Entity Linking

Bonan Min

bonanmin@gmail.com

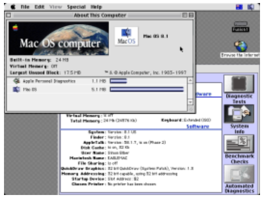
Some slides are based on class materials from Dan Roth, Heng Ji, Ralph Grishman, Thien Huu Nguyen

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".

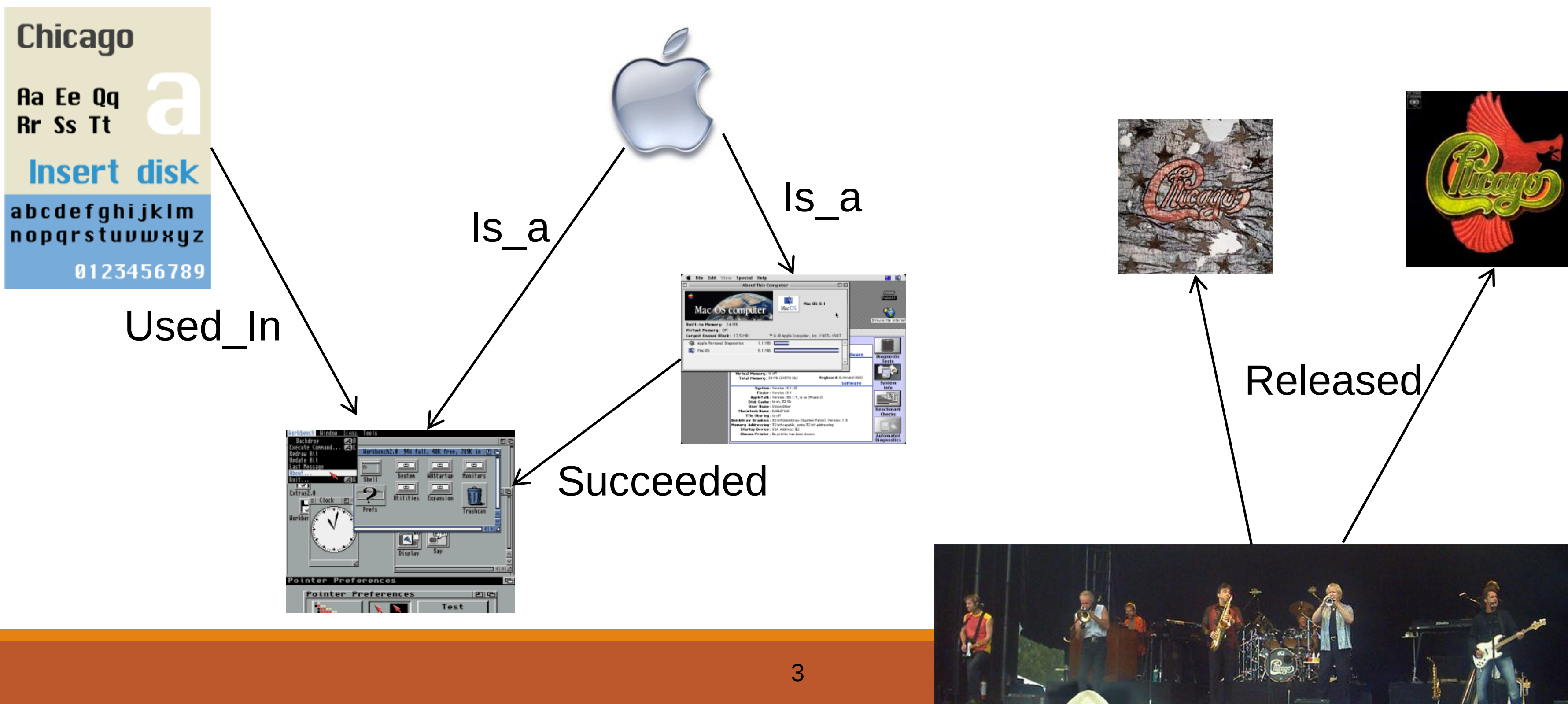
Chicago was used by default for ***Mac*** menus through ***MacOS 7.6***, and ***OS 8*** was released mid-1997..

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

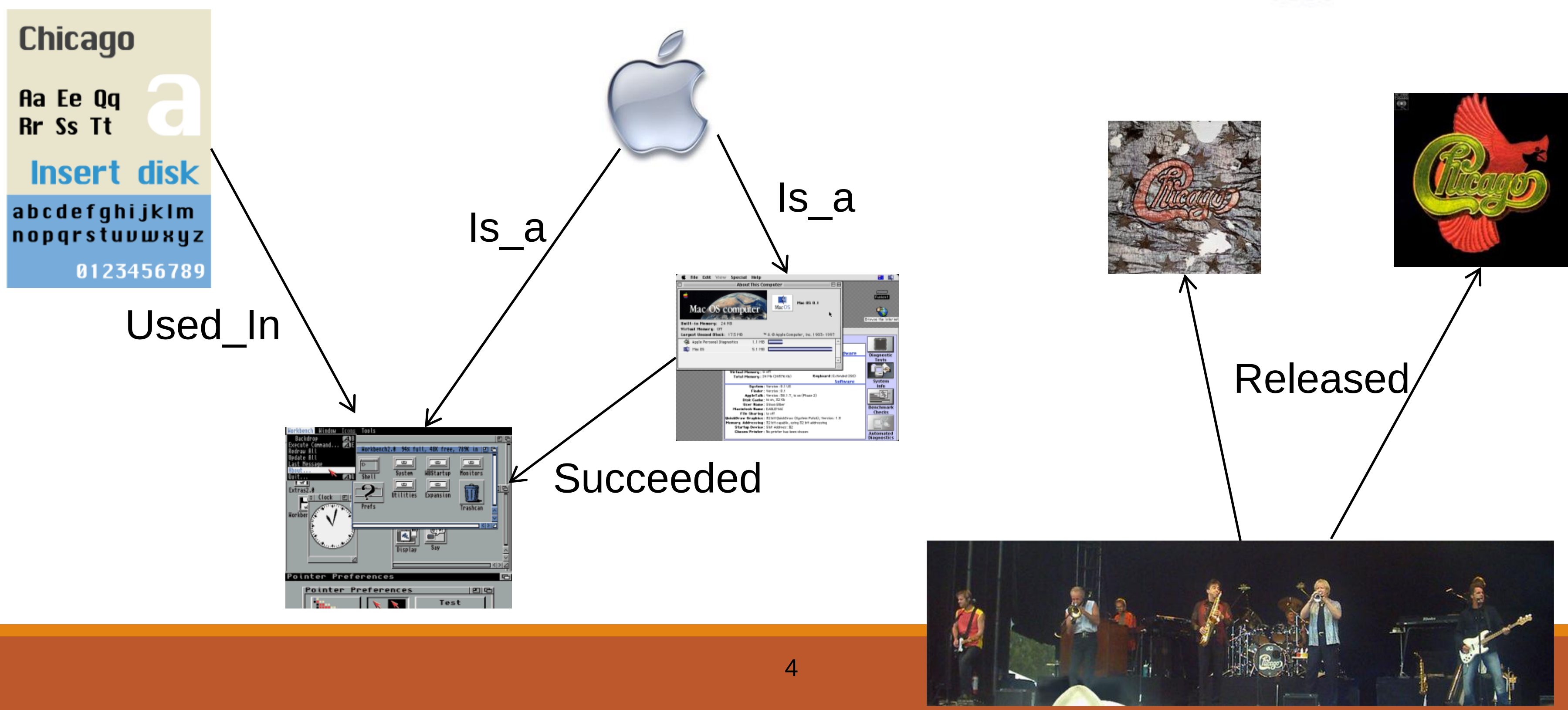


The “Reference” Collection has Structure

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***.



