Wikification and Beyond:

The Challenges of Entity and Concept Grounding

Dan Roth (UIUC), Heng Ji (RPI)

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

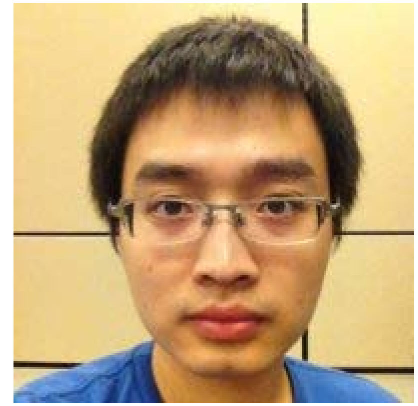
http://nlp.cs.rpi.edu/paper/wikificationtutorial.pdf [pptx]

http://L2R.cs.uiuc.edu/Talks/wikificationtutorial.pdf [pptx]

Thank You – Our Brilliant Wikifiers!

Xiaoman Pan

Jin Guang Zheng



Xiao Cheng

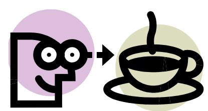
Lev Ratinov

Zheng Chen

Hongzhao Huang

Outline

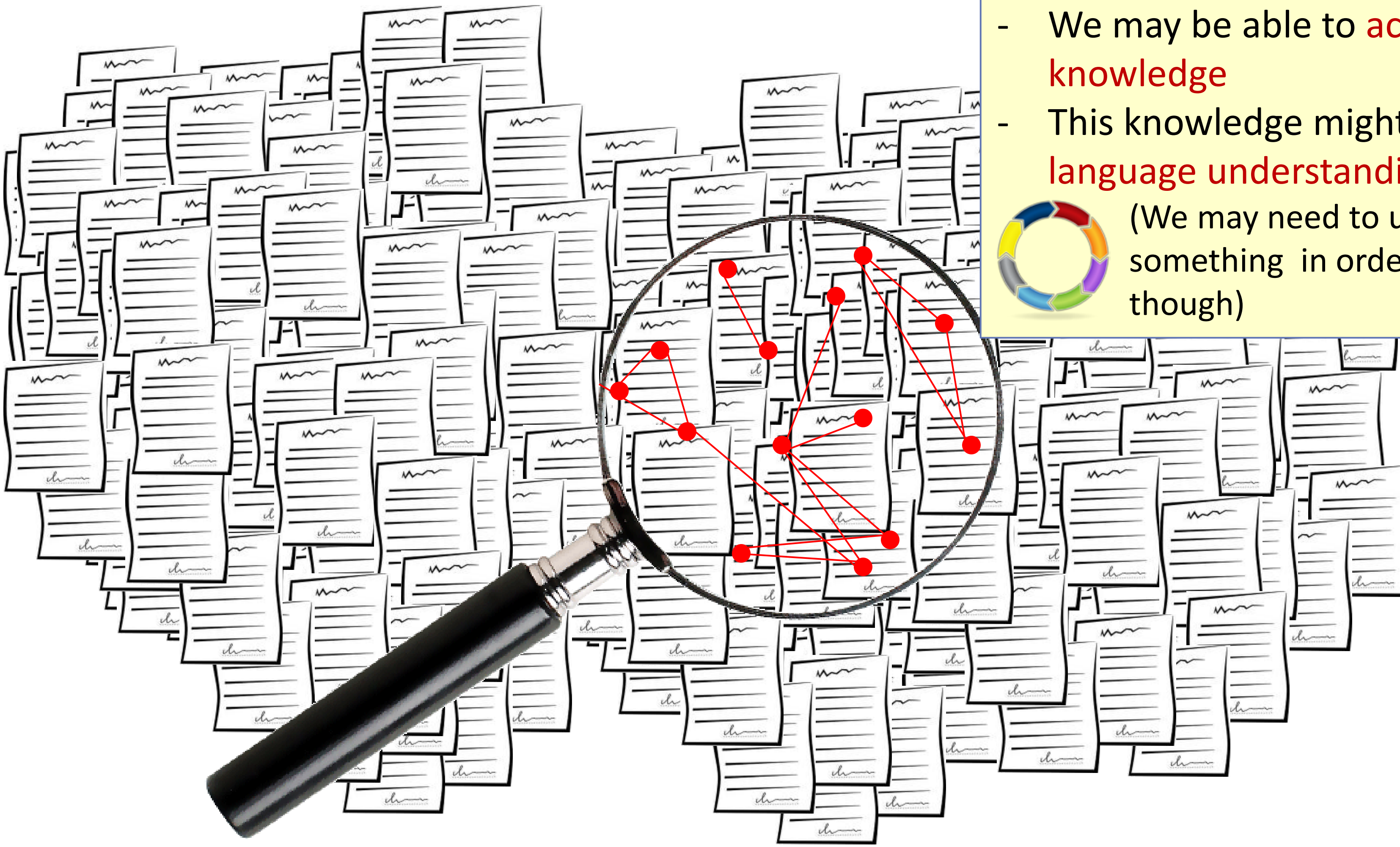
- Motivation and Definition)
- A Skeletal View of a Wikification System
  - High Level Algorithmic Approach
- Key Challenges



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

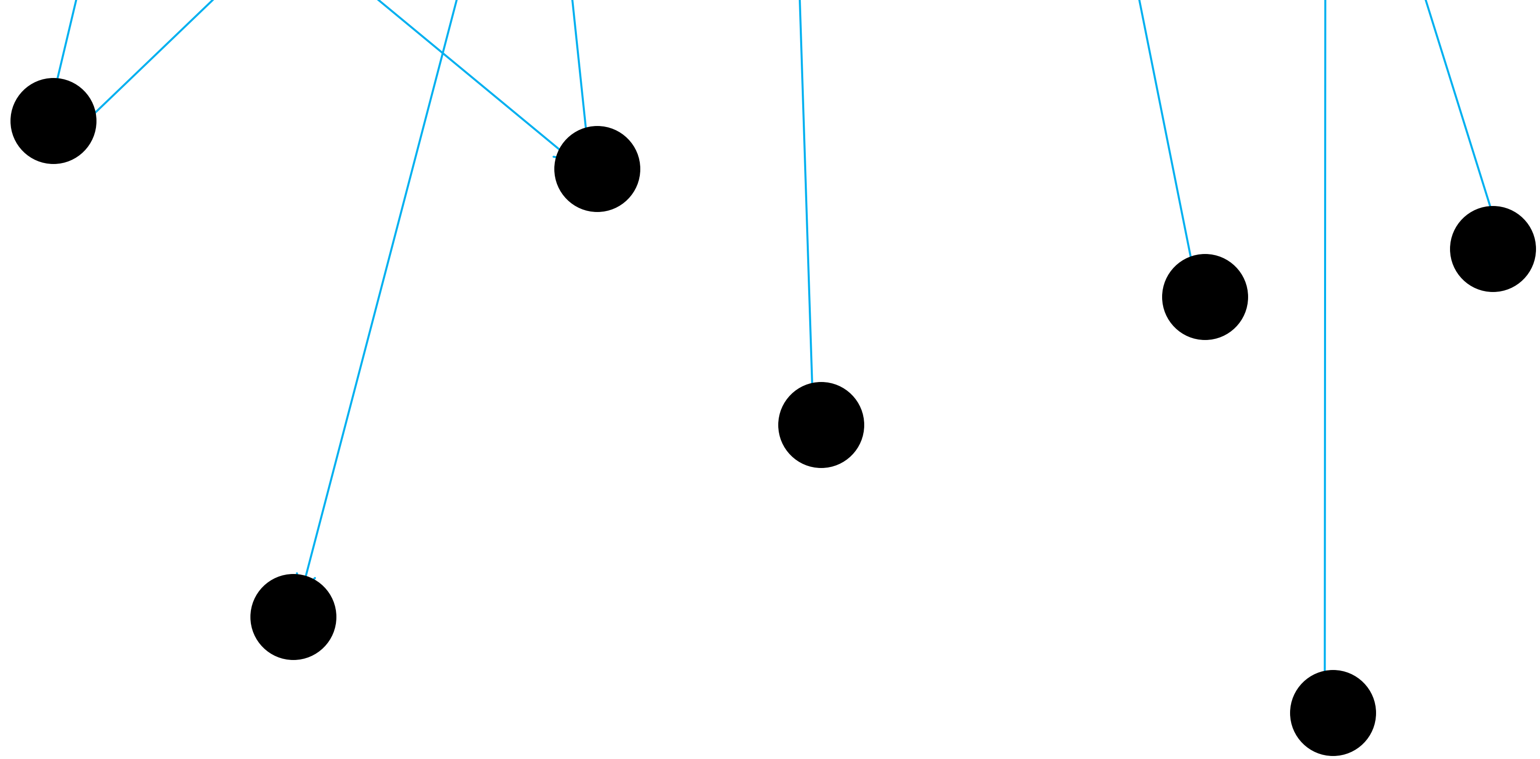
It's a version of ***Chicago*** – the standard classic Macintosh menu font, with that distinctive thick diagonal in the "N".	***Chicago*** was used by default for Mac menus through MacOS 7.6, and OS 8 was released mid-1997..	***Chicago VIII*** was one of the early 70s-era ***Chicago*** albums to catch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

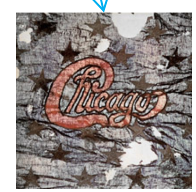
Chicago was used by default for ***Mac*** menus through ***MacOS 7.6***, and ***OS 8*** was released mid-1997..

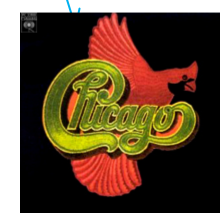
Chicago VIII was one of the early 70s-era ***Chicago*** albums to catch my ear, along with ***Chicago II***.

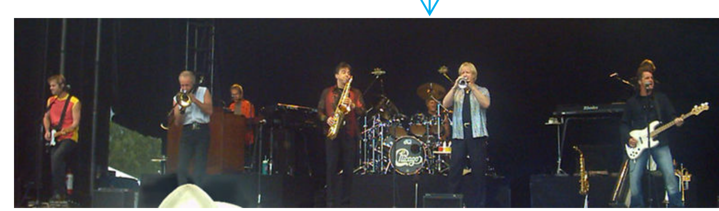


Reference resolution: (disambiguation to Wikipedia)

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***.

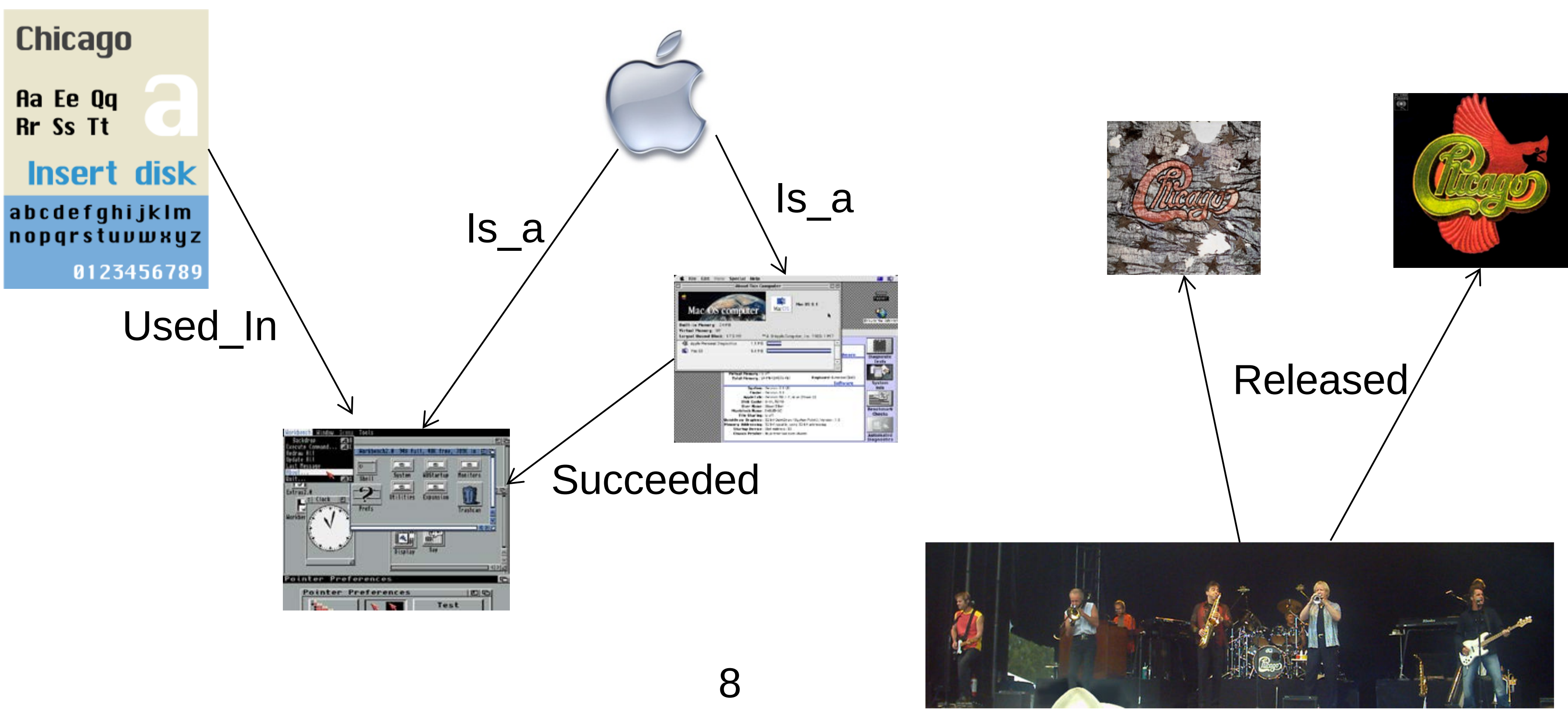






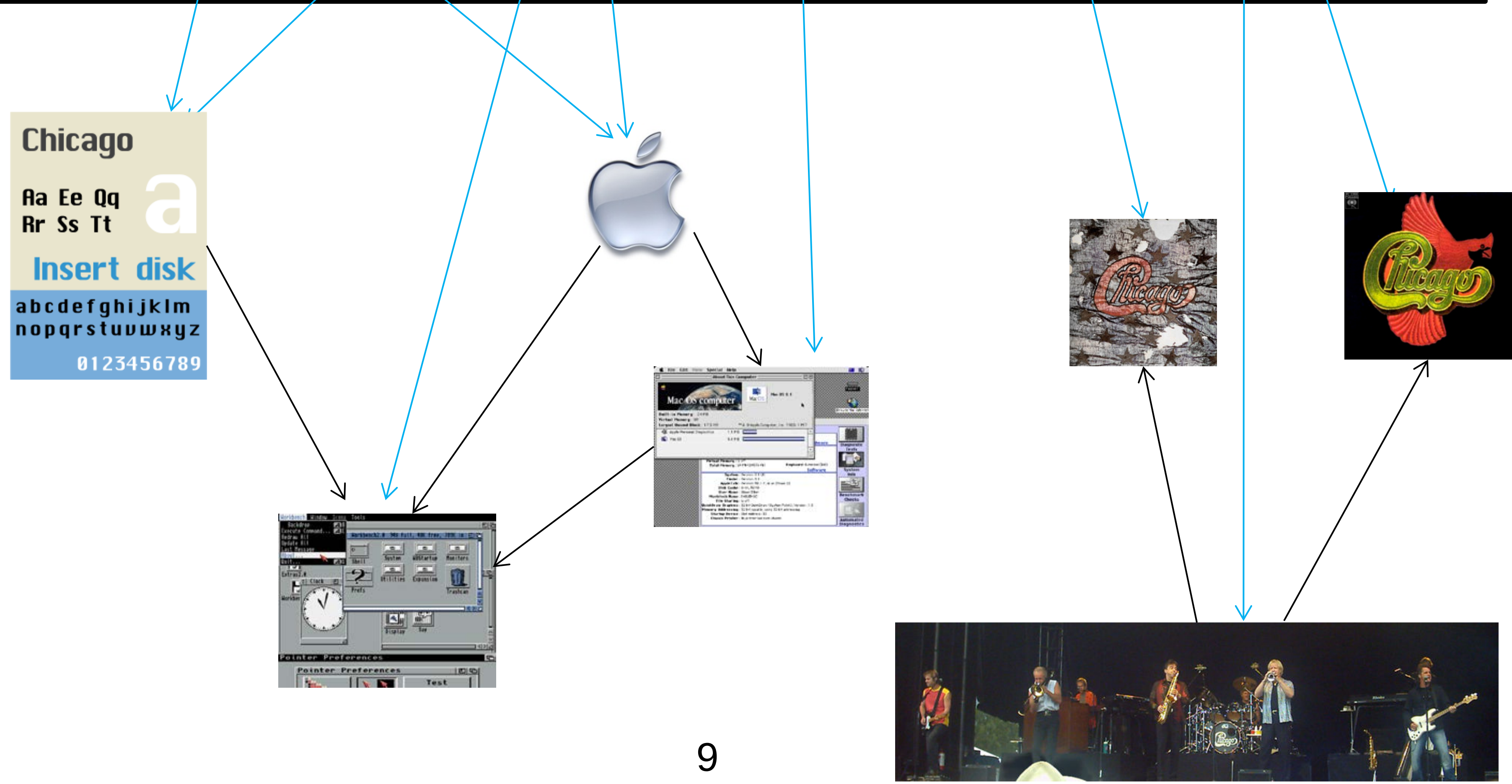
The “Reference” Collection has Structure

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***.

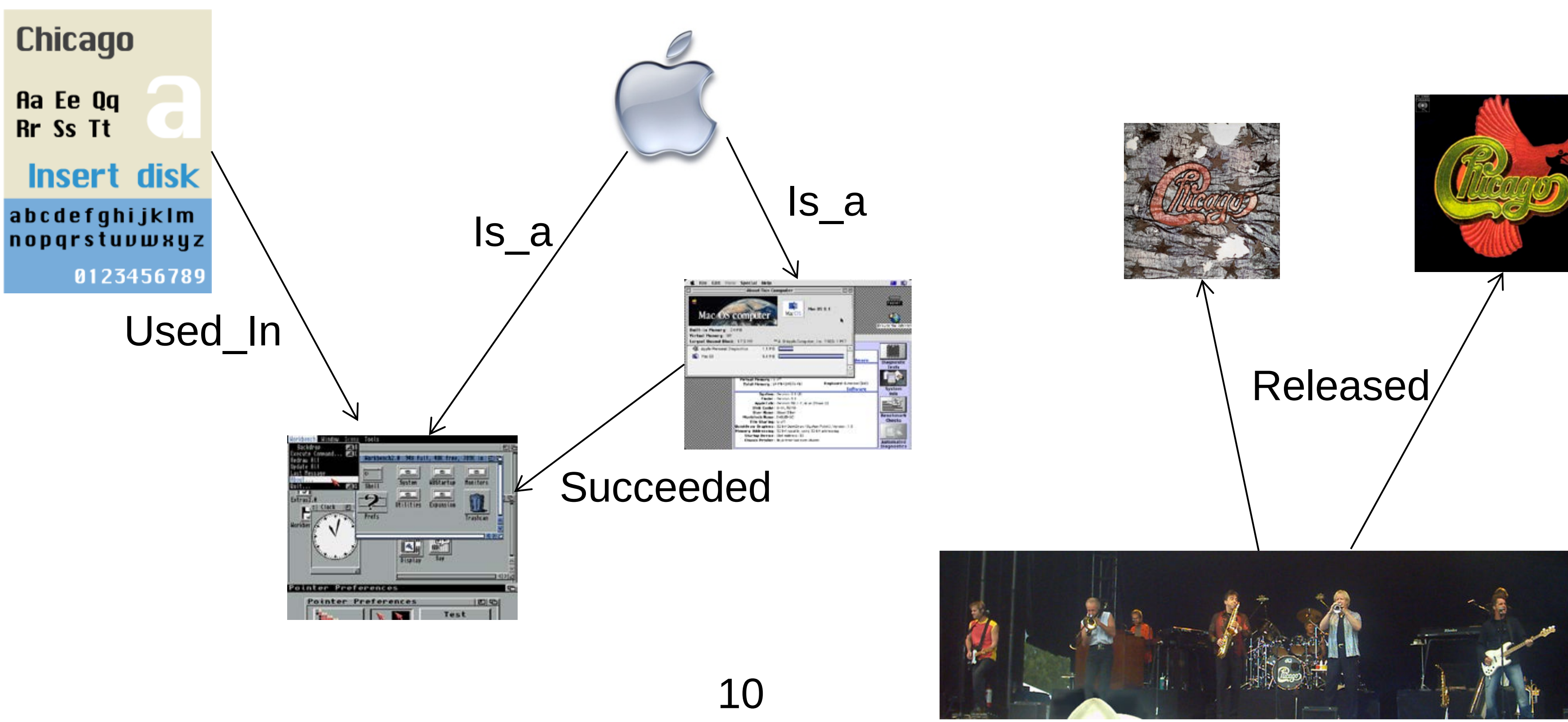


Analysis of Information Networks

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



Wikification: The Reference Problem

Cycles of Knowledge: Grounding for/using Knowledge

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.

Richard Blumenthal
From Wikipedia, the free encyclopedia

Democratic Party (United States)
From Wikipedia, the free encyclopedia

United States Senate
From Wikipedia, the free encyclopedia

[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].

Chris Dodd
From Wikipedia, the free encyclopedia

The New York Times
From Wikipedia, the free encyclopedia

Connecticut
From Wikipedia, the free encyclopedia

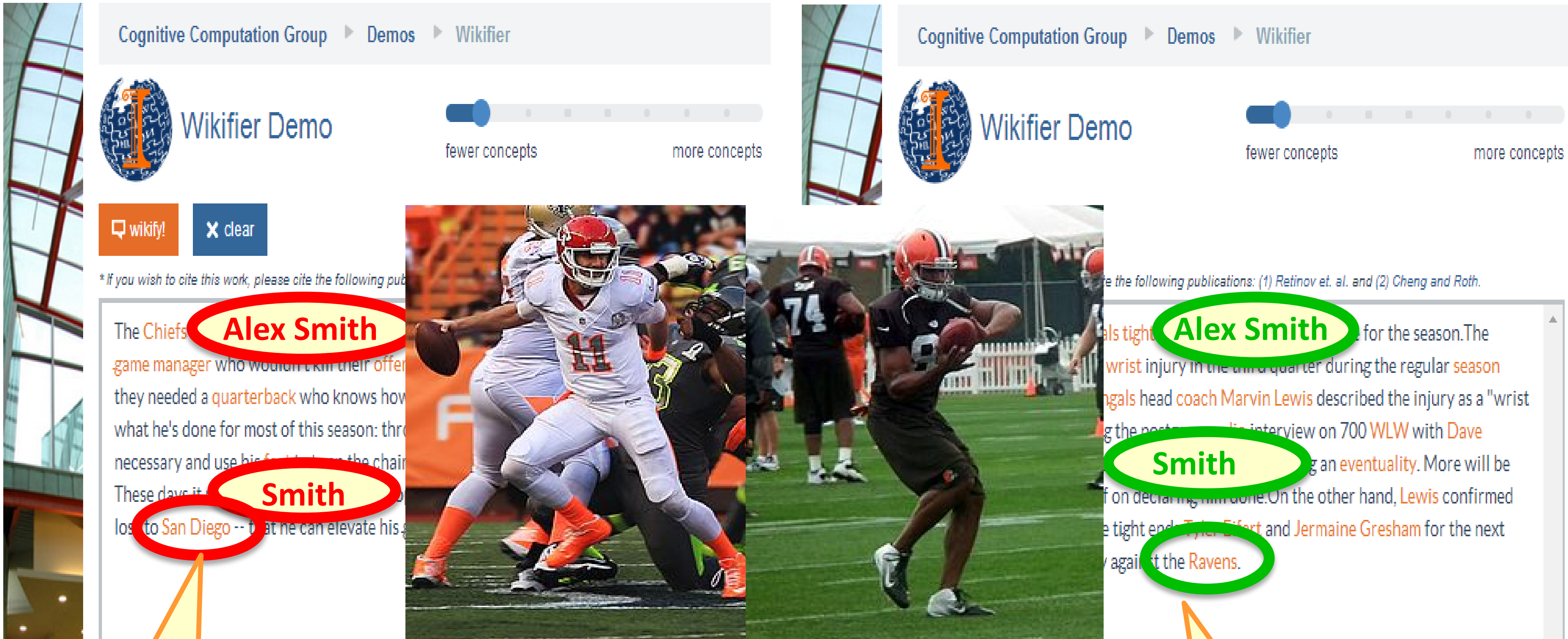
Motivation

- 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)

Who is Alex Smith?



Quarterback of the Kansas City Chief

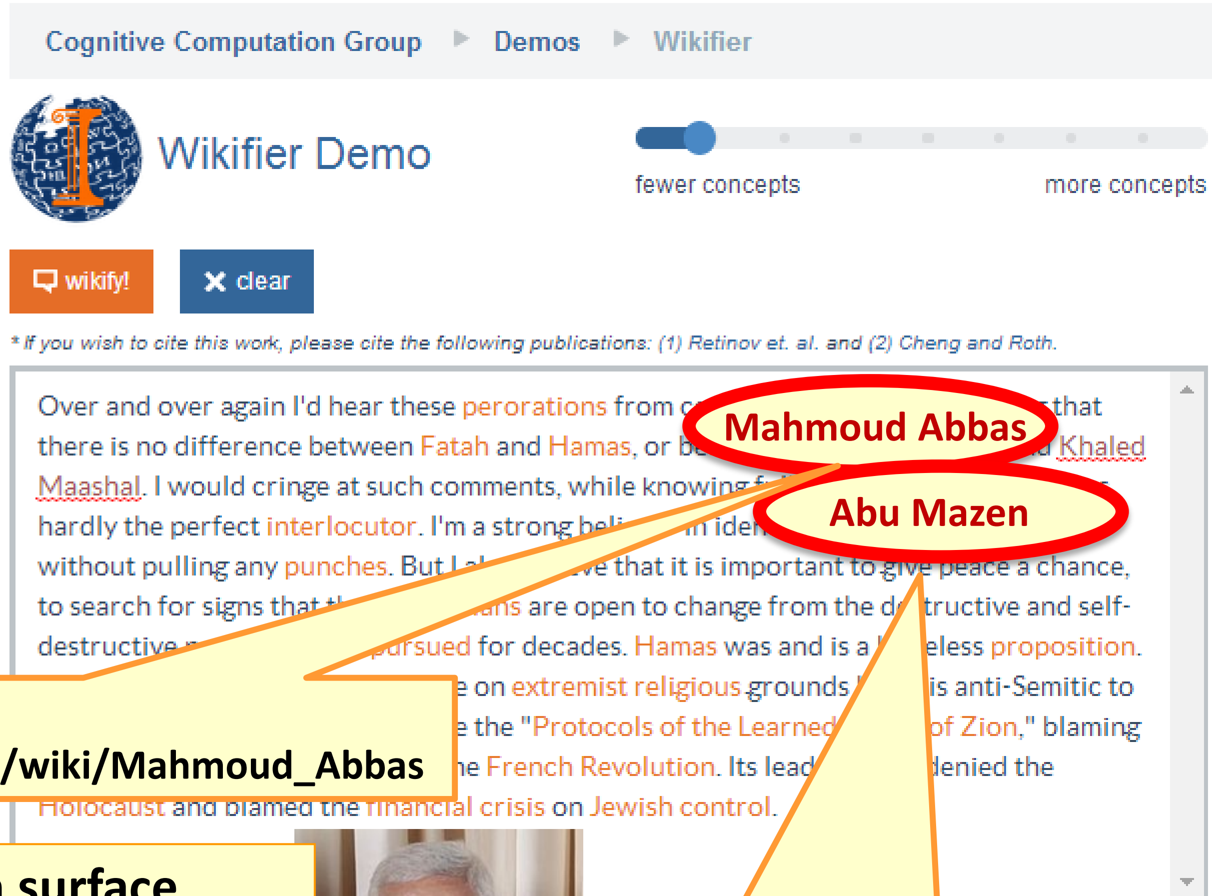
Tight End of the Cincinnati Bengals

San Diego: The San Diego Chargers (A Football team)

Contextual decision on **what is meant** by a given entity or concept. **WSD** with Wikipedia titles as categories.

