

Entity Linking - Part I

Pawan Goyal

CSE, IIT Kharagpur

Week 10, Lecture 1

Entity Linking: Introduction



- Entity Linking is the task of identifying all mentions in text of a specific entity from a database or ontology
- Researchers have used Wikipedia, YAGO, Freebase etc. as the databases
- Typically broken down into two main phases
 - ▶ Candidate Selection (entity annotation)
 - ▶ Reference disambiguation or entity resolution
- Entity linking needs to handle
 - ▶ Name variations (entities are referred to in many different ways)
 - ▶ Entity ambiguity (the same string can refer to more than one entity)

Entity Linking: Introduction

- We will take Wikipedia as the knowledge base to understand the task.
- With Wikipedia as the knowledge base, this task is commonly known as *Wikification*.


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 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 [matter](#) in a [p](#) [optical mode](#); emerged spin [interaction](#) of optical [metar](#) [optical antenna patterns](#). This [type](#) of metamaterial provides a route for spin-controlled

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...

What is 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-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 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

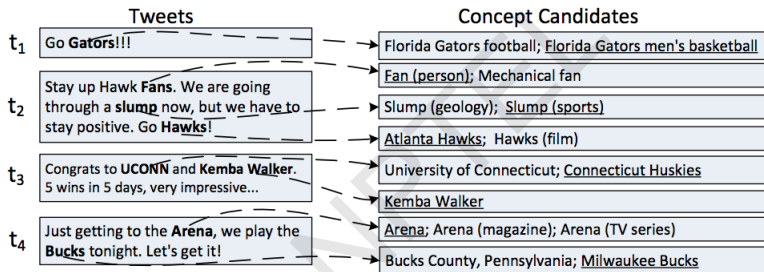
Baghdad

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](#)

ried as civil law detainees in the largest exchange

What is Entity Linking?



Entity Linking: Common Steps

Determine “linkable” phrases

mention detection - **MD**

Rank/Select candidate entity links

link generation - **LG**

Use “context” to disambiguate/filter/improve

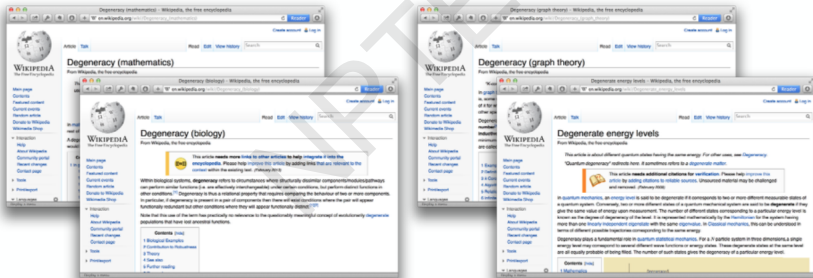
disambiguation - **DA**

Mention Detection (MD)

Q ... degeneracy is removed ...

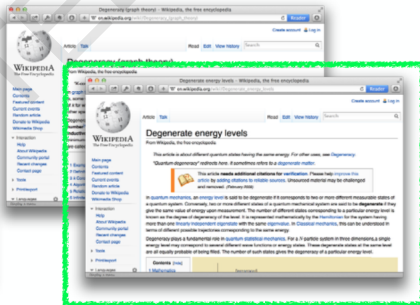
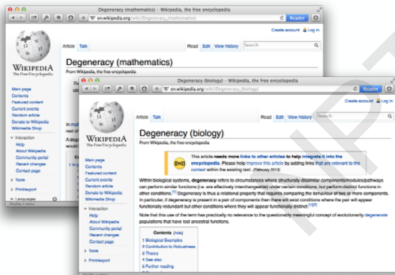
Link Generation (LG)

Q ... degeneracy ...



