Part I

Entity Linking

Outline

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

Introduction

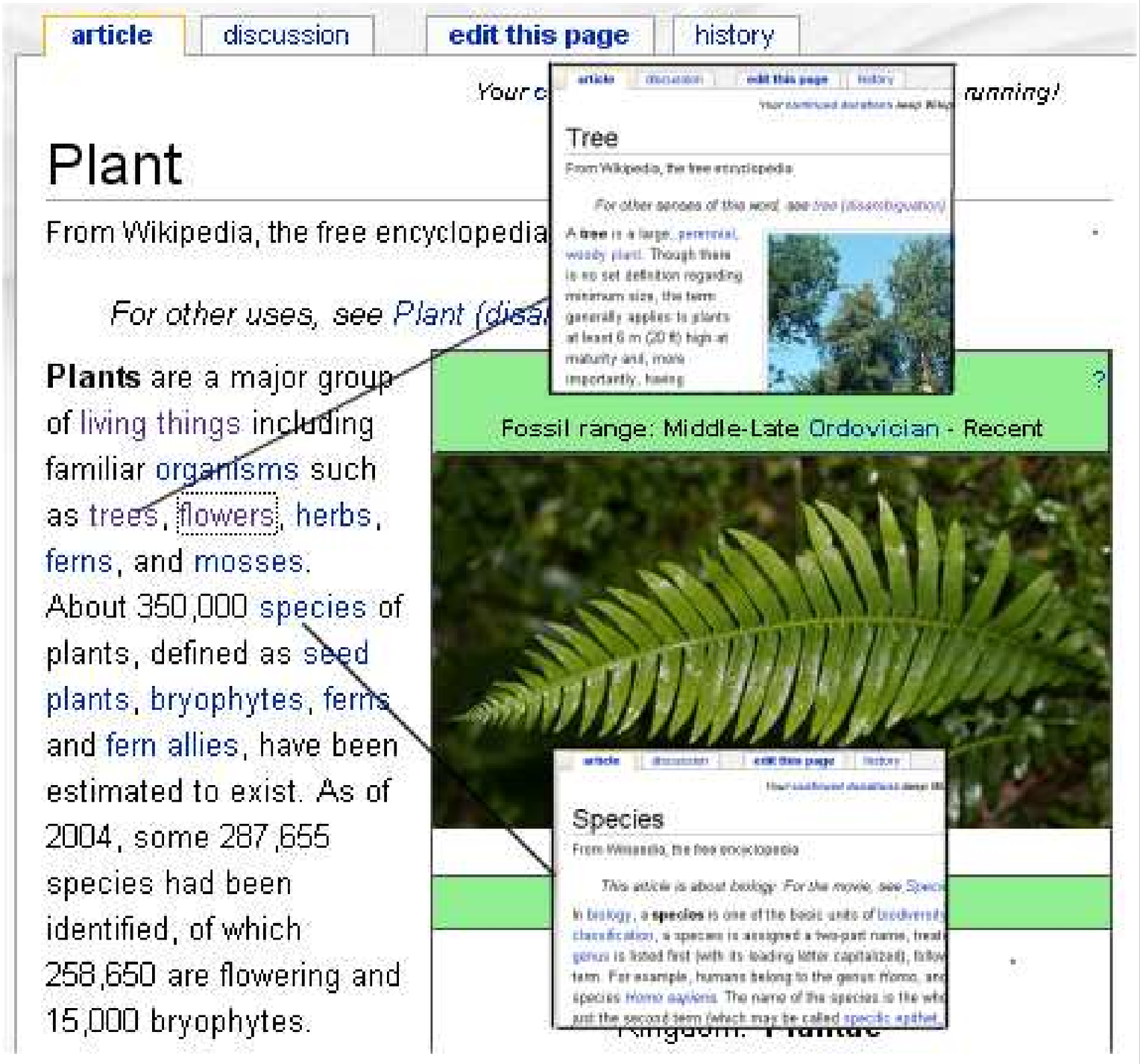


Image taken from Mihalcea and Csomai (2007). **Wikify!: linking documents to encyclopedic knowledge**. In CIKM '07.

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

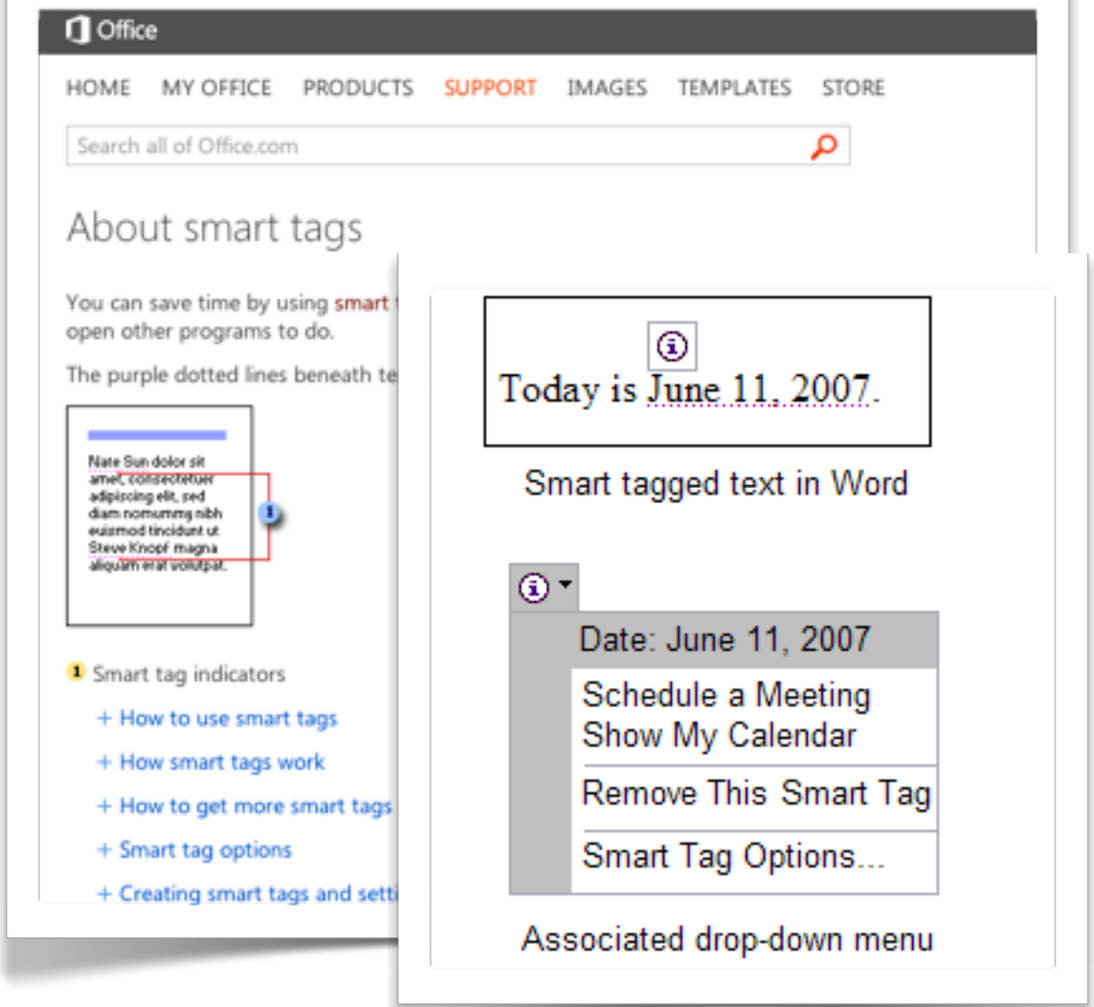
ABSTRACT | EDITOR'S SUMMARY

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.