Ravens: The Baltimore Ravens (A Football team)

Middle Eastern Politics



Mahmoud Abbas

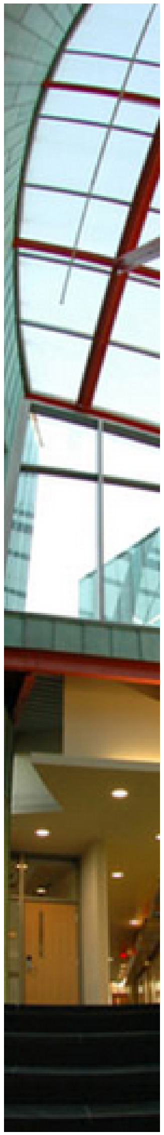
Abu Mazen

Mahmoud Abbas:
http://en.wikipedia.org/wiki/Mahmoud_Abbas

Getting away from **surface representations.**
Co-reference resolution within and across documents, **with grounding**

Abu Mazen:
http://en.wikipedia.org/wiki/Mahmoud_Abbas

Navigating Unfamiliar Domains



Cognitive Computation Group ▶ Demos ▶ Wikifier

Wikifier Demo

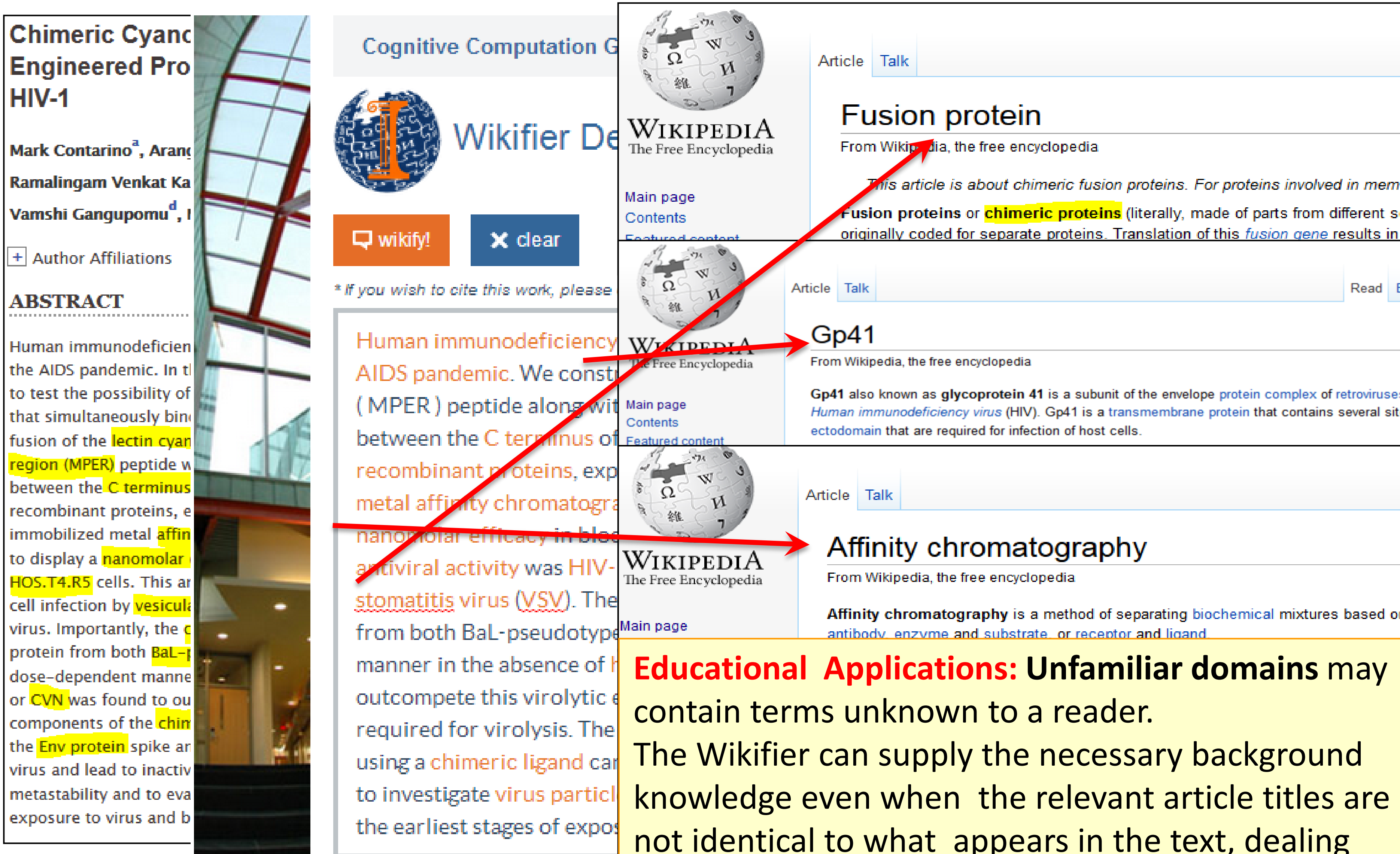
fewer concepts — more concepts

wikify! clear

** If you wish to cite this work, please cite the following publications: (1) Retinov et. al. and (2) Cheng and Roth.*

Human immunodeficiency virus (HIV) is the primary etiologic agent responsible for the AIDS pandemic. We constructed a fusion of the gp41 membrane-proximal external region (MPER) peptide along with a variable-length (Gly4Ser)x linker (where x is 4 or 8) between the C terminus of the former and N terminus of the latter. The His-tagged recombinant proteins, expressed in BL21(DE3)pLysS cells and purified by immobilized metal affinity chromatography followed by gel filtration, were found to display a nanomolar efficacy in blocking BaL-pseudotyped HIV-1 infection of HOS.T4.R5 cells. This antiviral activity was HIV-1 specific, since it did not inhibit cell infection by vesicular stomatitis virus (VSV). The chimeric proteins were found to release intraviral p24 protein from both BaL-pseudotyped HIV-1 and fully infectious BaL HIV-1 in a dose-dependent manner in the absence of host cells. The addition of either MPER or CVN was found to outcompete this virolytic effect, indicating that both components of the chimera are required for virolysis. The finding that engaging the Env protein spike and membrane using a chimeric ligand can destabilize the virus and lead to inactivation opens up a means to investigate virus particle metastability and to evaluate this approach for inactivation at the earliest stages of exposure to virus and before host cell encounter.

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**)

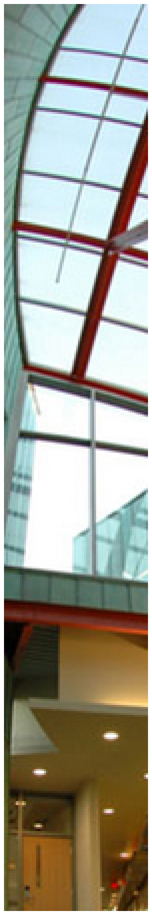
1. Mentions

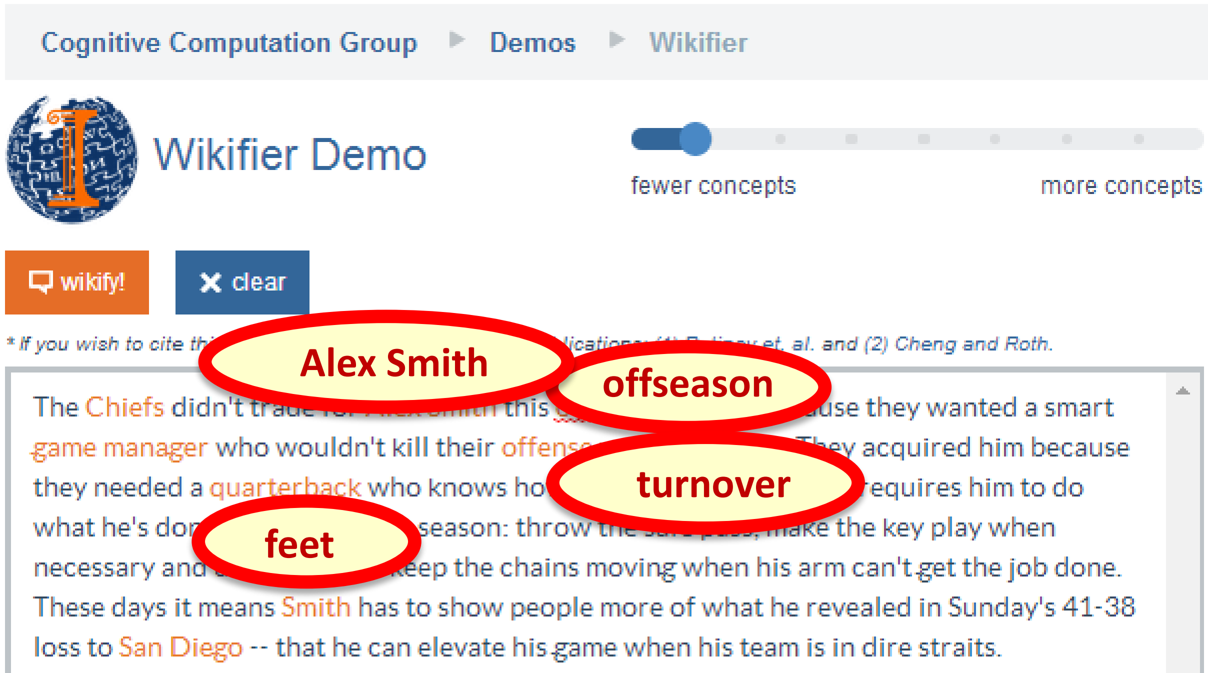
- A mention: a phrase used to refer to something in the world
  - Named entity (person, organization), object, substance, event, philosophy, mental state, rule …
- Task definitions vary across the definition of mentions
  - All N-grams (up to a certain size); Dictionary-based selection; Data-driven controlled 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)



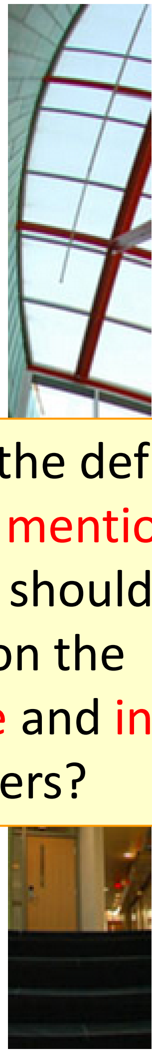


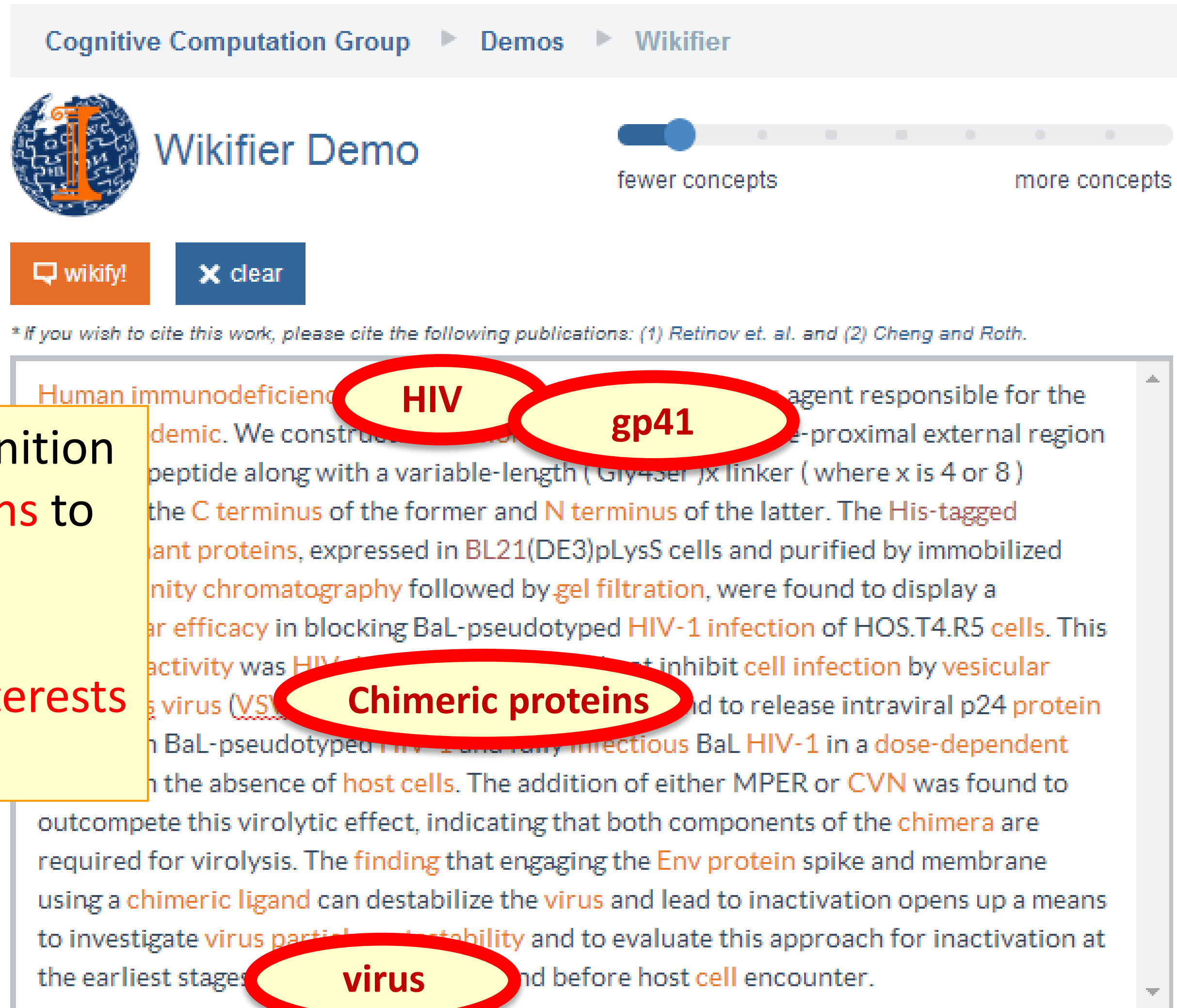
Some task definitions insist on dealing only with mentions that are named entities

How about: *Hosni Mubarak's wife?*

Both entities have a Wikipedia page

Examples of Mentions (3)





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 maintained KB, with significant amount of text 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. What to Link to?

- Often, there are multiple sensible links.

The veteran tight end suffered a wrist injury in the third quarter during the regular season finale against Baltimore. Bengals head coach Marvin Lewis described the injury as a "wrist dislocation".

Baltimore Raven: **Should the link be any different? Both?**

Baltimore: **The city? Baltimore Raven, the Football team? Both?**

The veteran tight e... suffered a wrist injury in the third quarter during the regular season finale against Baltimore Ravens. Bengals head coach Marvin Lewis described the injury as a "wrist dislocation".

Atmosphere: **The general term? Or the most specific one "Earth Atmosphere?**

Earth's biosphere then significantly altered the atmospheric and ... basic physical conditions, which enabled the proliferation of organisms.The atmosphere is composed of

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

- Entity Linking:
  - o Map Mentions to KB Titles
  - o 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]

Evaluation

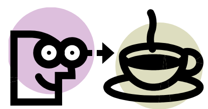
- In principle, evaluation on an application is possible, but hasn't been pursued [with some minor exceptions: NER, Coref]

Factors in Wikification/Entity-Linking Evaluation:

- Mention Selection
  - Are the mentions chosen for linking correct (R/P)
- Linking accuracy
  - Evaluate quality of links chosen per-mention
    - Ranking
    - Accuracy (including NIL)
- NIL clustering
  - Entity Linking: evaluate out-of-KB clustering (co-reference)
- Other (including IR-inspired) metrics
  - E.g. MRR, MAP, R-Precision, Recall, accuracy

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

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.

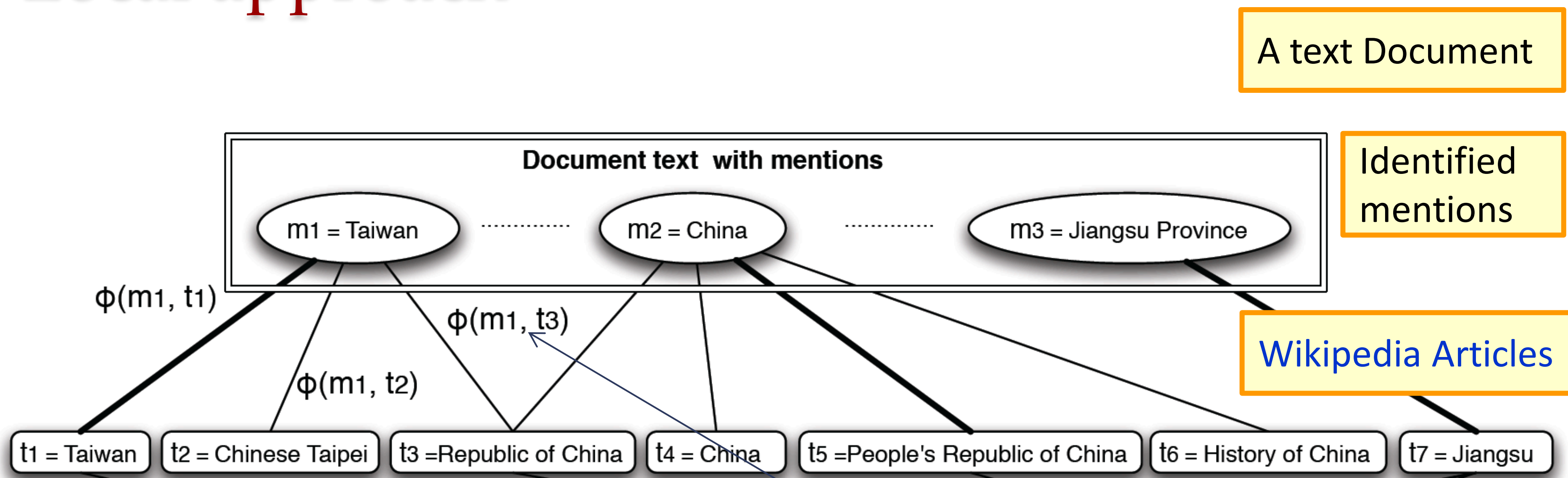
High-level Algorithmic Approach.

- Input: A text document d; Output: a set of pairs (m_i, t_i)
  - o 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
  - 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 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' may also be related]

Local approach



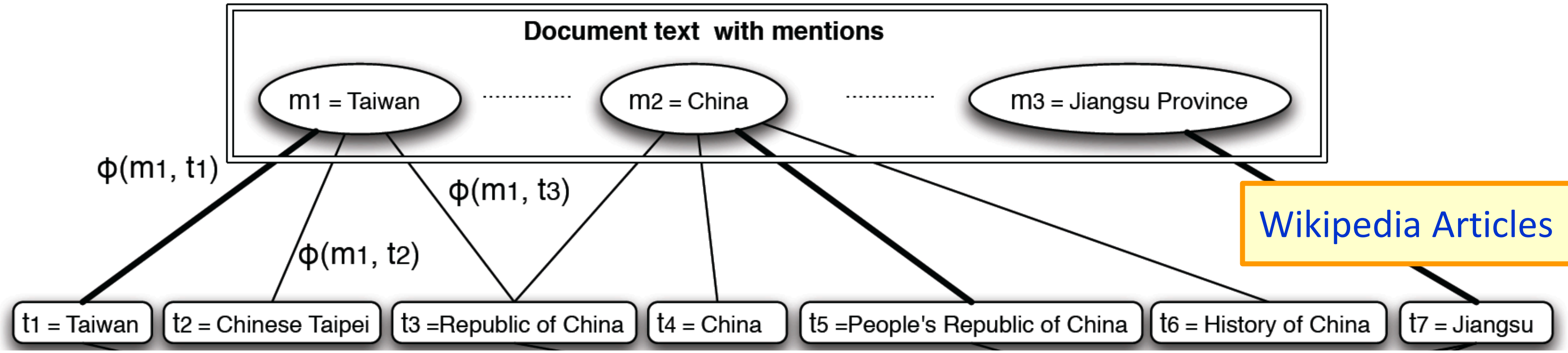
Local score of matching the mention to the title (decomposed by m_i)

- 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

Text Document(s)—News, Blogs,…



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

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)
  - m_i are mentions in d; t_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 of 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' 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)

Mention Identification (Mendes et al., 2012)

L Dictionary-Based chunking (LingPipe) using DBPedia Lexicalization Dataset (Mendes et al., 2011)

LNP Extends L with simple heuristic to isolate NP's

$NPL_{>k}$ Same as LNP but with Statistical NP Chunker

CW Extends L by filtering out common words (Daiber, 2011)

Kea Uses supervised key phrase extraction (Frank et al., 1999)

NER Based on OpenNLP 1.5.1

NER∪NP Augments NER with NPL

Method	P	R	Avg Time per mention
L>3	4.89	68.20	.0279
L>10	5.05	66.53	.0246
L>75	5.06	58.00	.0286
LNP*	5.52	57.04	.0331
NPL*>3	6.12	45.40	1.1807
NPL*>10	6.19	44.48	1.1408
NPL*>75	6.17	38.65	1.2969
CW	6.15	42.53	.2516
Kea	1.90	61.53	.0505
NER	4.57	7.03	2.9239
NER ∪ NP	1.99	68.30	3.1701

Need Mention Expansion

"Michael Jordon"

"His Airness"

"Jordanesque"

Michael Jordan

From Wikipedia, the free encyclopedia

"MJ23"

"Jordan, Michael"

"Michael J. Jordan"

"Corporate Counsel"

"Sole practitioner"

"Defense attorney"

Lawyer

From Wikipedia, the free encyclopedia

"Legal counsel"

"Litigator"

Trial lawyer

"Arizona"

"Alitalia"

AZ

From Wikipedia, the free encyclopedia

"Azerbaijan"

"Authority Zero"

"AstraZeneca"

"Assignment Zero"

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
  - o *“GA ADT 1, USDA, USAID, ADP, Turkish PRT, and the DAIL staff met to create the Wardak Agricultural Steering Committee. “*
  - o “*DST*” = “*District Stability Team*” or “*District Sanitation Technician*”?
  - o *“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"

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

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

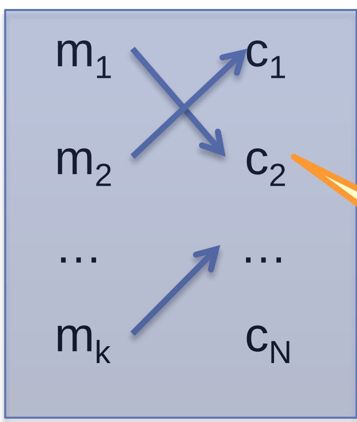
- Local: Mention-Concept Context Similarity
  - Use **similarity measure** to compare the context of the mention with the text associated with a candidate title (the text in the corresponding page)
- Global: Document-wide Conceptual Coherence
  - Use topical/semantic **coherence** measures between the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

