

# Entity Linking and Retrieval

**Edgar Meij** – @edgarmeij  
Yahoo! Research



**(Krisztian Balog, Daan Odijk)**





Minnesota Children's Museum **Deloitte.**

**GUIDANT**

 **NORTHWEST AIRLINES**

**ECOLAB**

**Yalspar**

  
**Donaldson**  
Filtration Solutions



  
**BEST BUY**

 **Carlson Companies**

  
**AMS**  
Solutions for Life®



  
**Medtronic**  
Alleviating Pain • Restoring Health • Extending Life

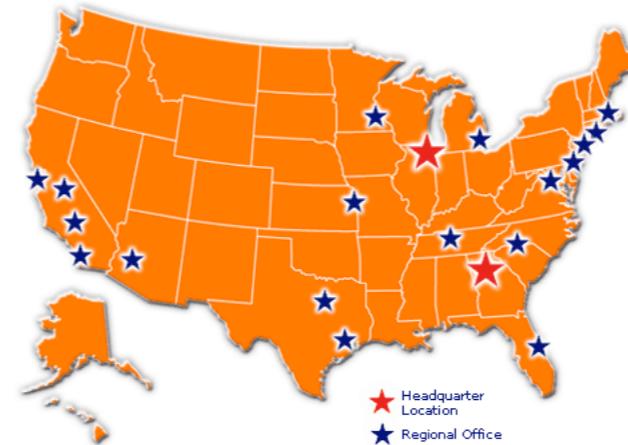
 **Thrivent Financial for Lutherans**

  
**MARVIN**  
Windows and Doors

 **POLARIS**  
The Way Out.

**3M** Worldwide

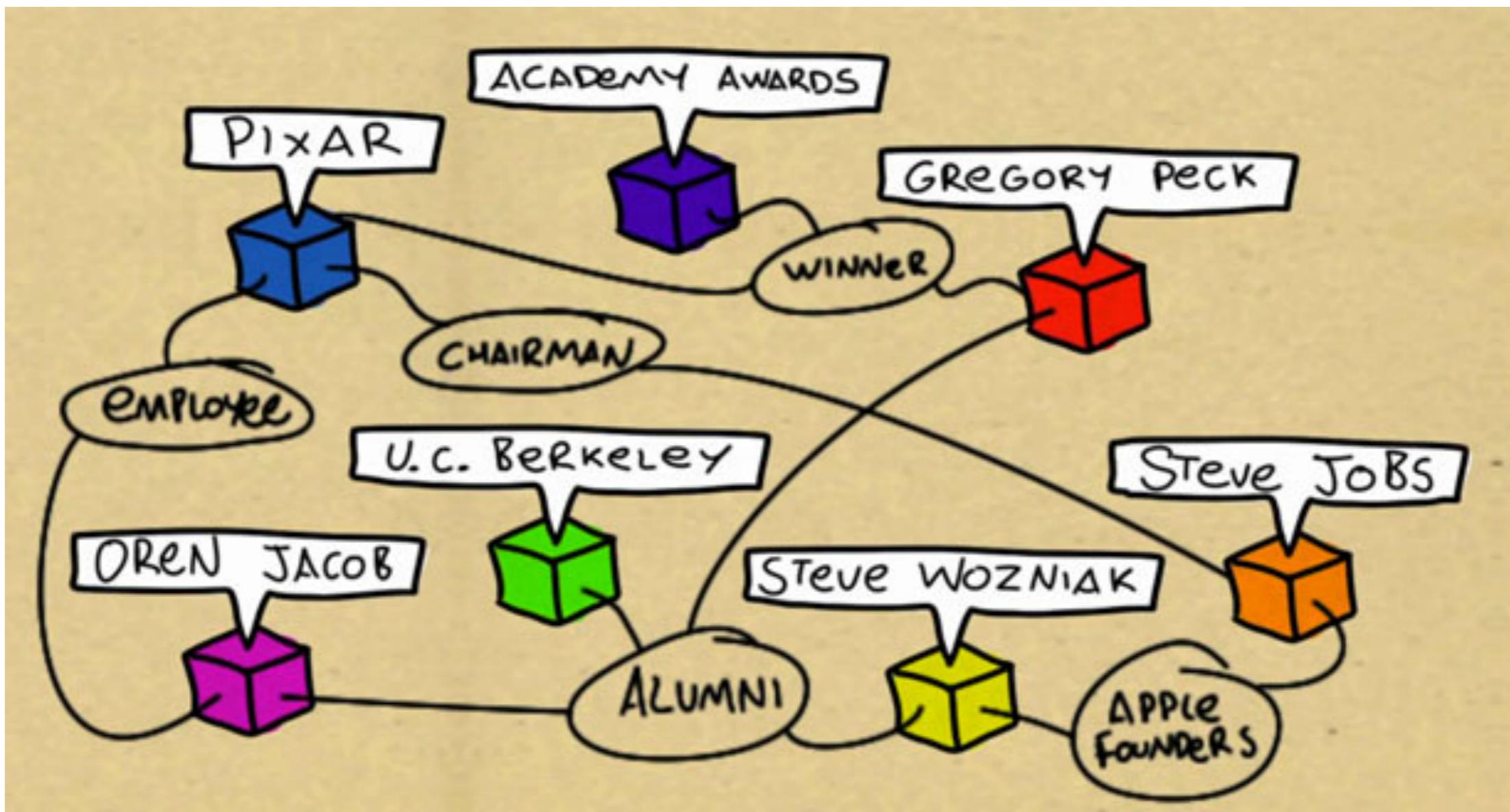
  
**GENERAL MILLS**



# **What is an entity?**

- Uniquely identifiable “thing” or “object”
  - “A thing with a distinct and independent existence”
- Properties:
  - ID
  - Name(s)
  - Type(s)
  - Attributes (/Descriptions)
  - Relationships to other entities

# What is an entity?



# Entity Linking

## Iranian POW negotiator holds talks with Iraqi ministers

The head of [Iran's prisoner of war](#) commission met with two [Iraqi](#) Cabinet ministers Saturday in a bid to glean information about thousands of Iranian POWs allegedly in Iraq, the official Iraqi News Agency reported.

Iraqi Foreign Minister [Mohammed Saeed al-Sahaf](#) told Abdullah al-Najafi that the two states needed to ``speed up the closure of what remains from the POW and Missing-In-Action file," INA said.

The issue of POWs and missing persons remains a stumbling block to normalizing relations between the two neighbors.

Iraq has long maintained that it has released all Iranian prisoners captured in the [1980-88 Iran-Iraq War](#). The countries accuse each other of hiding POWs and preventing visits by the [International Committee of the Red Cross](#) to prisoner camps.

The ICRC representative in [Baghdad](#), Manuel Bessler, told [The Associated Press](#) that his organization has had difficulty visiting POWs on both sides on a regular basis.

In April, Iran released 5,584 since [1990](#).

More than 1 million people w

### Baghdad

Baghdad is the capital of Iraq and of Baghdad Governorate. With a metropolitan area estimated at a population of 7,000,000, it is the largest city in Iraq. It is the second-largest city in the Arab world (after Cairo) and the second-largest city in southwest Asia (after Tehran).

[open in wikipedia](#)

fied as civil law detainees in the largest exchange

# Entity Retrieval

Indian restaurants in bangalore – Google Maps  
maps.google.co.in/maps?bav=on.2,or.r\_qf.&bvm=bv.48293060,d.bmk&biw=1158&bih=679&um=1&ie=UTF-8&q=indian+restaurants+in+bangal

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

Google Indian restaurants in bangalore SIGN IN

Get directions My places

Indian restaurants near Bangalore, Karnataka

A Serengeti  
Kanyakumari Rd, Bangalore, Karnataka 080 4000 3333 · zomato.com Category: Indian Restaurant 18 14 reviews · north indian food · jungle · main course "Good" -

B Woodys  
45/1, 5th Cross, 17th Main, J P Nagar, Bangalore, Karnataka 560078 080 2649 0888 · woodlands.in Category: Restaurants - South Indian 10 13 reviews · "Terrible place. We sat in a nominally air-conditioned place, but the AC ..." -

C Southindies  
Inner Ring Rd, Indira Nagar, Bangalore, Karnataka 560038 080 4163 6363 · thesouthindies.com Category: South Indian Restaurant 16 21 reviews · "good food but small portions. holier than thou

Indian restaurants near Bangalore, Karnataka

A Serengeti  
Kanyakumari Rd, Bangalore, Karnataka 080 4000 3333 · zomato.com Category: Indian Restaurant 18 14 reviews · north indian food · jungle · main course "Good" -

B Woodys  
45/1, 5th Cross, 17th Main, J P Nagar, Bangalore, Karnataka 560078 080 2649 0888 · woodlands.in Category: Restaurants - South Indian 10 13 reviews · "Terrible place. We sat in a nominally air-conditioned place, but the AC ..." -

C Southindies  
Inner Ring Rd, Indira Nagar, Bangalore, Karnataka 560038 080 4163 6363 · thesouthindies.com Category: South Indian Restaurant 16 21 reviews · "good food but small portions. holier than thou

# Menu

- Introduction
- Part 1 – Entity Linking
  - theory
  - hands-on
- Break
- Part 2 – Entity Retrieval
  - theory
  - hands-on

---

See <http://ejmeij.github.io/entity-linking-and-retrieval-tutorial/> or <http://bit.ly/yahoosummerschool> for the slides.

# References

---

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

# References

The screenshot shows a Mendeley group page with the following details:

- Title:** Entity Linking and Retrieval – Tutorial at WWW 2013 and SIGIR 2013
- Paper Count:** 44 papers
- Member Count:** 2/3 members
- Group Type:** Computer and Information Science
- Overview:** Shows the number of papers (1 - 20 of 44) and the date they were added.
- Papers:** A list of three research papers:
  - Analysis and Enhancement of Wikification for Microblogs with Context Expansion.** By Taylor Cassidy, Heng Ji, Lev-Arie Ratinov, Arkaitz Zubiaga, Hongzhao Huang in COLING 2012 (2012).  
Abstract: 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 in HT 2013 (2013).  
Abstract: 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 Zhicheng Zheng, Xiance Si, Fangtao Li, Edward Y. Chang, Xiaoyan Zhu in WI-IAT 2010 (2010).  
Abstract: 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...
- Top tags in this group:** 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, web.

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

# Outline

- Introduction
- Part 1 – Entity Linking
- Part 2 – Entity Retrieval

# **Part I**

# **Entity Linking**

# Outline

- Part 1 – Entity Linking
  - introduction
  - methods
  - evaluation
  - test collections
  - hands-on
  - 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**

Spin-Optical Metamaterial Route to Spin-Controlled Photonics

www.sciencemag.org/content/340/6133/724

Spin-Optical Metamaterial Route to Spin-Controlled Photonics

Science AAAS NEWS SCIENCE JOURNALS CAREERS BLOGS & COMMUNITIES MULTIMEDIA COLLECTIONS JOIN / SUBSCRIBE

Science The World's Leading Journal of Original Scientific Research, Global News, and Commentary.

Science Home Current Issue Previous Issues Science Express Science Products My Science About the Journal

Home > Science Magazine > 10 May 2013 > Shitrit et al., 340 (6133): 724–726

Article Views Abstract Full Text Full Text (PDF) Figures Only Supplementary Materials

Science 10 May 2013: Vol. 340 no. 6133 pp. 724–726 DOI: 10.1126/science.1234892

REPORT

## Spin-Optical Metamaterial Route to Spin-Controlled Photonics

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

\*Corresponding author. E-mail: [mehasman@technion.ac.il](mailto: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.  
Accepted for publication 13 March 2013.

ADVERTISEMENT

Science MOBILE Now Available for Android Phones

Scan the barcode to download from the Android Market.

ADVERTISEMENT

WOMEN IN SCIENCE forging new pathways in biology  
16 inspiring profiles

Display a menu for "http://oascentral.sciencemag.org/RealMedia/ads/click\_lx.ads/www.sciencemag.org/content/340/6133/724/L18/717884666/Right1/AAAS/SOL-CAR-HouseAdsAndroidApp101201/J-2957\_MobileApp-Android\_160x150b\_v1.html/306c72594d56474c34783441427"

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.

Input Text

Italiano English

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.

Many links

Few links

Reset

**TAGME!**

Tagged text

Topics

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 sp **Degenerate energy levels**  
In physics, two or more different quantum states are said to be degenerate if they are all at the same energy level. Statistically this means that they are all equally probable of being filled, and in...  
matter in a p [optical modes](#): emerged spin [interaction](#) of optical metar [ontical antenna patterns](#). This [tune](#) of metamaterial provides a route for spin-controlled

e [dispersion engineering](#) of a structured matter. We show that polarization-controlled [spatial inversion symmetry](#) is violated. The [emission](#) originates from the [spin-orbit](#) interaction based on [symmetry](#) restrictions in a spin-optical metasurface is obtained via [anisotropic](#)

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

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, 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  $\{\psi : H|\psi\rangle = \lambda|\psi\rangle\}$  (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 zero for all degenerate states is called a quantum number.

Office

HOME MY OFFICE PRODUCTS SUPPORT IMAGES TEMPLATES STORE

Search all of Office.com

About smart tags

You can save time by using smart tags to open other programs to do.

The purple dotted lines beneath text indicate where a smart tag has been applied.

Nate Sun dolor sit amet, consectetur adipiscing elit, sed diam nonummy nibh euismod tincidunt ut Steve Knopf magna aliquam erat volutpat.

① Smart tag indicators

- + How to use smart tags
- + How smart tags work
- + How to get more smart tags
- + Smart tag options
- + Creating smart tags and setting them up

Smart tagged text in Word

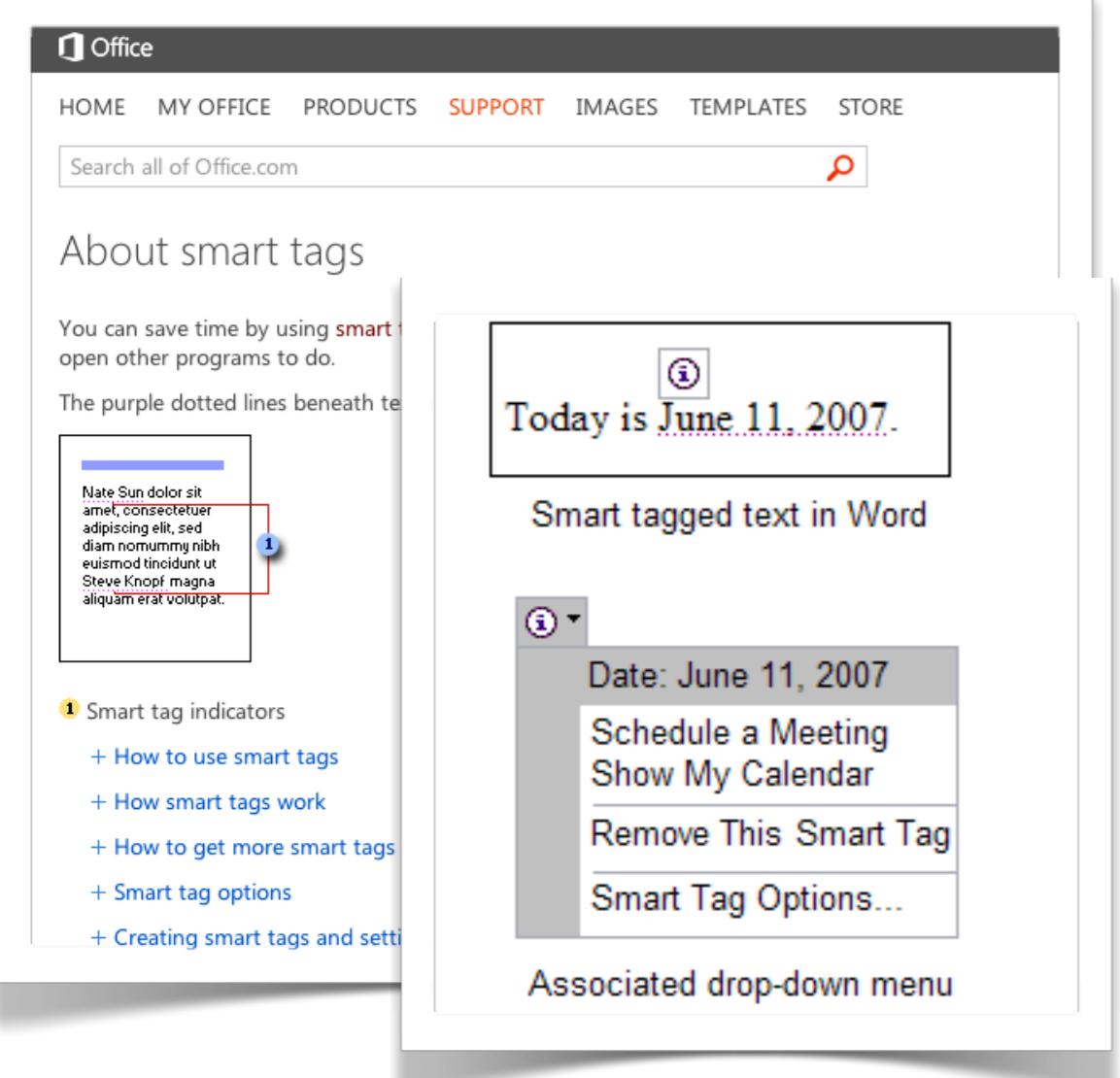
Date: June 11, 2007

Schedule a Meeting  
Show My Calendar

Remove This Smart Tag

Smart Tag Options...