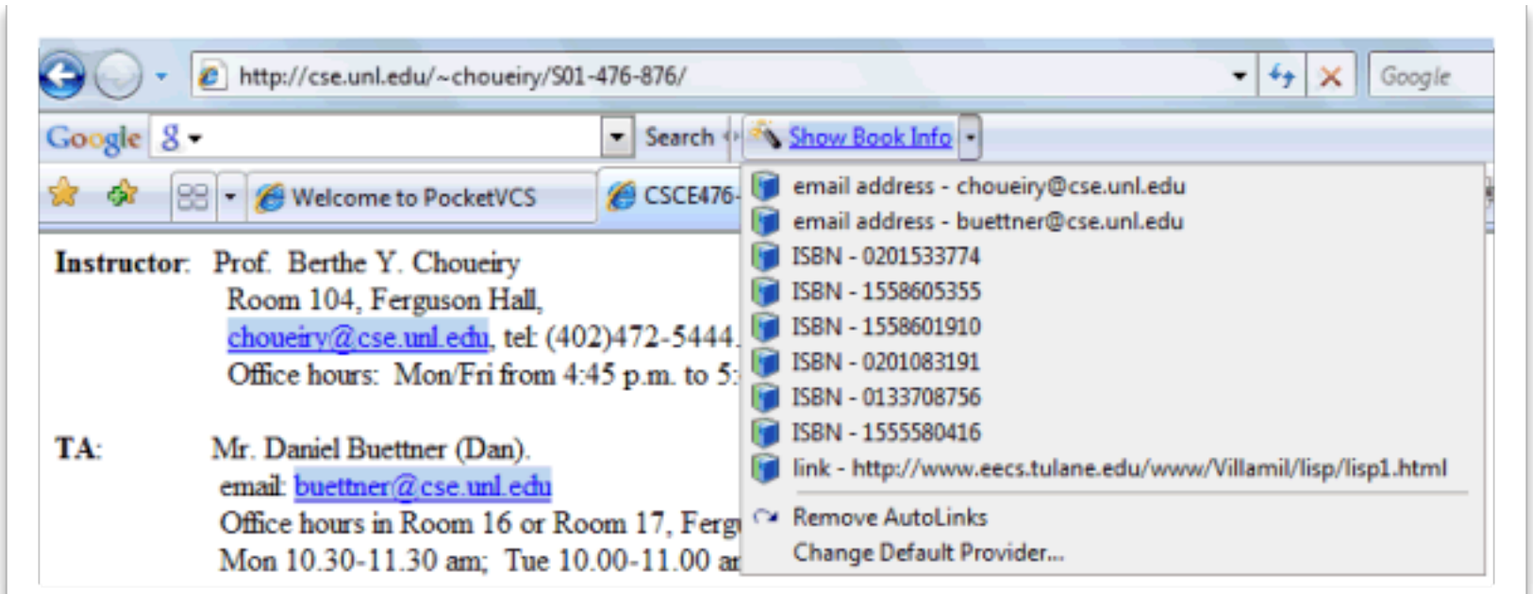
Received for publication 7 January 2013.