Context Similarity Measures

Determine assignment that maximizes pairwise similarity

$$\Gamma^* = \arg\max_{\Gamma} \sum_i \varphi(m_i, t_i)$$

Γ



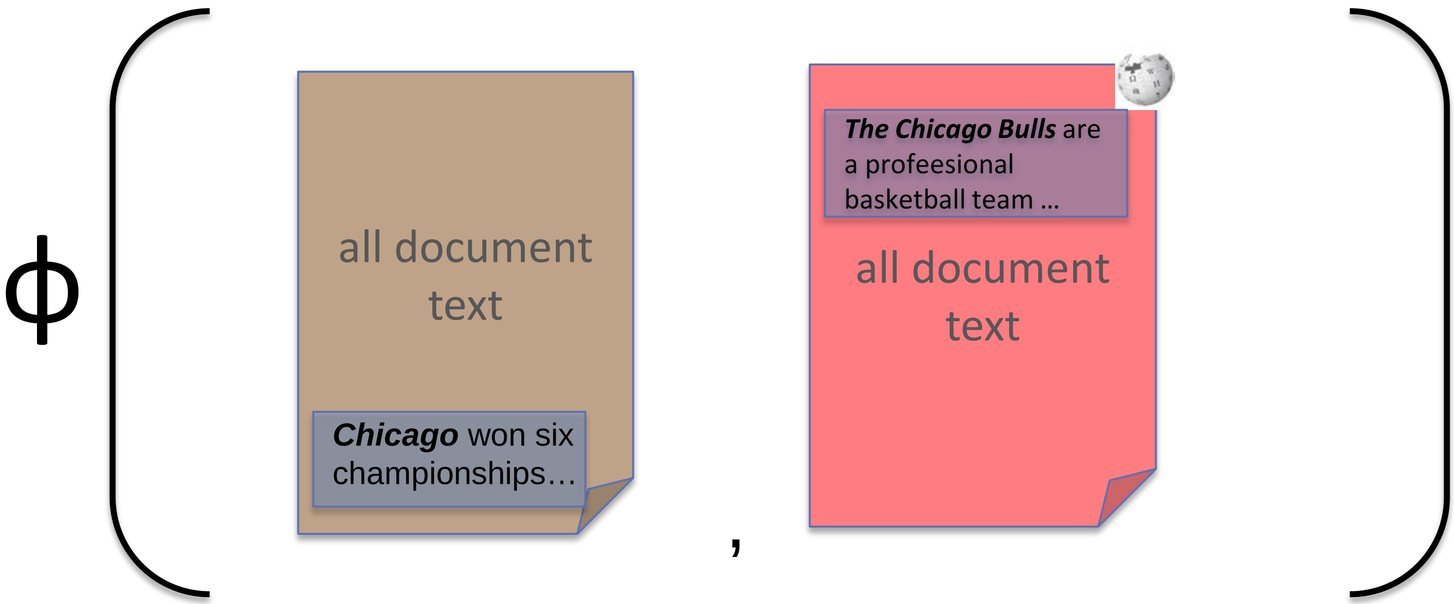
Mention-concept assignment

Mapping from mentions to titles

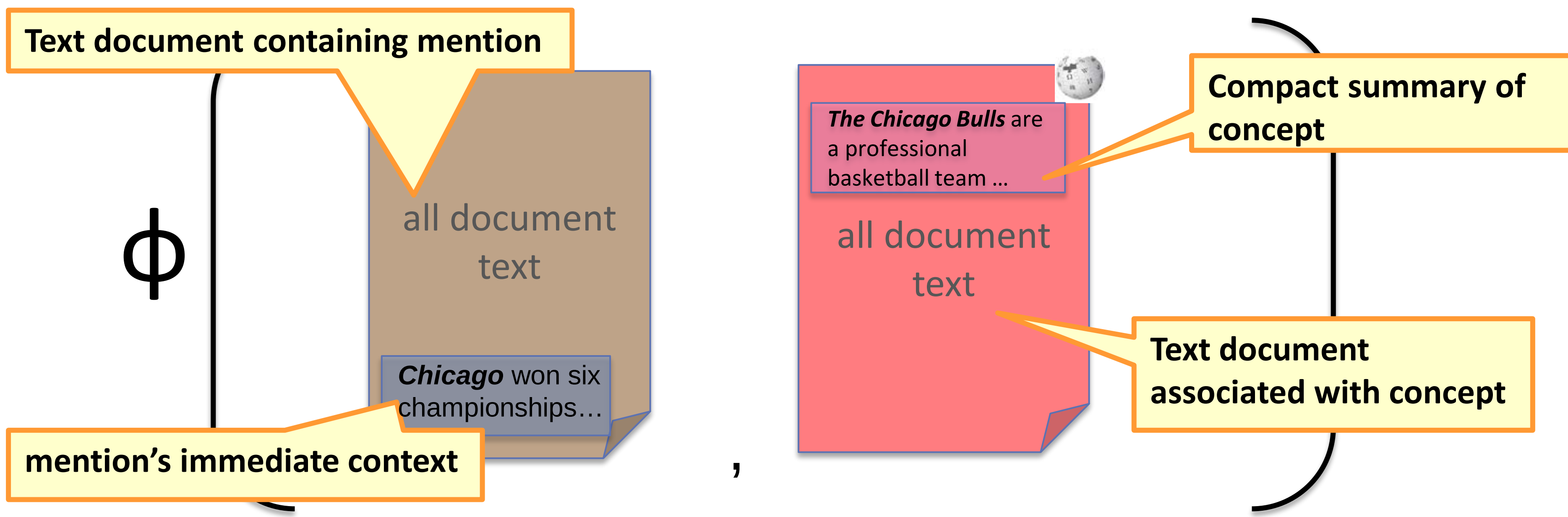
Feature vector to capture degree of contextual similarity

$$\varphi\left(\text{Mention, Title}\right)$$

Context Similarity Measures: *Context Source*

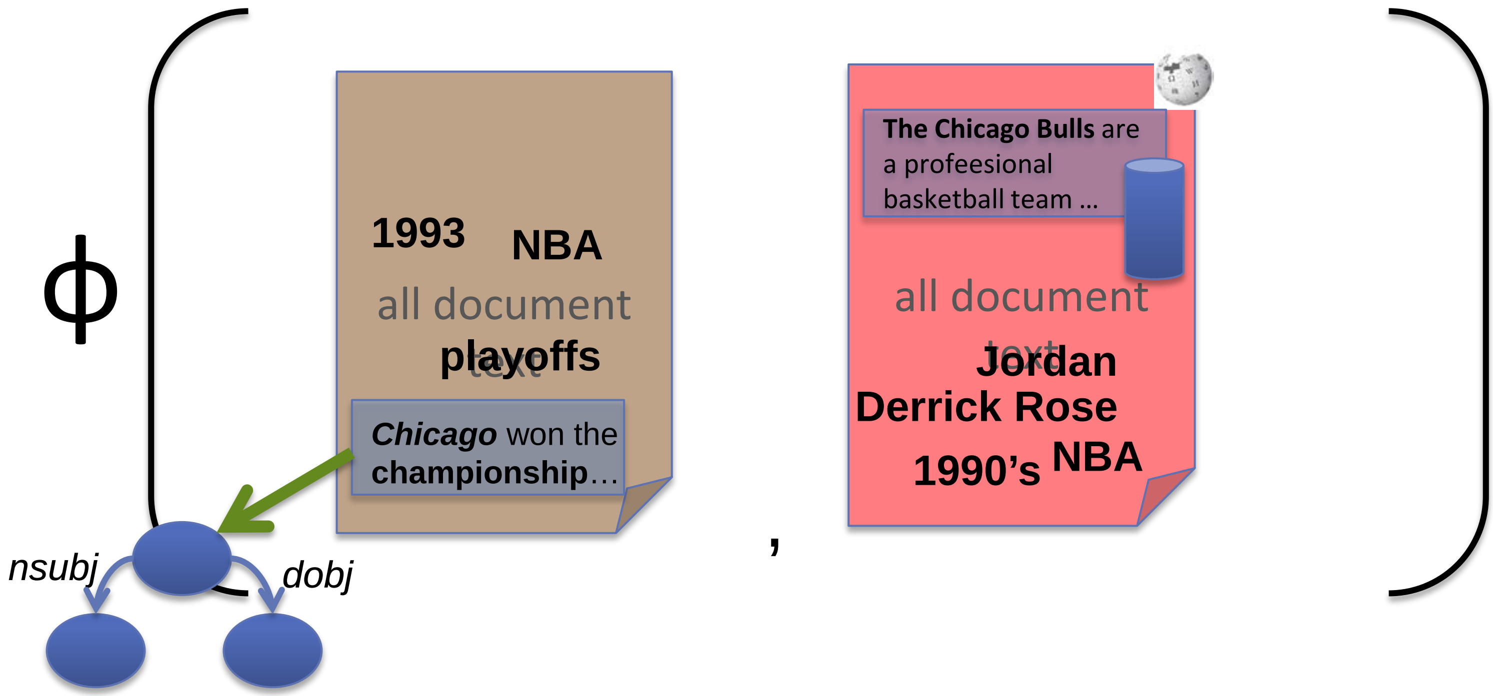


Context Similarity Measures: *Context Source*



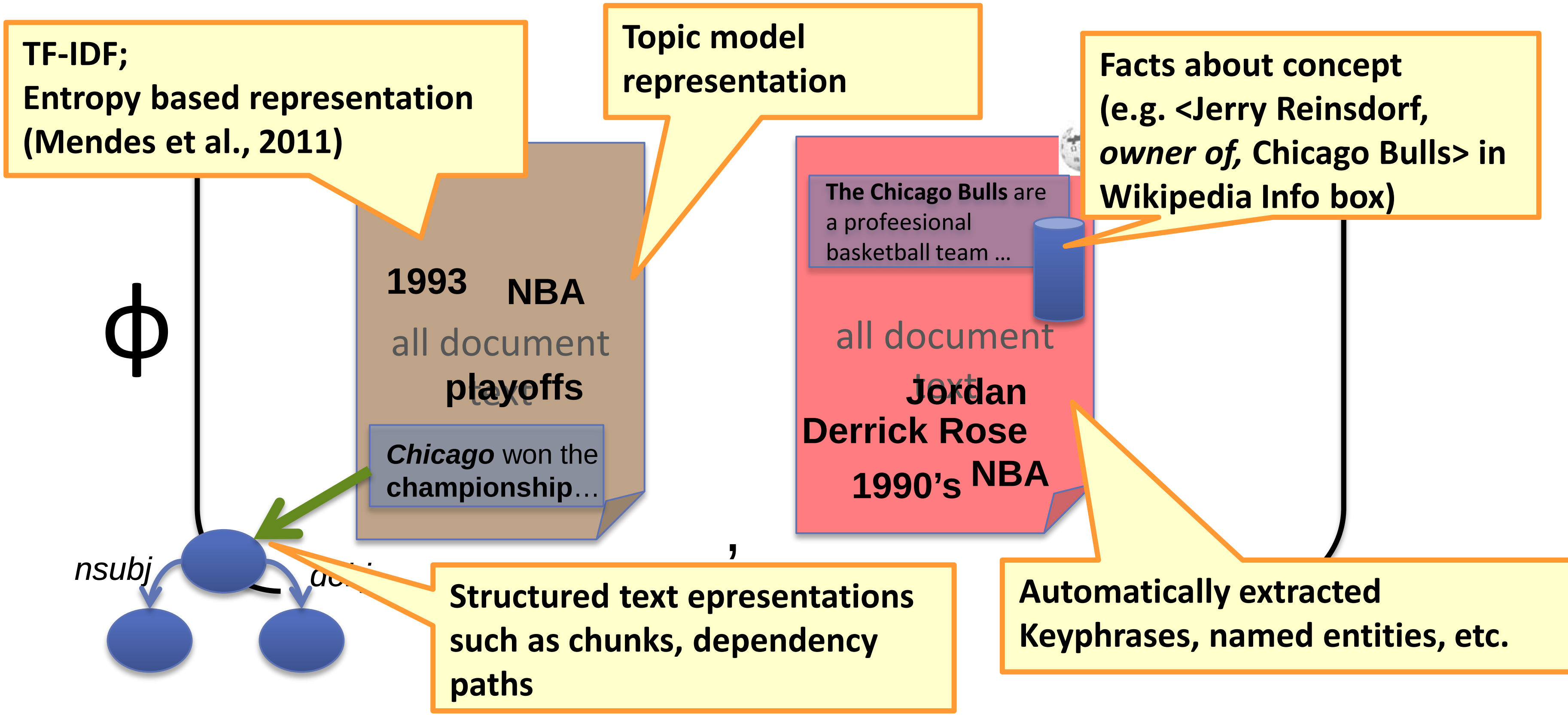
- Varying notion of distance between mention and context tokens
  - Token-level, discourse-level
- Varying granularity of concept description
  - Synopsis, entire document

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

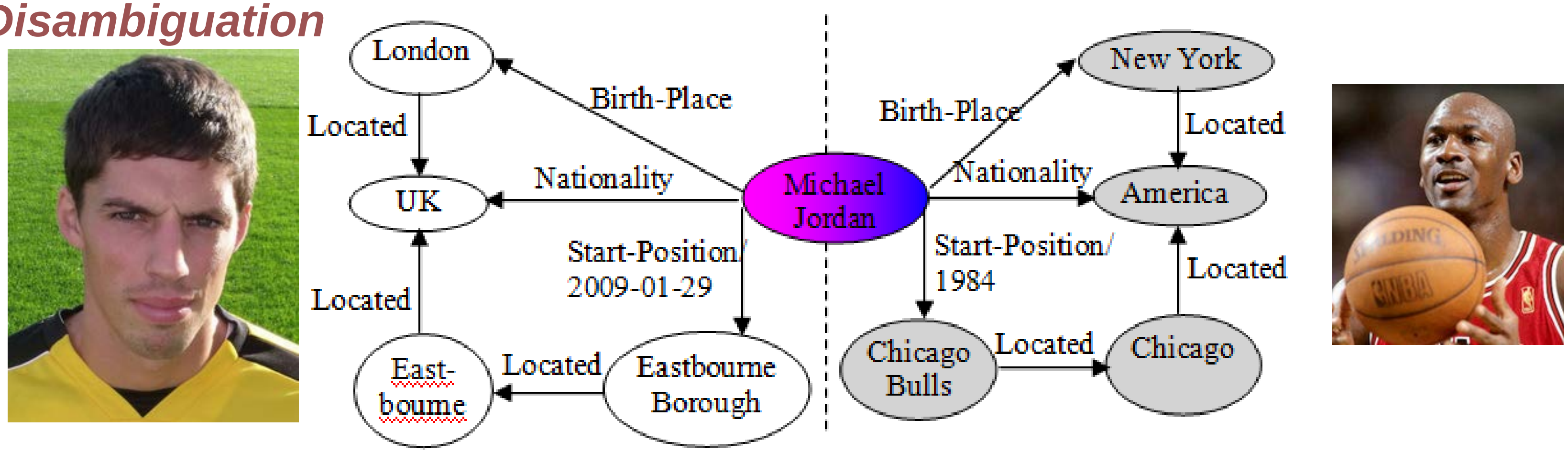
Typical Features for Ranking

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

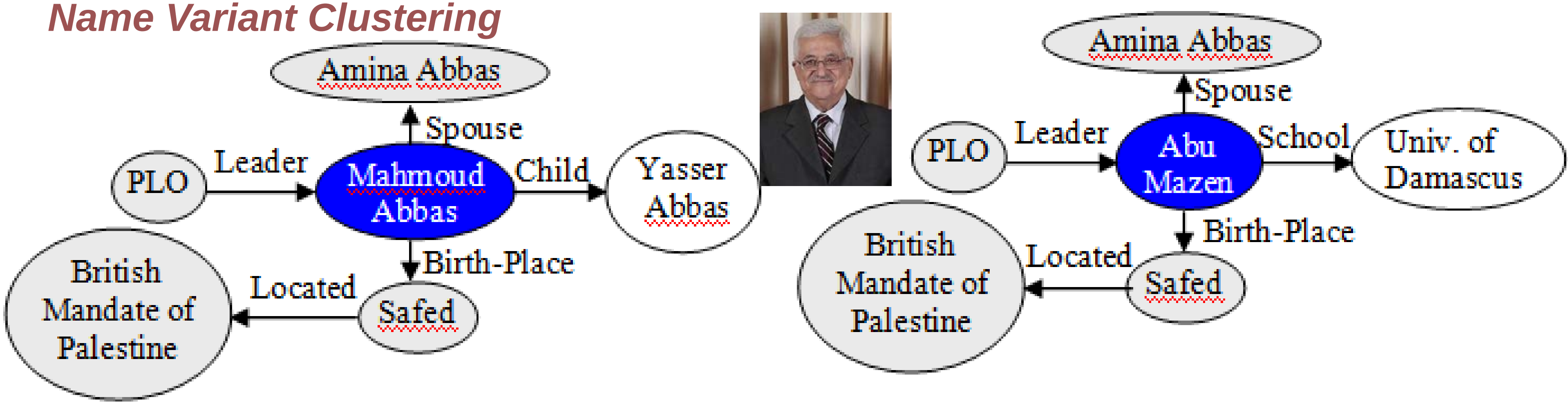
- (Ji et al., 2011; Zheng et al., 2010; Dredze et al., 2010; Anastacio et al., 2011)

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

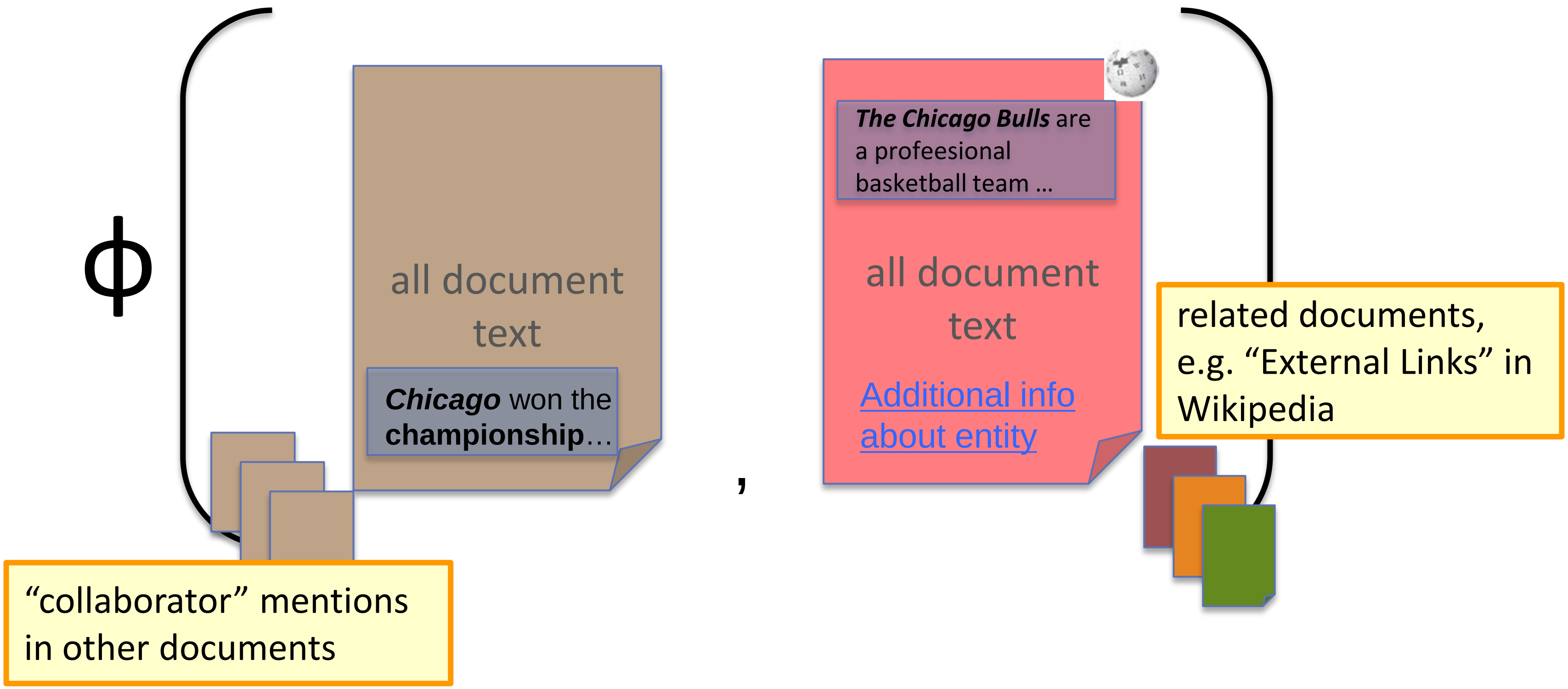
vice president

country

Prime minister

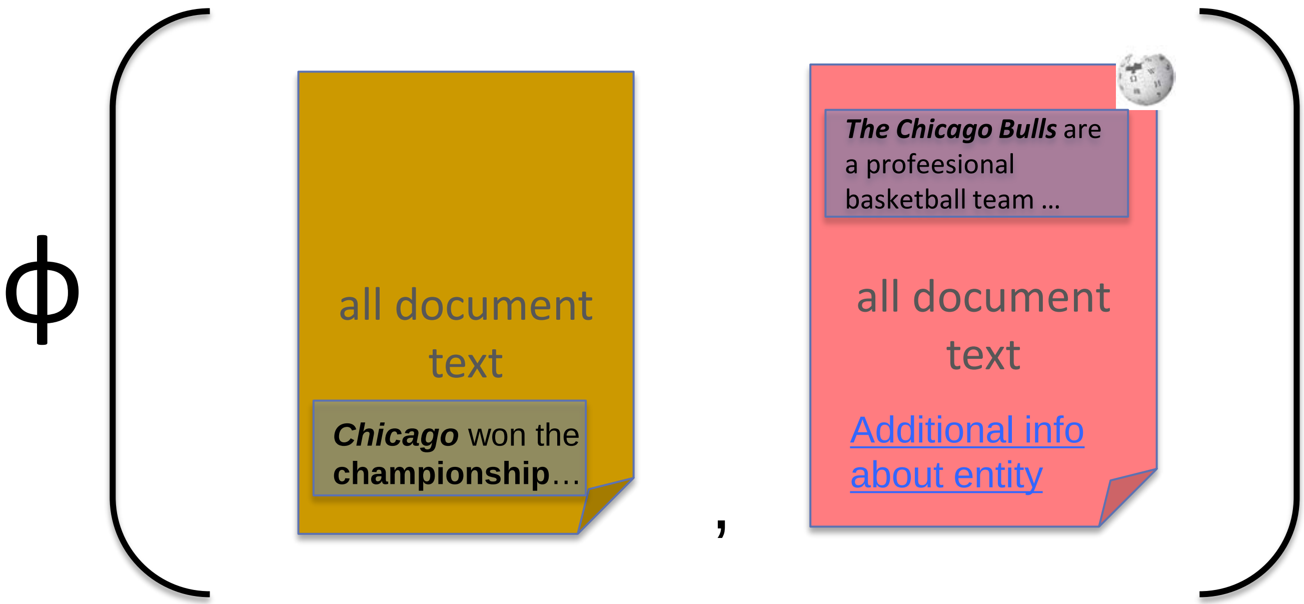
Topical features or topic based document clustering for context expansion (Milne and Witten, 2008; Syed et al., 2008; Srinivasan et al., 2009; Kozareva and Ravi, 2011; Zhang et al., 2011; Anastacio et al., 2011; Cassidy et al., 2011; Pink et al., 2013)

Context Similarity Measures: *Context Expansion*



- Obtain additional documents related to mention
  - Consider mention as information retrieval query
- KB may link to additional, more detailed information

Context Similarity Measures: *Computation*



- Cosine similarity (via TF-IDF)
- Other distance metrics (e.g. Jaccard)
- 2nd order vector composition (Hoffart et al., EMNLP2011)
- Mutual Information

Putting it All Together

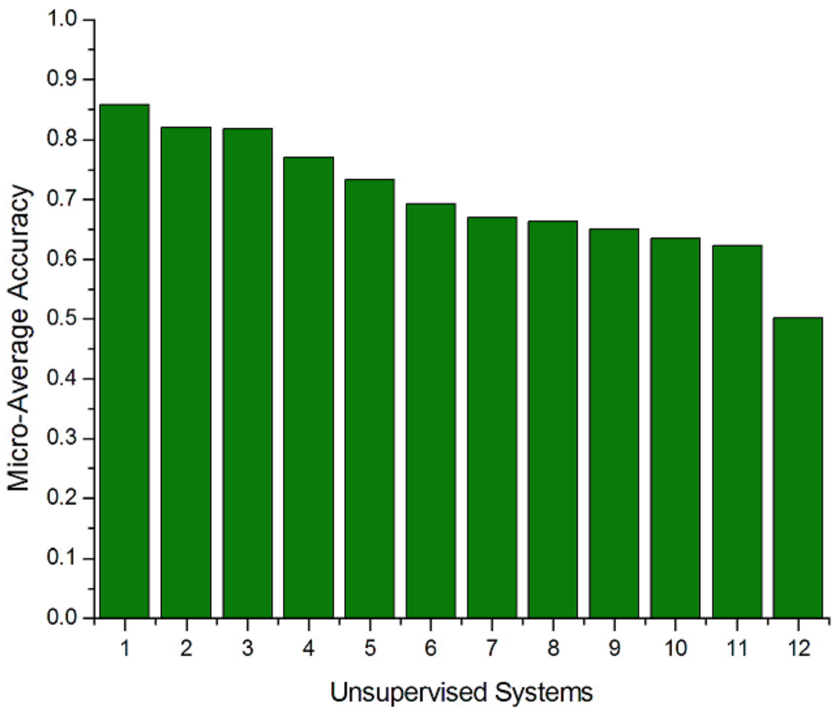
	Score Baseline	Score Context	Score Text
Chicago_city	0.99	0.01	0.03
Chicago_font	0.0001	0.2	0.01
Chicago_band	0.001	0.001	0.02

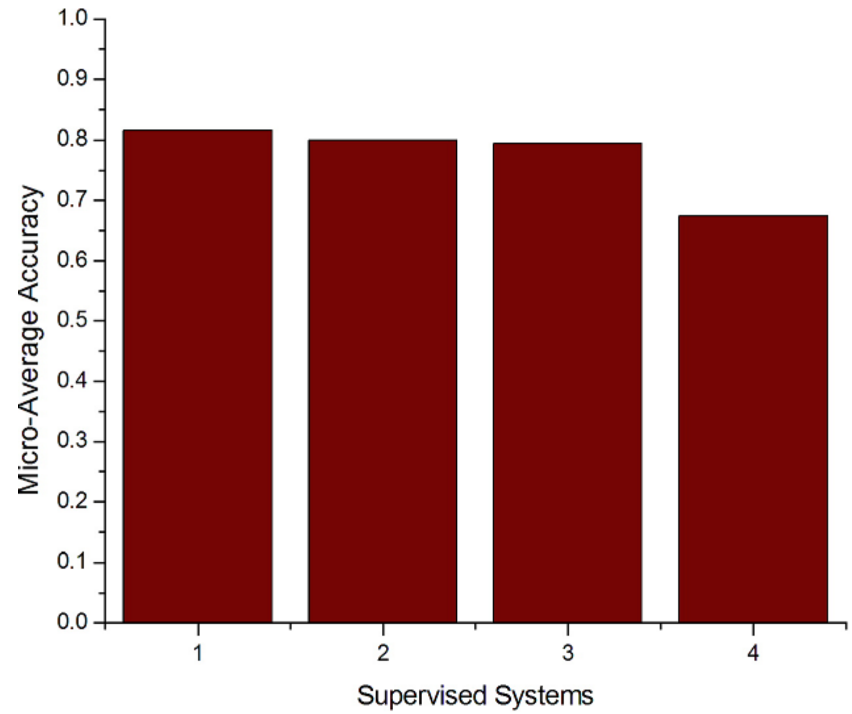
- Learning to Rank [Ratinov et. al. 2011]
  - Consider all pairs of title candidates
    - Supervision is provided by Wikipedia
  - Train a ranker on the pairs (learn to prefer the correct solution)
  - A Collaborative Ranking approach: outperforms many other learning approaches (Chen and Ji, 2011)

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 $t_i \in T(m_i)$

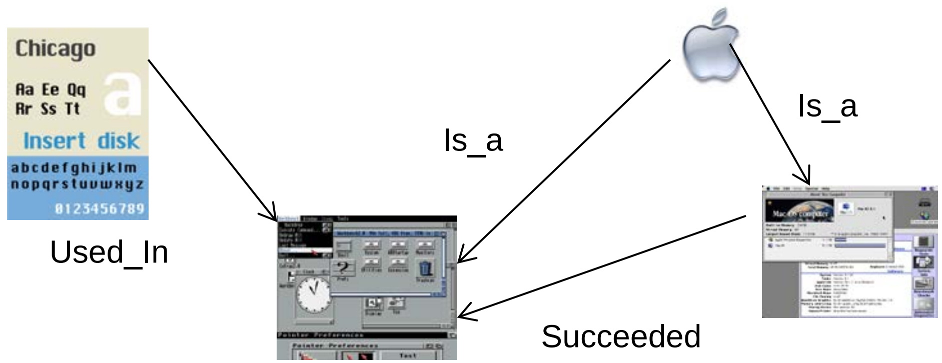
    [E.g., consider local statistics of edges (m_i, t_i), $(m_i\ *)$, and $(*, t_i)$ occurrences of in the Wikipedia graph]
- (3) Global Inference
  - o For each document d:
    - Consider all $m_i \in d$; and all $t_i \in T(m_i)$
    - Re-rank titles $t_i \in T(m_i)$

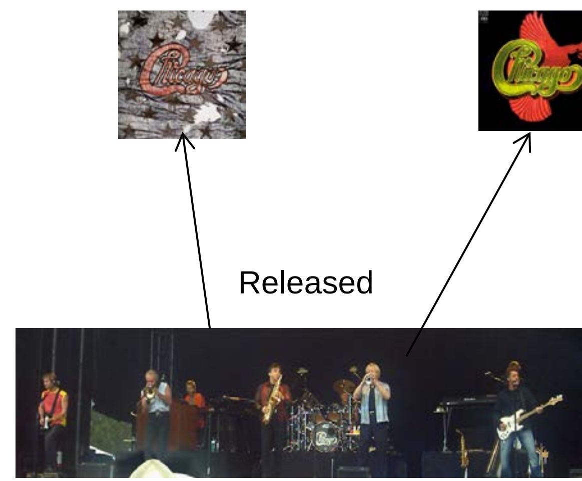
    [E.g., if m, m' are related by virtue of being in d, their corresponding titles t, t' should also be related]

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:
  - o Textual co-occurrence of concepts is reflected in the KB (Wikipedia)
- Incite:
  - o 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),

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

See García et al. (JAIR2014) for variational details

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

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: $\chi^2(a,b) = (\|in(b) \cap in(a)\| \cdot (\|W\| - \|in(b) \cup in(a)\|) + -\|in(b) \setminus in(a)\| \cdot \|in(a) \setminus in(b)\|)^2 \cdot \cdot \frac{\|W\|}{\|in(a)\| \cdot \|in(b)\| (\|W\| - \|in(a)\|)(\|W\| - \|in(b)\|)}$
$\chi^2_{\mathsf{out}}(a,b)$	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
$KL(\mathrm{c}\|e)$	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
$KL(\mathrm{e}\|c)$	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

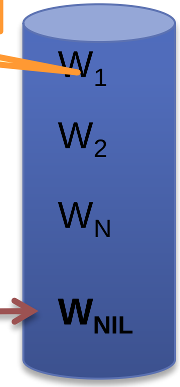
- The key difference between Wikification and Entity Linking is the way NIL are treated.
- In Wikification:
  - Local Processing
  - Each mention m_i that does not correspond to title t_i is mapped to NIL.
- In Entity Linking:
  - Global Processing
  - Cluster all mentions m_i that represent the same 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

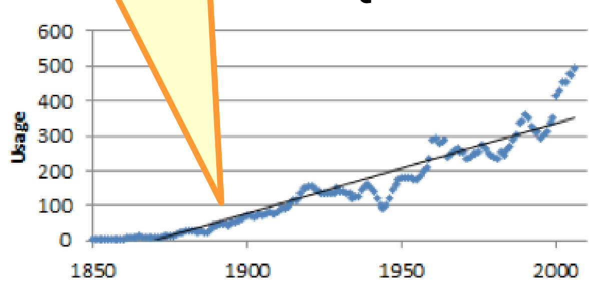
Is it an *entity*?