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

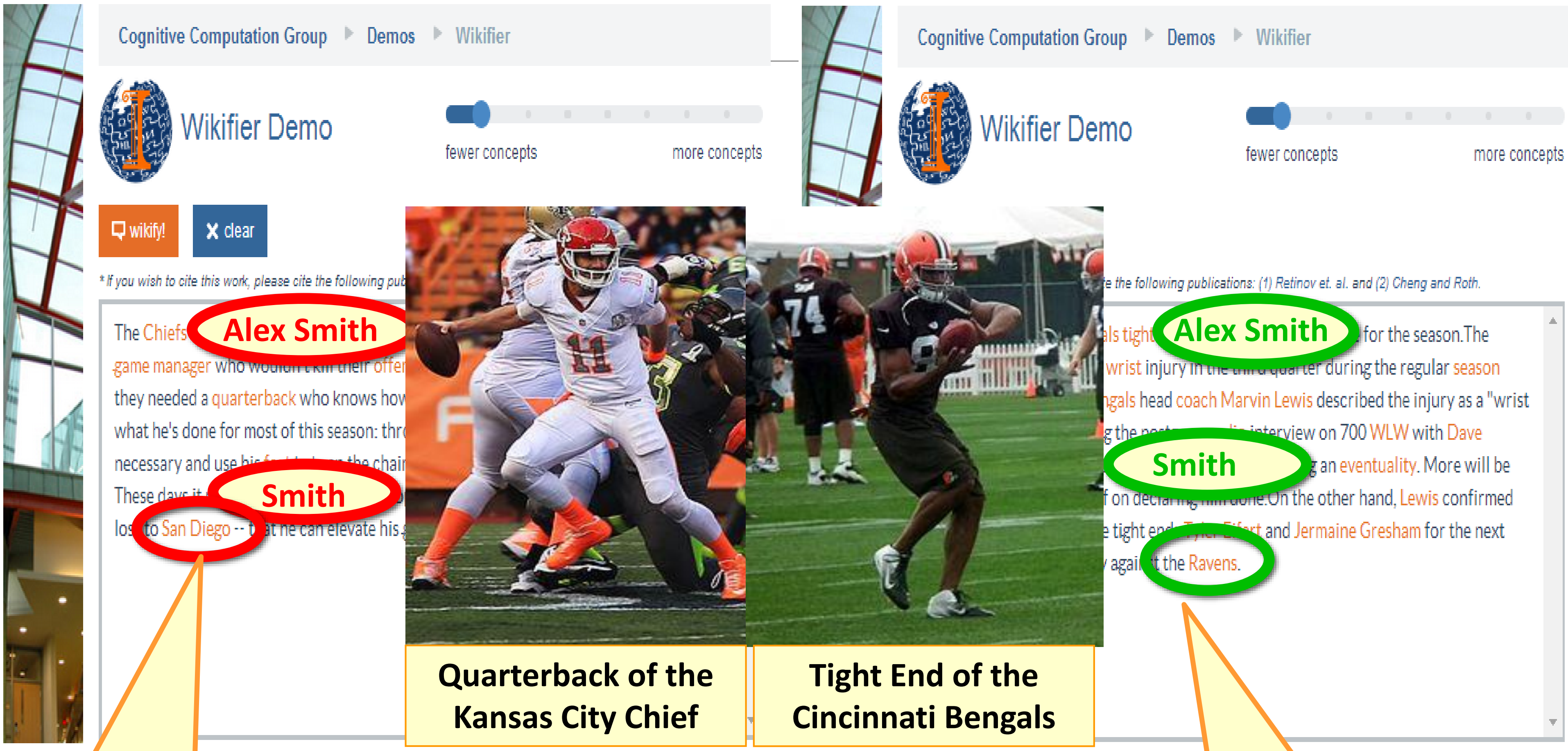
- A given concept could be expressed in many ways

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?

