

# Entity Linking

## [DAT640] Information Retrieval and Text Mining

Ivica Kostric & Krisztian Balog  
University of Stavanger

September 19, 2022



CC BY 4.0

## In this module

1. Entity linking
2. Entity linking and retrieval

## Entity linking

## Entity linking

- Task: recognizing entity mentions in text and linking them to the corresponding entries in a knowledge base (KB)
  - Limited to recognizing entities for which a target entry exists in the reference KB; each KB entry is a candidate
  - It is assumed that the document provides sufficient context for disambiguating entities

# Entity linking in action



The DBpedia Spotlight interface shows a search bar with the query "Berlin". Below the search bar are filtering options: Confidence (set to 0.5), Language (set to English), and a checkbox for "n-best candidates". To the right are buttons for "SELECT TYPES..." and "ANNOTATE". A detailed tooltip is displayed over the search results, providing historical context about Berlin's status as a capital city.

Confidence:  0.5

Language: English

n-best candidates

SELECT TYPES... ANNOTATE

First documented in the 13th century, Berlin was the capital of the Kingdom of Prussia (1701–1918), the German Empire (1871–1918), the Weimar Republic (1919–33) and the Third Reich (1933–45). Berlin in the 1920s was the third largest municipality in the world. After World War II, the city became divided into East Berlin -- the capital of East Germany -- and West Berlin, a West German enclave surrounded by the Berlin Wall from 1961–89. Following German reunification in 1990, the city regained its status as the capital of Germany, hosting 147 foreign embassies.

[BACK TO TEXT](#)

# Entity linking in action



The DBpedia Spotlight interface shows entity linking results for the text below. It includes a confidence slider set at 0.5, a language dropdown set to English, and an unchecked checkbox for 'n-best candidates'. Buttons for 'SELECT TYPES...' and 'ANNOTATE' are also present.

Confidence:  0.5

Language: English

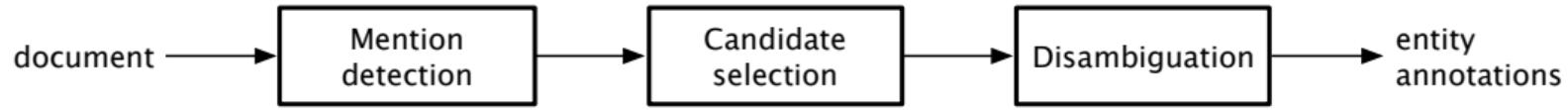
n-best candidates

SELECT TYPES... ANNOTATE

First documented in the 13th century, [Berlin](#) was the capital of the Kingdom of [Prussia](#) (1701–1918), the [German Empire](#) (1871–1918), the [Weimar Republic](#) (1919–33) and the [Third Reich](#) (1933–45). [Berlin](#) in the 1920s was the third largest [municipality](#) in the world. After [World War II](#), the city became divided into [East Berlin](#) -- the capital of [East Germany](#) -- and [West Berlin](#), a [West German enclave](#) surrounded by the [Berlin Wall](#) from 1961–89. Following [German reunification](#) in 1990, the city regained its status as the capital of [Germany](#), hosting 147 foreign embassies.

[BACK TO TEXT](#)

# Anatomy of an entity linking system

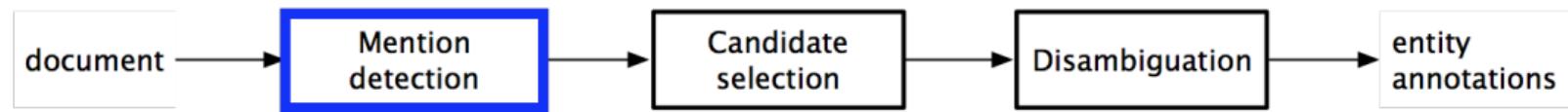


- **Mention detection:** Identification of text snippets that can potentially be linked to entities
- **Candidate selection:** Generating a set of candidate entities for each mention
- **Disambiguation:** Selecting a single entity (or none) for each mention, based on the context

# Entity linking

- Mention detection
- Candidate selection
- Disambiguation
- Evaluation

# Mention detection



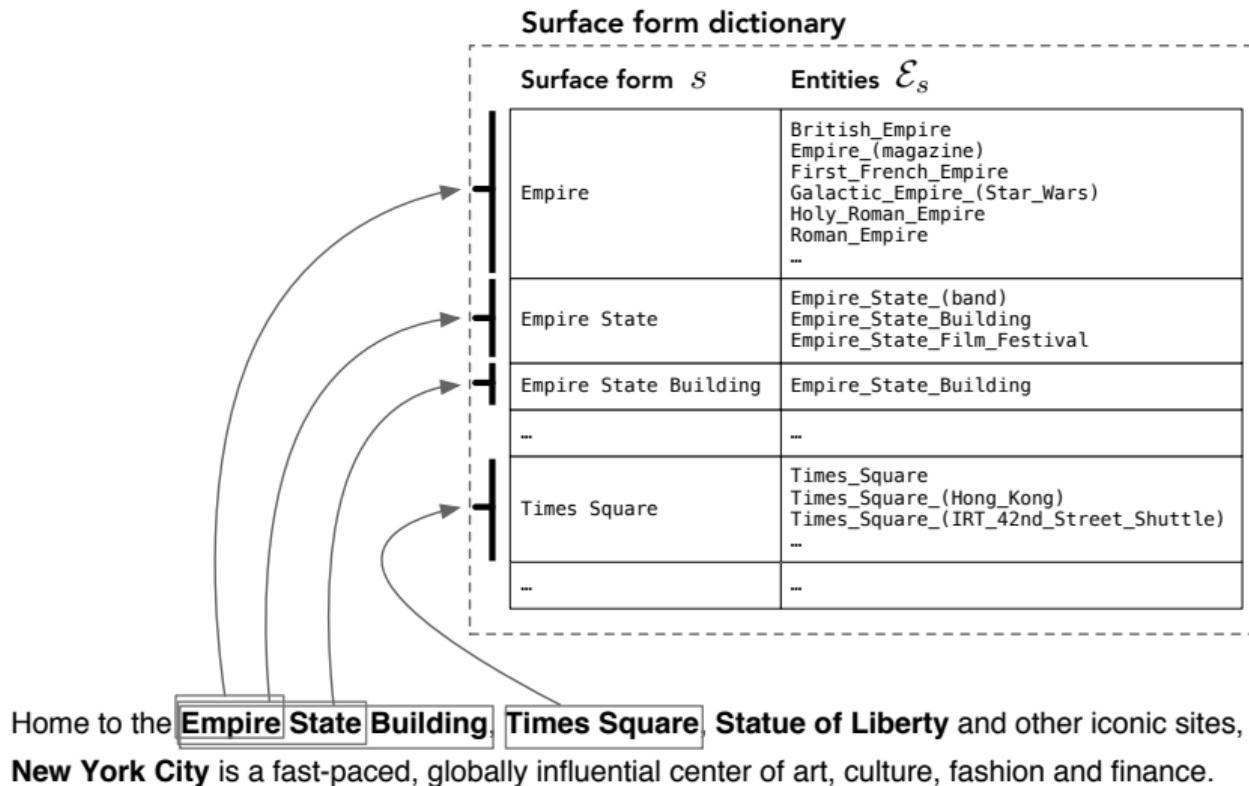
# Mention detection

- Goal: Detect all “linkable” phrases
- Challenges
  - Recall oriented
    - Do not miss any entity that should be linked
  - Find entity name variants
    - E.g. “jlo” is name variant of Jennifer Lopez
  - Filter out inappropriate ones
    - E.g. “new york” matches >2k different entities

## Common approach

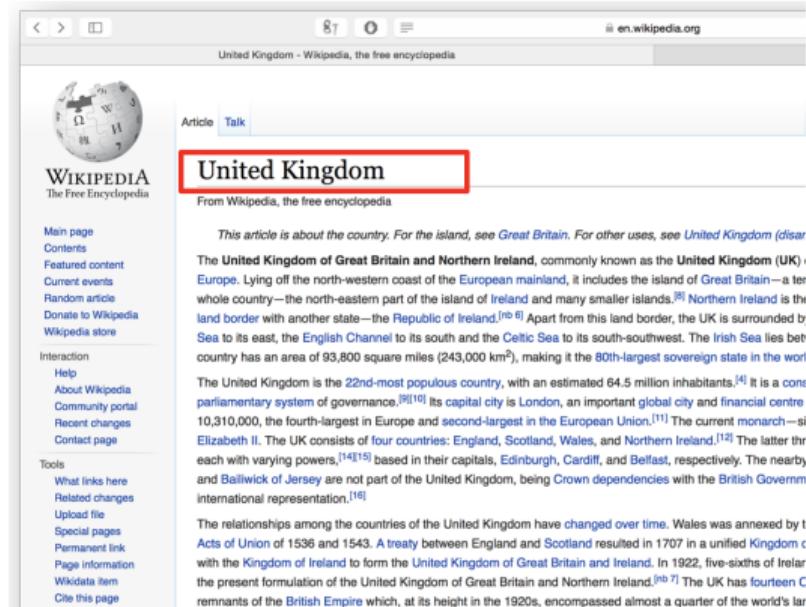
1. Build a dictionary of entity surface forms
  - o Entities with all names variants
2. Check all document n-grams against the dictionary
  - o The value of n is set typically between 6 and 8
3. Filter out undesired entities
  - o Can be done here or later in the pipeline

# Example



# Surface form dictionary construction from Wikipedia

- **Page title**
  - Canonical (most common) name of the entity



The screenshot shows a web browser displaying the English Wikipedia page for "United Kingdom". The title "United Kingdom" is highlighted with a red box. The page content discusses the United Kingdom of Great Britain and Northern Ireland, its location off the north-western coast of Europe, and its status as the 22nd-most populous country. It also mentions the current monarch, Queen Elizabeth II, and the four countries that make up the UK: England, Scotland, Wales, and Northern Ireland.

United Kingdom - Wikipedia, the free encyclopedia

Article Talk

# United Kingdom

From Wikipedia, the free encyclopedia

This article is about the country. For the island, see [Great Britain](#). For other uses, see [United Kingdom \(disambiguation\)](#).

