

# **Part I**

# **Entity Linking**

# Outline

- Part 1 – Entity Linking
  - introduction
  - methods
  - evaluation
  - test collections
  - toolkits
  - open challenges

# **Introduction**

article discussion edit this page history

You're running!

# Plant

From Wikipedia, the free encyclopedia

*For other uses, see Plant (disambiguation).*

**Plants** are a major group of living things including familiar organisms such as trees, flowers, herbs, ferns, and mosses.

About 350,000 species of plants, defined as seed plants, bryophytes, ferns and fern allies, have been estimated to exist. As of 2004, some 287,655 species had been identified, of which 258,650 are flowering and 15,000 bryophytes.

Tree

From Wikipedia, the free encyclopedia

For other senses of the word, see tree (disambiguation)

A tree is a large, perennial, woody plant. Though there is no set definition regarding minimum size, the term generally applies to plants at least 6 m (20 ft) high at maturity and, more importantly, having



Fossil range: Middle-Late Ordovician - Recent



Species

From Wikipedia, the free encyclopedia

This article is about biology. For the movie, see Species.

In biology, a species is one of the basic units of biodiversity. In classification, a species is assigned a two-part name; the genus is listed first (with its leading letter capitalized), followed by the species. For example, humans belong to the genus *Homo*, and species *Homo sapiens*. The name of the species is the whole, just the second term (which may be called *specific epithet*).

Image taken from Mihalcea and Csomai (2007). **Wikify!: linking documents to encyclopedic knowledge.** In CIKM '07.

**Let's learn something about  
Spin-Optical Metamaterial**

REPORT



# Spin–Optical Metamaterial Route to Spin–Controlled Photonics

Nir Shitrit, Igor Yulevich, Elhanan Maguid, Dror Ozeri, Dekel Veksler, Vladimir Kleiner, Erez Hasman\*

Author Affiliations

\*Corresponding author. E-mail: [mehasman@technion.ac.il](mailto:mehasman@technion.ac.il)

ADV

ABSTRACT

EDITOR'S SUMMARY

Spin optics provides a route to control light, whereby the photon helicity (spin angular momentum) degeneracy is removed due to a geometric gradient onto a metasurface. The alliance of spin optics and metamaterials offers the dispersion engineering of a structured matter in a polarization helicity-dependent manner. We show that polarization-controlled optical modes of metamaterials arise where the spatial inversion symmetry is violated. The emerged spin-split dispersion of spontaneous emission originates from the spin-orbit interaction of light, generating a selection rule based on symmetry restrictions in a spin-optical metamaterial. The inversion asymmetric metasurface is obtained via anisotropic optical antenna patterns. This type of metamaterial provides a route for spin-controlled nanophotonic applications based on the design of the metasurface symmetry properties.



WC  
IN SC  
forgi  
patl  
in h

Received for publication 7 January 2013.

Degenerate energy levels – Wikipedia, the free encyclopedia

W en.wikipedia.org/wiki/Degenerate\_energy\_level

Degenerate energy levels – Wikipedia, the free encyclopedia

Input Text

Italiano English

momentum) degeneracy of spin optical modes in a polarized matter emerged spin-split interaction of light optical metamaterials optical antenna pads nanophotonic applications

Tagged text Topics

Spin optics provides momentum) degeneracy of spin matter in a polarized optical modes emerged spin-split interaction of light optical metamaterials optical antenna pads

WIKIPEDIA The Free Encyclopedia

Main page Contents Featured content Current events Random article Donate to Wikipedia Interaction Help About Wikipedia Community portal Recent changes Contact Wikipedia Toolbox Print/export Languages العربية Deutsch Español Esperanto فارسی Français 한국어 עברית മലയാളം Nederlands 日本語 Norsk nynorsk Polski Português Русский

Article Talk Read Edit View history

## Degenerate energy levels

From Wikipedia, the free encyclopedia  
(Redirected from Degenerate energy level)

This article is about different quantum states having the same energy. For other uses, see Degeneracy.  
"Quantum degeneracy" redirects here. It sometimes refers to a degenerate matter.

 This article needs additional citations for verification. Please help improve this article by adding citations to reliable sources. Unsourced material may be challenged and removed. (February 2009)

In quantum mechanics, a branch of physics, two or more different states of a system are said to be degenerate if they are all at the same energy level. It is represented mathematically by the system having more than one linearly independent eigenstate with the same eigenvalue. Conversely, an energy level is said to be degenerate if it contains two or more different states at a particular energy level is called the level's degeneracy, and this phenomenon is generally known as a quantum degeneracy.

From the perspective of quantum statistical mechanics, several degenerate states at the same level are all equally probable of being filled.

Contents [hide]

- 1 Mathematics
- 2 Examples
- 3 Perturbation
- 4 See also
- 5 Further reading

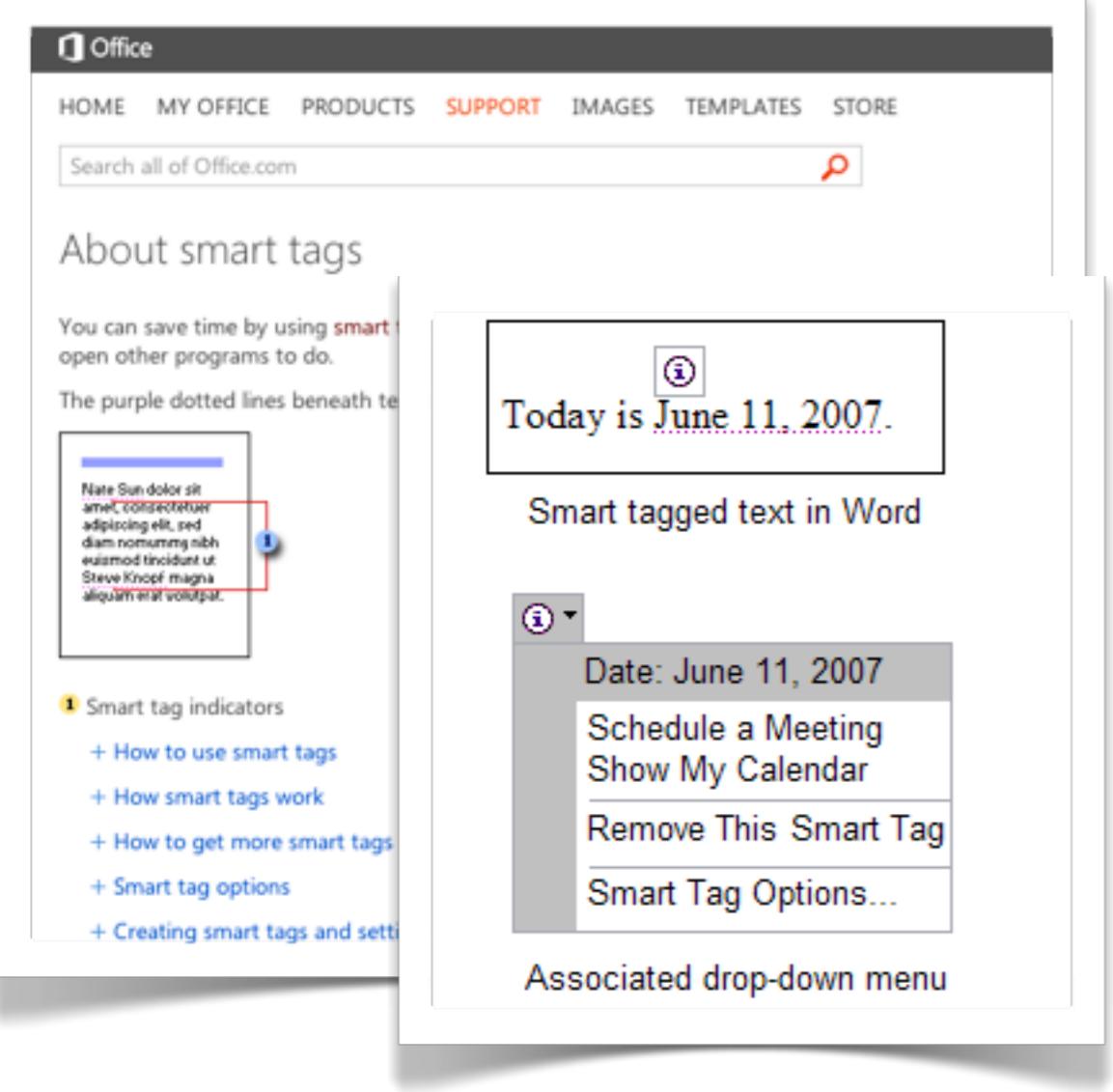
### Mathematics

The term comes from the fact that, for a point spectrum Hamiltonian  $H$ , degenerate eigenstates correspond to identical eigenvalues. Since eigenvalues correspond to roots of the characteristic polynomial, the word degeneracy here has the same meaning as the common mathematical usage of the word.

The eigenvalue  $\lambda$  is called nondegenerate (or simple) when its corresponding eigenvector is unique up to a constant factor, or, the same, the corresponding eigenspace is one-dimensional. Indeed, the eigenspace  $\{|\psi\rangle : H|\psi\rangle = \lambda|\psi\rangle\}$  (in bra-ket notation) is not necessarily one-dimensional. If there exist at least two linearly independent ket-vectors in it, then this eigenvalue is called degenerate. Its degree of degeneracy is then the dimension of the eigenspace, which is the same as the number of distinct (linearly independent) quantum states associated with it.

### Examples

In atomic physics, electron's energy levels are often degenerate, where different possible occupation states for particles may be related by symmetry. For example, in the hydrogen atom, for a given principal quantum number  $n$ , there exist several states which have that energy, but differ in the eigenvalues of angular momentum  $L^2$ , spin component  $S_z$  and so on. The eigenvalue of an operator which is invariant under a symmetry operation is called a quantum number.



# Microsoft Smart Tags

The screenshot shows a web browser window with the URL <http://cse.unl.edu/~choueiry/S01-476-876/>. The page content includes information about an instructor and TA. On the right side, a context menu is open over a link to "choueiry@cse.unl.edu", listing various actions such as email address, ISBN, and link options. The menu also includes "Remove AutoLinks" and "Change Default Provider...".

Instructor: Prof. Berthe Y. Choueiry  
Room 104, Ferguson Hall,  
[choueiry@cse.unl.edu](mailto:choueiry@cse.unl.edu), tel: (402)472-5444.  
Office hours: Mon/Fri from 4:45 p.m. to 5:30 p.m.

