### Entity Linking

### Pawan Goyal

CSE, IIT Kharagpur

August 17th, 2016

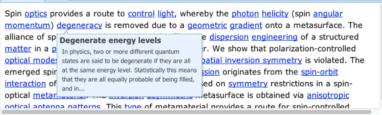
### What is Entity Linking?

#### 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 spinoptical 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.





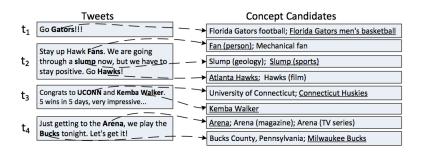


< ロ > < 同 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ >

#### 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 o ranian POWs allegedly in Iraq, the official Iraqi News Agency reported.				
	raqi Foreign Minister Mohammed Saeed al-Sahhaf 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 <b>1980-88 iran-iraq War</b> . 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 be			ganization has had difficulty visiting POWs on both	
sides on a regular basis.	Baghdad	X		
In April, Iran released 5,584			fied as civil law detainees in the largest exchange	
since 1990.	Baghdad is the capital of Iraq and of Bag	~		
More than 1 million people w	With a metropolitan area estimated a			
wore than I minor people w	7,000,000, it is the largest city in Iraq. It is			
	city in the Arab world (after Cairo) and the	e second-largest city		
	in southwest Asia (after Tehran).			
		open in wikipedia		



Determine "linkable" phrases

mention detection - MD

Determine "linkable" phrases

mention detection - MD

Rank/Select candidate entity links

link generation - LG

Determine "linkable" phrases

mention detection - MD

Rank/Select candidate entity links

link generation - LG

Use "context" to disambiguate/filter/improve

disambiguation - DA

**Q** ... degeneracy is removed ...

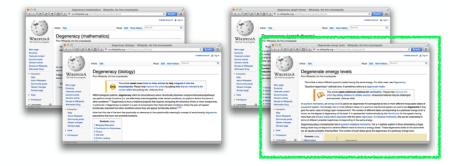
### Link Generation (LG)

### **Q** ... degeneracy ...



イロト イポト イヨト イヨト

### **Q** ... degeneracy ...



イロト イポト イヨト イヨト

- 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]]

00 01 + + w	en.wikipedia.org/wiki/United_States
WIKIPEDIA The Fire Incyclopedia	Article Talk United States From Wikipedia, the three encycloped
Main page Contents Featured content Current events Random article Donate to Waipedia	This article is about the United The United States of America (US states and a federal district. The lo Canada and Maxico. The state of J Pacific. The country also has five ; At 3.79 million square miles (0.83 i
Wikimedia Shop     Interaction     Help     About Wikipedia     Community portal     Recent chances	the third-largest by both land area a immigration from many countries. <sup>13</sup> are home to a wide variety of speci Paleoindians migrated from Asia to Europeans greatly reduced their po thirdeen Bitlish colorise located ale
Contact Wikipedia     Foolbox     Printlexport     Languages	On July 4, 1776, delegates from th The American Revolutionary War, successful war of independence and Amendments have since been add guarantee many fundamental civil
Aobh Agursóso • Afrikaans Alemannisch hir9CT Ængliso	Driven by the doctrine of manifest. This involved displacing native trib the United States. <sup>[14]</sup> By the end of largest. <sup>[14]</sup> The Spanish-American World War II as a global superpow the Cold War and the dissolution of
Ancues auxil Arapones DDDDD Arpetan webn	The United States is a developed of at purchasing-power parity, as of 28 inequality was also ranked highest developed infrastructure, <sup>[22]</sup> and hig manufacturers. <sup>[24]</sup> The country acc as a leader in scientific research as
Asturianu	Contents (hide

## Preliminaries: Disambiguation Pages

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

00	New York (disambiguation) -
	W en.wikipedia.org/wiki/New_York_(disambiguation)
WIKIPEDIA The Fire Encyclopedia	Artice Task New York (disambiguation) From Wapeds, the two exceptionals
Main page Contents Featured content Current events Random article Donate to Wikipedia Waimedia Shoo	New York is a state in the United States of America. New York may also refer to: Places [ext] England [ext]
<ul> <li>Interaction</li> <li>Help</li> <li>About Wikipedia</li> <li>Community portal</li> <li>Recent changes</li> <li>Contact page</li> </ul>	New York, North Yorkahe     New York, Type and Wear United States [ont]     New York (Drg, a city in New York State and the largest city in the United States     New York (Drg, a city in New York State and the largest city and ts su     New York Chrys, covering the same area as the New York Chy Isocoph of the
Tools     Printlexport	<ul> <li>Province of New York, a British colony preceding the state of New York</li> <li>New York Mills, Minnesota, a city in Otter Tail County.</li> </ul>
Languages     Atrikaans     Alemannisch     AEnglisc     Langl	New York, Texas, a community in Henderson County, Texas     New York, Texas, a community in Henderson County, Texas     West New York, Texas, a community in Henderson Rower Nore New York Context     New York, Missouri, a former community founded in 1850 or 1036 new Comme
ССССС Беларуская Български	Black Sea [edit] • The former Anglo-Saxon settlement of New York
Boarisch Bosanski	Media and entertainment [edt]
Brezhoneg Català Čeština Cymraeg Dansk Deutsch	Film [edit] • New York (Itim), 2009 Bollywood film directed by Kabir Khan • New York, Alexe York (Itim), by Martin Scoresee • New York: A Documentary Film, by Nic Burna Literature [edit]

