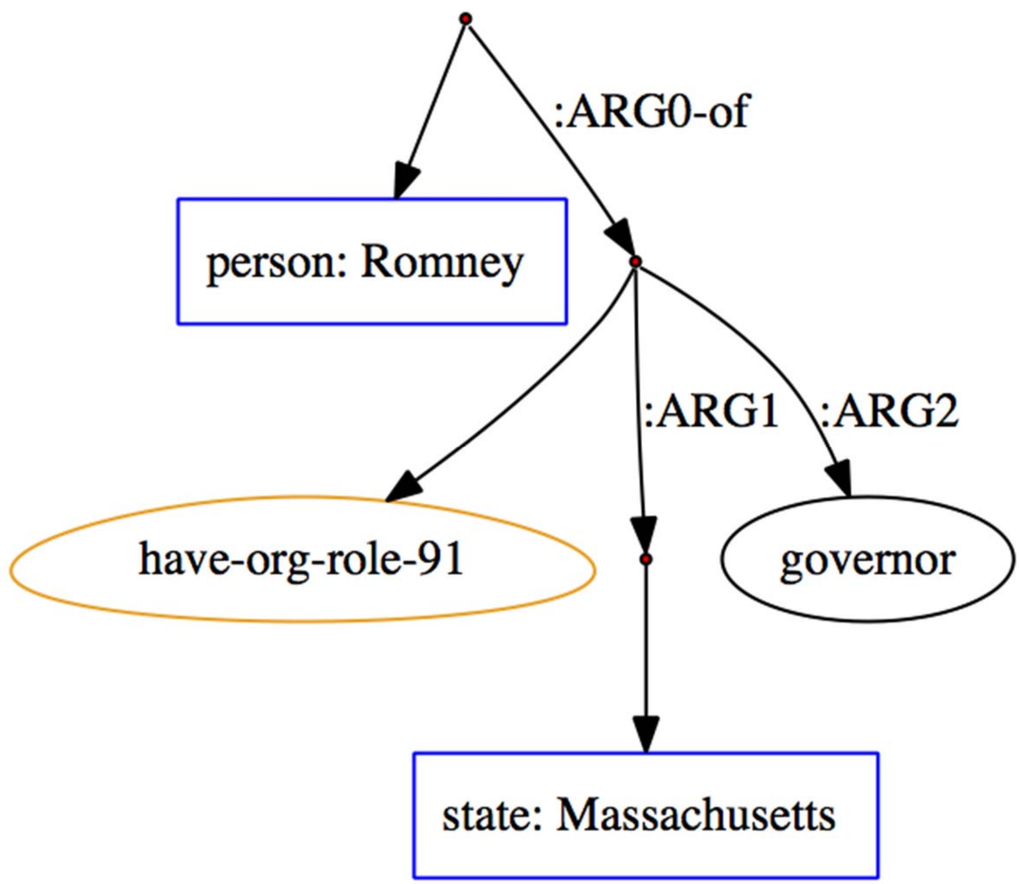
Abstract Meaning Representation

- I am cautiously anticipating the **GOP** nominee in 2012 not to be **Mitt Romney.**



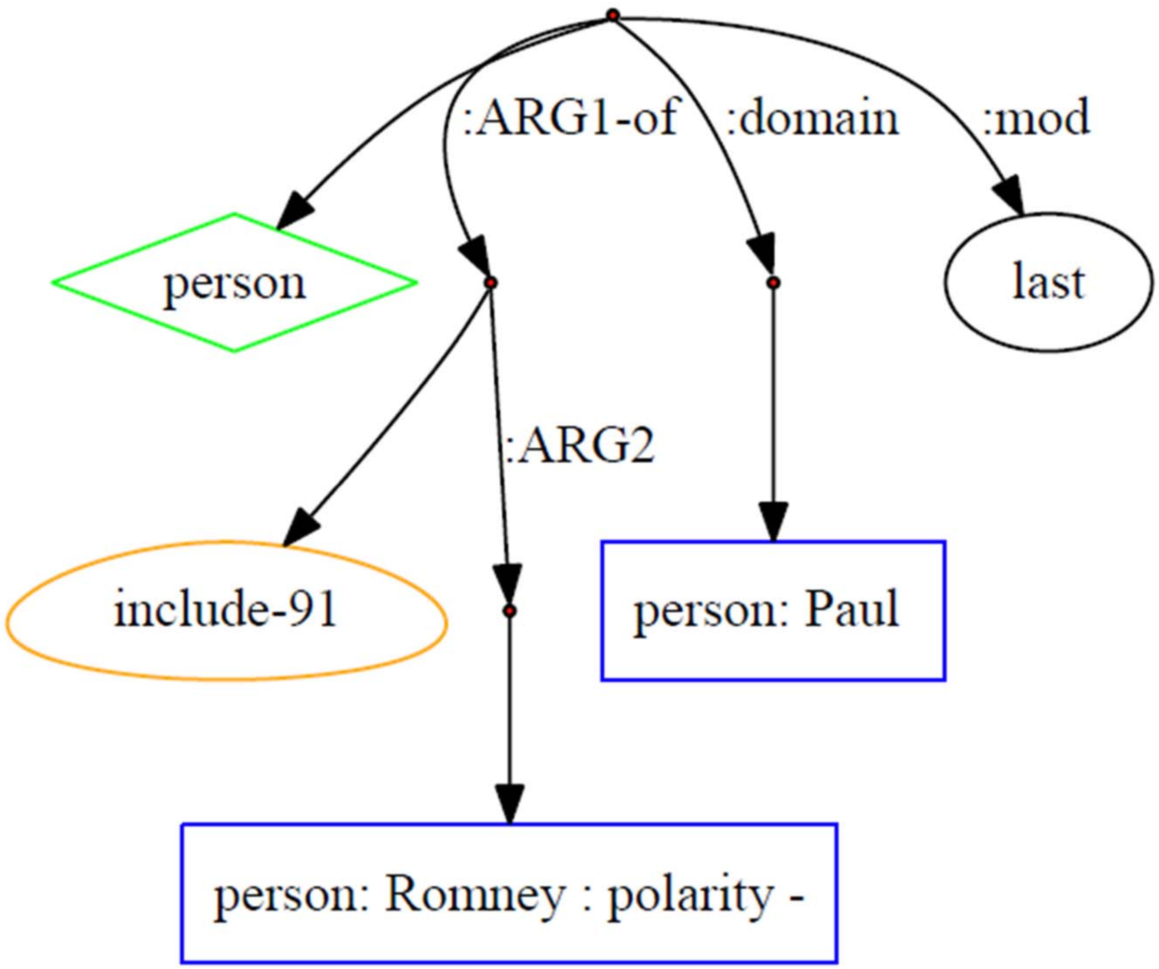
Abstract Meaning Representation

- When **Romney** was the Governor of **Massachusetts** …



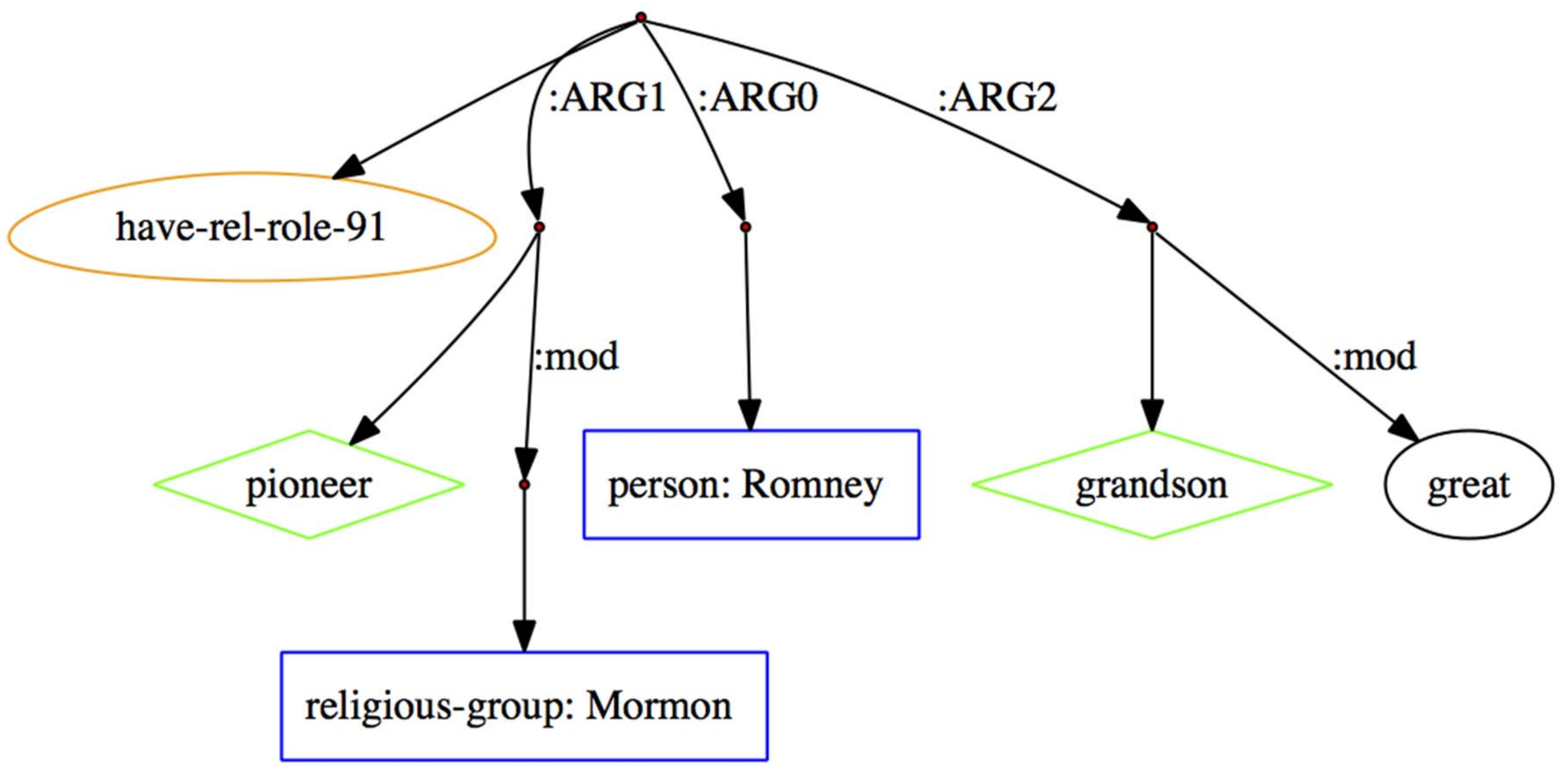
Abstract Meaning Representation

- I think **Paul** is going to be the last of the "Not **Romneys**".



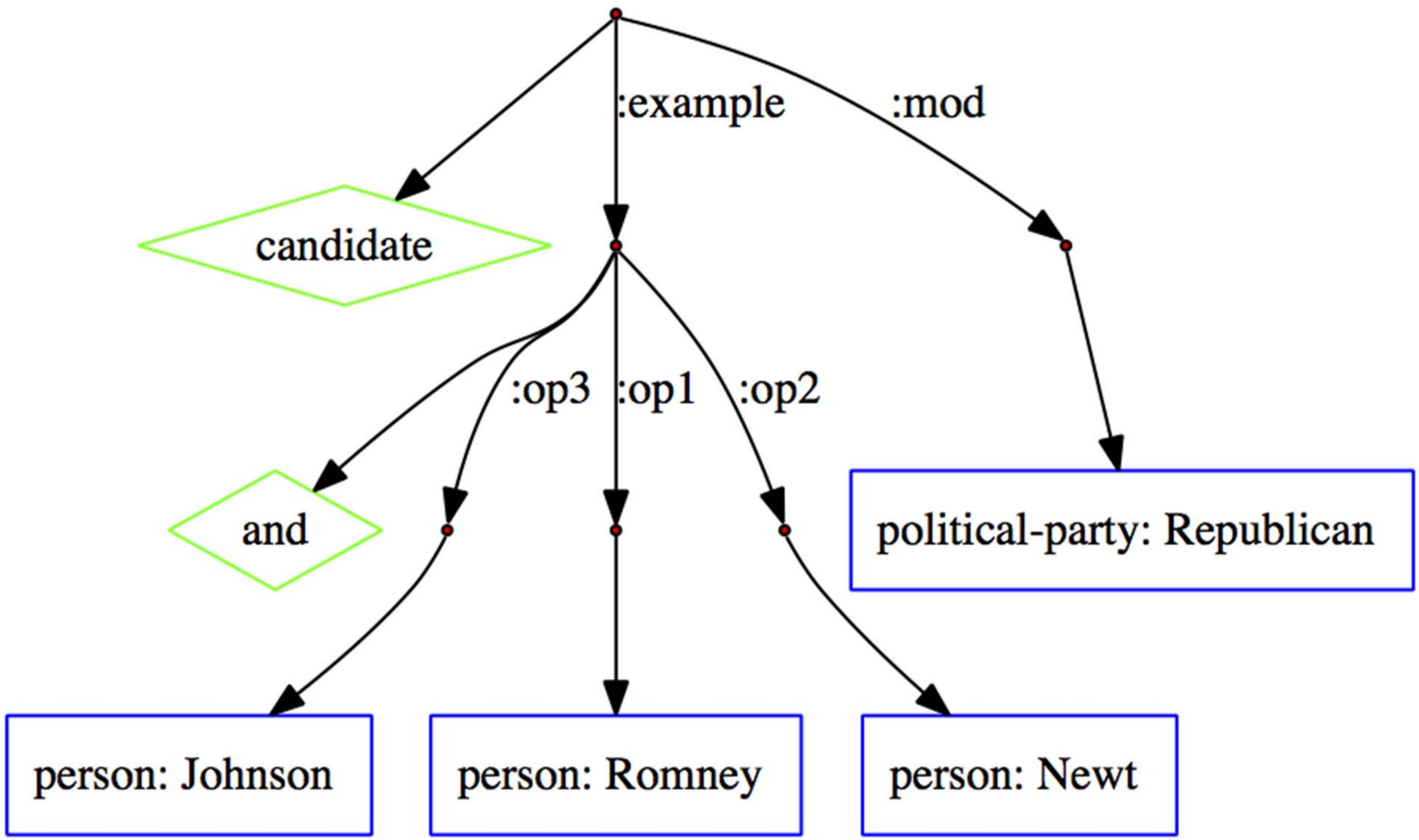
Abstract Meaning Representation

- **Romney** is the great-great-grandson of a **Mormon pioneer**.



Abstract Meaning Representation

- **Republican candidates** like **Romney**, **Newt**, and **Johnson** ...



Basic Idea: Collective Inference

- Construct a Knowledge Graph from Source
- Construct a Knowledge Graph from KBs
- Each Knowledge Graph contains a thematically homogeneous coherent story/context
- Semantic Matches of Knowledge Graphs using Graph based Measures to Match Three Criteria:
  - Ideally we want to align two graphs directly (Yan and Han, 2002), current simplified solutions→
  - **Similarity:** The mention and the concept should have high similarity
  - **Salience:** The concept should be salient and popular in KB
  - **Coherence:** The concept and its collaborators decided by the mention's collaborators should be strongly connected in KB

Constructing Knowledge Network from Source: Entity Nodes

- 8 main types:Person, Organization, Location, Facility, Event, Product, Publication, Natural object, Other
- Hundreds of fine-grained subtypes
  - e.g., Organization subtypes: company, government organization, military, criminal organization, political party, school, university, research institute, team and league
- Use AMR entity node annotation for Entity Linking
  - Name Expansion
  - Name Classification

Constructing Knowledge Network from Source: Roles

- Basic idea: use all other nodes which played the following roles in the same events as neighbors for the target entity mention
- Propbank/Nombank/Ontonotes Type Core Roles
  - Did Palin apologize to Giffords? He needs to conduct a beer summit between Palin and NBC.
  - Construct "Giffords" and "NBC" as neighbors for "Palin" because they are core roles involved in the "apologize" and "summit" events
- Special Roles
  - "ARG2" of "have-org-role-9" which indicates the title of the office held, such as President and Governor
  - "ARG2" and "ARG3" of "have-rel-role-91" are used to describe two entities with the same type, such as family and friendship relations
- Non-core Roles
  - domain, mod, cause, concession, condition, consist-of, extent, part, purpose, degree, manner, medium, instrument, ord, poss, quant, subevent, subset, topic

Constructing Knowledge Network from Source: Frames & Senses

Ok, my answer is no one and **Obama** wins the **GE**.
I think **Romney** wins big today and obviously stays in.
Santorum gets enough of a boost to do the **Huckabee** hangs around.
I think **Gingrich**'s sole win in **GA** is enough to hang it up and go back to making millions in the private sector.
I think **Mitt** drops out...
The only one with any reason to will be **Newt**, but I don't think that he will.