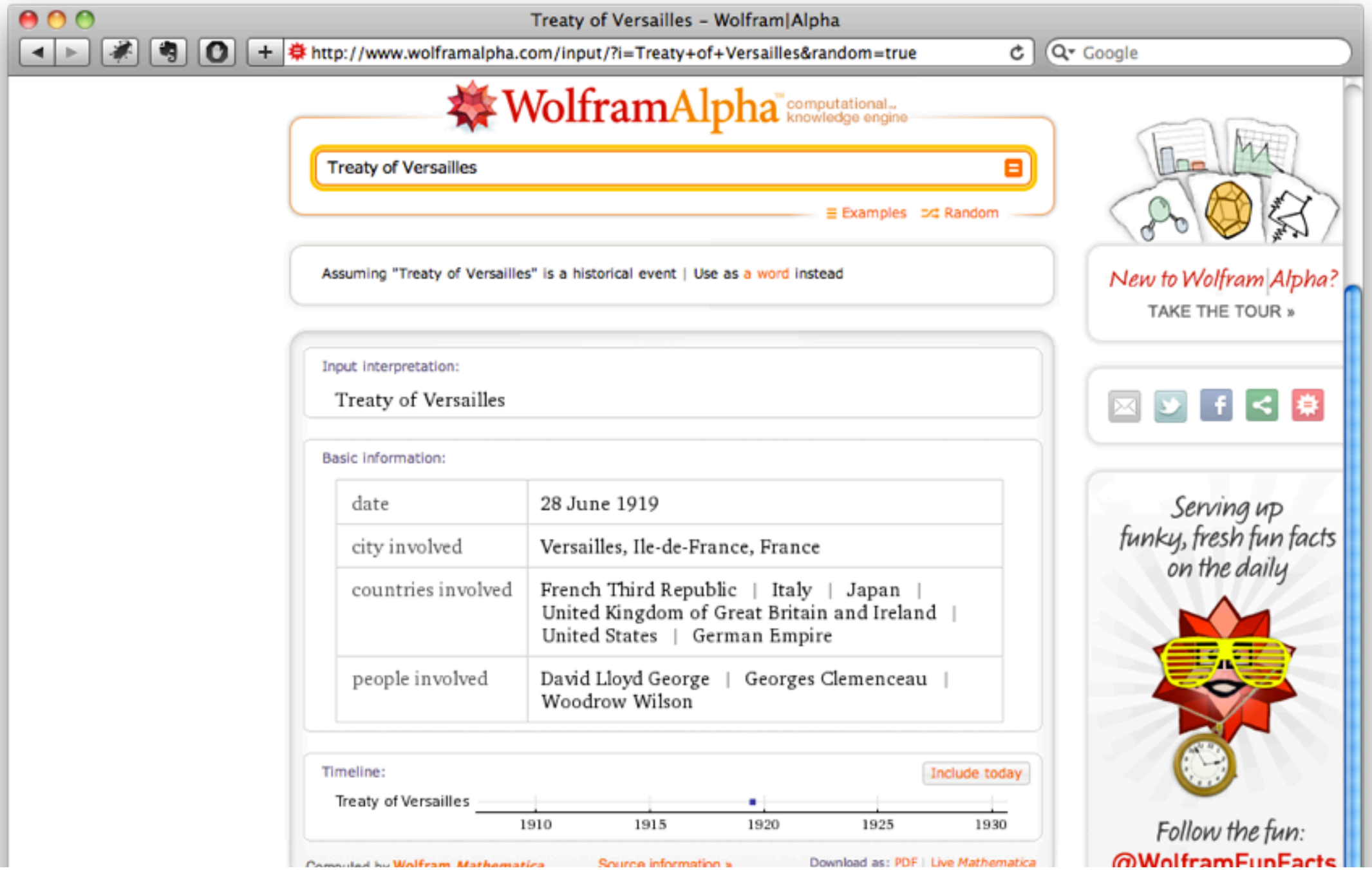
See http://tagme.di.unipi.it

Microsoft Smart Tags

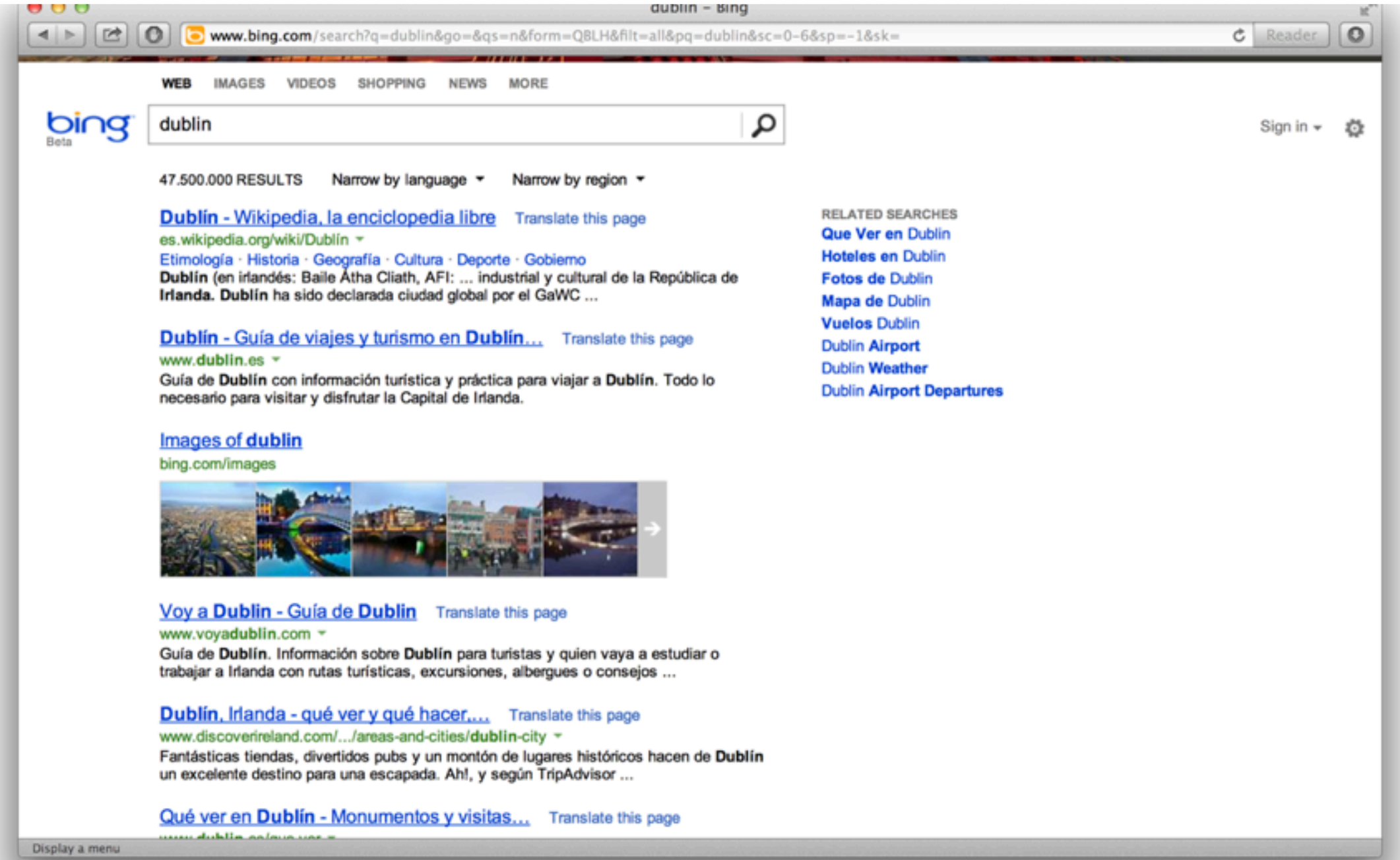


Google toolbar

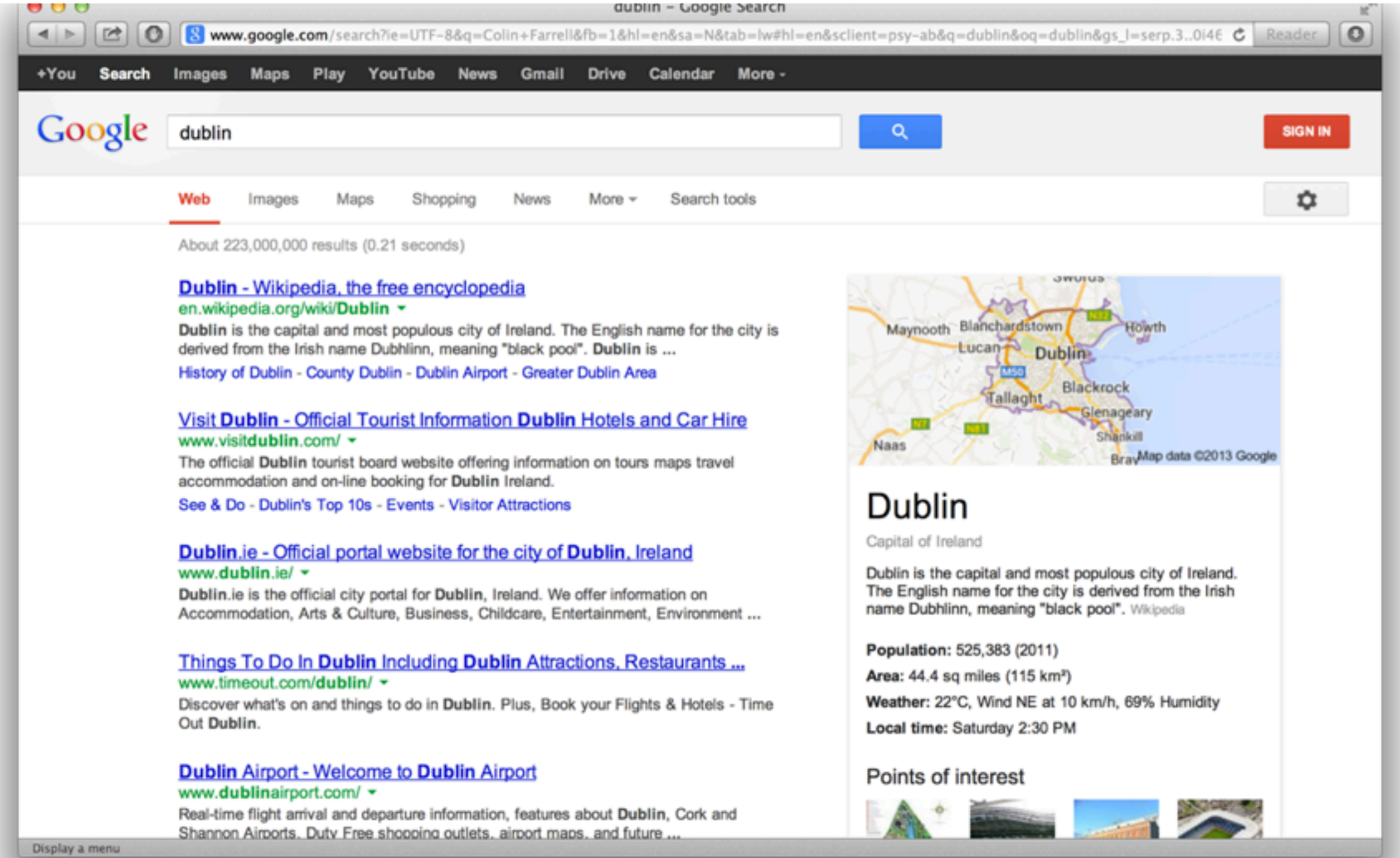




Wolfram Alpha


dublin - Bing
www.bing.com/search?q=dublin&go=&qs=n&form=QBLH&filt=all&pq=dublin&sc=0-6&sp=-1&sk=
Reader
WEB IMAGES VIDEOS SHOPPING NEWS MORE
bing
Beta
dublin
Sign in
47.500.000 RESULTS
Narrow by language
Narrow by region
Dublín - Wikipedia, la enciclopedia libre
Translate this page
es.wikipedia.org/wiki/Dublín
Etimología · Historia · Geografía · Cultura · Deporte · Gobierno
Dublín (en irlandés: Baile Átha Cliath, AFI: ... industrial y cultural de la República de Irlanda. Dublín ha sido declarada ciudad global por el GaWC ...
Dublín - Guía de viajes y turismo en Dublín...
Translate this page
www.dublin.es
Guía de Dublín 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
RELATED SEARCHES
Que Ver en Dublin
Hoteles en Dublin
Fotos de Dublin
Mapa de Dublin
Vuelos Dublin
Dublin Airport
Dublin Weather
Dublin Airport Departures
Voy a Dublin - Guía de Dublin
Translate this page
www.voyadublin.com
Guía de Dublín. Información sobre Dublín 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
Fantásticas tiendas, divertidos pubs y un montón de lugares históricos hacen de Dublín un excelente destino para una escapada. Ah!, y según TripAdvisor ...
Qué ver en Dublín - Monumentos y visitas...
Translate this page
Display a menu

