

# *Entity Linking*

# Entity Linking

- Entity linking is the task of identifying all mentions in text of a specific entity from a database or ontology
- Also referred to as entity disambiguation: there may be several different with the same name in the database
- Link mentions to the concept in the KB that best matches the meaning in the given context
- Do this efficiently for a KB with millions of concepts and with dozens or hundreds of concept candidates per mention

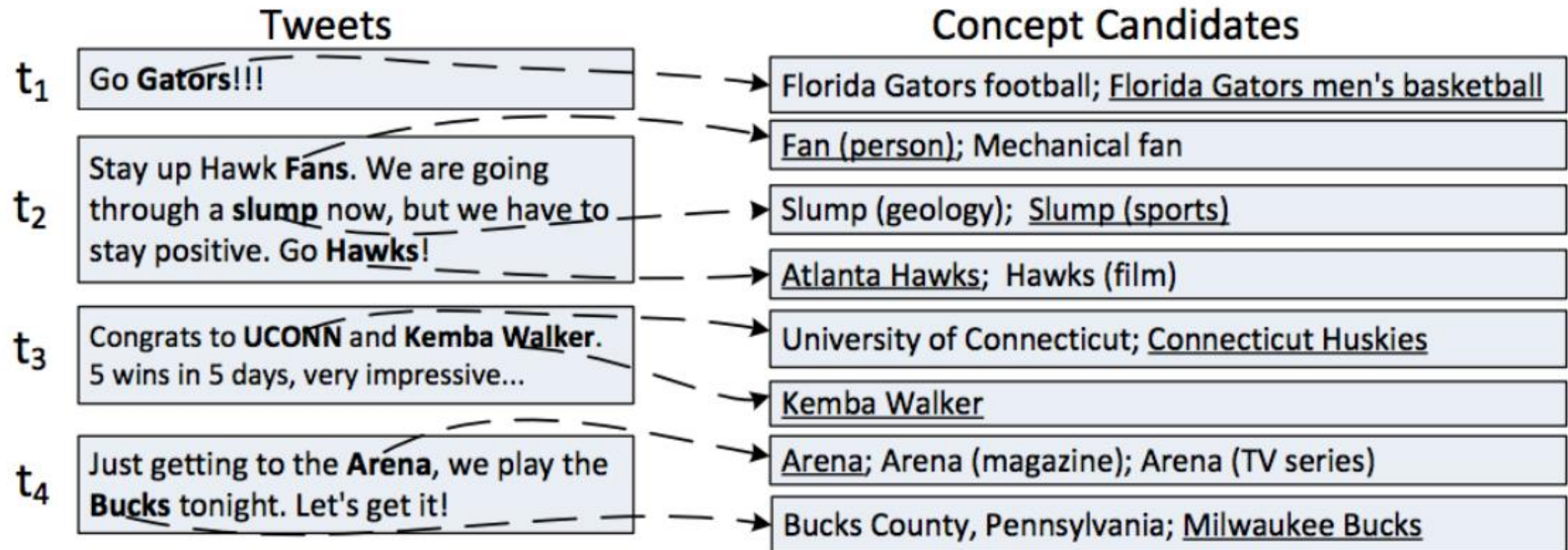
[https://en.wikipedia.org/wiki/Bat-and-ball\\_games](https://en.wikipedia.org/wiki/Bat-and-ball_games)

[https://en.wikipedia.org/wiki/Cricket\\_field](https://en.wikipedia.org/wiki/Cricket_field)

**Cricket** is a bat-and-ball game played between two teams of eleven players on a field at the centre of which is a 22-yard (20-metre) pitch with a wicket at each end, each comprising two bails balanced on three stumps. The batting side scores runs by striking the ball bowled at one of the wickets with the bat and then running between the

# What are we doing?

- Rather than just annotate the words “field” as a Place (NER), link it to a specific ontology instance (entity)
  - Differentiate between [Battle field](#), field (agriculture), field (sports) etc.
  - Ontologies tell us that this particular field is a sports field, which is a type of place. The surface of a field is most commonly composed of [sod \(grass\)](#), but may also be [artificial turf](#), [sand](#), [clay](#), [gravel](#), [concrete](#), or other materials, etc.
- These details are all helpful to disambiguate and link the mention (entity) in the text to the correct entity URI in the ontology
- Having identified the unique entity enables further tasks like relation detection, semantic search, etc.