Associated drop-down menu



# Microsoft Smart Tags

http://cse.unl.edu/~choueiry/S01-476-876/

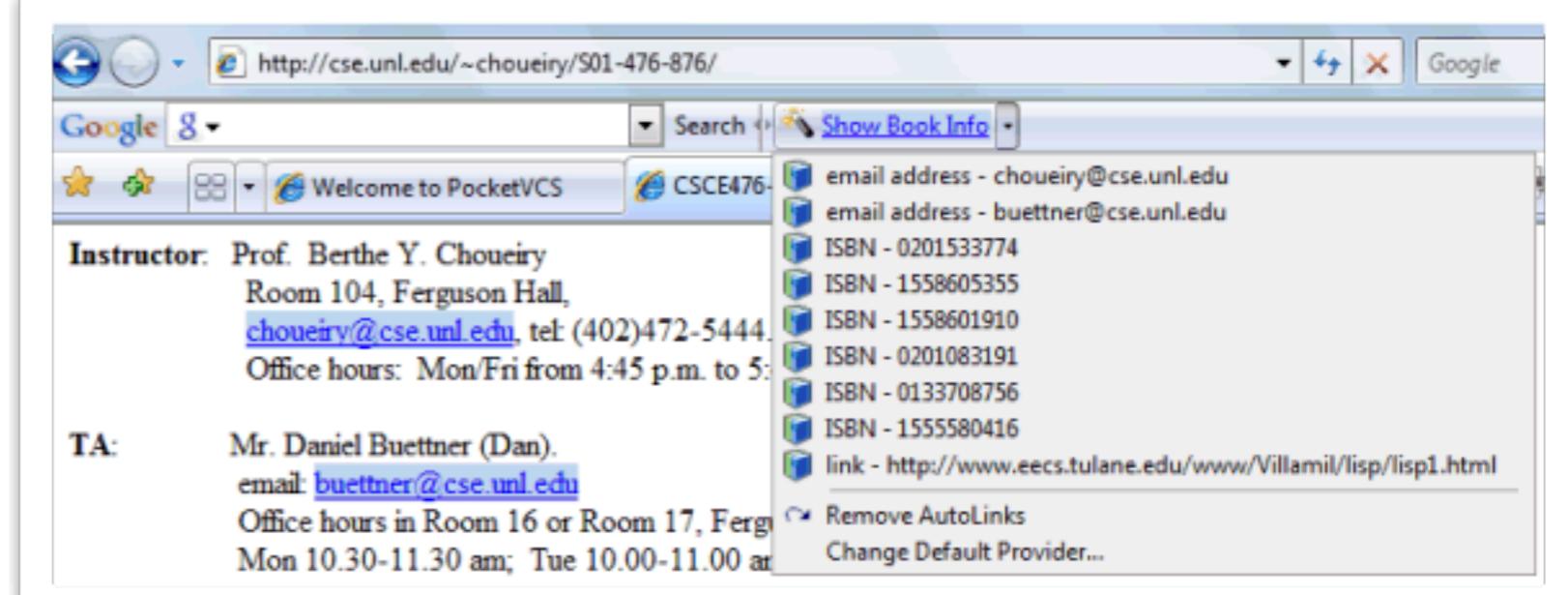
Google Welcome to PocketVCS CSCE476

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

email address - choueiry@cse.unl.edu  
email address - buettner@cse.unl.edu  
ISBN - 0201533774  
ISBN - 1558605355  
ISBN - 1558601910  
ISBN - 0201083191  
ISBN - 0133708756  
ISBN - 1555580416  
link - <http://www.eecs.tulane.edu/www/Villamil/lisp/lisp1.html>

Remove AutoLinks  
Change Default Provider...



# 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

Computed by Wolfram Mathematica      Source information      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**

**Bing**

bangalore, india - Bing

www.bing.com/search?q=bangalore%2C+india&go=&qs=n&form=QBRE&filt=all&pq=bangalore%2C+india&sc=8-15&sp=-1&sk=

Reader

WEB IMAGES VIDEOS NEWS MAPS MORE

bing Beta

Sign in

98,70,000 RESULTS Narrow by language ▾ Narrow by region ▾

[Images of bangalore, india](#)

bing.com/images

[Bangalore - Wikipedia, the free encyclopedia](#)

en.wikipedia.org/wiki/Bangalore ▾

Bangalore (or Bengaluru ['bengəluru]) is the capital city of the Indian state of Karnataka. Located on the Deccan Plateau in the south-eastern part of Karnataka ...

[Etymology](#) · [History](#) · [Geography](#) · [Civic administration](#) · [Economy](#) · [Transport](#)

[Bangalore India - Bangalore City - Bangalore Karnataka - Bangalore](#)

...  
www.bangaloreindia.org.uk/index.html ▾

Bangalore / Bengaluru city guide offering information on travel and tourism in Bangalore- the garden city of India.

[Bangalore travel guide - Wikitravel - Main Page - Wikitravel](#)

wikitravel.org/en/Bangalore ▾

[Get in](#) · [Do](#) · [Buy](#)  
Bangalore, also known as Bengaluru, is the capital of the Indian state of Karnataka. It is India's third-largest city with an estimated population of 8,474,970.

[News about Bangalore, India](#)

bing.com/news

[Bangalore University says no to bifurcation, favours new varsity](#)

Times of India · 16 hours ago

RELATED SEARCHES

Weather Bangalore India  
Hotels Bangalore India  
Bel India Bangalore  
3M India Bangalore  
Times of India Bangalore  
Bangalore Karnataka  
Goa India  
Bangalore India Map

# Bing

bangalore, india – Google Search

www.google.co.in/#sclient=psy-ab&q=bangalore%2C+india&oq=bangalore%2C+india&gs\_l=hp.3..0l4.7068.7068.0.8320.1.1.0.0.0.0.166.166.0j1. C Reader

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

SIGN IN

Web Images Maps News More Search tools ⚙

About 142,000,000 results (0.39 seconds)

**Bangalore - Wikipedia, the free encyclopedia**  
<https://en.wikipedia.org/wiki/Bangalore> ▾  
Bangalore is well-known as the hub of India's information technology sector. ... A succession of South Indian dynasties ruled the region of Bangalore until in ...  
[List of tourist attractions - History of Bangalore - Bengaluru International Airport](#)

**Bangalore - Maps of India**  
[www.mapsofindia.com/bangalore/](http://www.mapsofindia.com/bangalore/) ▾  
Mar 26, 2013 – Are you looking for information on Bangalore? Get detailed information on with Bangalore city covering Demographics, History, Transportation, ...

**Images for bangalore, india** - Report images

**Bengaluru (Bangalore), India - Travel Guide, Info & Bookings ...**  
[www.lonelyplanet.com/india/bengaluru-bangalore](http://www.lonelyplanet.com/india/bengaluru-bangalore) ▾  
Nov 10, 2008  
Bengaluru (Bangalore) travel recommendations and tips from Lonely Planet. Discover 83 things to do & 263 ...

**WHY BANGALORE IS SO FAMOUS CITY IN INDIA and IN THE ...**  
 | [www.youtube.com/watch?v=hDpsUTAi1JA](http://www.youtube.com/watch?v=hDpsUTAi1JA)

**Bangalore, Karnataka**

**Bangalore**  
City in India  
Bangalore is the capital city of the Indian state of Karnataka. Located on the Deccan Plateau in the south-eastern part of Karnataka, Bangalore is India's third most populous city and fifth-most populous urban agglomeration. Wikipedia

**Population:** 8.426 million (2011)  
**Area:** 741 km<sup>2</sup>  
**Founded:** 1537

# 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!**

bangalore, india

Search

bangalore india

bangalore india map

bangalore india time

bangalore india weather

bangalore india airport

bangalore india hotels

bangalore india news

bangalore india real estate

bangalore india postal code

bangalore india pictures

**BANGALORE, INDIA**  
04:43 PM (Asia/Kolkata). - Current local time

TOP RATED THINGS TO DO

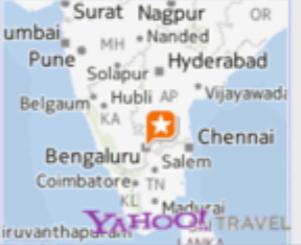
1. Group of Monuments at Hampi

BANGALORE OVERVIEW

Hotels

Flights





# Yahoo!

bangalore india – Yahoo! Search Results

Home Mail News Sports Finance Weather Games Groups Answers Flickr More ▾

**YAHOO!** bangalore india Search Sign In Mail Gear

Web Images Video Shopping Blogs More Anytime Past day Past week Past month

**City Guide** Hotels Flights

**Bangalore, India**  
travel.yahoo.com  
Tue Jun 25 7:16 pm IST

**WEATHER FORECAST**

Today	Showers/Wind Early	81°F   68°F
Tomorrow	Mostly Cloudy/Wind	85°F   68°F

See the extended weather forecast

A map of southern India and northern Sri Lanka. A red star marks the location of Bangalore. Labeled cities include Mumbai, Nagpur, Surat, Nanded, Hyderabad, Belgaum, Hubli, Vijayawada, Bengaluru, Coimbatore, Salem, Madurai, and Tiruvanthapuram. State abbreviations MH, KA, AP, TN, and SRI LANKA are also visible. The map is credited to Nokia.

**Bangalore - Wikipedia, the free encyclopedia**  
[en.wikipedia.org/wiki/Bangalore](http://en.wikipedia.org/wiki/Bangalore) Cached  
Etymology | History | Geography | Civic administration

Located on the Deccan Plateau in the south-eastern part of Karnataka, **Bangalore** is India's third most populous city and fifth-most populous urban agglomeration.

**Bangalore Tourism and Vacations: 175 Things to Do in ...**  
[www.tripadvisor.com/Tourism-g297628-Bangalore\\_Karnataka...](http://www.tripadvisor.com/Tourism-g297628-Bangalore_Karnataka.html) Cached  
Known as both the "Garden City" and "The Silicon Valley of India," **Bangalore** (officially "Bengaluru") is a techie's paradise, boasting the highest concentration of ...

**Bangalore India - Image Results**



VIP051  
YI Confidential [hide]

# Yahoo!

rio de janeiro – Yahoo! Search Results

Home Mail News Sports Finance Weather Games Groups Answers Flickr More ▾

YAHOO!  Search Sign In Mail

Web Images Video Shopping Blogs More Anytime Past day Past week Past month

**Rio De Janeiro, Brazil**  
travel.yahoo.com  
Tue Jun 25 10:48 am BST Fair, 77°F ☀  
Sitting on the southern shore of the magnificent Guanabara Bay, RIO DE JANEIRO has, without a shadow of a doubt, one of the most stunning settings in the world. Extending for 20km along an alluvial strip, between an azure sea and forest-clad mountains, the city's streets and buildings have been moulded ... [more](#)



**Rio de Janeiro - Wikipedia, the free encyclopedia**  
[en.wikipedia.org/wiki/Rio\\_de\\_Janeiro](http://en.wikipedia.org/wiki/Rio_de_Janeiro) Cached  
Geography | Climate | History | City districts  
Rio de Janeiro, commonly referred to simply as **Rio**, is the capital city of the State of Rio de Janeiro, the second largest city of Brazil, and the third largest ...

**Rio de Janeiro travel guide - Wikitravel - Main Page - Wikitravel**  
[wikitravel.org/en/Rio](http://wikitravel.org/en/Rio) Cached  
Rio de Janeiro is the second largest city in Brazil, on the South Atlantic coast. Rio is famous for its breathtaking landscape, its laidback beach culture and its ...

**Rio De Janeiro - Image Results**



Related Points Of Interest

 [Ipanema](#)  
 [Campo Grande, B...](#)  
 [Copacabana](#)  
 [Sugar Loaf Moun...](#)  
 [Resende, Brazil](#)  
 [Carnaval](#)  
 [Santana](#)  
 [Saúde](#)  
 [HSBC Arena](#)  
 [Rocha Miranda](#)

Ad

**Quikr - Free Classifieds**  
[Quikr.com/India-classifieds](http://Quikr.com/India-classifieds) Buy/Sell/Rent anything... Quick, easy, free! VIP051 YI Confidential [hide]

# Yahoo!

- [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 ▾



### 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°

[Quote & Tools](#) [Test H500](#)

# Yahoo! Homerun

- [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 supported President Barack Obama. [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 done? [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°

# Yahoo! Homerun



## 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](#)

London

## 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 look](#)



[Ad Feedback](#)

# Yahoo! Homerun

# **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)

# Why do we need entity linking?

- “Use as feature”
  - to improve
    - classification
    - retrieval
    - word sense disambiguation
    - semantic similarity
    - ...
  - dimensionality reduction (as compared to, e.g., term vectors)

# Why do we need entity linking?

- Enable
  - 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
- Knowledge base population
- Entity linking
  - D2W
  - Wikification
  - Semantic linking

# Entity Linking: main problem

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

# 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

- Wikipedia
- 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]]



# Disambiguation pages

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

The screenshot shows a web browser window displaying the Wikipedia disambiguation page for "United States". The title bar reads "United States (disambiguation)". The main content area is titled "United States (disambiguation)" and includes the subtext "From Wikipedia, the free encyclopedia". Below the title, there are several sections: "Countries" (with a link to edit), "Current" (with a link to edit), "Historical" (with a link to edit), "Proposed" (with a link to edit), and "Fictional" (with a link to edit). Each section contains a bulleted list of related topics. On the left side of the page, there is a sidebar with links to "Main page", "Contents", "Featured content", "Current events", "Random article", "Donate to Wikipedia", "Interaction" (with links to "Help", "About Wikipedia", "Community portal", "Recent changes", and "Contact Wikipedia"), "Toolbox", "Print/export", "Languages" (listing Arabic, Czech, Deutsch, Español, فارسی, Français, Italiano, עברית, Magyar, 日本語, Polski, Русский, Slovenčina, 中文), and "Edit links". At the bottom of the sidebar, there is a link "Display a menu". The overall layout is characteristic of the early 2000s Wikipedia interface.

# 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]

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

# Wikipedia-based measures

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

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



**Collection frequency**  
term  $w$

# Wikipedia-based measures

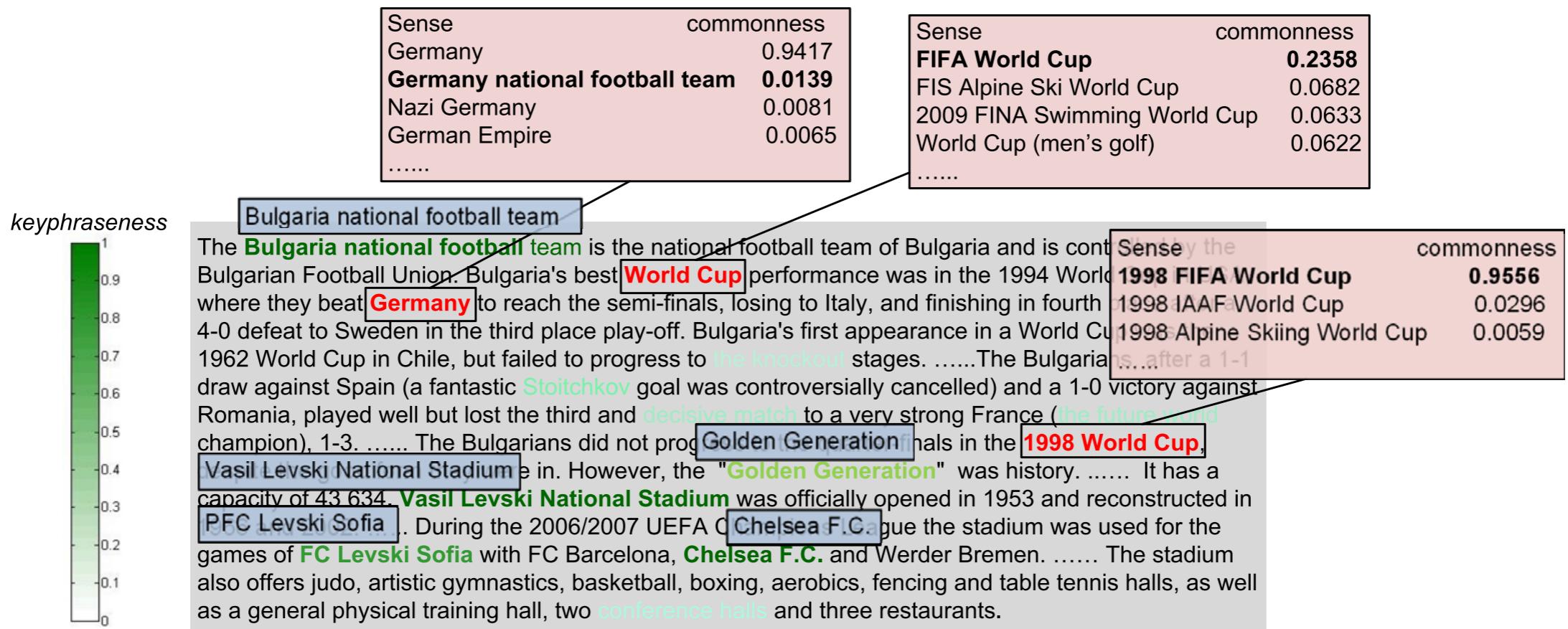
- commonness( $w, c$ ) [Medelyan et al. 2008]

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



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

# Commonness and keyphraseness



# Wikipedia-based measures

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

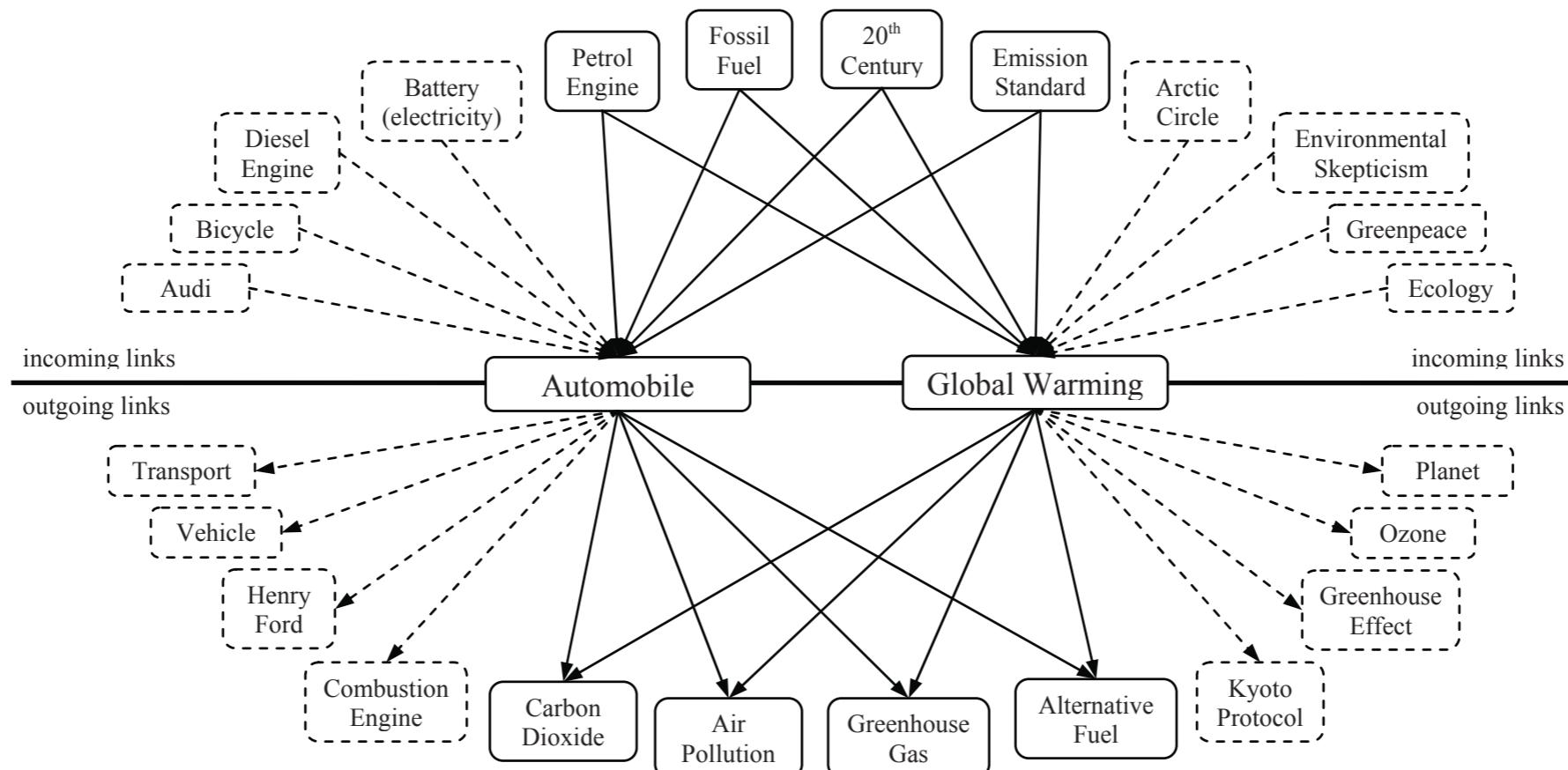


Image taken from Milne and Witten (2008a). An Effective, Low-Cost Measure of Semantic Relatedness Obtained from Wikipedia Links. In AAAI WikiAI Workshop.