TA: Mr. Daniel Buettner (Dan).  
email: [bueettner@cse.unl.edu](mailto:bueettner@cse.unl.edu)  
Office hours in Room 16 or Room 17, Ferguson Hall  
Mon 10.30-11.30 am; Tue 10.00-11.00 am

# Google toolbar

Treaty of Versailles – Wolfram|Alpha

http://www.wolframalpha.com/input/?i=Treaty+of+Versailles&random=true

Google

**WolframAlpha™ computational knowledge engine**

Treaty of Versailles

Examples Random

Assuming "Treaty of Versailles" is a historical event | Use as a word instead

Input interpretation: Treaty of Versailles

Basic information:

date	28 June 1919
city involved	Versailles, Ile-de-France, France
countries involved	French Third Republic   Italy   Japan   United Kingdom of Great Britain and Ireland   United States   German Empire
people involved	David Lloyd George   Georges Clemenceau   Woodrow Wilson

Timeline: Treaty of Versailles

Include today

1910 1915 1920 1925 1930

Download as: PDF | Live Mathematica

New to Wolfram|Alpha?  
TAKE THE TOUR »

Serving up funky, fresh fun facts on the daily



Follow the fun:  
**@WolframFunFacts**

# Wolfram Alpha

DUBLIN - Bing

www.bing.com/search?q=dublin&go=&qs=n&form=QBLH&filt=all&pq=dublin&sc=0-6&sp=-1&sk=

WEB IMAGES VIDEOS SHOPPING NEWS MORE

bing Beta dublin

Sign in ▾

47.500.000 RESULTS Narrow by language ▾ Narrow by region ▾

**Dublin - Wikipedia, la enciclopedia libre** [Translate this page](#)  
[es.wikipedia.org/wiki/Dublín](http://es.wikipedia.org/wiki/Dublín) ▾  
Etimología · Historia · Geografía · Cultura · Deporte · Gobierno  
Dublin (en irlandés: Baile Átha Cliath, AFI: ...) industrial y cultural de la República de Irlanda. Dublin ha sido declarada ciudad global por el GaWC ...

**Dublín - Guía de viajes y turismo en Dublín...** [Translate this page](#)  
[www.dublin.es](http://www.dublin.es) ▾  
Guía de Dublin con información turística y práctica para viajar a Dublín. Todo lo necesario para visitar y disfrutar la Capital de Irlanda.

**Images of dublin**  
[bing.com/images](http://bing.com/images)



**Voy a Dublin - Guía de Dublin** [Translate this page](#)  
[www.voyadublin.com](http://www.voyadublin.com) ▾  
Guía de Dublin. Información sobre Dublin para turistas y quien vaya a estudiar o trabajar a Irlanda con rutas turísticas, excursiones, albergues o consejos ...

**Dublín, Irlanda - qué ver y qué hacer...** [Translate this page](#)  
[www.discoverireland.com/.../areas-and-cities/dublin-city](http://www.discoverireland.com/.../areas-and-cities/dublin-city) ▾  
Fantásticas tiendas, divertidos pubs y un montón de lugares históricos hacen de Dublin un excelente destino para una escapada. Ah!, y según TripAdvisor ...

**Qué ver en Dublín - Monumentos y visitas...** [Translate this page](#)  
[www.visitdublin.com](http://www.visitdublin.com)

Display a menu

# Bing

DUBLIN - Google Search

www.google.com/search?ie=UTF-8&q=Colin+Farrell&fb=1&hl=en&sa=N&tab=lw#hl=en&sclient=psy-ab&q=dublin&oq=dublin&gs\_l=serp.3..0i46

+You Search Images Maps Play YouTube News Gmail Drive Calendar More

Google dublin

SIGN IN

Web Images Maps Shopping News More Search tools

About 223,000,000 results (0.21 seconds)

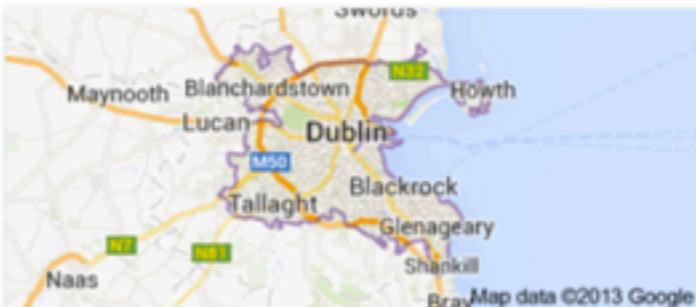
**Dublin - Wikipedia, the free encyclopedia**  
[en.wikipedia.org/wiki/Dublin](http://en.wikipedia.org/wiki/Dublin) ▾  
Dublin is the capital and most populous city of Ireland. The English name for the city is derived from the Irish name Dubhlinn, meaning "black pool". Dublin is ...  
History of Dublin - County Dublin - Dublin Airport - Greater Dublin Area

**Visit Dublin - Official Tourist Information Dublin Hotels and Car Hire**  
[www.visitdublin.com/](http://www.visitdublin.com/) ▾  
The official Dublin tourist board website offering information on tours maps travel accommodation and on-line booking for Dublin Ireland.  
See & Do - Dublin's Top 10s - Events - Visitor Attractions

**Dublin.ie - Official portal website for the city of Dublin, Ireland**  
[www.dublin.ie/](http://www.dublin.ie/) ▾  
Dublin.ie is the official city portal for Dublin, Ireland. We offer information on Accommodation, Arts & Culture, Business, Childcare, Entertainment, Environment ...

**Things To Do In Dublin Including Dublin Attractions, Restaurants ...**  
[www.timeout.com/dublin/](http://www.timeout.com/dublin/) ▾  
Discover what's on and things to do in Dublin. Plus, Book your Flights & Hotels - Time Out Dublin.

**Dublin Airport - Welcome to Dublin Airport**  
[www.dublinairport.com/](http://www.dublinairport.com/) ▾  
Real-time flight arrival and departure information, features about Dublin, Cork and Shannon Airports, Duty Free shopping outlets, airport maps, and future ...

  
Map data ©2013 Google

**Dublin**

Capital of Ireland

Dublin is the capital and most populous city of Ireland. The English name for the city is derived from the Irish name Dubhlinn, meaning "black pool". Wikipedia

Population: 525,383 (2011)  
Area: 44.4 sq miles (115 km<sup>2</sup>)  
Weather: 22°C, Wind NE at 10 km/h, 69% Humidity  
Local time: Saturday 2:30 PM

Points of interest



# Google

Yahoo! Search - Web Search

search.yahoo.com

Check out the new [Yahoo.com](#). Access Search, Mail and a virtually endless stream of content customized just for you. [Try it now!](#)

Web Images Video Local Shopping News More ▾

**YAHOO!**

dublin

**DUBLIN, IRELAND**  
01:31 PM (Europe/Dublin). - Current local time

**TOP RATED THINGS TO DO**

- 1. Dublin Mountains (The)
- 2. Ron Blacks
- 3. Judge Roy Beans

**DUBLIN OVERVIEW**

Hotels  
Restaurant Guide  
Flights

Glasgow  
UNITED KINGDOM  
Dublin  
IRELAND  
Liverpool  
London  
YAHOO! TRAVEL

Display a menu

# Yahoo!

dublin - Yahoo! Search Results

Home Mail News Sports Finance Weather Games Groups Answers Flickr More

YAHOO! dublin Search Edgar Mail

Web Images Video Shopping Blogs More

Anytime Past day Past week Past month

**Dublin, Ireland**  
travel.yahoo.com  
Sat Jul 20 2:32 pm IST Foggy, 62°F ☀  
More than a quarter of the Republic of Ireland's population of almost four million lives within the Greater Dublin area. Intensely proud of their city, Dubliners seem to possess an innate sense of its heritage and powerful literary culture, and can at times exhibit a certain snobbishness towards those ... [more](#)

**Dublin - Wikipedia, the free encyclopedia**  
en.wikipedia.org/wiki/Dublin Cached  
History | Government | Geography | Places of interest  
Dublin is the capital and most populous city of Ireland. The English name for the city is derived from the Irish name Dubhlann, meaning "black pool". Dublin is ...

**Visit Dublin - Official Tourist Information Dublin Hotels and...**  
www.visitdublin.com Cached  
The official Dublin tourist board website offering information on tours maps travel accommodation and on-line booking for Dublin Ireland.

**Dublin travel guide - Wikitravel**  
wikitravel.org/en/Dublin Cached  
Dublin is the capital city of Ireland. Its vibrancy, nightlife and tourist attractions are noteworthy, and it is the most popular entry point for international ...

Related Points Of Interest

Dublin Castle Trinity College... Guinness Storeh... O'Connell Stree... Glasnevin Cemet... Bull Island Irish Museum of... St Anne's Park Trinity Capital... National Botani...

Ads

**Dublin Hotels**  
www.ORBITZ.com/Dublin

Display a menu

# Yahoo!

- [Mail](#)
- [News](#)
- [Finance](#)
- [Sports](#)
- [Movies](#)
- [omgl!](#)
- [Shine](#)
- [Autos](#)
- [Shopping](#)
- [Travel](#)
- [Dating](#)
- [Jobs](#)

[More Y! Sites >](#)

Make **YAHOO!**  
your homepage

ADVERTISEMENT



The Hottest Gray  
Hair Trend 2013  
eSalon.com



## Writer under fire for slamming cheerleader's weight

A blogger says an Oklahoma City dancer has no business wearing a tiny outfit in front of an NBA crowd. [She politely fires back »](#)

1 - 5 of 55


[All Stories](#) [News](#) [Entertainment](#) [Sports](#) [Business](#) [More ▾](#)


## Court may limit use of race in college admission decisions

By Joan Biskupic WASHINGTON (Reuters) - Thirty-five years after the Supreme Court set the terms for boosting college admissions of African Americans and other minorities, the court may be about to issue a ruling that could restrict universities' [Reuters](#) 53 mins ago Education Society



## In a first, black voter turnout rate passes whites

WASHINGTON (AP) — America's blacks voted at a higher rate than other minority groups in 2012 and by most measures surpassed the white turnout for the first time, reflecting a deeply polarized presidential election in which blacks strongly [Associated Press](#)

## Dad Anticipates Tough Talks With His Teenage Daughters

DEAR ABBY: As a father of two teenage daughters, I have a question about couples living together. Do relationships that start this way have a higher failure rate than those that don't? What should be [Dear Abby](#)

## Trending Now

- |  |   |
|--|---|
| <a href="#">1 Eastwood age 105</a>         | <a href="#">6 Swift \$17 million man...</a> |
| <a href="#">2 10 band members die i...</a> | <a href="#">7 Tulsa 2024 Olympics</a>       |
| <a href="#">3 Michael Jordan marries</a>   | <a href="#">8 Rodney Allen Rippy</a>        |
| <a href="#">4 Cheerleader body found</a>   | <a href="#">9 N. Korea charges U.S....</a>  |
| <a href="#">5 NASCAR pit fight</a>         | <a href="#">10 FBI Boston boat</a>          |

[Watch the show »](#)

## YAHOO! AUTOS

Up-to-the-minute automotive news, reviews, and research.