<inventory lemma="win-v">
<sense group="1" n="1" name="beat, prevail or triumph" type="">
<commentary>
NP WIN PP
NP WIN NP [competition,activity,event]
</commentary>

General Electric ***United States presidential election, 2012***

Constructing Knowledge Network from Source: Time & Location

- AMR Event Time
  - time, year, month, day

Billy moved to the US in **1976**.

KB (Billy Thorpe): **Thorpe** also performed as a solo artist; he relocated to the United States from **1976** to 1996.

- AMR Event Location
  - Source, destination, path, location

I-55 will be closed in both directions between **Carondelet** and the 4500 block

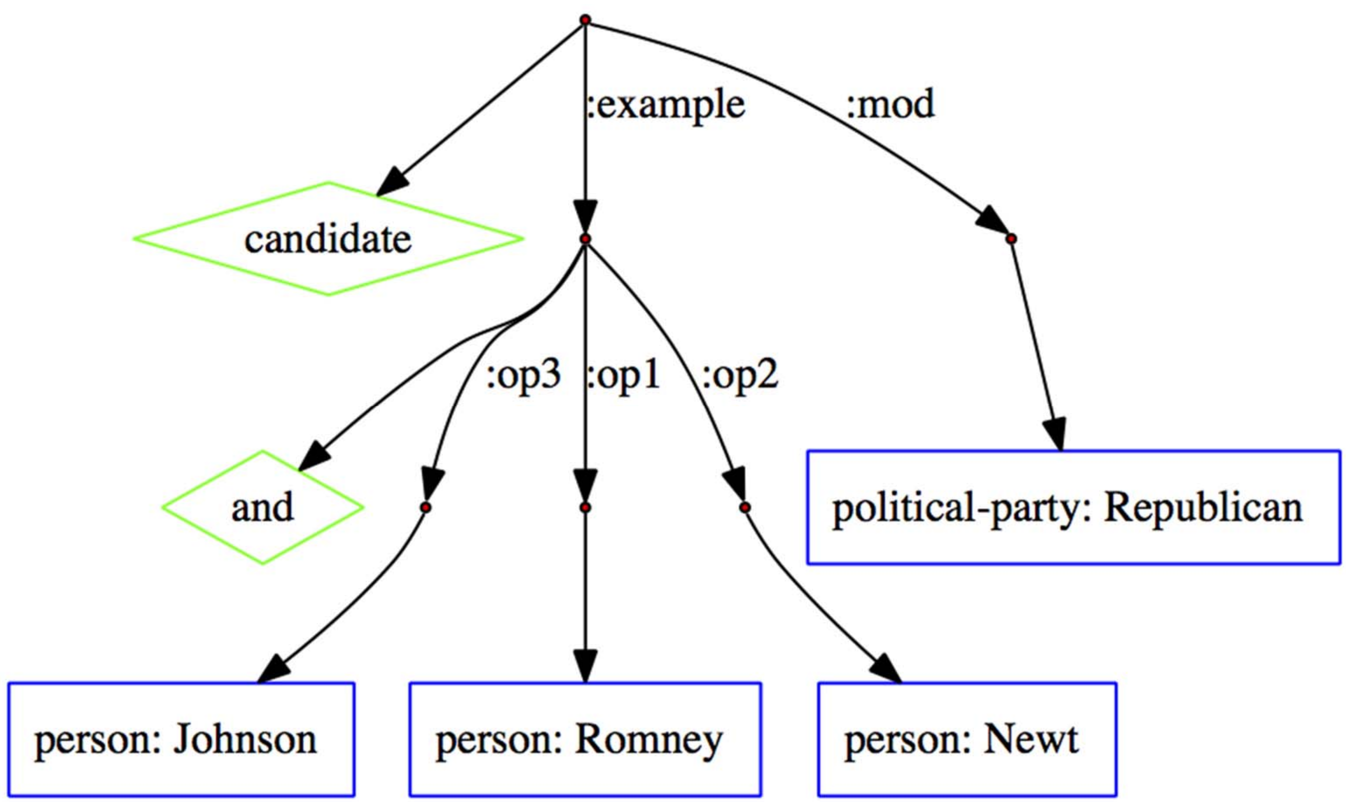
Carondelet is a neighborhood in the extreme southeastern portion of St. Louis, Missouri.

Interstate 55

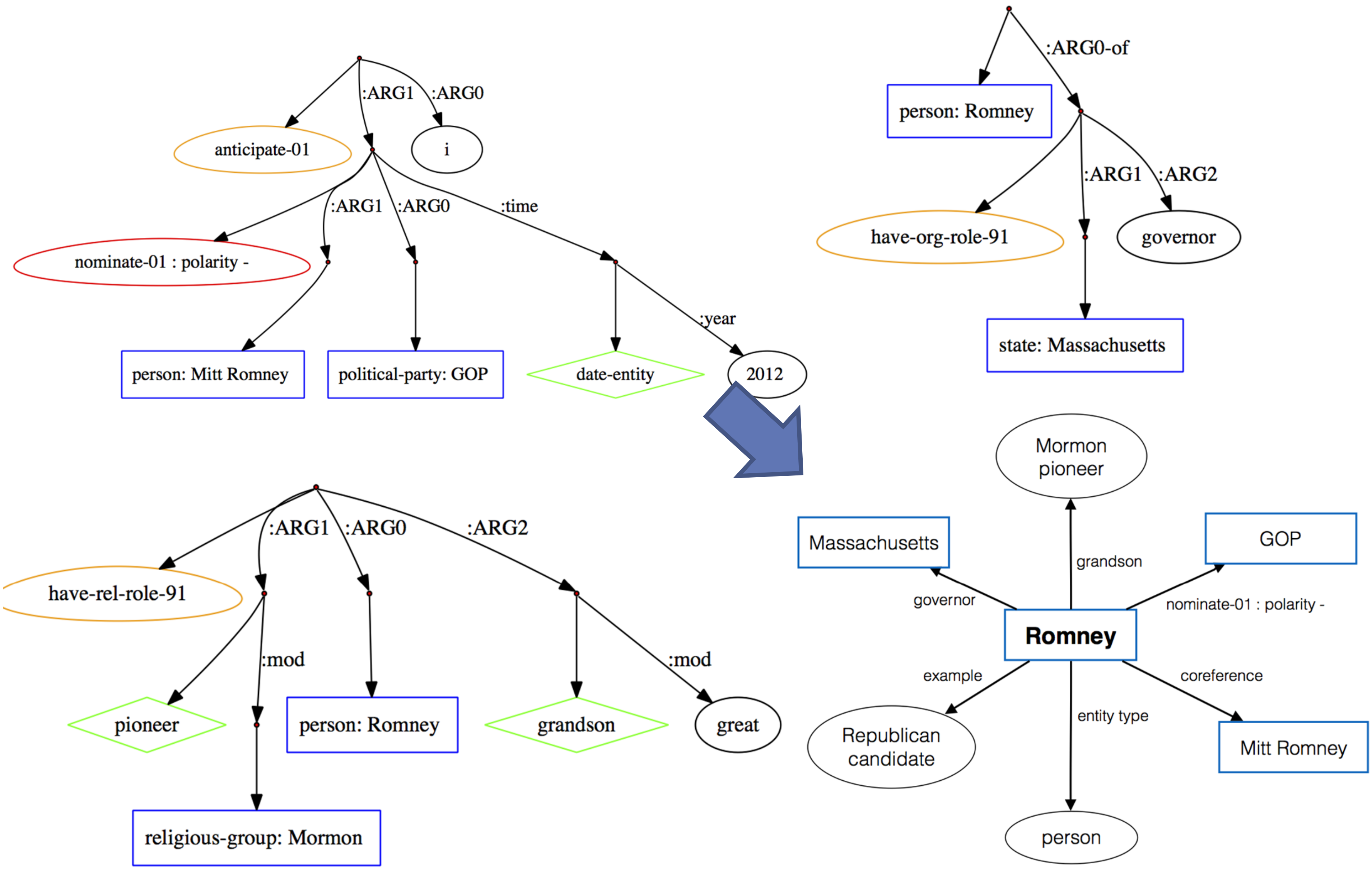
KB: Interstate 55 in Missouri

Collective Inference based on AMR

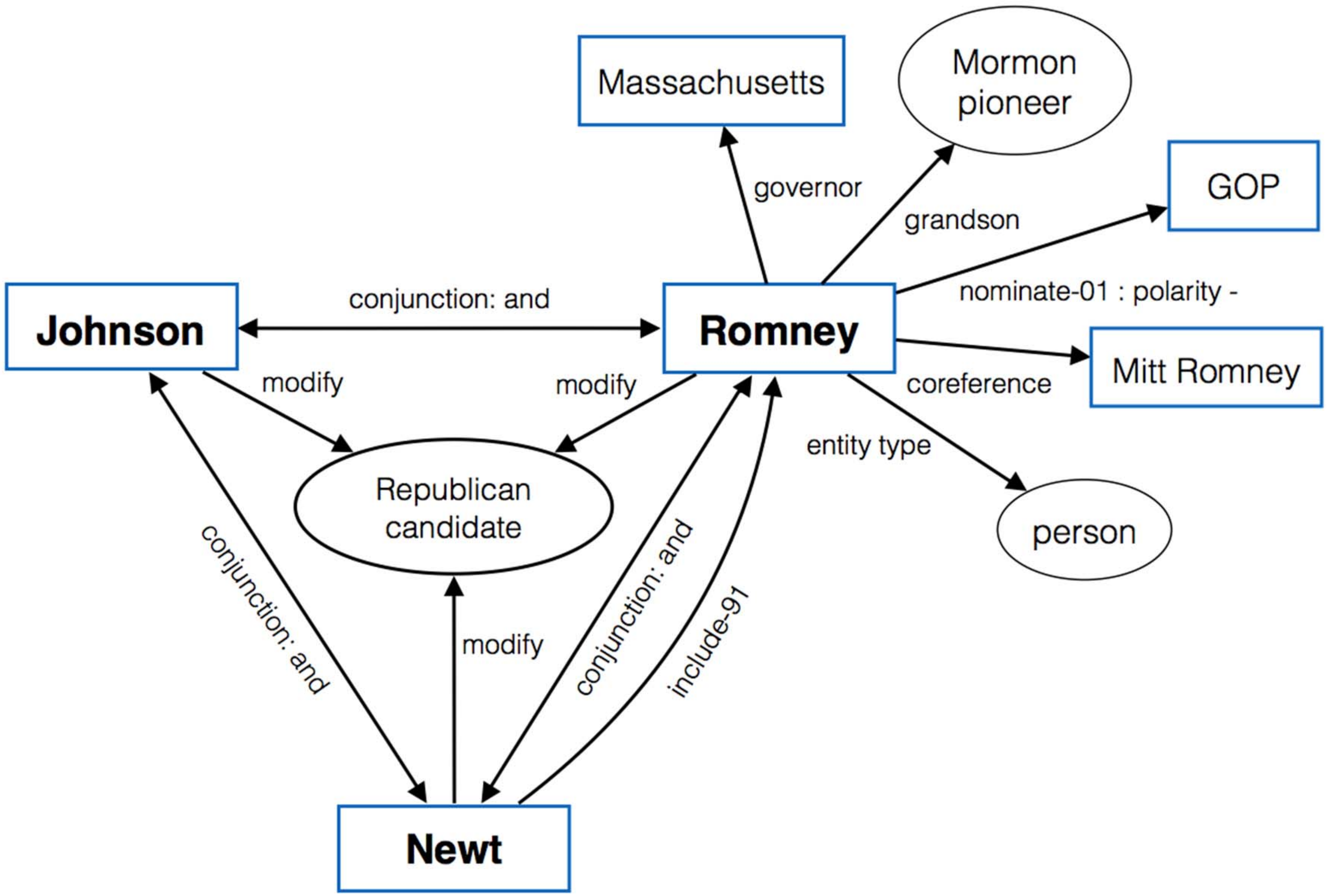
- Entity mentions involved in AMR conjunction relations should be linked jointly to KB; their candidates in KB should also be strongly connected to each other with high semantic relatedness
  - "and", "or", "contrast-01", "either", "compared to", "prep along with", "neither", "slash", "between" and "both"



Put Everything Together



Put Everything Together



Construct Knowledge Networks of Entity Candidates and their Collaborators in KB

- Wikipedia (Han and Zhao, 2009)
  - Wikipedia titles and their surface forms
  - Associative relation (internal page links), hierarchical relation and equivalence relation between concepts
  - Polysemy (disambiguation page) and synonymy (redirect page) between key terms
  - Templates (Zheng et al., 2014)
- DBPedia (Zheng et al., 2014)
  - Rich relational structures and hierarchies, fine-grained types

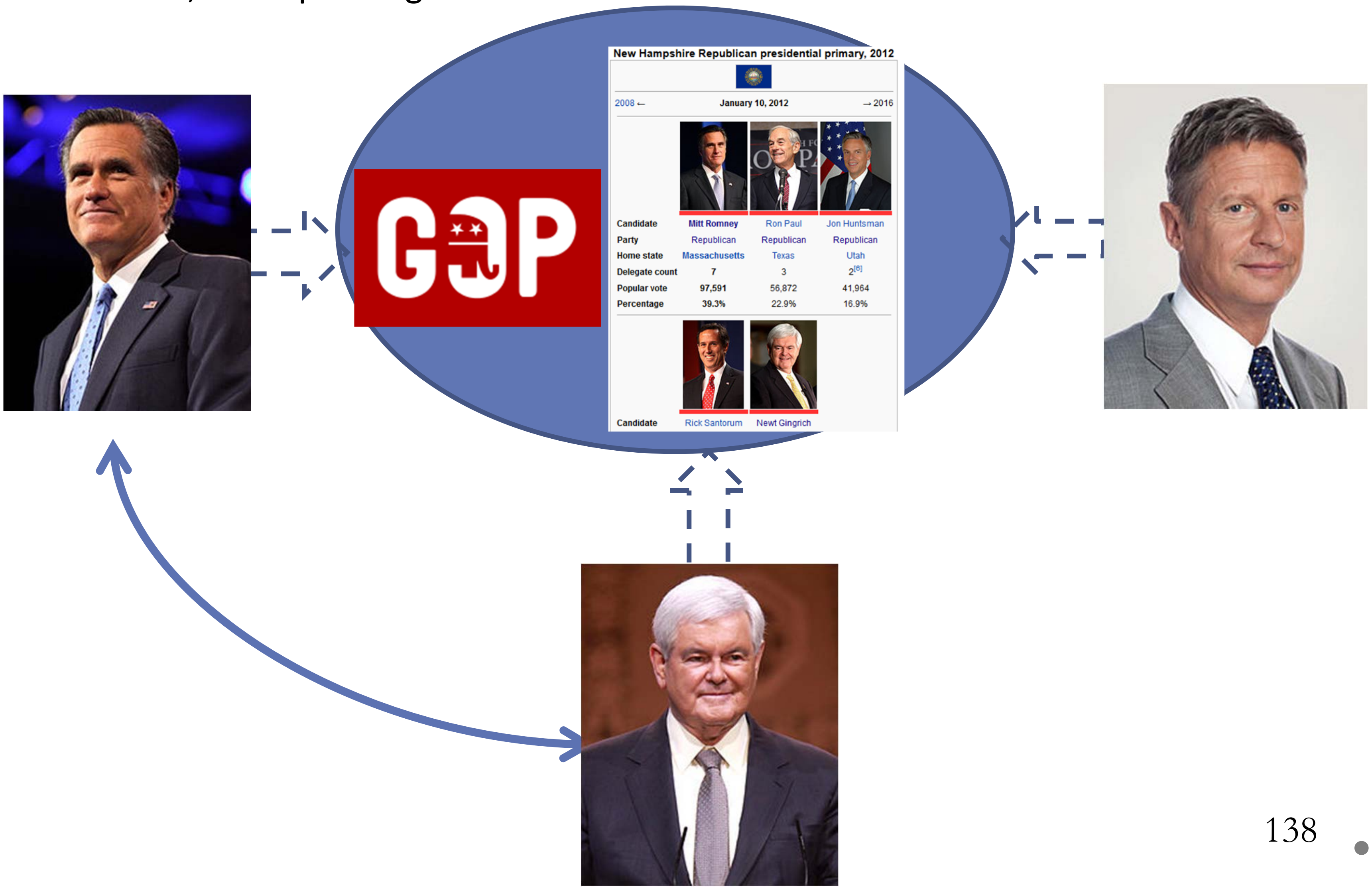
Willard Mitt Romney (born March 12, 1947) is an American businessman who was the *Republican Party's* nominee for President of the United States in the 2012 election. Before his presidential bid, he served as the *70th Governor of Massachusetts* from 2003 to 2007.