- 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"

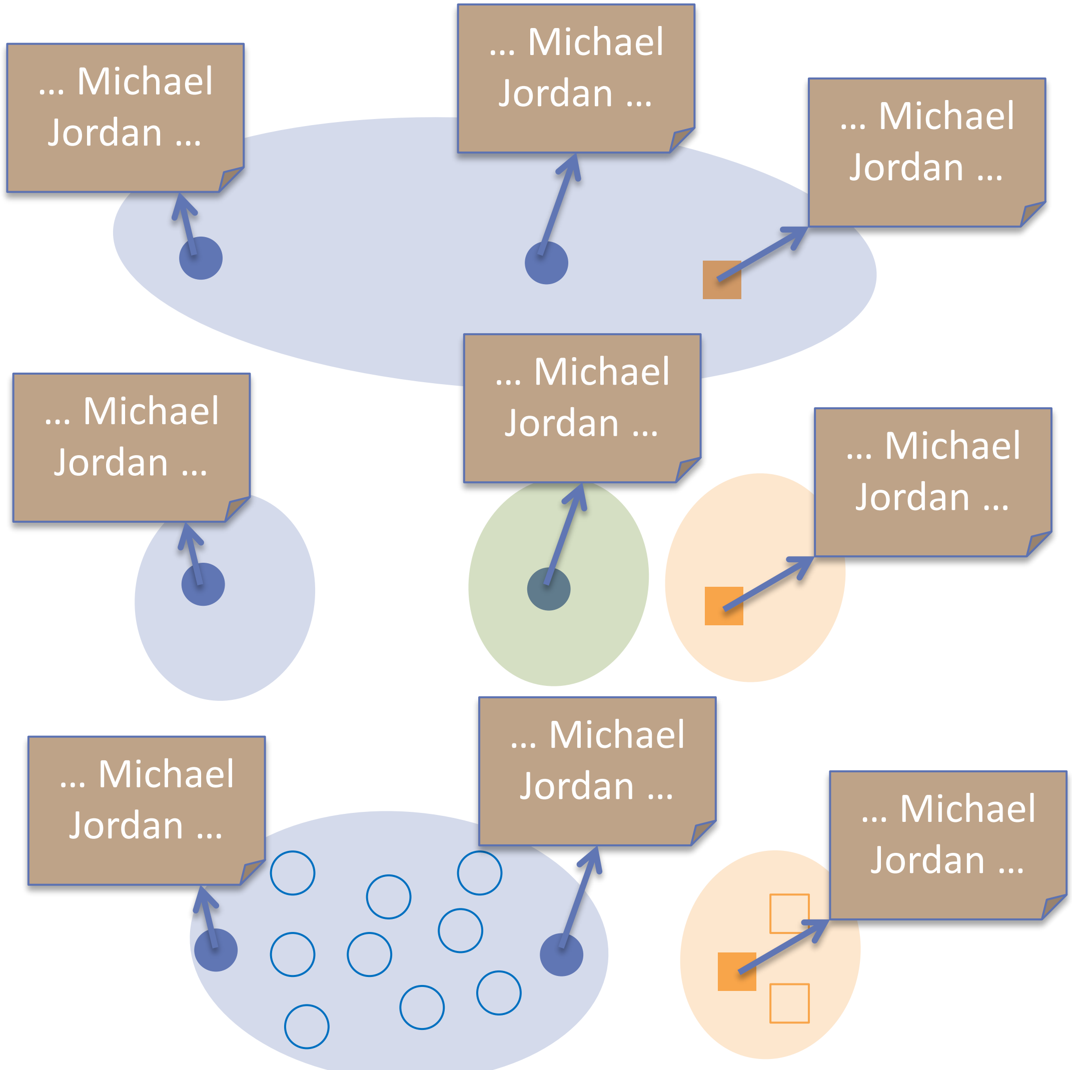
Simple string matching

"One in one"

Often difficult to beat!

Collaborative Clustering

Most effective when ambiguity is high



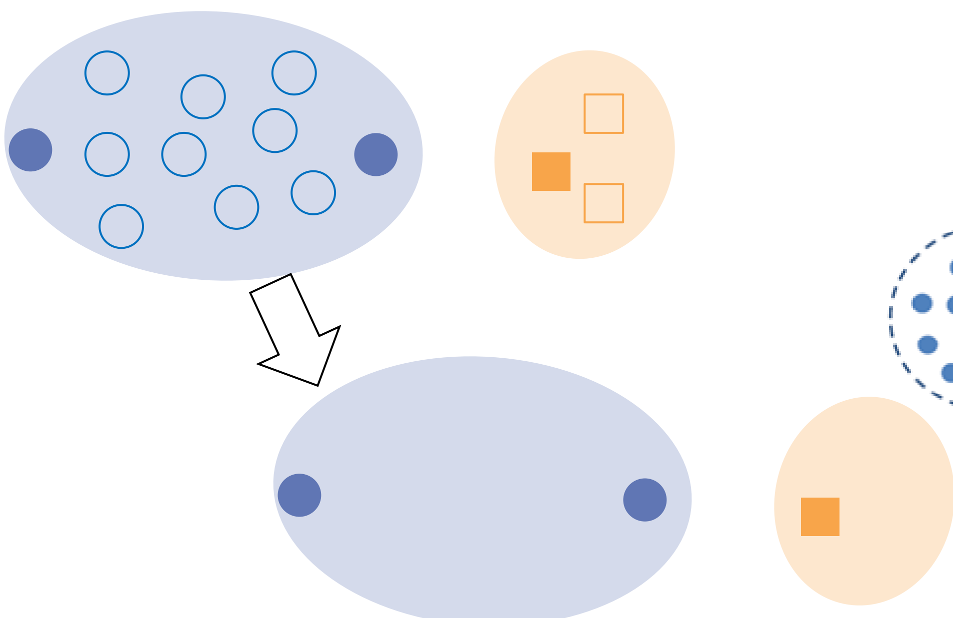
NIL Clustering Methods Comparison (Chen and Ji, 2011; Tamang et al., 2012)

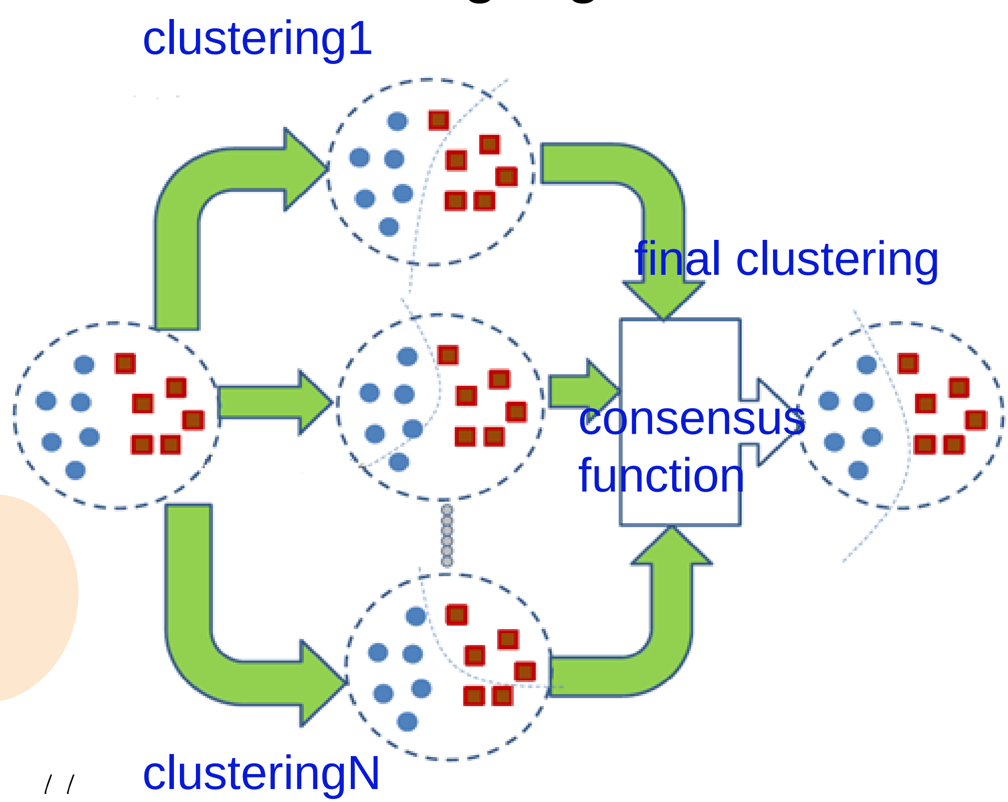
Algorithms		B-cubed+ F-Measure	Complexity
Agglomerative clustering	3 linkage based algorithms (single linkage, complete linkage, average linkage) (Manning et al., 2008)	85.4%-85.8%	$O(n^2)$ $O(n^2 \log n)$ n: the number of mentions
	6 algorithms optimizing internal measures cohesion and separation	85.6%-86.6%	$O(n^2 \log n)$ $O(n^3)$
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^3M 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

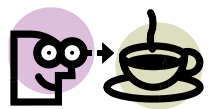




Outline

- 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

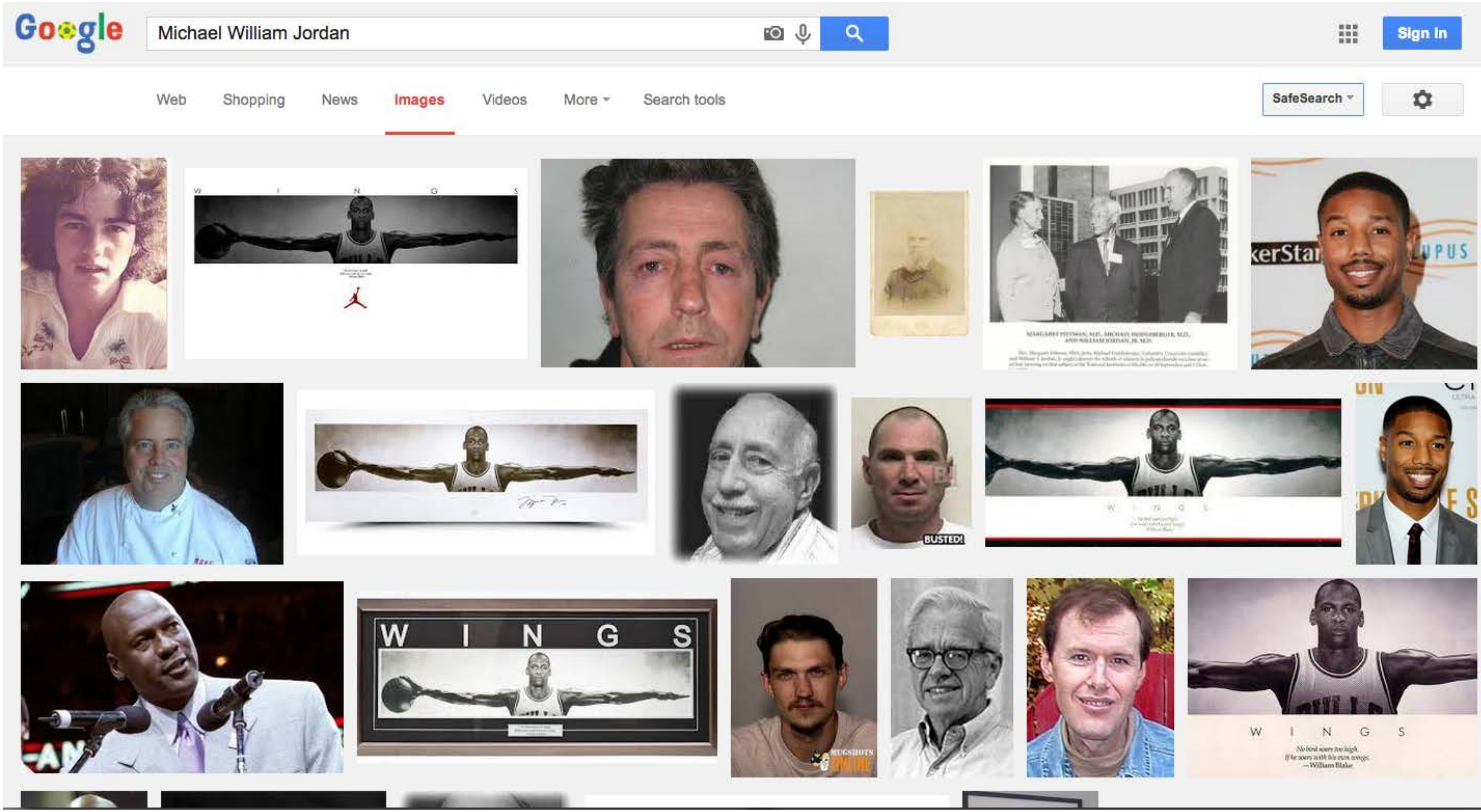
> Blumenthal (D) is a candidate for the U.S. Senate seat now held by Christopher Dodd (D), and he has held a commanding lead in the race since he entered it. But the Times report has the potential to fundamentally reshape the contest in the Nutmeg State.

- 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?

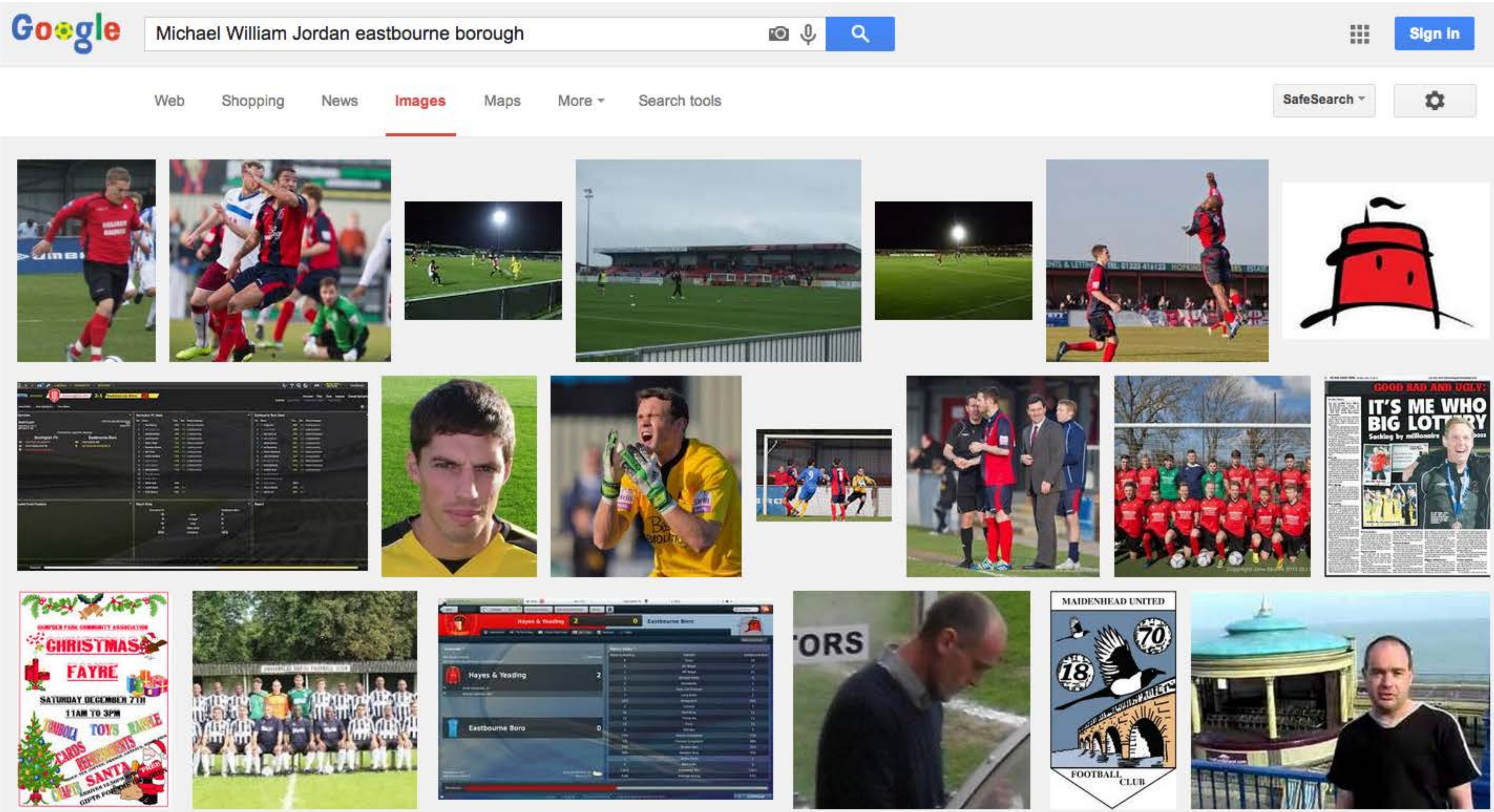
Challenges

- 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…



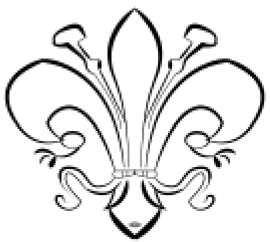
A Little Better…



Deep Semantic Knowledge

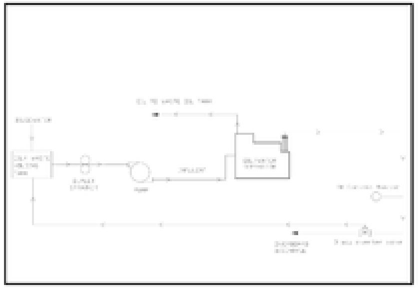
Local **OWS** activists were part of this **protest**.

Order of World Scouts



Occupy Wall Street

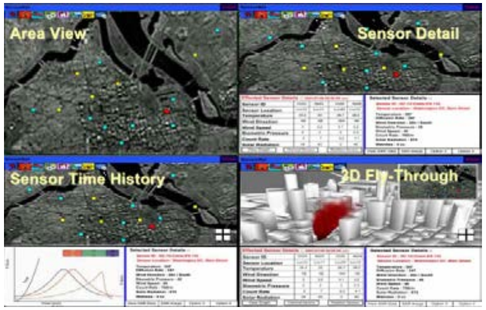
Oily Water Separator



Overhead Weapon Station

Open Window School

Open Geospatial Consortium



Deep Semantic Knowledge

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***

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.

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

From Wikipedia, the free encyclopedia

Hsiung Feng can refer to:

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

Hsiung Feng IIE

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
…

Background Knowledge

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

Interstate 55 in Missouri

Commonsense Knowledge

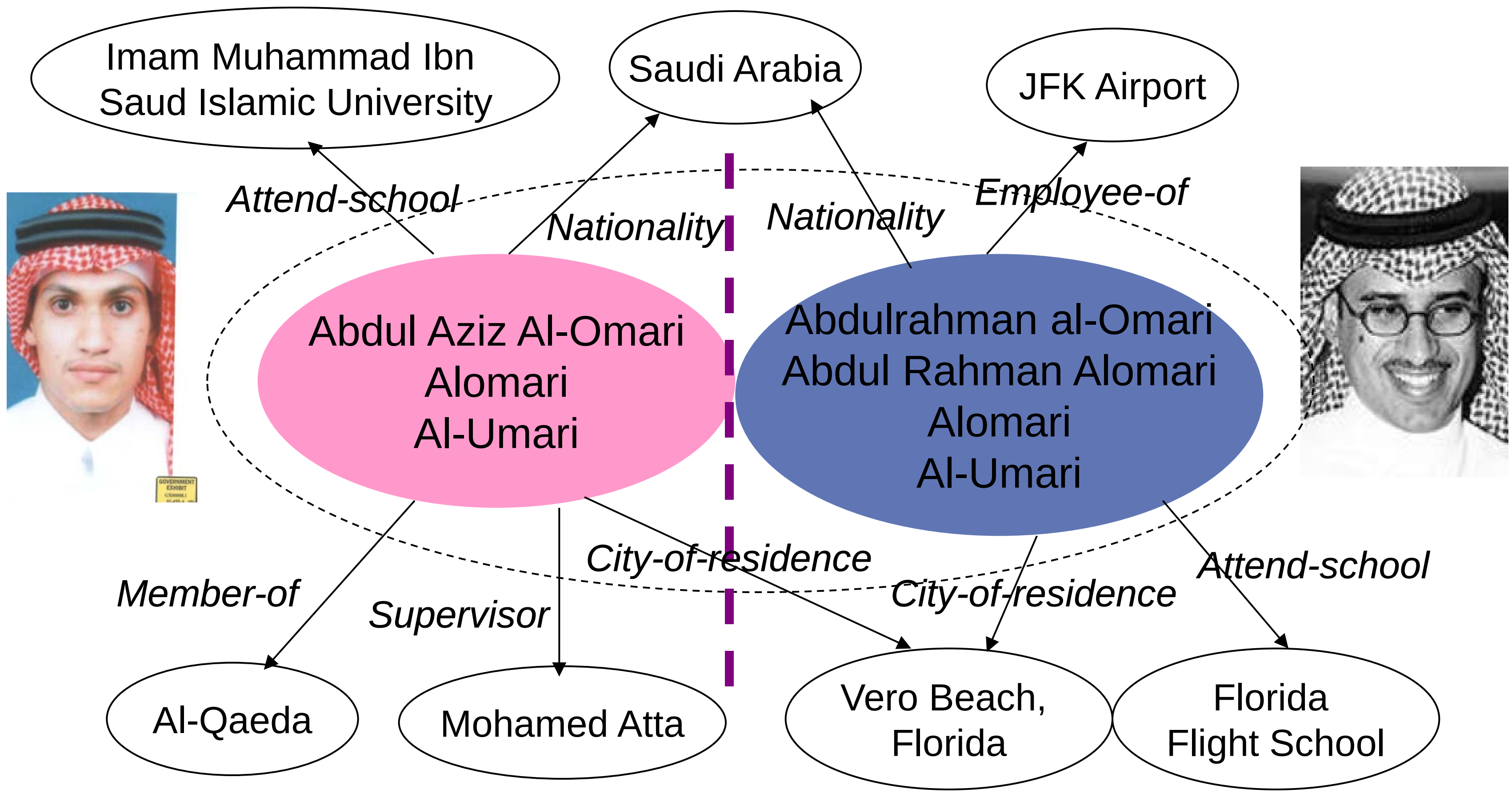
2008-07-26

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)



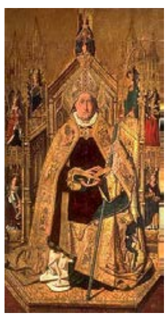
Rich Context

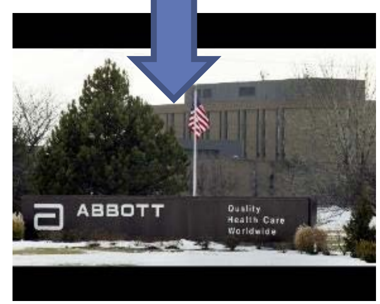
- “Supreme Court” (in Japan, China, U.S., Macedonia, etc.)
- “LDP (Liberty and Democracy Party)” (in Australia, Japan, etc.)
- “Newcastle University” can be located in UK or Australia
- Many person entities share the same common names such as “Albert”, “Pasha”, etc.
- “Ji county” can be located in “Shanxi”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





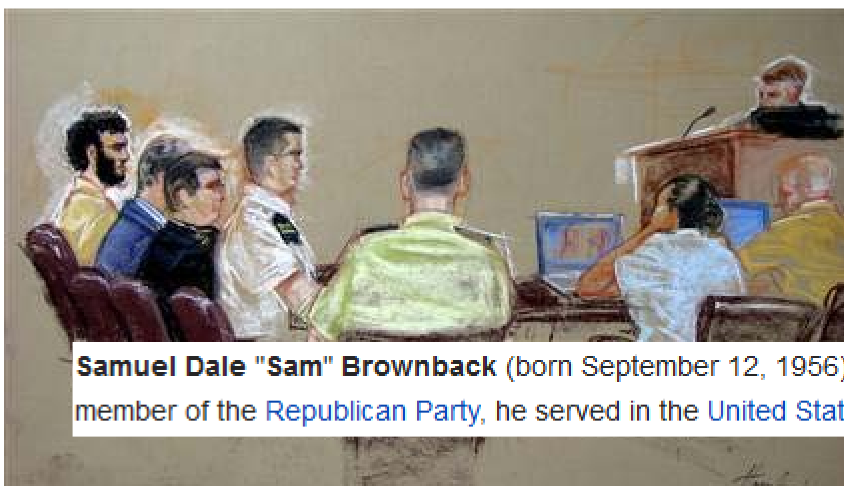
- Finding collaborators based cross-document entity clustering (Chen and Ji, 2011)