[Take a look](#)

[Ad Feedback](#)
[AdChoices](#)

## London

52°F Fair



Today  
52° 41°



Tomorrow  
59° 37°



Tuesday  
56° 38°

[Quotes](#)
[Test H500](#)

- Mail
- News
- Finance
- Sports
- Movies
- omg!
- Shine
- Autos
- Shopping
- Travel
- Dating
- Jobs
- [More Y! Sites >](#)

Make **YAHOO!**  
your homepage

## ADVERTISEMENT



The Hottest Gray  
Hair Trend 2013  
eSalon.com



## Writer under fire for slamming cheerleader's weight

A blogger says an Oklahoma City dancer has no business wearing a tiny outfit in front of an NBA crowd. [She politely fires back »](#)

1 – 5 of 55



[All Stories](#) [News](#) [Entertainment](#) [Sports](#) [Business](#) [More ▾](#)

Show me fewer stories about:

Story removed [Undo](#)

[Education](#) [Society](#) [Anthony Kennedy](#) [Abigail Fisher](#) [University](#) [Lewis F. Powell, Jr.](#)

[Edit content preferences](#)



## In a first, black voter turnout rate passes whites

WASHINGTON (AP) — America's blacks voted at a higher rate than other minority groups in 2012 and by most measures surpassed the white turnout for the first time, reflecting a deeply polarized presidential election in which blacks strongly Associated Press

## Dad Anticipates Tough Talks With His Teenage Daughters

DEAR ABBY: As a father of two teenage daughters, I have a question about couples living together. Do relationships that start this way have a higher failure rate than those that don't? What should be Dear Abby

## Trending Now

- |  |   |
|--|---|
| <a href="#">1 Eastwood age 105</a>         | <a href="#">6 Swift \$17 million man...</a> |
| <a href="#">2 10 band members die i...</a> | <a href="#">7 Tulsa 2024 Olympics</a>       |
| <a href="#">3 Michael Jordan marries</a>   | <a href="#">8 Rodney Allen Rippy</a>        |
| <a href="#">4 Cheerleader body found</a>   | <a href="#">9 N. Korea charges U.S....</a>  |
| <a href="#">5 NASCAR pit fight</a>         | <a href="#">10 FBI Boston boat</a>          |

Watch the show »

## YAHOO! AUTOS

Up-to-the-minute automotive news, reviews, and research.

[Take a look](#)



[Ad Feedback](#)

[AdChoices ▶](#)

## London

52°F Fair



Today  
52° 41°

Tomorrow  
59° 37°

Tuesday  
56° 38°

[Quicken Loans](#) [Test H500](#)



## Writer under fire for slamming cheerleader's weight

A blogger says an Oklahoma City dancer has no business wearing a tiny outfit in front of an NBA crowd. [She politely fires back »](#)

1 – 5 of 55



Blogger calls out cheerleader



Paltrow's dress defended



Paris Jackson with her mom



Progressive Insurance lady



Michael Jordan marries



All Stories News Entertainment Sports Business More ▾

Show me fewer stories about:

Story removed Undo

[Education](#)

[Society](#)

[Anthony Kennedy](#)

[Abigail Fisher](#)

[University](#)

[Lewis F. Powell, Jr.](#)

[Edit content preferences](#)

## Trending Now

- 1 [Eastwood age 105](#)
- 2 [10 band members die i...](#)
- 3 [Michael Jordan marries](#)
- 4 [Cheerleader body found](#)
- 5 [NASCAR pit fight](#)

**YAHOO**

Up-to-the-minute reviews, an

Take a



[Ad Feedback](#)

London

# **Goals of part I**

- Learn entity linking basics
- Get familiar with
  - terminology and essentials
  - seminal papers/methods
  - evaluation and datasets
- Obtain experience with
  - (publicly available) toolkits
  - evaluation

# Why do we need entity linking?

- (Automatic) document enrichment
  - go-read-here
  - assistance for (Wikipedia) editors
  - inline (microformats, RDFa)
- “Use as feature”
  - to improve
    - classification
    - retrieval
    - word sense disambiguation
    - semantic similarity
    - ...
  - dimensionality reduction (e.g., term vectors)

# Why do we need entity linking?

- Enable
  - entity retrieval / semantic search
  - advanced UI/UX
  - ontology learning, KB population
  - ...

# A bit of history

- Text classification
- NER
- WSD
- NED/NEN
  - {person name, geo, movie name, ...} disambiguation
  - (Cross-document) coreference resolution
  - Automatic link generation
- Entity linking

# Entity linking?

- NE normalization / canonicalization / sense disambiguation
- DB record linkage / schema mapping
  - (not the focus here, but see **[Demartini et al. 2013]**)
- Knowledge base population
- Entity linking
  - D2W
  - Wikification
  - Semantic linking

# Entity Linking: main problem

- Linking free text to entities
  - Any piece of text
    - news documents
    - blog posts
    - tweets
    - queries
    - ...
  - Entities (typically) taken from a knowledge base
    - Wikipedia
    - Freebase
    - ...

# Typical steps

1. Determine “linkable” phrases
  - mention detection – **MD**
2. Rank>Select candidate entity links
  - link generation – **LG**
  - may include NILs (null values, i.e., no target in KB)
3. (Use “context” to disambiguate/filter/improve)
  - disambiguation – **DA**

# **Methods**

# Preliminaries

- Knowledge bases...
- Wikipedia-based measures
  - commonness
  - relatedness
  - keyphraseness

# Wikipedia

- Basic element: article (proper)
  - But also
    - redirect pages
    - disambiguation pages
    - category/template pages
    - admin pages
  - Hyperlinks
    - use “unique identifiers” (URLs)
      - [[United States]] or [[United States|American]]
      - [[United States (TV series)]] or  
[[United States (TV series)|TV show]]



# Wikipedia style guidelines

- “the lead contains a quick summary of the topic's most important points, and each major subtopic is detailed in its own section of the article”
  - “The lead section (also known as the lead, introduction or intro) of a Wikipedia article is the section before the table of contents and the first heading. The lead serves as an introduction to the article and a summary of its most important aspects.”

# Disambiguation pages

- Senses of a phrase
- Short description
- (Possible) categorization
- Non-exhaustive

The screenshot shows a web browser displaying the Wikipedia disambiguation page for "United States". The title bar reads "United States (disambiguation)". The page features the Wikipedia logo and navigation links like "Main page", "Contents", and "Featured content". The main content area is titled "United States (disambiguation)" and includes sections for "Countries", "Current", "Historical", "Proposed", and "Fictional", each listing various entities or concepts. A sidebar on the left provides language links for other Wikipedia editions.

United States (disambiguation)

From Wikipedia, the free encyclopedia

The term **United States** currently usually refers to the [United States of America](#), a sovereign state in North America.

**Countries** [edit]

- The United Mexican States, the official name of [Mexico](#)

**Current** [edit]

- United States of Belgium, a confederation that existed during the year 1790
- Republic of the United States of Brazil ([Portuguese: República dos Estados Unidos do Brasil](#))
- United States of Central America (informal name), more properly known as the United Provinces of Central America, a 19th century federation of the nations of Central America
- United States of Colombia, name held by Colombia between 1863 and 1886
- United States of Indonesia, name of the country from 1949 to 1950
- United States of the Ionian Islands, former British protectorate from 1815 to 1864
- United States of Stellaland, a short-lived political union of Goshen and Stellaland provinces
- United States of Venezuela from 1864 to April 15, 1953
- Dutch Republic or United Provinces

**Historical** [edit]

- United States of Latin Africa, a political entity proposed by Barthélemy Boganda for Central Africa
- United States of China, a political concept of a federalized China modeled after the United States of America
- United States of Europe, a political concept of a single European state
- United States of Africa, a political concept/proposal, similar to the United States of Europe
- United States of South America, a proposed strong federation in South America or Latin America
- United States of Greater Austria was a successor state to the Austro-Hungarian Empire
- The United States of the West created by the admission of the European countries as states
- United States of Australia, a name suggested before the creation of the Commonwealth of Australia

**Proposed** [edit]

- United States of Southern Africa (USSA), successor state to South Africa in Africa
- United States of Japan in the TV series [Code Geass](#)

**Fictional** [edit]

- The United States of Southern Africa (USSA), successor state to South Africa in Africa
- United States of Japan in the TV series [Code Geass](#)

# Some statistics

- WordNet
  - 80k entity definitions
  - 115k surface forms
  - 142k senses (entity - surface form combinations)
- Wikipedia (only)
  - ~4M entity definitions
  - ~12M surface forms
  - ~24M senses

# **Wikipedia-based measures**

# Wikipedia-based measures

- keyphraseness( $w$ ) [Mihalcea & Csomai 2007]

$$\frac{\text{CF}(w_l)}{\text{CF}(w)} \longrightarrow \begin{array}{l} \textbf{Collection frequency} \\ \text{term } w \text{ as a link to another} \\ \text{Wikipedia article} \end{array}$$

↓

$$\begin{array}{l} \textbf{Collection frequency} \\ \text{term } w \end{array}$$

# Wikipedia-based measures

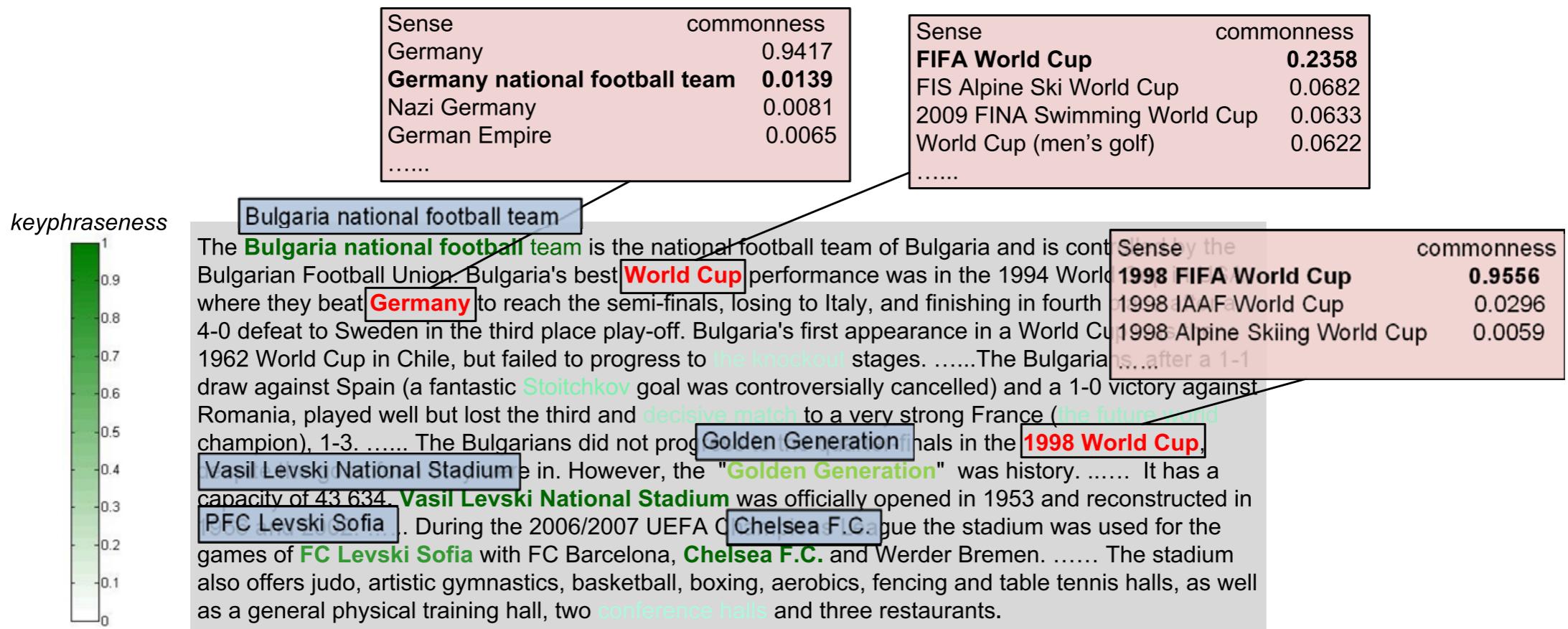
- commonness( $w, c$ ) [Medelyan et al. 2008]

$$\frac{|L_{w,c}|}{\sum_{c'} |L_{w,c'}|}$$



**Number of links**  
with target  $c'$  and anchor text  $w$

# Commonness and keyphraseness



# Wikipedia-based measures

- relatedness( $c, c'$ ) [Milne & Witten 2008a]