Religion	The Church of Jesus Christ of Latter-day Saints (Mormon)

Political party	Republican

Collaborators in KB

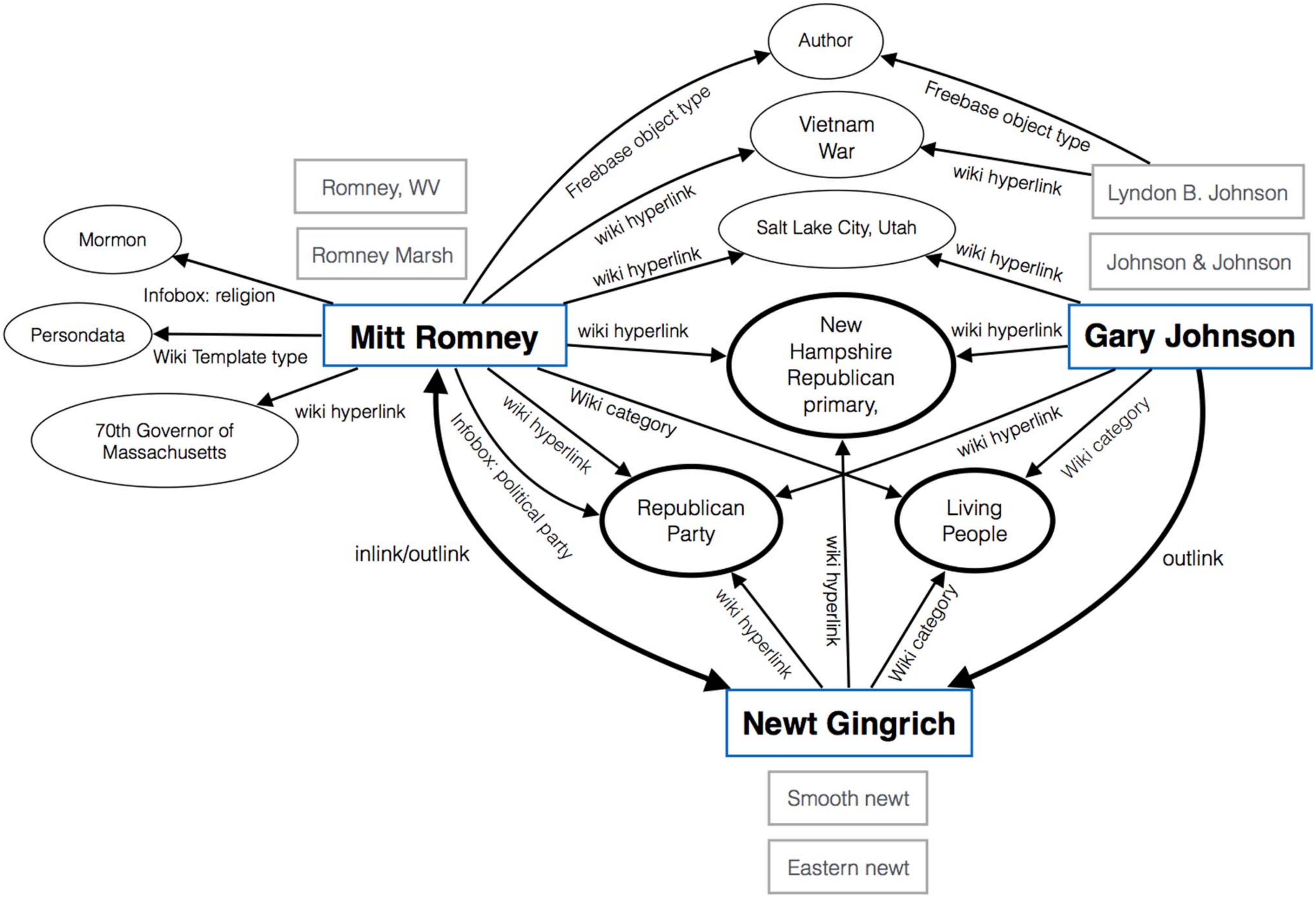
- Explicit inlinks / outlinks
- Implicit semantic relatedness: overlapped hyperlinks in Wikipedia titles, articles, infoboxes, concept categories and external links



Collaborators in KB

- Go beyond Wikipedia: Exploit rich structures in DBPedia, Freebase, YAGO, Ontologies
- Google Knowledge base: “people also search for”

Construct Knowledge Networks of Entity Candidates and their Collaborators in KB



Salience

$$Commonness(m \Rightarrow t) = \frac{count(m \to t)}{\sum_{t' \in W} count(m \to t')}$$

Typography

By default, a font called Charcoal is used to replace the similar Chicago typefa additional system fonts are also provided including Capitals, Gadget, Sand, Te operating system need to be provided, such as the Command key symbol, ⌘. I

Airlines and destinations

Although the population of Iceland is only about 300,000, there are scheduled flights to and from seven locations in the United States (Boston, Chicago, Minneapolis, New York, Orlando, Seattle, and Washington), three in Canada (Halifax, Toronto and Winnipeg) and 30 cities across Europe. The largest carriers at Keflavík are Icelandair and Iceland Express.

P(Title|"Chicago")

The Greatest Show on Earth were a British rock band, who recorded two albums for Harvest Records in 1970.

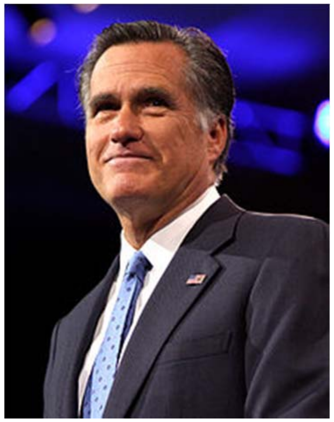
The band had been conceived by Harvest Records in an attempt to create a horn-based rock combo, such as Blood Sweat & Tears or Chicago.[1]

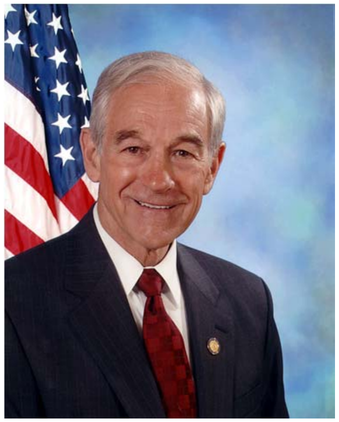
Salience based Ranking

- **Mitt Romney**
- Mitt Romney Presidential campaign, 2012
- George W. Romney
- Romney, West Virginia
- New Romney
- George Romney (painter)
- HMS Romney (1708)
- Romney family
- Romney Expedition
- Old Romney

- Paul McCartney
- **Ron Paul**
- Paul the Apostle
- St Paul's Cathedral
- Paul Martin
- Paul Klee
- Paul Allen
- Chris Paul
- Pauline epistles
- Paul I of Russia

- Lyndon B. Johnson
- Andrew Johnson
- Samuel Johnson
- Magic Johnson
- Jimmie Johnson
- Boris Johnson
- Randy Johnson
- Johnson & Johnson
- **Gary Johnson**
- Robert Johnson





Similarity and Coherence

- Similarity
  - M: entity mention m's neighbors
  - C: the neighbors of candidate entity c in KB
  - Compute similarity between M and C based on Jaccard Index

$$J(M, C) = \frac{M \cap C}{M \cup C}$$

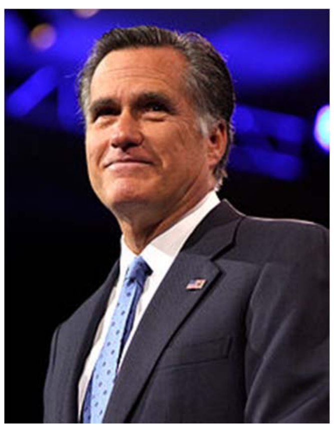
- Coherence
  - C1: the neighbors of candidate entity c1 in KB
  - C2: the neighbors of candidate entity c2 in KB
  - Compute similarity between C1 and C2 based on Jaccard Index

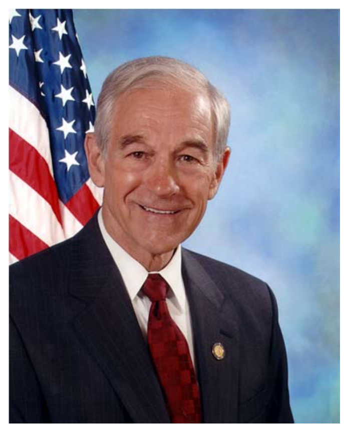
Similarity based Re-Ranking

- **Mitt Romney**
- George W. Romney
- Mitt Romney Presidential campaign, 2012
- Romney family
- George Romney (painter)
- Romney, West Virginia
- HMS Romney (1762)
- New Romney
- HMS Romney (1708)
- Old Romney

- **Ron Paul**
- Paul McCartney
- Paul the Apostle
- Paul Martin
- Paul Klee
- Paul Allen
- Chris Paul
- Paul I of Russia
- Paul
- Paul of Greece

- Andrew Johnson
- **Gary Johnson**
- Samuel Johnson
- Magic Johnson
- Jimmie Johnson
- Boris Johnson
- Randy Johnson
- Robert Johnson
- Jack Johnson (boxer)



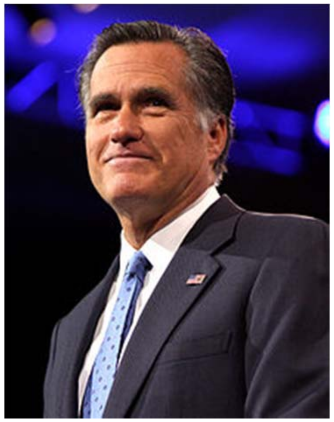


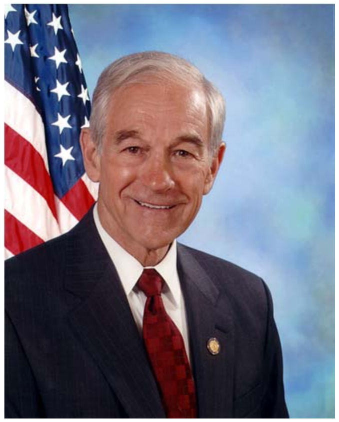
Coherence based Re-Ranking

- **Mitt Romney**
- Mitt Romney Presidential campaign, 2012
- George W. Romney
- Romney, West Virginia
- New Romney
- George Romney (painter)
- HMS Romney (1708)
- Romney family
- Romney Expedition
- Old Romney

- **Ron Paul**
- Paul McCartney
- Paul Martin
- Paul Allen
- Chris Paul
- Paul of Greece
- Paul I of Russia
- Gregory S. Paul
- Paul Klee
- Paul the Apostle

- **Gary Johnson**
- Lyndon B. Johnson
- Andrew Johnson
- Boris Johnson
- Magic Johnson
- Samuel Johnson
- B. S. Johnson
- Jimmie Johnson
- Robert Johnson
- Lawrence Alexander Sidney Johnson





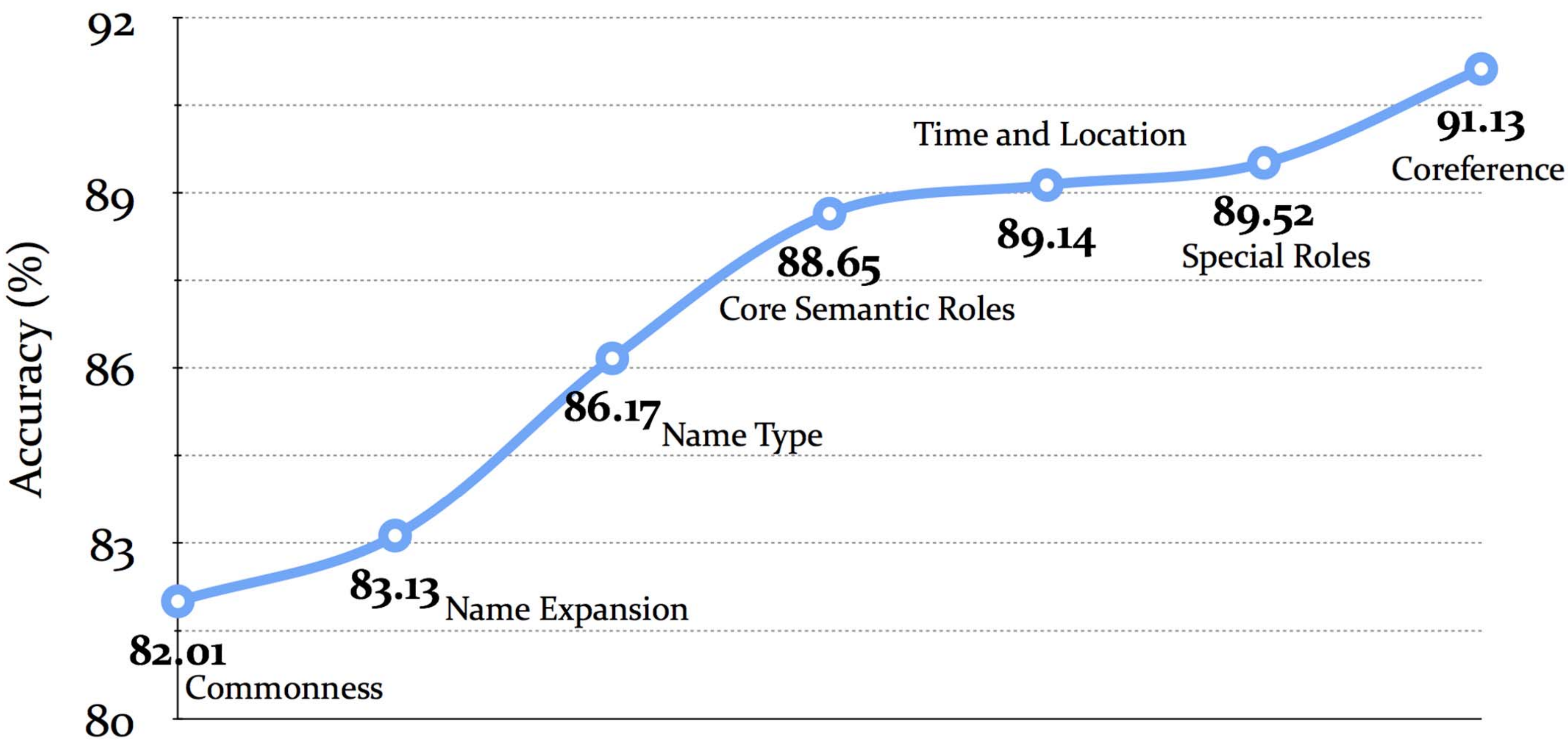
Linking Accuracy on AMR Corpus

(Pan et al., 2014 submission; collaboration with ISI)

- Tested on 1,613 entity mentions (394 persons, 316 organizations, 903 geo-political entities)
- Ranked top 2 in KBP2014; the top 1 system is fully supervised system trained from 20K queries, 5+ years feature engineering