Rich Context: Related Employer

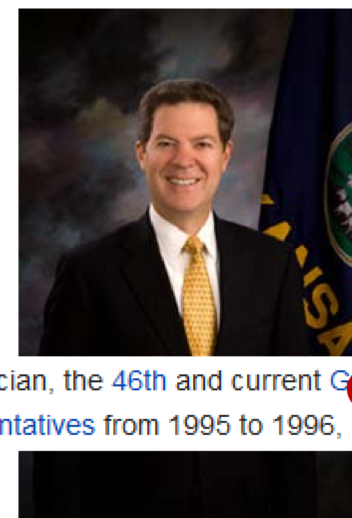
Hundreds of protesters from various groups converged on the state capitol in Topeka, Kansas today…

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



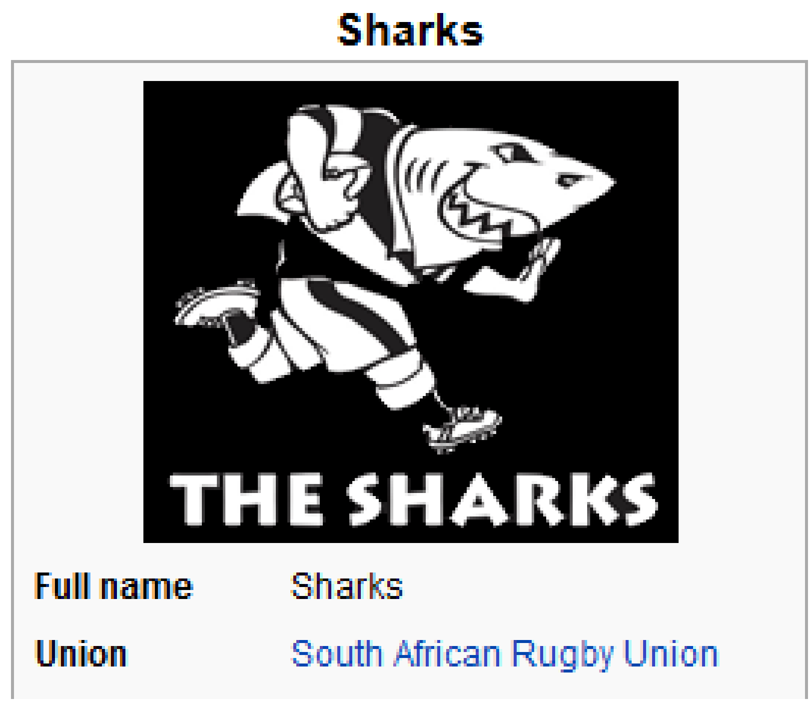
Sam Brownback



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 (i/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

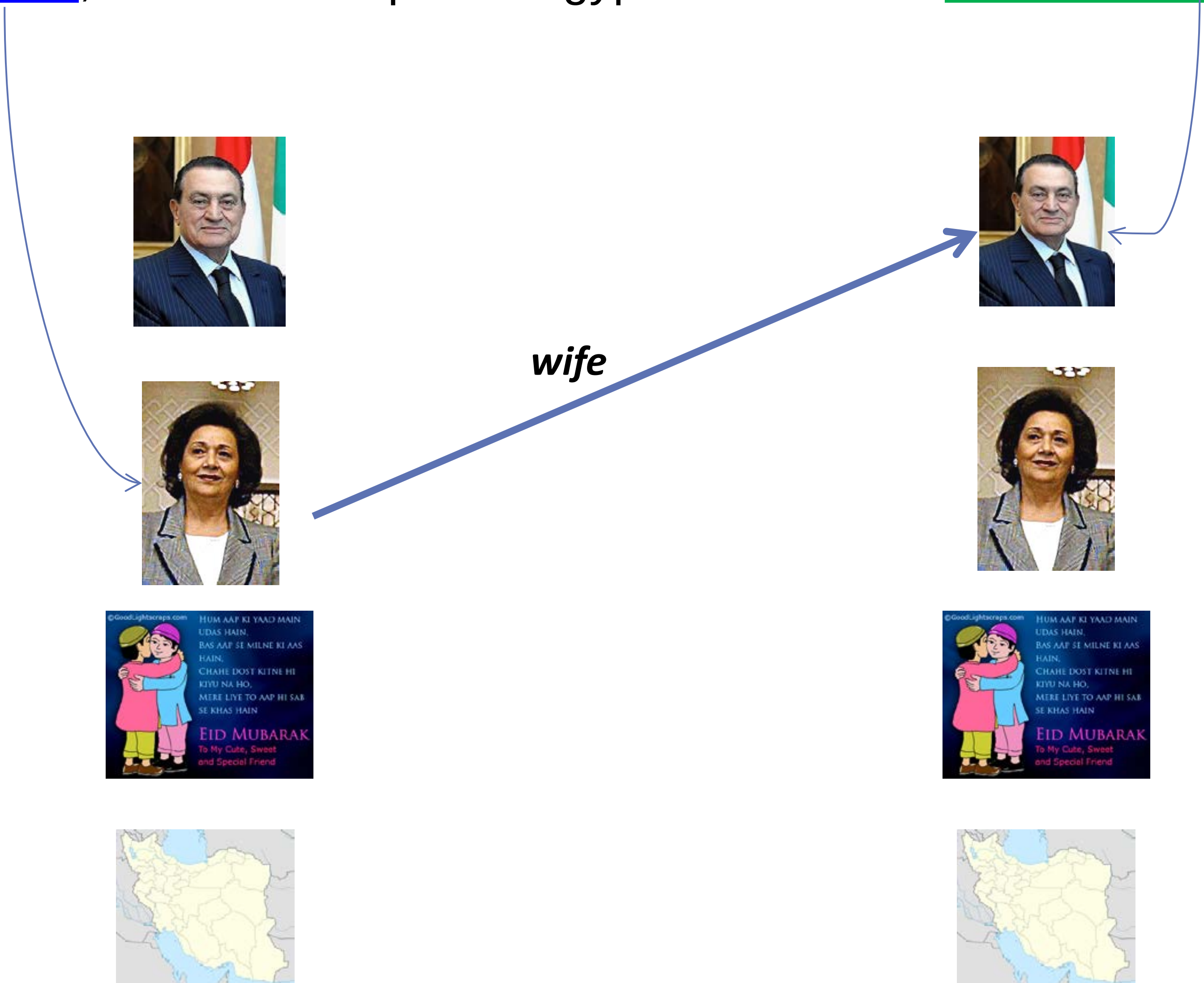
Rich Context: Related Colleagues

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.

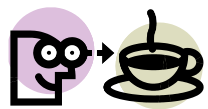
Rich Context: Related Family Members

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



Outline

- Motivation and Definition
- A Skeletal View of a Wikification System
    - 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.

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

Collaborative Title Collection on KB

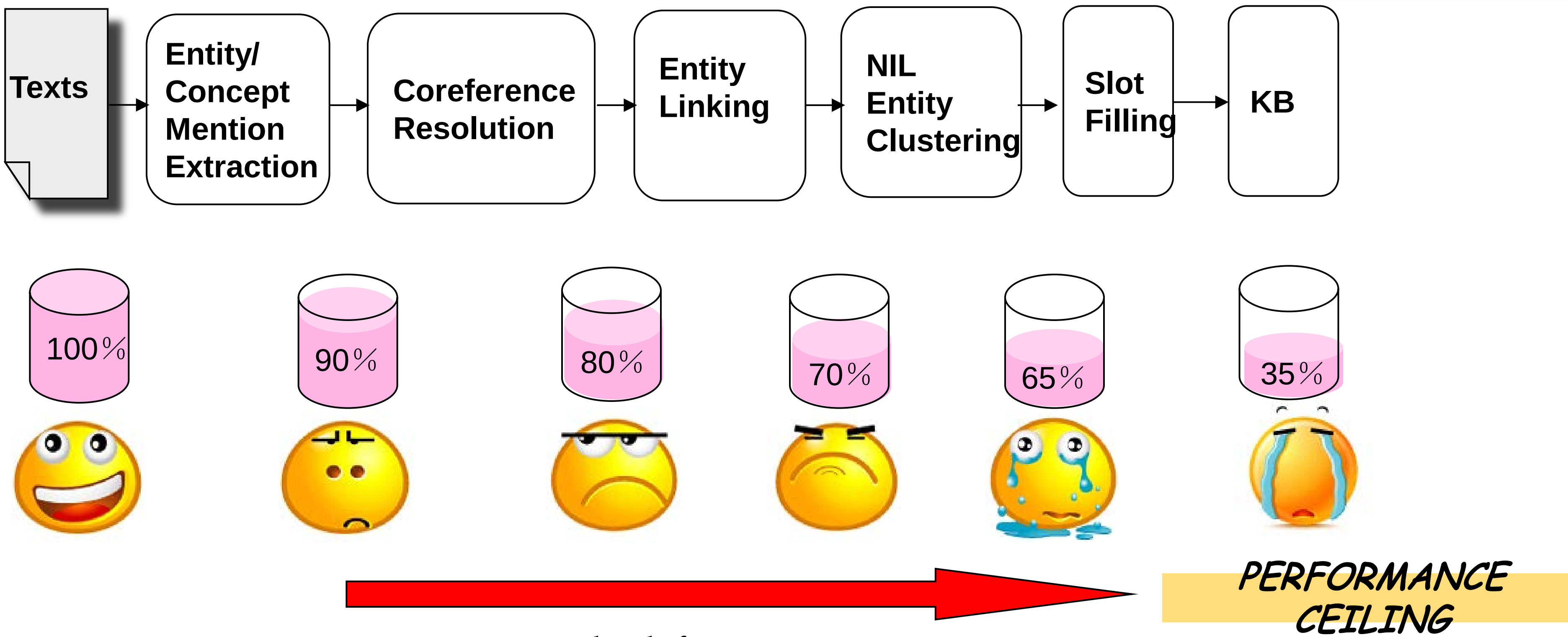
- Go beyond Wikipedia: Exploit rich structures in DBPedia, Freebase, YAGO, Ontologies
- Google Knowledge base: “people also search for”

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

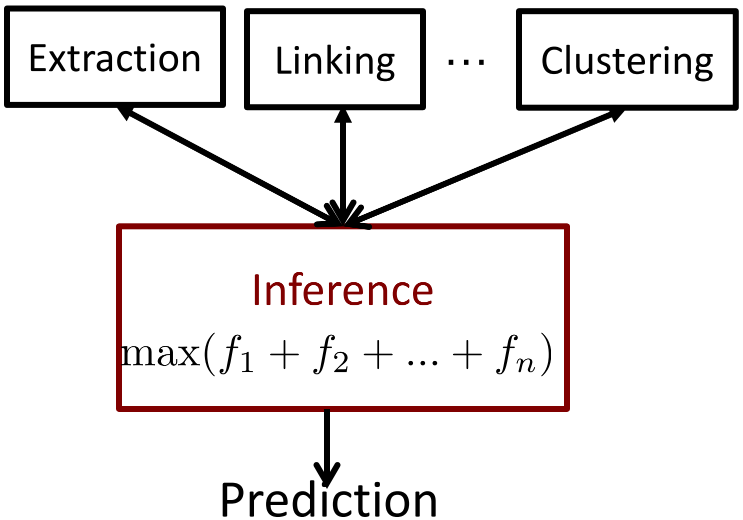


- Errors are compounded from stage to stage
- No interaction between individual predictions
- Incapable of dealing with global dependencies

Solution: Joint Extraction and Linking

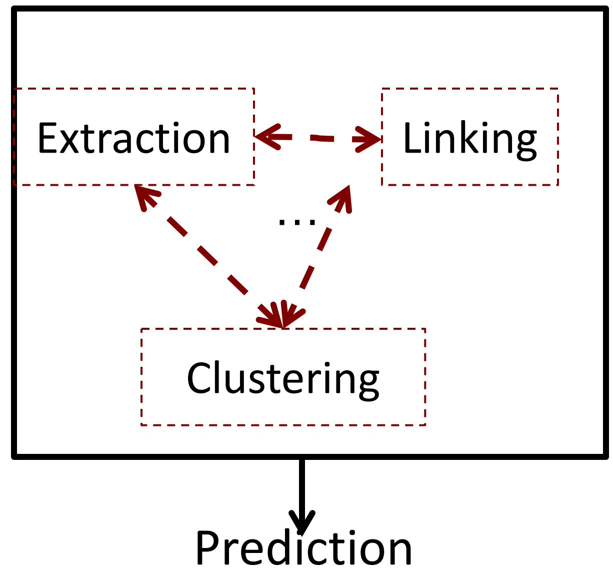
- [Blue Cross]$_{ORG}$ and [Blue Shield of Alabama]$_{ORG}$
- [Blue Cross and Blue Shield of Alabama]$_{ORG}$ → BCBS of Alabama

Joint Inference



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

Joint Modeling

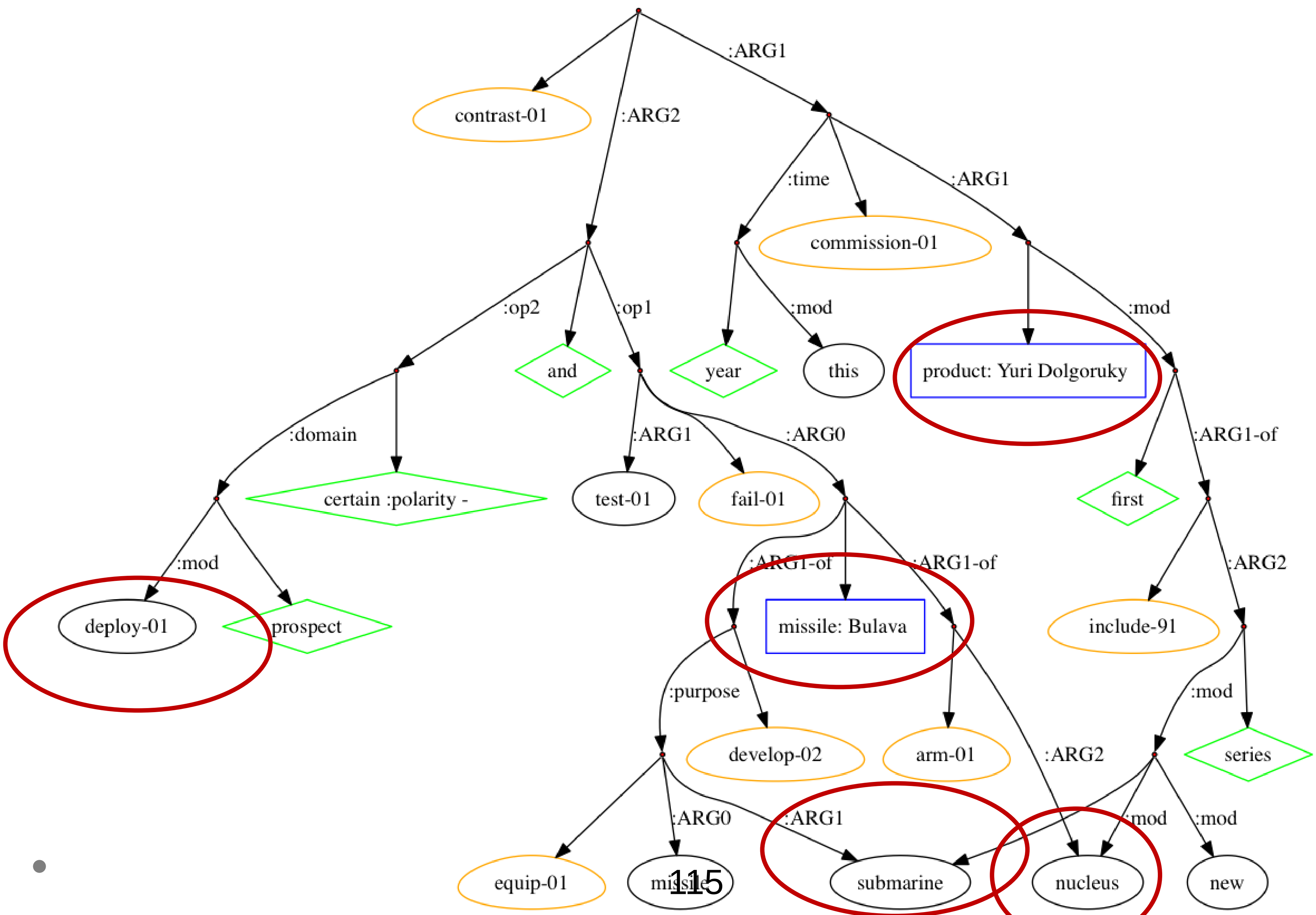


- Probabilistic Graphical Models [Sutton2004,Wick2012,**Wick2013**, Singh2013]
- Markov logic networks [Poon2007,Poon2010,Kiddon2012]
- Linking for extraction [**Meij2012,Guo2013,Fahrni2013,Huang2014**]

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



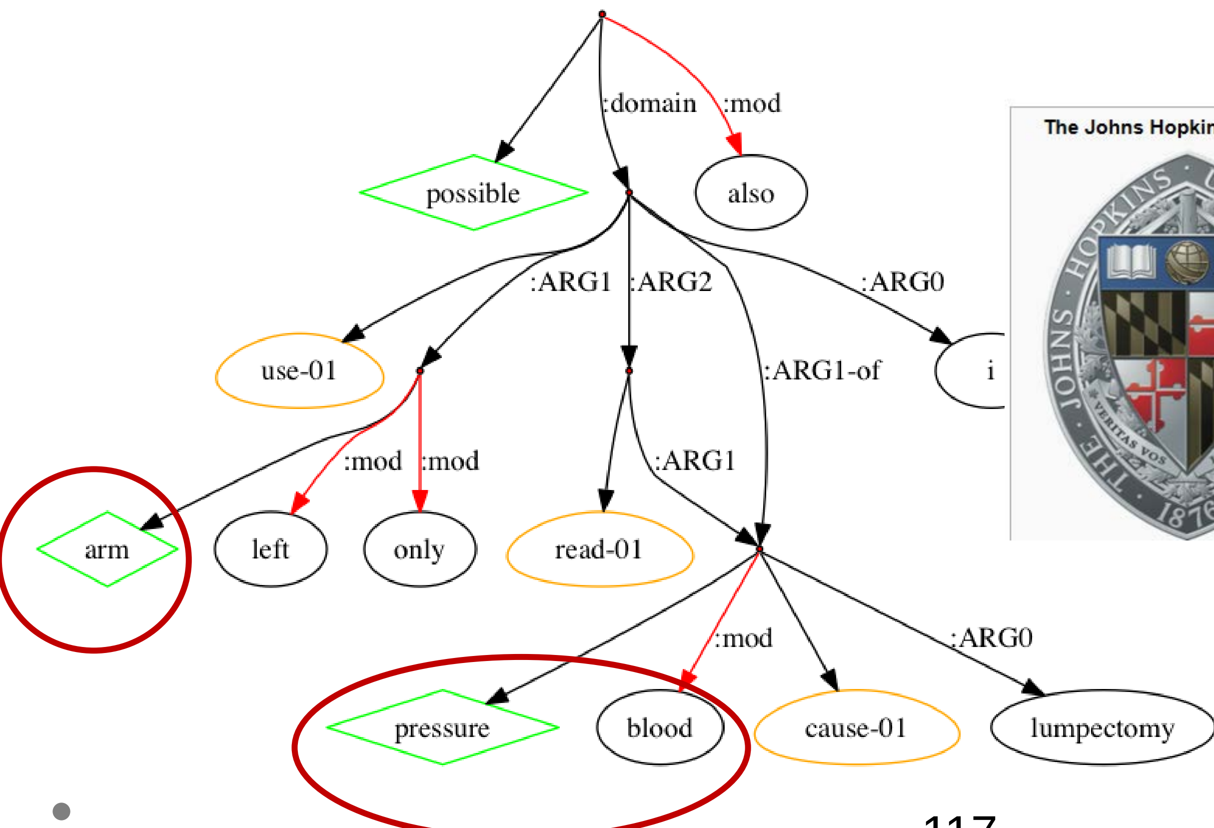
Joint Extraction and Linking

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_Depot Home_Depot

Indicates over-segmentation

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

[Home Depot]

Home_Depot

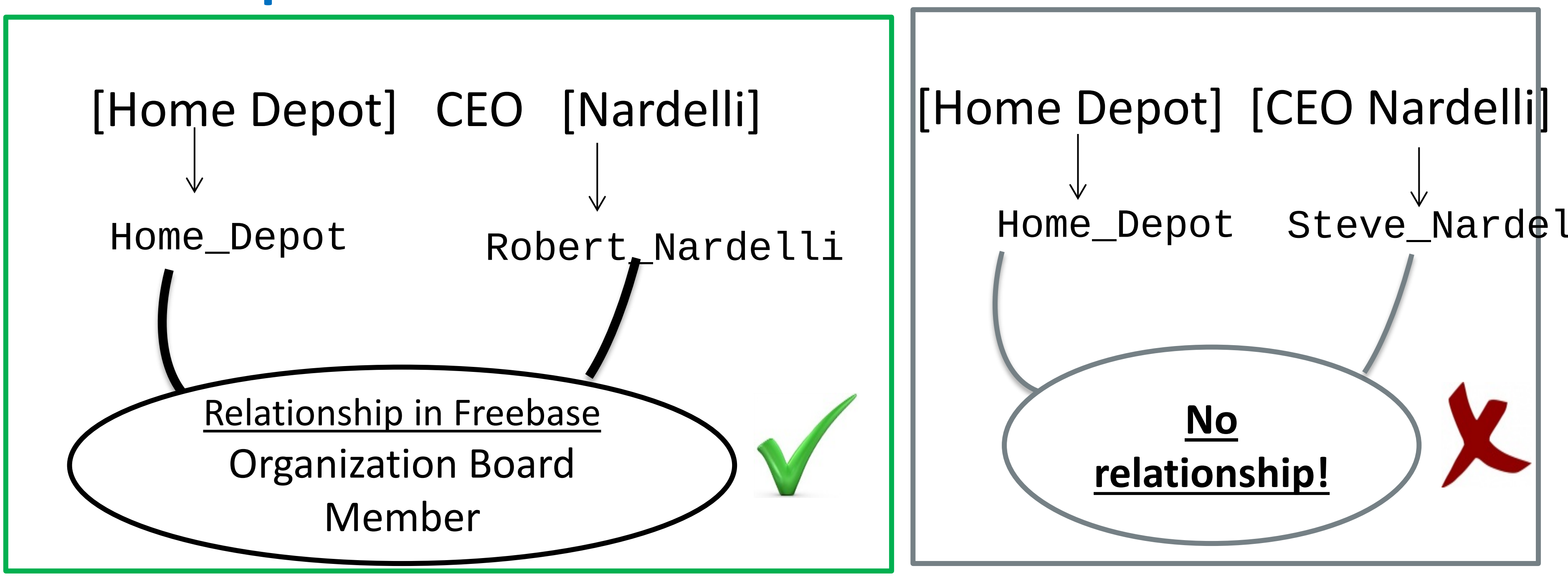
Indicates **correct segmentation**

#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]

$\downarrow$ Home_Depot
Type: Organization

$\downarrow$ Robert_Nardelli
Type: Org_Leader

[Home Depot] [CEO Nardelli]