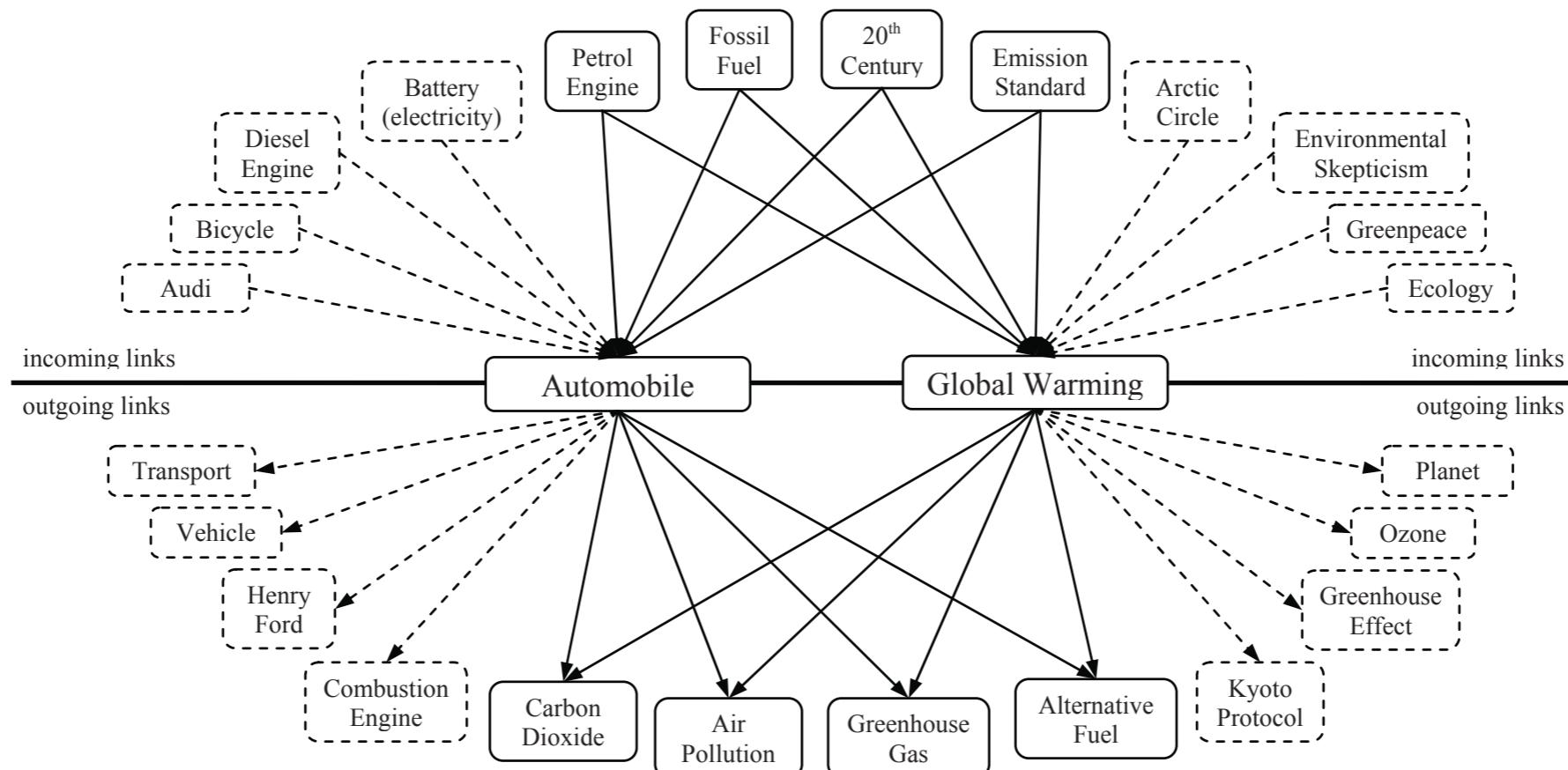


Image taken from Milne and Witten (2008a). **An Effective, Low-Cost Measure of Semantic Relatedness Obtained from Wikipedia Links.** In AAAI WikiAI Workshop.

# Wikipedia-based measures

- relatedness( $c, c'$ ) [Milne & Witten 2008a]

$$\frac{\log(\max(|L_c|, |L_{c'}|)) - \log(|L_c \cap L_{c'}|)}{\log(|WP|) - \log(\min(|L_c|, |L_{c'}|))}$$

Number of links  
with target  $c$

Intersection of inlinks  
with target  $c$  and  $c'$

Total number of  
Wikipedia articles

The diagram illustrates the formula for relatedness. It shows three main components: the top part involves the number of links to target  $c$ ; the bottom part involves the intersection of inlinks for both  $c$  and  $c'$ ; and the middle part is the total number of Wikipedia articles. Arrows indicate the flow from the formula terms to their corresponding descriptive labels.

# **Baseline methods**

# **Recall the steps**

- 1. mention detection – MD**
- 2. link generation – LG**
- 3. (disambiguation) – DA**

# Large-Scale Named Entity Disambiguation Based on Wikipedia Data

[Cucerzan 2007]

- Key intuition: leverage context links
  - **"Texas"** is a [[pop music]] band from [[Glasgow]], [[Scotland]], [[United Kingdom]]. They were founded by [[Johnny McElhone]] in [[1986 in music|1986]] and had their performing debut in [[March]] [[1988]] at ...
- Prune the candidates, keep only:
  - appearances in the first paragraph of an article, and
  - reciprocal links

# **Large-Scale Named Entity Disambiguation Based on Wikipedia Data**

**[Cucerzan 2007]**

- MD
  - NER; rule-based; co-ref resolution
  - not the focus...
- LG
  - Represent entities as vectors
    - context, categories
  - Same for all candidate entity links
  - Determine maximally coherent set

# Wikify!

[Mihalcea & Csomai 2007]

- First paper on actual entity linking, i.e., moving beyond disambiguation
- Identifies two steps
  1. identify important concepts in the text  
“keyword extraction” (mention detection - **MD**)
  2. link these to corresponding Wikipedia pages  
“word sense disambiguation” (link generation - **LG**)

# Wikify!

[Mihalcea & Csomai 2007]

- MD
  - tf.idf,  $\chi^2$ , keyphraseness
- LG
  1. Overlap between definition (Wikipedia page) and context (paragraph) [Lesk 1986]
  2. Naive Bayes [Mihalcea 2007]
    - context, POS, entity-specific terms
  3. Voting between (1) and (2)

# Topic Indexing with Wikipedia

[Medelyan et al. 2008]

- MD
  - keyphraseness [Mihalcea & Csomai 2007]
- LG
  - combination of average relatedness & commonness
- LG/DA
  - Naive Bayes
    - TF.IDF, position, length, degree, weighted keyphraseness

# Learning to Link with Wikipedia

[Milne & Witten 2008b]

- Key idea: disambiguation informs detection
  - start with unambiguous senses
  - compare each possible sense with its *relatedness* to the context sense candidates
  - So, first LG, then base MD on these results

# Learning to Link with Wikipedia

## [Milne & Witten 2008b]

### Depth-first search

From Wikipedia, the free encyclopedia



**Depth-first search (DFS)** is an algorithm for traversing or searching a tree structure or graph. One starts at the root (selecting some node as the root in the graph case) and explores as far as possible along each branch before backtracking.

Formally, DFS is an uninformed search that progresses by expanding the first child node of the search tree that appears and thus going deeper and deeper until a goal node is found, or until it hits a node that has no children. Then the search backtracks, returning to the most recent node it hadn't finished exploring. In a non-recursive implementation, all freshly expanded nodes are added to a LIFO stack for exploration.

sense	commonness
Tree	92.82%
Tree (graph theory)	2.94%
<b>Tree (data structure)</b>	<b>2.57%</b>
Tree (set theory)	0.15%
Phylogenetic tree	0.07%
Christmas tree	0.07%
Binary tree	0.04%
Family tree	0.04%
...	