# 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
- LG
  - Represent entities as vectors
    - context, categories
  - Same for all candidate entity links
  - Determine maximally coherent set

# 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
  - compare each possible sense with its *relatedness* to the context sense candidates
  - start with unambiguous senses

# 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	relatedness
Tree	92.82%	15.97%
Tree (graph theory)	2.94%	59.91%
**Tree (data structure)**	**2.57%**	**63.26%**
Tree (set theory)	0.15%	34.04%
Phylogenetic tree	0.07%	20.33%
Christmas tree	0.07%	0.0%
Binary tree	0.04%	62.43%
Family tree	0.04%	16.31%
...		

# **Learning to Link with Wikipedia**

**[Milne & Witten 2008b]**

- 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
- Combine

# **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

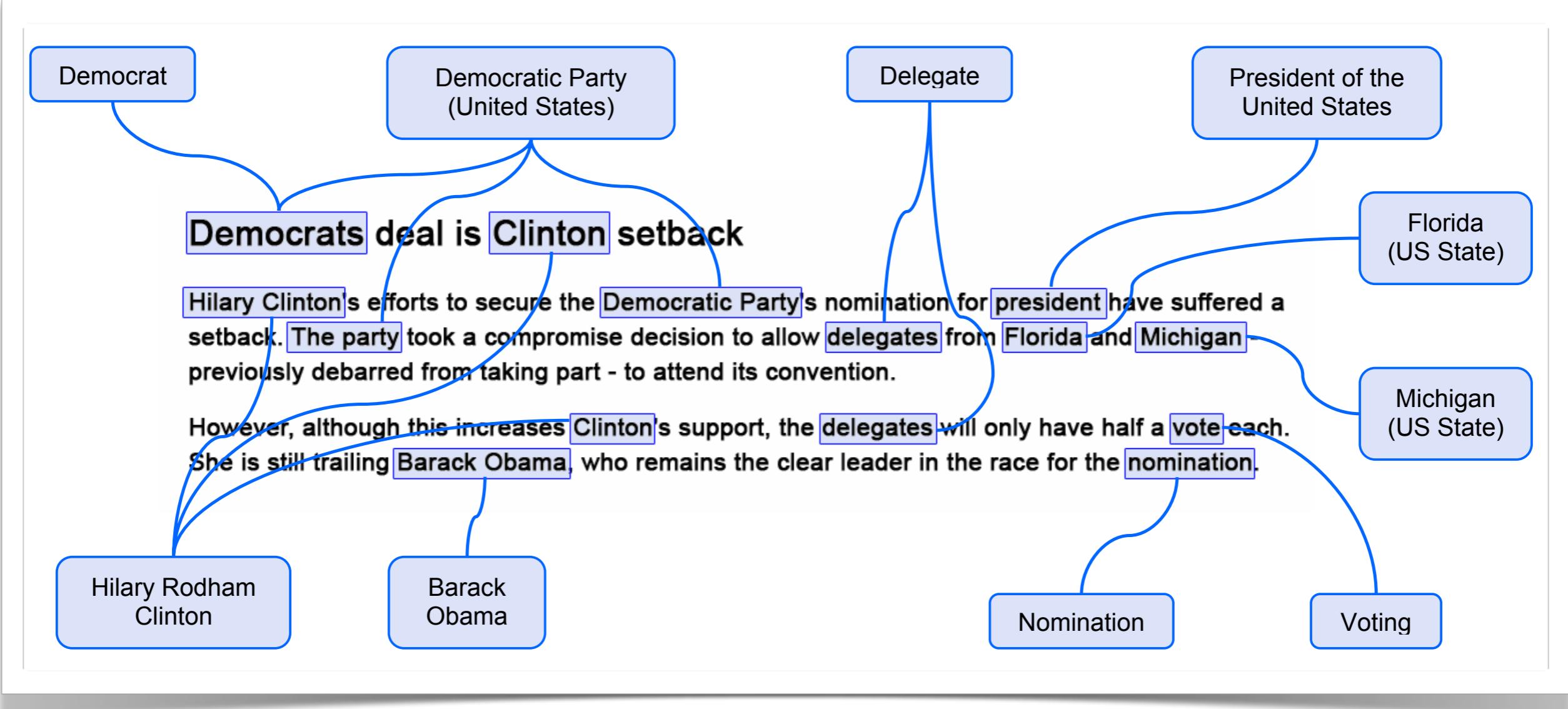


Image taken from Milne and Witten (2008b). Learning to Link with Wikipedia. In CIKM '08.

# Local and Global Algorithms for Disambiguation to Wikipedia

[Ratinov et al. 2011]

- Explicit focus on *global* versus *local* algorithms
  - “Global,” i.e., disambiguation of the candidate graph
  - NP-hard
- Optimization
  - reduce the search space to a “disambiguation context,” e.g.,
    - all plausible disambiguations [Cucerzan 2007]
    - unambiguous surface forms [Milne & Witten 2008b]

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

**[Ratinov et al. 2011]**

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

# **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, 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]

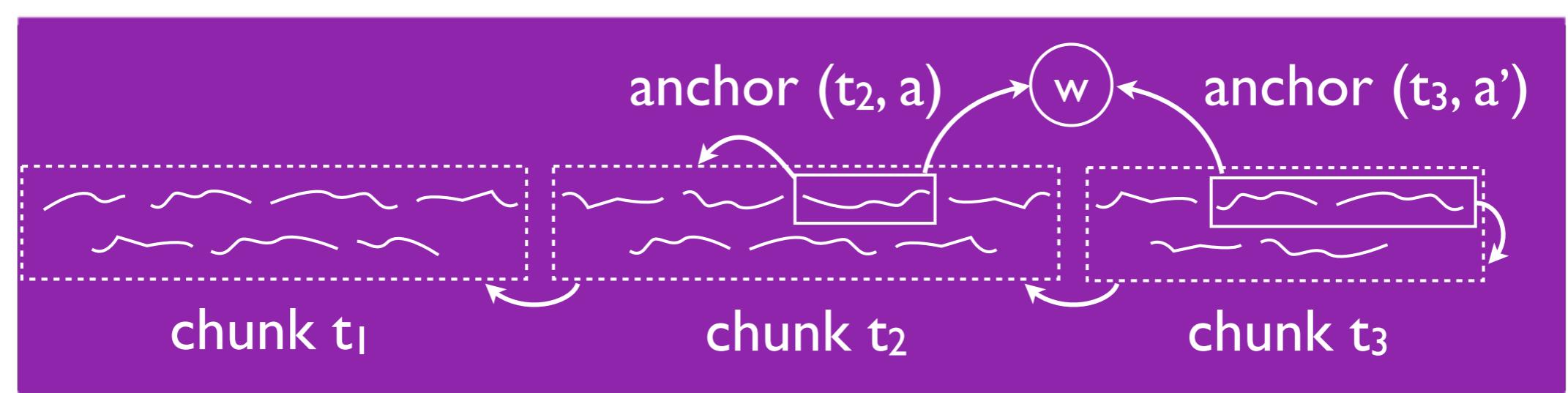
- MD
  - commonness (and others)
  - idea: obtain ranked list of **all** candidate entity links
- LG
  - use machine learning to determine which of the links to keep
    - ..., random forests, GBRT
    - big set of {text, entity, text+entity, context} features

# **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

# Recap

- Essential ingredients
  - MD
    - commonness
    - keyphraseness
  - LG
    - commonness
    - machine learning
  - DA
    - relatedness
    - machine learning

# Outline

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

# **Evaluation**

# DIY Entity Linking

- Ingredients
  - target KB (e.g., Wikipedia)
  - test collection
  - evaluation metrics

# DIY Entity Linking

- Ingredients
  - Target KB (Wikipedia)
    - wikipedia-miner
    - Google's Dictionaries for Linking Text, Entities and Ideas
  - Test collection
  - Evaluation metrics

# Measures

- Set-based (similar to WSD)
  - “How many correct links were retrieved?”
  - macro/micro precision, recall, F-measure
- Ranking-based

# 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 ranking-based metrics for entity linking**

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

# **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 2006 Wikipedia
  - 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
- Collected by correcting the output of a 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	
$ \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.

# DIY Entity Linking

- Target KB (Wikipedia)
  - wikipedia-miner
  - Google's Dictionaries for Linking Text, Entities and Ideas
- Test collection
- Evaluation metrics

# 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

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

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 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

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

# 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
- “Fuzzy” evaluation measures
- Main findings
  - Different systems perform well in different scenarios
  - AIDA and TagMe seem to be the winners overall

# Outline

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

# **Hands-on**



# **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
- 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/>

# DBpedia Spotlight

[Mendes et al., 2011]

- Open source
- Public web service
- LingPipe Exact Dictionary-Based Chunker
- Disambiguation in local context
  - Vector-space Model using Bag-of-Words
  - Cosine similarity
- See <http://spotlight.dbpedia.org>

# Illinois Wikifier

[Ratinov et al. 2011]

- Local install
- 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)

# AIDA

[Yosef et al. 2011]

- Open source
- (Public) web service
- Uses Stanford NER system
- 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, more than 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	?	

# Code Academy

- Contains some (Javascript) coding examples for entity linking and retrieval
  - <http://www.codecademy.com/courses/javascript-beginner-en-LkhDf/>

# Outline

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

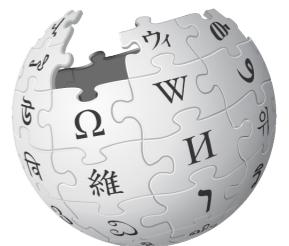
# **Open challenges**

# Open challenges

- Difficulty prediction
  - similar to ambiguity, but not the same
  - dependent on context, candidate links, ...
- Cross-lingual entity linking [**Wang et al. 2013**]
- Cross-KB entity linking (“Freebase”)
  - use Wikipedia as pivot
  - directly
    - lexical matching
    - machine learning (if annotators/training data available)

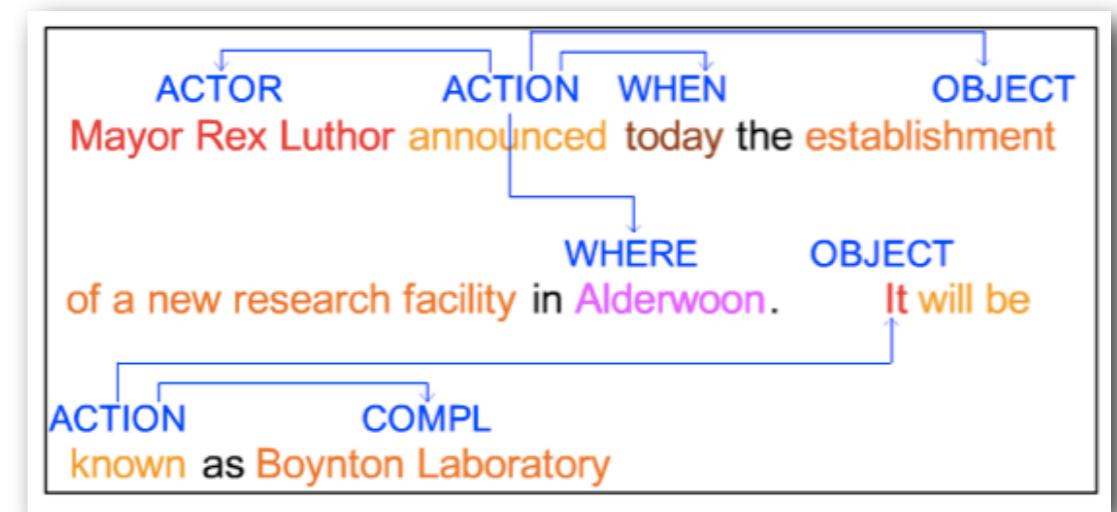
# Learning/Updating the KB

- 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? Evaluation?
    - TAC KBP?
    - set-based? ranking? known-item finding? top- $r$ ?
    - exhaustive linking? first mention only?
    - “aboutness”
- Moving beyond entities
  - events/news
  - concepts
  - relations

# Open challenges

- What if there is no/little textual evidence?
- Move beyond "ad hoc" entity linking:
  - incorporate contextual evidence in the task (and evaluation)
  - {users, history, profile, social, trending, ...}
- Stream-based entity linking

# Follow-up reading

- Detecting unlinkable entities [**Lin et al. 2012a**]
- Linking entities to any database [**Sil et al. 2012**]
- 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**]

# **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/>

# 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 44 papers. The first paper listed is "Analysis and Enhancement of Wikification for Microblogs with Context Expansion" by Taylor Cassidy, Heng Ji, Lev-Arie Ratinov, Arkaitz Zubiaga, Hongzhao Huang in COLING 2012 (2012). The second paper is "Microblog-genre noise and impact on semantic annotation accuracy" by Leon Derczynski, Diana Maynard, Niraj Aswani, Kalina Bontcheva in HT 2013 (2013). The third paper is "Entity Disambiguation with Freebase" by Zhicheng Zheng, Xiancse Si, Fangtao Li, Edward Y. Chang, Xiaoyan Zhu in WI-IAT 2013 (2013). 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/>

# **Part II**

# **Entity Retrieval**

# **Introduction**

# **Entity retrieval tasks**

- Ad-hoc entity retrieval
- List completion
- Question answering
  - Factual questions
  - List questions
  - Related entity finding
- Type-restricted variations
  - People, blogs, products, movies, etc.

total length of US highways - Wolfram|Alpha

http://www.wolframalpha.com/input/?i=total+length+of+US+highways

Google

# WolframAlpha™ computational knowledge engine

total length of US highways

Examples Random

Assuming "US" is a country | Use as a city instead

**Input Interpretation:**

United States | length of highways

**Result:**

271 900 km (kilometers) (2007 estimate)

Show non-metric

**History:**

(from 1990 to 2007)  
(in thousands of kilometers)

Road network:

	Show non-metric	Total number
motorways	75 111 km (kilometers) (2007 estimate)	
highways	271 900 km (kilometers) (2007 estimate)	
secondary roads	1.7 million km (kilometers) (2007 estimate)	
others	4.5 million km (kilometers) (2007 estimate)	
total	6.5 million km (kilometers) (2007 estimate)	

For the geek who has everything

Wolfram gear »

Related Wolfram|Alpha Queries

- = length of highways of North...
- = people killed in road accide...
- = length of highways of Canada
- = length of highways of Austra...

# expert finding

## EXPERTS

### language technology

Bogers, Drs. Toine M.	Arts
Bosch, Dr. Antal P.J. van den	Arts
Broeder, Dr. Peter	Arts
Canisius, Drs. Sander V.M.	Arts
Daelemans, Prof. dr. Walter M.P.	Arts
Geertzen, Jeroen	Arts
Keizer, Dr. ir. Simon	Arts
Marsi, Dr. Erwin C.	Arts
Reynaert, Dr. Martin W.C.	Arts
Sporleder, Dr. C.E.	Arts
Werf, Drs. Rintse van der	Arts

### See also:

[computer linguistics](#)  
[language technology and computers](#)



**EXPERTS**  
A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

**EXPERTISE**  
A B C D E F G H I J K L M N O P Q R S T U V W X Y Z



# expert profiling

## Antal P. J. van den Bosch

### Lecturer

Faculty of Arts  
Language and information science



Room D 343  
P.O. Box 90153  
NL-5000 LE Tilburg, The Netherlands

Phone +31 13 466 3117  
Fax +31 13 466 2892  
E-mail: [Antal.vdnBosch@uvt.nl](mailto:Antal.vdnBosch@uvt.nl)

### Present

	mon	tue	wed	thu	fri
morning	✓	✓	✓	✓	
afternoon	✓	✓	✓	✓	

[research](#)  
[study guide](#)  
[personal homepage](#)

### Expertise

My research is positioned in the intersection between artificial intelligence and linguistics. I am specialized in machine learning and language technology / computational linguistics. As for applications, I have professional experience with speech synthesis, the automatic syntactic and semantic analysis of text, text mining, dialogue systems, and spelling correction.