# Why is it hard?

- Entity linking needs to handle:
  - Name variations (entities are referred to in many different ways, including colloquial variants)
  - Entity ambiguity (the same string can refer to more than one entity)
  - Missing entities – there is no target entity in the entity knowledge base/database



# Data Sources for EL

- Entity Linking is based on a datasource/knowledge base to which to link (or several)
- Researchers have used Wikipedia (e.g. TAC KBP, WikipediaMiner), Linked Open Data (in particular DBpedia, YAGO, and Freebase) and the UMLS metathesaurus for medical concepts (Metamap, Bio-YODIE)
- Some datasources/knowledge bases have names in several languages (Wikipedia, UMLS), but coverage is often very different between languages

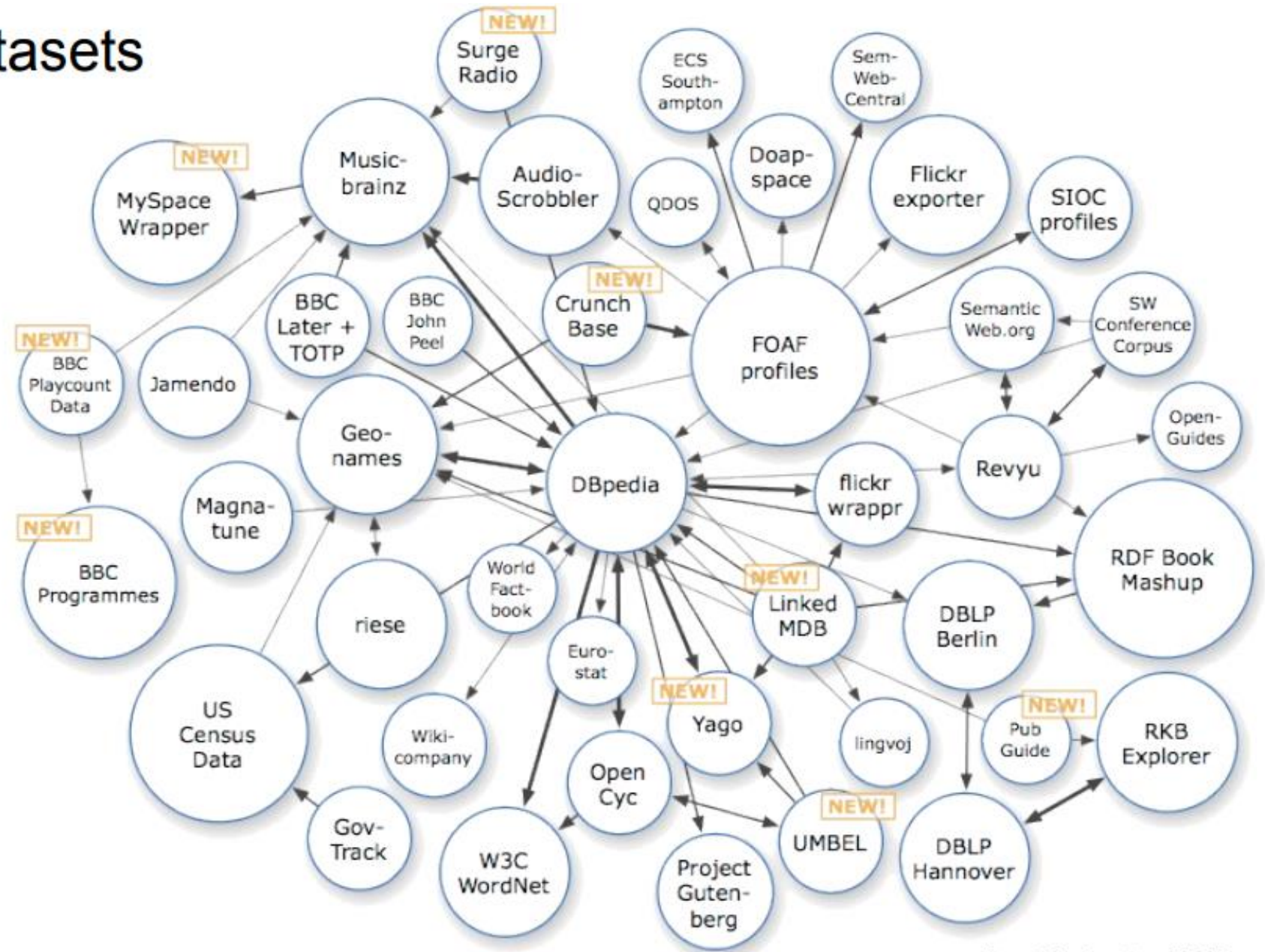
# Data Sources for EL

- The entity linking system can either return a matching entry from the target knowledge base or “NIL” to indicate there is no matching entry in the entity database
- Some entity linking systems make the closed world assumption (CWA) that there is always a target entity in the database
- Often still focused on entities of type PER, LOC, ORG and often focused on English documents