$\downarrow$ Home_Depot
Type: Organization

$\downarrow$ Steve_Nardelli
Type: Music_Artist

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

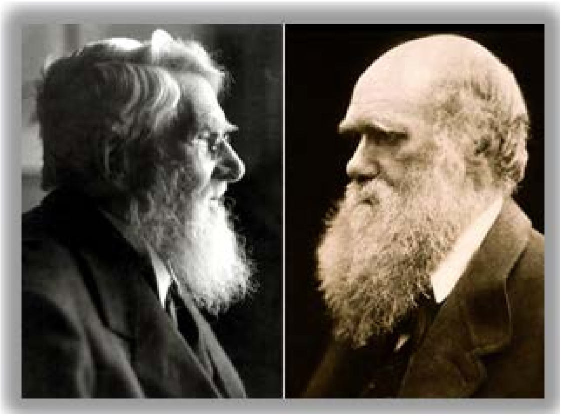
From Non-collective to Collective

- 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

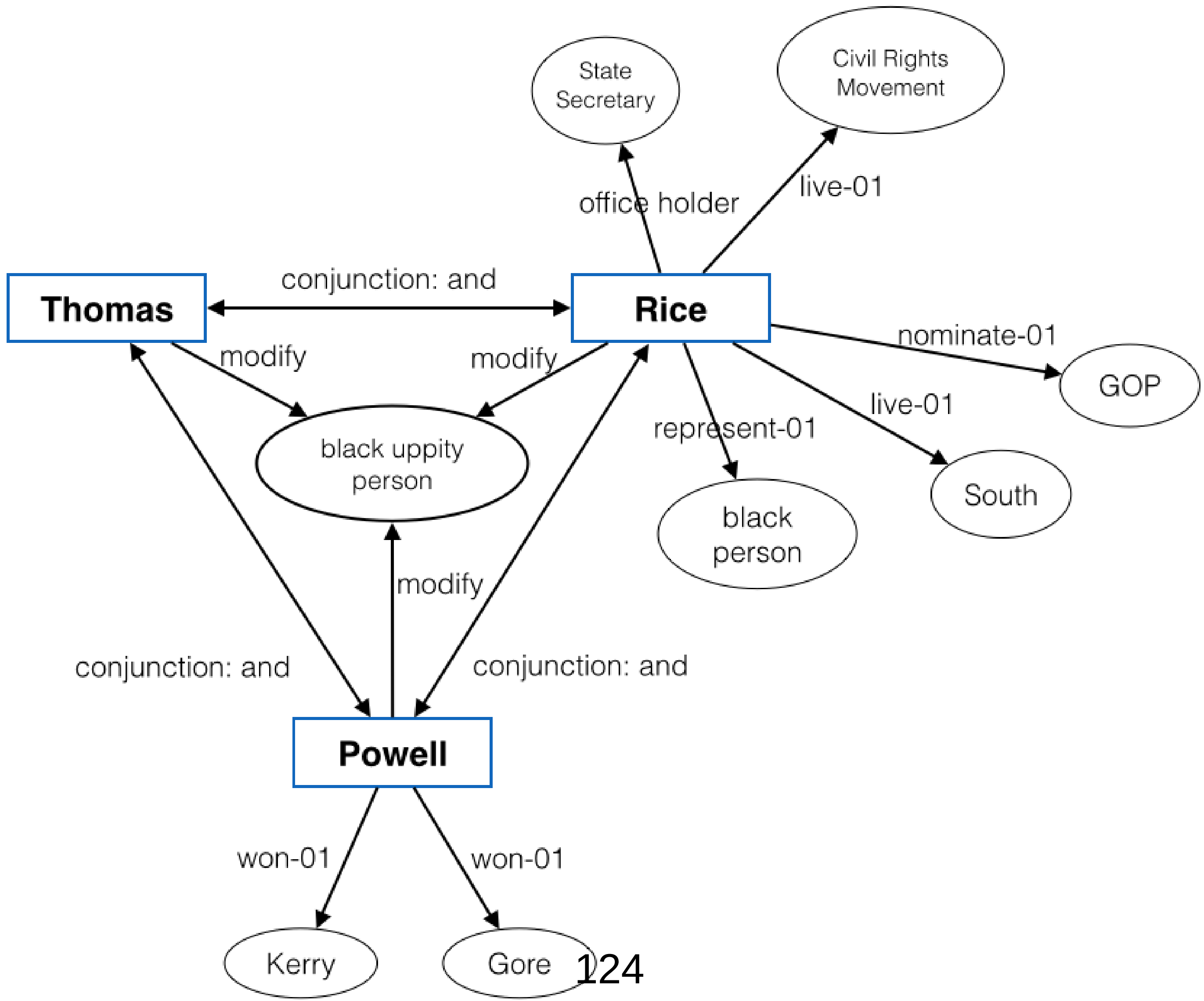


Great Minds Think Alike

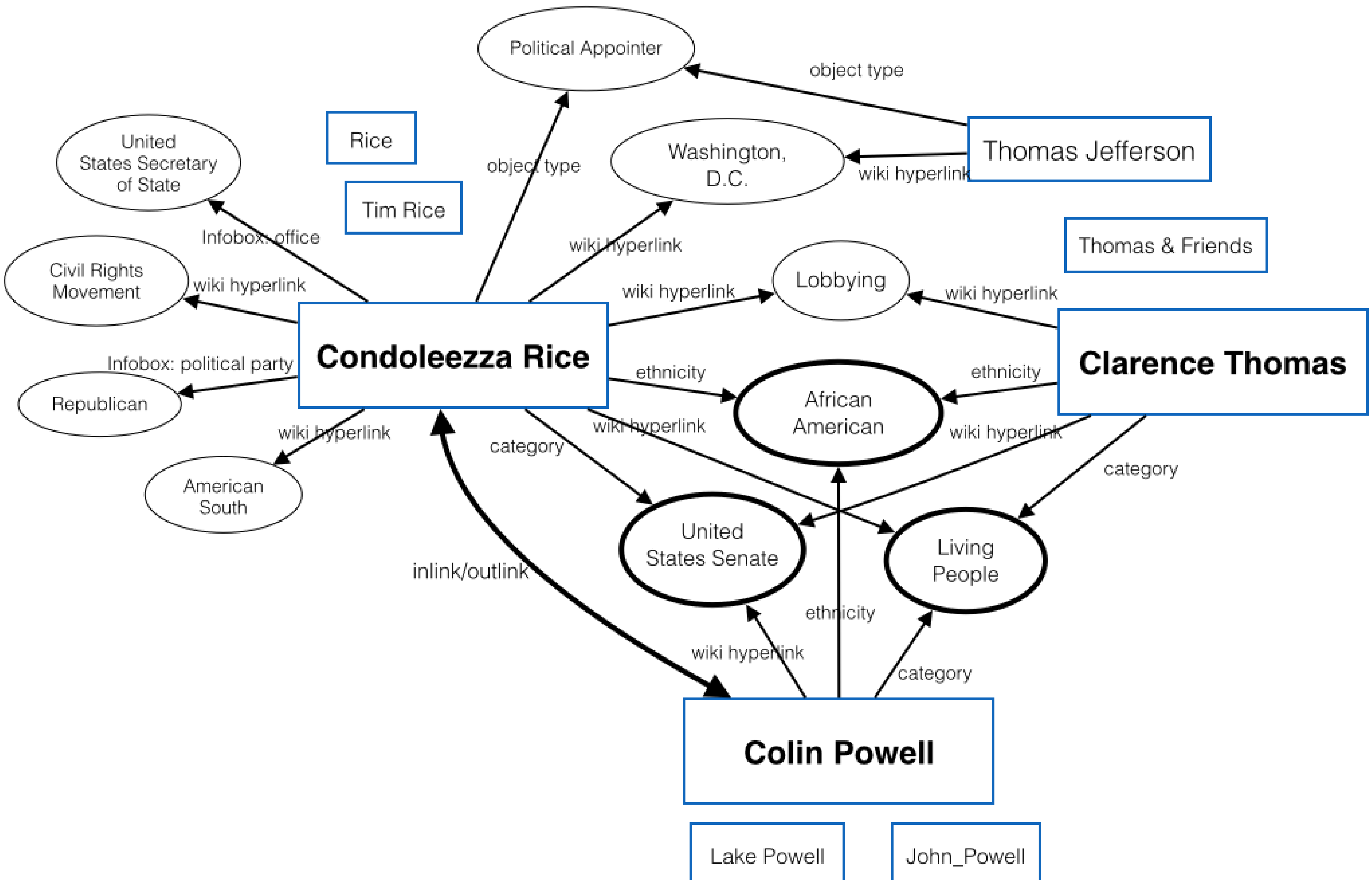
(1) Collective Inference: Basic Idea

- 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

Construct Knowledge Graph of Concept Mentions and their Collaborators



Construct Corresponding Knowledge Graph of Concept Candidates and their Collaborators



Put All Graph Measures Together

- Information Volume: Strong Connectivity among Important Concept Collaborators (Zheng et al., 2014)

$$Sim(m,c) = \propto \times I(c) + \beta \times \sum_{n \in \cap(\theta_m, \theta_c), p \in P} I(n) \times I(p)$$

Importance of C *Importance of C's neighbors* *Importance of property P*

$IV(G_v) = \sum_{c_i, c_j \in V, p \in E} (Sim(m_i, c_i) + Sim(m_j, c_j)) \times I(p) + \sum_{c_k \in V} Sim(m_k, c_k)$

- Select the concept embedded in the subgraph with the largest information volume

Linking Accuracy on AMR Corpus

Method		Acc@1	Acc@5	Acc@10
Baseline	Google Search	85.8%	90.1%	90.1%
Knowledge Graph Matching	(1). Knowledge Graph on Merged KBs	87.8%	93.1%	93.5%
	(1) + Human AMR on Source (Banarescu et al., 2013)	96.3%	98.8%	99.2%
	(1) + Automatic AMR Edges on Source (*Flanigan et al., 2014*)	**94.0%**	**97.8%**	**98.2%**
State-of-the-art Supervised Method		94.3%	97.7%	97.7%

B-cubed+ F-score on KBP Corpus

<table>
<tr><th colspan="2">Method</th><th>F</th></tr>
<tr><td rowspan="2">Knowledge Graph Matching</td><td>(1). Salience + Similarity</td><td>63.6%</td></tr>
<tr><td>(2). (1) + Coherence</td><td>70.9% (top 5)</td></tr>
<tr><td colspan="2">Top 1 Unsupervised System in KBP2013</td><td>63.2%</td></tr>
<tr><td colspan="2">Top 1 Supervised System in KBP2013</td><td>72.4%</td></tr>
</table>

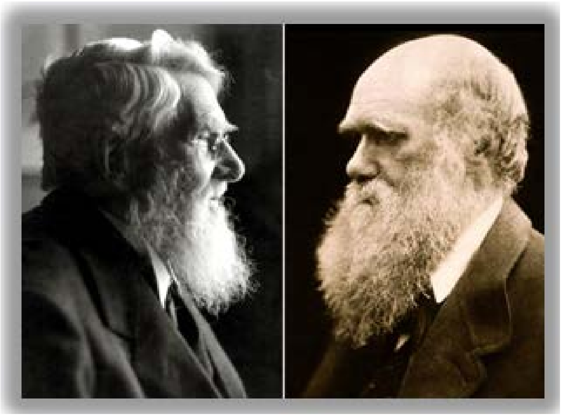
From Non-collective to Collective

- 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

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

Loss Function

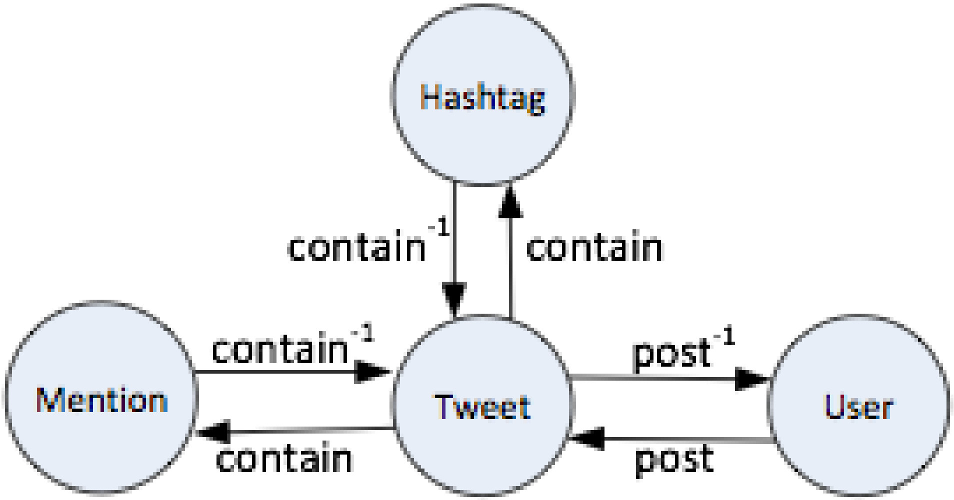
Regularizer

Collective Inference with Social Relations

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

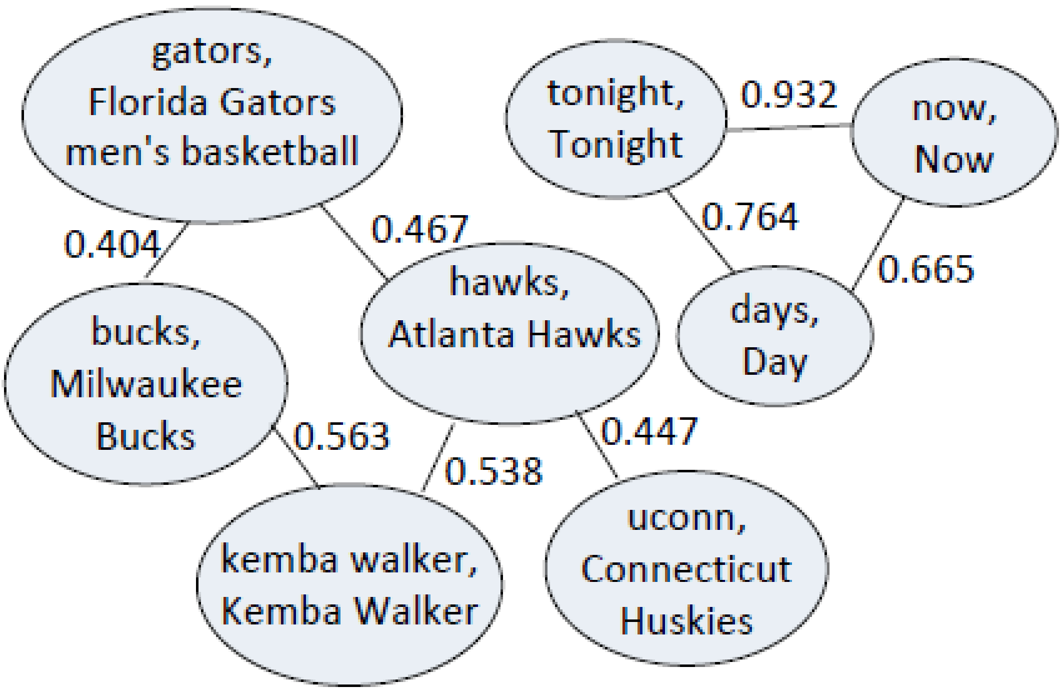
Meta Path

- 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
  - M-T-M
  - M-T-U-T-M
  - M-T-H-T-M
  - M-T-U-T-M-T-H-T-M
  - M-T-H-T-M-T-U-T-M

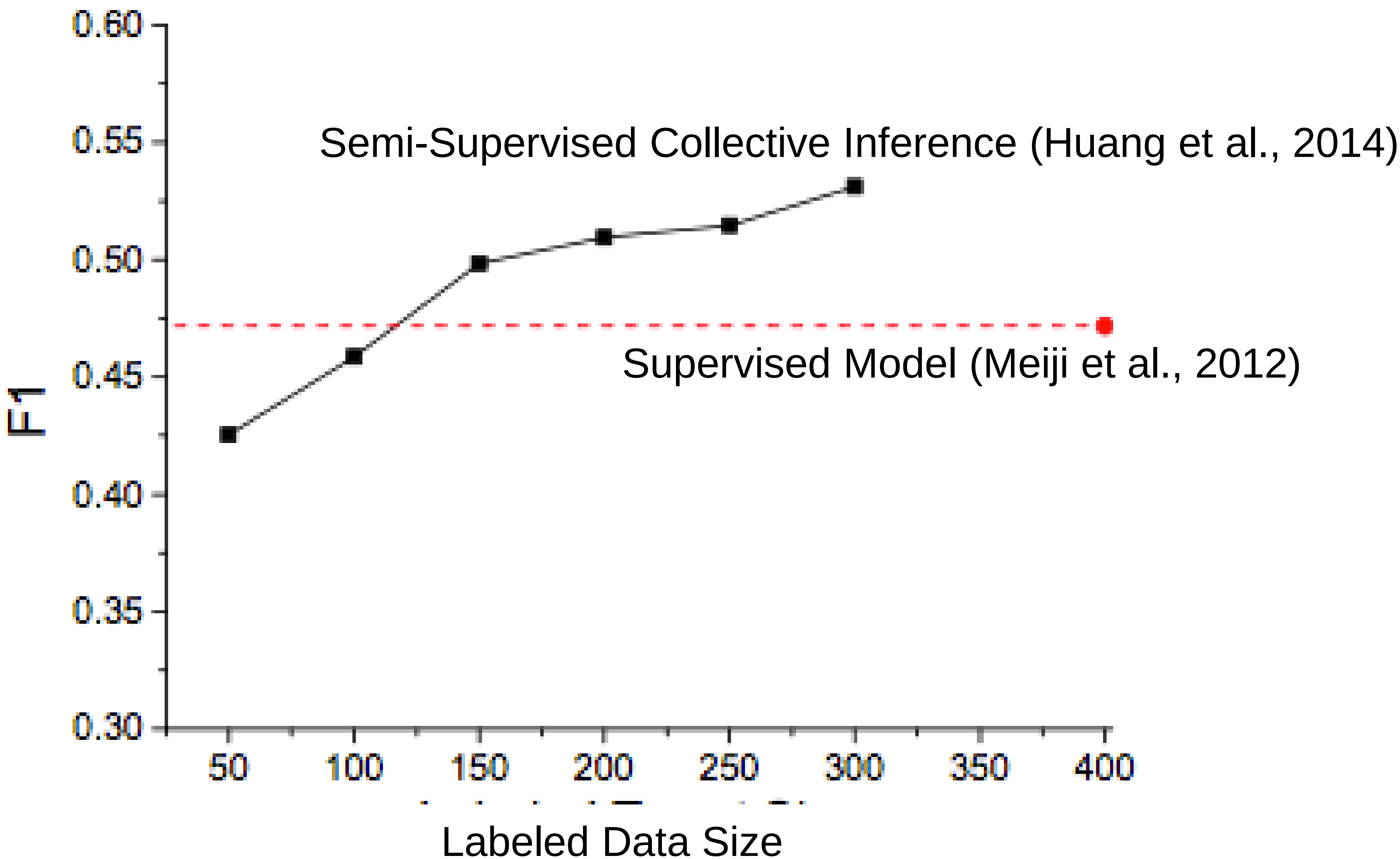


Relational Graph

- 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

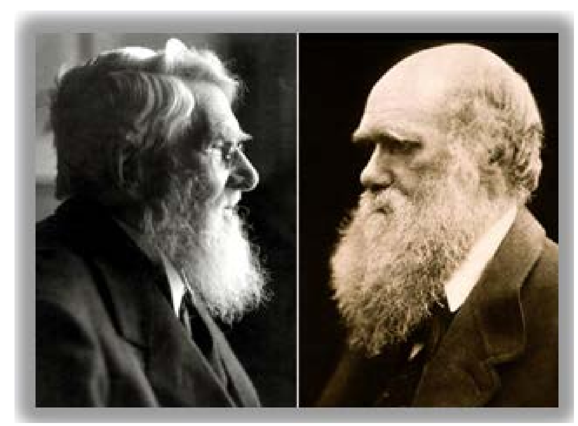
From Non-collective to Collective

- 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

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?**

Relational Inference

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

Relational Inference

Mubarak**, the wife of deposed** Egyptian President Hosni Mubarak,,....

- What are we missing with Bag of Words (BOW) models?
  - Who is Mubarak?
- Textual relations provide another dimension of text understanding
- Can be used to constrain interactions between concepts
  - $\boxed{\text{(Mubarak, wife, Hosni Mubarak)}}$
- Has impact **on several steps** in the Wikification process:
  - From candidate selection to ranking and global decision

Knowledge in Relational Inference

apposition

Coreference

possessive

...ousted long time Yugoslav President Slobodan Milošević in October. The Croatian parliament... Mr. Milošević's Socialist Party

- What concepts can "Socialist Party" refer to?
  - o Wikipedia link statistics is uninformative

Personal details	
Born	20 August 1941 Požarevac, Yugoslavia
Died	11 March 2006 (aged 64) The Hague, Netherlands
Nationality	Serbian
Political party	Socialist Party of Serbia (after 1990) League of Communists of Yugoslavia (until 1990)
Spouse(s)	Mirjana Marković
Children	Marko and Marija
Alma mater	University of Belgrade Faculty of Law
Religion	Atheist[1]
Signature	

Socialist Party of Serbia

Социјалистичка Партија Србије

Socijalistička Partija Srbije

President	Ivica Dačić
Founder	Slobodan Milošević
Founded	17 July 1990
Preceded by	League of Communists of Serbia

Candidate Generation

...ousted long time Yugoslav President Slobodan Milošević in October. Mr. Milošević's Socialist Party...

k	e^k_3
1	Slobodan_Milošević
2	Milošević_(surname)
3	Boki_Milošević
4	Alexander_Milošević
…	

k	e^k_4
1	Socialist_Party_(France)
2	Socialist_Party_(Portugal)
3	Socialist_Party_of_America
4	Socialist_Party_(Argentina)
…	

Candidate Ranking

...ousted long time Yugoslav President Slobodan Milošević in October. Mr. Milošević's Socialist Party...

k	e^k_3	s^k_3
1	Slobodan_Milošević	0.7
2	Milošević_(surname)	0.1
3	Boki_Milošević	0.1
4	Alexander_Milošević	0.05
…		

k	e^k_4	s^k_4
1	Socialist_Party_(France)	0.23
2	Socialist_Party_(Portugal)	0.16
3	Socialist_Party_of_America	0.07
4	Socialist_Party_(Argentina)	0.06
…		

- Local and global statistical features

Candidate Generation + Relations

...ousted long time Yugoslav President Slobodan Milošević in October. Mr. Milošević's Socialist Party...

k	e^k_3	s^k_3
1	Slobodan_Milošević	0.7
2	Milošević_(surname)	0.1
3	Boki_Milošević	0.1
4	Alexander_Milošević	0.05
…		

$r_{34}^{(1,21)} = 1$

k	e^k_4	s^k_4
1	Socialist_Party_(France)	0.23
2	Socialist_Party_(Portugal)	0.16
3	Socialist_Party_of_America	0.07
4	Socialist_Party_(Argentina)	0.06
…		
21	Socialist_Party_of_Serbia	0.0

- More robust candidate generation
  - Identified relations are verified against a knowledge base (DBPedia)
  - Retrieve relation arguments matching "(Milošević ,?,Socialist Party)" as our new candidates

Candidate Ranking + Relations

...ousted long time Yugoslav President Slobodan Milošević in October. Mr. Milošević's Socialist Party...

k	e^k_3	s^k_3
1	Slobodan_Milošević	0.7
2	Milošević_(surname)	0.1
3	Boki_Milošević	0.1
4	Alexander_Milošević	0.05
…		

k	e^k_4	s^k_4
1	Socialist_Party_(France)	0.23
2	Socialist_Party_(Portugal)	0.16
3	Socialist_Party_of_America	0.07
4	Socialist_Party_(Argentina)	0.06
…		
21	Socialist_Party_of_Serbia	0.0

$r_{34}^{(1,21)} = 1$

Relation query

Retrieved relation tuple

$$w = \frac{1}{Z} f(q, \sigma)$$

$$w_{34}^{(1,21)} = ?$$

Inference Formulation

Having some knowledge, and knowing how to use it to support decisions, facilitates the acquisition of additional knowledge.

- Goal: Promote concepts that are <u>coherent with textual relations</u>
- Formulate as an Integer Linear Program (ILP):

weight to output e_i^k

Whether to output kth candidate of the ith mention

$$\Gamma_D = \underline{\arg\max_{\Gamma} \sum_i \sum_k s_i^k e_i^k} + \sum_{i,j} \sum_{k,l} w_{ij}^{(k,l)} r_{ij}^{(k,l)}$$

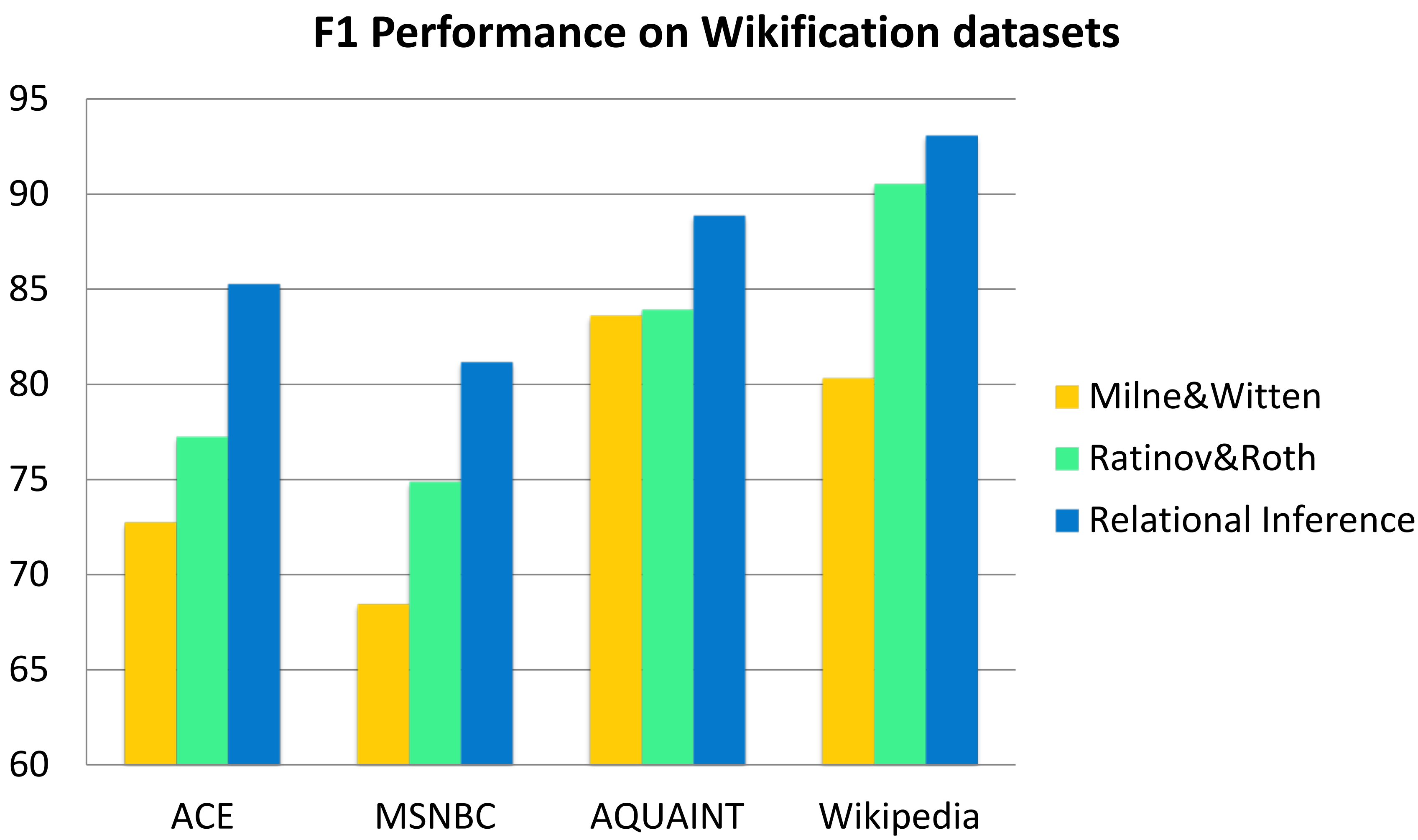
weight of a relation $r_{ij}{}^{(k,l)}$

Whether a relation exists between e_i^k and e_j^l

$s.t.$	$r_{ij}^{(k,l)} \in \{0,1\}$	Integral constraints
	$e_i^k \in \{0,1\}$	Integral constraints
	$\forall i \sum_k e_i^k = 1$	Unique solution
	$2r_{ij}^{(k,l)} \leq e_i^k + e_j^l$	Relation definition

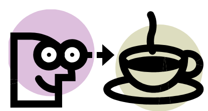
- If no relation exists, collapses to the non-structured decision

Wikification Performance Result [Cheng & Roth, EMNLP'13]



Outline

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

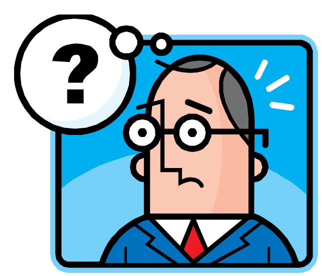


Coffee Break

- 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
  - Spatiotemporal Wikification
  - Handling emerging entities
  - Cross-lingual Entity Linking
  - Linking to general KB and ontologies
  - Fuzzy matching for candidates



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]

Motivation: Short and Noisy Text

- Microblogs are data gold mines!
  - Over 400M short tweets per day

- Many applications
  - Election results [Tumasjan et al., SSCR 10]
  - Disease spreading [Paul and Dredze, ICWSM 11]
  - Tracking product feedback and sentiment [Asur and Huberman, WI-IAT 10]
- Need more research
  - Stanford NER on tweets set achieves 44% F1 [Ritter et. al, EMNLP 2011]

Challenges for Social Media

- Messages are short, noisy and informal
  - Lack of rich context to compute context similarity and ensure topical coherence
- Lack of Labeled Data for Supervised Model
  - Lack of Context makes annotation more challenging
  - Need to search for more background information

who cares, nobody wanna see the spurs play. Remember they're boring…

What approach should we use?