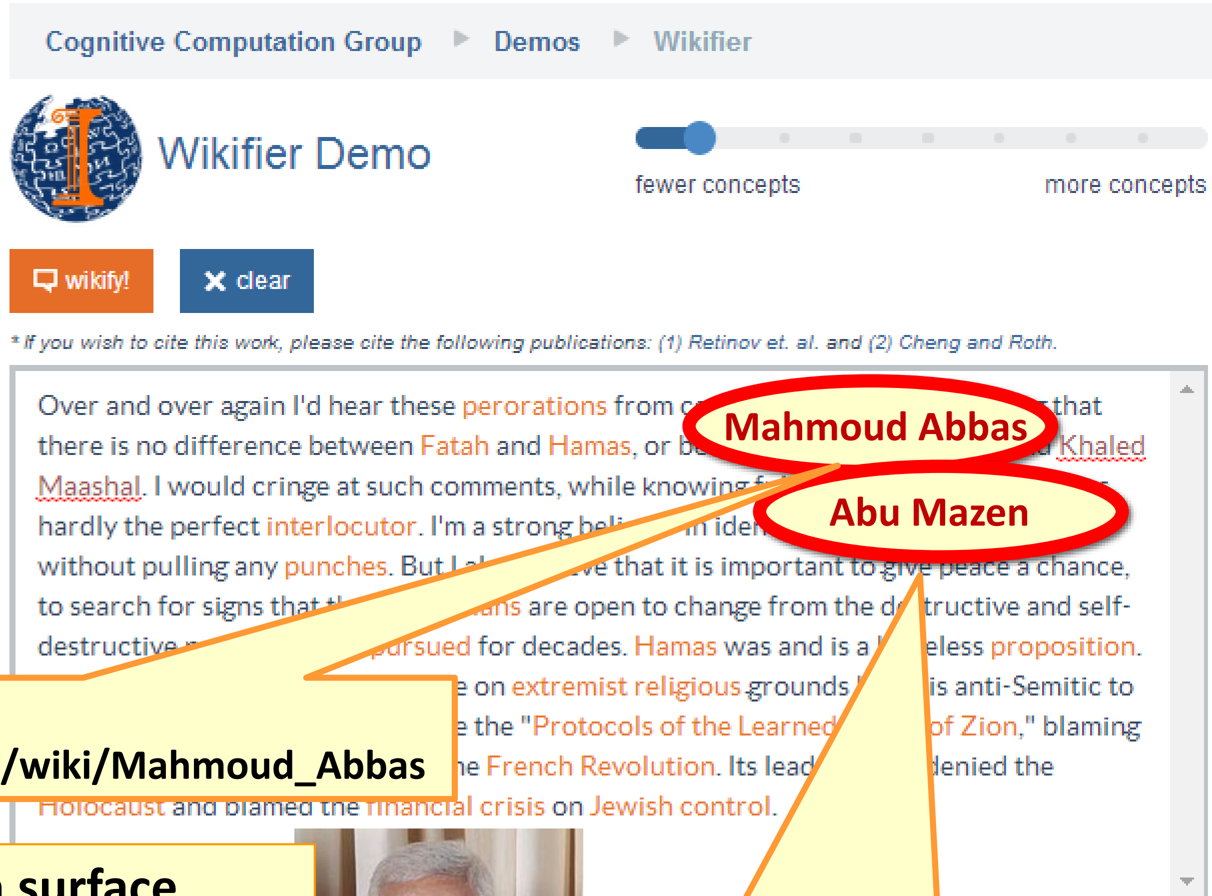


San Diego: **The San Diego Chargers (A Football team)**

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

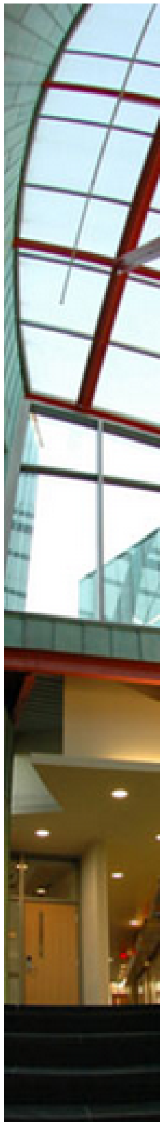
Abu Mazen

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

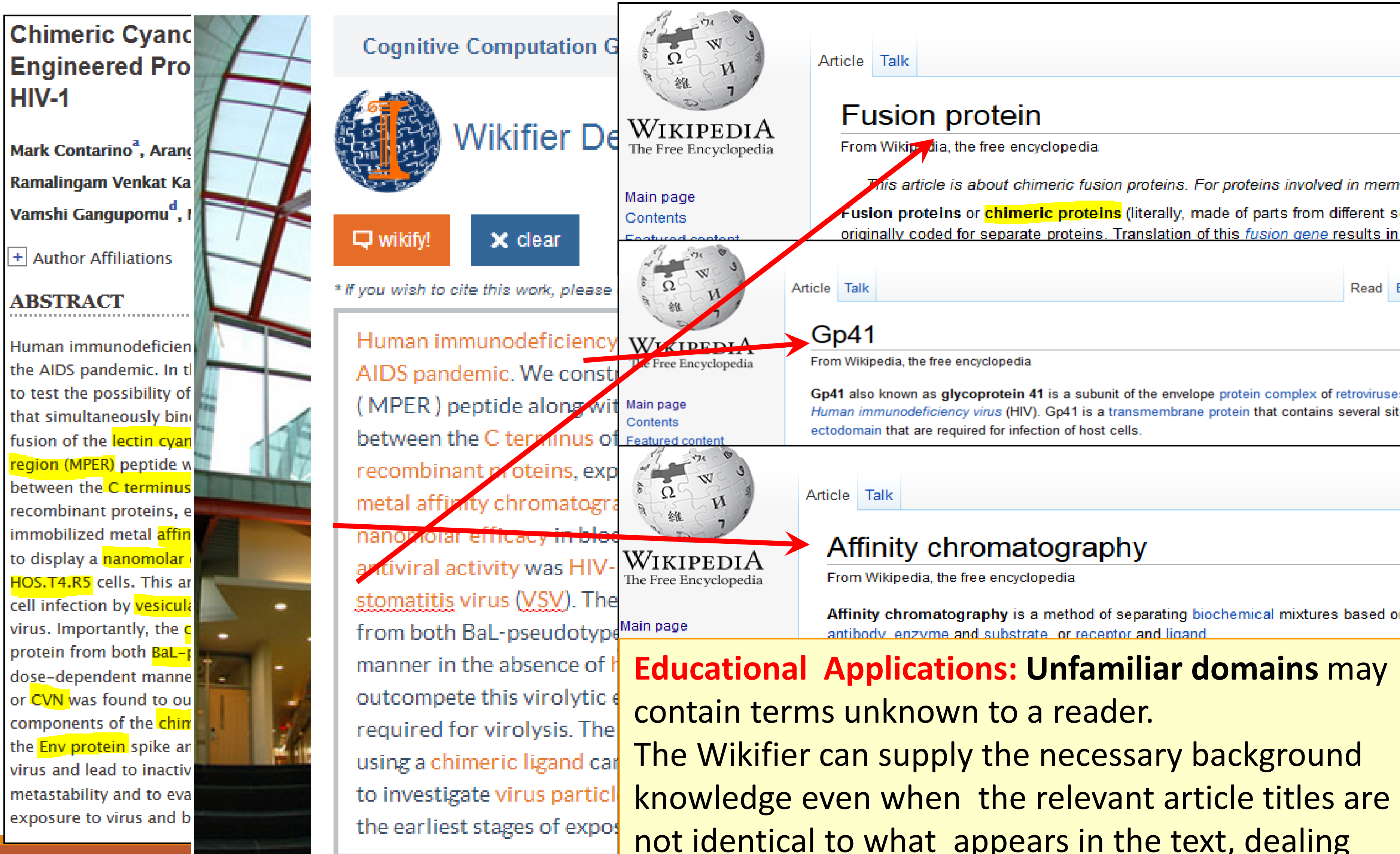
fewer concepts   more concepts

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.**

Entity Linking: 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**)

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

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.

- All type of pages?
  - Content pages
  - Disambiguation pages
  - List pages

What should happened to mentions that do not have entries in the target KB?

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

Null Links

Often, there are multiple sensible links.

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

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

Wikification: Simply map Dorothy Byrne → Null

Entity Linking: If multiple mentions in the given document(s) correspond to the same concept, which is outside KB

- First cluster relevant mentions as representing a single concept
- Map the cluster to Null

Naming Convention

Wikification:

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

Entity Linking:

- Map Mentions to KB Titles
- If multiple mentions in correspond to the same Title, which is outside KB:
  - First cluster relevant mentions as representing a single Title
  - Map the cluster to Null

If the set of target mentions only consists of named entities we call the task: Named Entity [Wikification, Linking]

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

Wikification: Subtasks

Wikification and Entity Linking requires addressing several sub-tasks:

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

High-level Algorithmic Approach

Input: A text document d; Output: a set of pairs (m_i, t_i)

- m_i are mentions in d; $t_j(m_i)$ are corresponding Wikipedia titles, or NIL.

(1) Identify mentions m_i in d

(2) Local Inference

- For each m_i in d:
  - Identify a set of relevant titles $T(m_i)$
  - Rank titles $t_i \in T(m_i)$

    [E.g., consider local statistics of edges [(m_i, t_i), $(m_i, *)$, and $(*, t_i)$] occurrences in the Wikipedia graph]

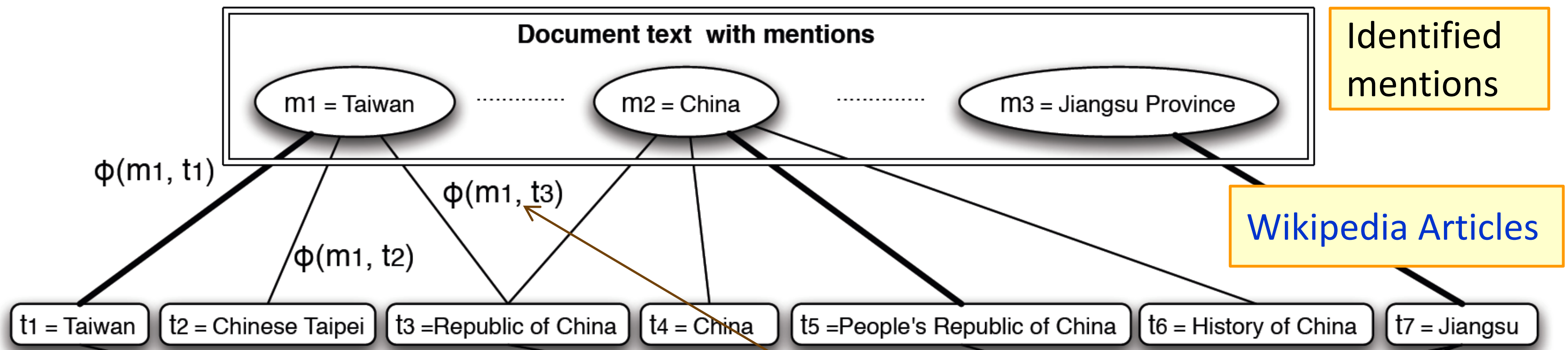
(3) Global Inference

- For each document d:
  - Consider all $m_i \in d$; and all $t_i \in T(m_i)$
  - Re-rank titles $t_i \in T(m_i)$

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

Local Approach

A text Document



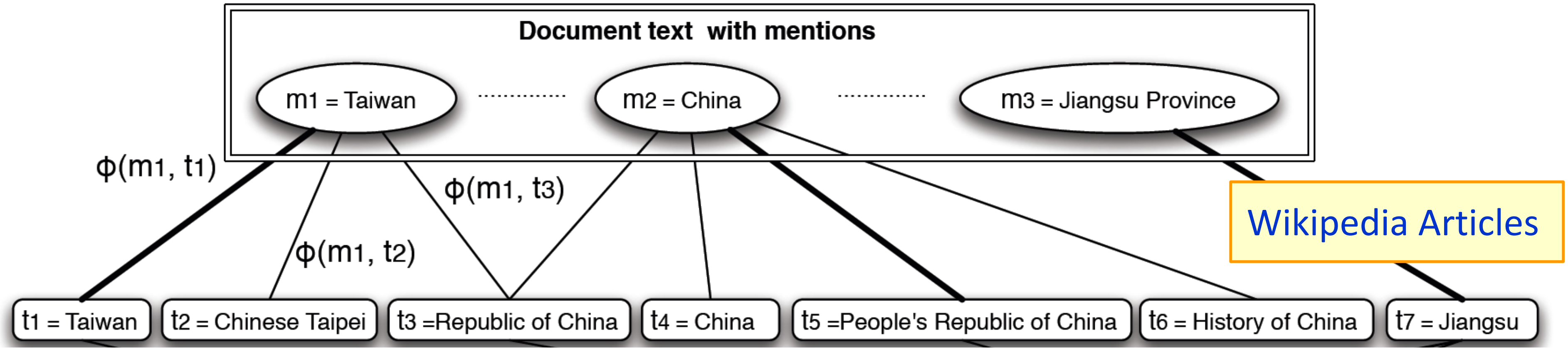
Local score of matching the mention to the title (decomposed by m_i)

- 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,…



Wikipedia Articles

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

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.