# Linked Open data (2008)

## 45 Datasets



As of September 2008



## 1014 Datasets



# Entity Linking: Common Steps

- **Mention Detection(MD):** Identify potential (linkable) entity mentions by matching known names of entities against text of documents/tweets/etc.
- **Link Generation (LG):** For each mention, get all entity candidates and information related to the entity, mention and entity/mention combination
- **Disambiguation(DA):** Score entity candidates based on contextual fit, mention type, congruence with other potential entities etc.
  - Select candidate using a machine learning approach that learns to combine the scores to find the best entity – Try to identify mentions that do not refer to anything in Dbpedia (NILs)

# Example

- Mention Detection

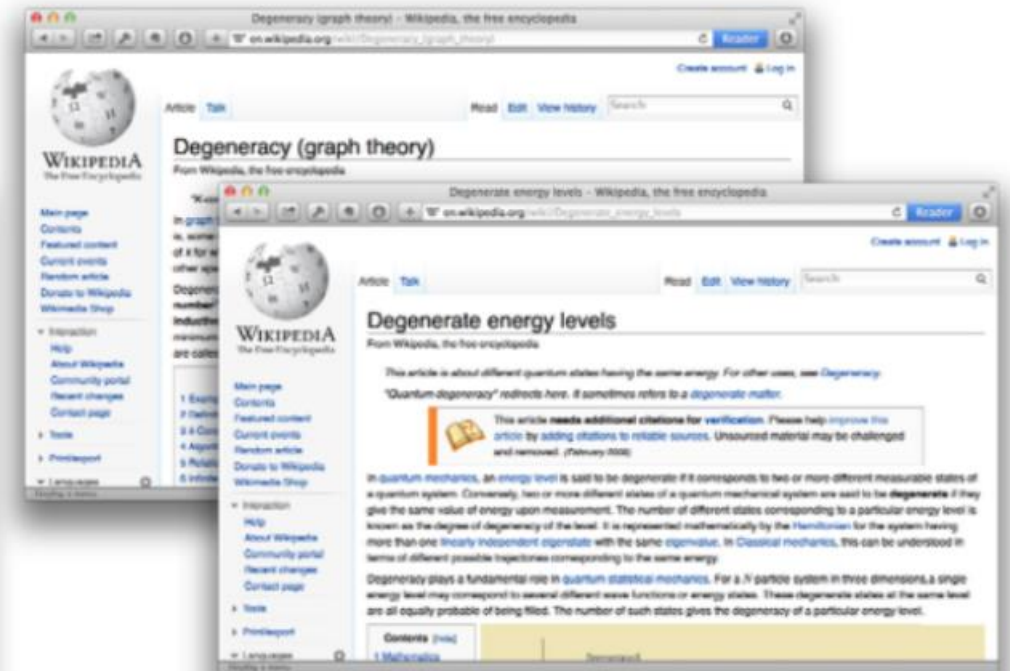
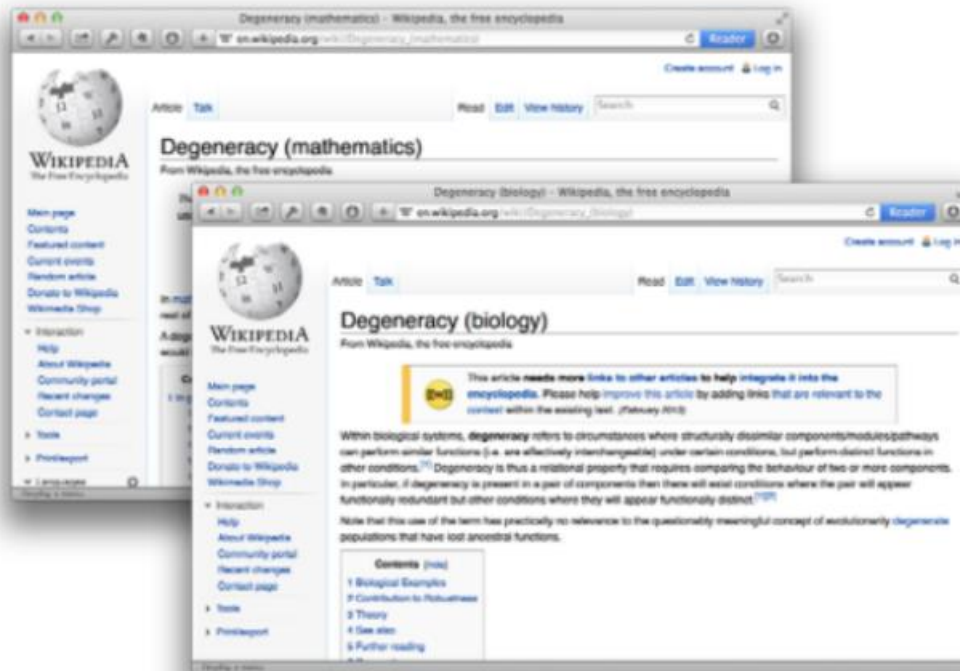
Q ... degeneracy is removed ...