- Task: Restrict mentions to Named Entities
  - Named entity Wikification

Mature Techniques

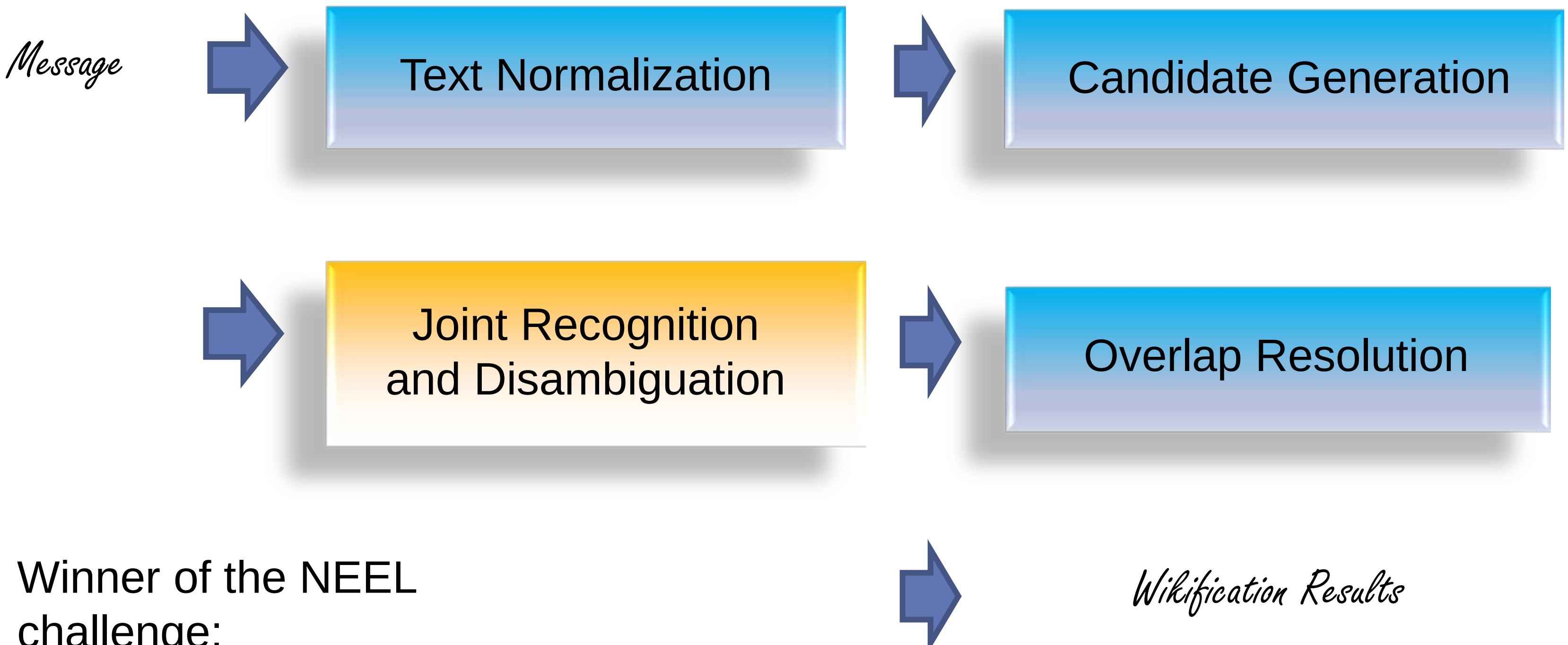
- Approach 1 (NER + Disambiguation):
  - Develop a named entity recognizer for target types
  - Link to entities based on the output of the first stage

Limited Types; **Adaptation**

- Approach 2 (End-to-end, Wikification):
  - Learn to jointly detect mention and disambiguate entities
  - Take advantage of Wikipedia information

A Simple End-to-End Wikification System

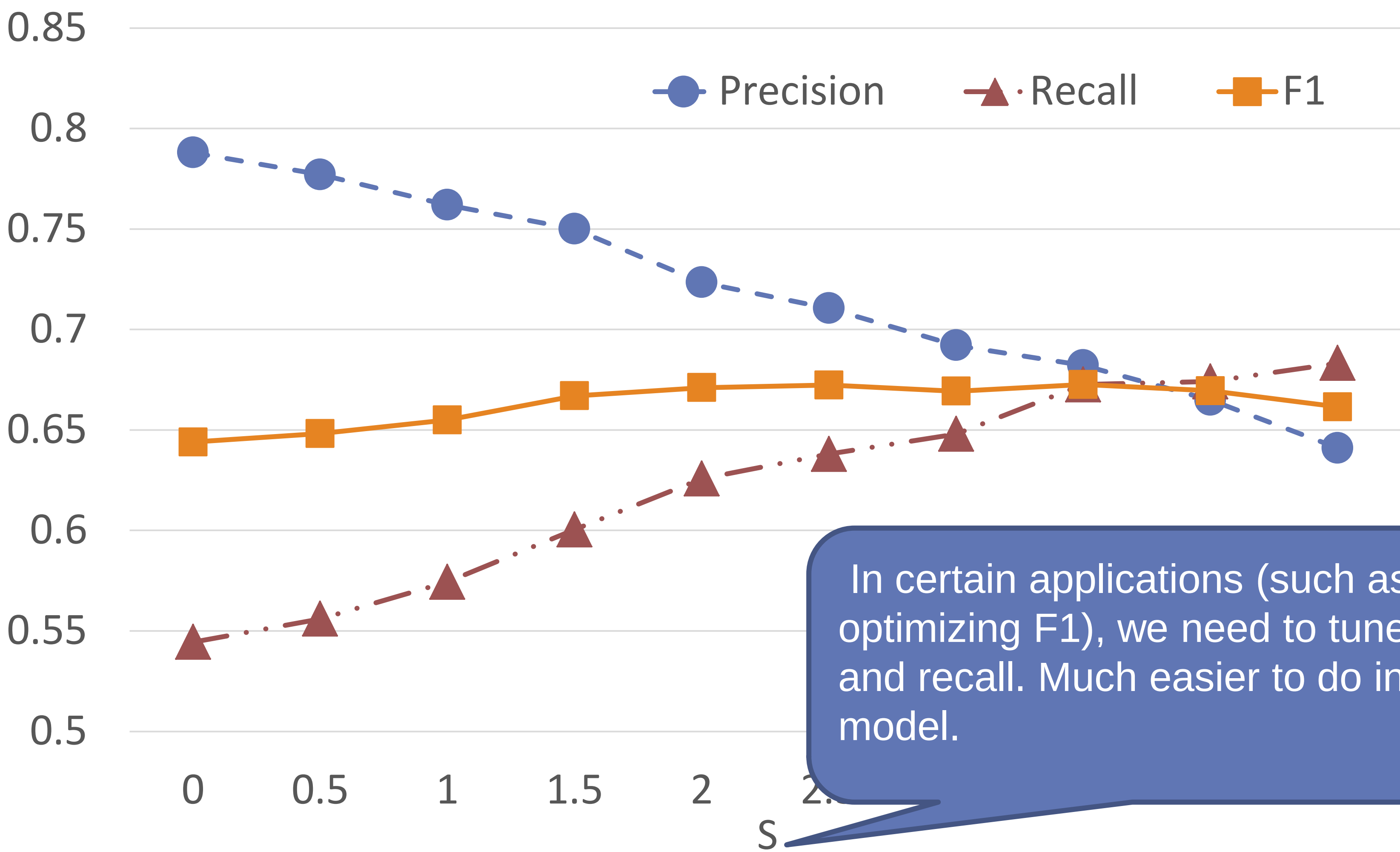
- [Guo, NAACL 13, Chang et. al. #Micropost 14]



Winner of the NEEL challenge;
<u>The best two systems all adopt the end-to-end approach</u>

Only matching; there is not mention detection stage

Balance the Precision and Recall



In certain applications (such as optimizing F1), we need to tune precisio and recall. Much easier to do in a joint model.

How Difficult is Disambiguation?

Data	#Tweets	#Cand	#Entities	P@1
Test 2	488	7781	332	89.6%

- Commoness Baseline [Guo et al., NAACL 13]
  - Gold mentions match the prior anchor text (e.g. the lexicon)
  - P@1 = the accuracy of the most popular entity
- The baseline for disambiguating entities is high
  - The overall entity linking performance is still low
    - Mention detection is challenging for tweets!
- The mention detection problem is even more challenging
  - The lexicon is not complete

Morphs in Social Media

 = 

"Conquer West King"
(平西王)

"Bo Xilai"
(薄熙来)

 = 

"Baby"
(宝宝)

"Wen Jiabao"
(温家宝)

Chris Christie → the Hutt

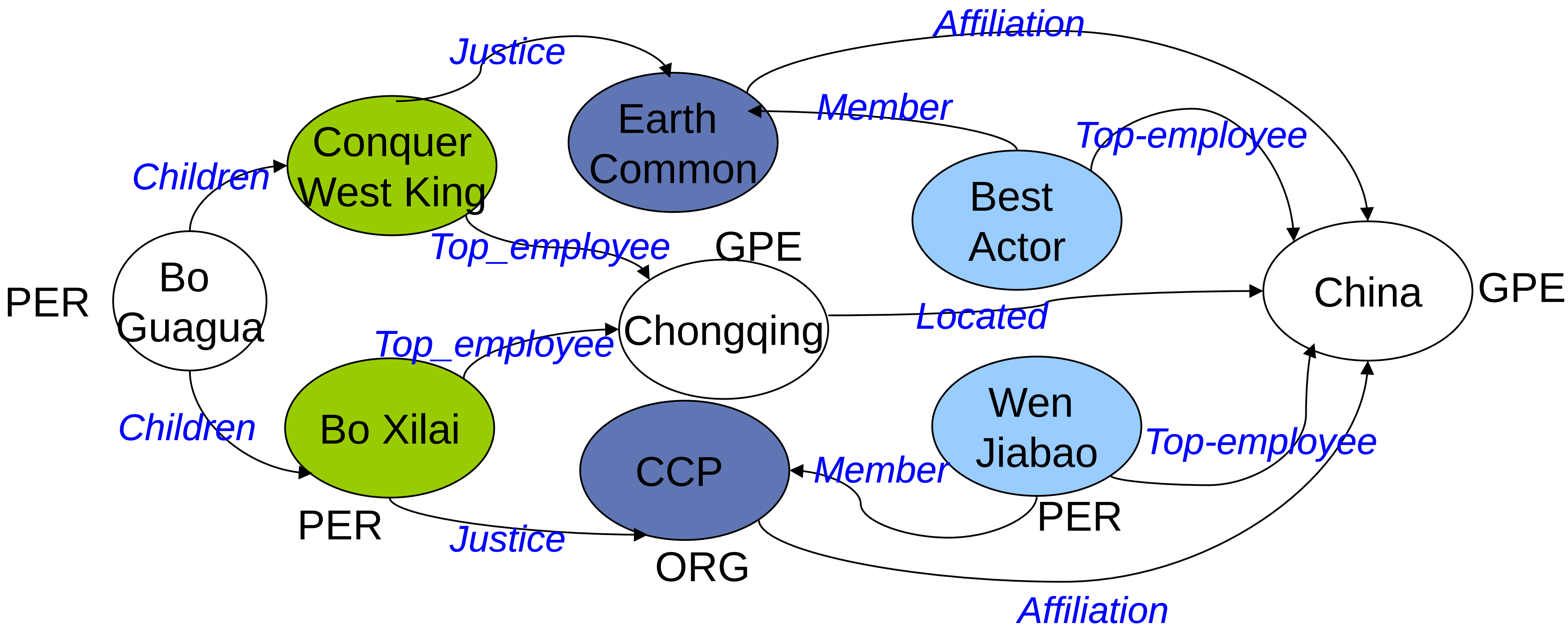
Morph Decoding

- Goal: automatically determine which term is used as a morph, and resolve it to its true target

- Then ***Wu Sangui*** helped the ***army of Qing dynasty*** invaded China, and became ***Conquer West King***.

- ***Conquer West King*** from ***Chongqing*** fell from power, still need to ***sing red songs***?

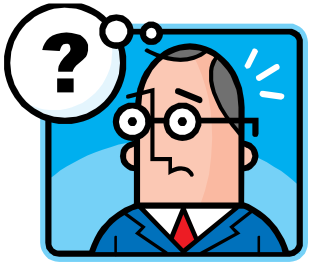
Morph Linking based on Information Networks Construction (Huang et al., 2013)



- Each node is an entity mention
- An edge: a semantic relation, event, sentiment, semantic role, dependency relation or co-occurrence, associated with confidence values
- Meta-path: a meta-level description of a existing or concatenated path between two object types (Sun et al., 2012)

New Trends

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



Spatiotemporal Signals

who cares, nobody wanna see the <u>spurs</u> play. Remember they're boring…

- In labeled data: [Fang and Chang, TACL 14]
  - In US, San Antonio Spurs accounts for 91% of "spurs"
  - In UK, San Antonio Spurs only accounts for 8 % of "spurs"

- It is important to use spatiotemporal signals
  - How to use? Indirect or direct?
    - Direct: associate entities with time and location
    - Indirect: assume entities around the same time are similar

Evaluating NE Wikification in Tweets

- Information Extraction and Information Retrieval Settings
  - IE: Given a collection of tweets, get all of the entities
  - IR: Given an entity, get all of the tweets that mentioning it
    - Use predefine keywords to get a set of tweets
    - Measure the classification performance

- IR setting is not easy, because of surface form ambiguity

	Query Entity
PER	Hillary Rodham Clinton
ORG	Big Bang (South Korea Band)
LOC	Washington (state)

Bill Clinton

Big Bang Theory

DC, University,…

- IR setting is important for market research

Spatiotemporal Signals can help

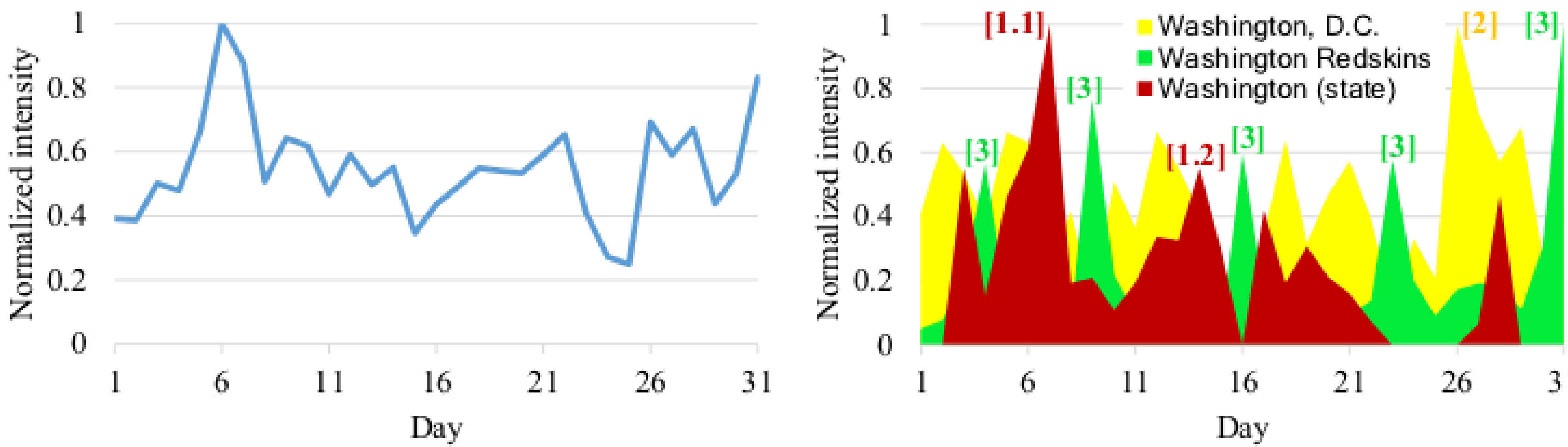
- The state of the art baseline
  - (At that time)
- Using both time and location
  - Helps even more

	IE	IR
Base	57.0	58.4
+T	64.9	71.4
+T+L	68.6	79.0

- Error analysis:
  - Entities are time and location sensitive

Tweet	Example entity	Posting time	User location
#1	*person*: Colin Kaepernick	during his game	*(not useful)*
#2	*org*: Cali. State U. Fullerton	in campus emergency	California
#3	*loc*: Los Angeles	*(not useful)*	California
#4	*event*: New Year's Eve	on 31 December	*(not useful)*

IR: Entity Linking VS Keyword

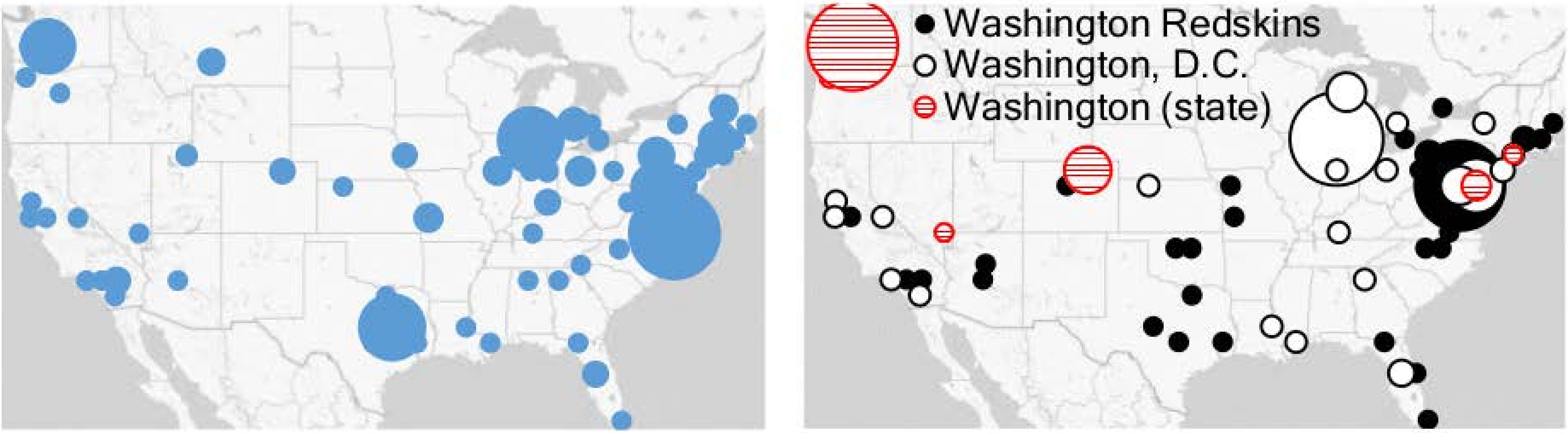


(a) with keywords ("washington")

(b) with entity linking

- Retrieve all tweets for Washington (state)
  - Entity Linking can capture the temporal behavior of entities
- Keyword is the state of the art
  - Naïve keyword: design is simple, but keyword is ambiguous
  - Complicate keyword: low recall: Hard to scale

IR: Entity Linking vs Keyword

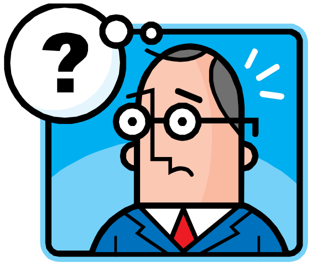


(a) with keywords ("washington") (b) with entity linking

- Entity Linking can recover geo pattern as well
  - We can still spot some mistakes by just looking at the results
- Future research
  - Retweet behavior needs to be reanalysis
  - People have done spatiotemporal analysis for language usage
    - More research is needed on spatiotemporal analysis for entities

New Trends

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



Hoffart et. al, WWW 2014

theguardian

Your search terms | Search

News | Sport | Comment | Culture | Business | Money | Life & style | Travel | Environment | Tech | TV | Video | Dating | Offers | Jobs

News > World news > The NSA files

Laura Poitras and **Glenn Greenwald** in Hong Kong,
theguardian.com , Sunday 9 June 2013

"**Washington**'s **Prism** program was revealed by the whistleblower **Snowden.**"

DBpedia

Identifying Emerging Entities

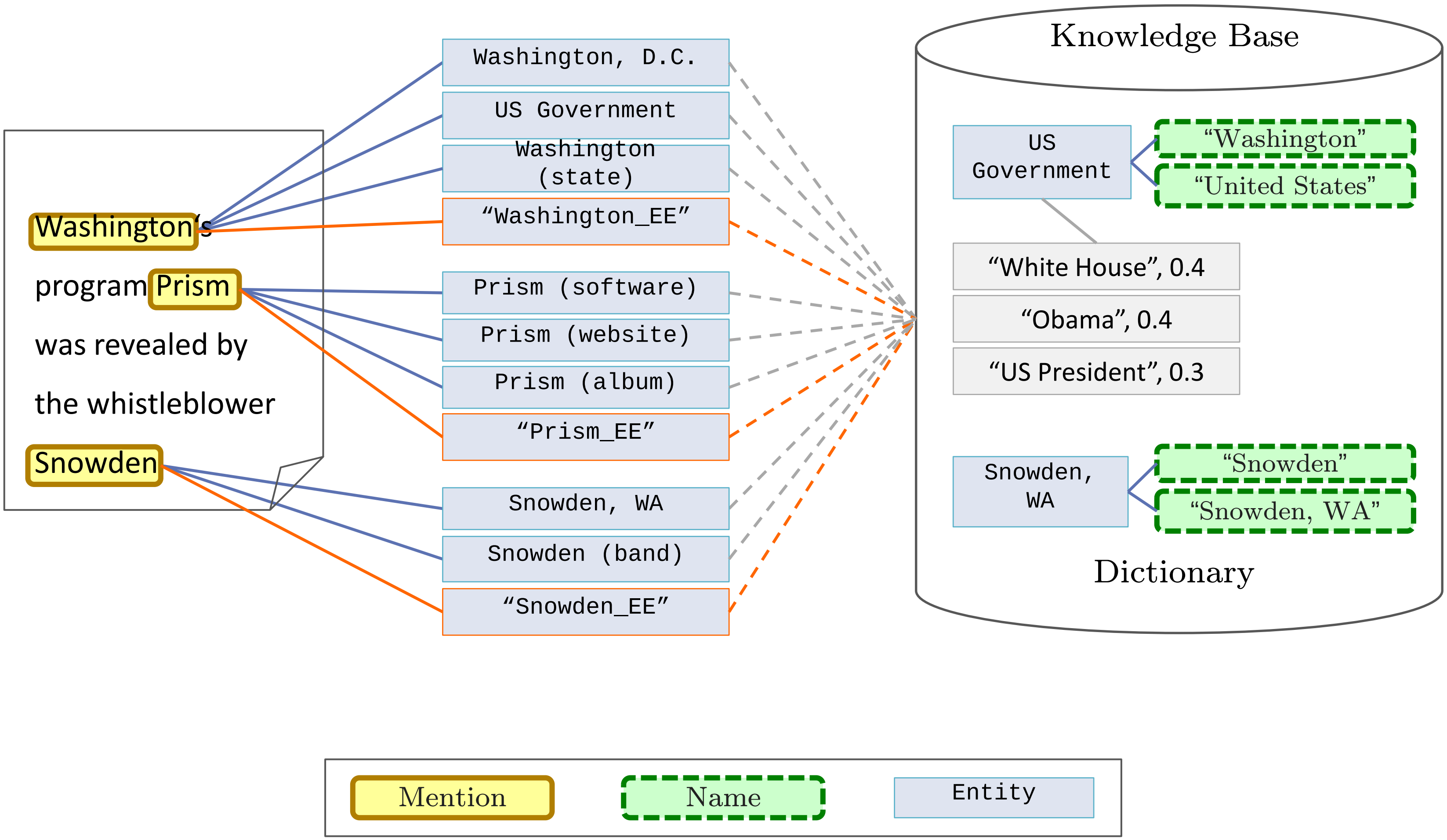
Wikipedia-derived knowledge bases
have **lexicon** of **(name, entity)** pairs

	New Entity	Existing Entity
New Name	assumption	☐
Existing Name	☐	disambiguation

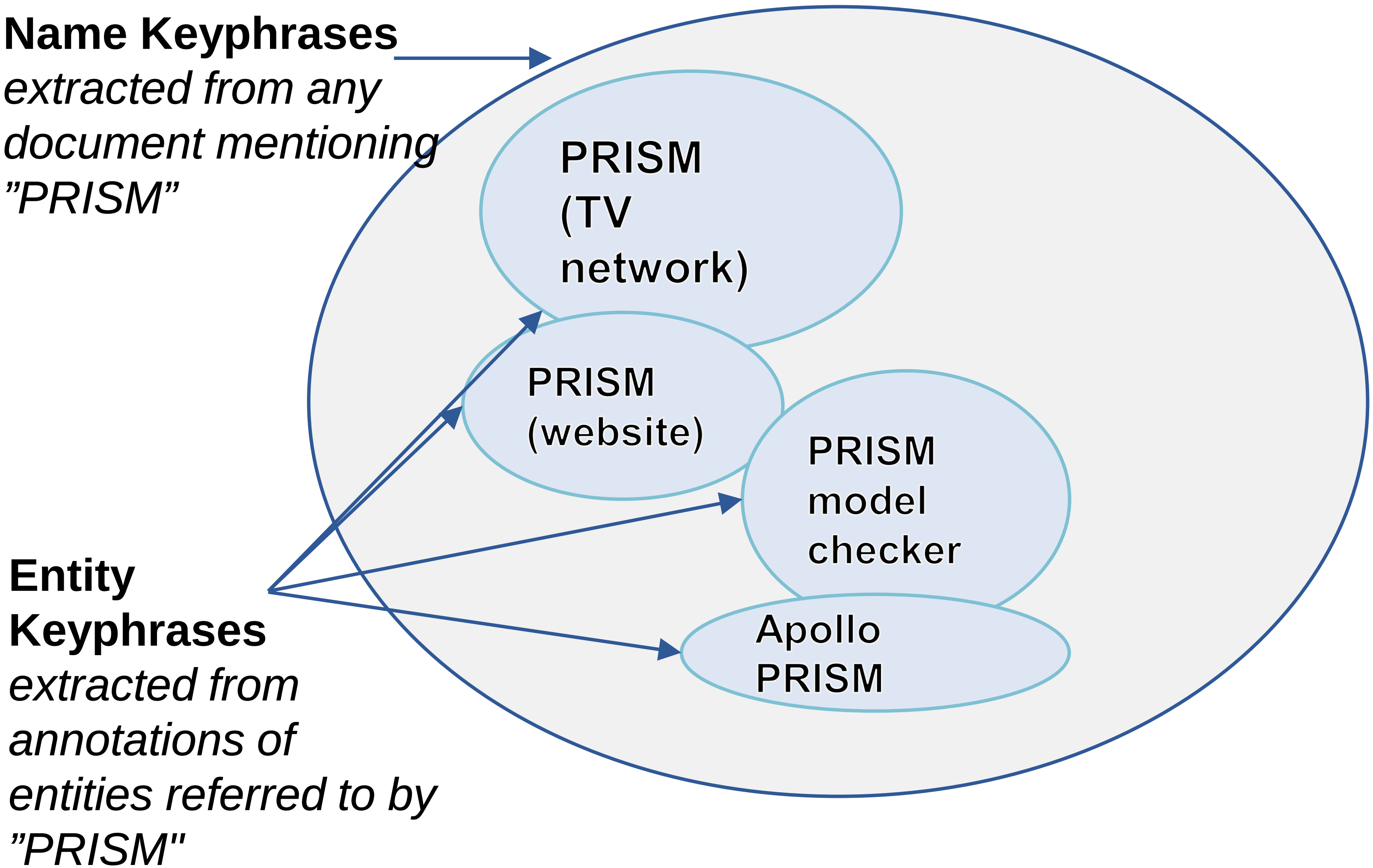
Key idea: Profile Emerging Entities from the Web