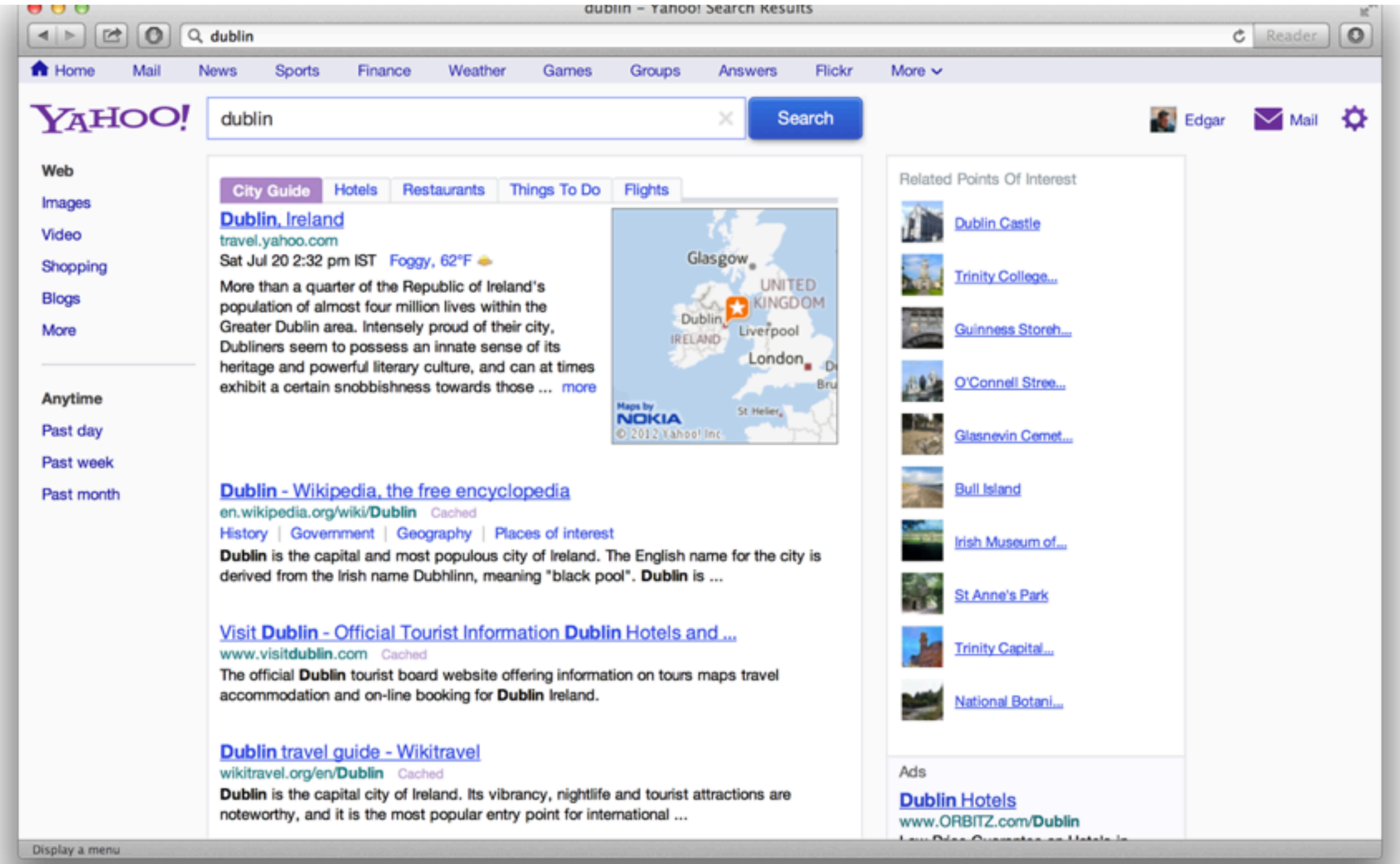
Bing


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..0i4€
Reader
+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
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/
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/
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/
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/
Real-time flight arrival and departure information, features about Dublin, Cork and Shannon Airports, Duty Free shopping outlets, airport maps, and future ...
Maynooth
Blanchardstown
Howth
Lucan
Dublin
Tallaght
Blackrock
Glenageary
Shankill
Naas
Bray
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 km²)
Weather: 22°C, Wind NE at 10 km/h, 69% Humidity
Local time: Saturday 2:30 PM
Points of interest
Display a menu

Google


Yahoo! Search – Web Search
search.yahoo.com
Reader
Check out the new Yahoo.com. Access Search, Mail and a virtually endless stream of content customized just for you. Try it now!
YAHOO!
Web Images Video Local Shopping News More
dublin
Search
dublin
dublin ireland
dublin ohio
dublin airport
dublin ca
dublin city schools
dublin dr pepper
dublin hotels
dublin bus
dublin irish festival
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
dublin
Reader
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
City Guide
Hotels
Restaurants
Things To Do
Flights
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
Glasgow
UNITED KINGDOM
Dublin
IRELAND
Liverpool
London
Maps by NOKIA
© 2012 Yahoo! Inc
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 Dubhlinn, 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!

YAHOO!

Web Search

Sign In Mail

- Mail
- News
- Finance
- Sports
- Movies
- omg!
- Shine
- Autos
- Shopping
- Travel
- Dating
- Jobs

More Y! Sites ›

Make YAHOO! your homepage

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

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

Watch the show »

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

London

52°F Fair

Today	Tomorrow	Tuesday
52° 41°	59° 37°	56° 38°

YAHOO!

Web Search

Sign In Mail

- Mail
- News
- Finance
- Sports
- Movies
- omg!
- Shine
- Autos
- Shopping
- Travel
- Dating
- Jobs

More Y! Sites ›

Make YAHOO! your homepage

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

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

Watch the show »

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

London
52°F Fair

Today	Tomorrow	Tuesday
52° 41°	59° 37°	56° 38°

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

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**

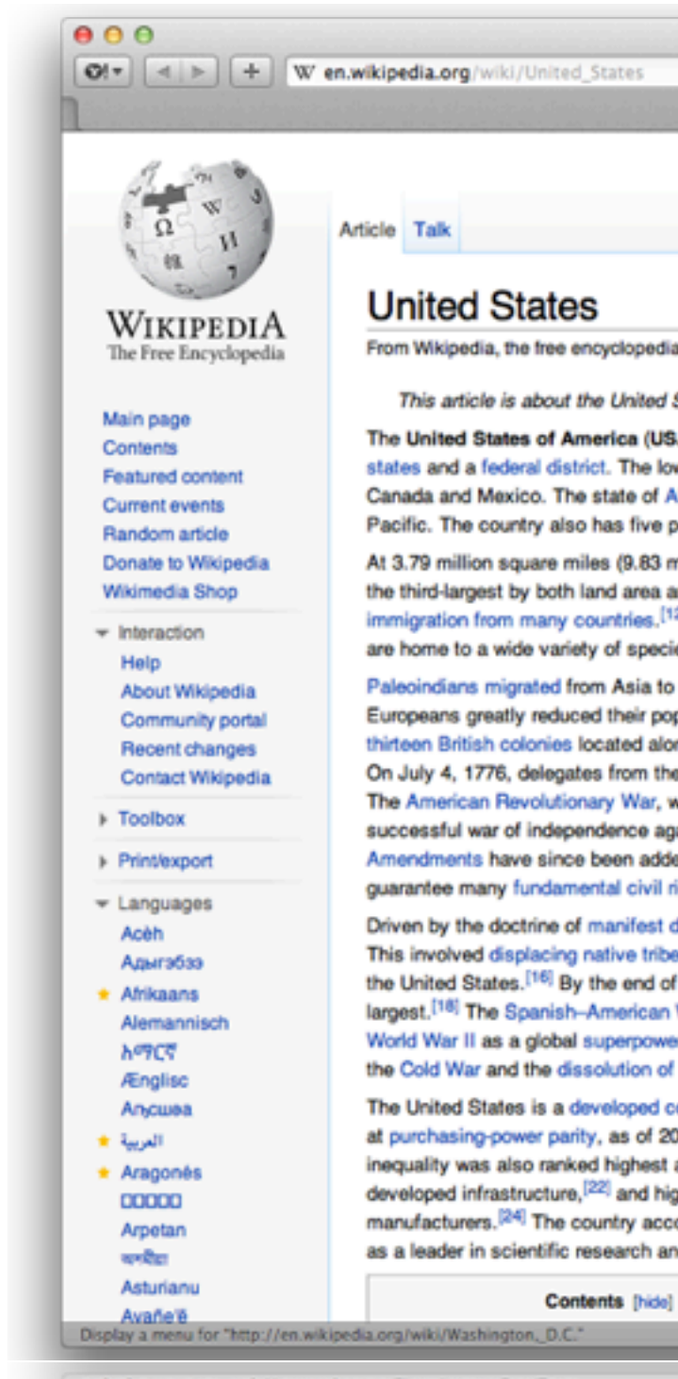
Methods

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.”