# Example

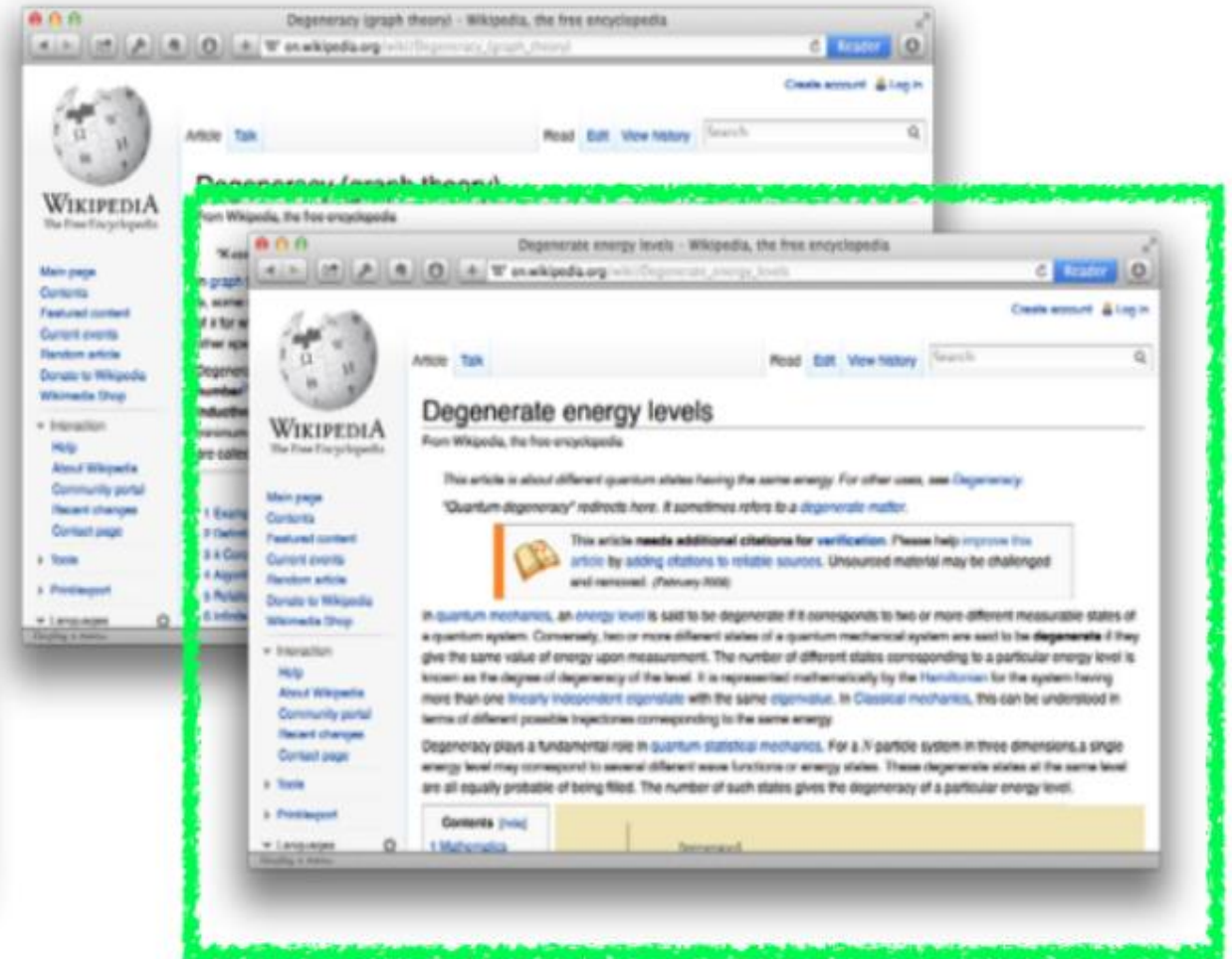
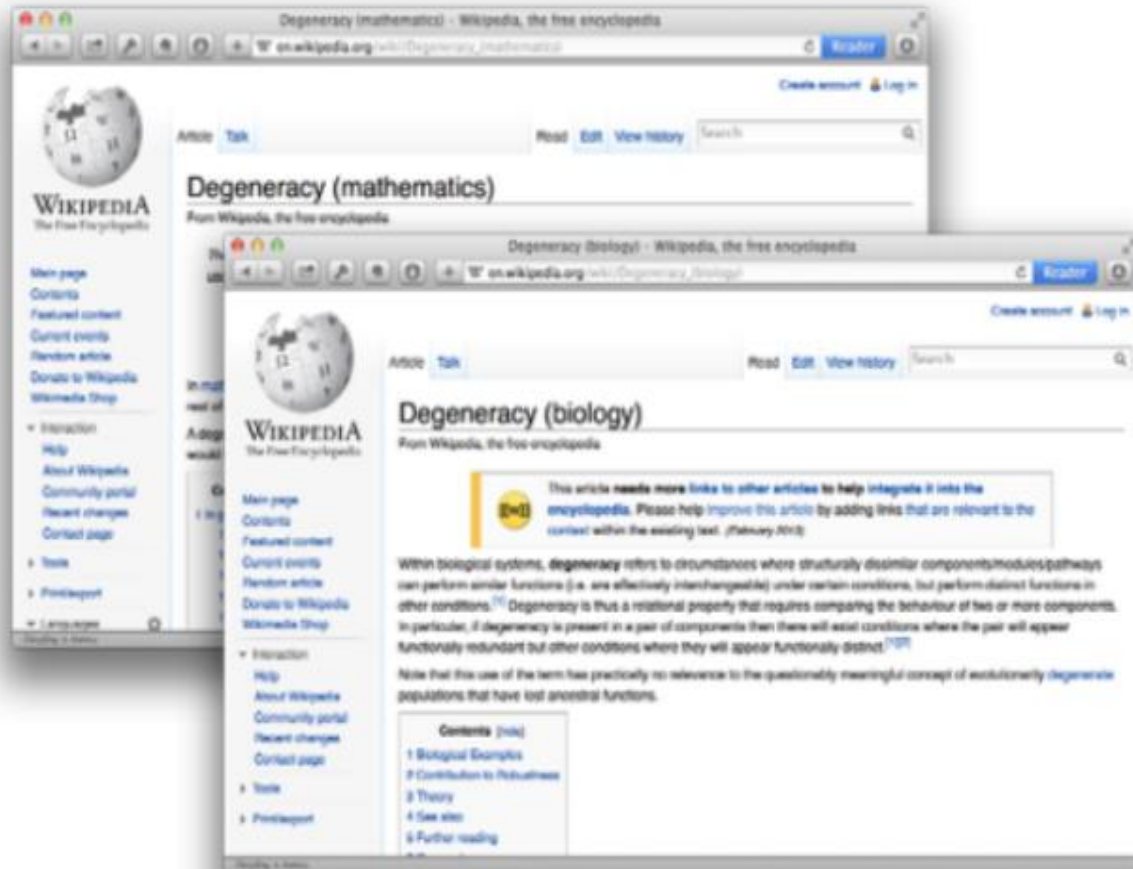
## Link Generation