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)
  - 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;)
- 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)

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"

AZ

From Wikipedia, the free encyclopedia

"Alitalia"

"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

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

Mention Expansion

Co-reference resolution
- Each mention in a co-referential cluster should link to the same concept
- Canonical names are often less ambiguous
- Correct types: “*Detroit*” = “*Red Wings*”; “*Newport*” = “*Newport-Gwent Dragons*”

Known Aliases
- KB link mining (e.g., Wikipedia “re-direct”) (Nemeskey et al., 2010)
- Patterns for Nicknames/ Acronym mining (Zhang et al., 2011; Tamang et al., 2012)

  “full-name” (acronym) or “acronym (full-name)”, “city, state/country”

  Statistical models such as weighted finite state transducer (Friburger and Maurel, 2004)

  - CCP = “Communist Party of China”; “MINDEF” = “Ministry of Defence”

  Ambiguity drops from 46.3% to 11.2% (Chen and Ji, 2011; Tamang et al., 2012).

Local Inference: 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)

Local Inference: 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.

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(Title|”Chicago”)

P(t|m): “Commonness”

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

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

“Commonness” Not robust across domains

Formal Genre

Ratinov et al. (2011)

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

Tweets

Meij et al. (2012)

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

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

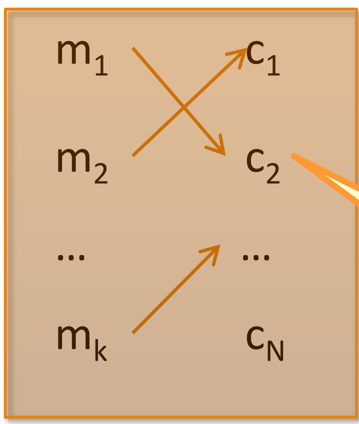
- Use topical/semantic **coherence** measures between the set of referent concepts for all mentions in a document

Context Similarity Measures

Determine assignment that maximizes pairwise similarity

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

Γ



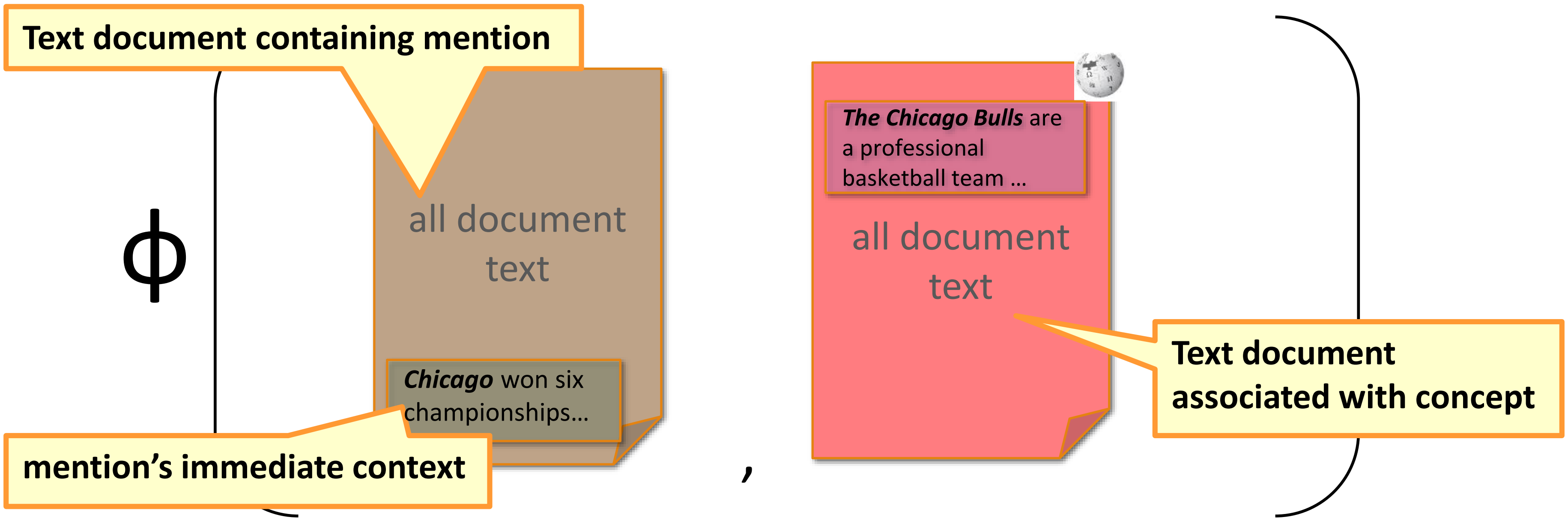
Mention-concept assignment

Mapping from mentions to titles

Feature vector to capture degree of contextual similarity

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

Context Similarity Measures: *Context Source*



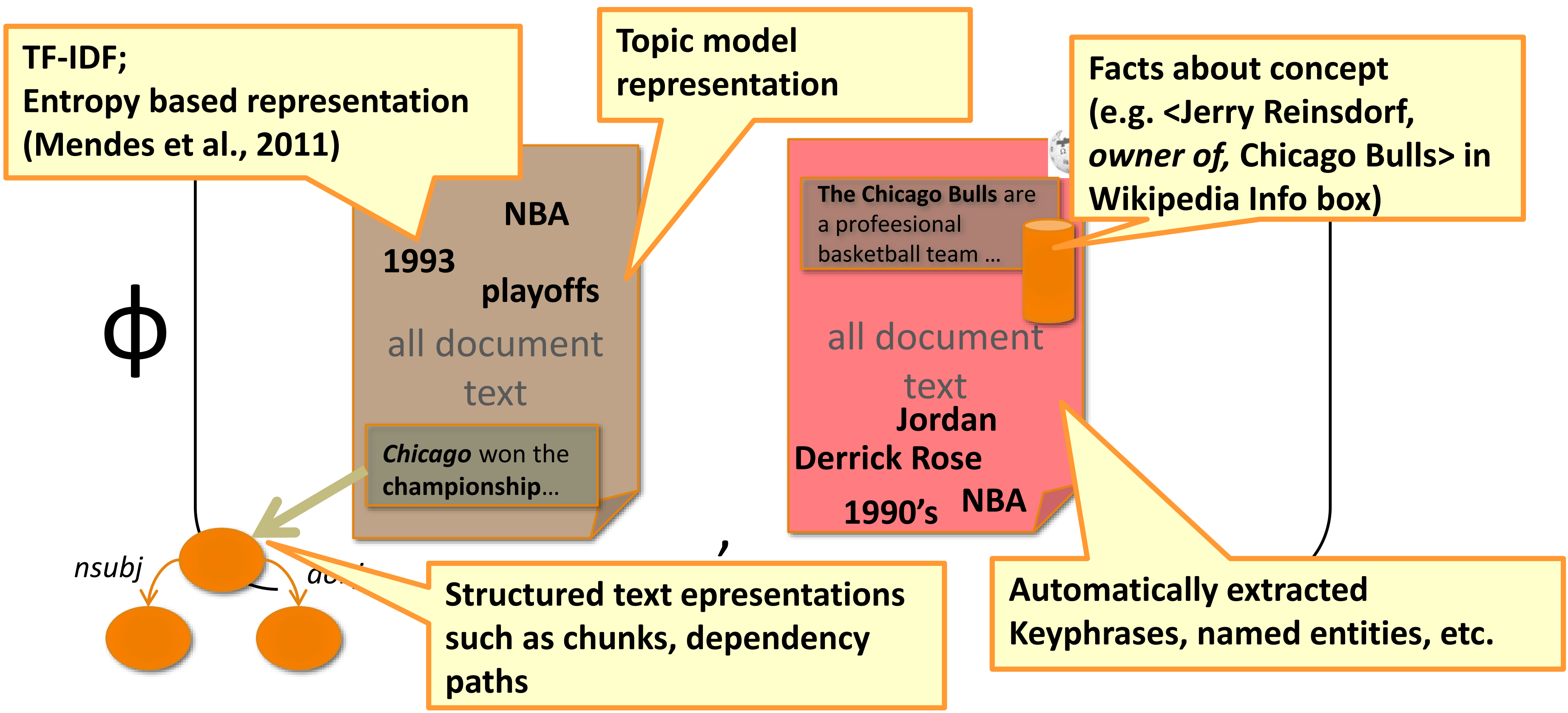
Varying notion of distance between mention and context tokens