The United Kingdom of Great Britain and Northern Ireland, commonly known as the United Kingdom (UK) or Europe. Lying off the north-western coast of the European mainland, it includes the island of Great Britain—a ter whole country—the north-eastern part of the island of Ireland and many smaller islands.<sup>[8]</sup> Northern Ireland is the land border with another state—the Republic of Ireland.<sup>[nb 6]</sup> Apart from this land border, the UK is surrounded by sea. To its east, the English Channel to its south and the Celtic Sea to its south-southwest. The Irish Sea lies betw country has an area of 93,800 square miles (243,000 km<sup>2</sup>), making it the 80th-largest sovereign state in the world. The United Kingdom is the 22nd-most populous country, with an estimated 64.5 million inhabitants.<sup>[4]</sup> It is a cons parliamentary system of governance.<sup>[9][10]</sup> Its capital city is London, an important global city and financial centre with 10,310,000, the fourth-largest in Europe and second-largest in the European Union.<sup>[11]</sup> The current monarch—si Elizabeth II. The UK consists of four countries: England, Scotland, Wales, and Northern Ireland.<sup>[12]</sup> The latter thr each with varying powers,<sup>[14][15]</sup> based in their capitals, Edinburgh, Cardiff, and Belfast, respectively. The nearby and Bailiwick of Jersey are not part of the United Kingdom, being Crown dependencies with the British Government international representation.<sup>[16]</sup>

The relationships among the countries of the United Kingdom have changed over time. Wales was annexed by t Acts of Union of 1536 and 1543. A treaty between England and Scotland resulted in 1707 in a unified Kingdom c with the Kingdom of Ireland to form the United Kingdom of Great Britain and Ireland. In 1922, five-sixths of Ireland the present formulation of the United Kingdom of Great Britain and Northern Ireland.<sup>[nb 7]</sup> The UK has fourteen C remnants of the British Empire which, at its height in the 1920s, encompassed almost a quarter of the world's la

# Surface form dictionary construction from Wikipedia

- Page title
- Redirect pages
  - Alternative names that are frequently used to refer to an entity



The screenshot shows a web browser displaying the Wikipedia article for "United Kingdom". The page title is "United Kingdom" and the sub-page name is "(Redirected from UK)". A red box highlights the "(Redirected from UK)" link. The page content discusses the United Kingdom of Great Britain and Northern Ireland, its location, and its status as a sovereign state. It also mentions the Republic of Ireland and the English Channel.

United Kingdom

(Redirected from UK)

This article is about the country. For the island, see [Great Britain](#). For other uses, see [United Kingdom \(disambiguation\)](#).

The United Kingdom of Great Britain and Northern Ireland, commonly known as the United Kingdom (UK), is a sovereign state in Europe. Lying off the north-western coast of the European mainland, it includes the island of Great Britain—a term which also covers the whole country—plus the north-eastern part of the island of Ireland and many smaller islands.<sup>[8]</sup> Northern Ireland is the only part of the United Kingdom that shares a land border with another state—the Republic of Ireland.<sup>[nb 6]</sup> Apart from this land border, the UK is surrounded by the North Sea to its east, the English Channel to its south and the Celtic Sea to its south-southwest. The Irish Sea lies between Northern Ireland and the island of Great Britain. The United Kingdom has an area of 93,800 square miles (243,000 km<sup>2</sup>), making it the 80th-largest sovereign state in the world.

The United Kingdom is the 22nd-most populous country, with an estimated 64.5 million inhabitants.<sup>[4]</sup> It is a constitutional monarchy and a unitary parliamentary state of governance.<sup>[9][10]</sup> Its capital city is London, an important global city and financial centre with 10,310,000, the fourth-largest in Europe and second-largest in the European Union.<sup>[11]</sup> The current monarch is Queen Elizabeth II. The UK consists of four countries: England, Scotland, Wales, and Northern Ireland.<sup>[12]</sup> The latter three each with varying powers,<sup>[14][15]</sup> based in their capitals, Edinburgh, Cardiff, and Belfast, respectively. The nearby Isle of Wight and the Bailiwick of Jersey are not part of the United Kingdom, being Crown dependencies with the British Government's international representation.<sup>[16]</sup>

The relationships among the countries of the United Kingdom have changed over time. Wales was annexed by the Acts of Union of 1536 and 1543. A treaty between England and Scotland resulted in 1707 in a unified Kingdom of Great Britain, with the Kingdom of Ireland to form the United Kingdom of Great Britain and Ireland. In 1922, five-sixths of Ireland left the United Kingdom to form the Irish Free State, leaving Northern Ireland in the United Kingdom.<sup>[nb 7]</sup> The UK has fourteen Crown dependencies.

# Surface form dictionary construction from Wikipedia

- Page title
- Redirect pages
- **Disambiguation pages**
  - List of entities that share the same name



The screenshot shows a web browser window displaying the Wikipedia disambiguation page for "United Kingdom". The URL in the address bar is [en.wikipedia.org](https://en.wikipedia.org). The page title is "United Kingdom (disambiguation)". The sidebar on the left contains links such as Main page, Contents, Featured content, Current events, Random article, Donate to Wikipedia, Wikipedia store, Help, About Wikipedia, Community portal, Recent changes, Contact page, and Tools. The main content area includes a DNA helix icon and the text "European Science Photo Competition 2015". Below the header, it says "From Wikipedia, the free encyclopedia". The main text block discusses the United Kingdom as a sovereign state and lists various entities that share the name "United Kingdom". A "See also" section at the bottom lists related topics like Terminology of the British Isles, United Kingdoms, and British (disambiguation).

United Kingdom (disambiguation)

From Wikipedia, the free encyclopedia

The **United Kingdom** is a sovereign state located off the north-western coast of continental Europe.

United Kingdom may also refer to:

- [United Kingdom \(album\)](#)
- [Political union](#), a state formed out of smaller states
  - [United Kingdom of Great Britain and Ireland](#), the United Kingdom from 1801 to 1922
  - [United Kingdom of Great Britain](#), a sovereign state from 1707 to 1800
  - [United Kingdom of England](#), a state from the 10th century to 1707
  - [Kingdom of the Netherlands](#), a sovereign state with territory in western Europe and the Caribbean
    - [United Kingdom of the Netherlands](#), the Kingdom of the Netherlands from 1815 to 1830
  - [United Kingdom of Libya](#), a sovereign state from 1951 to 1969
  - [Kingdom of Portugal](#), a state from 1139 to 1910
    - [United Kingdom of Portugal, Brazil and the Algarves](#), the Kingdom of Portugal from 1815 to 1825

See also [edit]

- [Terminology of the British Isles](#)
- [United Kingdoms](#)
- [British \(disambiguation\)](#)

# Surface form dictionary construction from Wikipedia

- Page title
- Redirect pages
- Disambiguation pages
- **Anchor texts**
  - of links pointing to the entity's Wikipedia page

The screenshot shows a Mac OS X desktop environment with a window for the English Wikipedia article "Ireland". The window title bar says "en.wikipedia.org". The main content area displays the "Ireland" article, which starts with a brief introduction about the island in Europe. Below the intro, there is a detailed paragraph about Ireland's political division, mentioning the Republic of Ireland, Northern Ireland, and the United Kingdom. A red rectangular box highlights the word "United Kingdom". The sidebar on the left contains a navigation menu with links like Main page, Contents, Featured content, and various interaction and tools options.

Ireland

From Wikipedia, the free encyclopedia

This article is about the island in Europe. For the sovereign state of the same name, [Ireland](#). For other uses, see [Ireland \(disambiguation\)](#).

**Ireland** (Irish: *Éire* [eːɾə] (listen); Ulster-Scots: *Airlann* [ɑːrlən]) is an island separated from Great Britain to its east by the North Channel, the Irish Sea, and St George's Channel. It is the second-largest island of the British Isles, the third-largest in Europe, and the twentieth-largest island in the world. Politically, Ireland is divided between the Republic of Ireland (officially named Ireland), which covers most of the island, and Northern Ireland, which is part of the United Kingdom, in the northeast corner. The population of Ireland was about 6.4 million, ranking it the second-most populous island in the world. Just under 4.6 million live in the Republic of Ireland and just over 1.8 million live in Northern Ireland. The island's geography comprises relatively low-lying mountains surrounding a central plain, with rivers extending inland. The island has lush vegetation, a product of its mild but changeable climate. Thick woodlands covered the island until the Middle Ages. As that is wooded in Ireland is about 11% of the total, compared with a European average of 20%. Ireland has twenty-six extant mammal species native to Ireland.<sup>[10]</sup> The Irish climate is very moderate, with temperatures ranging from 5°C to 15°C. As a result, winters are milder than expected for such a northerly area. However, those in Continental Europe. Rainfall and cloud cover are abundant. The earliest evidence of human presence in Ireland is dated at 10,500 BC.<sup>[12]</sup> Gaelic Ireland

# Surface form dictionary construction from Wikipedia

- Page title
- Redirect pages
- Disambiguation pages
- Anchor texts
- **Bold texts from first paragraph**
  - generally denote other name variants of the entity



## Surface form dictionary construction from other sources

- Anchor texts from external web pages pointing to Wikipedia articles
- Problem of *synonym discovery*
  - Expanding acronyms
  - Leveraging search results or query-click logs from a web search engine
  - ...