Q ... degeneracy ...



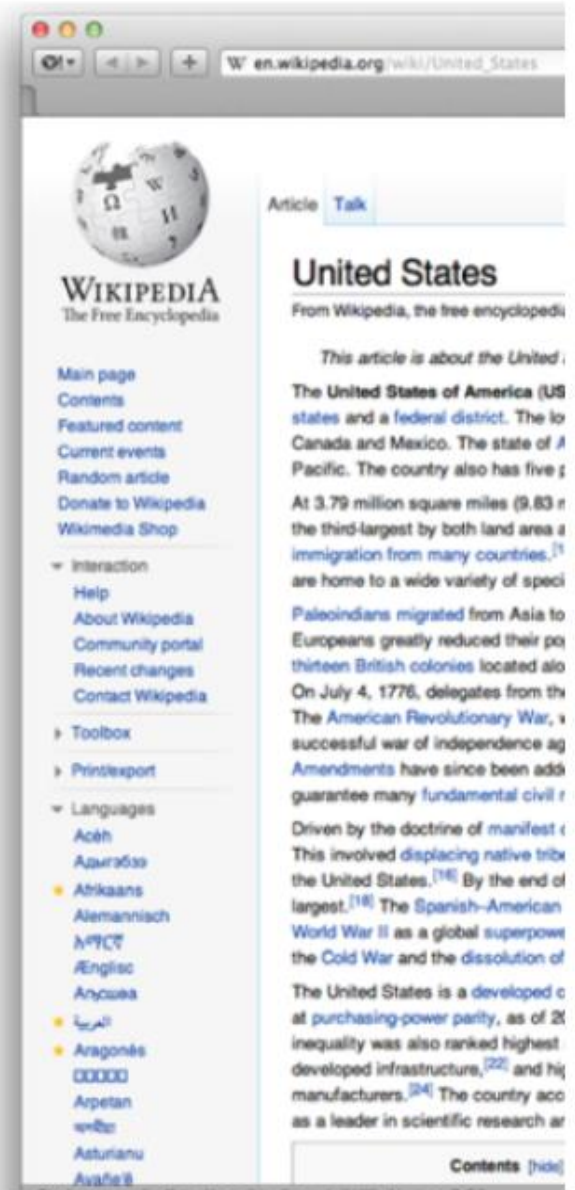
# Example

## Disambiguation



# 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]]





# Disambiguation Page

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

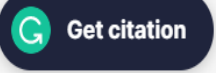


WIKIPEDIA  
The Free Encyclopedia

[Main page](#)  
[Contents](#)  
[Current events](#)  
[Random article](#)  
[About Wikipedia](#)  
[Contact us](#)  
[Donate](#)