- Token-level, discourse-level

Varying granularity of concept description

- Synopsis, entire document

Context Similarity Measures: *Context Analysis*



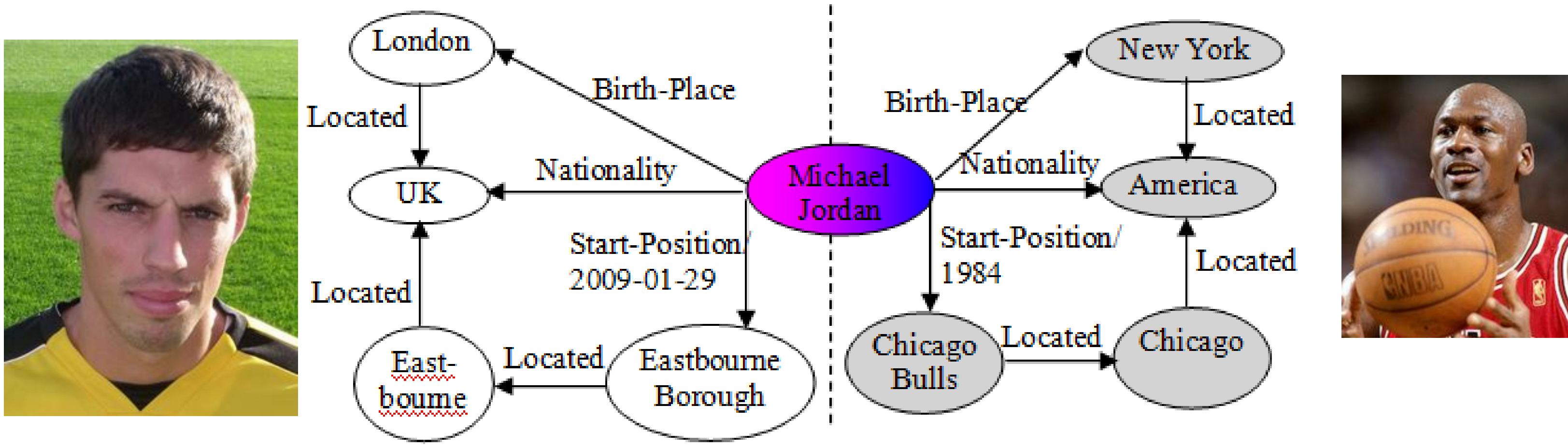
- Context is processed and represented in a variety of ways

Typical Features for Ranking

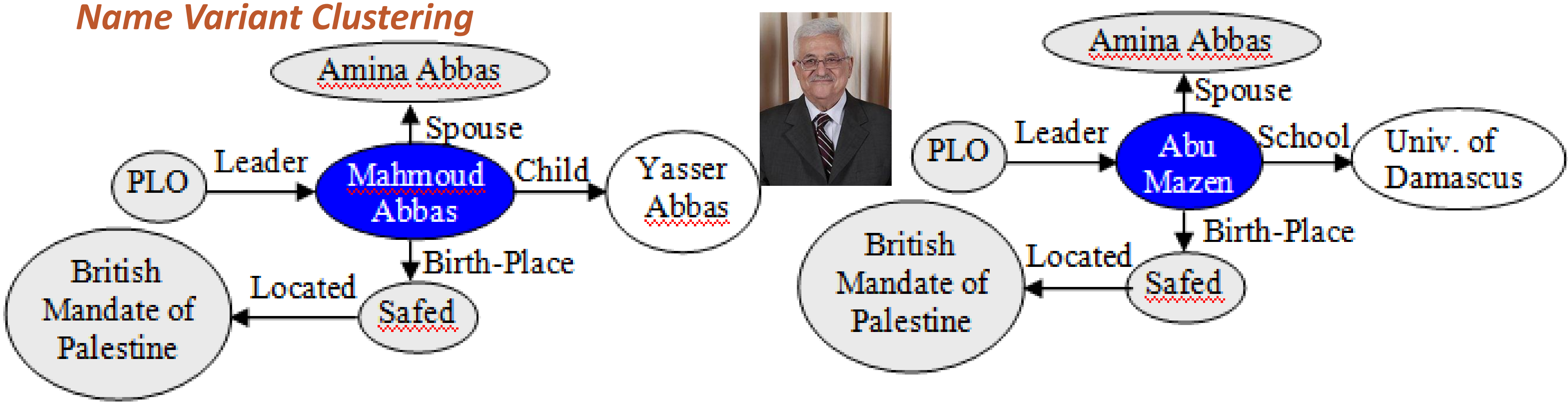
Mention/Concept Attribute		Description
Name	Spelling match	Exact string match, acronym match, alias match, string matching…
	KB link mining	Name pairs mined from KB text redirect and disambiguation pages
	Name Gazetteer	Organization and geo-political entity abbreviation gazetteers
Document surface	Lexical	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