## Filtering mentions

- Objective is to filter our mentions that are unlikely to be linked to any entity
- **Keyphraseness**

$$P(\text{keyphrase}|m) = \frac{|D_{\text{link}}(m)|}{|D(m)|}$$

- $|D_{\text{link}}(m)|$  is the number of Wikipedia articles where  $m$  appears as an anchor text of a link
- $|D(m)|$  is the number of Wikipedia articles that contain  $m$

## Filtering mentions (cont'd)

- **Link probability**

$$P(\text{link}|m) = \frac{\text{link}(m)}{\text{freq}(m)}$$

- $\text{link}(m)$  is the number of times mention  $m$  appears as an anchor text of a link
- $\text{freq}(m)$  is the total number of times mention  $m$  occurs in Wikipedia (as a link or not)

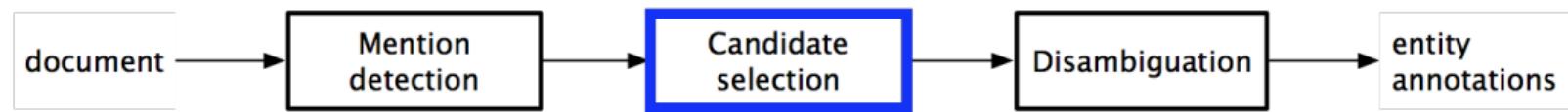
## Overlapping entity mentions

- Dealing with them in this phase
  - E.g., by dropping a mention if it is subsumed by another mention
- Keeping them and postponing the decision to a later stage (candidate selection or disambiguation)

# Entity linking

- Mention detection
- Candidate selection
- Disambiguation
- Evaluation

## Candidate selection



## Candidate selection

- Goal: Narrow down the space of disambiguation possibilities
- Balances between precision and recall (effectiveness vs. efficiency)
- Often approached as a ranking problem
  - Keeping only candidates above a score/rank threshold for downstream processing

## Commonness

- Perform the ranking of candidate entities based on their overall popularity, i.e., “most common sense”

$$P(e|m) = \frac{n(m, e)}{\sum_{e' \in \mathcal{E}} n(m, e')}$$

- $n(m, e)$  the number of times entity  $e$  is the link destination of mention  $m$
- Can be pre-computed and stored in the entity surface form dictionary
- Follows a power law with a long tail of extremely unlikely senses; entities at the tail end of the distribution can be safely discarded
  - E.g., 0.001 is a sensible threshold

## Example

Entity	Commonness
$e$	$P(e m)$

Times_Square	0.940
Times_Square_(film)	0.017
Times_Square_(Hong_Kong)	0.011
Times_Square_(IRT_42nd_Street_Shuttle)	0.006
...	...



Home to the **Empire State Building**, **Times Square**, **Statue of Liberty** and other iconic sites,  
**New York City** is a fast-paced, globally influential center of art, culture, fashion and finance.

## Example #2

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 ...

Entity	Commonness
FIFA_World_Cup	0.2358
FIS_Apline_Ski_World_Cup	0.0682
2009_FINA_Swimming_World_Cup	0.0633
World_Cup_(men's_golf)	0.0622
...	

Entity	Commonness
1998_FIFA_World_Cup	0.9556
1998_IAAF_World_Cup	0.0296
1998_Alpine_Skiing_World_Cup	0.0059
...	

Entity	Commonness
Germany	0.9417
Germany_national_football_team	0.0139
Nazi_Germany	0.0081
German_Empire	0.0065
...	

- Commonness works in many of the cases, but not in all
- Other entities help to disambiguate which entity is being referred to

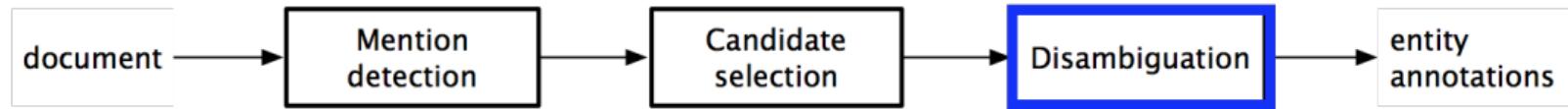
## Exercise

### E7-1 Entity linking

# Entity linking

- Mention detection
- Candidate selection
- Disambiguation
- Evaluation

# Disambiguation



# Disambiguation

- Baseline approach: most common sense
- Consider additional types of evidence
  - **Prior importance** of entities and mentions
  - **Contextual similarity** between the text surrounding the mention and the candidate entity
  - **Coherence** among all entity linking decisions in the document
- Combine these signals
  - Using supervised learning or graph-based approaches
- Optionally perform pruning
  - Reject low confidence or semantically meaningless annotations

## Prior importance features

- **Context-independent features**
  - Neither the text nor other mentions in the document are taken into account
- Keyphraseness
- Link probability
- Commonness

## Prior importance features (cont'd)

- **Link prior**

- Popularity of the entity measured in terms of incoming links

$$P_{link}(e) = \frac{|\mathcal{L}_e|}{\sum_{e' \in \mathcal{E}} |\mathcal{L}_{e'}|}$$

- $|\mathcal{L}_e|$  is the total number of incoming links entity  $e$  has

- **Page views**

- Popularity of the entity measured in terms traffic volume

$$P_{pageviews}(e) = \frac{\text{pageviews}(e)}{\sum_{e' \in \mathcal{E}} \text{pageviews}(e')}$$

- $\text{pageviews}(e)$  is the total number of page views (measured over a certain time period)

## Contextual features

- Compare the surrounding *context* of a mention with the (textual) representation of the given candidate entity
- Context of a mention
  - Window of text (sentence, paragraph) around the mention
  - Entire document
- Entity's representation
  - Wikipedia entity page, first description paragraph, terms with highest TF-IDF score, etc.
  - Entity's description in the knowledge base

## Contextual similarity

- Commonly: bag-of-words representation
- **Cosine similarity**

$$sim_{cos}(m, e) = \frac{\vec{d}_m \cdot \vec{d}_e}{\|\vec{d}_m\| \|\vec{d}_e\|}$$

- Many other options for measuring similarity
  - Dot product, KL divergence, Jaccard similarity
- Representation does not have to be limited to bag-of-words
  - Concept vectors (named entities, Wikipedia categories, anchor text, keyphrases, etc.)

## Entity-relatedness features

- It can reasonably be assumed that a document focuses on one or at most a few topics
- Therefore, entities mentioned in a document should be topically related to each other
- Capturing *topical coherence* by developing some measure of *relatedness* between (linked) entities
  - Defined for pairs of entities

## Wikipedia Link-based Measure (WSM)

- Often referred to simply as *relatedness*
- A close relationship is assumed between two entities if there is a large overlap between the entities linking to them

$$WLM(e, e') = 1 - \frac{\log(\max(|\mathcal{L}_e|, |\mathcal{L}_{e'}|)) - \log(|\mathcal{L}_e \cap \mathcal{L}_{e'}|)}{\log(|\mathcal{E}|) - \log(\min(|\mathcal{L}_e|, |\mathcal{L}_{e'}|))}$$

- $\mathcal{L}_e$  is the set of entities that link to  $e$
- $|\mathcal{E}|$  is the total number of entities

# Wikipedia Link-based Measure (WSM)

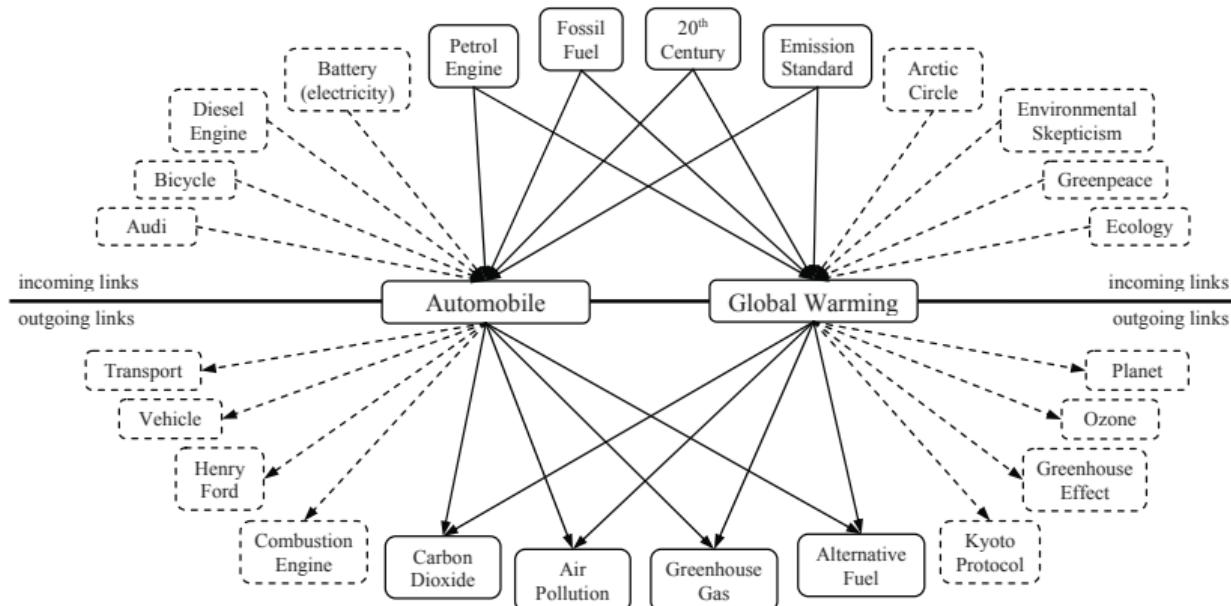


Figure: Image taken from Milne and Witten (2008). An Effective, Low-Cost Measure of Semantic Relatedness Obtained from Wikipedia Links. In: AAAI WikiAI Workshop.