[Contribute](#)  
[Help](#)  
[Learn to edit](#)  
[Community portal](#)  
[Recent changes](#)  
[Upload file](#)

[Tools](#)  
[What links here](#)  
[Related changes](#)  
[Special pages](#)  
[Permanent link](#)  
[Page information](#)

 **Get citation**

[Print/export](#)

Article [Talk](#)

## New York

From Wikipedia, the free encyclopedia

**New York** most commonly refers to:

- [New York City](#), the most populous city in the United States, located in the state of New York
- [New York \(state\)](#), a state in the northeastern United States

**New York** may also refer to:

### Film and television

- [New York \(1916 film\)](#), a lost American silent comedy drama by George Fitzmaurice
- [New York \(1927 film\)](#), an American silent drama by Luther Reed
- [New York \(2009 film\)](#), a Bollywood film by Kabir Khan
- [New York: A Documentary Film](#), a film by Ric Burns
- "New York" (*Glee*), an episode of *Glee*

### Literature

- [New York \(Burgess book\)](#), a 1976 work of travel and observation by Anthony Burgess
- [New York \(Morand book\)](#), a 1930 travel book by Paul Morand
- [New York \(novel\)](#), a 2009 historical novel by Edward Rutherfurd
- [New York \(magazine\)](#), a bi-weekly magazine founded in 1968

### Music

## Music

- [New York EP](#), a 2012 EP by Angel Haze
  - "New York" ([Angel Haze song](#))
- [New York \(album\)](#), a 1989 album by Lou Reec
- "New York" ([Eskimo Joe song](#)) (2007)
- "New York" ([Ja Rule song](#)) (2004)
- "New York" ([Paloma Faith song](#)) (2009)
- "New York" ([St. Vincent song](#)) (2017)
- "New York" ([Snow Patrol song](#)) (2011)
- "New York" ([U2 song](#)) (2000)
- [New York](#), a 2006 album by [Antti Tuisku](#)
- "New York", a 1977 song by the Sex Pistols fr

## Places

### United Kingdom

- [New York, Lincolnshire](#)
- [New York, North Yorkshire](#)
- [New York, Tyne and Wear](#)

### United States

#### New York state

- [New York metropolitan area](#), the region encomp
- [New York County](#), covering the same area as t
  - New York, the [US Postal Service address](#)
- [Province of New York](#), a British colony precedi



# Candidate Ambiguity is High- Tough Task

	TAC-KBP				
	PER	LOC	ORG	UKN	TOTAL
Entities	89	361	141	274	865
Avg. number of tokens	1.91	1.20	2.12	1.87	1.78
Candidate URIs	9,427	9,553	9,502	14,649	43,131
Avg. number cand. URIs	105.02	26.46	67.39	53.46	49.86
Unambig. candidates	3	10	3	43	59

- Ambiguity depends on what and how we match
- Case, spelling variations
- Alternate known names, short names
- ...

# Some Statistic- Knowledge Base

## *WordNet*

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

## *Wikipedia*

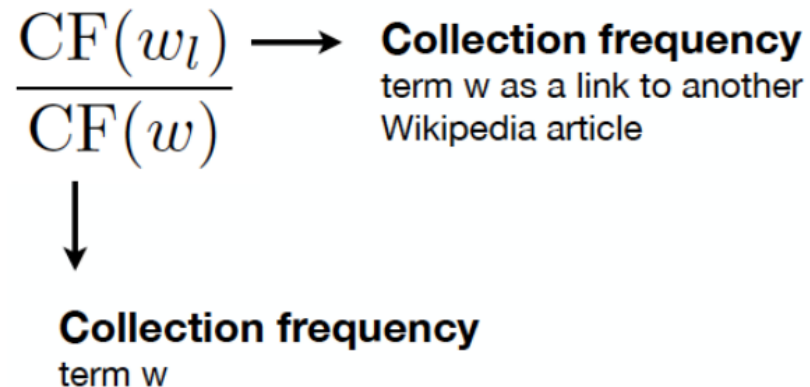
- 4M entity definitions
- 24M senses

# MD: Wikipedia based methods

- What can be good measure for Mention Detection

*keyphraseness(w)*

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



# DA: Wikipedia based methods

- What can be good measure for Disambiguation

*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**

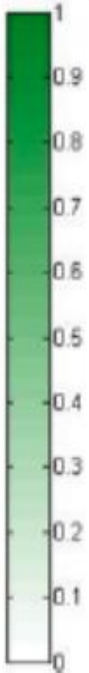
with target  $c'$  and anchor text  $w$

# Commonness and Keyphraseness

Sense	commonness
Germany	0.9417
<b>Germany national football team</b>	<b>0.0139</b>
Nazi Germany	0.0081
German Empire	0.0065
.....	