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

singl

final

gain

half

female

Li Na

Pakistan

relation

vice president

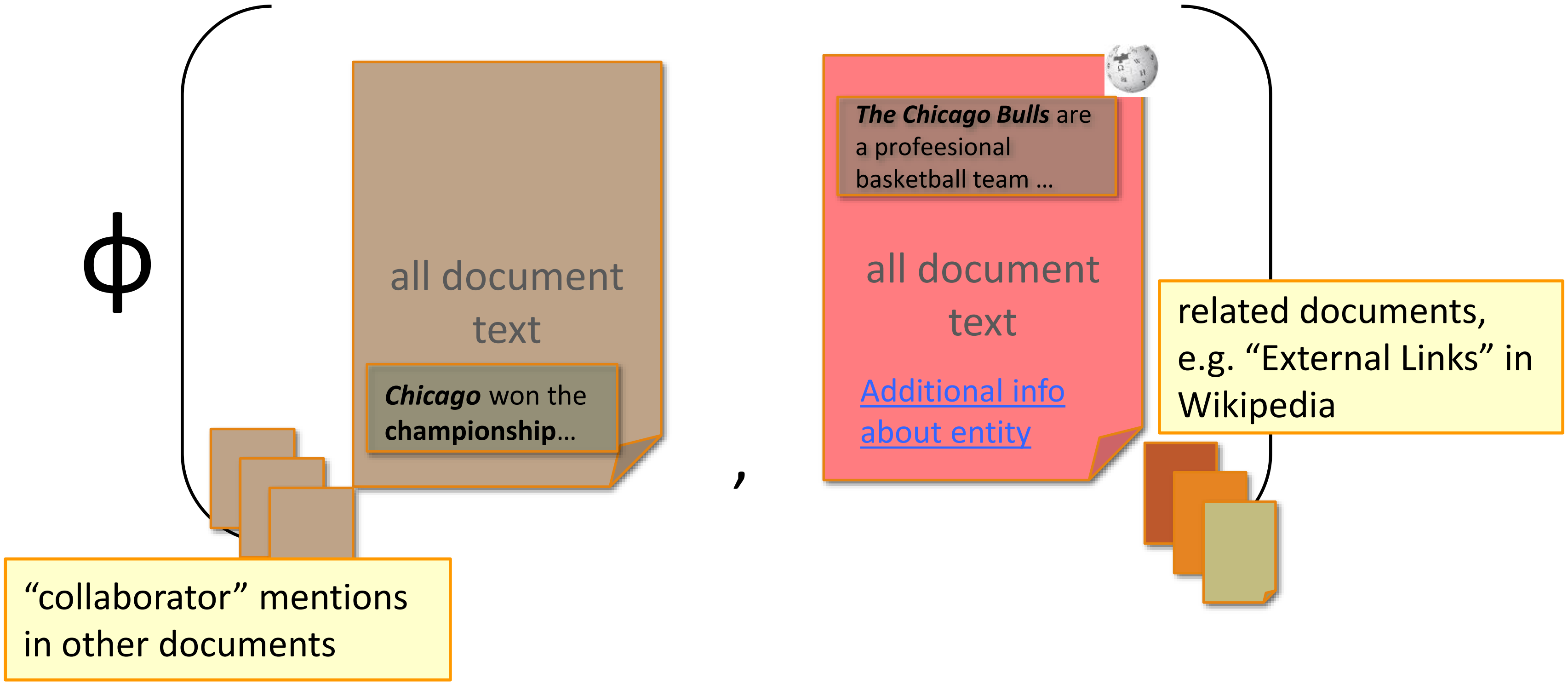
express

country

Prime minister

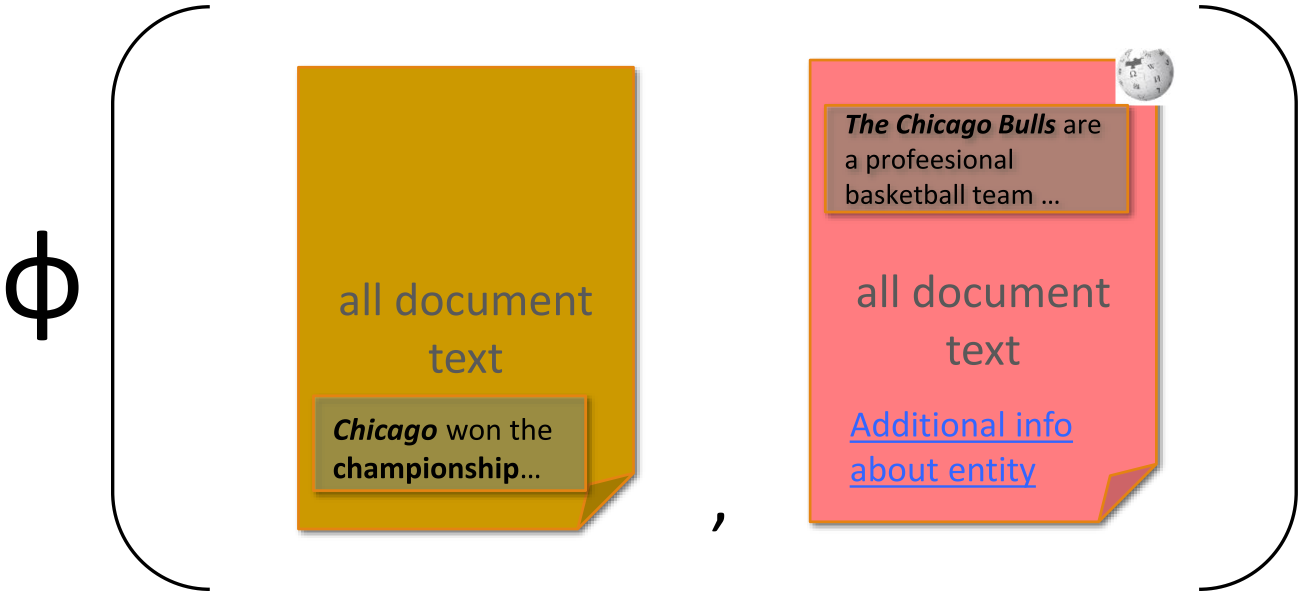
Topical features or topic based document clustering for context expansion (Milne and Witten, 2008; Syed et al., 2008; Srinivasan et al., 2009; Kozareva and Ravi, 2011; Zhang et al., 2011; Anastacio et al., 2011; Cassidy et al., 2011; Pink et al., 2013)

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*

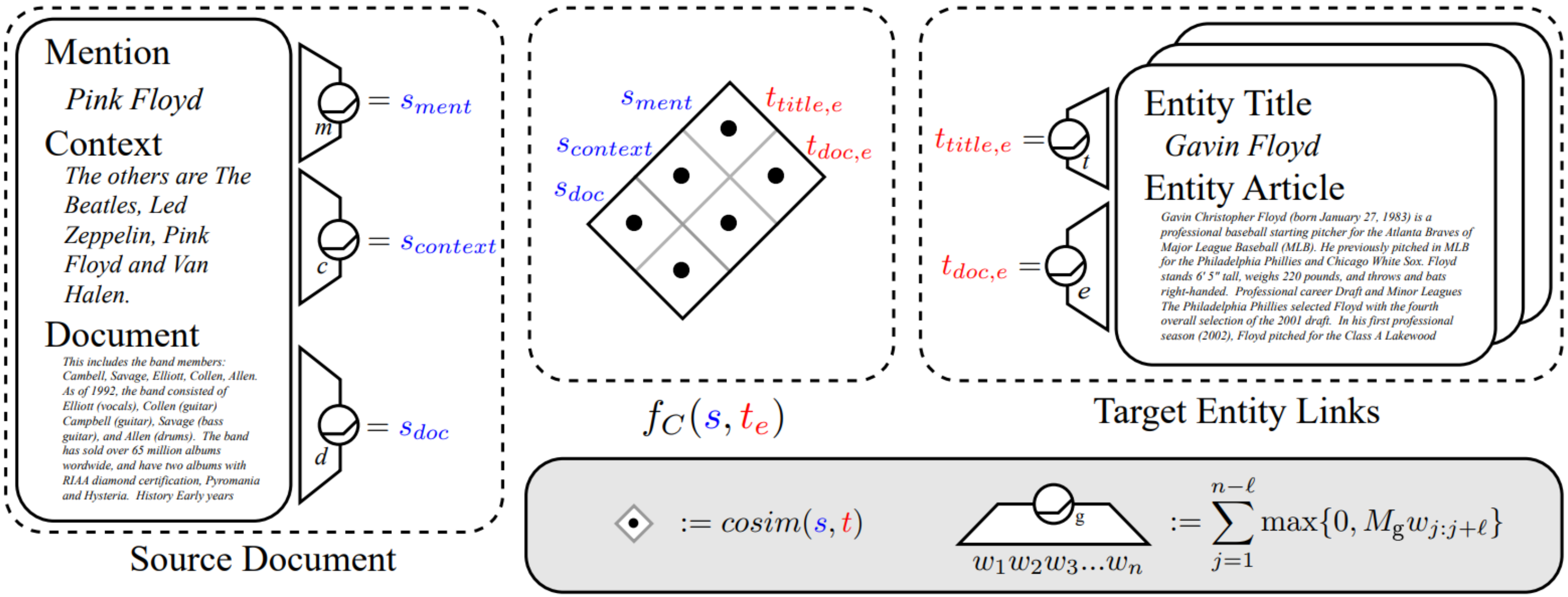


- Cosine similarity (via TF-IDF)
- Other distance metrics (e.g. Jaccard)
- 2nd order vector composition (Hoffart et al., EMNLP2011)
- Mutual Information

NN for Context Similarity

Extraction of convolutional vector space features $f_C(s, t_e)$, Use CNN for

- Three types of information from the input document
- two types of information from the proposed title



Alternative context representation: BERT

Matthew Francis-Landau, Greg Durrett and Dan Klein. Capturing Semantic Similarity for Entity Linking with Convolutional Neural Networks. NAACL-HLT 2016

Samuel Broscheit. Investigating Entity Knowledge in BERT with Simple Neural End-To-End Entity Linking. CoNLL 2019.

Unsupervised vs. Supervised Ranking

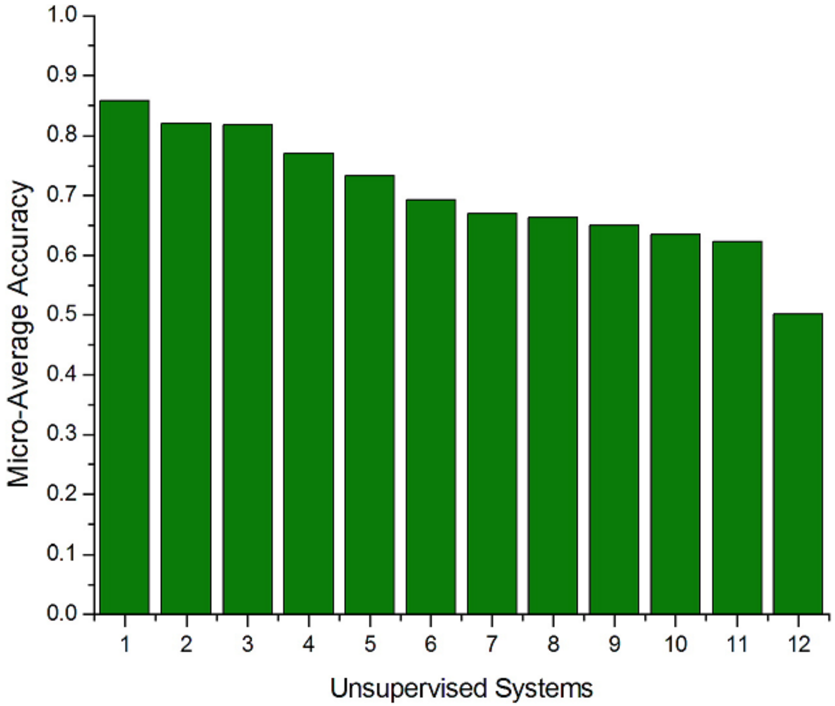
Unsupervised or weakly-supervised learning (Ferragina and Scaiella, 2010)