### WordNet

- 80k entity definitions
- 142k senses (entity surface forms)

### WordNet

- 80k entity definitions
- 142k senses (entity surface forms)

### Wikipedia

- 4M entity definitions
- 24M senses

What can be a good measure for MD?

What can be a good measure for MD?

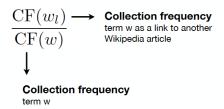
keyphraseness(w)

Number of Wikipedia articles that use it as an anchor, divided by the numer of articles that mention it at all.

What can be a good measure for MD?

### *keyphraseness(w)*

Number of Wikipedia articles that use it as an anchor, divided by the numer of articles that mention it at all.



What can be a good measure for DA?

What can be a good measure for DA?

*commonness*(*w*,*c*)

The fraction of times, a particular sense is used as a destination in Wikipedia.

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

Number of links with target c' and anchor text w

0.9417
0.0139
0.0081
0.0065

	monness
FIFA World Cup	0.2358
FIS Alpine Ski World Cup	0.0682
2009 FINA Swimming World Cup	0.0633
World Cup (men's golf)	0.0622
FIS Alpine Ski World Cup 2009 FINA Swimming World Cup World Cup (men's golf)	0.0682 0.063 0.062

#### keyphrasenes

**D** 1 1

hraseness	Bulgaria national tootball team	
<b>1</b>	The Bulgaria national football team is the national football team of Bulgaria and is cont Sensey the	commonness
0.9	Bulgarian Football Uniger-Bulgaria's best World Cup performance was in the 1994 World 1998 FIFA World Cup	0.9556
0.8	where they beat Germany to reach the semi-finals, losing to Italy, and finishing in fourth 1998 IAAF World Cup	0.0296
	4-0 defeat to Sweden in the third place play-off. Bulgaria's first appearance in a World Cu 1998 Alpine Skiing World Cu	up 0.0059
0.7	1962 World Cup in Chile, but failed to progress to stagesThe Bulgarians. after a 1-1	
0.6	draw against Spain (a fantastic Stoltchkov goal was controversially cancelled) and a 1-0 victory against	
0.5	Romania, played well but lost the third and decreased to a very strong France (	
10.5	champion), 1-3, The Bulgarians did not prog Golden Generation hals in the 1998 World Cup,	
· · 0.4	Vasil Levski National Stadium in. However, the "Golden Generation" was history It has a	
- 0.3	capacity of 43 634, Vasil Levski National Stadium was officially opened in 1953 and reconstructed in	
	PFC Levski Sofia . During the 2006/2007 UEFA C Chelsea F.C. gue the stadium was used for the	
0.2	games of FC Levski Sofia with FC Barcelona, Chelsea F.C. and Werder Bremen The stadium	
0.1	also offers judo, artistic gymnastics, basketball, boxing, aerobics, fencing and table tennis halls, as well	
	as a general physical training hall, two and concerning and three restaurants.	
*		

### Depth-first search

From Wikipedia, the free encyclopedia

Depth-first search (DFS) is an algorithm for traversing or searching a tree 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%

### Depth-first search

From Wikipedia, the free encyclopedia

Depth-first search (DFS) is an algorithm for traversing or searching a tree 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%
$\backslash$	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%

### Using Relatedness: Basic Idea

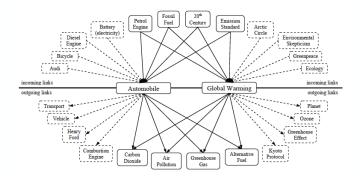
- In a sufficiently long text, one finds terms that do not require disambiguation at all.
- Use every unambiguous link in the document as context to disambiguate ambiguous ones.

### Finding Relatedness: A link-based measure

### Finding Relatedness: A link-based measure

relatedness(c, c')

Using the intersection among incoming as well as outgoing links of two Wikipedia pages



• Each candidate sense and context term is represented by a single Wikipedia article.