Assumption: for one name, one emerging entity

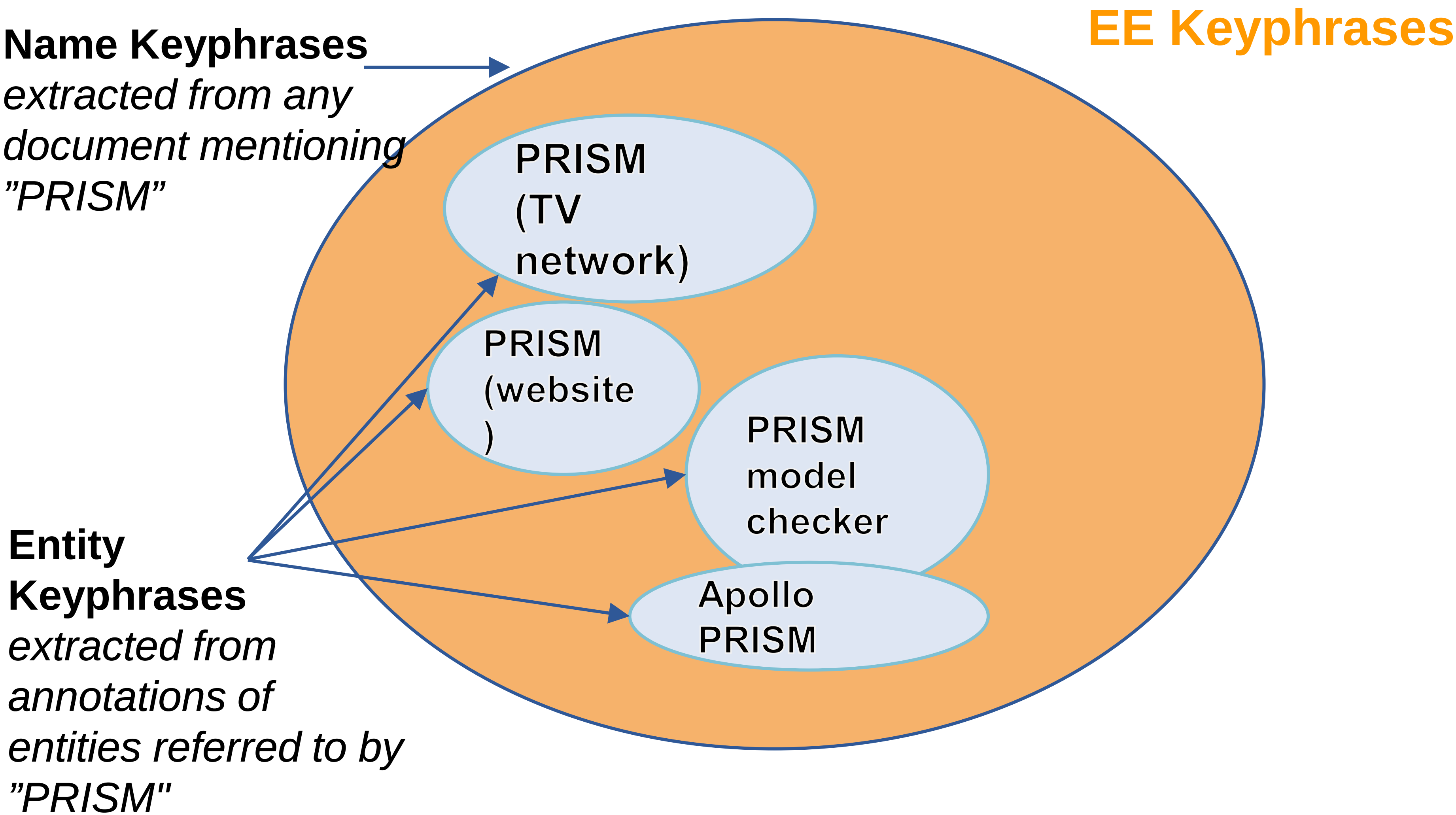
From NED to NED-EE



Harvesting EE Keyphrases



Harvesting EE Keyphrases



Modeling New Entities

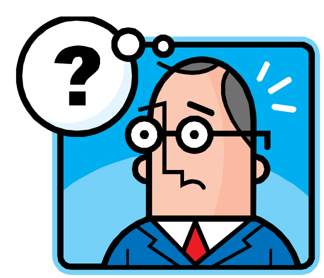
...
The *PRISM* program collects a wide range of data from a number of companies, e.g. Google and Facebook. The leaked National Security Agency (NSA) document obtained by the Guardian claims it operates with the "assistance of communications providers in the US".

...

keyphrases defined by POS pattern filters for **named entities** and **technical terms**

New Trends

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



Cross-lingual Entity Linking (CLEL)

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

Ang Lee

Ang Lee, 2009

Chinese name	李安 (Traditional)
Chinese name	李安 (Simplified)
Pinyin	Lǐ Ān (Mandarin)
Born	October 23, 1954 (age 56) Chaochou, Pingtung, Taiwan)
Years active	1992 – present
Spouse(s)	Jane Lin (1983–)
Children	Haan Lee (b.1984) Mason Lee (b.1990)

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

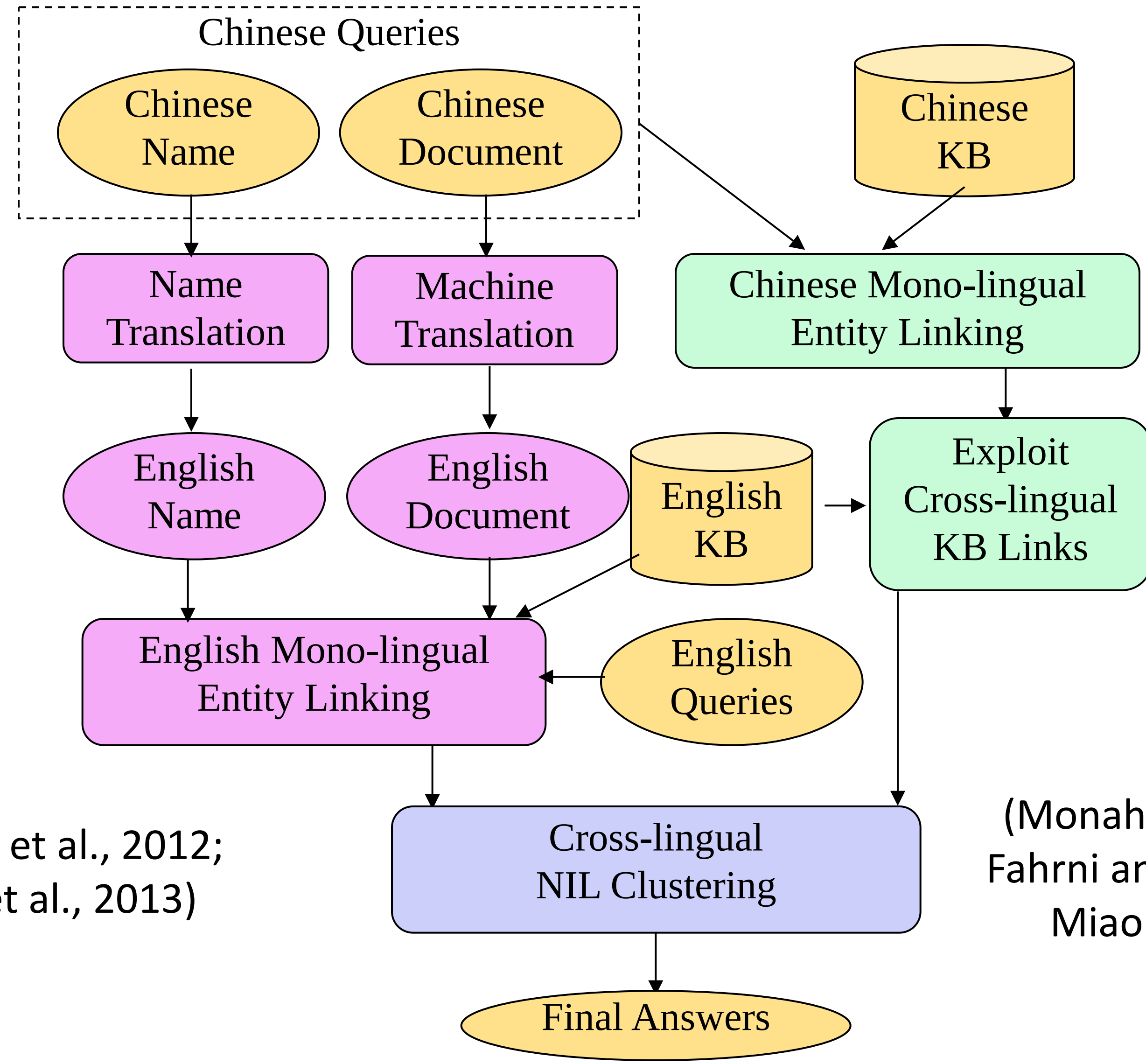
李安曾言，住在花莲的八年，乃其北上就读艺专前最快乐的一段学习岁月。十岁之前的李安在花莲念了两所小学，接受的是美式开放教育，来到台南，又念了两所小学，面对语言习惯不同国语一台语，头一次经验到文化冲击。

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



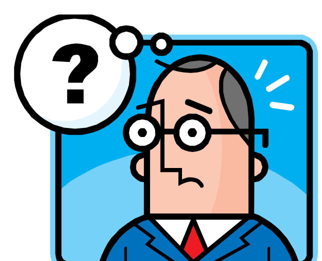
(Cassidy et al., 2012; Miao et al., 2013)

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

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

New Trends

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



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

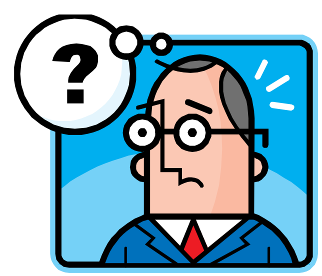
- Wikipedia is not enough
  - Wikification trained only from news and Wikipedia
  - ➔ 20% end-to-end extraction and linking F-measure

- We could learn from Ontologies
  - Semantic relations among concepts in the ontologies (e.g. subClassOf) + collective inference technologies ➔ 84.5%

- Another approach:
  - Wikipedia + Ontologies

New Trends

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



Fuzzy Matching for Candidates

- For current Wikification systems all require
  - A lexicon that maps a surface form to an entity is needed
- Limitations
  - Lexicon size is fixed; Computer memory is limited
  - Countless surface form variations.
- Example 1: Misspellings
  - E.g., in queries, OCR and Speech Rec. errors, …
- Example 2: How to link hashtags to the entities?
  - #TheCloneWar

Link Hashtags

- URL/word/Hashtag Breaker [Wang et. al, WWW 11]
  - o By using the a language model that is trained a very large corpus (Microsft N-gram), we could break hashtags into the words by maximizing the probability of the individual tokens

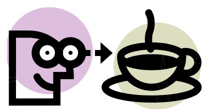
- Building a tire data structure that could get words with similar spellings efficiently [Duan et al. WWW11]

“theCloneWar” → $\boxed{\arg\max_{c;q:q=\bar{q}\cdots} p(q\,|\,c)\,p(c)}$ → *The clone war*

q c

Outline

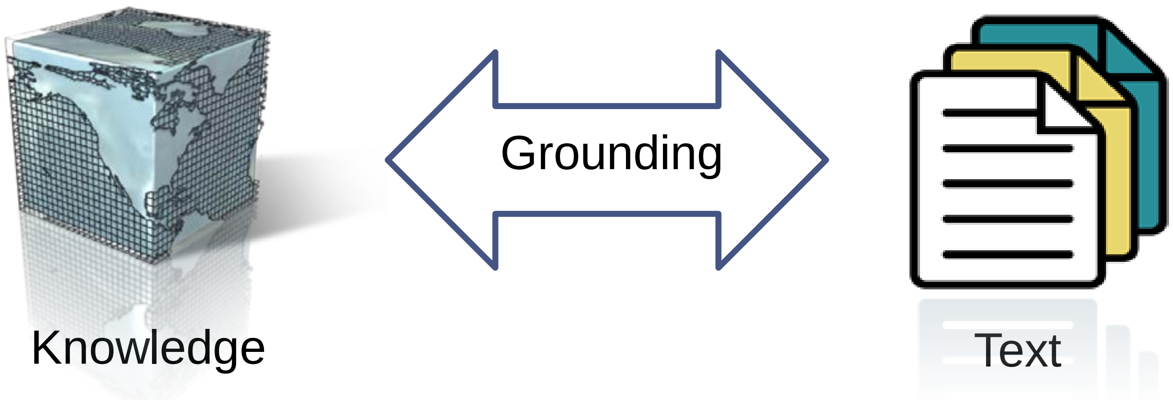
- Motivation and Definition
- A Skeletal View of a Wikification System
  - o High Level Algorithmic Approach
- Key Challenges



Coffee Break

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

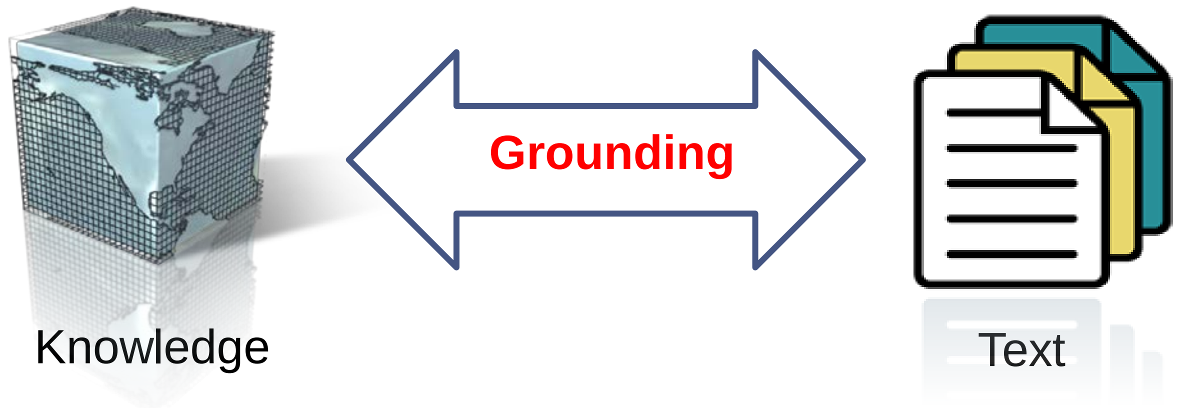
What's Next? Now..



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

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

Entity Grounding to Knowledge Grounding

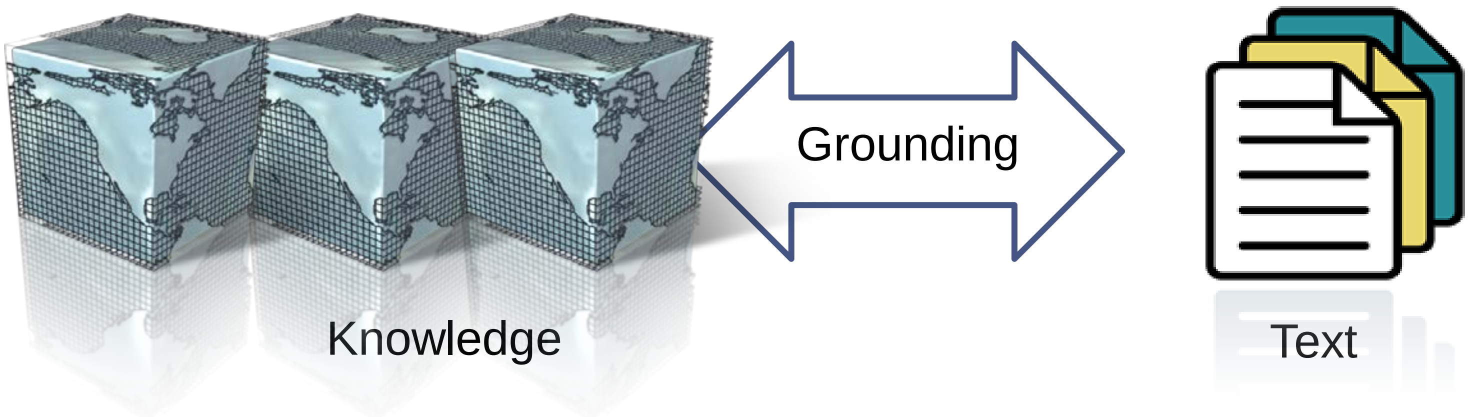


- Knowledge does not only contain entities
  - o Relations: Freebase or Dbpedia
- Large scale semantic parsing
  - o Semantic parsing + Entity Linking?

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

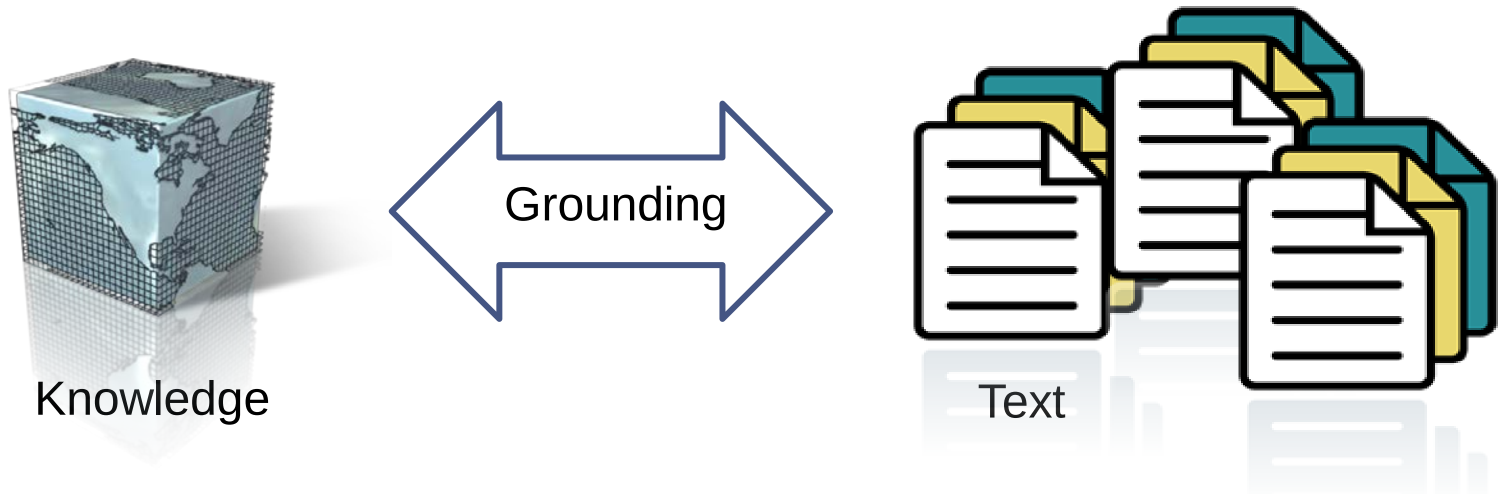
    <u>The lexical matching problem in semantic parsing is entity linking</u>
- Should we jointly ground entities and relations?

Multiple Knowledge Resources



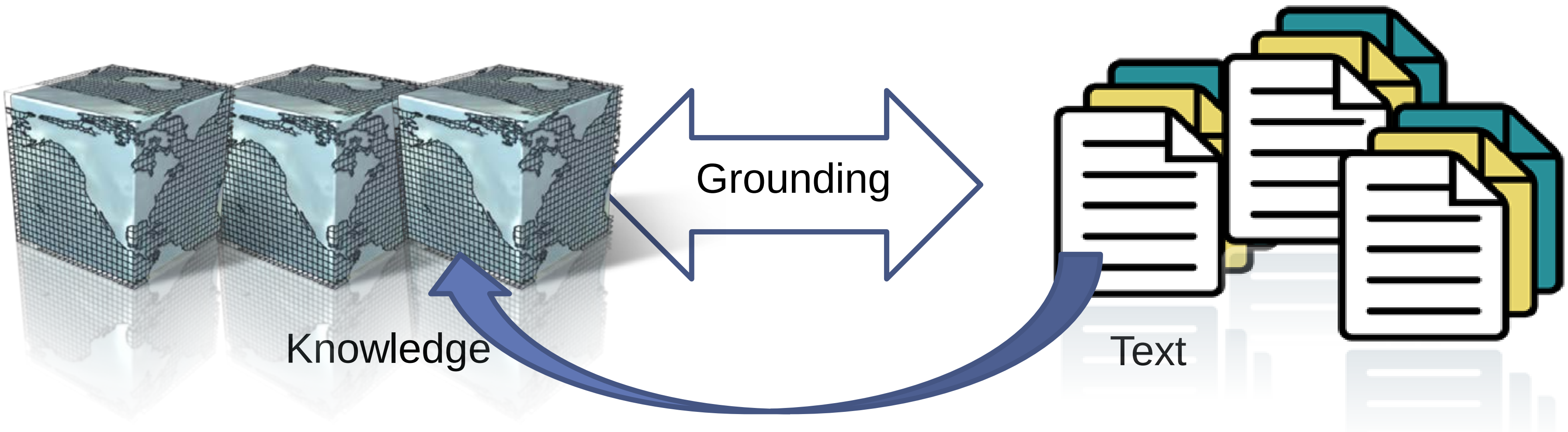
- We have: Wikipedia; Freebase; Customized databases; IMDB…
- How can we have one unified id for all databases?
  - Entity Linking is related to DB Integration
- Different Knowledge bases contain different resources
  - How to utilized them for better grounding?
- How can we use Wikipedia together with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

Handling Various Types of Documents



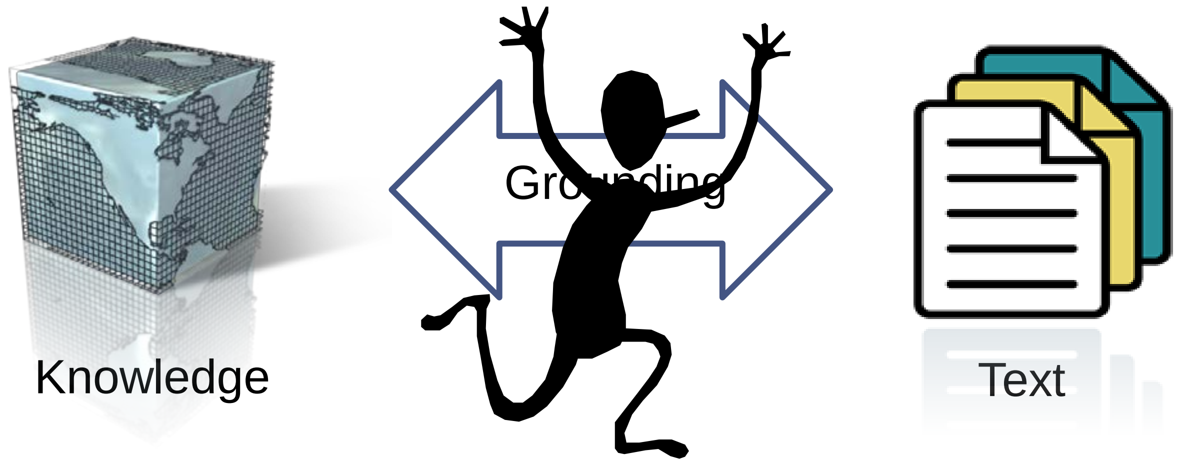
- It is not only text
  - Webpage, queries, tweets, …. All with meta information
  - How can we make use of the meta information?
- Noise
  - How can we develop robust linking techniques on noisy text?
  - Tweet, table + text, broken format, html
- Text in different languages

Machine Reading



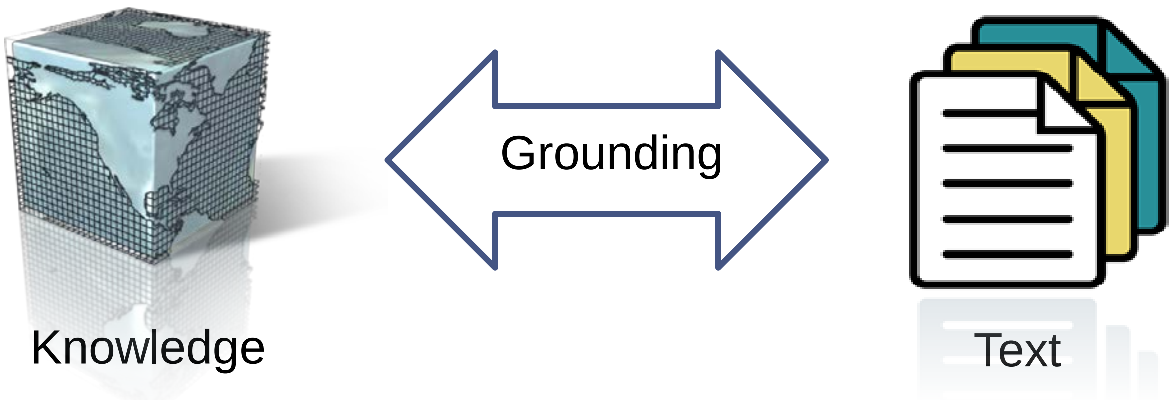
- Can we automatically extract large amounts of knowledge by reading text?
- Grounding should play an important role
  - Grounding → Better understanding of Text → Better Extraction
- How to put it back? Trustworthiness [Dong, et. Al 2014]
- How to handle probabilistic knowledge bases?

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?

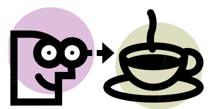
Knowledge Grounding



- Exciting time
  - We only touched the surface of an emerging field
  - Machines can do a much better job remembering knowledge
    - Machines should really be our assistant
  - 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 (Evaluation: September)
  - 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

Demo

• • •

•

•

ERD 2014 Evaluation Service

- http://web-ngram.research.microsoft.com/erd2014

The Governator, as Schwarzenegger came to be known, helped bring about the state's primary election system.

DOC1 0 14 /m/0tc7 The Governator 0.99 0 DOC1 19 33 /m/0tc7 Schwarzenegger 0.97 0

- Encourage teams to build webservice
  - Easy to share, easy to compare
  - Evaluate time as well
  - Easy for future teams to collaborate

- We will continue to keep the service

Demos

- UIUC Wikification Demo
  - http://cogcomp.cs.illinois.edu/demo/wikify/?id=25
- RPI Wikification Demo
  - http://orion.tw.rpi.edu/~zhengj3/wod/search.php
  - http://orion.tw.rpi.edu/~zhengj3/wod/link.php

Thank You – Our Brilliant Wikifiers!