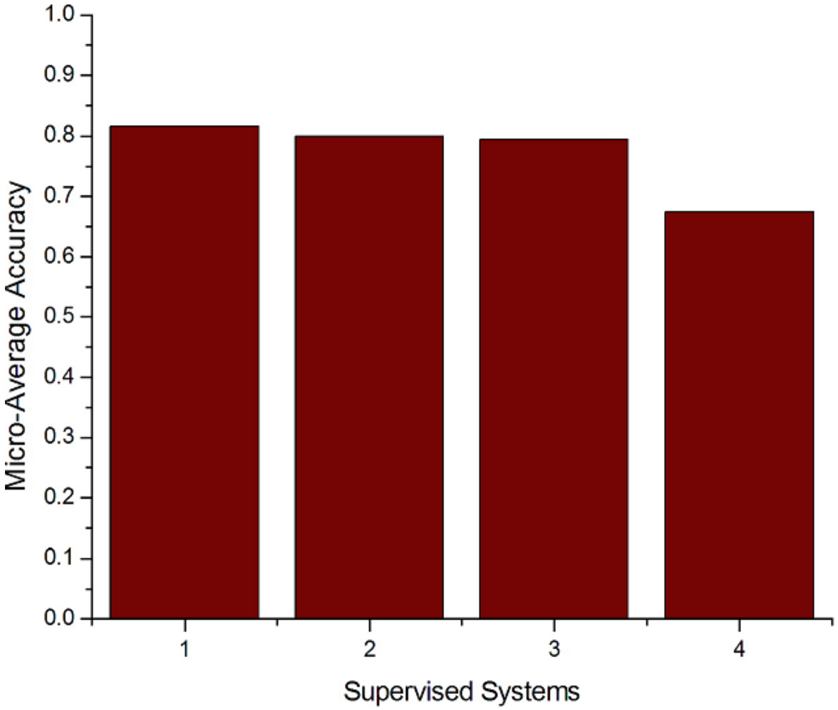
- Annotated data is minimally used to tune thresholds and parameters
- The similarity measure is largely based on the unlabeled contexts

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)

Unsupervised vs. Supervised Ranking



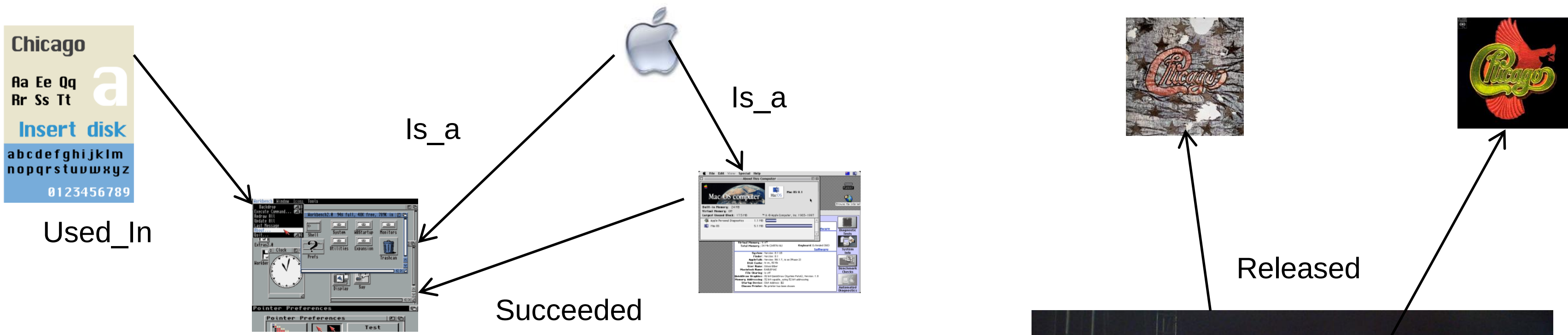


KBP2010 Entity Linking Systems (Ji et al., 2010)

Conceptual Coherence

Recall: The reference collection (might) have structure.

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



Hypothesis:

- Textual co-occurrence of concepts is reflected in the KB (Wikipedia)

Incite:

- Preferred disambiguation Γ contains structurally coherent concepts

Co-occurrence(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.

The city senses of Boston and Chicago appear together often.

Rock music and albums appear together often

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

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

Global Ranking

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

How to approximate the “global semantic context” in the document”?

- It is possible to only use non-ambiguous mentions as a way to approximate it.

How to define relatedness between two titles? (What is Ψ?)

Title Coherence & Relatedness

Let c, d be a pair of titles …

Let C and D be their sets of incoming (or outgoing) links

- Unlabeled, directed link structure

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

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

See García et al. (JAIR2014) for variational details

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

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

Let C and D ∈{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($a \rightarrow b$)	equals 1 if a links to b, and 0 otherwise.
$P(a \rightarrow b)$	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: $\log \frac{P(b\|a)}{P(b)} = \log \frac{P(a\|b)}{P(a)} = \log \frac{\|in(b) \cap in(a)\|\|W\|}{\|in(b)\|\|in(a)\|}$

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.

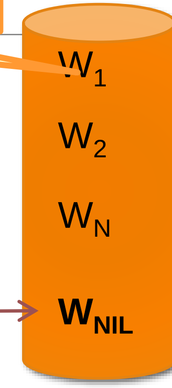
Mapping to NIL is challenging in both cases

NIL Detecti

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?

{ , *<u>NIL</u>* }



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

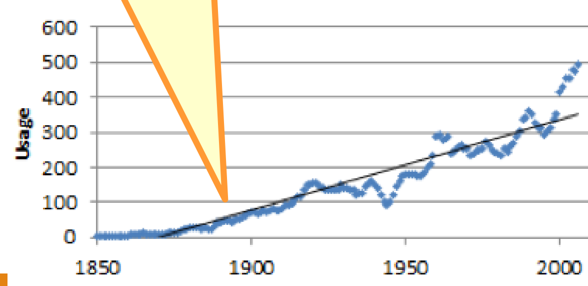
Is it an *entity*?

Concept Mention Identification (above)

Not all NP's are linkable

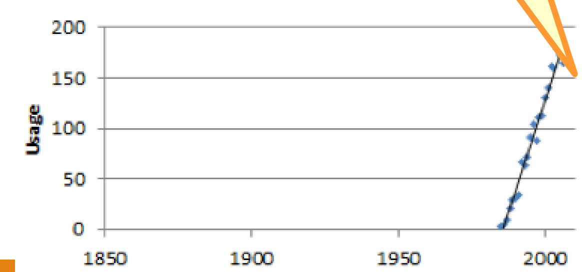
No spike: Not an entity

"Prices Quoted"



Sudden Google Books frequency spike: Entity

"Soluble Fiber"



NIL Clustering

"All in one"