See http://en.wikipedia.org/wiki/Wikipedia:Summary_style

Disambiguation pages

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

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]**

$$\frac{\mathrm{CF}(w_l)}{\mathrm{CF}(w)}$$

Collection frequency
term w as a link to another Wikipedia article

Collection frequency
term w

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

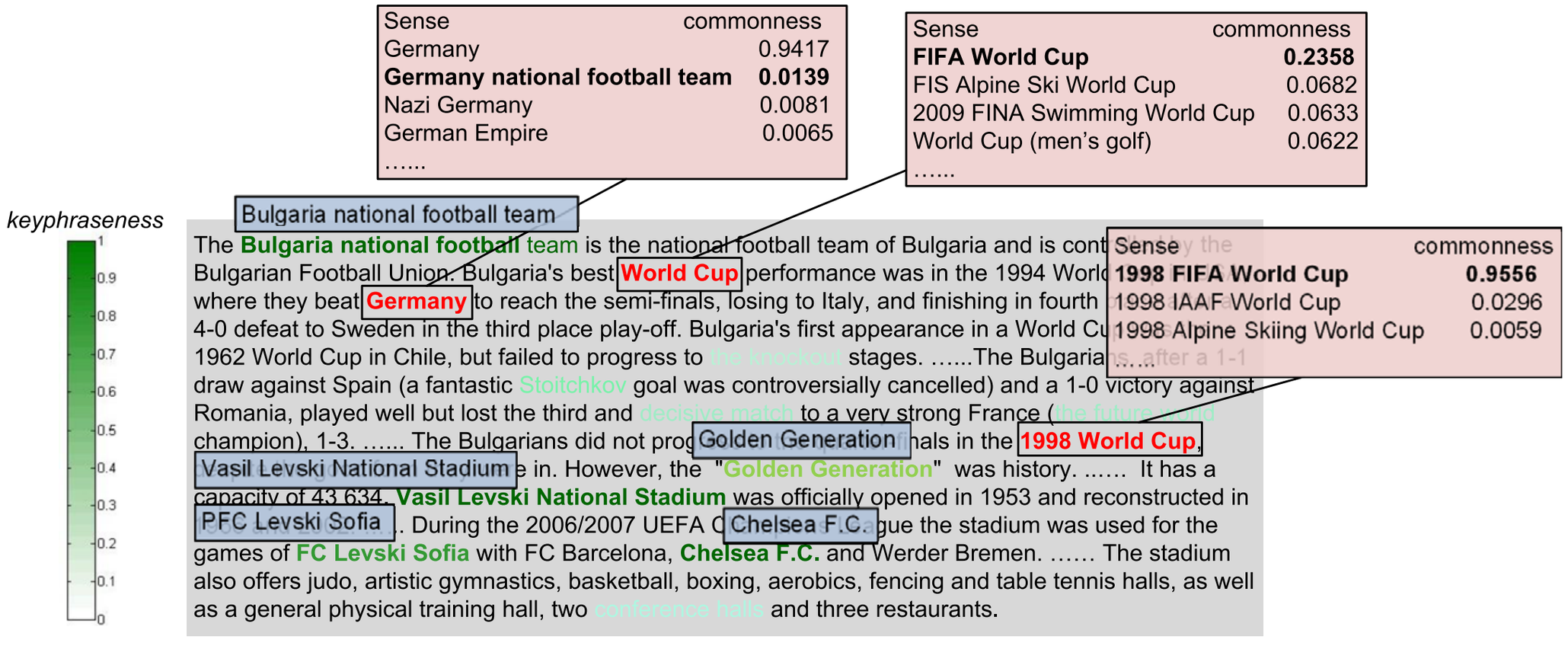


Image taken from Li et al. (2013). **TSDW: Two-stage word sense disambiguation using Wikipedia**. In JASIST 2013.

Wikipedia-based measures

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

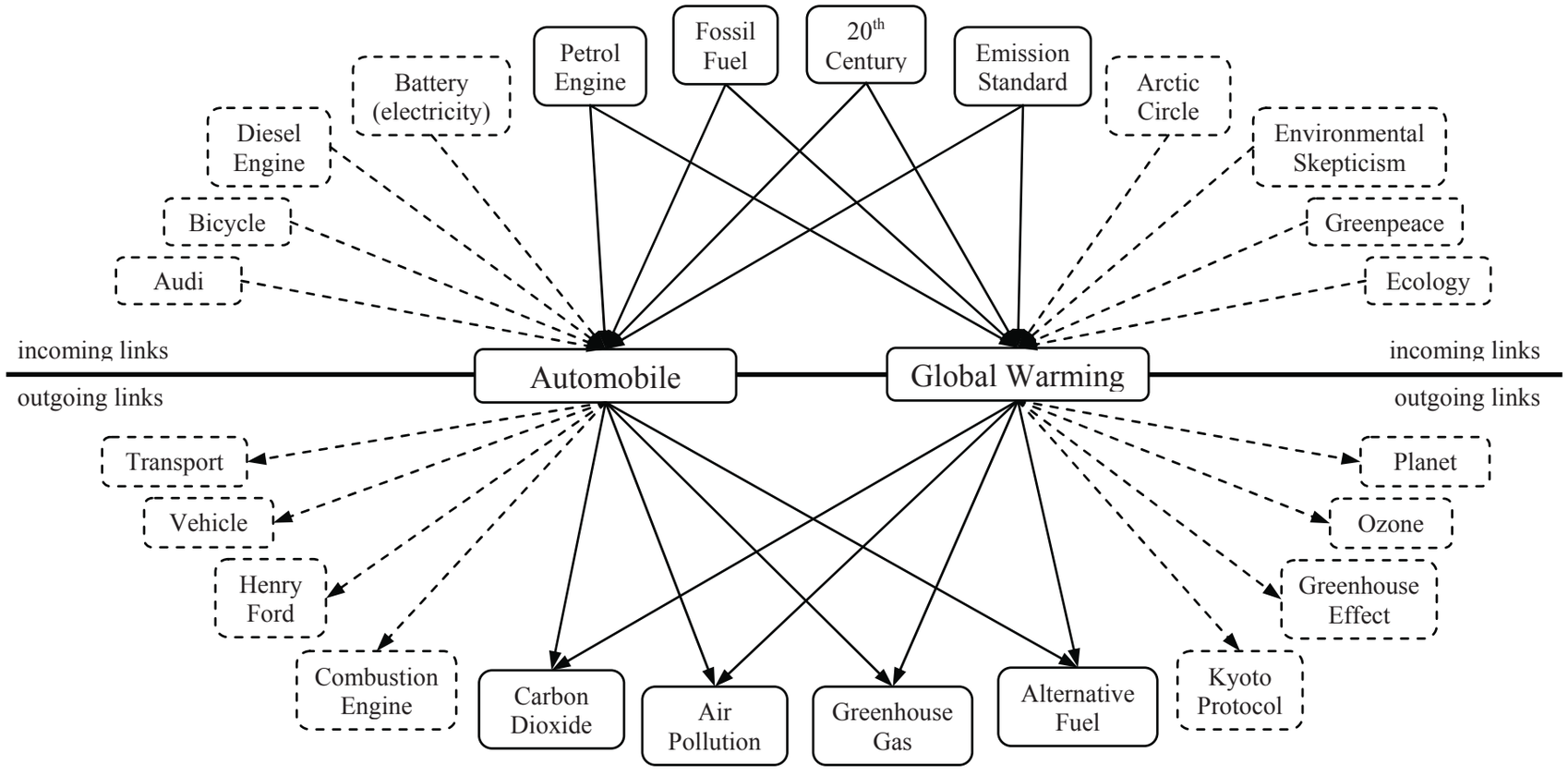


Image taken from Milne and Witten (2008a). **An Effective, Low-Cost Measure of Semantic Relatedness Obtained from Wikipedia Links**. In AAAI WikiAI Workshop.