# **Learning to Link with Wikipedia**

**[Milne & Witten 2008b]**

- MD
  - ...
- LG
  - Machine learning
    - keyphraseness, average relatedness, sum of average weights

# Learning to Link with Wikipedia

[Milne & Witten 2008b]

- MD
  - Machine learning
    - link probability, relatedness, **confidence of LG**, generality, frequency, location, spread
- LG
  - Machine learning
    - keyphraseness, average relatedness, sum of average weights

# Learning to Link with Wikipedia

[Milne & Witten 2008b]

- Some heuristics
  - filter non-informative, non-ambiguous candidates (e.g., “the”)
    - based on keyphraseness, i.e., link probability
  - filter non-central candidates
    - based on average relatedness to all other context senses

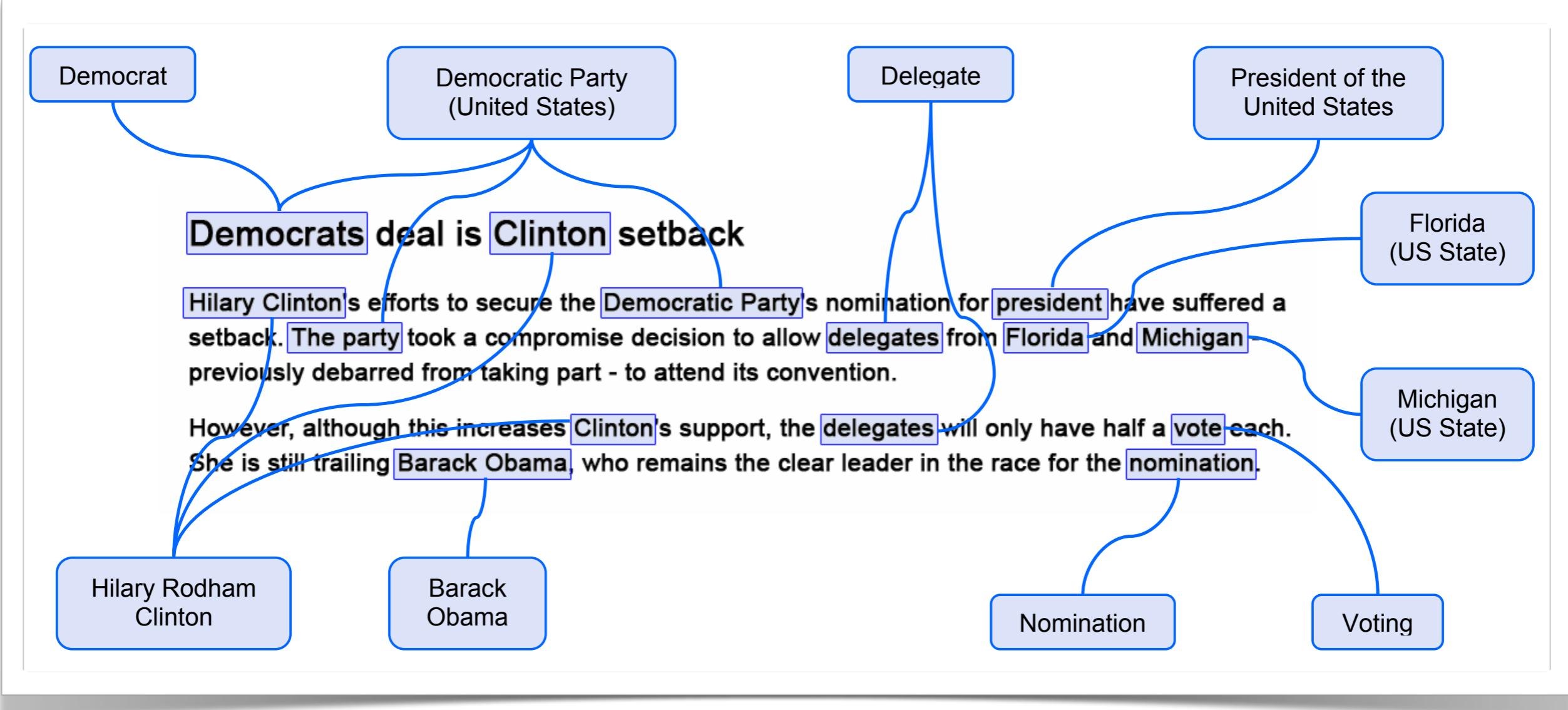


Image taken from Milne and Witten (2008b). Learning to Link with Wikipedia. In CIKM '08.

# **Local versus global context**

# Local versus global context

- “Global”
  - i.e., disambiguation of the candidate entity graph
  - NP-hard
- Optimization
  - reduce the search space to a “disambiguation context”
    - all plausible (reciprocal) disambiguations **[Cucerzan 2007]**
    - unambiguous surface forms, pair-wise comparisons, and/or averages **[Milne & Witten 2008b]**
    - hill-climbing, integer linear programs **[Kulkarni et al. 2009]**
    - hybrid + ML **[Ratinov et al. 2011, Ferragina & Scaiella 2010]**

# **Collective annotation of Wikipedia entities in web text**

**[Kulkarni et al. 2009]**

- Contribution
  - determine a collective score based on trade-off between local compatibility and global topical coherence between candidate entities
  - use ILP or Hill-climbing (ILP beats HC, but is slower)
- Also
  - new test collection (web pages), including NILs

# **Local and Global Algorithms for Disambiguation to Wikipedia**

**[Ratinov et al. 2011]**

- Main contribution, in steps – MD + DA
  1. use “local” approach (e.g., commonness) to generate a disambiguation context
  2. apply “global” machine learning approach on pairs
    - relatedness, PMI
      - {inlinks, outlinks} in various combinations (c and c')
      - {avg, max}
- (Apply another round of machine learning) – LG

# **TAGME: On-the-fly Annotation of Short Text Fragments**

**[Ferragina & Scaiella 2010]**

- MD
  - keyphraseness **[Mihalcea & Csomai 2007]**
- LG
  - use “local” approach to generate a disambiguation context, very similar to **[Ratinov et al. 2011]**
  - Heavy pruning
    - mentions; candidate links; coherence
- Accessible at <http://tagme.di.unipi.it>

# Adding semantics to microblog posts

[Meij et al. 2012]

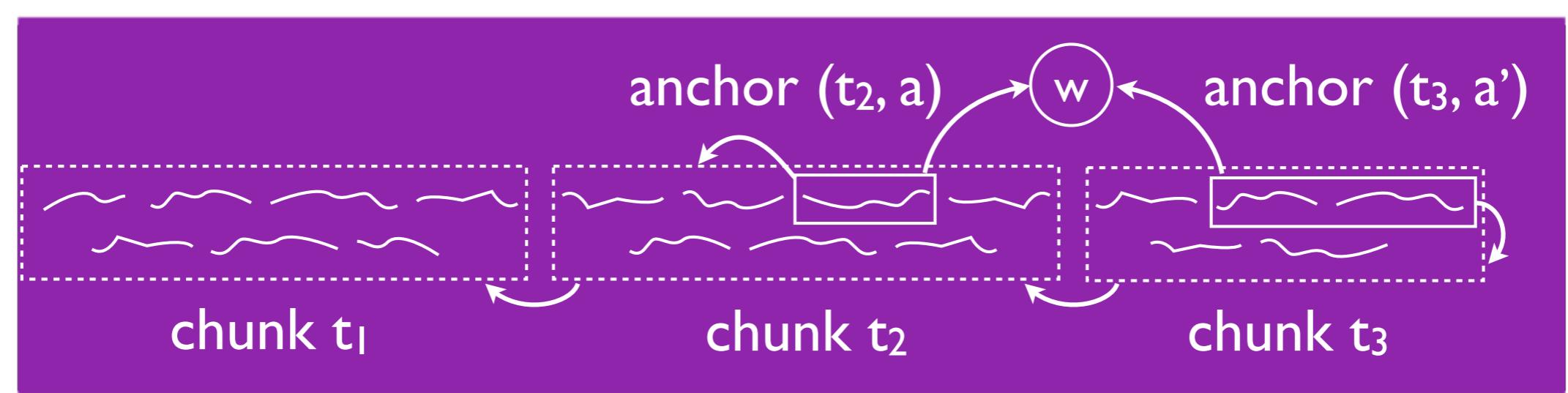
- MD
  - commonness (and others)
  - idea: obtain ranked list of **all** candidate entity links
- LG
  - use point-wise LeToR to determine which links to keep
    - ..., random forests, GBRT
    - {text, entity, text+entity, context} features
    - (**[Guo et al. 2013]** use a similar approach)
- Relatively “clean” test collection, see **[Derczynski et al. 2012]** and **[Cassidy et al. 2012]**

# **Graph-based methods**

# Feeding the Second Screen: Semantic Linking based on Subtitles

[Odijk et al. 2013]

- Setting: entity linking on closed captions
  - streaming, high-precision, real-time
- Graph information as additional features
  - Idea: maintain a (coherent) tripartite context graph
    - entities
    - chunks
    - anchors



# Feeding the Second Screen: Semantic Linking based on Subtitles

[Odijk et al. 2013]

## *Context features*

$DEGREE(w, G)$	Number of edges connected to the node representing Wikipedia article $w$ in context graph $G$ .
$DEGREE - CENTRALITY(w, G)$	Centrality of Wikipedia article $w$ in context graph $G$ , computed as the ratio of edges connected to the node representing $w$ in $G$ .
$PAGERANK(w, G)$	Importance of the node representing $w$ in context graph $G$ , measured using PageRank.

# **A Graph-based Method for Entity Linking**

[Guo et al. 2011]

- MD
  - rule-based; prefer longer links
  - generate a disambiguation context
- LG
  - (weighted interpolation of) in- and outdegree in disambiguation context to select entity links
    - edges defined by wikilinks
- Evaluation on TAC KBP