Method		Acc@1
Popularity	Commonness	82.01%
	Google Search	84.12%
Supervised (trained from 20,000 entity mentions)	Ground-truth was created based on correcting output from this model	91.01%
Unsupervised	Human AMR	**91.13%**
	System AMR	**89.14%**
	Document-level Co-occurrence	80.90%
	Human Ontonotes-type Core-role Semantic Role Labeling	85.24%

AMR Knowledge Learning Curve

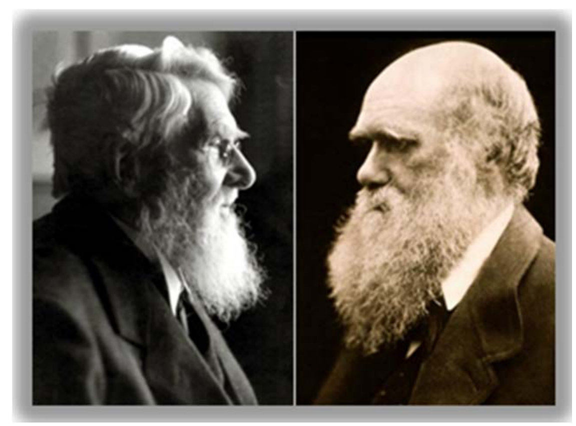


From Non-collective to Collective

- Intuition: Promote semantically coherent pairs of titles
  - (1) Enrich text with external KB knowledge
    - Use graph metrics (as described earlier)
  - (2) Enrich text with (some) gold titles
    - Use graph propagation algorithms
  - (3) Enrich Text with (local) relational information
    - Use global inference methods

Collaborative Learning

Collective Animal Behavior



Great Minds Think Alike

Enrichment with (some) Gold Titles:
Inference with Graph Regularization (Huang et al., 2014)

- Relational Graph Construction with semantic relations
  - Perform collective inference to identify and link a set of semantically related mentions
  - Make use of manifold (cluster) structure and need less training data
- Semi-supervised Graph Regularization
  - Loss Function: ensure the refined labels is not too far from the initial labels
  - Regularizer: smooth the refined labels over the constructed graph
  - Both closed and iterative form solutions exist

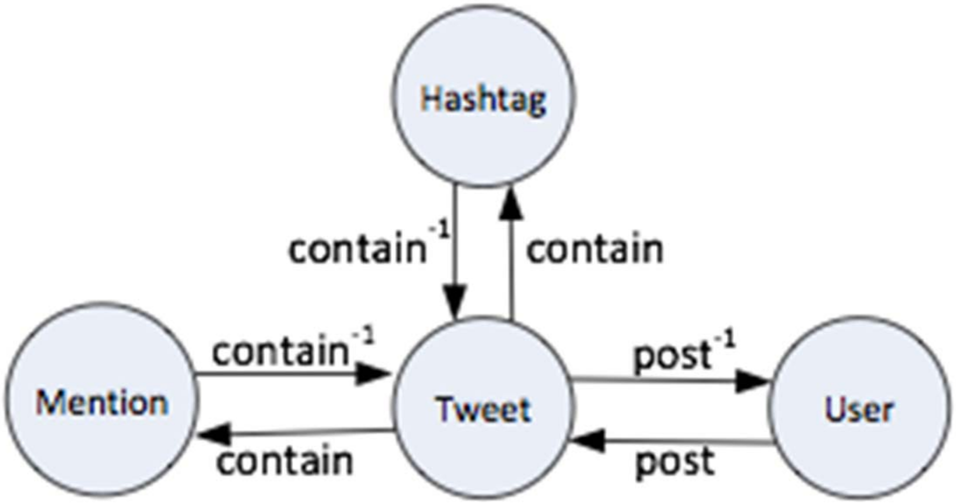
$$\mathcal{Q}(\mathcal{Y}) = \mu \underbrace{\sum_{i=l+1}^{n} (y_i - y_i^0)^2}_{\text{Loss Function}} + \underbrace{\frac{1}{2} \sum_{i,j} W_{ij} (y_i - y_j)^2}_{\text{Regularizer}}.$$

Collective Inference with Social Relations

- Stay up *Hawk Fans*.
- We are going through a ***slump***,
- but we have to stay positive. Go *Hawks*!

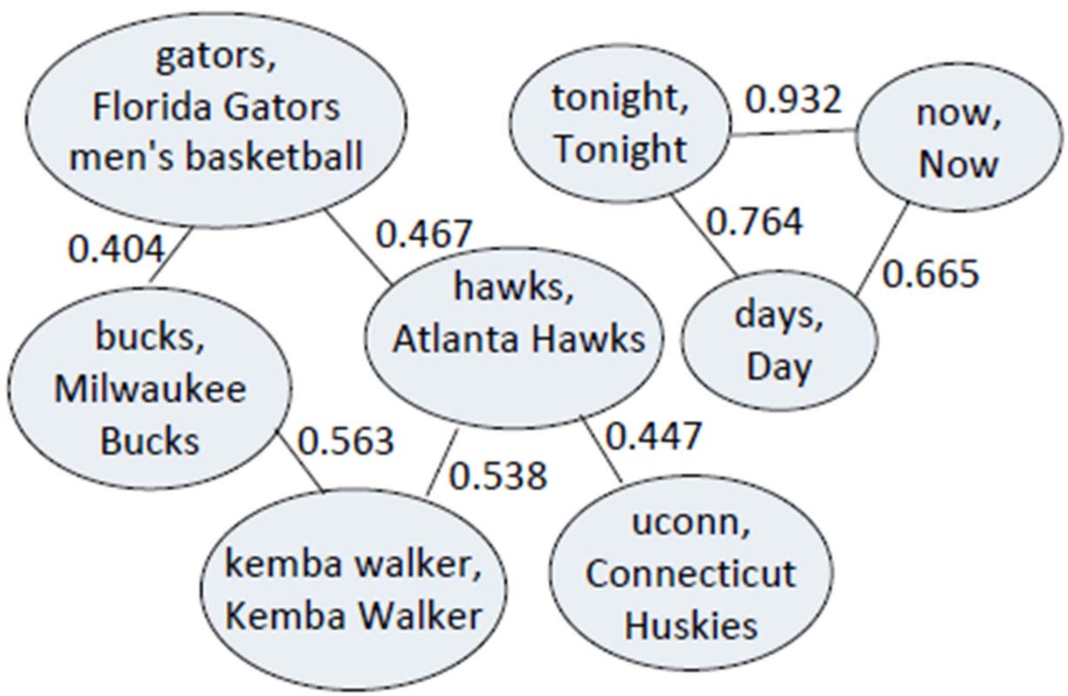
Meta Path

- A meta-path is a path defined over a network and composed of a sequence of relations between different object types (Sun et al., 2011)
- Meta paths between mentions
  - o M-T-M
  - o M-T-U-T-M
  - o M-T-H-T-M
  - o M-T-U-T-M-T-H-T-M
  - o M-T-H-T-M-T-U-T-M

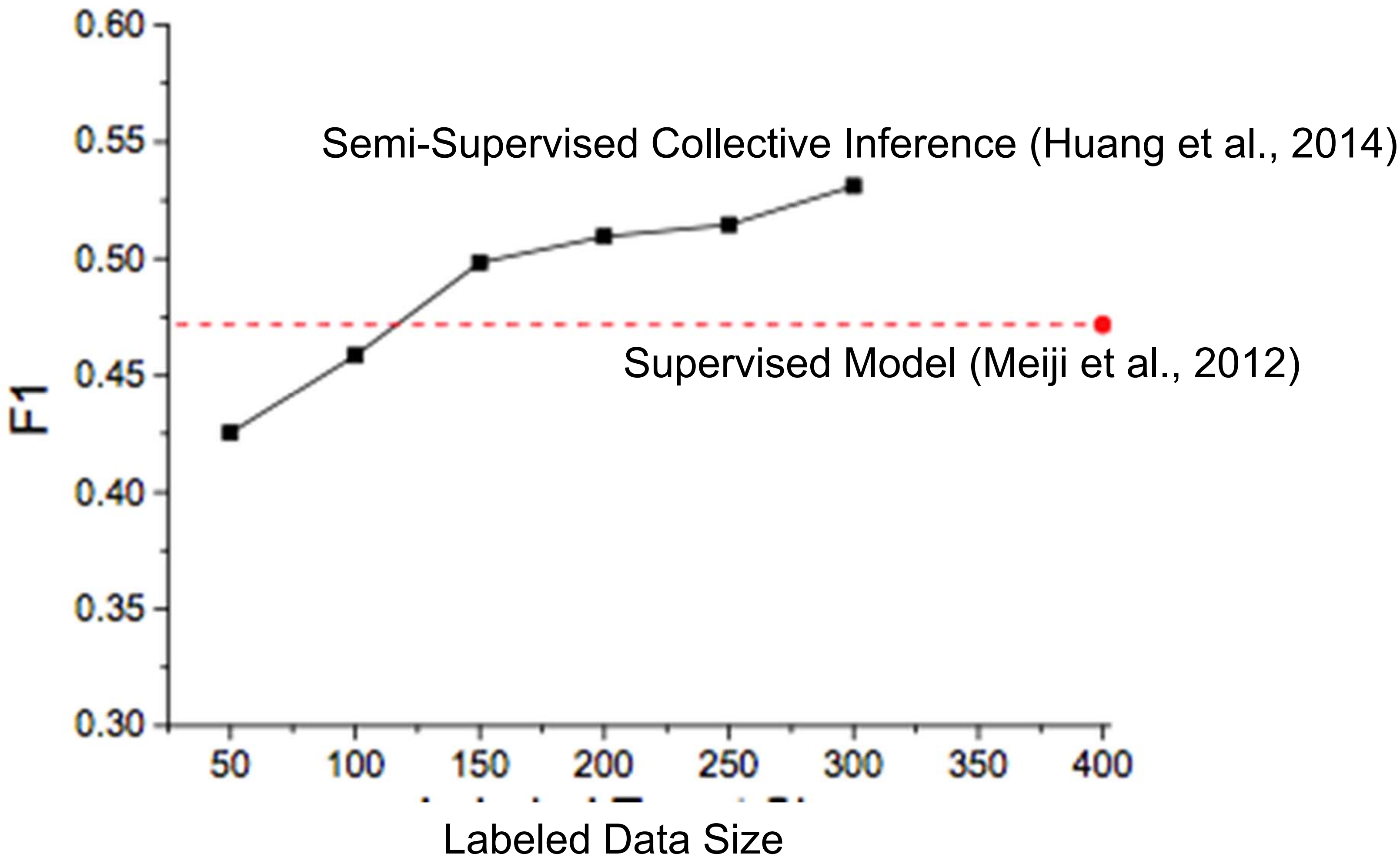


Relational Graph

- Each pair of mention m and concept c as a node
    - m is linkable, and c is the correct concept, <m, c> should be assigned label 1, otherwise 0



Performance Comparison



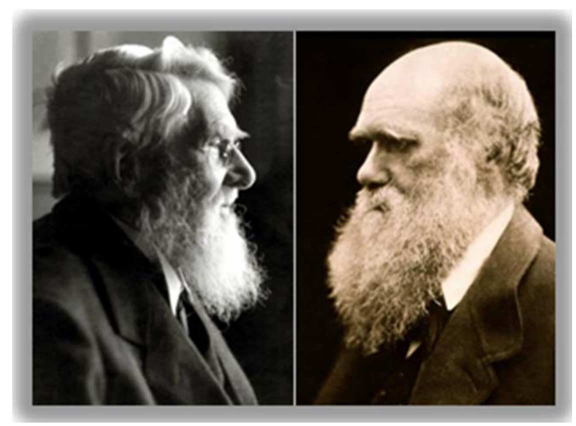
Semi-supervised collective inference with 30% labeled data achieves comparable performance with the state-of-the-art supervised model

From Non-collective to Collective

- Intuition: Promote semantically coherent pairs of titles
  - (1) Enrich text with external KB knowledge
    - Use graph metrics (as described earlier)
  - (2) Enrich text with (some) gold titles
    - Use graph propagation algorithms
  - (3) Enrich Text with (local) relational information
    - Use global inference methods

Collaborative Learning

Collective Animal Behavior



Great Minds Think Alike

Modeling Topical Coherence with DNN and KGs (Huang & Heck, MSR-TR-2014-108, 2014)

- State-of-the-art Approache
  Relatedness
  - (Milne and Witten, 2008):
    method

$$SR(c_i, c_j) = 1 - \frac{\log\max(|C \quad \ldots}{\log(|C \quad \ldots}$$

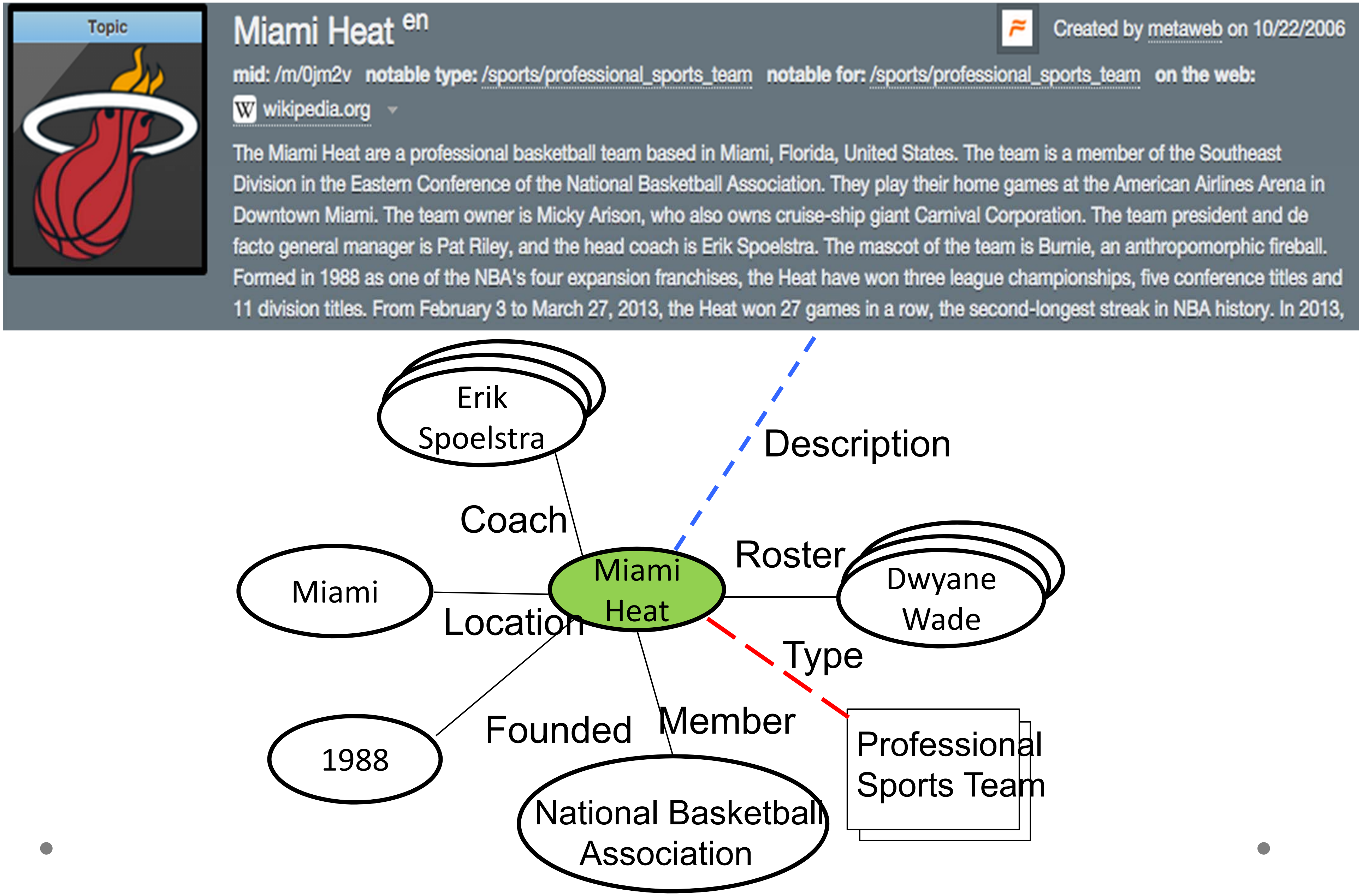
- C: the set of entities in W
- C_i: the set of incoming links to c_i
  - Supervised Method (Ceccarelli et al., 2013)
    - Formulate as a learning-t
    - Explore a set of link-base