## Entity-relatedness features

- Numerous ways to define relatedness
  - Consider not only incoming, but also outgoing links or the union of incoming and outgoing links
  - Jaccard similarity, Pointwise Mutual Information (PMI), or the Chi-square statistic, etc.
- A relatedness function does not have to be symmetric
  - E.g., the relatedness of the UNITED STATES given NEIL ARMSTRONG is intuitively larger than the relatedness of NEIL ARMSTRONG given the UNITED STATES
  - **Conditional probability**

$$P(e'|e) = \frac{|\mathcal{L}_{e'} \cap \mathcal{L}_e|}{|\mathcal{L}_e|}$$

- Having a single relatedness function is preferred, to keep the disambiguation process simple
- Various relatedness measures can effectively be combined into a single score using a machine learning approach

# Disambiguation approaches

- Consider *local compatibility* (including prior evidence) and *coherence* with the other entity linking decisions
- Overall objective function:

$$\Gamma^* = \arg \max_{\Gamma} \left( \sum_{(m,e) \in \Gamma} \phi(m, e) + \psi(\Gamma) \right)$$

- $\phi(m, e)$  is the local compatibility between the mention and the assigned entity
- $\psi(\Gamma)$  is the coherence function for all entity annotations in the document
- $\Gamma$  is a solution (set of mention-entity pairs)
- **This optimization problem is NP-hard!**
  - Need to resort to approximation algorithms and heuristics

# Disambiguation strategies

- **Individually**, one-mention-at-a-time
  - Rank candidates for each mention, take the top ranked one (or NIL)
  - Interdependence between entity linking decisions may be incorporated in a pairwise fashion

$$\Gamma(m) = \arg \max_{e \in \mathcal{E}_m} score(e, m)$$

- **Collectively**, all mentions in the document jointly

## Disambiguation approaches

<b>Approach</b>	<b>Context</b>	<b>Entity interdependence</b>
Most common sense	none	none
Individual local disambiguation	text	none
Individual global disambiguation	text & entities	pairwise
Collective disambiguation	text & entities	collective

## Individual local disambiguation

- Early entity linking approaches
- Local compatibility score can be written as a linear combination of features

$$\phi(e, m) = \sum_i \lambda_i f_i(e, m)$$

- $f_i(e, m)$  can be either a context-independent or a context-dependent feature
- Learn the “optimal” combination of features from training data using machine learning

## Individual global disambiguation

- Consider what other entities are mentioned in the document
- True global optimization would be NP-hard
- Good approximation can be computed efficiently by considering pairwise interdependencies for each mention independently
  - Pairwise entity relatedness scores need to be aggregated into a single number (how coherent the given candidate entity is with the rest of the entities in the document)

## TAGME (Ferragina & Scaiella, 2010)

- Combine the two most important features (*commonness* and *relatedness*) using a voting scheme
- The score of a candidate entity for a particular mention:

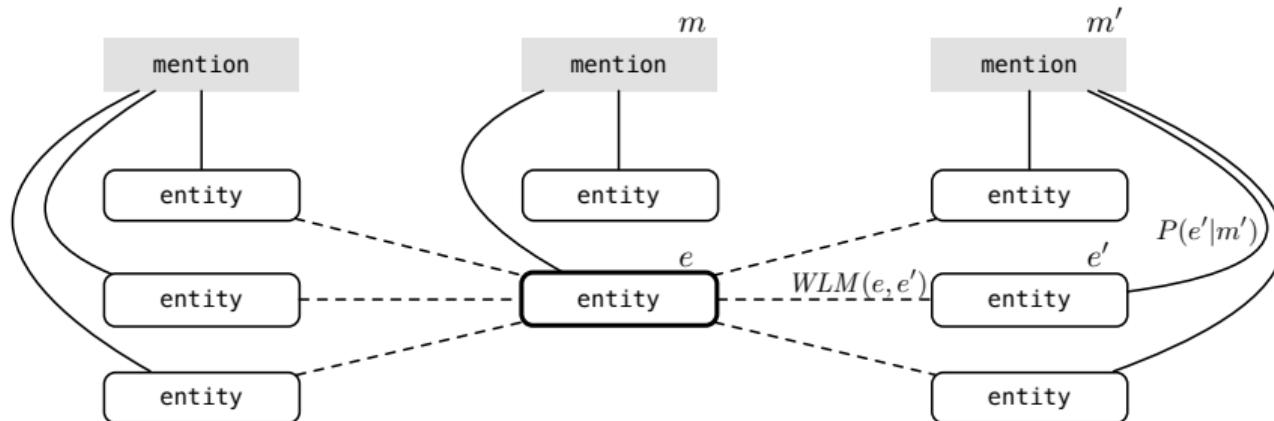
$$score(e, m) = \sum_{\substack{m' \in \mathcal{M}_d \\ m' \neq m}} vote(m', e)$$

- The vote function estimates the agreement between  $e$  and all candidate entities of all other mentions in the document

# TAGME (voting mechanism)

- Average relatedness between each possible disambiguation, weighted by its commonness score

$$vote(m', e) = \frac{\sum_{e' \in \mathcal{E}_{m'}} WLM(e, e') P(e'|m')}{|\mathcal{E}_{m'}|}$$



## TAGME (final score)

- Final decision uses a simple but robust heuristic
  - The top entities with the highest score are considered for a given mention and the one with the highest commonness score is selected

$$\Gamma(m) = \arg \max_{e \in \mathcal{E}_m} \{P(e|m) : e \in \text{top}_\epsilon[\text{score}(e, m)]\}$$

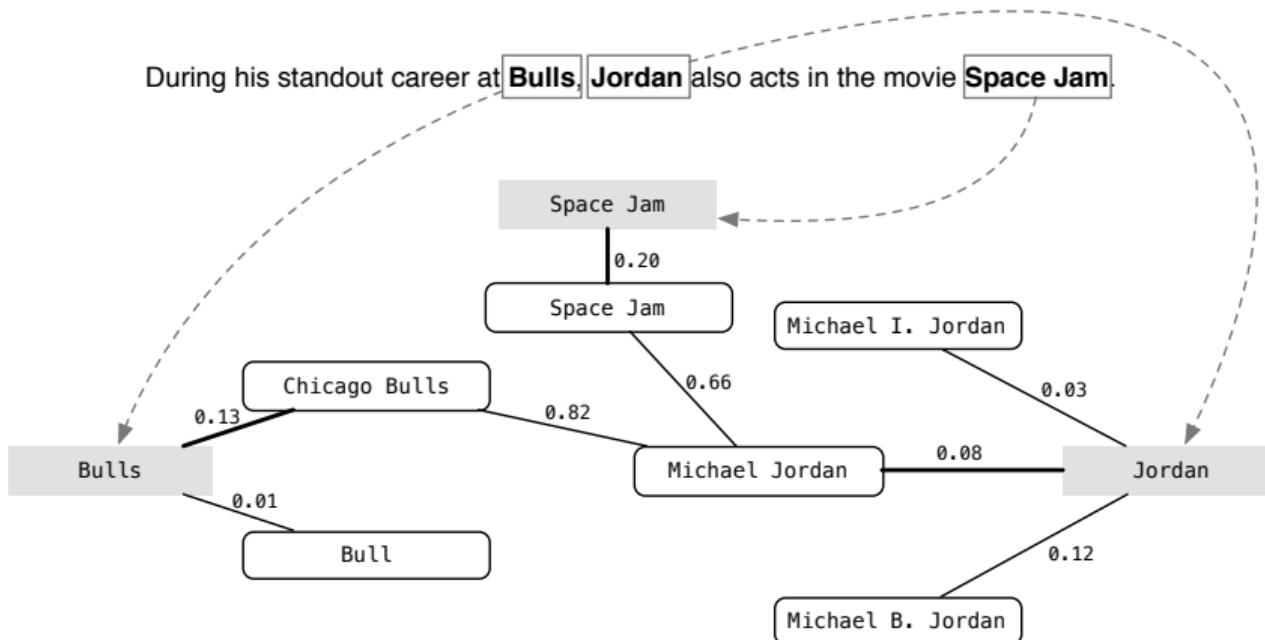
- Note that *score* merely acts as a filter
  - Only entities in the top  $\epsilon$  percent of the scores are retained ( $\epsilon = 0.3$ )
  - Out of the remaining entities, the most common sense of the mention will be finally selected

## Collective disambiguation

- Graph-based representation
- **Mention-entity edges** capture the local compatibility between the mention and the entity
  - Measured using a combination of context-independent and context-dependent features
- **Entity-entity edges** represent the semantic relatedness between a pair of entities
  - Common choice is *relatedness* (WLM)
- Use these relations jointly to identify a single referent entity (or none) for each of the mentions

# Example

During his standout career at **Bulls**, **Jordan** also acts in the movie **Space Jam**.



## AIDA (Hoffart et al., 2011)

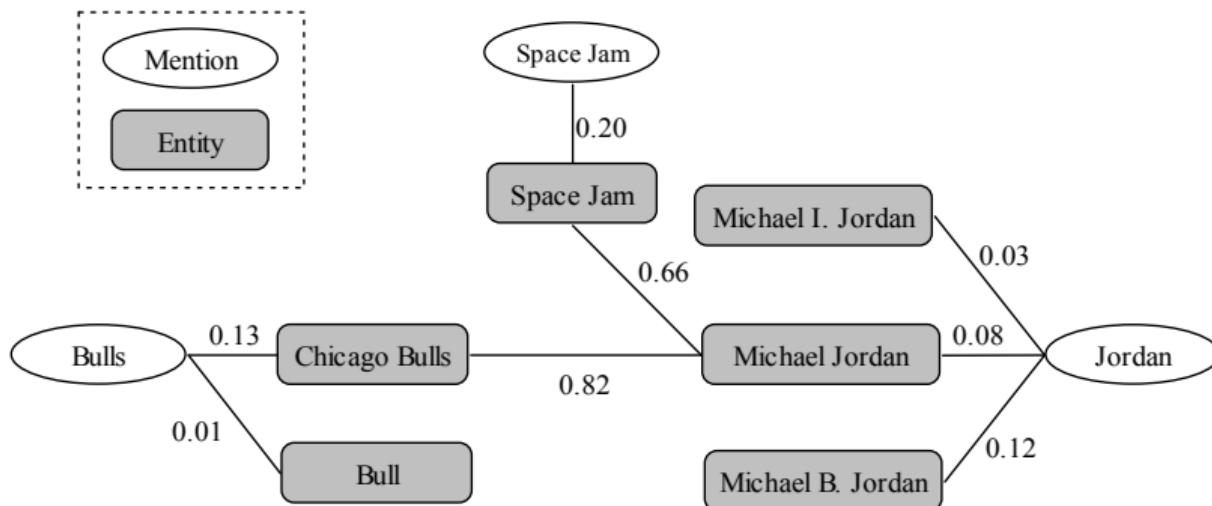
- Problem formulation: find a dense subgraph that contains all mention nodes and exactly one mention-entity edge for each mention
- Greedy algorithm iteratively removes edges