# **Graph-based named entity linking with Wikipedia**

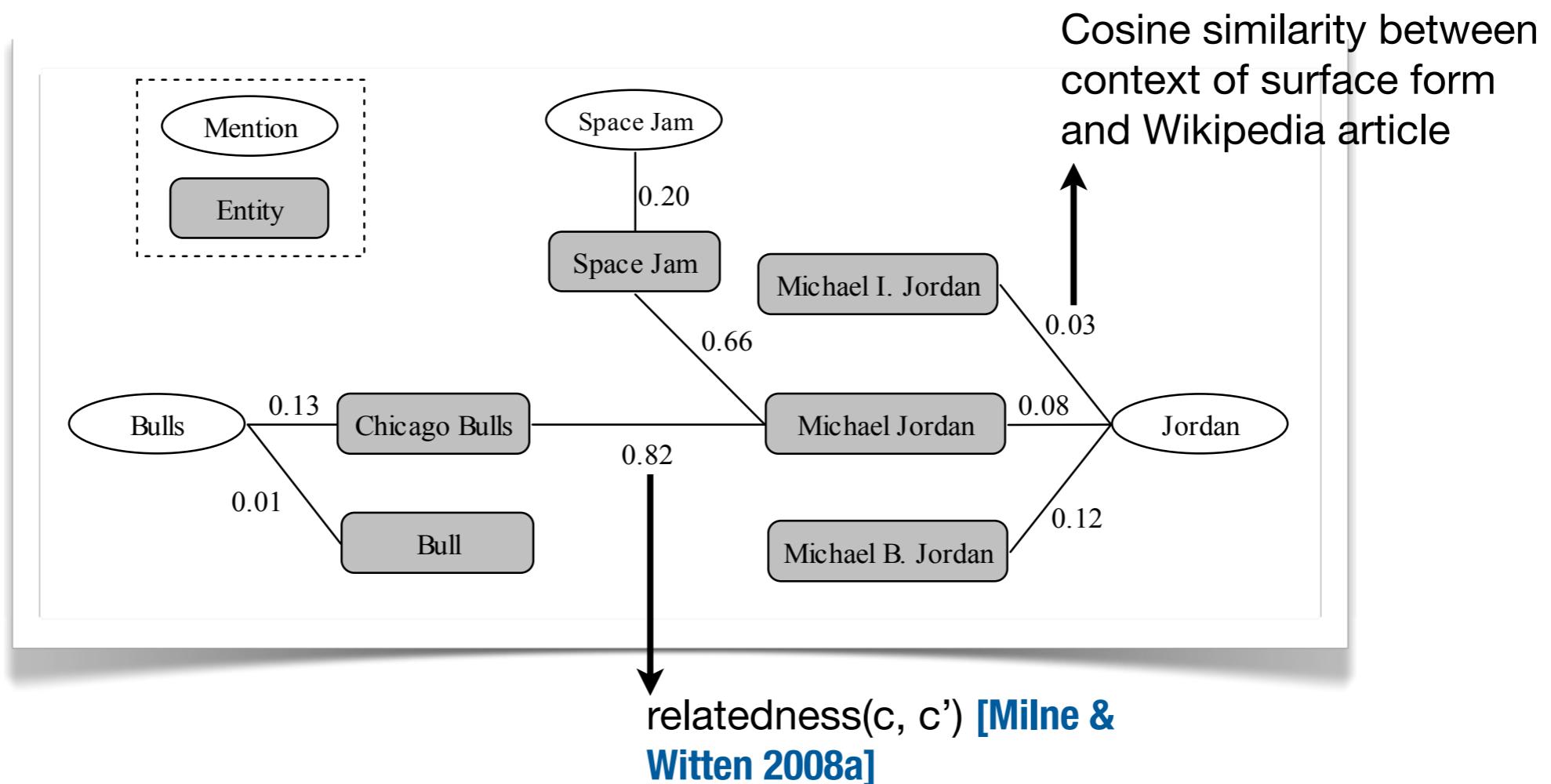
**[Hachey et al. 2011]**

- MD
  - generate disambiguation context
    - based on unambiguous entity links
  - edges defined by wikilinks (articles & categories)
    - max step size: 2 (articles), 3 (categories)
- LG
  - use degree centrality and PageRank to reweigh cosine-based similarity scores
- Evaluation on TAC KBP

# Collective Entity Linking in Web Text: A Graph-Based Method

[Han et al. 2011]

- Global approach, main contribution
  - random walk on the *referent graph*, defined by the intra-Wikipedia links



# Collective Entity Linking in Web Text: A Graph-Based Method

[Han et al. 2011]

- Evaluation
  - IITB test collection [Kulkarni et al. 2009]
  - LC = using only cosine similarity

	Precision	Recall	F1
<i>Wikify!</i>	0.55	0.28	0.37
<i>Cucerzan</i>	0.71	0.33	0.45
<i>M&amp;W</i>	0.80	0.38	0.52
<i>CSAW</i>	0.65	0.73	0.69
<i>Our Method(LC)</i>	0.52	0.34	0.41
<i>Our Method</i>	0.69	0.76	0.73

Table 3. The overall results on IITB data set

# **Topic modeling methods**

# From Names to Entities using Thematic Context Distance

[Pilz et al. 2011]

- Main contribution
  - “extend” previous BOW approaches for disambiguation with LDA topics
  - compare topic distributions of source document with candidate entities

**Table 1: Topics for entities with name *John Taylor* (excerpt) with associated probability value**

disambiguation term	$i$	$p(t_i)$	Important words (titles) of the topics
South Carolina governor	109	0.3805	unit state, state senat, lieuten governor, hous repres, elect governor, ...
	120	0.2477	north carolina, south carolina, unit state, west virginia, civil war, ...
athlete	80	0.4190	summer olymp, gold medal, world record, silver medal, world championship, ...
	135	0.1047	unit state, rhode island, baltimore maryland, new hampshire, georg washington, ...
racing driver	129	0.7407	grand prix, race driver, motor race, formula, race team, sport car, ...
jazz	141	0.5781	jazz musician, big band, new york, duke ellington, jazz band, ...
bass guitarist	18	0.2964	rock band, solo album, play guitar, band member, rock roll, ...
	70	0.1594	album releas, studio album, debut album, record label, music video, ...

# Collective Entity Linking in Web Text: A Graph-Based Method

[Han et al. 2011]

- Evaluation
  - IITB test collection [Kulkarni et al. 2009]
  - LC = using only cosine similarity

	Precision	Recall	F1
<i>Wikify!</i>	0.55	0.28	0.37
<i>Cucerzan</i>	0.71	0.33	0.45
<i>M&amp;W</i>	0.80	0.38	0.52
<i>CSAW</i>	0.65	0.73	0.69
<b><i>Our Method(LC)</i></b>	<b>0.52</b>	<b>0.34</b>	<b>0.41</b>
<b><i>Our Method</i></b>	<b>0.69</b>	<b>0.76</b>	<b>0.73</b>

Table 3. The overall results on IITB data set

# An Entity-topic Model for Entity Linking

[Han & Sun 2012]

- Main contribution
  - “extend” previous BOW approaches for disambiguation with LDA topics
  - add/correlate with global disambiguation

# An Entity-topic Model for Entity Linking

[Han & Sun 2012]

- Main contribution
  - “extend” previous BOW approaches for disambiguation with LDA topics
  - add/correlate with global disambiguation

	Precision	Recall	F1
<i>Wikify!</i>	0.55	0.28	0.37
<i>EM-Model</i>	0.82	0.48	0.61
<i>M&amp;W</i>	0.80	0.38	0.52
<i>CSAW</i>	0.65	0.73	0.69
<i>EL-Graph</i>	0.69	0.76	0.73
<b><i>Our Method</i></b>	<b>0.81</b>	<b>0.80</b>	<b>0.80</b>

Table 1. The overall results on IITB data set

# **Entity Disambiguation with Hierarchical Topic Models**

**[Kataria et al. 2011]**

- Semi-supervised hierarchical topic modeling
  - one topic per entity: need heavy pruning
  - restrict possible topics for words in a WP article based on occurrences of surface forms
- Semi-supervised
  - bias topic-word distributions based on commonness
- Hierarchical
  - capture word and topic correlations: use Wikipedia categories for co-occurrence (with pruning)

# Recap

- Essential ingredients
  - MD
    - commonness
    - keyphraseness
  - LG
    - commonness
    - machine learning
  - DA
    - relatedness
    - machine learning
    - topic modeling
    - graph-based methods

# Outline

- Part 1 – Entity Linking
  - introduction
  - methods
  - evaluation
  - test collections
  - toolkits
  - open challenges

# **Evaluation**

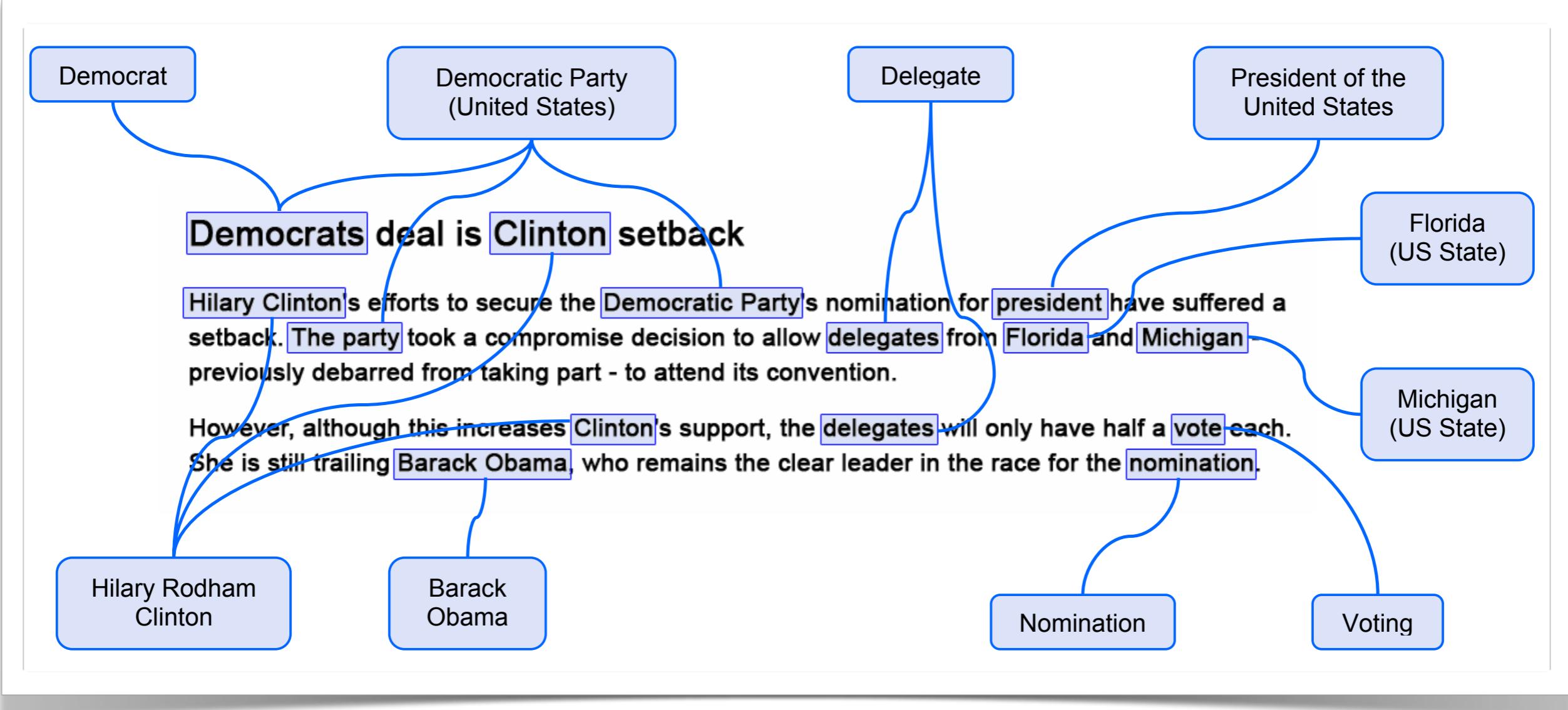


Image taken from Milne and Witten (2008b). Learning to Link with Wikipedia. In CIKM '08.

# **DIY Entity Linking**

- Ingredients
  - knowledge bases
  - evaluation metrics
  - test collection

# **Knowledge bases**

# **DBpedia**

- Extract structured information from Wikipedia
  - infoboxes, categories, and more
  - crowd-sourced community effort
- Open source
  - written in Scala, Java and VSP
  - Virtuoso Universal Server Operating system
- See <http://dbpedia.org/About>

# Freebase

- Initially seeded from high-quality open data
  - now composed mainly by community
  - harvested from many sources
    - Wikipedia, MusicBrainz, and others.
- Acquired by Google in 2010 (GKG)
- See <http://www.freebase.com/>

# Wikipedia vs Freebase

- Freebase 5x larger than Wikipedia  
(in terms of the number of entities)
- Geared towards entertainment
- For 85% of Freebase entities there's no text...
  - but, there are Wikipedia-Freebase links  
(for some entities)
  - initial work trying to ameliorate this problem **[Zheng et al. 2012]**

# **YAGO**

- Accuracy manually evaluated
  - confirmed accuracy of 95%
  - relation is annotated with its confidence value.
- Anchored in Time and Space
  - Thematic domains (e.g. "music" or "science")
- Includes WordNet
- See <http://www.mpi-inf.mpg.de/yago-naga/yago/>

# Sense of Scale

- YAGO: 10 million entities and 120 million facts
- Freebase: 37 million topics, 1,998 types, and more than 30,000 properties
- DBpedia: 3.77 million things
  - 2.35 million classified in Ontology, including:
    - 764,000 persons, 573,000 places,
    - 333,000 creative works, 192,000 organizations,
    - 202,000 species and 5,500 diseases.
  - 111 languages, together 20.8 million things

# **DIY Entity Linking**

- Ingredients
  - knowledge base
  - evaluation metrics
  - test collection

# Evaluation metrics

# Evaluation metrics

- What is the task?