### Subjects

[artificial intelligence](#)  
[computer linguistics](#)



Indian restaurants in bangalore



SIGN IN

Get directions

My places



## Indian restaurants near Bangalore, Karnataka

**Serengeti**Kanyakumari Rd, Bangalore, Karnataka  
080 4000 3333 · [zomato.com](#)

Category: Indian Restaurant

18 14 reviews ·

north indian food · jungle · main course  
"Good" -**Woodys**45/1, 5th Cross, 17th Main, J P Nagar, Bangalore,  
Karnataka 560078  
080 2649 0888 · [woodlands.in](#)

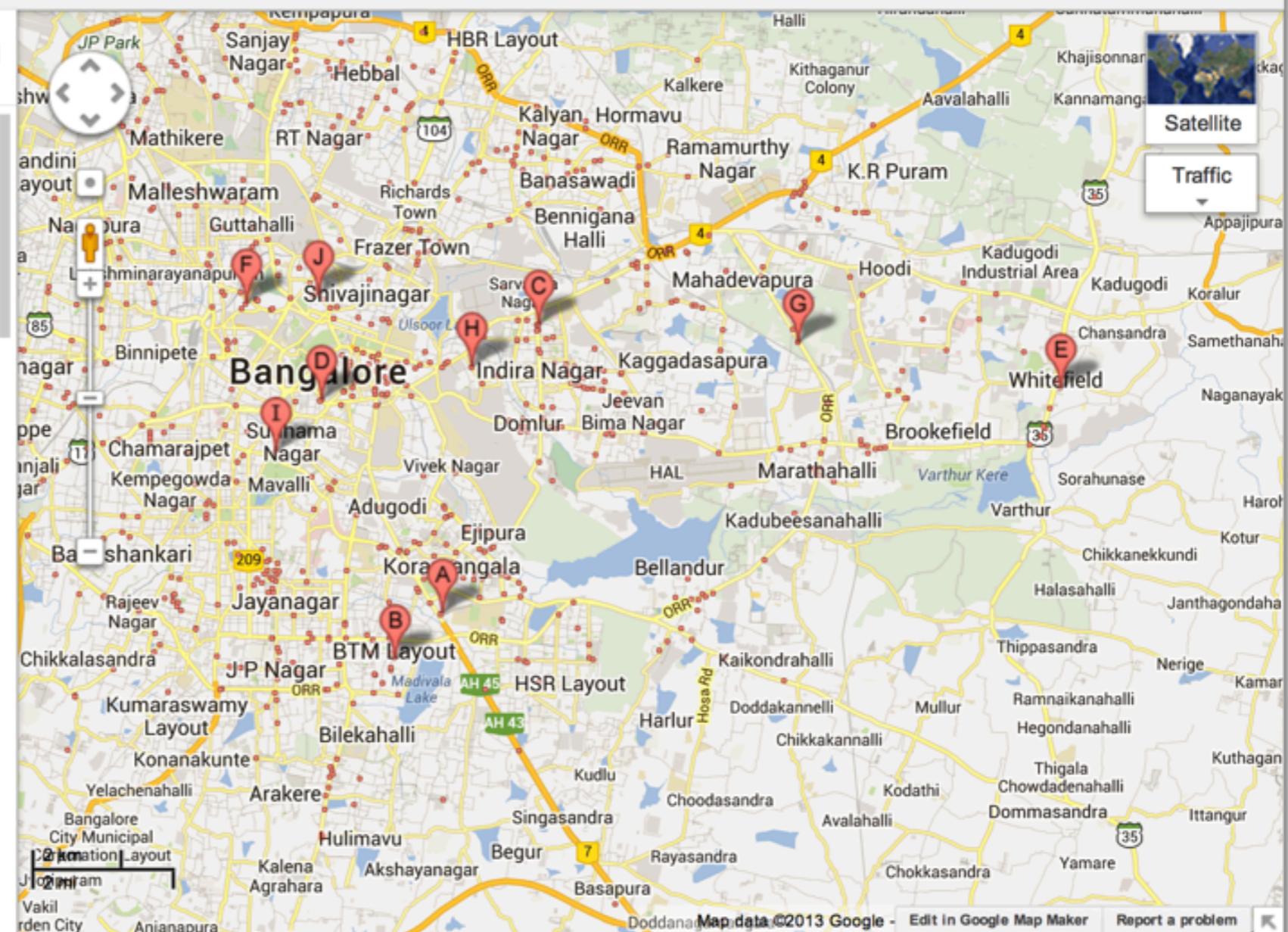
Category: Restaurants - South Indian

10 13 reviews ·

"Terrible place. We sat in a nominally air-conditioned  
place, but the AC ..." -**Southindies**Inner Ring Rd, Indira Nagar, Bangalore, Karnataka  
560038  
080 4163 6363 · [thesouthindies.com](#)

Category: South Indian Restaurant

16 21 reviews ·

"good food but small portions. holier than thou  
service. overpriced. i am not ..." -

# **What's so special about it?**

- Entities are not always directly represented
  - Recognize and disambiguate entities in text
  - Collect and aggregate information about a given entity from multiple documents and even multiple data collections
  - ~ entity linking
- More structure than document-based IR
  - Types (from some taxonomy)
  - Attributes (from some ontology)
  - Relationships to other entities (“typed links”)

# In this Part

- Focus on the ad-hoc entity retrieval task
- Mainly probabilistic models
  - Specifically, Language Models

# Basics

- Probability of an event

$$P(A)$$

- Conditional probability

$$P(A|B)$$

- Joint probability

$$P(A, B)$$

# Conditional dependence

- Independent events

$$P(A, B) = P(A) \cdot P(B)$$

$$P(A, B|C) = P(A|C) \cdot P(B|C)$$

- Conditionally dependent events

$$P(A, B) = P(A|B) \cdot P(B)$$

$$P(A, B|C) = P(A|B, C) \cdot P(B|C)$$

# Bayes' theorem

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

# Bayes' theorem

$$P(A|B) = \frac{\underbrace{P(B|A)}_{\text{Likelihood}} \cdot \underbrace{P(A)}_{\text{Prior}}}{P(B)}$$

↓  
**Posterior**

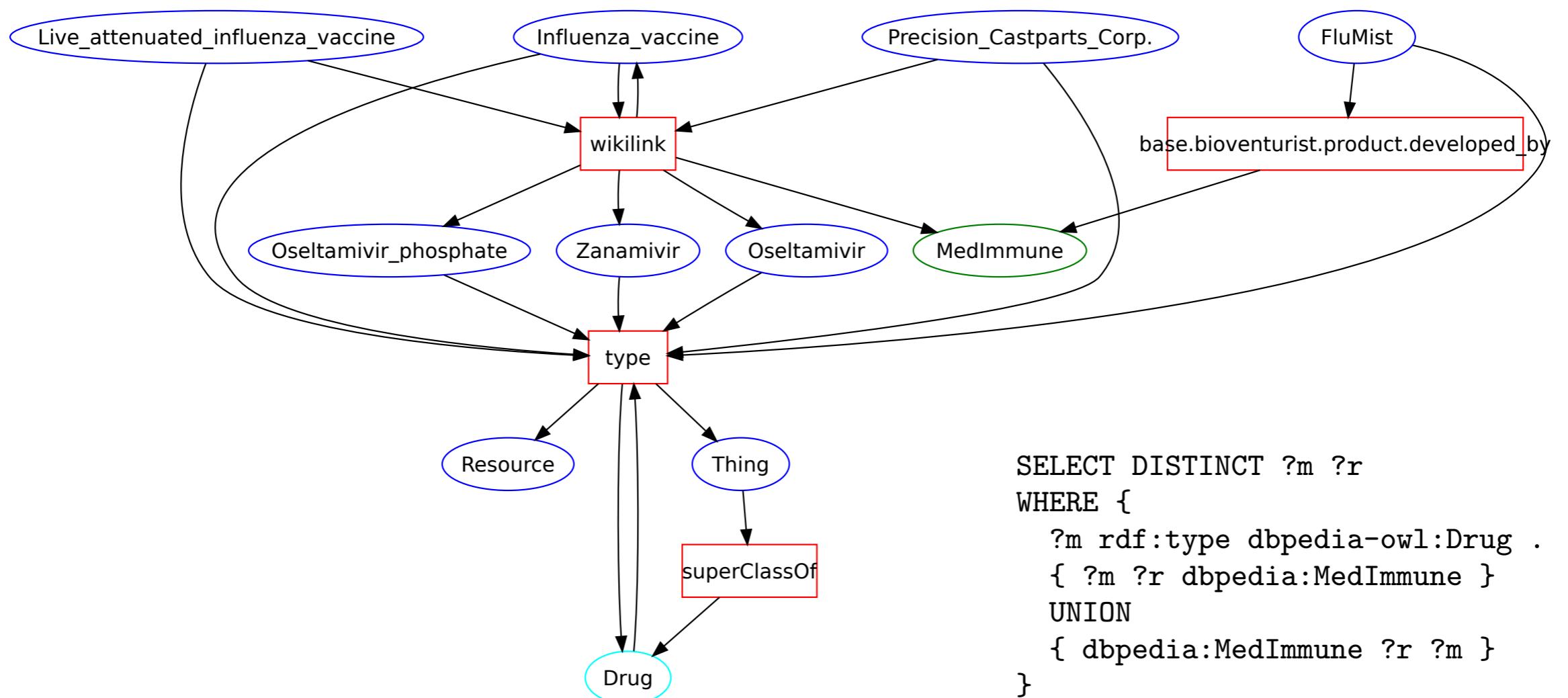
# Outline

- Part 2 – Entity Retrieval
  - introduction
  - ranking with ready-made entity descriptions
  - ranking without explicit entity representations
  - test collections
  - hands-on
  - open challenges

# Ad-hoc entity retrieval

- **Input:** unconstrained natural language query
  - “telegraphic” queries (neither well-formed nor grammatically correct sentences or questions)
- **Output:** ranked list of entities
- **Collection:** unstructured and/or semi-structured documents

# This is not...



# This is not...

*User interface*

**Title:** The Da Vinci Code  
**Author:** Dan Brown, 1964  
**Year:** 2003

*Application*

**SPARQL**

Select ?title ?year ...  
Select ?name ?year WHERE .....

**Books record**

<b>URI</b>	<a href="http://openlibrary.org/works/OL76837W">http://openlibrary.org/works/OL76837W</a>
<b>Title</b>	The Da Vinci Code
<b>Author</b>	<a href="http://viaf.org/viaf/102403515">http://viaf.org/viaf/102403515</a>
<b>Year</b>	2003

**Authors record**

<b>URI</b>	<a href="http://viaf.org/viaf/102403515">http://viaf.org/viaf/102403515</a>
<b>Name</b>	Dan Brown
<b>Year</b>	1964

# This is not...

*User interface*

**Title:** The Da Vinci Code  
**Author:** Dan Brown, 1964  
**Year:** 2003

*Application*

**SQL**

Select title, year from books  
Select name, year from authors where books.author=authors.id

*Database*

**Books record**

**ID** 1289

**Title** The Da Vinci Code

**Author** 456

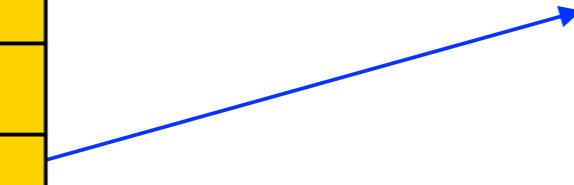
**Year** 2003

**Authors record**

**ID** 456

**Name** Dan Brown

**Year** 1964



# **Ranking with ready-made entity descriptions**

**This is not unrealistic...**

# This is not unrealistic...

Bangalore – Wikipedia, the free encyclopedia

en.wikipedia.org/wiki/Bangalore

Create account Log in

Article Talk Read Edit View history Search

## Bangalore

From Wikipedia, the free encyclopedia

Coordinates: 12°58'N 77°34'E

For other uses, see [Bangalore \(disambiguation\)](#).

**Bangalore** (or **Bengaluru** [bəŋgəluːru] (help·info)) is the capital city of the Indian state of [Karnataka](#). Located on the [Deccan Plateau](#) in the south-eastern part of Karnataka, Bangalore is India's [third most populous city](#) and [fifth-most populous urban agglomeration](#). Bangalore is well known as the hub of India's information technology sector. The city is amongst the top 10 preferred entrepreneurial locations in the world.<sup>[5]</sup> As a growing metropolitan city in a developing country, Bangalore confronts substantial pollution and other logistical and socio-economic problems.<sup>[6][7]</sup>

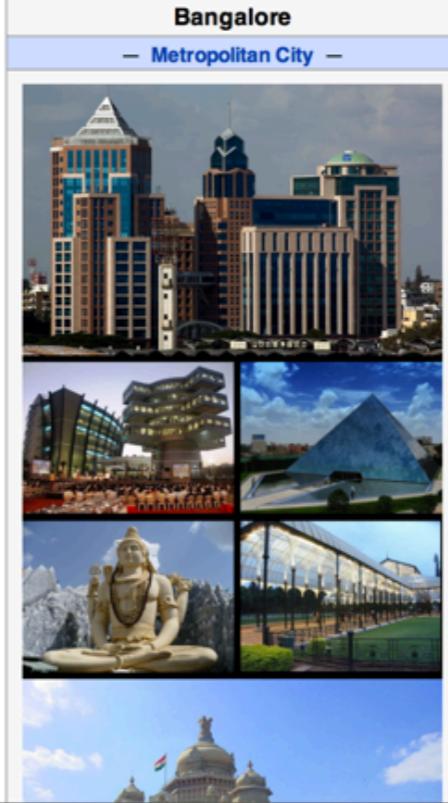
A succession of South Indian dynasties ruled the region of Bangalore until in 1537 CE, [Kempé Gowdā](#) — a feudatory ruler under the [Vijayanagara Empire](#) — established a [mud fort](#) considered to be the foundation of modern Bangalore. Following transitory occupation by the Marāthās and Mughals, the city remained under the [Mysore kingdom](#), which is now a part of the Indian state of Karnataka. Bangalore continued to be a cantonment of the British and a major city of the [Princely State of Mysore](#) which existed as a nominally sovereign entity of the [British Raj](#). Following the [independence of India](#) in 1947, Bangalore became the capital of [Mysore state](#), and remained capital when the new Indian state of Karnataka was formed in 1956. With a [Gross domestic product](#) of US\$83 billion, Bangalore is listed 4th among the top 15 cities contributing to [India's overall GDP](#).<sup>[8]</sup>

Bangalore is home to many well-recognised educational and research institutions in India. Numerous public sector [heavy industries](#), technology companies, [aerospace](#), telecommunications, and [defence organisations](#) are located in the city. Bangalore is known as the [Silicon Valley of India](#) because of its position as the nation's leading IT exporter.<sup>[9][10][11][12]</sup> A demographically diverse city, Bangalore is a major economic and cultural hub and the second-fastest growing major metropolis in India.<sup>[13]</sup>

Contents [hide]

- 1 Etymology
- 2 History
- 3 Geography
  - 3.1 Climate

**Bangalore**  
— Metropolitan City —



# This is not unrealistic...

Bangalore – Wikipedia, the free encyclopedia

W en.wikipedia.org/wiki/Bangalore

Reader Create account Log in

Man of Steel (2013) – IMDb

www.imdb.com/title/tt0770828/

Reader

WIKIPEDIA The Free Encyclopedia

Main page Contents Featured content Current events Random article Donate to Wikipedia Interaction Help About Wikipedia Community Recent changes Contact Wikipedia Toolbox Print/export Languages Afrikaans العربية বাংলা Беларуская беларускай (тарашкевіца) Български Brezhoneg Català Česky

IMDb Find Movies, TV shows, Celebrities and more... All

Movies TV News Showtimes Community IMDbPro Apps Your Watchlist

**Man of Steel (2013)**

PG-13 143 min - Action | Adventure | Fantasy - 14 June 2013 (USA)

Your rating: ★★★★★★★★★★ 8.0 /10 Ratings: 8.0/10 from 121,569 users Metascore: 55/100 Reviews: 1,298 user | 460 critic | 47 from Metacritic.com

A young itinerant worker is forced to confront his secret extraterrestrial heritage when Earth is invaded by members of his race.

Director: Zack Snyder  
Writers: David S. Goyer (screenplay), David S. Goyer (story), 3 more credits »  
Stars: Henry Cavill, Amy Adams, Michael Shannon | See full cast and crew

+ Watchlist Watch Trailer Share...

1 win & 2 nominations. See more awards »

**Videos**

on IMDb 13:01 on IMDb 00:40 Featurette Promo

**Photos**

Watch: 'Chronicle' Writer Max Landis on 'Man of Steel' and Why Superhero Movies Need Better Heroes just now | Movies.com

Warner Bros. and Legendary Pictures to Part Ways

Quick Links

Full Cast and Crew Plot Summary  
Trivia Parents Guide  
Quotes User Reviews  
Awards Release Dates  
Message Board Company Credits

Explore More

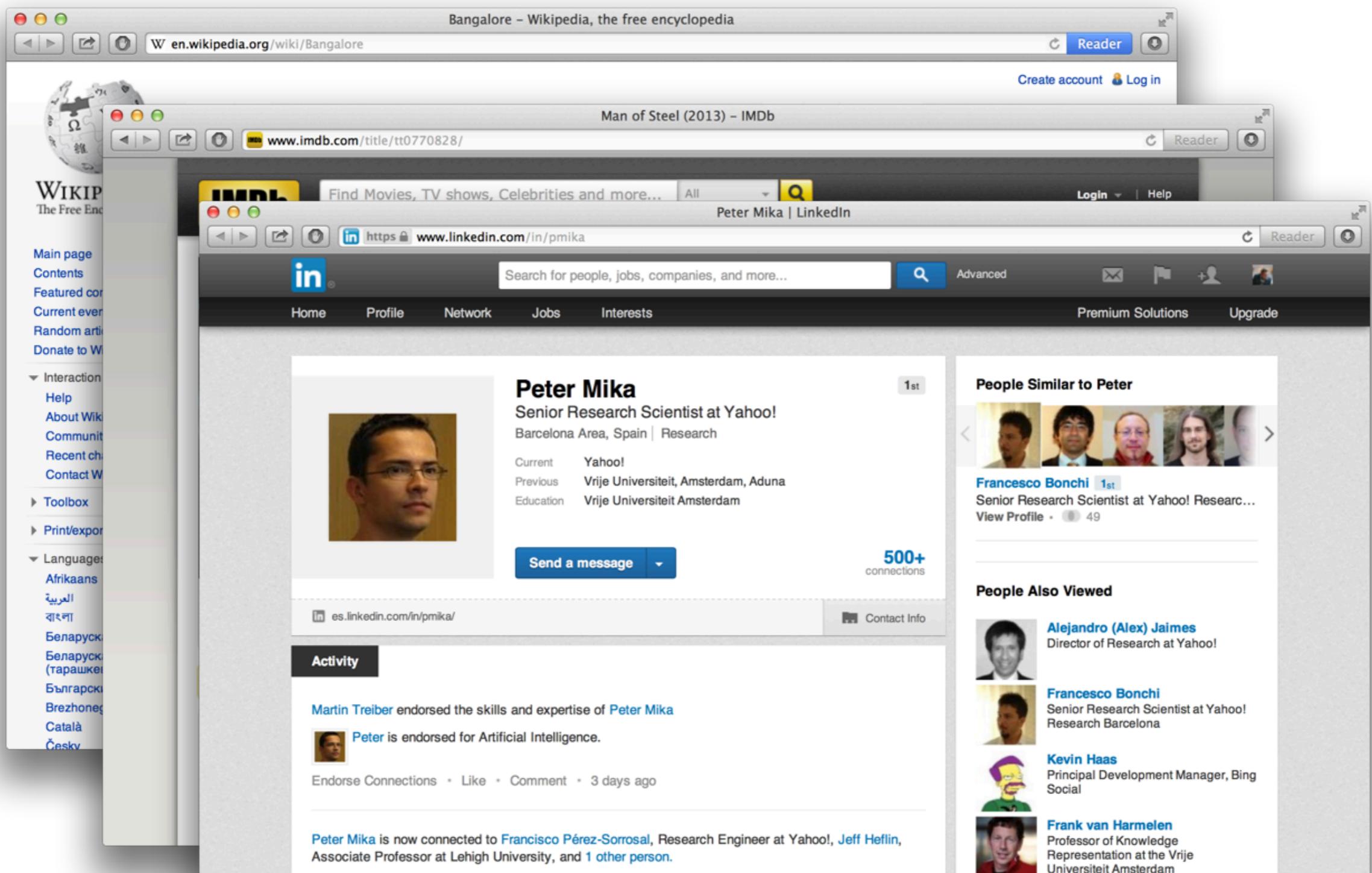
Like 85,616 people like this. Sign Up to see what your friends like.