## AIDA algorithm

- Start with the full graph
- Iteratively remove the entity node with the lowest *weighted degree* (along with all its incident edges), provided that each mention node remains connected to at least one entity
  - Weighted degree of an entity node is the sum of the weights of its incident edges
- The graph with the highest *density* is kept as the solution
  - The density of the graph is measured as the minimum weighted degree among its entity nodes

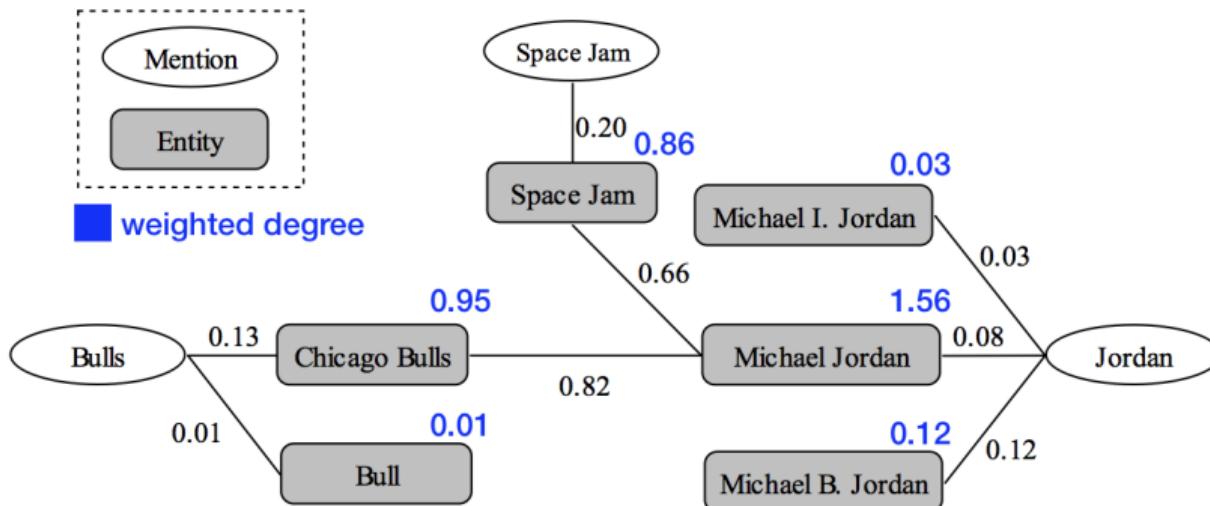
## Example iteration #1

- Which entity should be removed first?



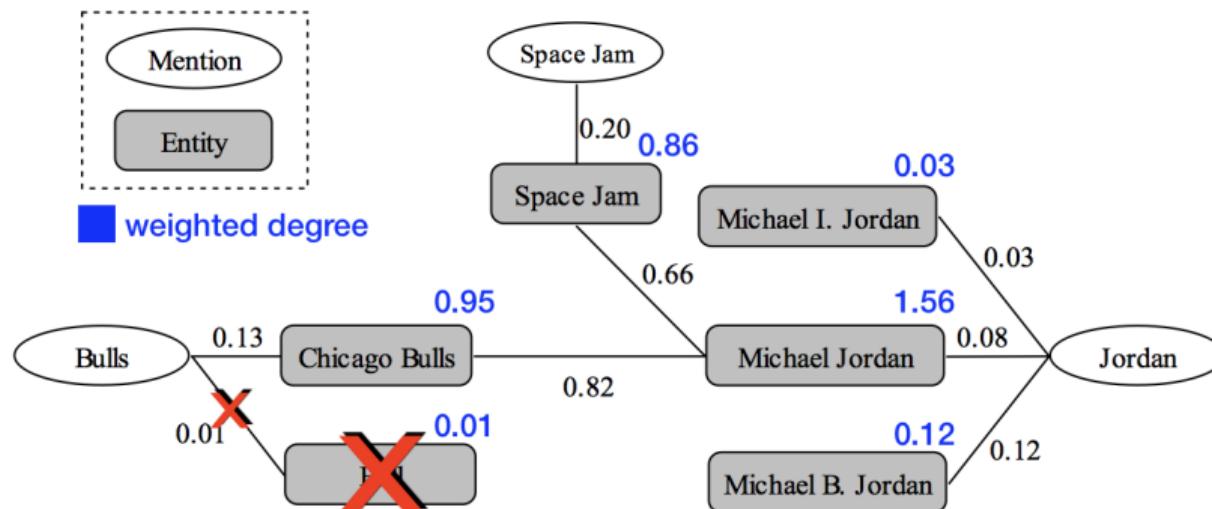
# Example iteration #1

- Which entity should be removed first?



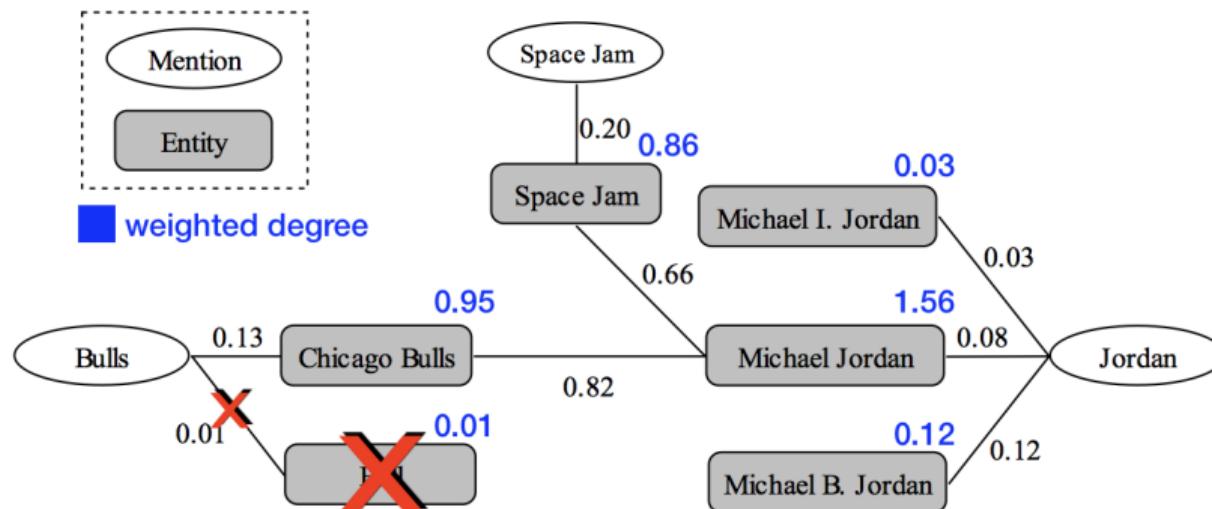
# Example iteration #1

- Which entity should be removed first?



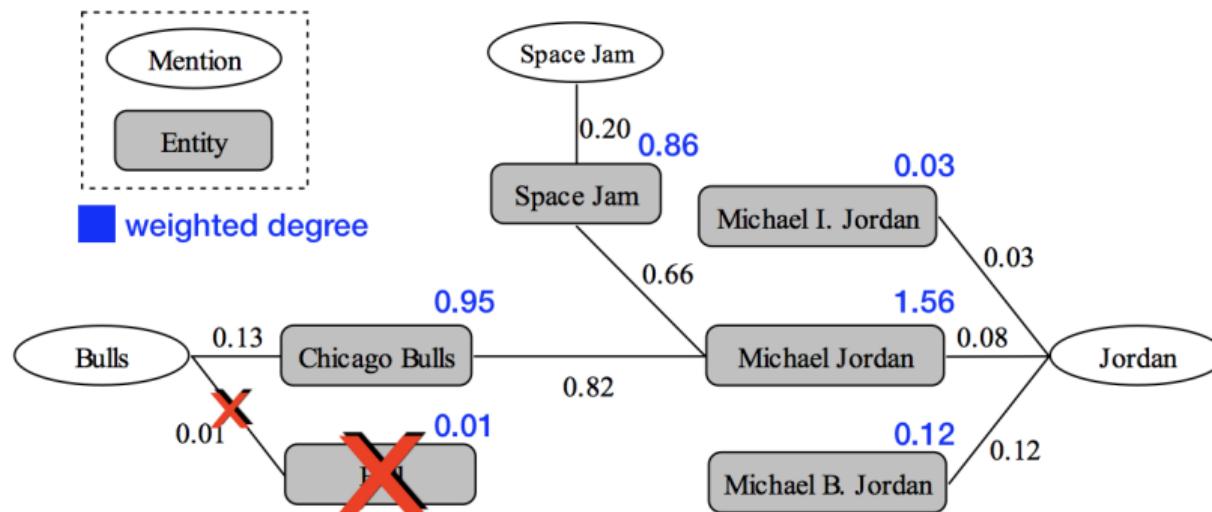
# Example iteration #1

- What is the density of the graph?