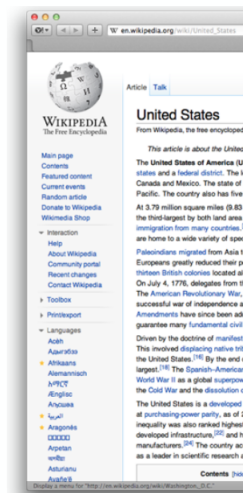
Disambiguation (DA)

Q ... degeneracy ...



Preliminaries: 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]]



Preliminaries: Disambiguation Pages

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



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?

NPTEL

What can be a good measure for MD?

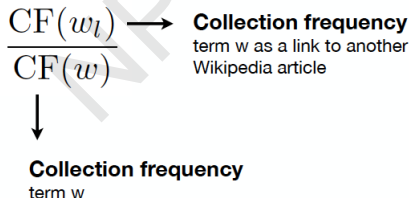
keyphraseness(w)

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



What can be a good measure for DA?

NPTEL

What can be a good measure for DA?

commonness(w, c)

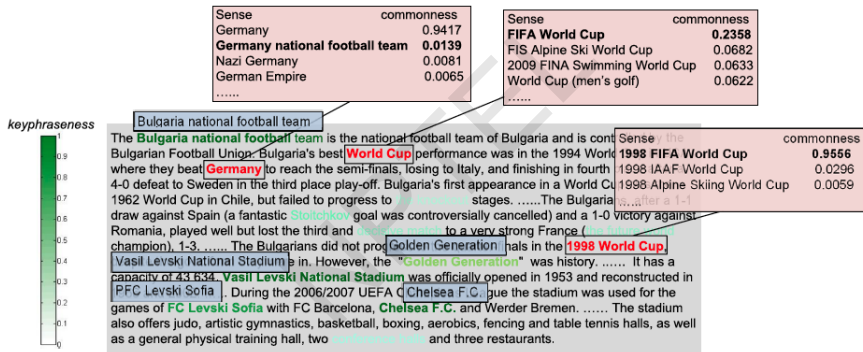
The fraction of times, a particular sense is used as a destination in Wikipedia.

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

Number of links

with target c' and anchor text w

Commonness and keyphraseness



Always the best decision?

NPTEL

Always the best decision?

- This can never help you build an accurate system, because you will always give some wrong links.
- Need to use the context.

Keyphraseness and Commonness: Always the best decision?

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%
...		

Keyphraseness and Commonness: Always the best decision?

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%
...		

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.

- 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

NPTEL

Weighting the 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

Weighting the 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

Weighting the 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.

Can we improve mention detection with this approach?

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

NPTEL

Can we improve mention detection with this approach?

- 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?*

Can we improve mention detection with this approach?

- 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.

Can we improve mention detection with this approach?

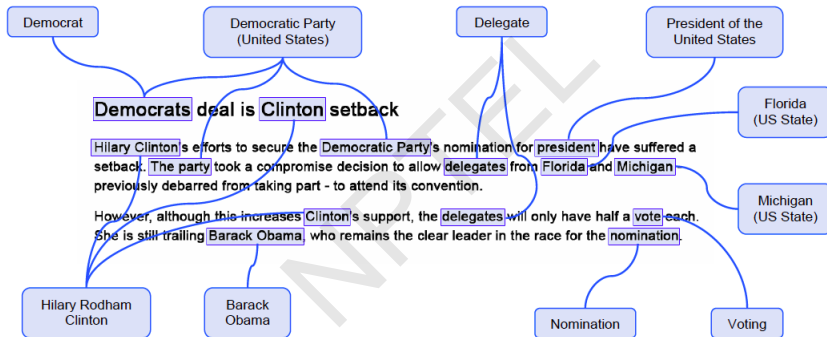
- The link detection process starts by gathering all n-grams in the document, and retaining those whose probability exceeds a very low threshold. *Is it the best method?*
- All the remaining phrases are disambiguated using the approach mentioned earlier.
- This results in a set of associations between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

Can we improve mention detection with this approach?

- 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?

Example



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.

What are the features?

- **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.