# Evaluation metrics

- Set-based (similar to WSD)
  - “How many correct links were retrieved?”
    - precision, recall, F-measure
- Rank-based
  - “Was the correct link(s) retrieved with a high score?”
- macro/micro
  - per anchor phrase
  - per tweet, query, sentence, document

# Common set-based metrics

- Accuracy

$$A = \frac{|\{\mathcal{C}_{i,0} | \mathcal{C}_{i,0} = \mathcal{G}\}|}{N}$$

- Precision

$$P_{\mathcal{C}} = \frac{|\{\mathcal{C}_i | \mathcal{C}_i \neq \emptyset \wedge \mathcal{G}_i \in \mathcal{C}_i\}|}{|\{\mathcal{C}_i | \mathcal{C}_i \neq \emptyset\}|}$$

- Recall

$$R_{\mathcal{C}} = \frac{|\{\mathcal{C}_i | \mathcal{G}_i \neq \text{NIL} \wedge \mathcal{G}_i \in \mathcal{C}_i\}|}{|\{\mathcal{G}_i | \mathcal{G}_i \neq \text{NIL}\}|}$$

$N$	Number of queries in data set
$\mathcal{G}$	Gold standard annotations for data set ( $ \mathcal{G}  = N$ )
$\mathcal{G}_i$	Gold standard for query $i$ (KB ID or NIL)
$\mathcal{C}$	Candidate sets from system output ( $ \mathcal{C}  = N$ )
$\mathcal{C}_i$	Candidate set for query $i$
$\mathcal{C}_{i,j}$	Candidate at rank $j$ for query $i$ (where $\mathcal{C}_i \neq \emptyset$ )

# Common rank-based metrics

- Recall @ k
- Precision @ k
- R-precision
- Mean average precision
- Mean reciprocal rank
- Precision @ 1

# **DIY Entity Linking**

- Ingredients
  - knowledge base
  - evaluation metrics
  - test collection

# **Test collections**

# Entity linking test collections

- Wikipedia
- MSNBC
- AQUAINT
- ACE
- Twitter
- AIDA (CoNLL)
- IITB (web data)
- INEX link-the-wiki
- TREC knowledge base acceleration (KBA)
- TAC knowledge base population (KBP)

# **Wikipedia (for evaluation)**

- Widely used
- Pros
  - cheap and easy; the links are already provided
- Cons
  - biased (style guides!)
  - specific scenario
  - unbalanced

# **MSNBC**

**[Cucerzan 2007]**

- 20 news articles
  - linked to EN Wikipedia from 2006
  - 756 total links; 127 of these are NIL
- Focus: disambiguate entities after NER and co-reference resolution
  - all mentions of all the detected entities are linked
- Con
  - collected by correcting the output of Cucerzan's system

# AQUAINT

[Milne & Witten 2008]

- 50 news articles
  - 449 links, obtained using Amazon mechanical turk
- Subset of AQUAINT newswire, annotated to mimic Wikipedia hyperlink structure
  - only first mentions of “important” titles were linked
  - uninteresting and redundant mentions of the same title not linked

# **ACE**

**[Ratinov et al. 2011]**

- Subset of ACE co-reference data set
  - mentions and their types are given
  - co-references resolved
- First nominal mentions of each co-reference chain are linked
  - Amazon mechanical turk
  - accuracy of majority vote ~85%
  - manually corrected

# Twitter

[Meij et al. 2012]

- Tweets taken from “verified accounts,” so relatively clean
- ~500 tweets, manually linked to Wikipedia
  - ~2 entity links per tweet on average

Task	Name	Year	Source	All Mentions	Instances
CDCR	John Smith	1998	News	✗	197
CDCR	WePS 1	2007	Web	✗	3,489
CDCR	Day et al.	2008	News	✓	3,660
CDCR	WePS 2	2008	Web	✗	3,432
CDCR	WePS 3	2009	Web	✗	31,950
wikify	Mihalcea	2007	Wiki	✓	7,286
wikify	Kulkarni	2009	Web	✓	17,200
wikify	Milne	2010	Wiki	✓	11,000
NEL	Cucerzan	2007	News	✓	797
NEL	TAC 09	2009	News	✗	3,904
NEL	Fader	2009	News	✗	500
NEL	TAC 10	2010	News, Blogs	✗	3,750
NEL	Dredze	2010	News	✗	1,496
NEL	Bentivogli	2010	News, Web, Transcripts	✓	16,851
NEL	Hoffart	2011	News	✓	34,956

Table taken from Hachey et al. (2013). **Evaluating Entity Linking with Wikipedia**. In AI '13.

# TAC

[McNamee et al. 2010]

- Target: KB – from Wikipedia (~800k instances)
  - infoboxes; article text; type
- “Query”
  - document ID (news, web, blog)
  - mention string (occurring at least once in that doc)
- Focus on ambiguous mentions
  - collected by cherry-picking ‘interesting’ mentions, rather than systematically annotating all mentions
- Explicit NILs (> 50% of the queries)

	TAC 2009 test		TAC 2010 train		TAC 2010 test	
$ Q $	3,904		1,500		2,250	
KB	1,675	(43%)	1,074	(72%)	1,020	(45%)
NIL	2,229	(57%)	426	(28%)	1,230	(55%)
PER	627	(16%)	500	(33%)	751	(33%)
ORG	2710	(69%)	500	(33%)	750	(33%)
GPE	567	(15%)	500	(33%)	749	(33%)
News	3904	(100%)	783	(52%)	1500	(67%)
Web	0	(0%)	717	(48%)	750	(33%)
Acronym	827	(21%)	173	(12%)	347	(15%)

$ \mathcal{E} $	560		—		871	
KB	182	(33%)	462	(—)	402	(46%)
NIL	378	(67%)	—	(—)	469	(54%)
PER	136	(24%)	—	(—)	334	(38%)
ORG	364	(65%)	—	(—)	332	(38%)
GPE	60	(11%)	—	(—)	205	(24%)

Table taken from Hachey et al. (2013). **Evaluating Entity Linking with Wikipedia**. In AI '13.

# Evaluation - recap

- Even with so many test collections to choose from, there's still quite some variation
- People create their own “extracts” from WP
- Same method, same test collection, but different results in different papers
  - tokenization, normalization, ...
- We need meta-evaluations...

# Meta-evaluations

- [Hachey et al. 2013]
- [Cornolti et al. 2013]

# Evaluating Entity Linking with Wikipedia

[Hachey et al. 2013]

- Named entity linking, a.k.a., “NEL”
  - include NILs
  - Wikipedia articles not always named entities
- Explicit focus on separating “search” (LG) and “disambiguation” (DA)
- Reimplement and evaluate three NEL systems
  - [Bunescu & Pasă 2006]
  - [Cucerzan 2007]
  - [Varna et al. 2009] (TAC system paper)

System	Extractor	Condition	Searcher						Disambiguator	
			Title	Redirect	Link	Truncated	Bold	DABTitle		
Bunescu and Pașca (2006)	NER	NA	✓	✓				✓	NA	SVM rank over cosine and mention context word×category features
Cucerzan (2007)	NER, coreference expansion	NA	✓	✓	✗	✓		✓	NA	Scalar product between candidate category/term vector and document-level vector
Varma et al. (2009)	NER, acronym expansion	if acronym								Cosine between candidate article term vector and mention context vector
		if expandable	✓							
		else	✓	✓			✓	✓	NA	
		else								
		search 1	✓							
		if no candidates	✓	✓			✓	✓	NA	

Table taken from Hachey et al. (2013). **Evaluating Entity Linking with Wikipedia**. In AI '13.

Alias	Source	$\langle C \rangle$	$P_{\mathcal{C}}^{\infty}$	$R_{\mathcal{C}}^{\infty}$	$P_{\emptyset}$	$R_{\emptyset}$
Title		0.2	<b>83.5</b>	37.2	68.1	96.5
Redirect		0.1	74.6	20.0	62.1	96.2
Link		4.2	55.7	<b>80.1</b>	<b>88.6</b>	59.5
Bold		1.6	45.1	48.8	71.7	67.2
Hatnote		0.0	42.6	1.2	57.7	<b>99.9</b>
Truncated		1.2	37.8	24.5	62.2	78.6
DABTitle		3.5	34.2	29.3	58.7	65.1
DABRedirect		2.7	34.0	18.9	57.9	77.3

**LG**

Alias Source	$\langle C \rangle$	$P_{\mathcal{C}}^{\infty}$	$R_{\mathcal{C}}^{\infty}$	$P_{\emptyset}$	$R_{\emptyset}$
Title	0.2	83.5	37.2	68.1	96.5
+Redirect	0.3	79.4	54.6	75.0	92.6
+Link	2.4	56.2	76.5	87.6	63.8
+Bold	2.4	55.8	77.1	88.2	62.9
+Hatnote	2.4	55.8	77.1	88.2	62.9
+Truncated	2.4	55.8	77.1	88.2	62.9
+DABTitle	2.4	55.8	77.1	88.2	62.9
+DABRedirect	2.4	55.4	77.1	88.1	62.2

**LG**

System	Number of NIL queries in the test collection		
	$A$	$A_C$	$A_\emptyset$
NIL Baseline	57.1	0.0	100.0
Title Baseline	71.0	37.2	96.5
+ Redirect Baseline	76.3	54.6	92.6
Bunescu and Paşa	77.0	67.8	83.8
Cucerzan	78.3	71.3	83.5
Varma et al. Replicated	80.1	72.3	86.0
TAC 09 Median	71.1	63.5	78.9
TAC 09 Max (Varma)	82.2	76.5	86.4

**DA**

# Meta-evaluations

- [Hachey et al. 2013]
- [Cornolti et al. 2013]

# A Framework for Benchmarking Entity-Annotation Systems

[Cornolti et al. 2013]

- Compare five publicly available entity linkers
  - [Hoffart et al. 2007] (AIDA)
  - [Ratinov et al. 2011]
  - [Ferragina & Scaiella 2010] (TAGME)
  - [Milne & Witten 2008] (wikipedia-miner)
  - DBpedia Spotlight
- And also investigate parameter/cut-off settings