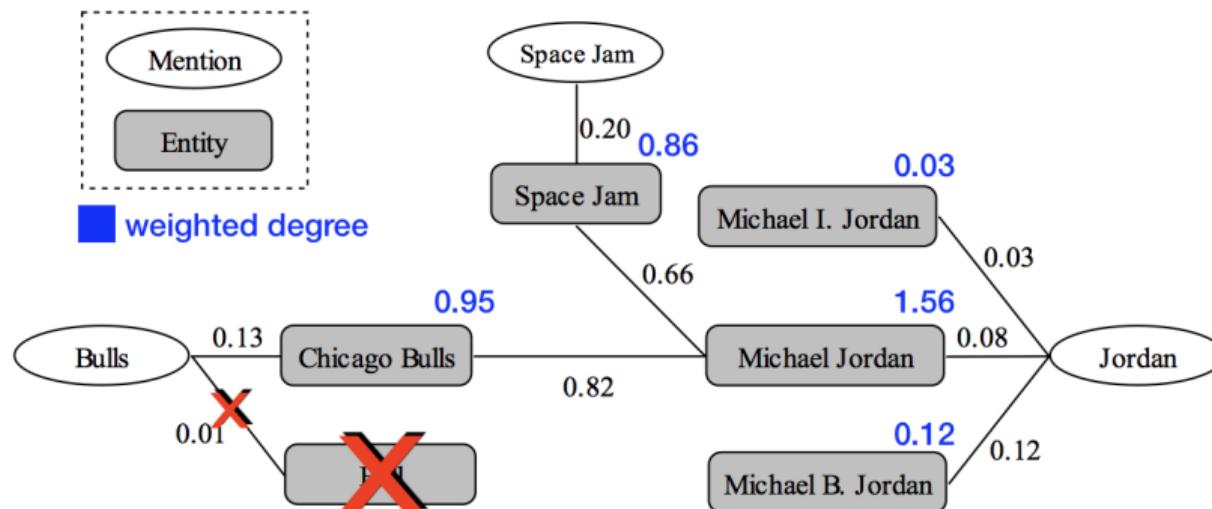
## Example iteration #1

- What is the density of the graph? **0.03**



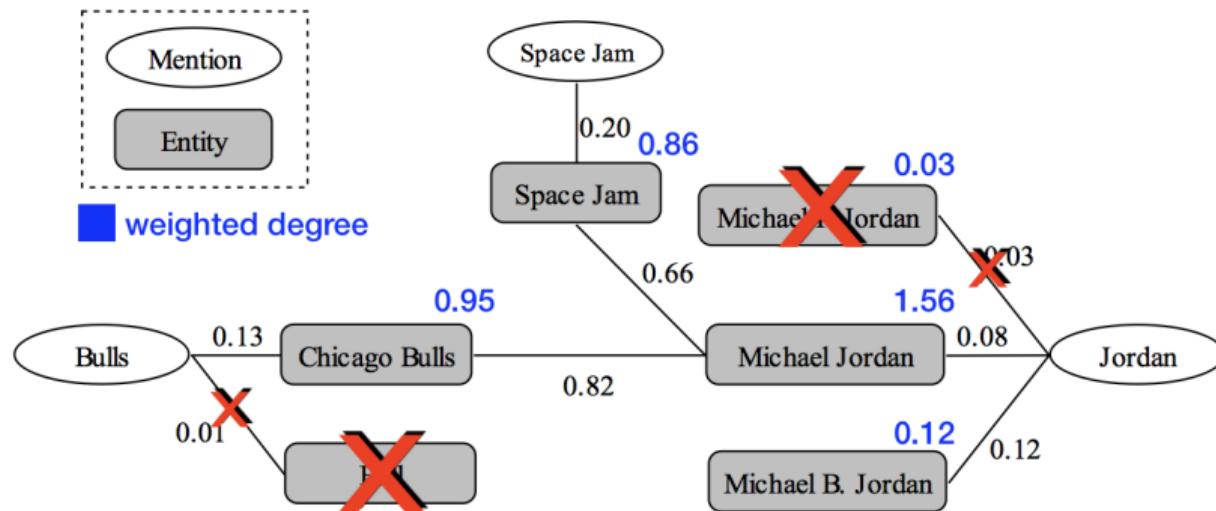
## Example iteration #2

- Which entity should be removed next?



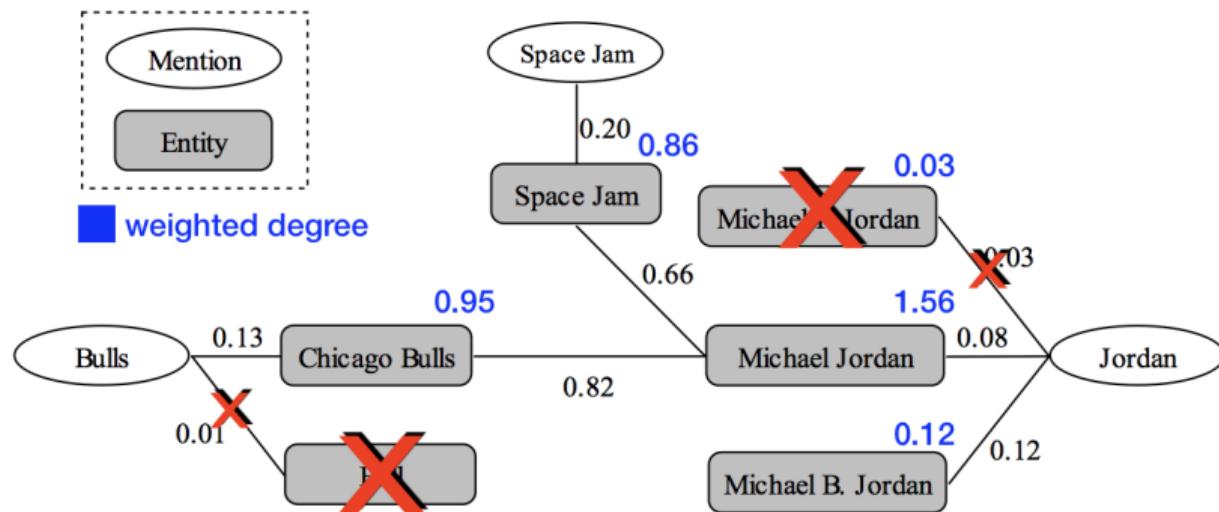
## Example iteration #2

- Which entity should be removed next?



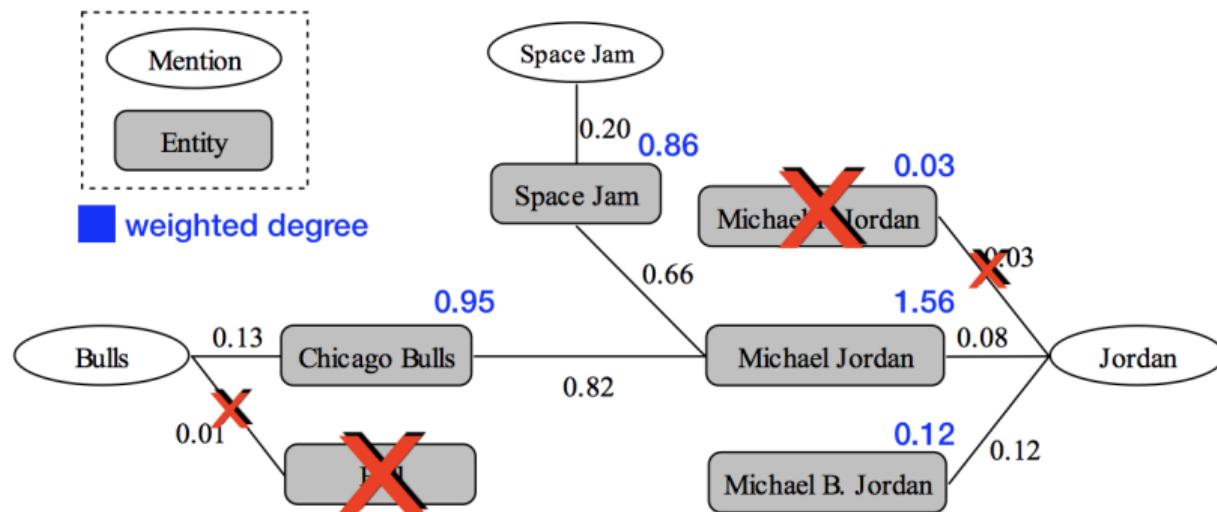
## Example iteration #2

- What is the density of the graph?



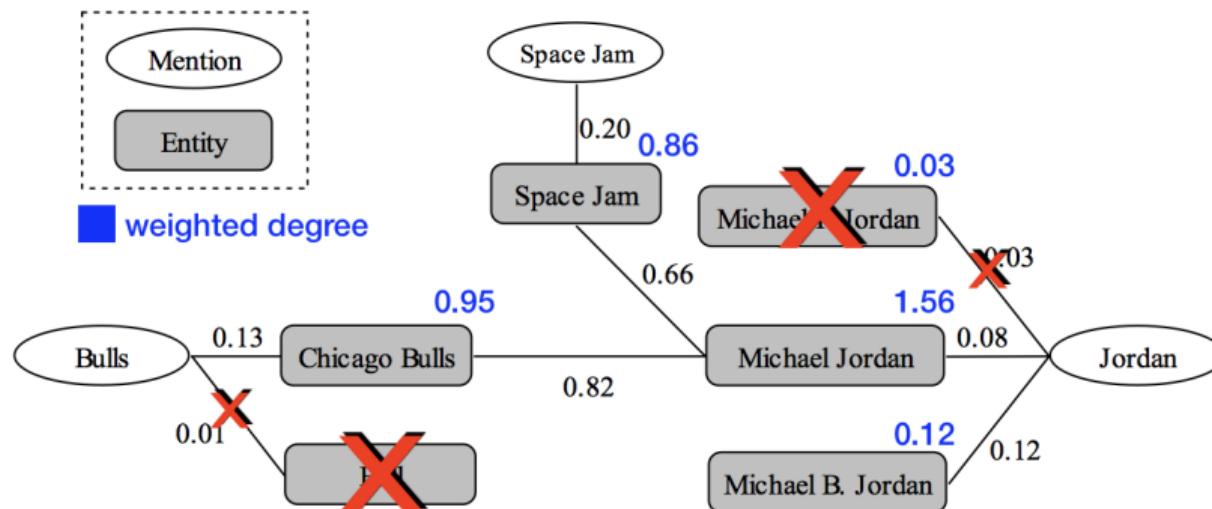
## Example iteration #2

- What is the density of the graph? **0.12**



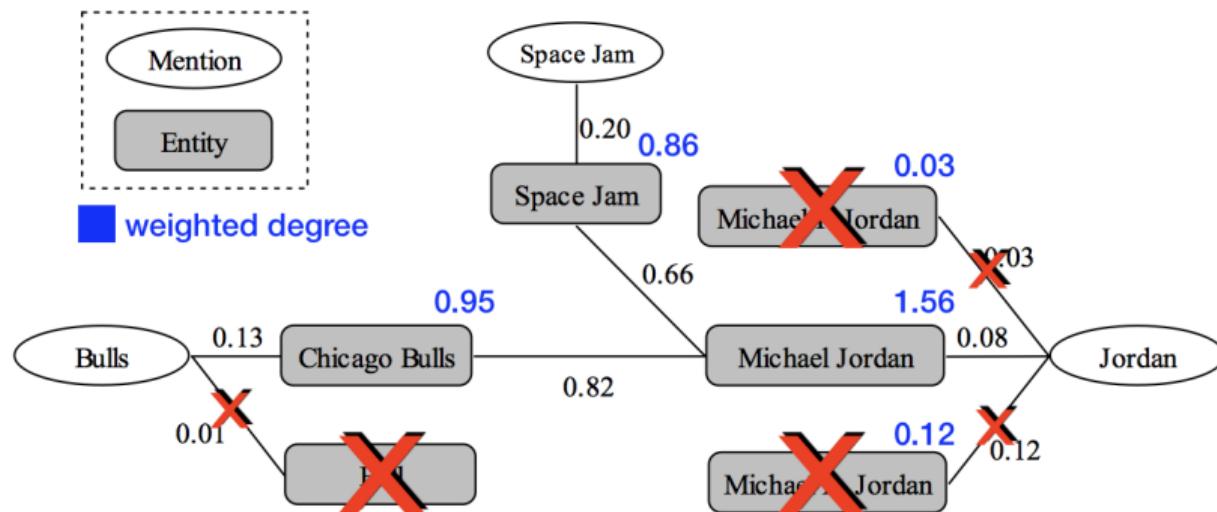
## Example iteration #3

- Which entity should be removed next?