Wikipedia-based measures

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

Number of links with target c

Intersection of inlinks with target c and c'

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

Total number of Wikipedia articles

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, X^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]

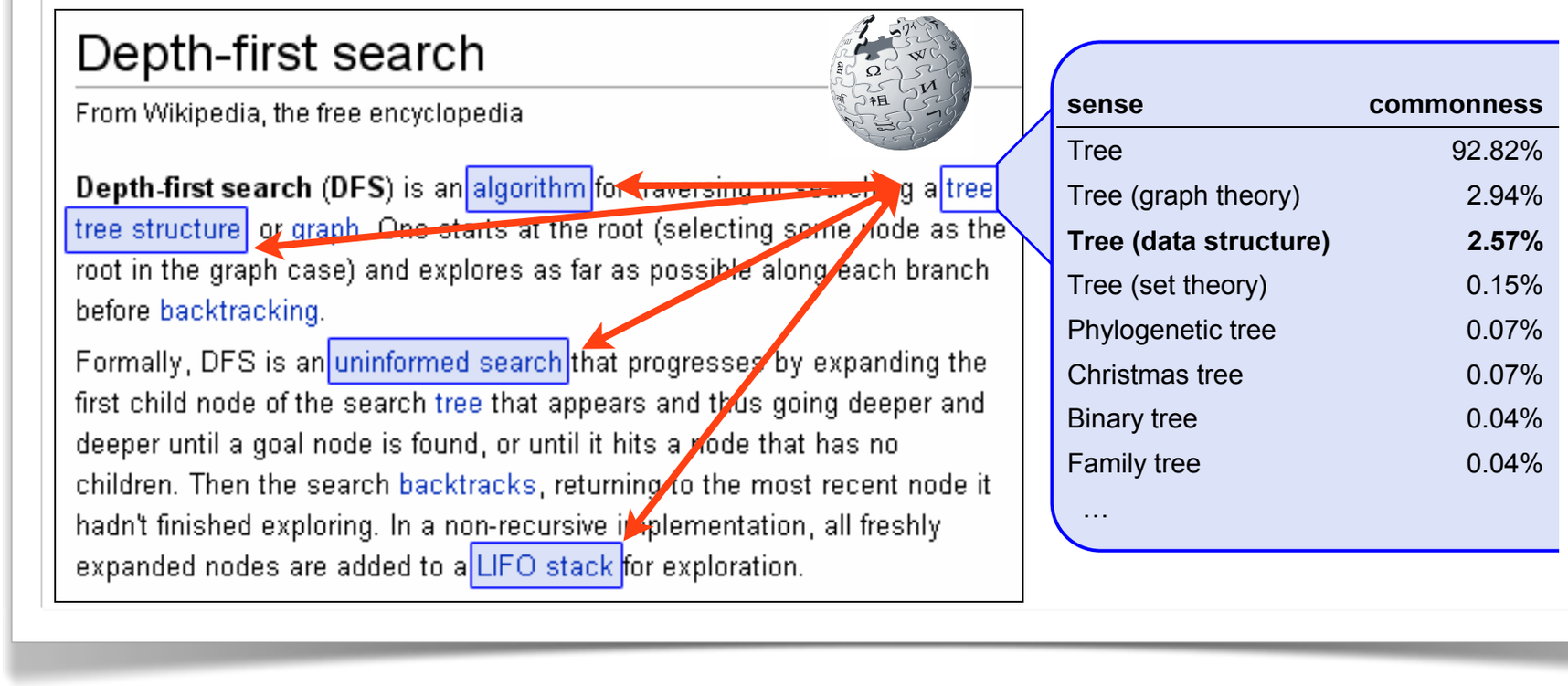


Image taken from Milne and Witten (2008b). **Learning to Link with Wikipedia**. In CIKM '08.

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

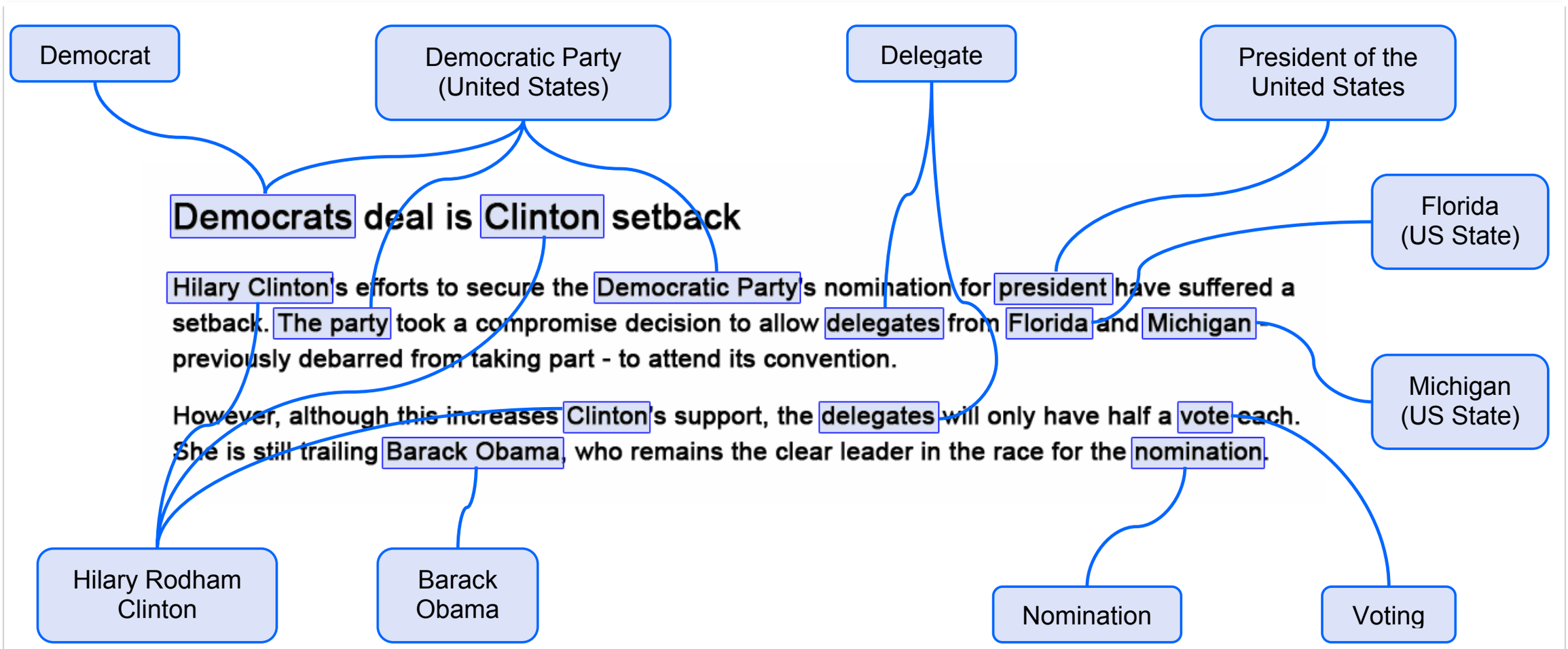


Image taken from Milne and Witten (2008b). **Learning to Link with Wikipedia**. In CIKM '08.

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]

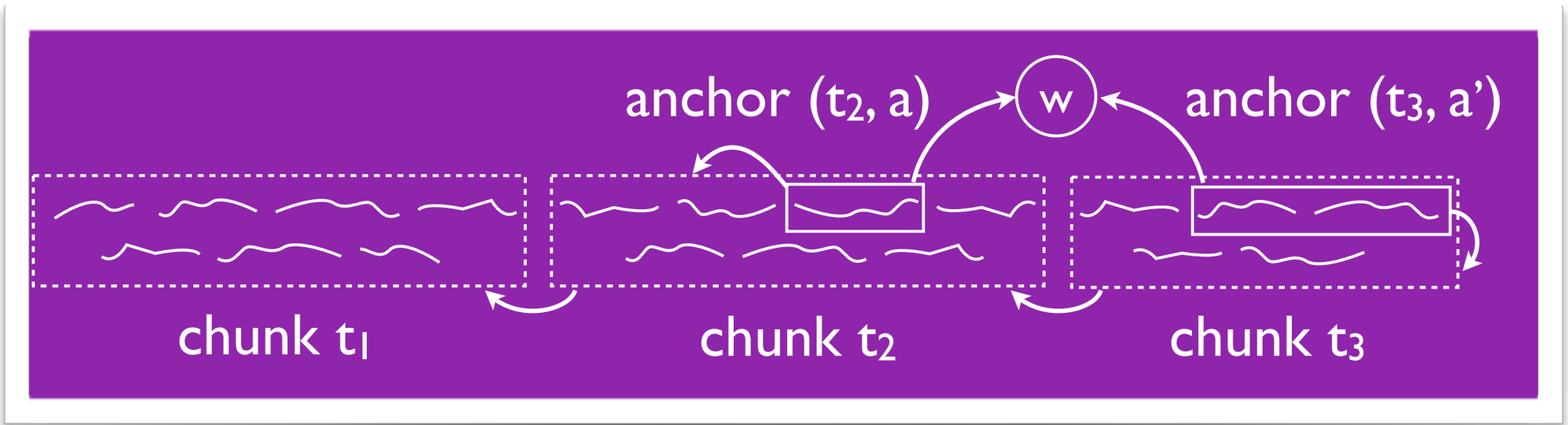
- MD
  - commonness (and others)
  - idea: obtain ranked list of **all** candidate entity links
- LG
  - use point-wise LeToR to determine which links to keep
    - ..., random forests, GBRT
    - {text, entity, text+entity, context} features
    - ([**Guo et al. 2013**] use a similar approach)
- Relatively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