# A Framework for Benchmarking Entity-Annotation Systems

[Cornolti et al. 2013]

- On five publicly available test collections
  - AIDA **[Hoffart et al. 2007]**
    - based on CoNLL 2003: noun annotations
    - 1393 Reuters newswire articles
    - hand-annotated all nouns with entities in YAGO2
  - AQUAINT **[Milne & Witten 2008]**
  - MSNBC **[Cucerzan 2007]**
  - IITB **[Kulkarni et al. 2010]** (web data)
  - Twitter **[Meij et al. 2012]**

# A Framework for Benchmarking Entity-Annotation Systems

[Cornolti et al. 2013]

- Benchmarking framework
- Introduces “fuzzy” evaluation measures
- Main findings
  - different systems perform well in different scenarios
  - AIDA and TagMe seem to be the winners overall

# **DIY Entity Linking**

- Ingredients
  - knowledge bases
  - evaluation metrics
  - test collection

# DIY Entity Linking – footnotes

- ClueWeb annotated with Freebase (FACC1)
  - TREC Web topics too
- Dictionaries for Linking Text, Entities and Ideas:

concept: “soccer”			
football <i>and</i>			
Football			
Soccer <i>and</i>			
soccer			
Association football	サッカー	piłkarz	
fútbol <i>and</i>	축구	voetbalclub	
Fútbol	footballeur	ฟุตบอล	
footballer	Fußballspieler	bóng đá	
Futbol <i>and</i>	sepak bola	voetbal	
futbol	足球	Foutbaal	
Fußball	فوتبال	futebolista	
futebol	футболист	لعبة كرة القدم	
futbolista	כדורגל	fotbal	

# Outline

- Part 1 – Entity Linking
  - introduction
  - methods
  - evaluation
  - test collections
  - toolkits
  - open challenges

# Toolkits



# **Public Toolkits and Web Services for Entity Linking**

- Wikipedia Miner
- TagMe
- DBpedia Spotlight
- Illinios Wikifier
- AIDA
- (OpenCalais)

# Wikipedia Miner

[Milne & Witten 2008b]

- Open source
- (Public) web service
  - Java
  - Hadoop preprocessing pipeline
- Lexical matching + machine learning
- See <http://wikipedia-miner.cms.waikato.ac.nz>

# TagMe

[Ferragina & Scaiella 2010]

- Web service only (demo + API)
- Approach similar to Wikipedia Miner
- Voting for disambiguation
  - based on all possible bindings
  - heuristics to select best target
- Designed for short texts
- See <http://tagme.di.unipi.it/>

# Illinois Wikifier

[Ratinov et al. 2011]

- Local install + online demo
  - uses Illinois NER system
- Disambiguation as weighted sum of features
  - Textual similarity
  - Global coherence based on link structure
- See [http://cogcomp.cs.illinois.edu/page/  
software view/33](http://cogcomp.cs.illinois.edu/page/software_view/33)

# DBpedia Spotlight

[Mendes et al., 2011]

- Open source
- Public web service
- Disambiguation in local context
  - vector-space model using bag-of-words and cosine similarity
  - (actually, Lucene)
- See <http://spotlight.dbpedia.org>

# AIDA

[Yosef et al. 2011]

- Open source
  - uses Stanford NER system
- (Public) web service, API
- Links to YAGO2
- Disambiguation in 3 variants
  - PriorOnly: link to most common target
  - Local: disambiguate individual links with local features
  - CocktailParty: collective disambiguation maximizing coherence using iterative graph-based approach

# **OpenCalais**

- Only on public content
  - does not keep a copy of content
  - keeps a copy of the metadata it extracts
- Free for up to 50,000 documents per day
- Early adopters:
  - CBS Interactive / CNET, Huffington Post, Al Jazeera, The White House
  - more than 30,000 developers & 50 publishers

	Programming Language	Service	Available Languages	Open Source
Wikipedia Miner	Java	Web API, Application	any WP	✓
TagMe	Java	Web API	EN, IT	✗
DBpedia Spotlight	Java	Web API, Application	EN + any WP	✓
Illinois Wikifier	Java	Application	EN	✓
AIDA	Java	Web API	EN	✓
OpenCalais	?	Web API	EN, FR, SP	✗

	Matching	Target KB	Context	Comment
Wikipedia Miner	Lexical	Wikipedia	ML on Relatedness	
TagMe	Lexical	Wikipedia	Vote on Relatedness	Focus on Short texts
DBpedia Spotlight	Lexical?	DBpedia	Cosine Similarity	Structure
Illinois Wikifier	NER	Wikipedia	Global Coherence	
AIDA	NER	YAGO2	Multiple	Structure
OpenCalais	?	Calais	?	

# Outline

- Part 1 – Entity Linking
  - introduction
  - methods
  - evaluation
  - test collections
  - toolkits
  - open challenges

# **Open challenges**

# Open challenges

- Difficulty prediction
  - similar to ambiguity, but not the same
  - dependent on context, candidate links, ...
- Multi/Cross-lingual entity linking
  - **[Wang et al. 2013]**
  - CrossLink-2 (NTCIR-9), CJK - EN **[Tang et al. 2013]**
  - TAC...
- Cross-KB entity linking (“Freebase”)
  - directly? use Wikipedia as pivot?

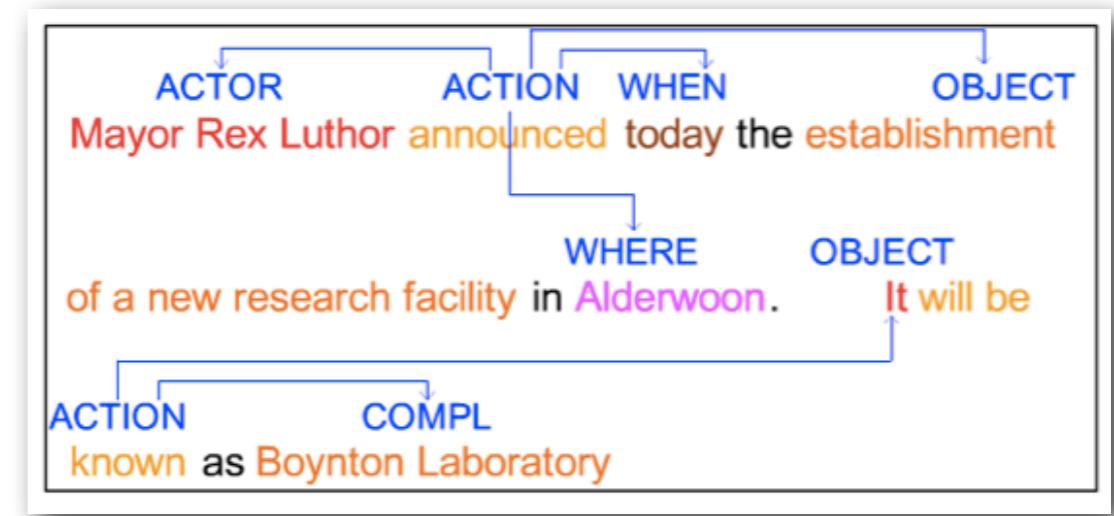
# Learning/Updating the KB

- Link parallel, continuous streams of items
  - news, tweets, blogs, status updates
  - queries, clicks
  - web pages, RDFa/schema.org
  - etc.
- Given an entity
  - “What is new?” What do I need to know now?”
  - Add: personal
  - Add: social
- TREC KBA/KBP/KBx, TREC TS



# Learning/Updating the KB: ingredients?

- Accurate entity linking
  - real-time
  - cross-item
  - cross-genre
  - cross-vertical
- What is being said?
  - aspects, attributes, relations, events
- Correlate with already known facts
- Detect bursts, events



# Open challenges

- Generic test collections
  - What's the task? User model? Evaluation?
    - TAC? set-based? ranking? known-item finding? top- $k$ ?
    - exhaustive linking? first mention only?
    - "aboutness"
- Moving beyond entities
  - events/news, concepts, relations
- Moving beyond "ad hoc" entity linking:
  - incorporate contextual evidence in the task (and evaluation)
  - {users, history, profile, social, trending, ...}

# Follow-up reading

- Detecting **unlinkable** entities [[Lin et al. 2012a](#)]
- Linking entities to **any database** [[Sil et al. 2012](#)]
- Hyperlinking for **multimedia** data [[Eskevich et al. 2013](#)]
- Automatically generating **Wikipedia articles**  
[\[Sauper & Barzilay 2009\]](#)
- **Scaling up** to the web [[Lin et al. 2012b](#)]
- Serendipitous **suggestions** based on **personalized** entity links [[Bordino et al. 2013](#)]
- **Actionable** entities/queries [[Lin et al. 2012](#)]

# References – Entity linking

The screenshot shows a Mendeley group page titled "Entity Linking and Retrieval – Tutorial at WWW 2013 and SIGIR 2013". The page displays three research papers:

- Analysis and Enhancement of Wikification for Microblogs with Context Expansion.** By Taylor Cassidy, Heng Ji, Lev-Arie Ratinov, Arkatz Zubaga, Hongzhao Huang. Published in COLING 2012 (2012).

Disambiguation to Wikipedia (D2W) is the task of linking mentions of concepts in text to their corresponding Wikipedia entries. Most previous work has focused on linking terms in formal texts (e.g. newswire) to Wikipedia. Linking terms in short...
- Microblog-genre noise and impact on semantic annotation accuracy** By Leon Derczynski, Diana Maynard, Niraj Aswani, Kalina Bontcheva. Published in HT 2013 (2013).

Using semantic technologies for mining and intelligent information access to microblogs is a challenging, emerging research area. Unlike carefully authored news text and other longer content, tweets pose a number of new challenges, due to their...
- Entity Disambiguation with Freebase** By Zicheng Zheng, Xiancse Si, Fangtao Li, Edward Y. Chang, Xiaoyan Zhu. Published in WIAT 2013 (2013).

Freebase is a large-scale knowledge base that contains millions of entities and their relationships. Entity disambiguation is the task of identifying the correct entity from a set of candidates based on context. This is a challenging problem because many entities can have similar names or descriptions. In this paper, we propose a novel approach for entity disambiguation using Freebase. Our approach consists of two main steps: (1) extracting features from the input sentence; (2) using a support vector machine (SVM) to classify the entity based on the extracted features. We evaluate our approach on several datasets and show that it outperforms state-of-the-art methods.

The right side of the page shows "Top tags in this group" including entity linking, Wikipedia, TAC, commonness, SVM, graph, relatedness, naive bayes, pagerank, keyphraseness, Twitter, centrality, meta evaluation, NER, word sense disambiguation, random forests, Freebase, tagme, local, and web.

<http://www.mendeley.com/groups/3339761/entity-linking-and-retrieval-tutorial-at-www-2013-and-sigir-2013/papers/added/0/tag/entity+linking/>