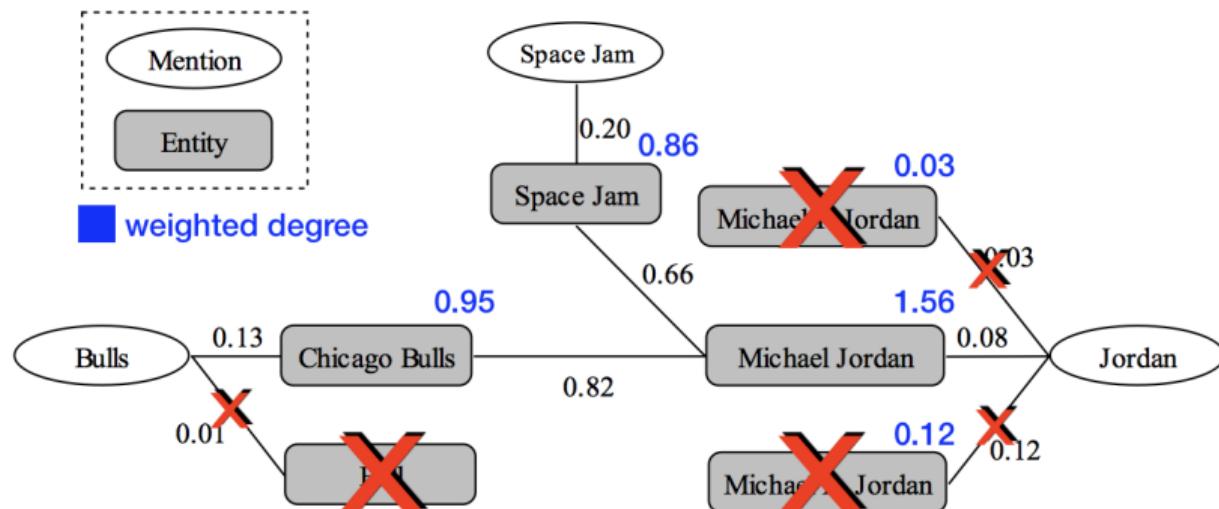
## Example iteration #3

- Which entity should be removed next?



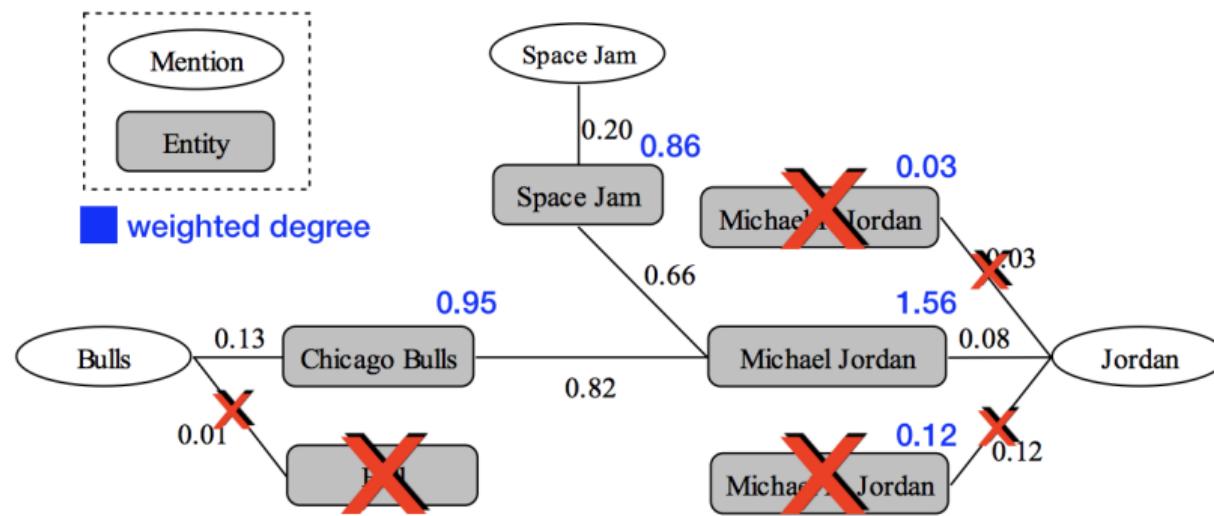
## Example iteration #3

- What is the density of the graph?



## Example iteration #3

- What is the density of the graph? **0.86**



## AIDA pre- and post-processing

- Pre-processing phase: remove entities that are “too distant” from the mention nodes
- At the end of the iterations, the solution graph may still contain mentions that are connected to more than one entity; deal with this in post-processing
  - If the graph is sufficiently small, it is feasible to exhaustively consider all possible mention-entity pairs
  - Otherwise, a faster local (hill-climbing) search algorithm may be used

# Pruning

- Discarding meaningless or low-confidence annotations produced by the disambiguation phase
- Simplest solution: use a confidence threshold
- More advanced solutions
  - Machine learned classifier to retain only entities that are “relevant enough” (human editor would annotate them)
  - Optimization problem: decide, for each mention, whether switching the top ranked disambiguation to NIL would improve the objective function

# Entity linking

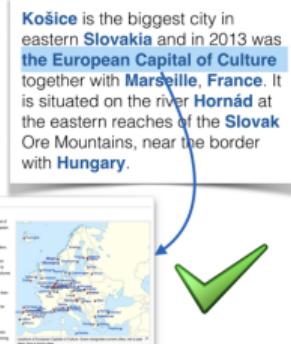
- Mention detection
- Candidate selection
- Disambiguation
- Evaluation

## Evaluation (end-to-end)

- Comparing the system-generated annotations against a human-annotated gold standard
- Evaluation criteria
  - **Perfect match:** both the linked entity and the mention offsets must match
  - **Relaxed match:** the linked entity must match, it is sufficient if the mention overlaps with the gold standard

# Evaluation with relaxed match

## Example #1

ground truth	system annotation
<p>Košice is the biggest city in eastern Slovakia and in 2013 was the European Capital of Culture together with Marseille, France. It is situated on the river Hornád at the eastern reaches of the Slovak Ore Mountains, near the border with Hungary.</p> 	<p>Košice is the biggest city in eastern Slovakia and in 2013 was the European Capital of Culture together with Marseille, France. It is situated on the river Hornád at the eastern reaches of the Slovak Ore Mountains, near the border with Hungary.</p>  

## Example #2

ground truth	system annotation
<p>Košice is the biggest city in eastern Slovakia and in 2013 was the European Capital of Culture together with Marseille, France. It is situated on the river Hornád at the eastern reaches of the Slovak Ore Mountains, near the border with Hungary.</p>  <p>Slovak Ore Mountains Mountain range in Slovakia</p>   	<p>Košice is the biggest city in eastern Slovakia and in 2013 was the European Capital of Culture together with Marseille, France. It is situated on the river Hornád at the eastern reaches of the Slovak Ore Mountains, near the border with Hungary.</p>  <p>Slovakia Country in Europe</p>   

## Evaluation metrics

- Set-based metrics
  - **Precision:** fraction of correctly linked entities that have been annotated by the system
  - **Recall:** fraction of correctly linked entities that should be annotated
  - **F-measure:** harmonic mean of precision and recall
- Metrics are computed over a collection of documents
  - Micro-averaged: aggregated across mentions
  - Macro-averaged: aggregated across documents

# Evaluation metrics

- **Micro-averaged**

$$P_{mic} = \frac{|\mathcal{A}_{\mathcal{D}} \cap \hat{\mathcal{A}}_{\mathcal{D}}|}{|\mathcal{A}_{\mathcal{D}}|}$$

$$R_{mic} = \frac{|\mathcal{A}_{\mathcal{D}} \cap \hat{\mathcal{A}}_{\mathcal{D}}|}{|\hat{\mathcal{A}}_{\mathcal{D}}|}$$

- $\mathcal{A}_{\mathcal{D}}$  include all annotations for a set  $\mathcal{D}$  of documents
- $\hat{\mathcal{A}}_{\mathcal{D}}$  is the collection of reference annotations for  $\mathcal{D}$

- **Macro-averaged**

$$P_{mac} = \frac{1}{|\mathcal{D}|} \sum_{d \in \mathcal{D}} \frac{|\mathcal{A}_d \cap \hat{\mathcal{A}}_d|}{|\mathcal{A}_d|}$$

$$R_{mac} = \frac{1}{|\mathcal{D}|} \sum_{d \in \mathcal{D}} \frac{|\mathcal{A}_d \cap \hat{\mathcal{A}}_d|}{|\hat{\mathcal{A}}_d|}$$

- $\mathcal{A}_d$  are the annotations generated by the entity linking system
- $\hat{\mathcal{A}}_d$  denote the reference (ground truth) annotations for a single document  $d$

- **F1 score**

$$F1 = \frac{2 P R}{P + R}$$

## Component-based evaluation

- The pipeline architecture makes the evaluation of entity linking systems especially challenging
  - The main focus is on the disambiguation component, but its performance is largely influenced by the preceding steps
- Fair comparison between two approaches can only be made if they share all other elements of the pipeline

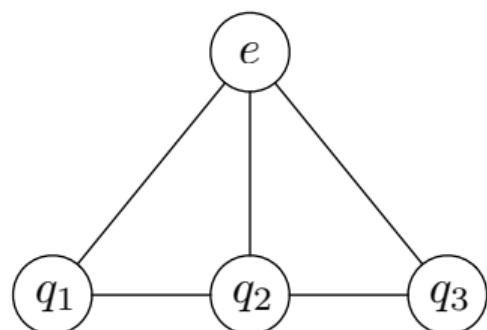
## Exercise

### E7-2 Entity linking evaluation

## Entity linking and retrieval

## Recap - Entity retrieval

- Two instantiations of Markov random field (MRF) models for modeling term dependence
  - SDM - Unstructured retrieval model
  - FSDM - Fielded retrieval model
- The approaches use unigram and bigram matches between terms
  - So far it is all term based.