Limitation I: Fail to capture deep semantics of entities (e.g., SR(NBA, Chicago Bull) = 0.59 < SR(NBA, Chicago) = 0.83)

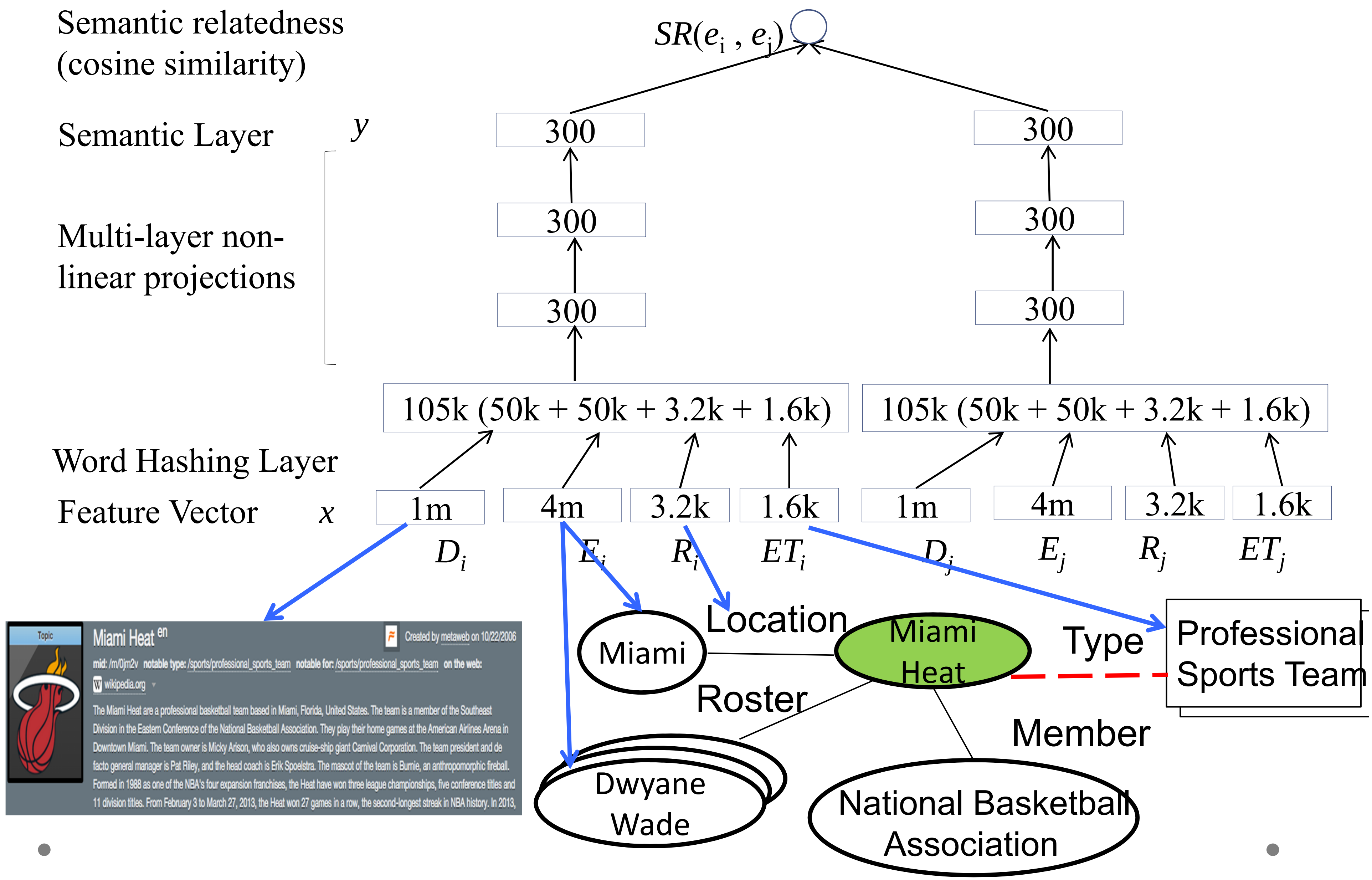
Limitation II: Ingore the knowledge from the rich Knowledge Graphs

Limitation III: what if we donot have anchor links?

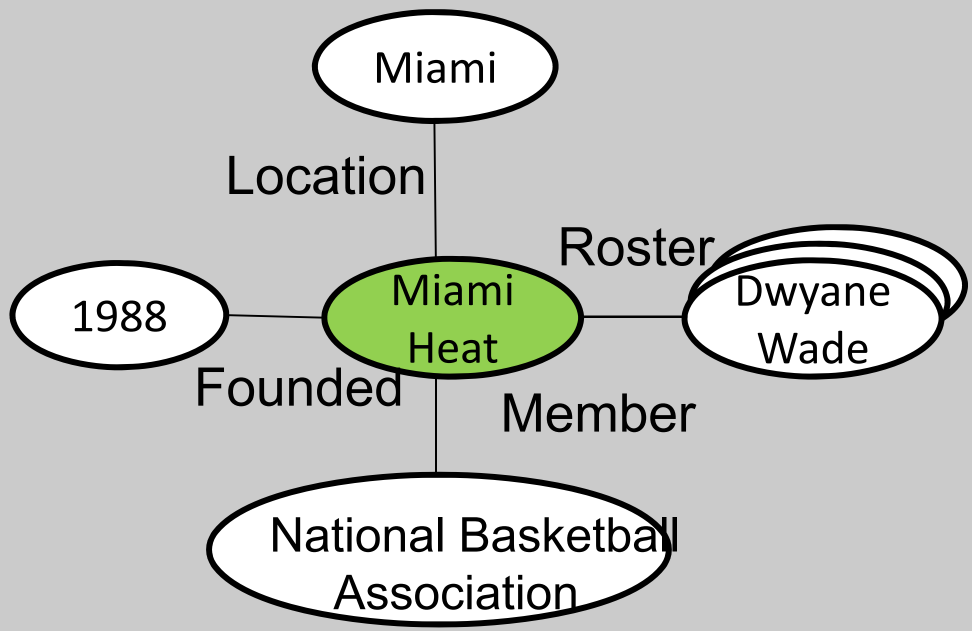
Semantic Knowledge Graphs



A Deep Semantic Relatedness Model (Huang et al., 2014submission)



Encoding Knowledge from KGs

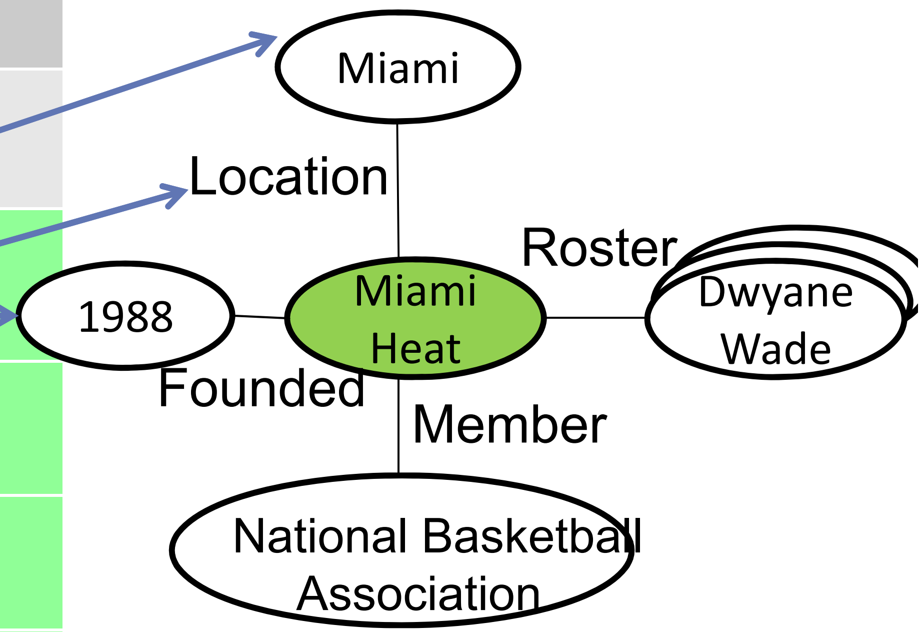
Knowledge	Representation	Example
Description	Letter tri-gram vector	dog = <#do, dog, og#> <0,…,1,1,…,0,1,…,0>
Entity Type	1-of-V vector	<0,….,0,…,1,…,0,…>
Subgraph	1-of-V vector for relation Letter tri-gram for entities	

Models for Comparison

- For news documents
  - **Kul_sp**: A collective inference approach with rounding integer linear programs (Kulkarni et al., 2009).
  - **Shirak**: A naive Bayes probabilistic model with a probabilistic taxonomy (Shirakawa et al., 2011).
  - **AIDA**: A graph-based collective inference approach based on dense subgraph finding (Hoffart et al., 2011).
- For Tweets
  - **TagMe**: an unsupervised model based on prior popularity and semantic relatedness of a single message (Ferragina and Scaiella, 2010)
  - **Meij**: the state-of-the-art supervised approach based on the random forest model (Meij et al., 2012)

Overall Performance on News Documents

Method	Micro P@1	Macro P@1
Kul_sp	72.87%	76.74%
Shirak	81.40%	83.57%
AIDA	82.29%	82.02
GraphRegu + (Milne and Witten, 2008)	82.23%	81.10%
GraphRegu + DSRM + Connected Entities	84.17%	83.30%
+ Relations	85.33%	83.94%
+ Entity Type (e.g., film)	84.91%	83.56%
+ Entity Descriptions	**85.82%**	**84.41%**



- 20% error rate reduction over AIDA
- 20% error rate reduction over the standard method to compute semantic relatedness (Milne and Witten, 2008)

Overall Performance on Tweets

Method	Micro P@1	Macro P@1
TagMe (unsupervised)	61.03%	60.46%
Meiji (5 fold cross-validation)	68.33%	69.19%
GraphRegu + (Milne and Witten, 2008)	65.13%	66.20%
GraphRegu + DSRM + Connected Entities	69.19%	69.04%
+ Relations	70.22%	69.61%
+ Entity Type (e.g., film)	71.50%	70.92%
+ Entity Descriptions	**71.89%**	**71.72%**

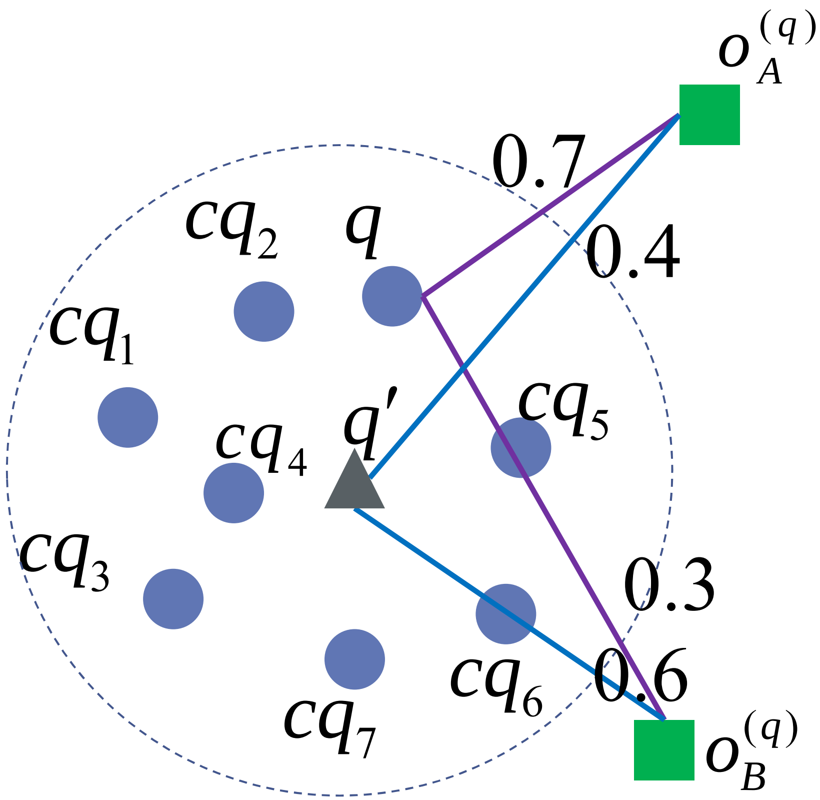
- 28% error rate reduction over TagMe
- 19% error rate reduction over the standard method to compute semantic relatedness (Milne and Witten, 2008)

Comparision of Semantic Relatedness Methods

Method	M&W	DSRM
New York City	0.92	0.22
New York Knicks	0.78	0.79
Washington, D.C.	0.80	0.30
Washington Wizards	0.60	0.85
Atlanta	0.71	0.39
Atlanta Hawks	0.53	0.83
Houston	0.55	0.37
Houston Rockets	0.49	0.80

Semantic relatedness scores between a sample of entities and the entity "**National Basketball Association**" in *sports* domain.

Other Inference Approaches (mostly, local information)



correct rank : $A \prec B$

$f\left(q, o_A^{(q)}\right) = 0.7, f\left(q, o_B^{(q)}\right) = 0.3, A \succ B$

$\Longrightarrow \; g_1(\cdot_A) = 0.4, g_2(\cdot_B) = 0.6, A \prec B$

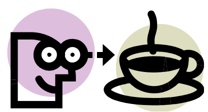
- Relational Inference (Chan and Roth, 2013)
- Collaborative Ranking and Collaborative Clustering (Chen and Ji, 2011)
  - Construct a collaborative network for the target entity based on graph-based clustering
  - Rank multiple decisions from collaborative entities (micro) and algorithms (macro)
  - 7% absolute improvement
- Disambiguate all entities in the contexts with PageRank
  - (Fernandez et al., 2010; Pink et al., 2013; Radford et al., 2010; Cucerzan, 2011; Guo et al., 2011; Han and Sun, 2011; Han et al., 2011; Kozareva et al., 2011; Fahrni et al., 2012; Shen et al., 2013; Liu et al., 2013; Dalton and Dietz, 2013)
- Re-formulate mentions using contexts (Gottipati and Jiang, 2010)
  - "Cambridge" → "Cambridge, Massachusetts, the United States"
  - "Cambridge" → "Cambridge, Ontario, Canada"
- Model contexts using Wikipedia page concepts, and computed linkability scores iteratively (Lehmann et al., 2010, PIoch et al., 2011; Moro et al., 2014)

A Summary: Milestones

- **2006**: The first definition of Wikification task (Bunescu and Pasca, 2006)
- **2009**: TAC-KBP Entity Linking launched (McNamee and Dang, 2009)
- **2008-2012**: Supervised learning-to-rank with diverse levels of features such as entity profiling, various popularity and similarity measures were developed (Gao et al., 2010; Chen and Ji, 2011; Ratinov et al., 2011; Zheng et al., 2010; Dredze et al., 2010; Anastacio et al., 2011)
- **2008-2013**: Collective Inference, Coherence measures were developed (Milne and Witten, 2008; Kulkarni et al., 2009; Ratinov et al., 2011; Chen and Ji, 2011; Ceccarelli et al., 2013; Cheng and Roth, 2013)
- **2012**: Various applications(e.g., Knowledge Acquisition (via grounding), Coreference resolution (Ratinov and Roth, 2012) and Document classification (Vitale et al., 2012; Song and Roth, 2014; Gao et al., 2014)
- **2014**: TAC-KBP Entity Discovery and Linking (end-to-end name tagging, cross-document entity clustering, entity linking)
- **2012-2014**: Many different versions of international evaluations were inspired from TAC-KBP; more than 130 papers have been published

Outline

- Motivation and Definition
- A Skeletal View of a Wikification System
  - o High Level Algorithmic Approach
- Key Challenges