Summer Entertainment Guide

IMDb SUMMER MOVIES AND TV EXPLORE NOW »

Related News

# This is not unrealistic...



# This is not unrealistic...



# **Document-based entity representations**

- Most entities have a “home page”
- I.e., each entity is described by a document
- Ranking entities much like ranking documents
  - Unstructured
  - Semi-structured

# Standard Language Modeling approach

- Rank documents  $d$  according to their likelihood of being relevant given a query  $q$ :  $P(d|q)$

$$P(d|q) = \frac{P(q|d)P(d)}{P(q)} \propto P(q|d)P(d)$$

# Standard Language Modeling approach

- Rank documents  $d$  according to their likelihood of being relevant given a query  $q$ :  $P(d|q)$

$$P(d|q) = \frac{P(q|d)P(d)}{P(q)} \propto P(q|d)P(d)$$

**Query likelihood**  
Probability that query  $q$  was “produced” by document  $d$

**Document prior**  
Probability of the document being relevant to *any* query

$$P(q|d) = \prod_{t \in q} P(t|\theta_d)^{n(t,q)}$$

# **Standard Language Modeling approach (2)**

$$P(q|d) = \prod_{t \in q} P(t|\theta_d)^{n(t,q)}$$

# Standard Language Modeling approach (2)

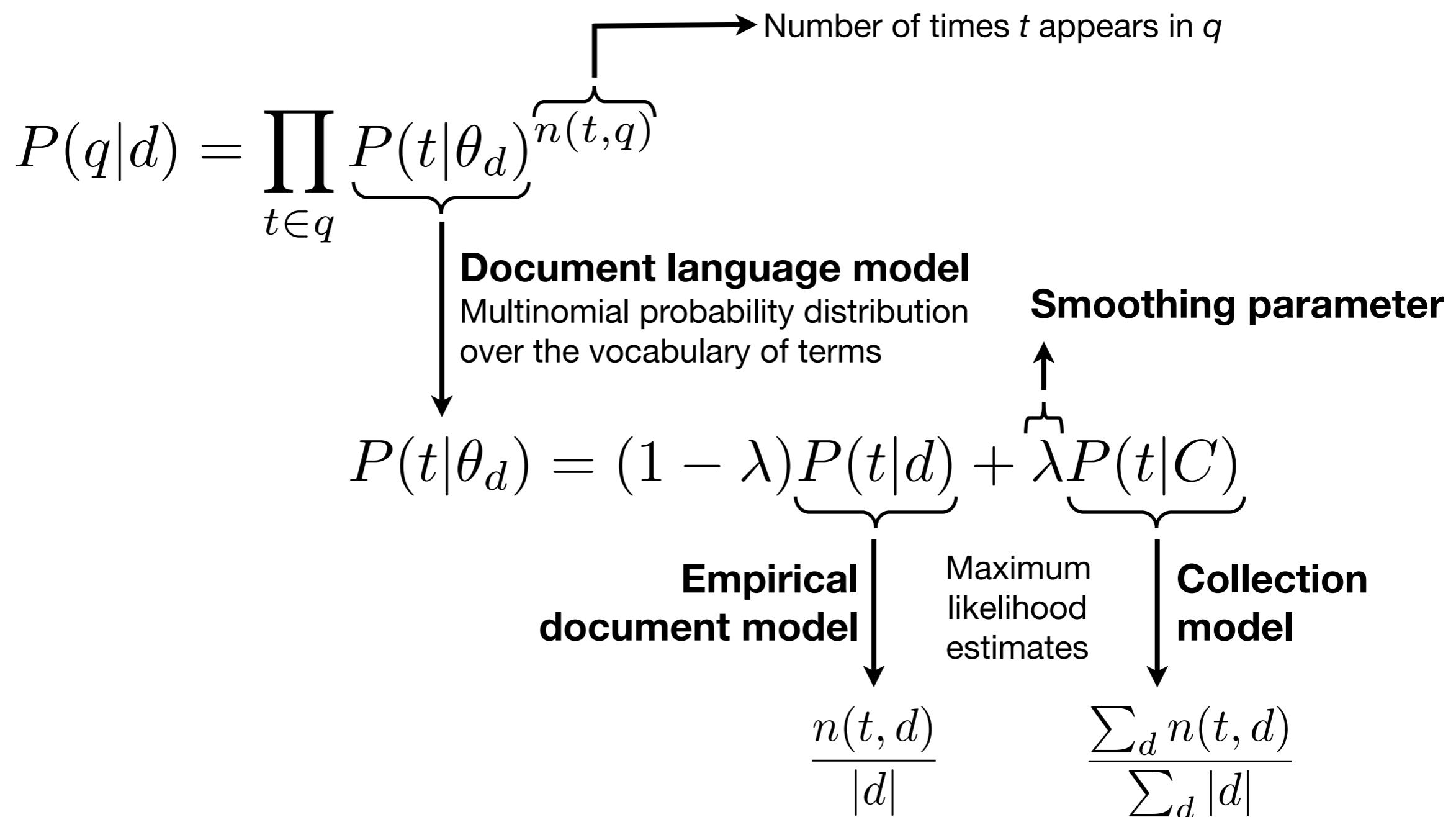
$$P(q|d) = \prod_{t \in q} \underbrace{P(t|\theta_d)}_{\text{Document language model}}^{n(t,q)}$$

Number of times  $t$  appears in  $q$

$P(t|\theta_d) = (1 - \lambda)P(t|d) + \lambda P(t|C)$

**Document language model**  
Multinomial probability distribution over the vocabulary of terms

# Standard Language Modeling approach (2)



**Here, documents=entities, so**

$$P(e|q) \propto P(e)P(q|\theta_e) = P(e) \prod_{t \in q} P(t|\theta_e)^{n(t,q)}$$

# Here, documents=entities, so

$$P(e|q) \propto P(e)P(q|\theta_e) = \underbrace{P(e)}_{\text{Entity prior}} \prod_{t \in q} \underbrace{P(t|\theta_e)^{n(t,q)}}_{\text{Entity language model}}$$

**Entity prior**  
Probability of the entity being relevant to *any* query

**Entity language model**  
Multinomial probability distribution over the vocabulary of terms

# **Semi-structured entity representation**

- Entity description documents are rarely unstructured
- Representing entities as
  - Fielded documents – the IR approach
  - Graphs – the DB/SW approach



# Audi A4

From Wikipedia, the free encyclopedia

The **Audi A4** is a line of compact executive cars produced since late 1994 by the German car manufacturer Audi, a subsidiary of the Volkswagen Group.

The A4 has been built in four generations and is based on Volkswagen's B platform. The first generation A4 succeeded the [Audi 80](#). The automaker's internal numbering treats the A4 as a continuation of the Audi 80 lineage, with the initial A4 designated as the B5-series, followed by the B6, B7, and the current B8. The B8 A4 is built on the [Volkswagen Group MLB platform](#) shared with many other Audi models and potentially one Porsche model within Volkswagen Group.<sup>[2]</sup>

The Audi A4 automobile layout consists of a longitudinally oriented engine at the front, with transaxle-type transmissions mounted at the rear of the engine. The cars are front-wheel drive, or on some models, "quattro" all-wheel drive.

The A4 is available as a saloon/sedan and estate/wagon. The second (B6) and third generations (B7) of the A4 also had a convertible version, but the B8 version of the convertible became a variant of the [Audi A5](#) instead as Audi got back into the compact executive coupé segment. The facebook fans of the Audi A4 page are more than 870,000.

**Contents** [\[show\]](#)

## Audi A4



<b>Manufacturer</b>	Audi
<b>Production</b>	1994–present
<b>Assembly</b>	Ingolstadt, Germany Changchun, China <sup>[1]</sup> Tokyo, Japan (AMA; B5 only) Jakarta, Indonesia (Garuda Mataram Motor; B5 & B8) Solomonovo, Ukraine (Eurocar; B7 only) Aurangabad, India
<b>Predecessor</b>	Audi 80
<b>Class</b>	Compact executive car (globally)
<b>Layout</b>	front-engine, front-wheel-drive front-engine, four-wheel-drive
<b>Platform</b>	Volkswagen Group B



# Audi A4

From Wikipedia, the free encyclopedia

The Audi A4 is a line of compact executive cars produced since late 1994 by the German car manufacturer Audi, a subsidiary of the Volkswagen Group.

The A4 has been built in four generations and is based on Volkswagen's B platform. The first generation A4 succeeded the Audi 80. The automaker's internal numbering treats the A4 as a continuation of the Audi 80 lineage, with the initial A4 designated as the B5-series, followed by the B6, B7, and the current B8. The B8 A4 is built on the Volkswagen Group MLB platform shared with many other Audi models and potentially one Porsche model within Volkswagen Group.<sup>[2]</sup>

**Audi A4**



Manufacturer Audi

## dbpedia:Audi\_A4

**foaf:name**

Audi A4

**rdfs:label**

Audi A4

**rdfs:comment**

The Audi A4 is a compact executive car produced since late 1994 by the German car manufacturer Audi, a subsidiary of the Volkswagen Group. The A4 has been built [...]

**1994**

**2001**

**2005**

**2008**

**rdf:type**

[dbpedia-owl:MeanOfTransportation](#)

[dbpedia-owl:Automobile](#)

[dbpedia:Audi](#)

[dbpedia:Compact\\_executive\\_car](#)

[freebase:Audi\\_A4](#)

[dbpedia:Audi\\_A5](#)

[dbpedia:Cadillac\\_BLS](#)

**dbpedia-owl:manufacturer**

**dbpedia-owl:class**

**owl:sameAs**

is [dbpedia-owl:predecessor](#) of

is [dbpprop:similar](#) of

# Mixture of Language Models

[Ogilvie & Callan, 2003]

- Build a separate language model for each field
- Take a linear combination of them

$$P(t|\theta_d) = \sum_{j=1}^m \mu_j P(t|\theta_{d_j})$$

**Field weights**

$$\sum_{j=1}^m \mu_j = 1$$

**Field language model**

Smoothed with a collection model built from all document representations of the same type in the collection

# Setting field weights

- Heuristically
  - Proportional to the length of text content in that field, to the field's individual performance, etc.
- Empirically (using training queries)
- Problems
  - Number of possible fields is huge
    - It is not possible to optimise their weights directly
- Entities are sparse w.r.t. different fields
  - Most entities have only a handful of predicates

# Predicate folding

- **Idea:** reduce the number of fields by grouping them together
- Grouping based on (BM25F and)
  - type **[Pérez-Agüera et al. 2010]**
  - manually determined importance **[Blanco et al. 2011]**

# Hierarchical Entity Model

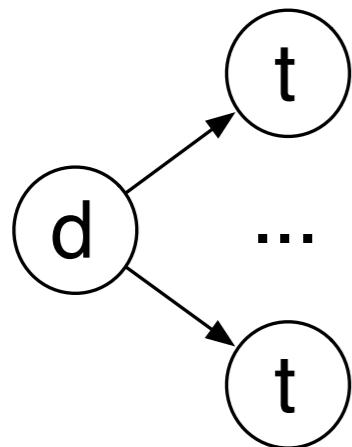
[Neumayer et al. 2012]

- Organize fields into a 2-level hierarchy
  - Field types (4) on the top level
  - Individual fields of that type on the bottom level
- Estimate field weights
  - Using training data for field types
  - Using heuristics for bottom-level types

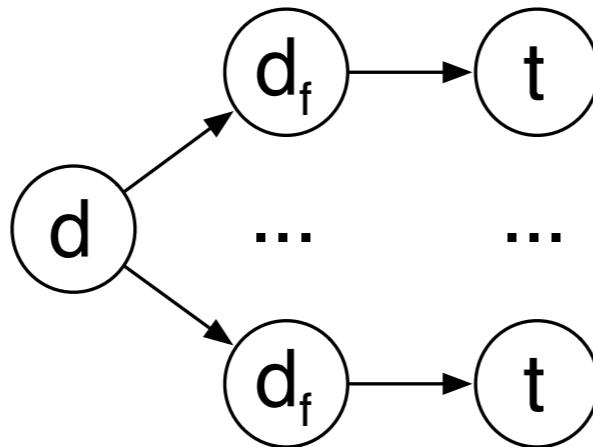
# Two-level hierarchy

<b>Name</b>	{	foaf:name rdfs:label rdfs:comment	Audi A4 Audi A4 The Audi A4 is a compact executive car produced since late 1994 by the German car manufacturer Audi, a subsidiary of the Volkswagen Group. The A4 has been built [...]
<b>Attributes</b>	{	dbpprop:production	1994 2001 2005 2008
<b>Out-relations</b>	{	rdf:type dbpedia-owl:manufacturer dbpedia-owl:class owl:sameAs	<a href="#">dbpedia-owl:MeanOfTransportation</a> <a href="#">dbpedia-owl:Automobile</a> <a href="#">dbpedia:Audi</a> <a href="#">dbpedia:Compact_executive_car</a> <a href="#">freebase:Audi_A4</a>
<b>In-relations</b>	{	is <a href="#">dbpedia-owl:predecessor</a> of is <a href="#">dbpprop:similar</a> of	<a href="#">dbpedia:Audi_A5</a> <a href="#">dbpedia:Cadillac_BLS</a>

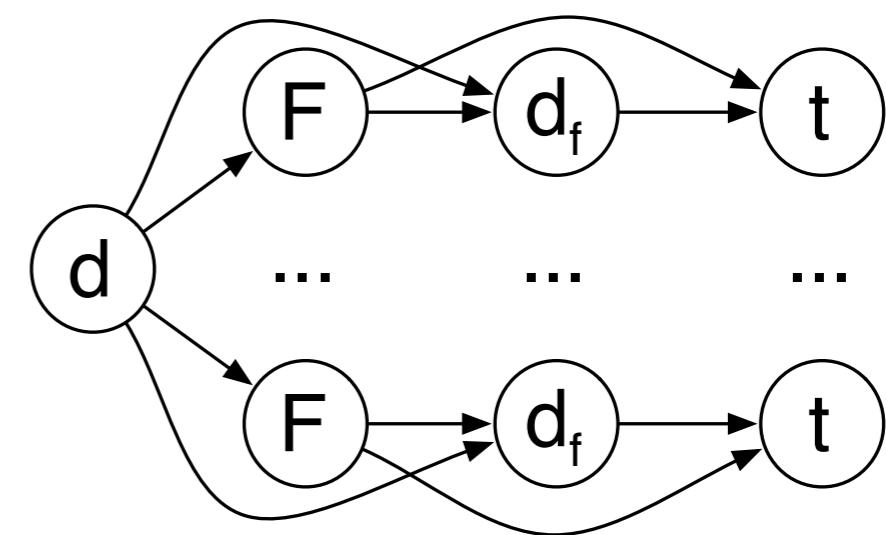
# Comparison of models



**Unstructured  
document model**



**Fielded  
document model**



**Hierarchical  
document model**

# Probabilistic Retrieval Model for Semistructured data

[Kim et al. 2009]

- Extension to the Mixture of Language Models
- Find which document field each query term may be associated with

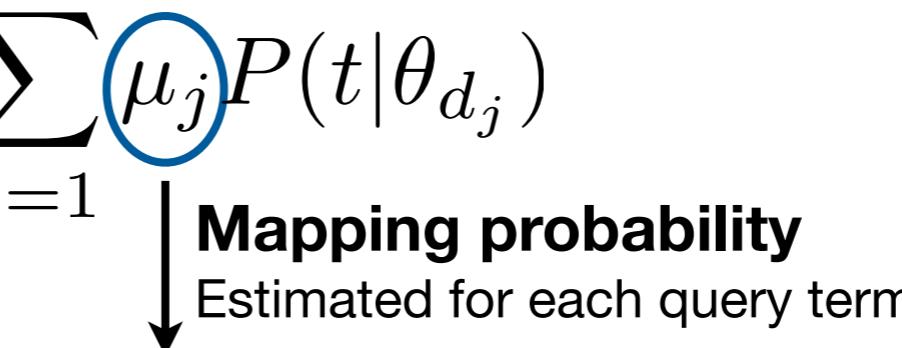
$$P(t|\theta_d) = \sum_{j=1}^m \mu_j P(t|\theta_{d_j})$$

# Probabilistic Retrieval Model for Semistructured data

[Kim et al. 2009]

- Extension to the Mixture of Language Models
- Find which document field each query term may be associated with

$$P(t|\theta_d) = \sum_{j=1}^m \mu_j P(t|\theta_{d_j})$$

  
**Mapping probability**  
Estimated for each query term

$$P(t|\theta_d) = \sum_{j=1}^m \overbrace{P(d_j|t)} P(t|\theta_{d_j})$$

# **Estimating the mapping probability**

$$P(d_j|t) = \frac{P(t|d_j)P(d_j)}{P(t)}$$

# Estimating the mapping probability

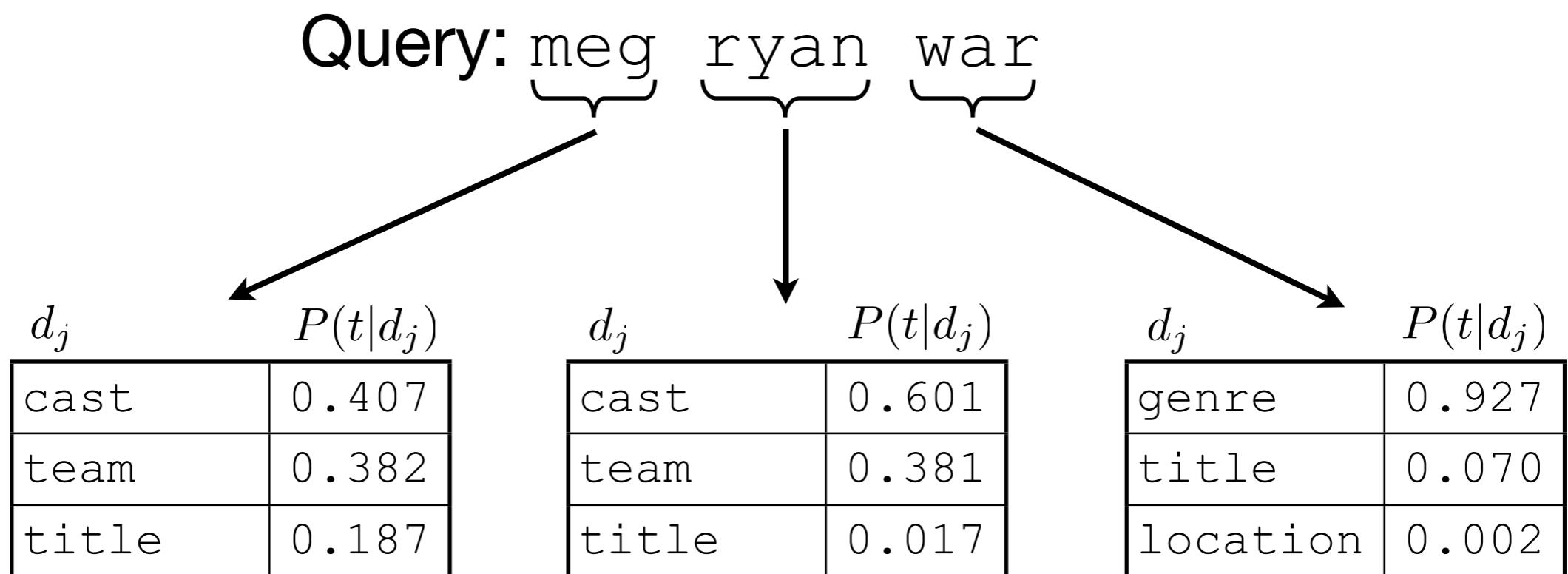
$$P(t|C_j) = \frac{\sum_d n(t, d_j)}{\sum_d |d_j|}$$