Sense	commonness
<b>FIFA World Cup</b>	<b>0.2358</b>
FIS Alpine Ski World Cup	0.0682
2009 FINA Swimming World Cup	0.0633
World Cup (men's golf)	0.0622
.....	

keyphraseness



Bulgaria national football team

The **Bulgaria national football team** is the national football team of Bulgaria and is controlled by the Bulgarian Football Union. Bulgaria's best **World Cup** performance was in the 1994 World Cup where they beat **Germany** to reach the semi-finals, losing to Italy, and finishing in fourth place after a 4-0 defeat to Sweden in the third place play-off. Bulgaria's first appearance in a World Cup was at the 1962 World Cup in Chile, but failed to progress to the **knockout** stages. ....The Bulgarians drew against Spain (a fantastic **Stoichkov** goal was controversially cancelled) and a 1-0 victory against Romania, played well but lost the third and **decisive match** to a very strong France (**the future world champion**), 1-3. .... The Bulgarians did not progress to the **Golden Generation** finals in the **1998 World Cup**. **Vasil Levski National Stadium** is in. However, the "**Golden Generation**" was history. .... It has a capacity of 43 634. **Vasil Levski National Stadium** was officially opened in 1953 and reconstructed in 2005. **PFC Levski Sofia**. During the 2006/2007 UEFA Champions League the stadium was used for the games of **FC Levski Sofia** with FC Barcelona, **Chelsea F.C.** and Werder Bremen. .... The stadium also offers judo, artistic gymnastics, basketball, boxing, aerobics, fencing and table tennis halls, as well as a general physical training hall, two **conference halls** and three restaurants.

Sense	commonness
<b>1998 FIFA World Cup</b>	<b>0.9556</b>
1998 IAAF World Cup	0.0296
1998 Alpine Skiing World Cup	0.0059
.....	



# Always the best way?

## 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%
<b>Tree (data structure)</b>	<b>2.57%</b>	<b>63.26%</b>
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%
...		