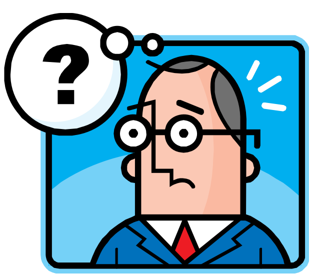


Coffee Break

- Recent Advances
- New Tasks, Trends and Applications
- What's Next?
- Resources, Shared Tasks and Demos

New Trends

- Wikification Until now: Solving Wikification Problems in
  - Standard settings; Long documents
- Extending the Wikification task to new settings
  - Social media Wikification
  - Cross-lingual Entity Linking
  - Linking to general KB and ontologies



Morphs in Social Media

苏亚雷斯(Suarez)

↓

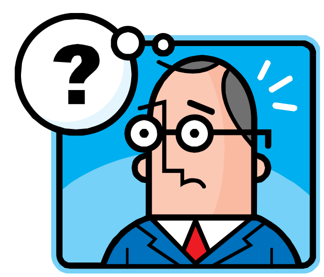
苏牙(Su-tooth)

Chris Christie ⟶ the Hutt

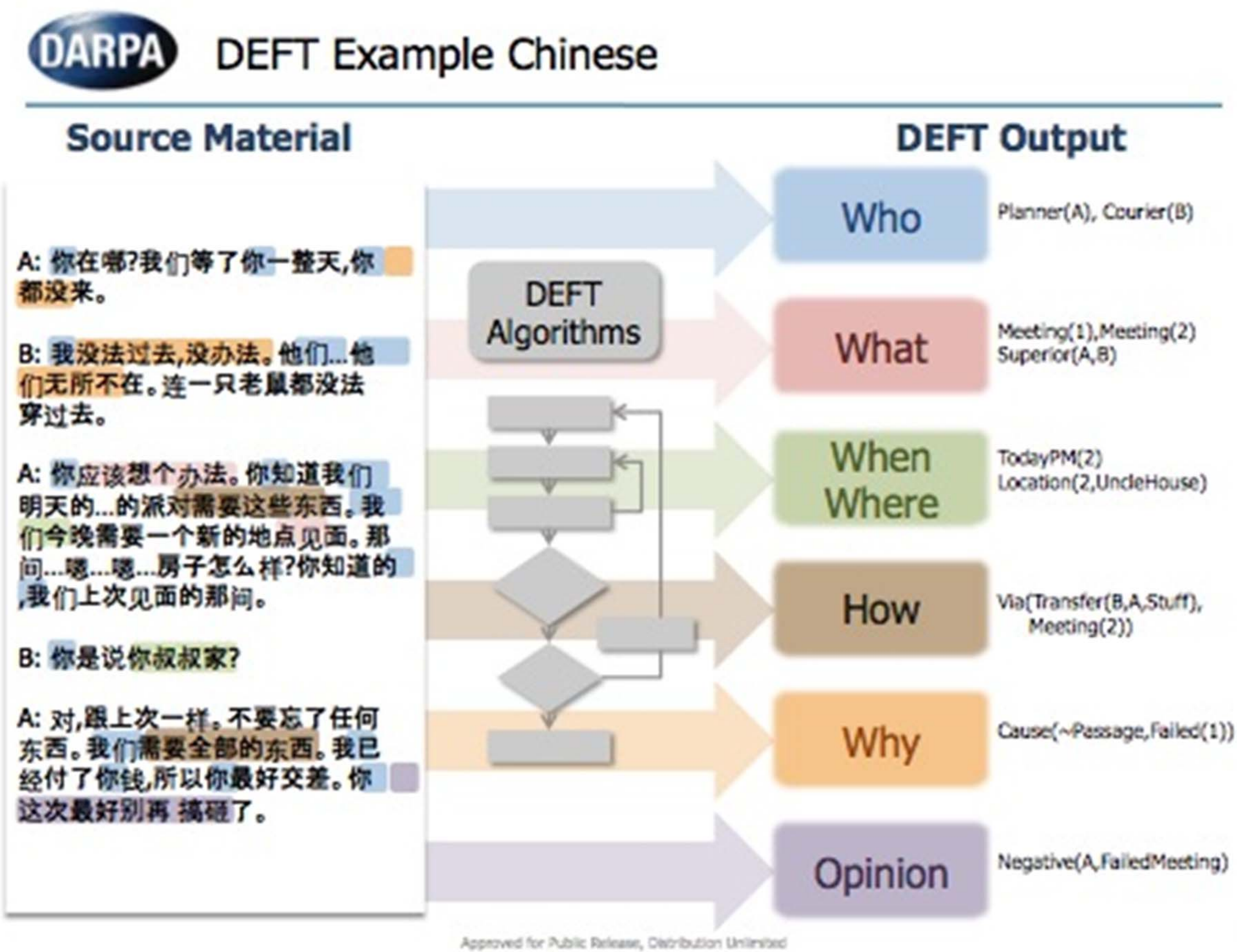
- Morph Encoding (Zhang et al., ACL14)
- Morph Decoding (Huang et al., ACL13)

New Trends

- Wikification Until now: Solving Wikification Problems in
  - Standard settings; Long documents
- Extending the Wikification task to new settings
  - Social media Wikification
  - Cross-lingual Entity Linking
  - Linking to general KB and ontologies
  - Fuzzy matching for candidates



Cross-lingual Knowledge Base Population



Cross-lingual Entity Linking (CLEL)

<query id="SF114">
<name>李安</name>
<docid>XIN20030616.0130.0053</docid>
</query>

李安 － 简介 纠错 | 编辑本段

Parent: Li Sheng

李安，台湾著名导演，祖籍江西省九江市德安县，生于台湾屏东县，父亲李升。李安高中原就读台南二中，后转学考进了台南第一志愿——台南一中。对于读书，李安一点兴趣都没有，心里只想着当导演。大学考试落榜两次，后来准备专科考试，进了国立台湾艺专（今国立台湾艺术大学）影剧科，从此改变了李安的一生。

Birth-place: Taiwan Pindong City

Residence: Hua Lian

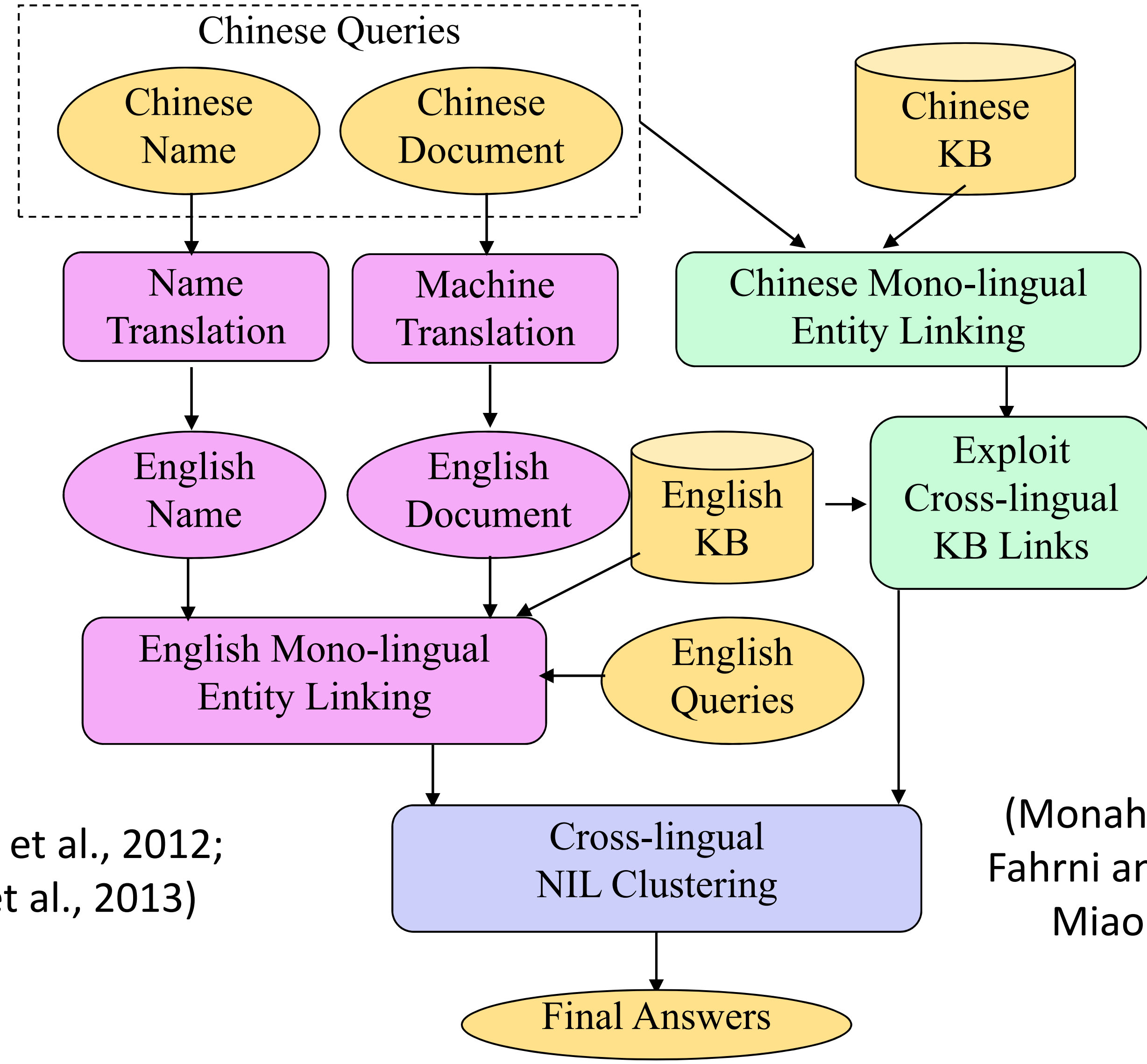
李安曾言，住在花莲的八年，乃其北上就读艺专前最快乐的一段学习岁月。十岁之前的李安在花莲念了两所小学，接受的是美式开放教育，来到台南，又念了两所小学，面对语言习惯不同国语一台语，头一次经验到文化冲击。

Attended-School: NYU

李安于1979年赴美就读伊利诺大学香槟分校戏剧系取得学士学位，后于1981年至纽约大学就读电影制作研究所，取得硕士学位。李安的妻子林惠嘉是伊利诺大学香槟分校生物学博士，现任纽约医学院病理学研究员。

[Cassidy et. al, 2011]

General CLEL System Architecture



(Cassidy et al., 2012; Miao et al., 2013)

(Monahan et al., 2011; Fahrni and Strube, 2011; Miao et al., 2013)

- Major Challenge: Name Translation (Ji et al., 2011)

KBP2015 Tri-lingual EDL Task

- Input: Source Collection: English, Chinese, Spanish
- KB: English only (Chinese KB and Spanish KB are disallowed)
  - Discourage using Inter-lingual Wikipedia links → rapid KB construction for low-density languages
- Output: Entity clusters presented in English, some have links to English KB
  - Some clusters are from single languages; and some are from multiple languages
  - May need to normalize NIL mention translations for ground-truth
- A typical system should extract entity mentions from all three languages, link them to English KB, cluster and translate NIL mentions
- Query: English, Chinese, Corpus
  - Some queries will be from single language only
  - Some queries will exist in multiple languages to form cross-lingual entity clusters

KBP2015 Tri-lingual EDL Task

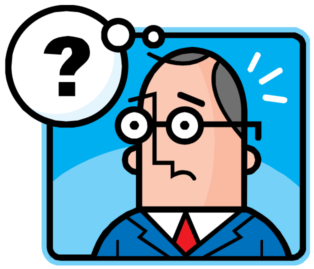
- Diagnostic task
  - Perfect mentions (queries) are given
  - Query: English, Chinese, Spanish
    - Some queries will be from single language only
    - Some queries will exist in multiple languages to form cross-lingual clusters
- Source Collection
  - Some KBA web streaming data in English, Chinese, Spanish
  - Some social media data with code-switch
  - Some formal comparable newswire
  - Some discussion forum posts
  - Include KBP2014 EDL corpora

Timeline

- Release training data in May 2015
- A pilot study in May 2015
  - You can submit manual runs!
- Evaluation: September/October 2015
- We will provide rich annotations in the source collection and marriage opportunities with Spanish teams☺
- Possible extension to Tri-lingual slot filling
- Coordinated by Heng Ji
- Please join us – it's a fun task!

New Trends

- Wikification Until now: Solving Wikification Problems in
  - Standard settings; Long documents
- Extending the Wikification task to new settings
  - Social media Wikification
  - Cross-lingual Entity Linking
  - Linking to general KB and ontologies



Link to General Database

- Many entities are not in Wikipedia
  - In "The Great Gatsby" movie
    - 8 characters listed; only one has a Wikipedia page
  - Many movies are missing

- Question: How can we link open-ended Databases?
  - Challenge: no new labeled data for the new knowledge base
  - Similarity-based features are domain independent [Sil, et. al 12]:
  - Train on the labeled examples based on a sports database
    - Test on the documents with a movie database
  - Very simple approach. Outperform the oracle wikifier on the movie domain

Link to Ontologies

- Wikification as a "Reading Assistant" for Scientific Literature

<table>
<tr><th>KB1</th><th colspan="3">nuclear factor kappa-light-chain-enhancer of activated B cells</th></tr>
<tr><td rowspan="4">KB2</td><td rowspan="4">nuclear factor of kappa light polypeptide gene enhancer in B-cells inhibitor</td><td>KB2-1</td><td>alpha</td></tr>
<tr><td>KB2-2</td><td>beta</td></tr>
<tr><td>KB2-3</td><td>eta</td></tr>
<tr><td>KB2-4</td><td>gamma</td></tr>
<tr><td>KB3</td><td colspan="3">B-cell lymphoma 3-encoded protein</td></tr>
<tr><td>KB4</td><td colspan="3">carboxyl-terminus</td></tr>
</table>

In resting cells, **p50–65 heterodimers** ***[KB1]*** (referred herein as **NF-kB** ***[KB1]***) are sequestered in the cytoplasm by association with members of another family of proteins called **IkB** ***[KB2]***: This family of proteins includes **IkBa** ***[KB2-1]***; **IkBb** ***[KB2-2]***; **IkBe** ***[KB2-3]*** **IkBg** ***[KB2-4]*** and **Bcl-3** ***[KB3]***, but also **p105** ***[NIL1]*** and **p100** ***[NIL2]***, due to their **C-terminal** ***[KB4]*** ankyrin-repeat regions have homologous functions to **IkB** ***[KB2]***.