# Semantically enriched entity retrieval

- Working definition: semantics == structure
- Goal: enrich term-based representations by leveraging structured information about entities
- Important: this semantic enrichment needs to happen both on the query and on the entity side, as well as reflected in the matching (scoring function)

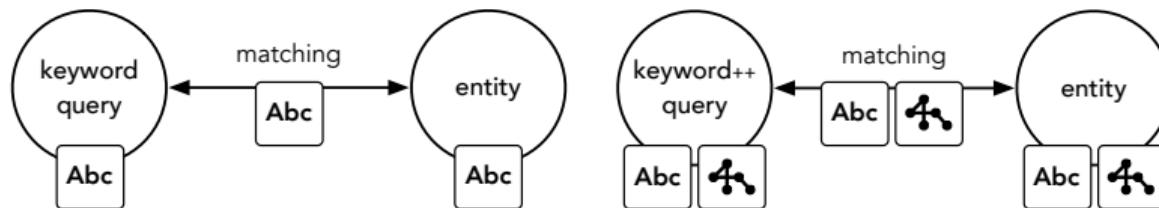


Figure: (Left) Ranking entities using term-based representations. (Right) Semantically enriched entity ranking by incorporating structure. Illustration is taken from Balog (2018) [Fig 4.1].

## Question

How to incorporate entity linking information in entity retrieval?

# Dual term-based and entity-based entity representations

Idea: preserve these entity references by employing a *dual entity representation*: on top of the traditional term-based representation, each entity is additionally represented by means of its associated entities.

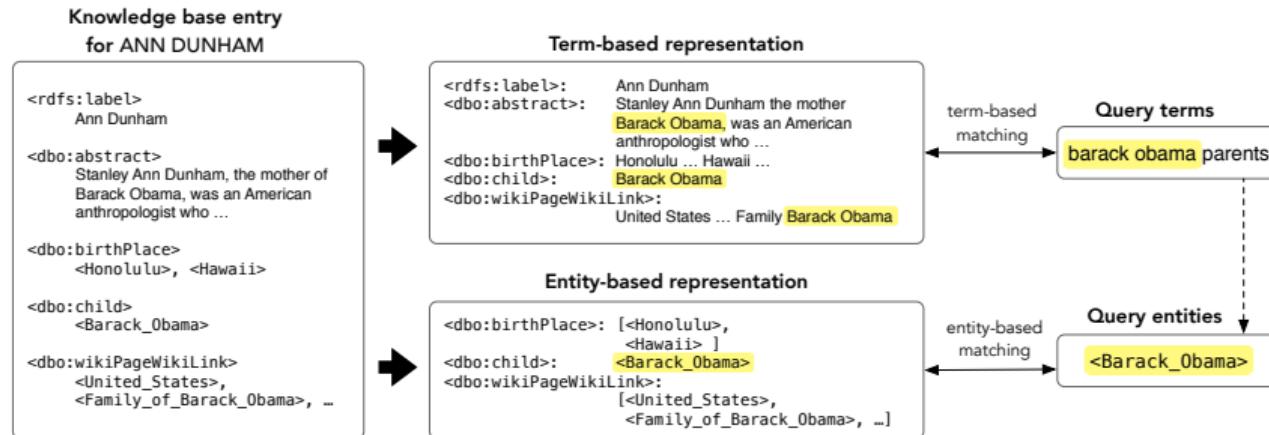
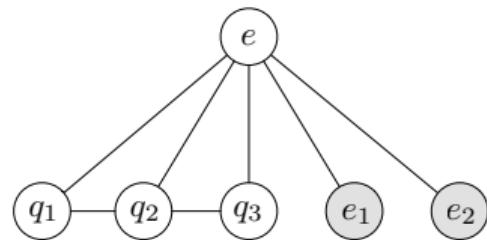


Figure: Illustration is based on (Hasibi et al., 2016).

# Entity linking incorporated retrieval (ELR)

- Entity-specific extension applied on top of the MRF framework
  - May be applied on top of any term-based retrieval model that can be instantiated in the MRF framework, but here we will focus on SDM
- The underlying graph representation of SDM is extended with query entity nodes
  - It is assumed that entities have already been identified (linked) in queries and in the descriptions of entities
- New type of clique: 2-cliques between the given entity (that is being scored) and the query entities



## Ranking function SDM+ELR

- The SDM+ELR ranking function is given by a weighted combination of four feature functions
  - Query terms ( $f_T$ )
  - Exact match of query bigrams ( $f_O$ )
  - Unordered match of query bigrams ( $f_U$ )
  - Entity matches ( $f_E$ )**

$$\begin{aligned} \text{score}(e, q) = & \lambda_T \sum_{i=1}^n f_T(q_i, e) + \lambda_O \sum_{i=1}^{n-1} f_O(q_i, q_{i+1}, e) \\ & + \lambda_U \sum_{i=1}^{n-1} f_U(q_i, q_{i+1}, e) + \lambda_E \sum_{i=1}^m f_E(e_i; e) \end{aligned}$$

- Issue: the number of entities ( $m$ ) varies per query! How can  $\lambda_E$  be trained?

## Parameterizing weights

- Rewriting  $\lambda$  parameters as parameterized functions over each clique:

$$\lambda_T(q_i) = \lambda_T \frac{1}{n} ,$$

$$\lambda_O(q_i, q_{i+1}) = \lambda_O \frac{1}{n-1} ,$$

$$\lambda_U(q_i, q_{i+1}) = \lambda_U \frac{1}{n-1} ,$$

$$\lambda_E(e_i) = \lambda_E \frac{w(e_i, q)}{\sum_{j=1}^m w(e_j, q)} .$$

- The weight  $w(e_i, q)$  reflects the confidence in that entity annotation (entity linker's confidence score)

## Ranking function SDM+ELR

- Final ranking function using the parameterized weights:

$$\begin{aligned} \text{score}(e, q) = & \frac{\lambda_T}{n} \sum_{i=1}^n f_T(q_i; e) \\ & + \frac{\lambda_O}{n-1} \sum_{i=1}^{n-1} f_O(q_i, q_{i+1}; e) \\ & + \frac{\lambda_U}{n-1} \sum_{i=1}^{n-1} f_U(q_i, q_{i+1}; e) \\ & + \frac{\lambda_{\mathcal{E}}}{\sum_{j=1}^m w(e_j, q)} \sum_{i=1}^m w(e_i, q) f_{\mathcal{E}}(e_i; e) \end{aligned}$$

- Default setting:  $\lambda_T = 0.8$ ,  $\lambda_O = 0.05$ ,  $\lambda_U = 0.05$ , and  $\lambda_{\mathcal{E}} = 0.1$

## Feature function for entity matches

- Differences from term-based scoring
  - We assume that each entity appears at most once in each field
  - If a query entity appears in a field, then it shall be regarded as a “perfect match,” independent of what other entities are present in the same field
  - Unlike for terms-based representations, predicate folding is not performed

$$f_{\mathcal{E}}(e_i; e) = \log \sum_{f \in \tilde{\mathcal{F}}} w_f^{\mathcal{E}} \left( (1 - \lambda) \mathbb{1}(e_i, f_{\tilde{e}}) + \lambda \frac{\sum_{e' \in \mathcal{E}} \mathbb{1}(e_i, f_{\tilde{e}'})}{|\{e' \in \mathcal{E} : f_{\tilde{e}'} \neq \emptyset\}|} \right)$$

- $\tilde{e}$  denote the entity-based representation of entity  $e$
- $\tilde{\mathcal{F}}$  denotes the set of fields in the entity-based representation
- $\mathbb{1}(e, f_{\tilde{e}})$  is a binary indicator function, which is 1 if  $e_i$  is present in the entity field  $f_{\tilde{e}}$  and otherwise 0
- $\lambda$  is the smoothing parameter (default: 0.1)
- Field weights  $w_f^{\mathcal{E}}$  may be set manually or via dynamic mapping using PRMS

## Exercise

E7-3 Entity linking incorporated retrieval

# Summary

- The entity linking task
- Canonical entity linking architecture and main components
- Mention detection
  - Entity surface form dictionary and its construction from Wikipedia
  - Filtering mentions (keyphraseness and link probability)
- Candidate selection (using commonness)
- Disambiguation
  - Prior importance features (incl. keyphraseness, link probability, commonness, link prior, page views)
  - Contextual features (contextual similarity and entity-relatedness features, incl. WSM, a.k.a. relatedness)
  - TAGME method
  - AIDA algorithm
- Evaluation with perfect/relaxed match, evaluation measures
- Dual entity representation, Entity Linking incorporated Retrieval (ELR) model

# Reading

- Entity-Oriented Search (Balog)
  - Chapter 4: Sections 4.1 and 4.2.2
  - Chapter 5, until 5.6 (inclusive) + Section 5.8.1