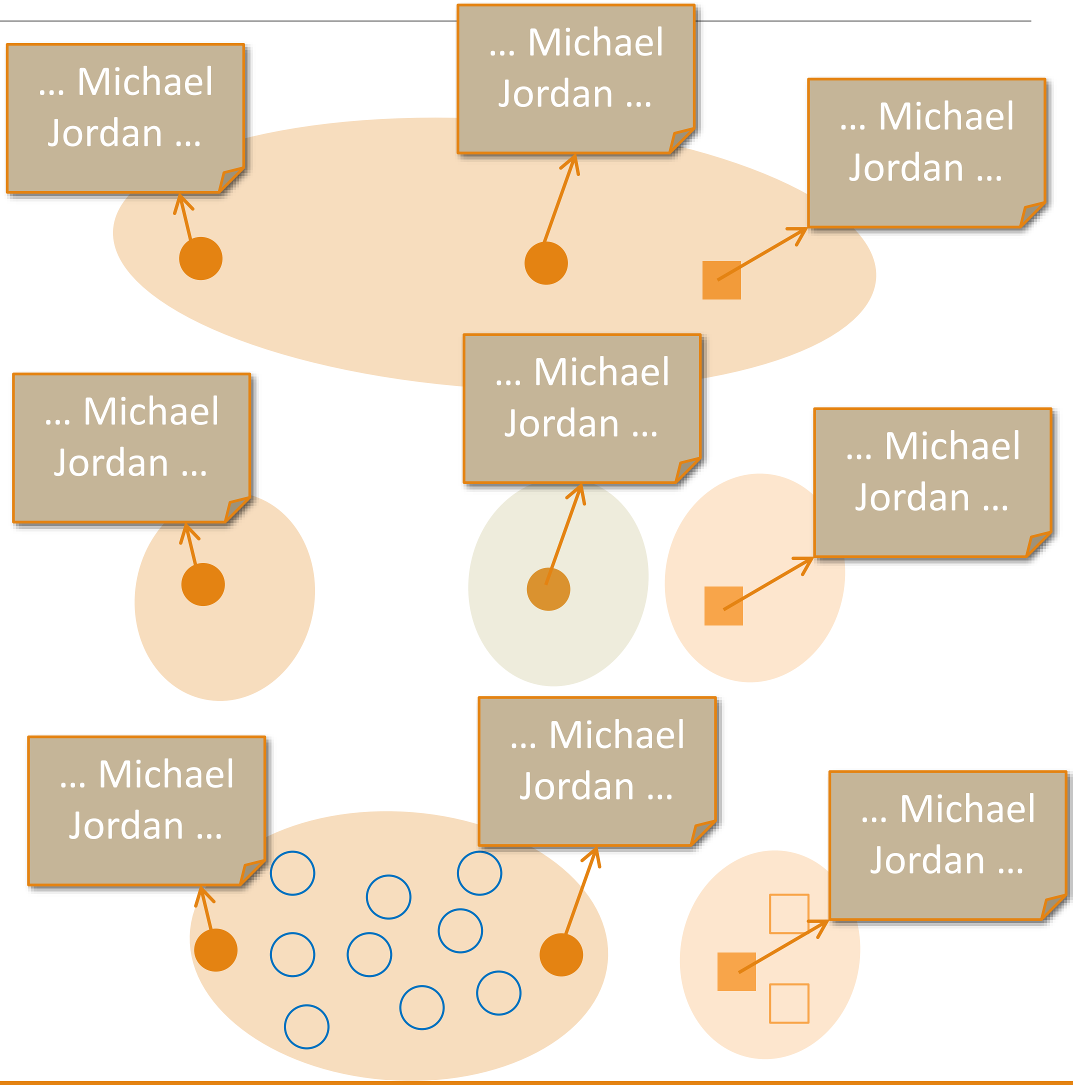
Simple string matching

"One in one"

Often difficult to beat!

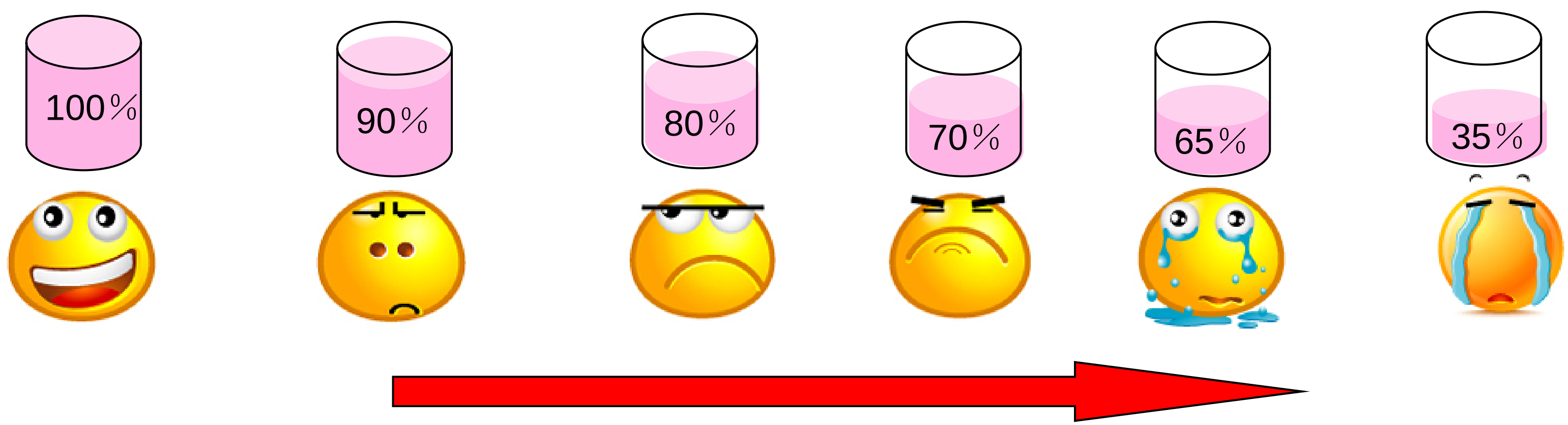
Collaborative Clustering

Most effective when ambiguity is high



End-to-end Wikification: Pipeline Approach

Texts → Entity/ Concept Mention Extraction → Coreference Resolution → Entity Linking → NIL Entity Clustering → Slot Filling → KB



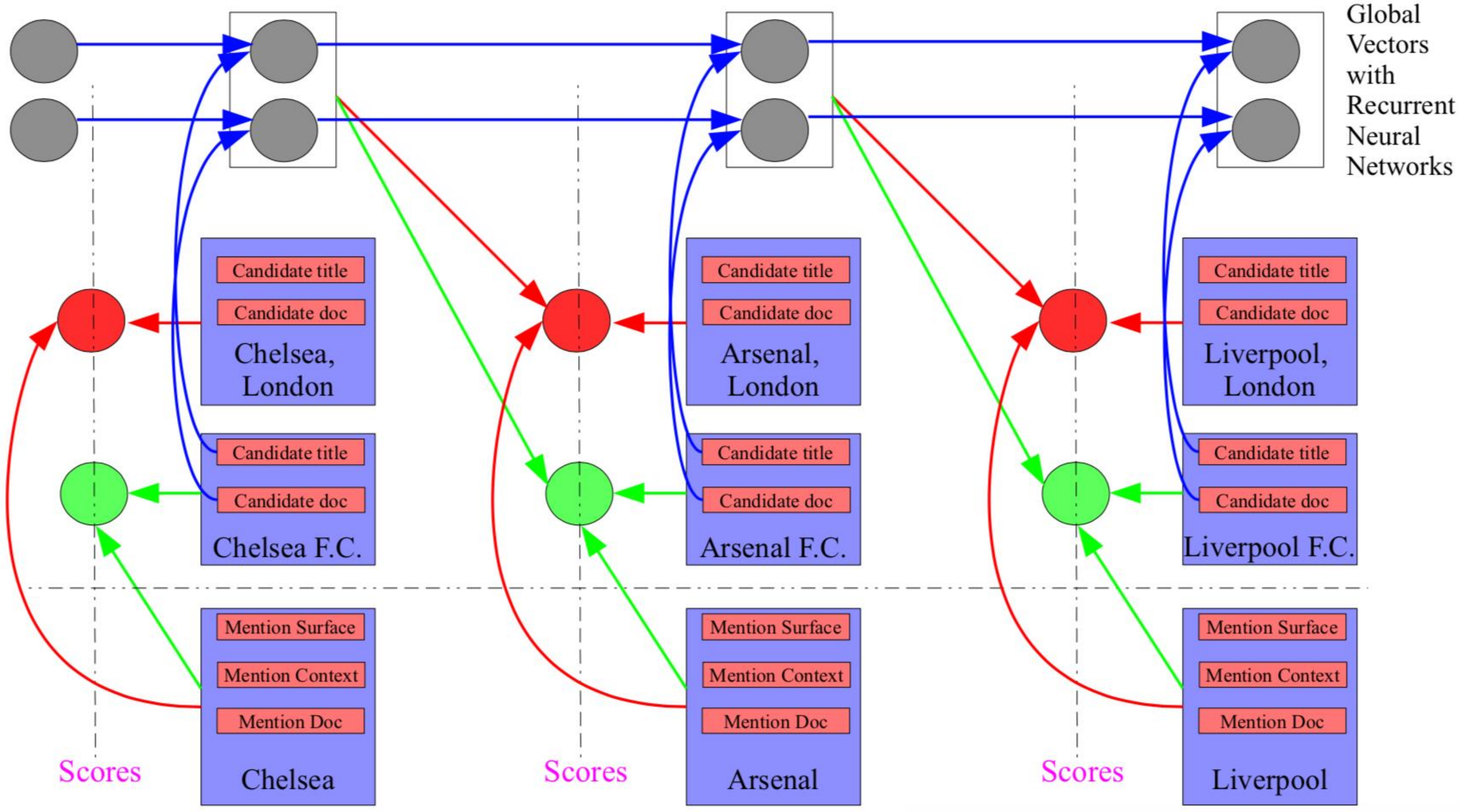
Errors are compounded from stage to stage

No interaction between individual predictions

Incapable of dealing with global dependencies

End-to-end NN joint models can help!

NN Joint Learning for Entity Linking



Thien Huu Nguyen, Nicolas Fauceglia, Mariano Rodriguez Muro, Oktie Hassanzadeh, Alfio Massimiliano Gliozzo and Mohammad Sadoghi. Joint Learning of Local and Global Features for Entity Linking via Neural Networks. COLING 2016.

Scaling Up

Potential scale for cross-doc coref much larger

- collection may have 10^7 documents with 10-100 entities each: 10^9 document-level entities
- computing all pairwise similarities infeasible
- use hierarchical approach to divide set
  - analog of entity-mention representation within a document
  - potentially with multiple levels ('sub-entities')