Link to Biomedical Ontologies

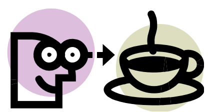
- Wikipedia is not enough
  - o Wikification trained only from news and Wikipedia
  - ➔ 20% end-to-end extraction and linking F-measure

- We can take advantage of the structures of the ontologies
  - o Semantic relations among concepts in the ontologies (e.g. subClassOf) + collective inference technologies ➔ 84.5% (Zheng et al., 2014)

- Another approach:
  - o Wikification: Wikipedia + Ontologies

Outline

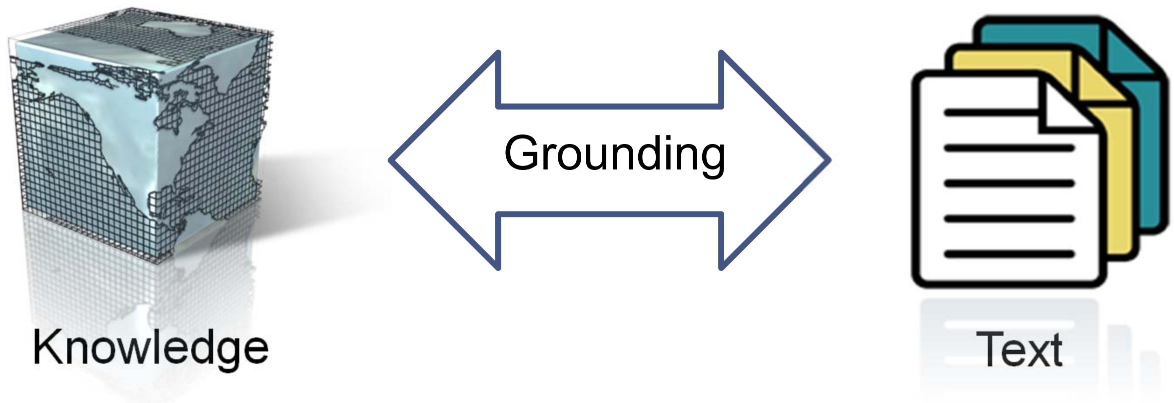
- Motivation and Definition
- A Skeletal View of a Wikification System
  - High Level Algorithmic Approach
- Key Challenges



Coffee Break

- Recent Advances
- New Tasks, Trends and Applications
- What’ s Next?
- Resources, Shared Tasks and Demos

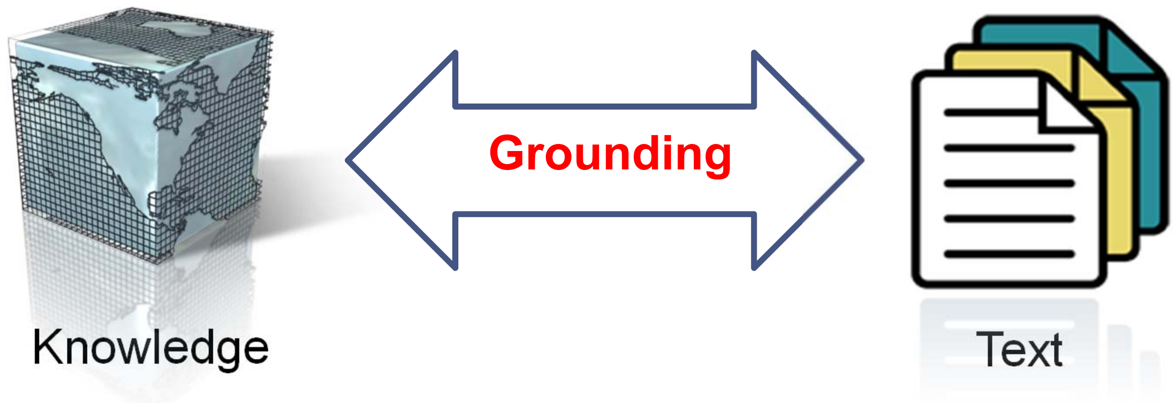
What's Next? Now..



- Wikification & Entity Linking
  - o Understand the semantic of text by "linking/grounding"
- Right now:
  - o Knowledge = (almost) Wikipedia entities
  - o Text = Text-based Documents; News Documents

- How can we bring text understanding to the next level?

Entity Grounding to Knowledge Grounding

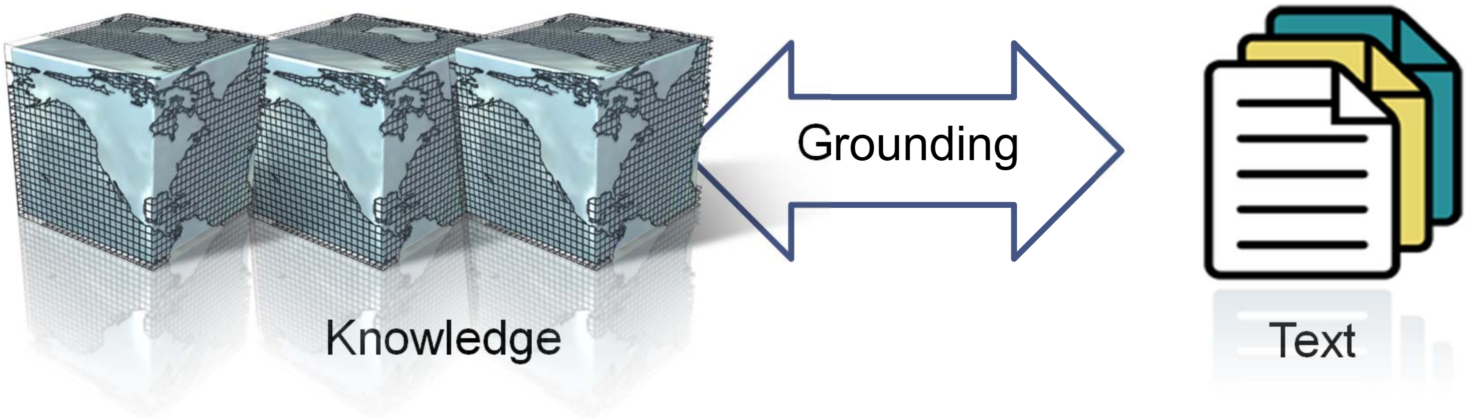


- Knowledge does not only contain entities
  - o Relations: Freebase or Dbpedia
- Large scale semantic parsing
  - o Semantic parsing + Entity Linking?

    Which university did Obama go to? [Berant, et. al, ACL 14]

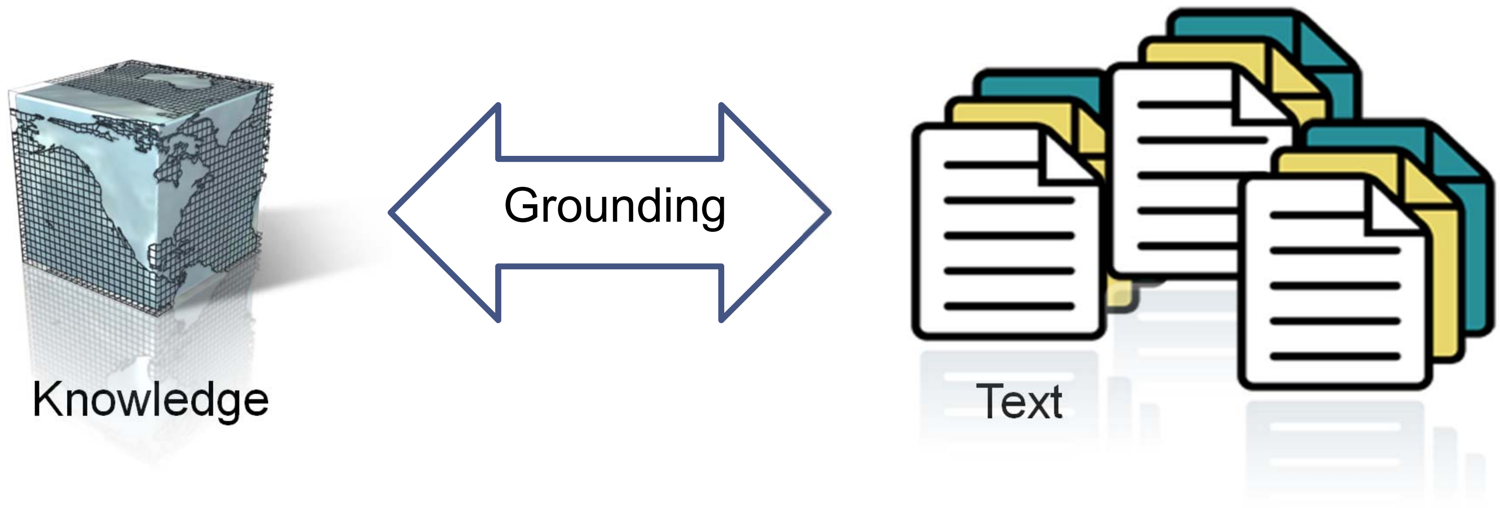
    <u>The lexical matching problem in semantic parsing is entity linking</u>
- Should we jointly ground entities and relations?

Multiple Knowledge Resources



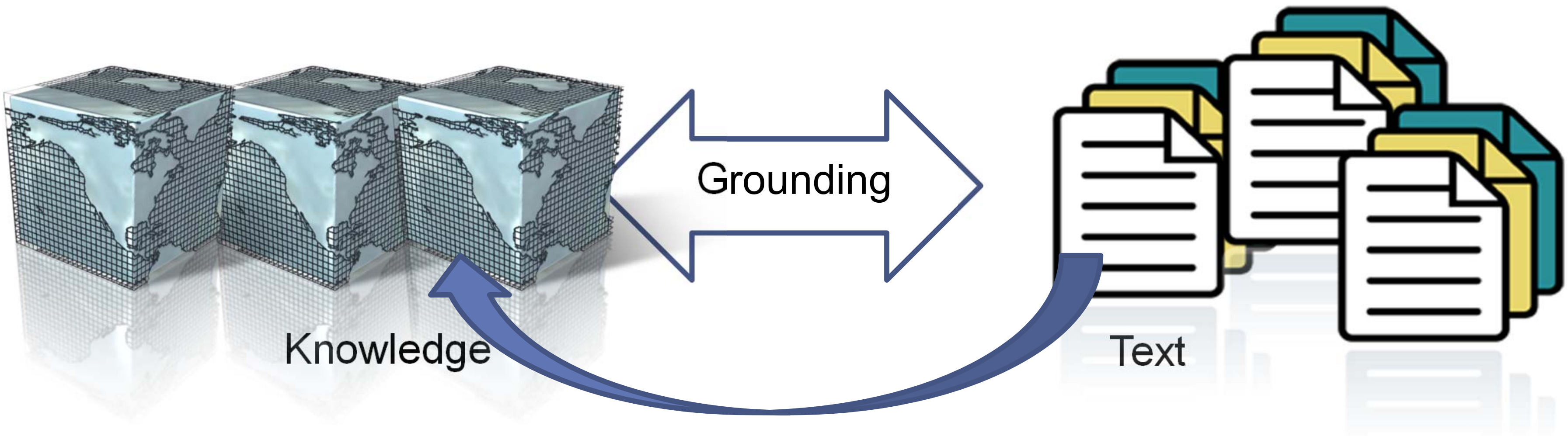
- We have: Wikipedia; Freebase; Customized databases; IMDB…
- How can we have one unified id for all databases?
  - Entity Linking is related to DB Integration
- Different Knowledge bases contain different resources
  - How to utilized them for better grounding?
- How can we use Wikipedia together with another knowledge base?

Handling Various Types of Documents



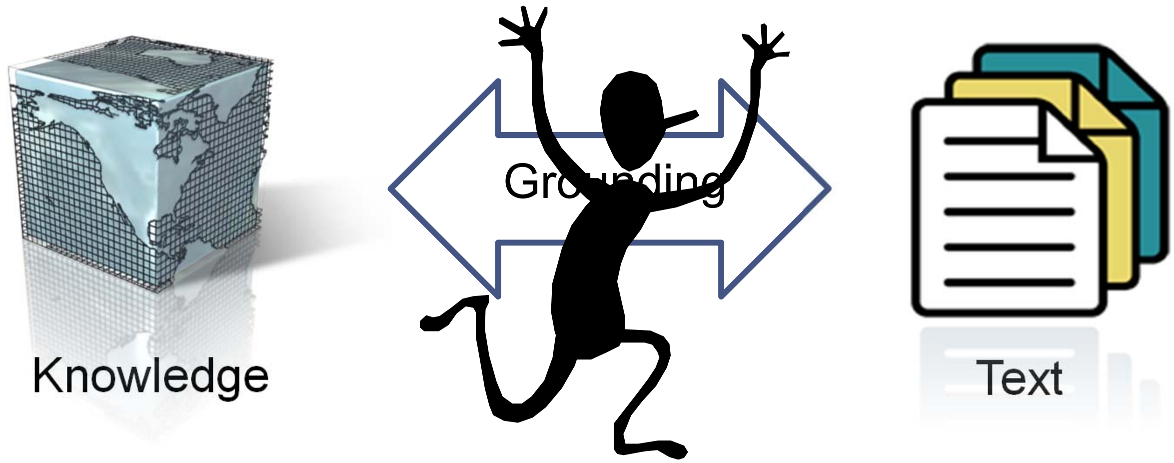
- It is not only text
  - o Webpage, queries, tweets, …. All with meta information
  - o How can we make use of the meta information?
- Noise
  - o How can we develop robust linking techniques on noisy text?
  - o Tweet, table + text, broken format, html
- Text in different languages

Machine Reading



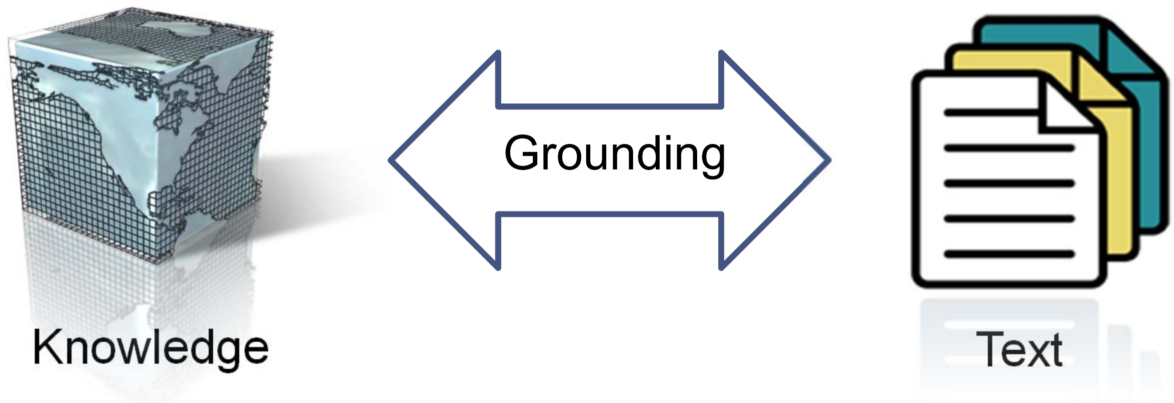
- Can we automatically extract large amounts of knowledge by reading text?
- Grounding should play an important role
    - Grounding ➔ Better understanding of Text ➔ Better Extraction
- How to put it back? Trustworthiness [Dong, et. Al 2014]
- How to handle probabilistic knowledge bases?

Getting Human in the Loop



- How can we apply knowledge grounding to better help human?
  - To understand text better?
  - To query knowledge bases in a better way?
  - Personalize assistant? Educational purposes?