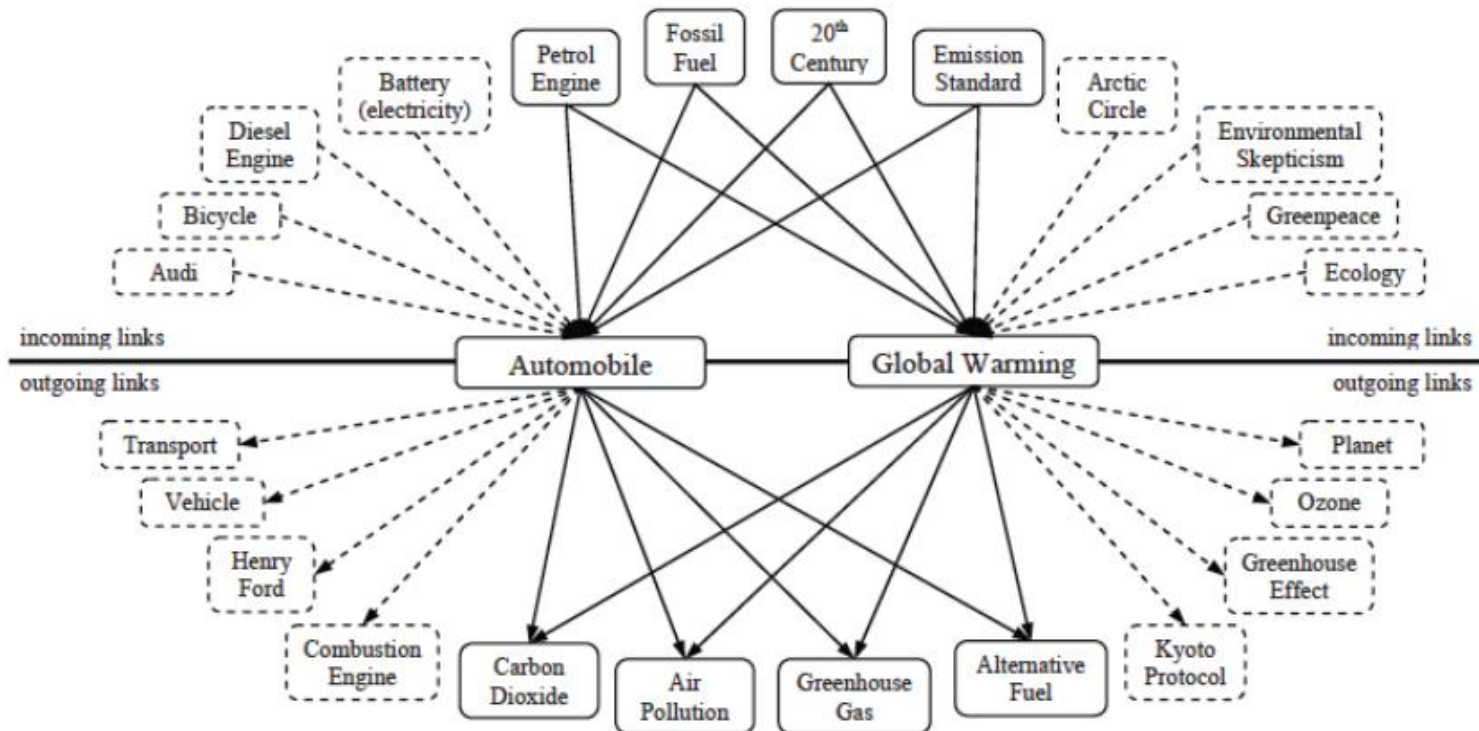
## *Using Relatedness: Basic Idea*

- 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.

# Relatedness: A link-based measure

$relatedness(c, c')$

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





# Computing Relatedness

- 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

- **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

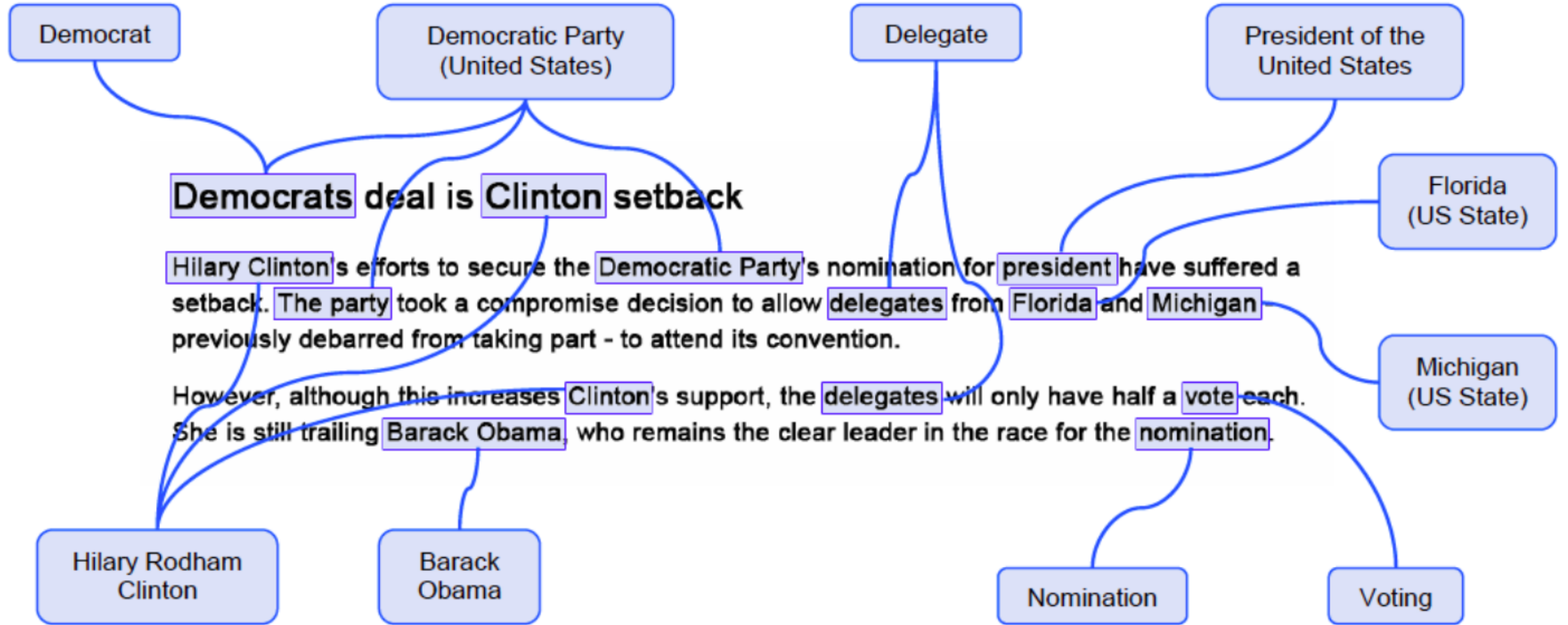
*These two variables - link probability and relatedness - are averaged to provide a weight for each context.*

# 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. *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.
- 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.