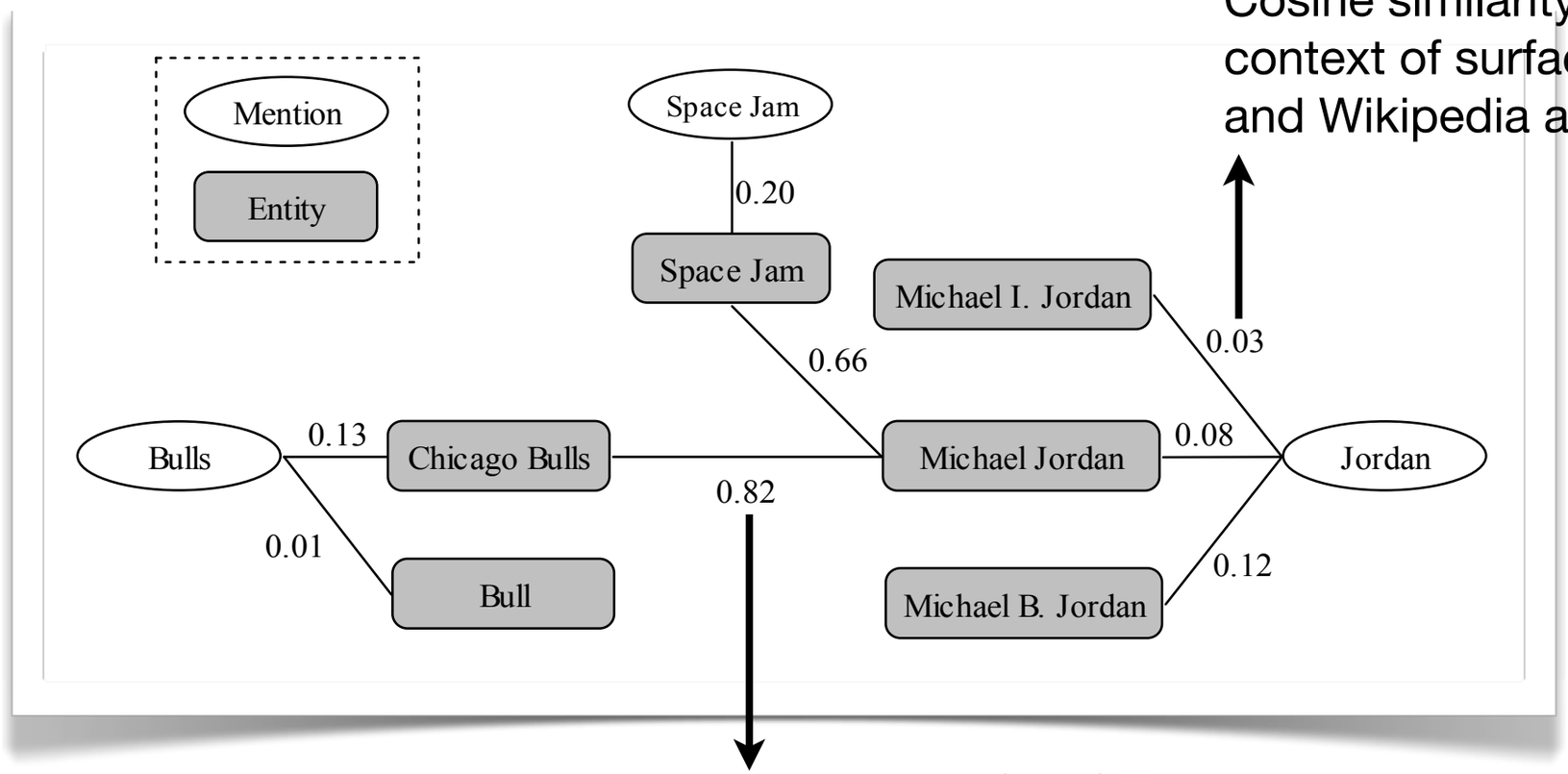
[Hachey et al. 2011]

- 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



Cosine similarity between context of surface form and Wikipedia article

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

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
Wikify!	*0.55*	*0.28*	*0.37*
Cucerzan	*0.71*	*0.33*	*0.45*
M&W	*0.80*	*0.38*	*0.52*
CSAW	*0.65*	*0.73*	*0.69*
Our Method(LC)	***0.52***	***0.34***	***0.41***
Our Method	***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, baltimor maryland, new hampshir, 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
Wikify!	*0.55*	*0.28*	*0.37*
Cucerzan	*0.71*	*0.33*	*0.45*
M&W	*0.80*	*0.38*	*0.52*
CSAW	*0.65*	*0.73*	*0.69*
Our Method(LC)	***0.52***	***0.34***	***0.41***
Our Method	***0.69***	***0.76***	***0.73***

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**
Wikify!	*0.55*	*0.28*	*0.37*
EM-Model	*0.82*	*0.48*	*0.61*
M&W	*0.80*	*0.38*	*0.52*
CSAW	*0.65*	*0.73*	*0.69*
EL-Graph	*0.69*	*0.76*	*0.73*
Our Method	***0.81***	***0.80***	***0.80***

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

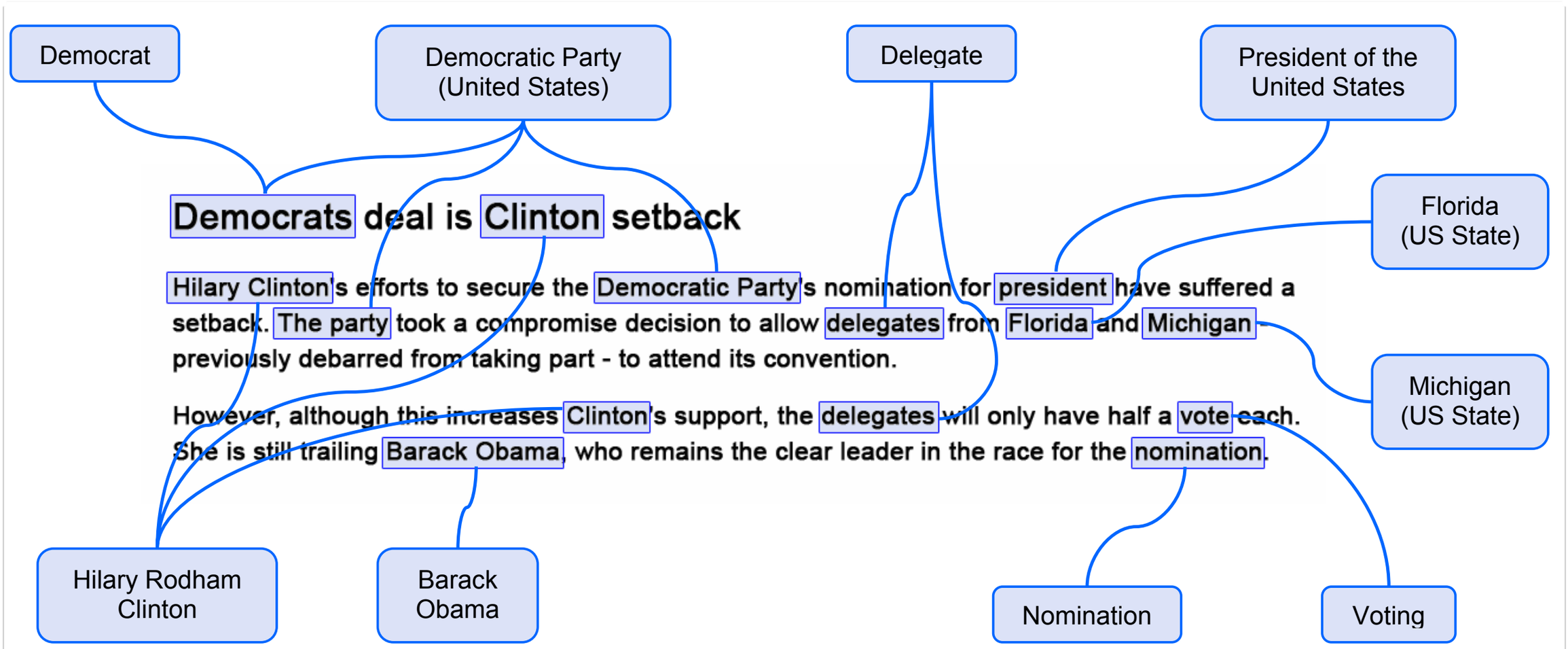


Image taken from Milne and Witten (2008b). **Learning to Link with Wikipedia**. In CIKM '08.