**Term likelihood**  
Probability of a query term occurring in a given field type

**Prior field probability**  
Probability of mapping the query term to this field before observing collection statistics

$$P(d_j|t) = \frac{P(t|d_j)P(d_j)}{P(t)}$$
$$\sum_{d_k} P(t|d_k)P(d_k)$$

# Example



# The usual suspects from document retrieval...

- Priors
  - HITS, PageRank
  - Document link indegree **[Kamps & Koolen 2008]**
- Pseudo relevance feedback
  - Document-centric vs. entity-centric **[Macdonald & Ounis 2007; Serdyukov et al. 2007]**
    - sampling expansion terms from top ranked documents and/or (profiles of) top ranked candidates
  - Field-based **[Kim & Croft 2011]**

# **So far...**

- Ranking (fielded) documents...
- What is special about entities?
  - Type(s)
  - Relationships with other entities

# Entity types

`rdf:type`

`dbpedia-owl:MeanOfTransportation`  
`dbpedia-owl:Automobile`

Categories: Audi vehicles | Compact executive cars | Euro NCAP large family cars | Sedans | Station wagons | Convertibles  
| Vehicles with CVT transmission | All-wheel-drive vehicles | Front-wheel-drive vehicles | Vehicles introduced in 1994  
| 1990s automobiles | 2000s automobiles | 2010s automobiles | Hybrid electric cars

Freebase Find... Browse Query Help Sign In or Sign Up English ▾

Audi A4 en

Created by metaweb on 10/22/2006

`id: /guid/9202a8c04000641f800000000305a7c mid: /m/030qmx notable type: /automotive/model notable for: /automotive/model on the web: W wikipedia.org`

The Audi A4 is a line of compact executive cars produced since late 1994 by the German car manufacturer Audi, a subsidiary of the Volkswagen Group. The A4 has been built in four generations and is based on Volkswagen's B platform. The first generation A4 succeeded the Audi 80. The automaker's internal numbering treats the A4 as a continuation of the Audi 80 lineage, with the initial A4 designated as the B5-series, followed by the B6, B7, and the current B8. The B8 A4 is built on the Volkswagen Group MLB platform shared with many other Audi models and potentially one Porsche model within Volkswagen Group. The Audi A4 automobile layout consists of a longitudinally oriented engine at the front, with transaxle-type transmissions mounted at the rear of the engine. The cars are front-wheel drive, or on some models, "quattro" all-wheel drive. The A4 is available as a saloon/sedan and estate/wagon. The second and third generations of the A4 also had a convertible version, but the B8 version of the convertible became a variant of the Audi A5 instead as Audi got back into the compact executive coupé segment. Wikipedia [-]

Properties I18n Keys Links

View and edit specific domains, types, or properties

Filter options:  Show all domains and properties

Common /common

•Topic /common/topic

Also known as /common/topic/alias

Freebase Commons

Types:

Common

Topic

Automotive

Automobile Model

# **Using target types**

**Assuming they have been identified...**

- Constraining results
  - Soft/hard filtering
  - Different ways to measure type similarity (between target types and the types associated with the entity)
    - Set-based
    - Content-based
    - Lexical similarity of type labels
- Query expansion
  - Adding terms from type names to the query
- Entity expansion
  - Categories as a separate metadata field

# Modeling terms and categories

[Balog et al. 2011]

$$P(e|q) \propto P(q|e)P(e)$$
$$P(q|e) = (1 - \lambda) \underbrace{P(\theta_q^T | \theta_e^T)} + \lambda \underbrace{P(\theta_q^C | \theta_e^C)}$$

# Modeling terms and categories

[Balog et al. 2011]

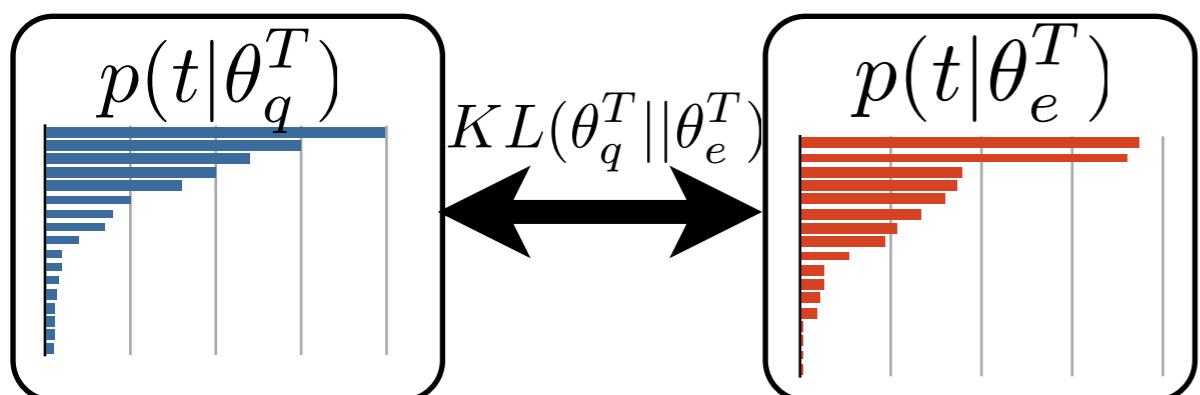
$$P(e|q) \propto P(q|e)P(e)$$

$$P(q|e) = (1 - \lambda) \underbrace{P(\theta_q^T | \theta_e^T)} + \lambda \underbrace{P(\theta_q^C | \theta_e^C)}$$

Term-based representation

Query model

Entity model



# Modeling terms and categories

[Balog et al. 2011]

$$P(e|q) \propto P(q|e)P(e)$$

$$P(q|e) = (1 - \lambda) \underbrace{P(\theta_q^T | \theta_e^T)}_{\text{Term-based representation}} + \lambda \underbrace{P(\theta_q^C | \theta_e^C)}_{\text{Category-based representation}}$$

Term-based representation

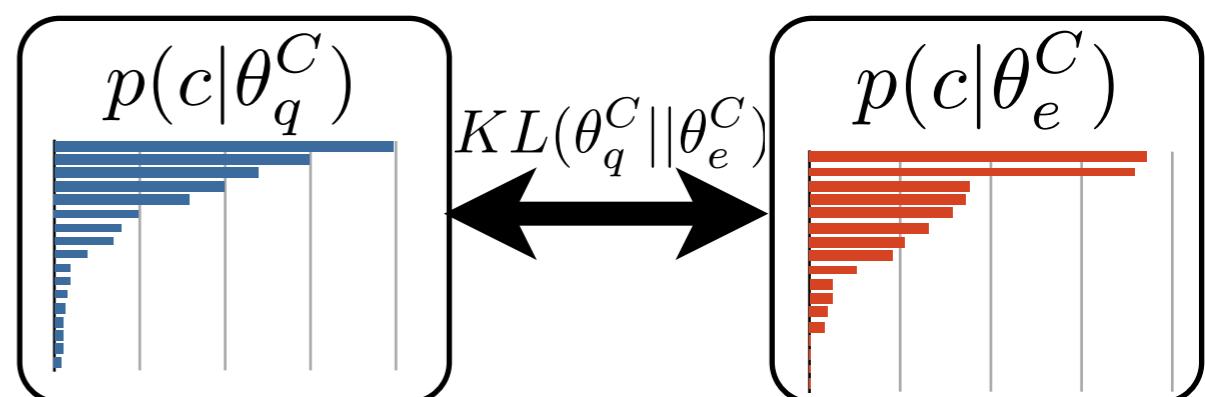
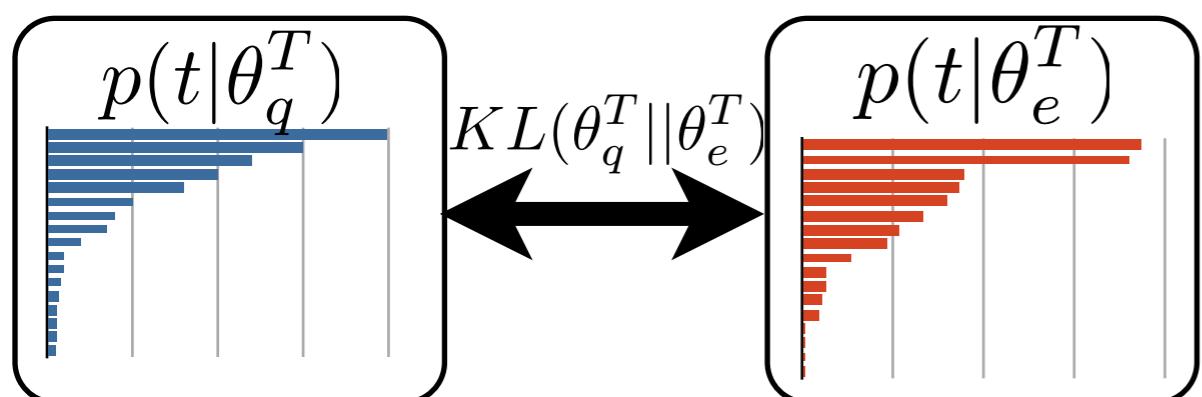
Query model

Entity model

Category-based representation

Query model

Entity model



# Identifying target types

- Types of top ranked entities **[Vallet & Zaragoza 2008]**
- Direct term-based vs. indirect entity-based representations **[Balog & Neumayer 2012]**
- Hierarchical case is difficult... **[Sawant & Chakrabarti 2013]**

# Expanding target types

- Pseudo relevance feedback
- Based on hierarchical structure
- Using lexical similarity of type labels

# Outline

- Part 2 – Entity Retrieval
  - introduction
  - ranking with ready-made entity descriptions
  - ranking without explicit entity representations
  - test collections
  - hands-on
  - open challenges

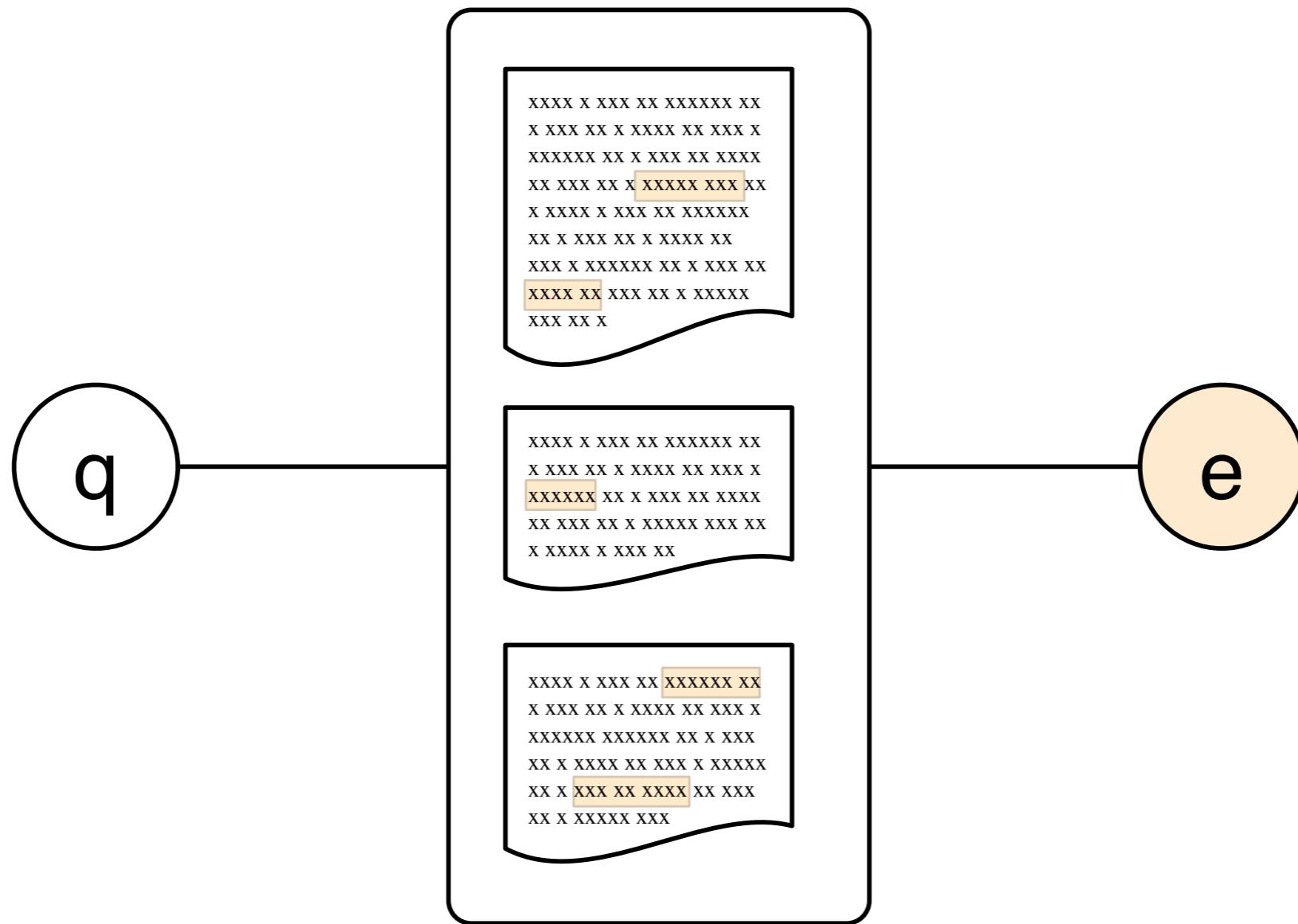
# **Ranking without explicit entity representations**

# Scenario

- Entity descriptions are not readily available
- Entity occurrences are annotated
  - manually
  - automatically (~entity linking)