- Each candidate sense and context term is represented by a single Wikipedia article.
- Thus the problem is reduced to selecting the sense article that has most in common with all of the context articles.

- Each candidate sense and context term is represented by a single Wikipedia article.
- Thus the problem is reduced to selecting the sense article that has most in common with all of the context articles.
- Comparison of articles is facilitated by the Wikipedia Link-based measure, which measures the semantic similarity of two Wikipedia pages by comparing their incoming and outgoing links.

- Each candidate sense and context term is represented by a single Wikipedia article.
- Thus the problem is reduced to selecting the sense article that has most in common with all of the context articles.
- Comparison of articles is facilitated by the Wikipedia Link-based measure, which measures the semantic similarity of two Wikipedia pages by comparing their incoming and outgoing links.
- The relatedness of a candidate sense is the weighted average of its relatedness to each context article.

- Each candidate sense and context term is represented by a single Wikipedia article.
- Thus the problem is reduced to selecting the sense article that has most in common with all of the context articles.
- Comparison of articles is facilitated by the Wikipedia Link-based measure, which measures the semantic similarity of two Wikipedia pages by comparing their incoming and outgoing links.
- The relatedness of a candidate sense is the weighted average of its relatedness to each context article.

*How to give different weights to the context terms?* 

# Weighting the Context Terms

A B + A B +
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A

-

• **link probability:** Use the ones that are almost always used as a link within the articles where they are found, and always link to the same destination

- **link probability:** Use the ones that are almost always used as a link within the articles where they are found, and always link to the same destination
- relatedness: We can determine how closely a term relates to the central document by calculating its average semantic relatedness to all other context terms

- **link probability:** Use the ones that are almost always used as a link within the articles where they are found, and always link to the same destination
- relatedness: We can determine how closely a term relates to the central document by calculating its average semantic relatedness to all other context terms

*These two variables - link probability and relatedness - are averaged to provide a weight for each context.* 

• The link detection process starts by gathering all n-grams in the document, and retaining those whose probability exceeds a very low threshold.

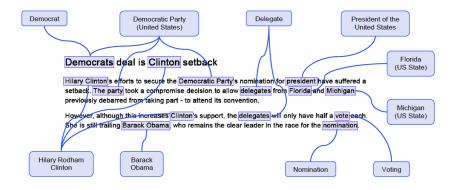
• The link detection process starts by gathering all n-grams in the document, and retaining those whose probability exceeds a very low threshold. *Is it the best method?* 

- The link detection process starts by gathering all n-grams in the document, and retaining those whose probability exceeds a very low threshold. *Is it the best method?*
- All the remaining phrases are disambiguated using the approach mentioned earlier.

- The link detection process starts by gathering all n-grams in the document, and retaining those whose probability exceeds a very low threshold. *Is it the best method?*
- All the remaining phrases are disambiguated using the approach mentioned earlier.
- This results in a set of associations between terms in the document and the Wikipedia articles that describe them.

- The link detection process starts by gathering all n-grams in the document, and retaining those whose probability exceeds a very low threshold. *Is it the best method?*
- All the remaining phrases are disambiguated using the approach mentioned earlier.
- This results in a set of associations between terms in the document and the Wikipedia articles that describe them.

Can you use this to learn – which concepts should be linked?



(日)

# The Learning Problem: Which topics should be linked?

• The automatically identified Wikipedia articles provide training instances for a classifier.

# The Learning Problem: Which topics should be linked?

- The automatically identified Wikipedia articles provide training instances for a classifier.
- Positive examples are the articles that were manually linked to, while negative ones are those that were not.

# The Learning Problem: Which topics should be linked?

- The automatically identified Wikipedia articles provide training instances for a classifier.
- Positive examples are the articles that were manually linked to, while negative ones are those that were not.
- Features of these articles and the places where they were mentioned are used to inform the classifier about which topics should and should not be linked.

- Link Probability: Average as well as maximum of link probability of the link locations (e.g. Hillary Clinton and Clinton)
- **Relatedness:** Topics which relate to the central thread of the document are more likely to be linked
- **Disambiguation Confidence:** The confidence score of the classifier for disambiguation
- **Generality:** Defined as the minimum depth at which it is located in Wikipedia's category tree. More useful for the readers to provide links for specific topics.
- Location and Spread: Where are these mentioned? First occurrence, last occurrence and the spread.

- Mihalcea, Rada, and Andras Csomai. "Wikify!: linking documents to encyclopedic knowledge." Proceedings of the sixteenth ACM conference on information and knowledge management. ACM, 2007.
- Milne, David, and Ian H. Witten. "Learning to link with wikipedia." Proceedings of the 17th ACM conference on Information and knowledge management. ACM, 2008.