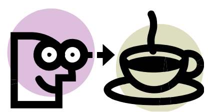
Knowledge Grounding



- Exciting time
  - o We only touched the surface of an emerging field
  - o Machines can do a much better job remembering knowledge
    - Machines should really be our assistant
  - o Research + Engineering (Speed, Scale, Accuracy)

Outline

- Motivation and Definition
- A Skeletal View of a Wikification System
    - o High Level Algorithmic Approach
- Key Challenges



Coffee Break

- Recent Advances
- New Tasks, Trends and Applications
- What's Next?
- Resources, Shared Tasks and Demos

Dataset

Dataset – Long Text

- KBP Evaluations (can obtain all data sets after registration)
  - http://nlp.cs.rpi.edu/kbp/2014/

- CoNLL Dataset
  - http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/aida/downloads/

- Emerging Entity Recognition
  - http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/aida/downloads/

Dataset - Short Text

- Micropost Challenge
  - http://www.scc.lancs.ac.uk/microposts2014/challenge/index.html
- Dataset for “Adding semantics to microblog posts”
  - http://edgar.meij.pro/dataset-adding-semantics-microblog-posts/
- Dataset for “Entity Linking on Microblogs with Spatial and Temporal Signals”
  - http://research.microsoft.com/en-us/downloads/84ac9d88-c353-4059-97a4-87d129db0464/
- Query Entity Linking
  - http://edgar.meij.pro/linking-queries-entities/

Dataset Summary Angela Fahrni (2014)

Data Set	Task	Language Information		Mention Information		Corpus Information		Inventory	Annotation Information		Usage
Name		Setting	Lang.	Definition	Tokens	Source	Texts	Version	Strategy	Agreement	Shared Task
ACE 2005 Bentivogli et al. (2010)	concept and entity disambiguation, recognition of NILs	monolingual	en	ACE mentions (common and proper nouns)	29,300 92.8% in KB 7.2% NILs	broadcast news, newspapers, newswire reports, internet sources, transcribed audio data	597	Online version of Wikipedia 2010 (February - April; August)	Annotated by humans, partly by two annotators	0.85 (Dice coefficient with respect to annotated concepts and entities; before reconciliation)	no
ACE 2004 Ratinov et al. (2011)	concept and entity disambiguation, recognition of NILs	monolingual	en	ACE mentions (common and proper nouns)	306 (84.0% in KB, 16% NILs)	newswire, broadcast news	36	Wikipedia 2011 (?)	mechanical turk, only first mention in coreference chain is annotated	0.85 (agreement, then corrected)	no
IITB Kulkarni et al. (2009)	concept and entity disambiguation, recognition of NILs	monolingual	en	as much as possible, identified by people (including common and proper nouns)	17,200 (60% in KB; 40% NILs)	collection of web pages (sports, entertainment, science and technology, health)	107	Wikipedia dump from August 2008	annotated by humans, partly by two annotators; candidate mentions and tokens were suggested by the system	0.80 (agreement)	no
NewsSc Turdakov & Lizorkin (2009)	concept and entity disambiguation, recognition of NILs	monolingual	en	identified by humans (as many as possible, including common and proper nouns)	8,236 (80.6% in KB, 19.4% NILs)	news articles, scientific papers	131	Wikipedia dump from October 2008	annotated by humans	n.a.	no
MSNBC Cucerzan (2007)	entity disambiguation, recognition of NILs	monolingual	en	proper nouns recognized by a system	756 (83.2% in KB, 16.8% NILs)	MSNBC news (Business, US politics, Entertainment, Health, Sports, Tech & Science, Travel TV news, U.S. News, World News)	20	Wikipedia version from the 11.9.2006	post-hoc evaluation of system output	n.a.	no

Resources

- Reading List
  - o http://nlp.cs.rpi.edu/kbp/2014/elreading.html
- Tool List
  - o http://nlp.cs.rpi.edu/kbp/2014/tools.html
- Shared Tasks
  - o KBP 2014
    - http://nlp.cs.rpi.edu/kbp/2014/
  - o ERD 2014
    - http://web-ngram.research.microsoft.com/erd2014
  - o #Micropost Challenge (for tweets)
    - http://www.scc.lancs.ac.uk/microposts2014/challenge/index.html
  - o Chinese Entity Linking Task at NLPCC2014
    - http://tcci.ccf.org.cn/conference/2014/dldoc/evatask3.pdf

Task and Evaluation

• • •

ERD 2014

- Given a document, recognize all of the mentions and the entities;
  - No target mention is given
- An entity snapshot is given
  - Intersection of Freebase and Wikipedia

- Input: Webpages
- Output: Byte-offset based predictions

- Webservice-driven; Leaderboard

NIST TAC Knowledge Base Population (KBP)

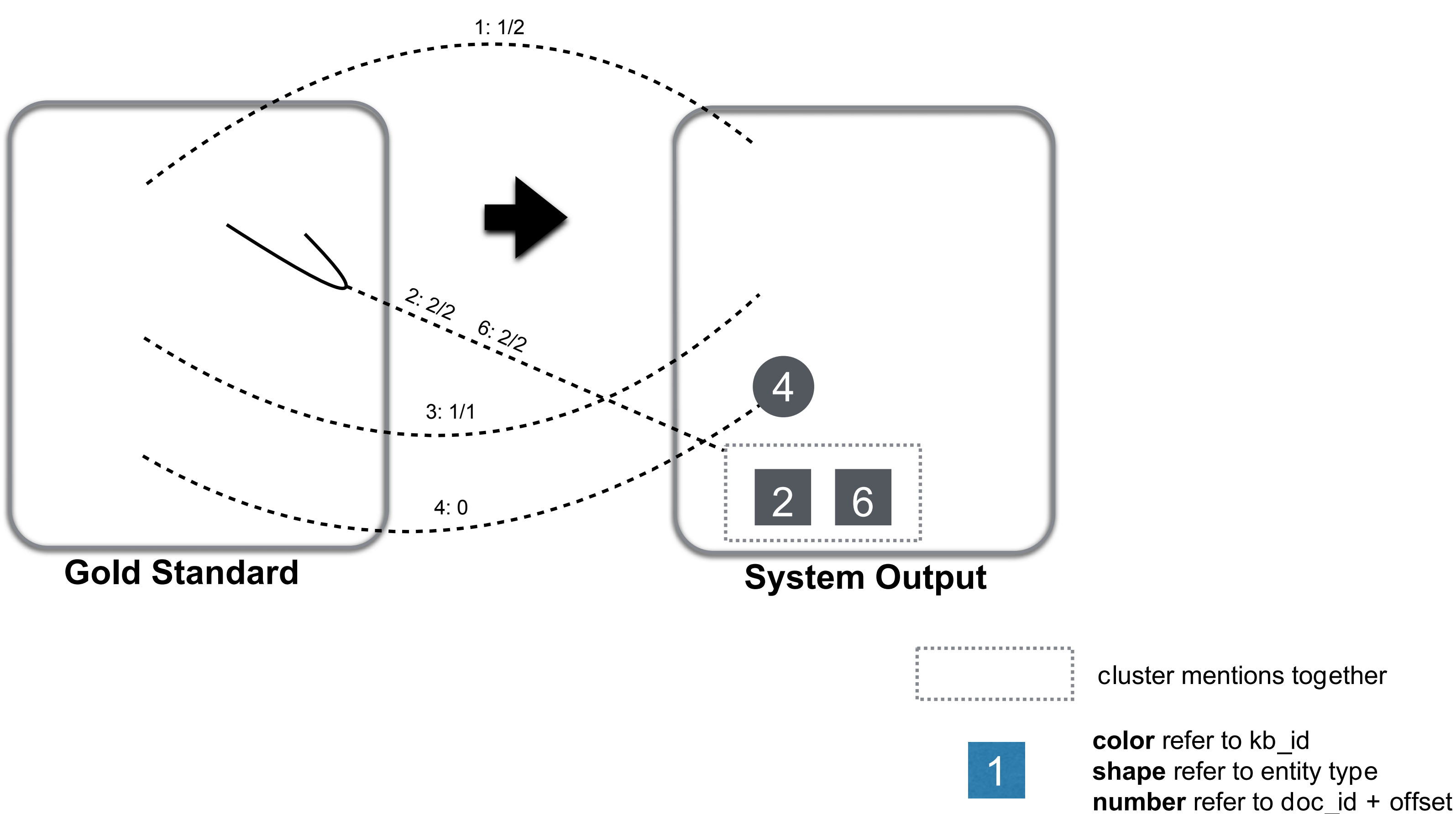
- KBP2009-2010 Entity Linking(Ji et al., 2010)
  - Entity mentions are given, Link to KB or NIL, Mono-lingual
- KBP2011-2013 (Ji et al., 2011)
  - Added NIL clustering and cross-lingual tracks
- KBP2014 Entity Discovery and Linking
  - http://nlp.cs.rpi.edu/kbp/2014/
  - Given a document source collection (from newswire, web documents and discussion forums), an EDL system is required to automatically extract (identify and classify) entity mentions ("queries"), link them to the KB, and cluster NIL mentions
  - English Mono-lingual track
  - Chinese-to-English Cross-lingual track
  - Spanish-to-English Cross-lingual track

Evaluation Metrics

- Concept/Entity Extraction
  - F-Measure, Clustering
- Linking
  - Accuracy @ K (K=1, 5, 10…)
- End-to-end Concept/Entity Extraction + Linking + NIL Clustering
  - B-cubed
  - CEAF
  - Graph Edit Distance
- How should we handle the mention boundary?
  - KBP2014 Rules: Extraction for Population
  - Fuzzy F1-Measure (ERD 2014)
  - Full credit if the predicted boundary overlaps with the gold boundary

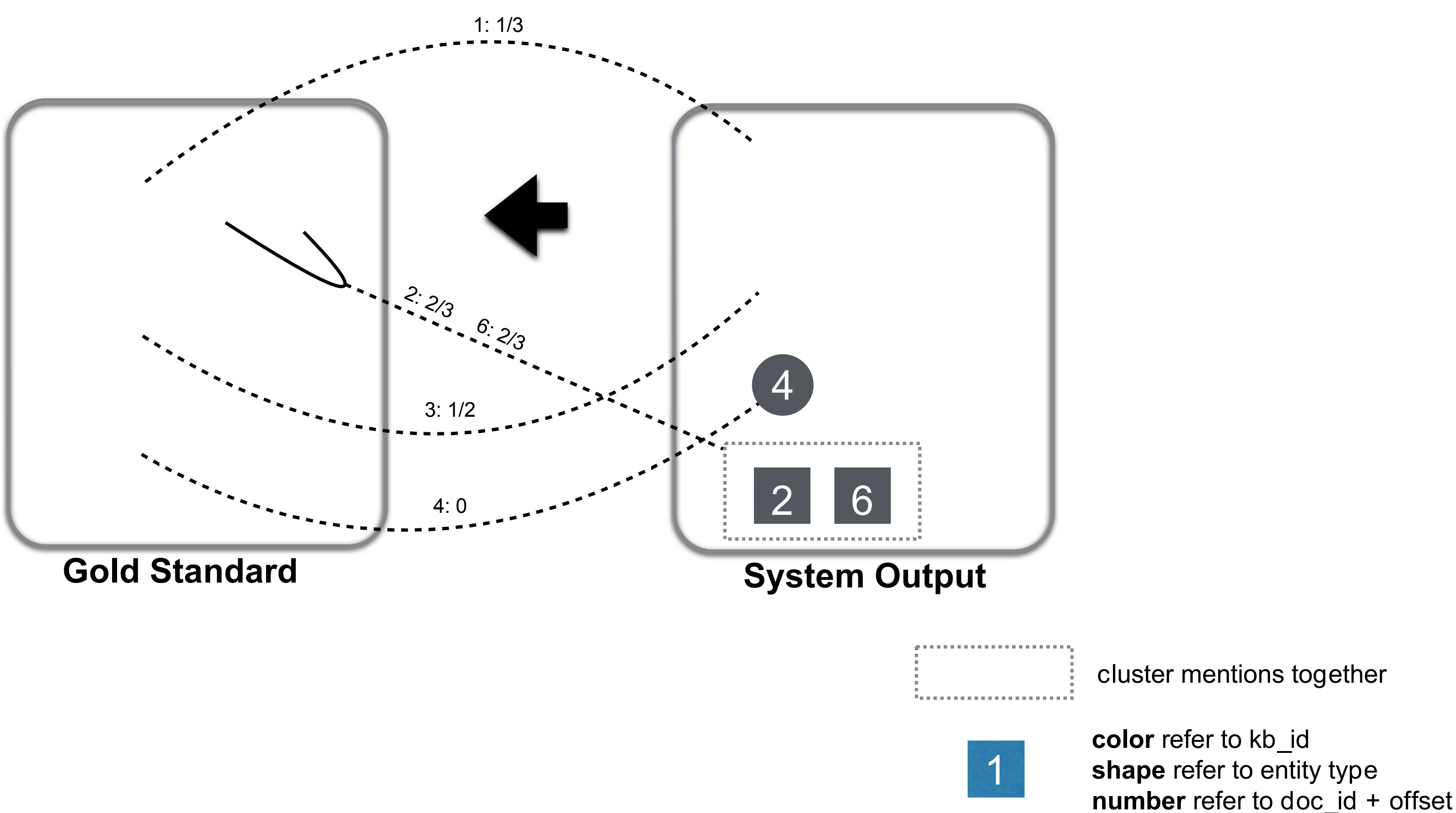
B^3: Precision

- Precision = sum mention credits / #system-output-mentions
  = (1/2 + 2/2 + 2/2 +1/1 + 0)/6 = 0.583



B^3: Recall

- Recall = sum mention credits / #gold-standard-mentions
  $= (1/3 + 2/3 + 2/3 + 1/2)/6 = 0.361$

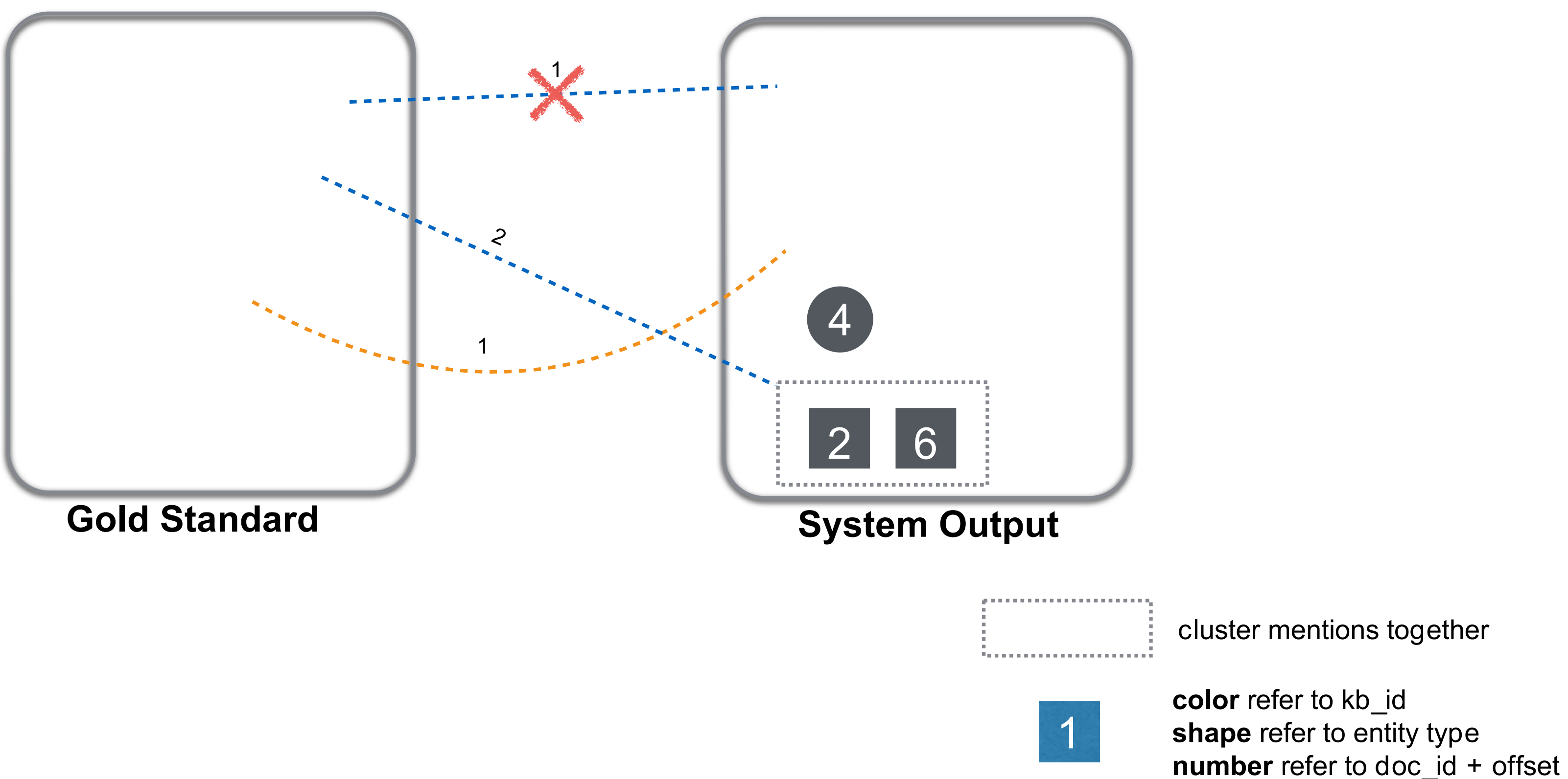


CEAF (Luo, 2005)

- Idea: a mention or entity should not be credited more than once
- Formulated as a bipartite matching problem
  - A special ILP problem
  - Efficient algorithm: Kuhn-Munkres

CEAFm: Example

- Solid: best 1-1 alignment
- $\phi(G_i, S_i) = |G_i \bigcap S_i|$
- Recall=#common / #mentions-in-key = (2+1)/6 = 1/2
- Precision= #common / #mentions-in-response = (2+1)/6 = 1/2



Demo

- • •

Demos

- RPI Wikification Demo
  - Entity Discovery and Linking:
  - http://orion.tw.rpi.edu/~zhengj3/wod/wikify.php

  - AMR based Entity Linker:
  - http://blender02.cs.rpi.edu:3300/

- UIUC Wikification Demo
  - http://cogcomp.cs.illinois.edu/demo/wikify/?id=25