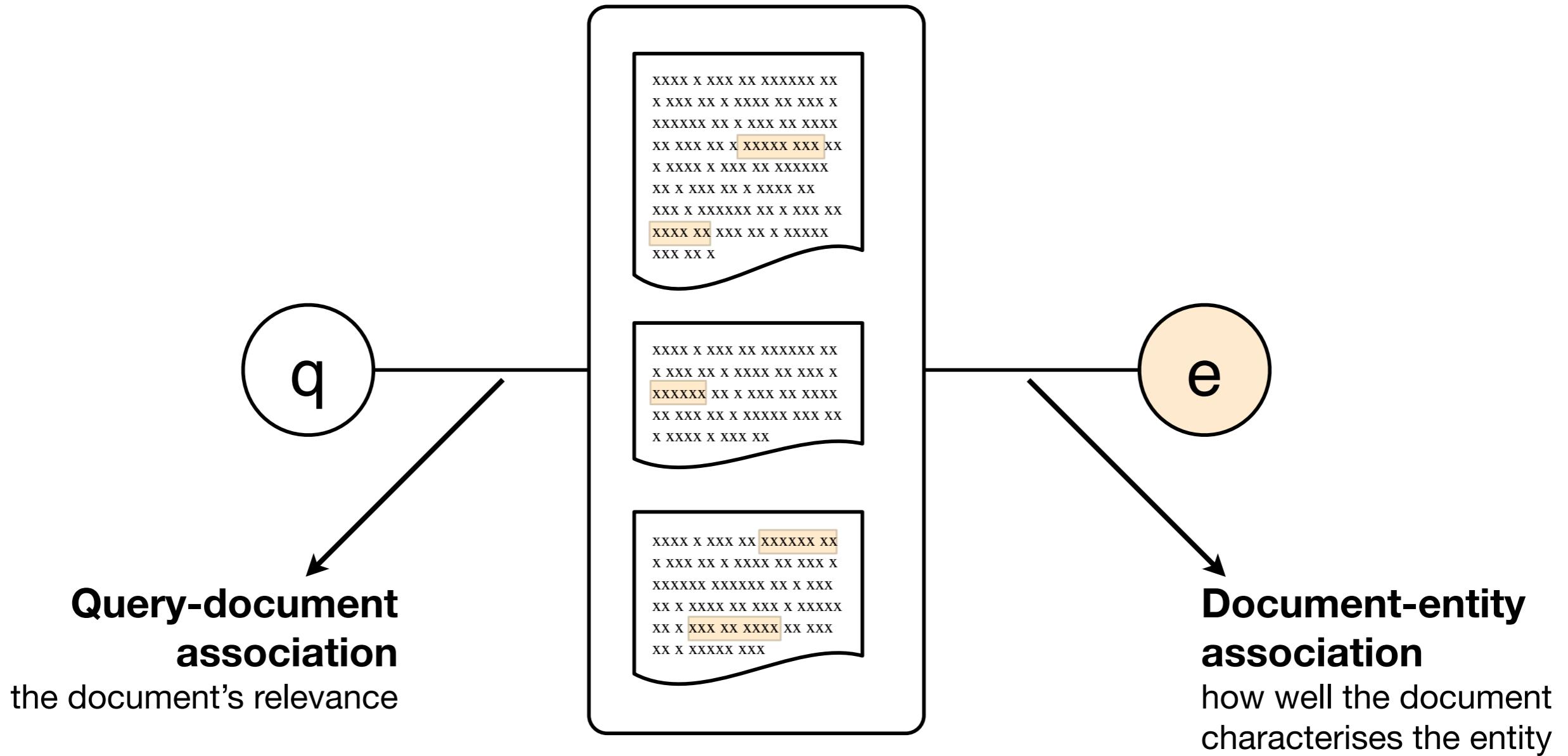
# The basic idea

Use documents to go from queries to entities



# The basic idea

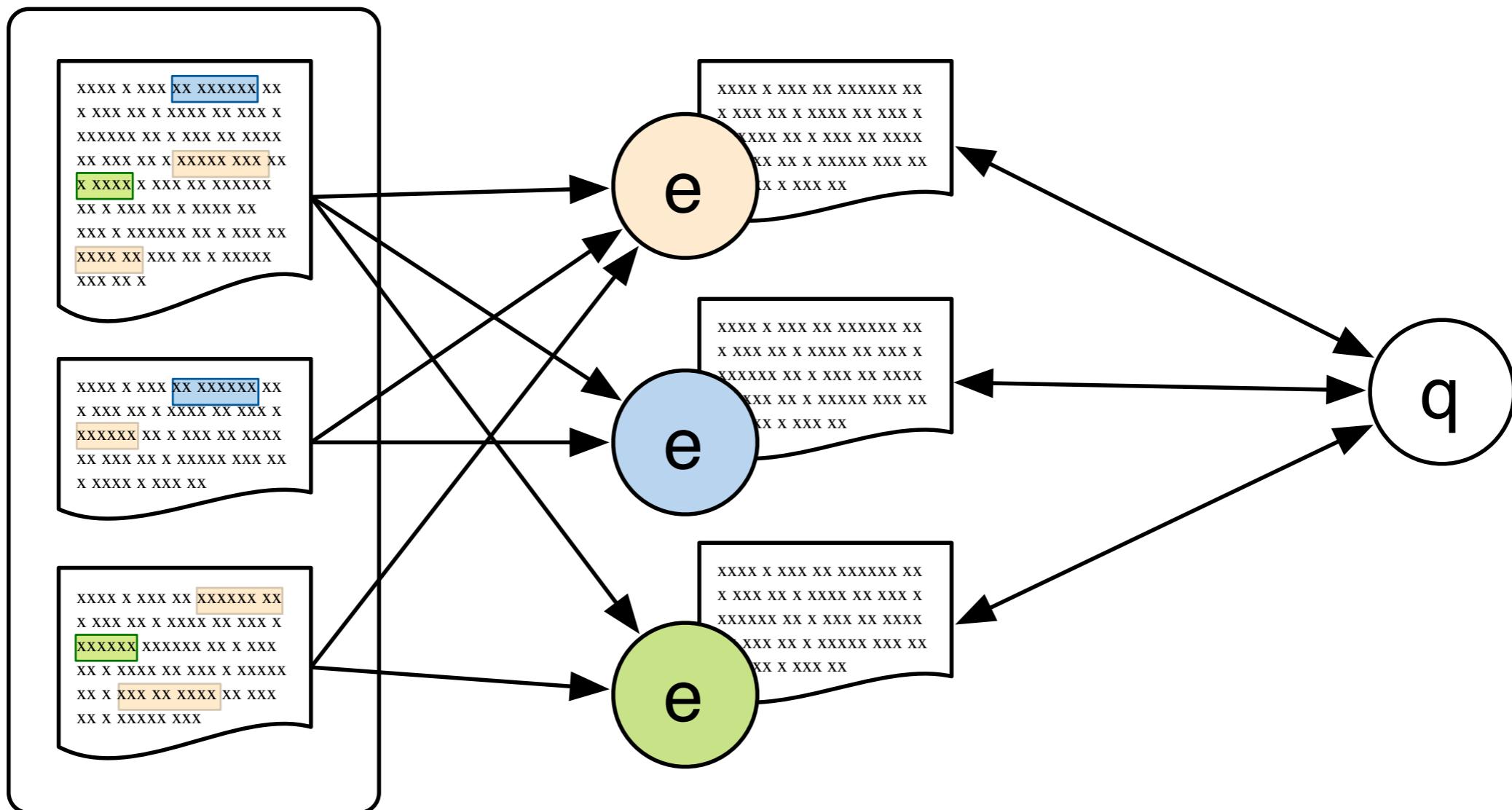
Use documents to go from queries to entities



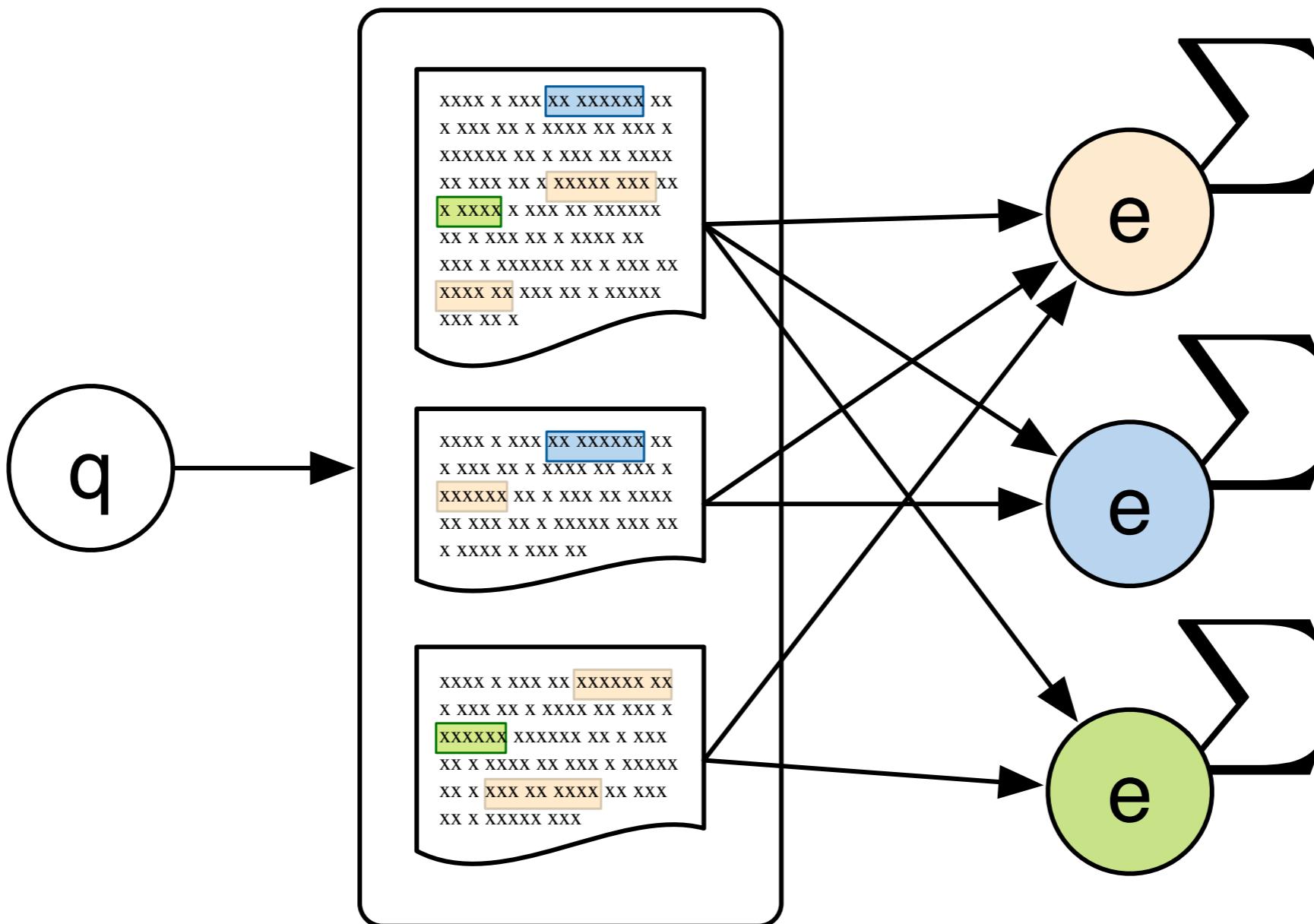
# Two principal approaches

- **Profile-based** methods
  - Create a textual profile for entities, then rank them (by adapting document retrieval techniques)
- **Document-based** methods
  - Indirect representation based on mentions identified in documents
  - First ranking documents (or snippets) and then aggregating evidence for associated entities

# Profile-based methods



# Document-based methods



# **Many possibilities in terms of modeling**

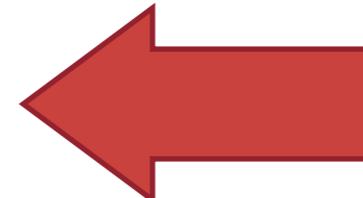
- Generative probabilistic models
- Discriminative probabilistic models
- Voting models
- Graph-based models

# Generative probabilistic models

- Candidate generation models ( $P(e|q)$ )
  - Two-stage language model
- Topic generation models ( $P(q|e)$ )
  - Candidate model, a.k.a. Model 1
  - Document model, a.k.a. Model 2
  - Proximity-based variations
- Both families of models can be derived from the Probability Ranking Principle **[Fang & Zhai 2007]**

# Generative probabilistic models

- Candidate generation models ( $P(e|q)$ )
  - Two-stage language model
- Topic generation models ( $P(q|e)$ )
  - Candidate model, a.k.a. Model 1
  - Document model, a.k.a. Model 2
  - Proximity-based variations
- Both families of models can be derived from the Probability Ranking Principle **[Fang & Zhai 2007]**



# Candidate models (“Model 1”)

[Balog et al. 2006]

$$P(q|\theta_e) = \prod_{t \in q} P(t|\theta_e)^{n(t,q)}$$

# Candidate models (“Model 1”)

[Balog et al. 2006]

$$P(q|\theta_e) = \prod_{t \in q} \underbrace{P(t|\theta_e)}_{\text{Smoothing}}^{n(t,q)}$$

↓

**Smoothing**  
With collection-wide background model

$$(1 - \lambda)P(t|e) + \lambda P(t)$$

# Candidate models (“Model 1”)

[Balog et al. 2006]

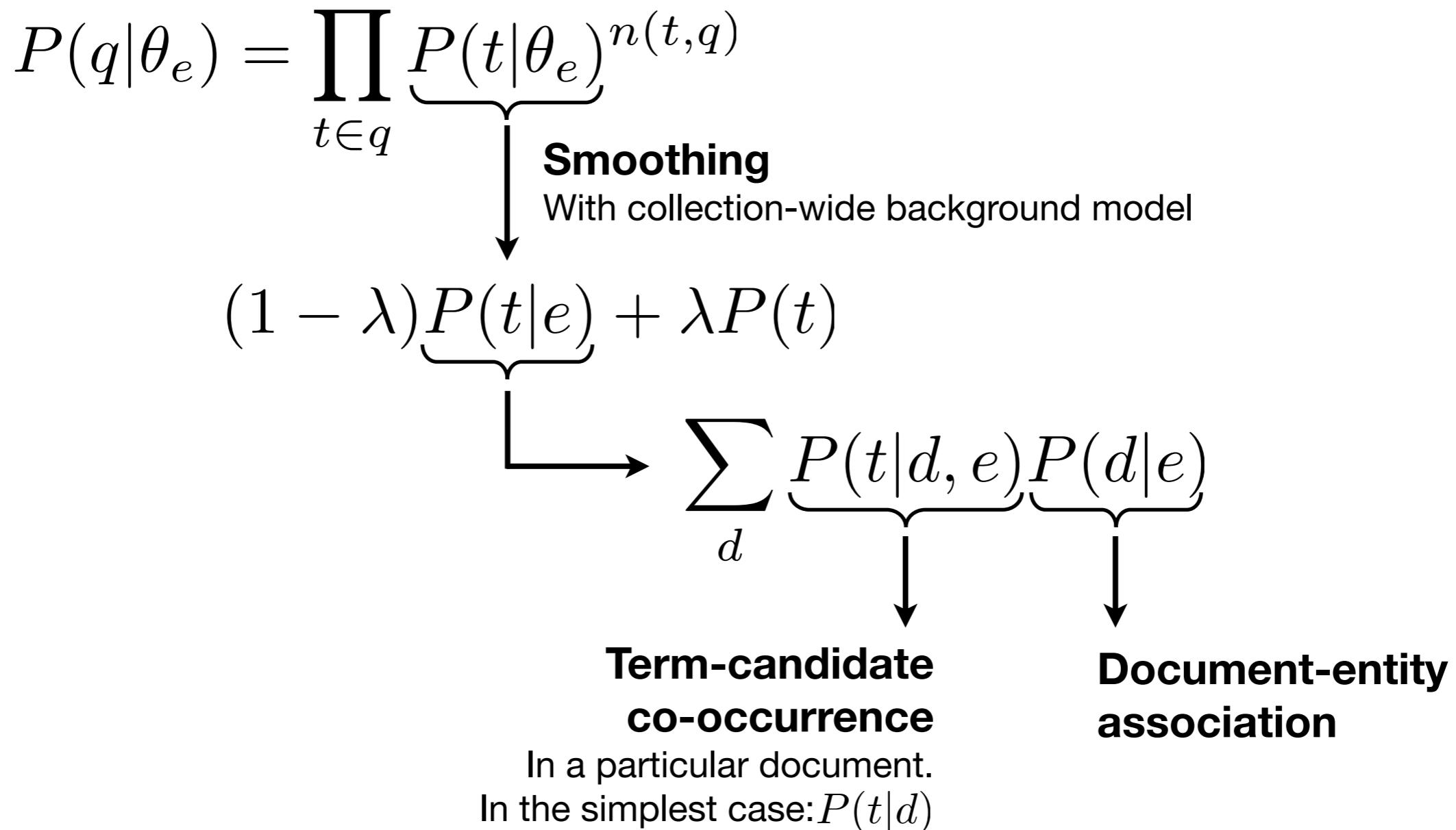
$$P(q|\theta_e) = \prod_{t \in q} P(t|\theta_e)^{n(t,q)}$$

**Smoothing**  
With collection-wide background model

$$(1 - \lambda) \underbrace{P(t|e)}_{\longrightarrow} + \lambda P(t)$$
$$\sum_d P(t|d, e) P(d|e)$$

# Candidate models (“Model 1”)

[Balog et al. 2006]



# Document models (“Model 2”)

[Balog et al. 2006]

$$P(q|e) = \sum_d P(q|d, e)P(d|e)$$

# Document models (“Model 2”)

[Balog et al. 2006]

$$P(q|e) = \sum_d P(q|d, e) P(d|e)$$

**Document relevance**  
How well document  $d$  supports the claim that  $e$  is relevant to  $q$

**Document-entity association**

$$\prod_{t \in q} P(t|d, e)^{n(t,q)}$$

# Document models (“Model 2”)

[Balog et al. 2006]

$$P(q|e) = \sum_d P(q|d, e) P(d|e)$$

**Document relevance**  
How well document  $d$  supports the claim that  $e$  is relevant to  $q$

**Document-entity association**

$$\prod_{t \in q} \underbrace{P(t|d, e)}_{\text{Simplifying assumption}}^{n(t,q)}$$

( $t$  and  $e$  are conditionally independent given  $d$ )

$$P(t|\theta_d)$$

# Document-entity associations

- Boolean (or set-based) approach
- Weighted by the confidence in entity linking
- Consider other entities mentioned in the document

# Proximity-based variations

- So far, conditional independence assumption between candidates and terms when computing the probability  $P(t|d,e)$
- Relationship between terms and entities that in the same document is ignored
  - Entity is equally strongly associated with everything discussed in that document
- Let's capture the dependence between entities and terms
  - Use their distance in the document

# Using proximity kernels

[Petkova & Croft 2007]

$$P(t|d, e) = \frac{1}{Z} \sum_{i=1}^N \delta_d(i, t) k(t, e)$$

# Using proximity kernels

[Petkova & Croft 2007]

$$P(t|d, e) = \frac{1}{Z} \sum_{i=1}^N \underbrace{\delta_d(i, t)}_{\text{Indicator function}} \underbrace{k(t, e)}_{\text{Proximity-based kernel}}$$

**Normalizing  
constant**

**Indicator function**

1 if the term at position i is t,  
0 otherwise

**Proximity-based kernel**

- constant function
- triangle kernel
- Gaussian kernel
- step function

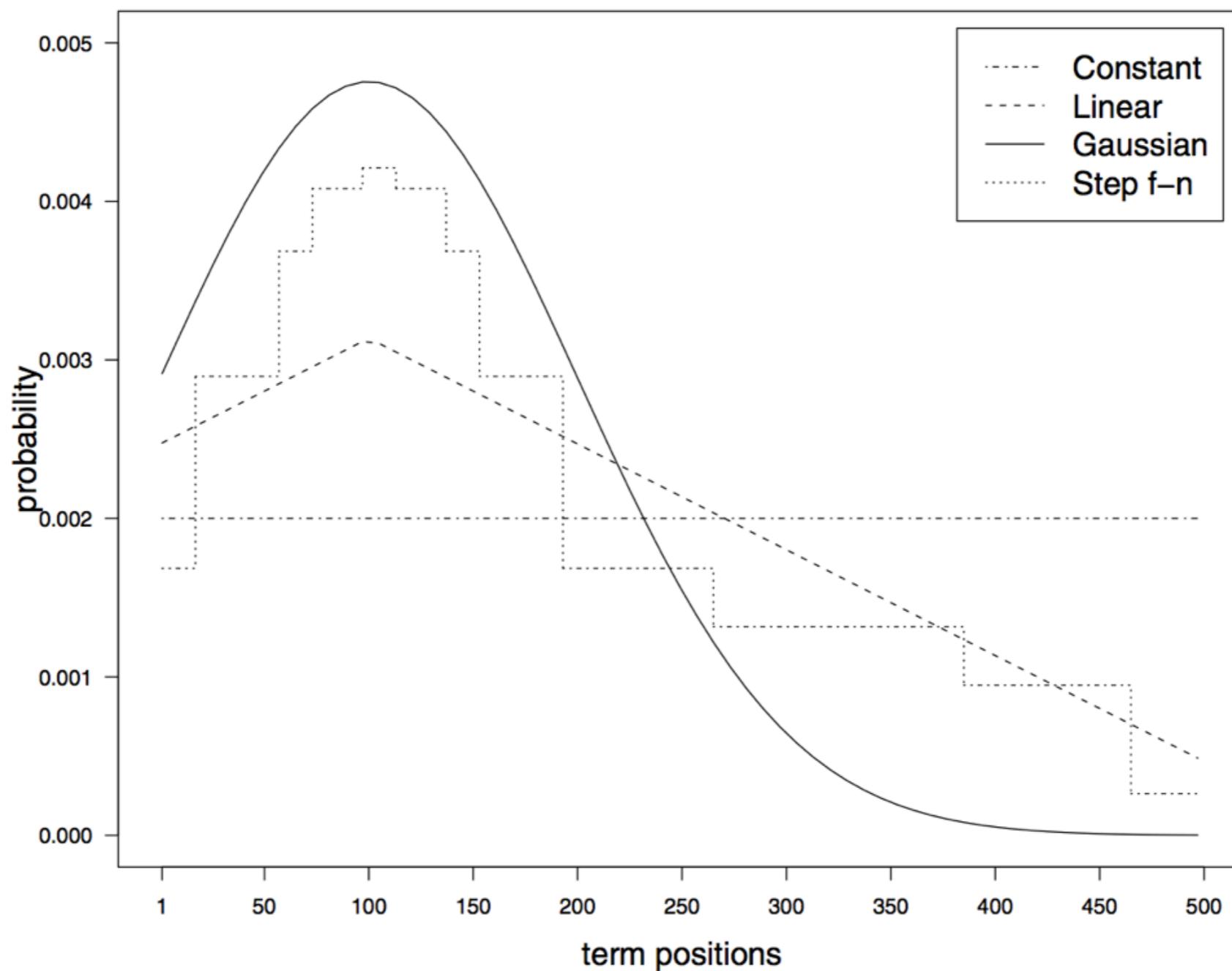


Figure taken from D. Petkova and W.B. Croft. **Proximity-based document representation for named entity retrieval.** CIKM'07.