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

See http://research.microsoft.com/en-us/um/people/silviu/WebAssistant/TestData/

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

---

See http://cogcomp.cs.illinois.edu/page/resources/data

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

See http://cogcomp.cs.illinois.edu/page/resources/data

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

See http://edgar.meij.pro/dataset-adding-semantics-microblog-posts/

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	
$\|\mathcal{Q}\|$	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 & Pasça 2006]**
  - **[Cucerzan 2007]**
  - **[Varna et al. 2009]** (TAC system paper)

System	Extractor	Searcher: Condition	Title	Redirect	Link	Truncated	Bold	DABTitle	Filter	Disambiguator
Bunescu and Paşca (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	✔						in KB	
		else	✔	✔			✔	✔	NA	
		else								
		search 1	✔						in KB	
		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	**83.5**	37.2	68.1	96.5
Redirect	0.1	74.6	20.0	62.1	96.2
Link	4.2	55.7	**80.1**	**88.6**	59.5
Bold	1.6	45.1	48.8	71.7	67.2
Hatnote	0.0	42.6	1.2	57.7	**99.9**
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

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	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

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

Number of NIL queries in the test collection

System	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 Paşca	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

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

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

See http://acube.di.unipi.it/ (not available yet).

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 *and* Football		
	Soccer *and* soccer		
	Association football	サッカー	piłkarz
	fútbol *and* Fútbol	축구	voetbalclub
	footballer	footballeur	**ฟุตบอล**
	Futbol *and* futbol	Fußballspieler	bóng đá
	Fußball	sepak bola	voetbal
	futebol	足球	Foutbaal
	futbolista	فوتبال	futebolista
		футболист	لعبة كرة القدم
		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

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

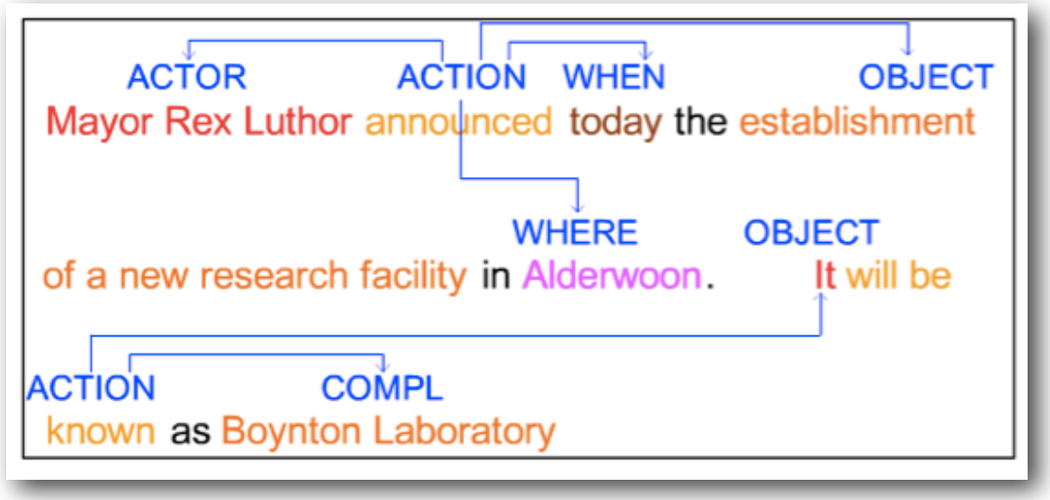
- Difficulty prediction
  - similar to ambiguity, but not the same
  - dependent on context, candidate links, ...
- Multi/Cross-lingual entity linking
  - **[Wang et al. 2013]**
  - CrossLink-2 (NTCIR-9), CJK - EN **[Tang et al. 2013]**
  - TAC...
- Cross-KB entity linking (“Freebase”)
  - directly?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

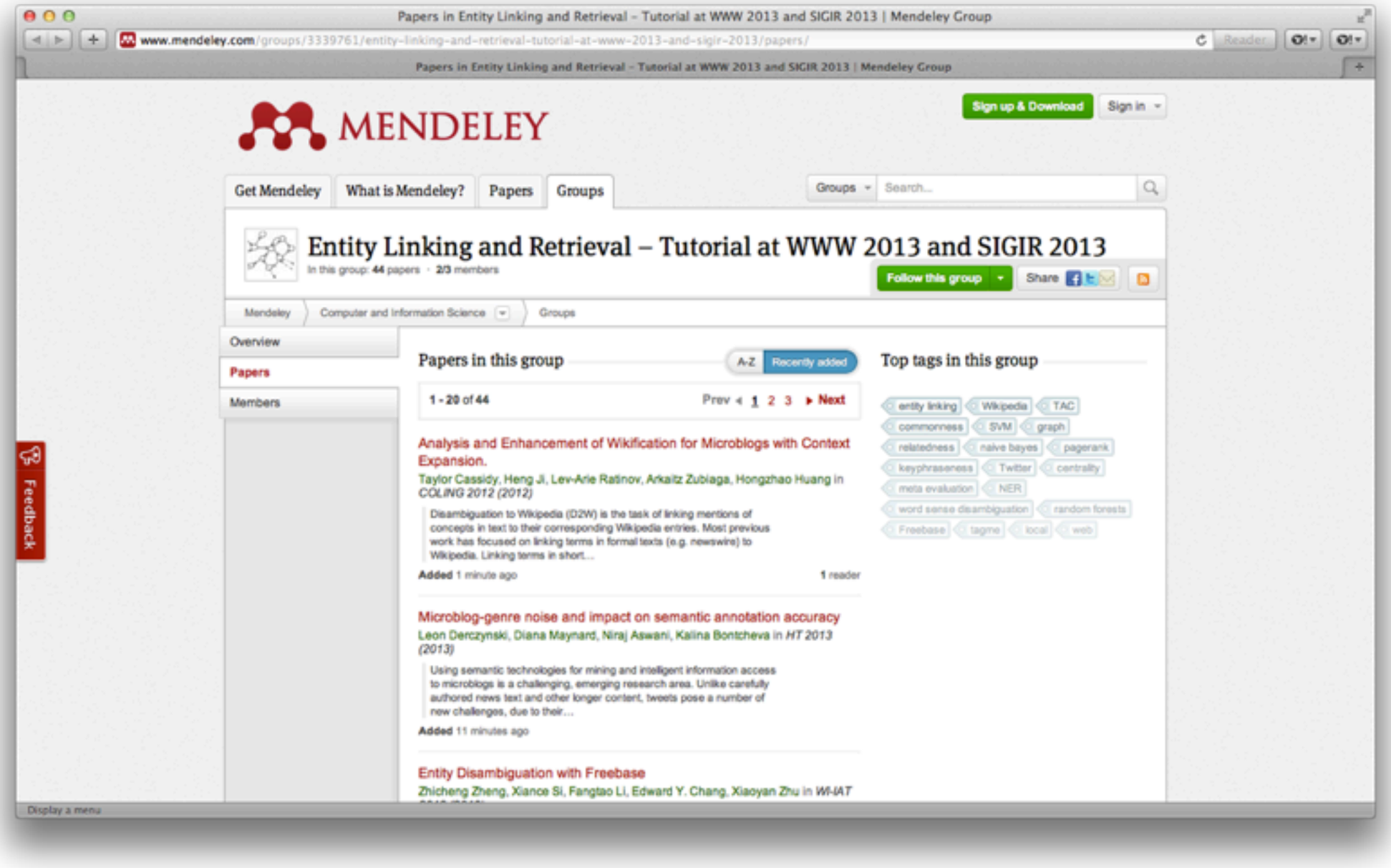
Open challenges

- 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



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