# **Many possibilities in terms of modeling**

- Generative probabilistic models
- Discriminative probabilistic models
- Voting models
- Graph-based models

# Discriminative models

- Vs. generative models:
  - Fewer assumptions (e.g., term independence)
  - “Let the data speak”
    - Sufficient amounts of training data required
  - Incorporating more document features, multiple signals for document-entity associations
  - Estimating  $P(r=1|e,q)$  directly (instead of  $P(e,q|r=1)$ )
  - Optimization can get trapped in a local maximum/minimum

# **Arithmetic Mean Discriminative (AMD) model**

**[Yang et al. 2010]**

$$P_{\theta}(r = 1|e, q) = \sum_d P(r_1 = 1|q, d)P(r_2 = 1|e, d)P(d)$$

# Arithmetic Mean Discriminative (AMD) model

[Yang et al. 2010]

$$P_{\theta}(r = 1|e, q) = \sum_d \underbrace{P(r_1 = 1|q, d)}_{\text{Query-document relevance}} \underbrace{P(r_2 = 1|e, d)}_{\text{Document-entity relevance}} \underbrace{P(d)}_{\text{Document prior}}$$

# Arithmetic Mean Discriminative (AMD) model

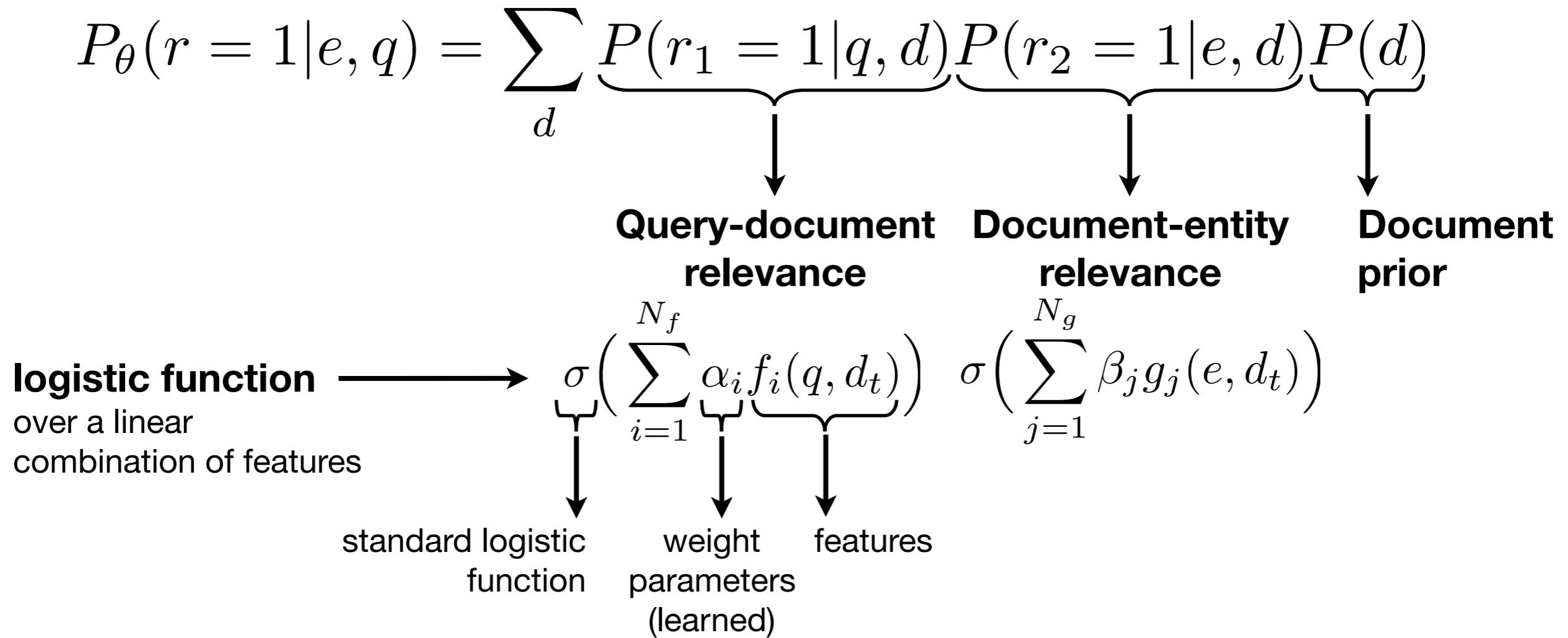
[Yang et al. 2010]

$$P_{\theta}(r = 1|e, q) = \sum_d \underbrace{P(r_1 = 1|q, d)}_{\text{Query-document relevance}} \underbrace{P(r_2 = 1|e, d)}_{\text{Document-entity relevance}} \underbrace{P(d)}_{\text{Document prior}}$$

**logistic function**  $\longrightarrow \sigma\left(\sum_{i=1}^{N_f} \alpha_i f_i(q, d_t)\right) \quad \sigma\left(\sum_{j=1}^{N_g} \beta_j g_j(e, d_t)\right)$   
over a linear combination of features

# Arithmetic Mean Discriminative (AMD) model

[Yang et al. 2010]



# Learning to rank && entity retrieval

- Pointwise
  - AMD, GMD **[Yang et al. 2010]**
  - Multilayer perceptrons, logistic regression **[Sorg & Cimiano 2011]**
  - Additive Groves **[Moreira et al. 2011]**
- Pairwise
  - Ranking SVM **[Yang et al. 2009]**
  - RankBoost, RankNet **[Moreira et al. 2011]**
- Listwise
  - AdaRank, Coordinate Ascent **[Moreira et al. 2011]**

# Voting models

[Macdonald & Ounis 2006]

- Inspired by techniques from data fusion
  - Combining evidence from different sources
- Documents ranked w.r.t. the query are seen as “votes” for the entity

# Voting models

**Many different variants, including...**

- Votes

- Number of documents mentioning the entity

$$Score(e, q) = |M(e) \cap R(q)|$$

- Reciprocal Rank

- Sum of inverse ranks of documents

$$Score(e, q) = \sum_{\{M(e) \cap R(q)\}} \frac{1}{rank(d, q)}$$

- CombSUM

- Sum of scores of documents

$$Score(e, q) = |\{M(e) \cap R(q)\}| \sum_{\{M(e) \cap R(q)\}} s(d, q)$$

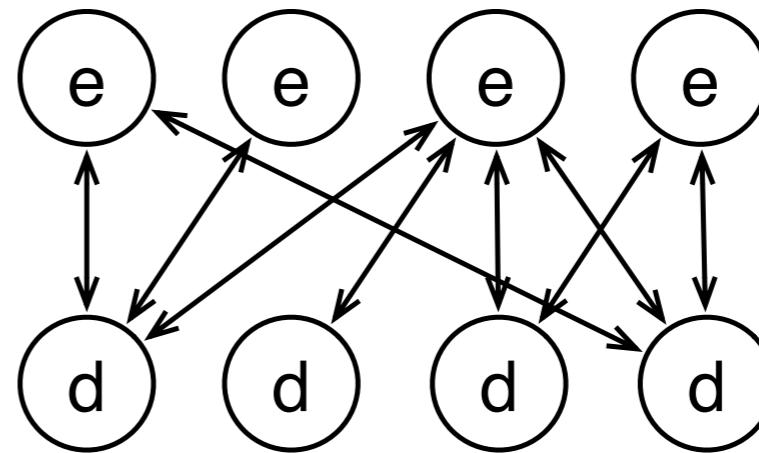
# Graph-based models

[Serdyukov et al. 2008]

- One particular way of constructing graphs
  - Vertices are documents and entities
  - Only document-entity edges
- Search can be approached as a random walk on this graph
  - Pick a random document or entity
  - Follow links to entities or other documents
  - Repeat it a number of times

# Infinite random walk

[Serdyukov et al. 2008]



$$P_i(d) = \lambda P_J(d) + (1 - \lambda) \sum_{e \rightarrow d} P(d|e) P_{i-1}(e),$$

$$P_i(e) = \sum_{d \rightarrow e} P(e|d) P_{i-1}(d),$$

$$P_J(d) = P(d|q),$$

# Outline

- Part 2 – Entity Retrieval
  - introduction
  - ranking with ready-made entity descriptions
  - ranking without explicit entity representations
  - test collections
  - hands-on
  - open challenges

# **Test collections**

# Test collections

Campaign	Task	Collection	Entity repr.	#Topics
TREC Enterprise (2005-08)	Expert finding	Enterprise intranets (W3C, CSIRO)	Indirect	99 (W3C) 127 (CSIRO)
TREC Entity (2009-11)	Rel. entity finding	Web crawl (ClueWeb09)	Indirect	120
	List completion			70
INEX Entity Ranking (2007-09)	Entity search	Wikipedia	Direct	55
	List completion			
SemSearch Chall. (2010-11)	Entity search	Semantic Web crawl (BTC2009)	Direct	142
	List search			50
INEX Linked Data (2012-13)	Ad-hoc search	Wikipedia + RDF (Wikipedia-LOD)	Direct	100 ('12) 144 ('13)

# **Test collections (2)**

- Entity search as Question Answering
  - TREC QA track
  - QALD-2 challenge
  - INEX-LD Jeopardy task

# Entity search in DBpedia

[Balog & Neumayer 2013]

- Synthesising queries and relevance assessments from previous eval. campaigns
- From short keyword queries to natural language questions
- 485 queries in total
- Results are mapped to DBpedia

# Outline

- Part 2 – Entity Retrieval
  - introduction
  - ranking with ready-made entity descriptions
  - ranking without explicit entity representations
  - test collections
  - hands-on
  - open challenges

# **Hands-on**



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

- YAGO
- Freebase
- DBpedia
- EARS
- Sindice & SIREn

# 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 the WordNet class hierarchy
- See <http://www.mpi-inf.mpg.de/yago-naga/yago/>

# 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
  - Google Knowledge Graph
- See <http://www.firebaseio.com/>

# DBpedia

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

# 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

# **EARS**

- **Entity and Association Retrieval System**
  - Open source, built on top of Lemur in C++
    - Not actively maintained
- **Entity-topic association finding models**
  - Suited for other tasks, e.g. blog distillation
  - Focuses on two entity-related tasks:
    - Finding entities:
      - "Which entities are associated with topic X?"
    - Profiling entities:
      - "What topics is an entity associated with?"
- See <https://code.google.com/p/ears/>

# Sindice/SIREn

- Handling of semi-structured data
  - Efficient, large scale
  - Typically based on DBMS backends
  - Apache Lucene plugin for semi-structured search
- Search engine features: top-k query processing, real time updates, full text search, distributed indexes over shards, etc.
- Open source

# Code Academy

- Contains some (Javascript) coding examples for entity linking and retrieval
  - <http://www.codecademy.com/courses/javascript-beginner-en-LkhDf/>

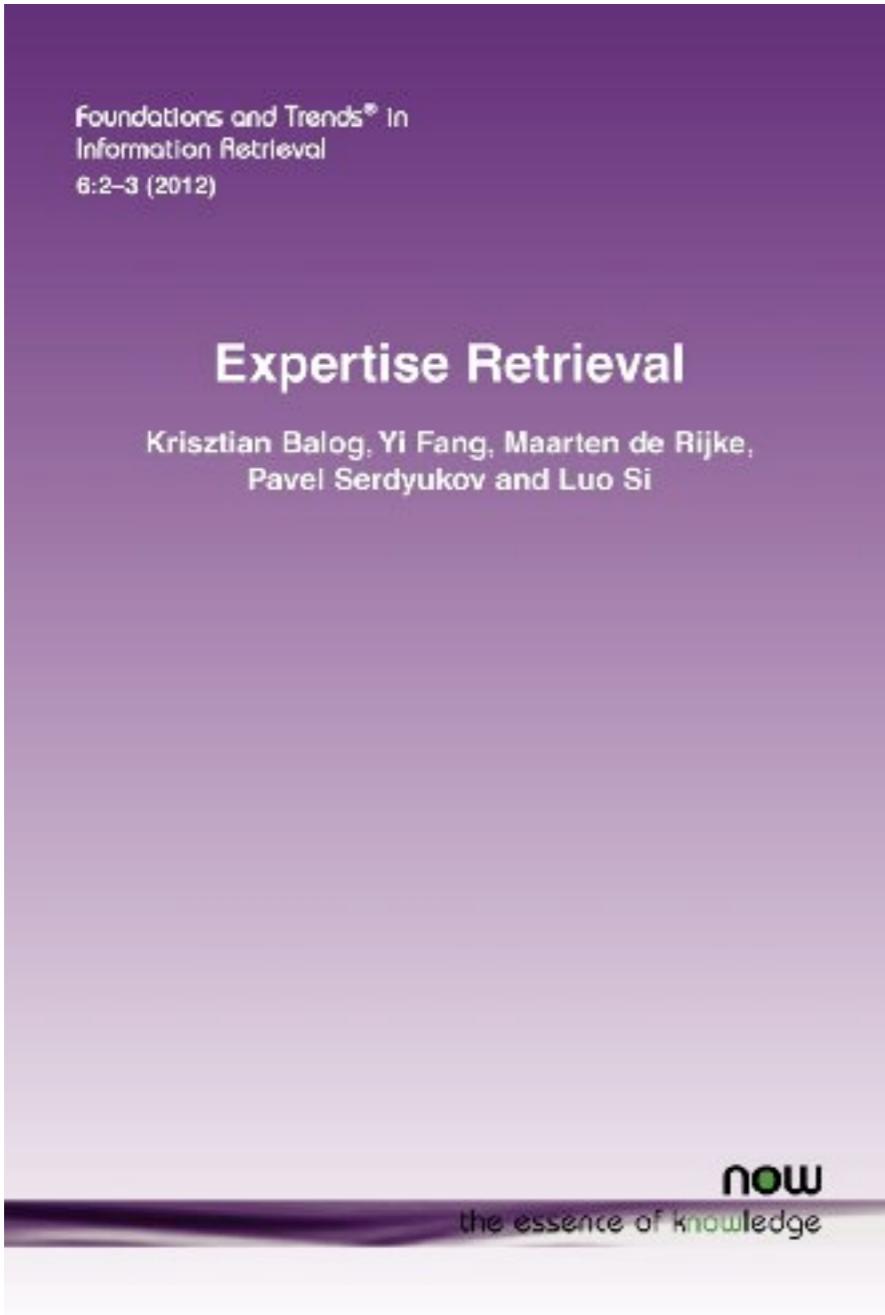
# Outline

- Part 2 – Entity Retrieval
  - introduction
  - ranking with ready-made entity descriptions
  - ranking without explicit entity representations
  - test collections
  - hands-on
  - open challenges

# Open challenges

- Combining text and structure
  - Knowledge bases and unstructured Web documents
- Query understanding and modeling **[Sawant & Chakrabarti 2013]**
- UI/UX/Result presentation
  - How to interact with entities
- Hyperlocal
  - Siri/Google Now/...

# Follow-up reading



K. Balog, Y. Fang, M. de Rijke, P. Serdyukov, and L. Si.  
**Expertise Retrieval.** *FnTIR'12.*

# **References – Entity retrieval**

---

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

# References – Entity retrieval

The screenshot shows a Mendeley group page titled "Entity Linking and Retrieval – Tutorial at WWW 2013 and SIGIR 2013". The page displays 44 papers. The first paper listed is "Analysis and Enhancement of Wikification for Microblogs with Context Expansion" by Taylor Cassidy, Heng Ji, Lev-Arie Ratinov, Arkaitz Zubiaga, Hongzhao Huang in COLING 2012 (2012). The second paper is "Microblog-genre noise and impact on semantic annotation accuracy" by Leon Derczynski, Diana Maynard, Niraj Aswani, Kalina Bontcheva in HT 2013 (2013). The third paper is "Entity Disambiguation with Freebase" by Zhicheng Zheng, Xiance Si, Fangtao Li, Edward Y. Chang, Xiaoyan Zhu in WI-IAT 2013 (2013). The page also features a sidebar with a "Feedback" button and a "Top tags in this group" section.

Papers in Entity Linking and Retrieval – Tutorial at WWW 2013 and SIGIR 2013 | Mendeley Group

www.mendeley.com/groups/3339761/entity-linking-and-retrieval-tutorial-at-www-2013-and-sigir-2013/papers/

Papers in Entity Linking and Retrieval – Tutorial at WWW 2013 and SIGIR 2013 | Mendeley Group

MENDELEY

Get Mendeley What is Mendeley? Papers Groups Sign up & Download Sign in

Groups Search...

Entity Linking and Retrieval – Tutorial at WWW 2013 and SIGIR 2013

In this group: 44 papers · 2/3 members Follow this group Share f t e

Mendeley Computer and Information Science Groups

Overview Papers Members

1 - 20 of 44 Prev 1 2 3 Next

Analysis and Enhancement of Wikification for Microblogs with Context Expansion.

Taylor Cassidy, Heng Ji, Lev-Arie Ratinov, Arkaitz Zubiaga, Hongzhao Huang 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...

Added 1 minute ago 1 reader

Microblog-genre noise and impact on semantic annotation accuracy

Leon Derczynski, Diana Maynard, Niraj Aswani, Kalina Bontcheva 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...

Added 11 minutes ago

Entity Disambiguation with Freebase

Zhicheng Zheng, Xiance Si, Fangtao Li, Edward Y. Chang, Xiaoyan Zhu in WI-IAT 2013 (2013)

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 web

Feedback

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

# Tutorial Resources

- Complete tutorial material  
<http://bit.ly/yahoosummerschool>
- Coding exercises  
see link above under “**CodeAcademy Course**”
- References  
<http://www.mendeley.com/groups/3339761/entity-linking-and-retrieval-tutorial-at-www-2013-and-sigir-2013/>