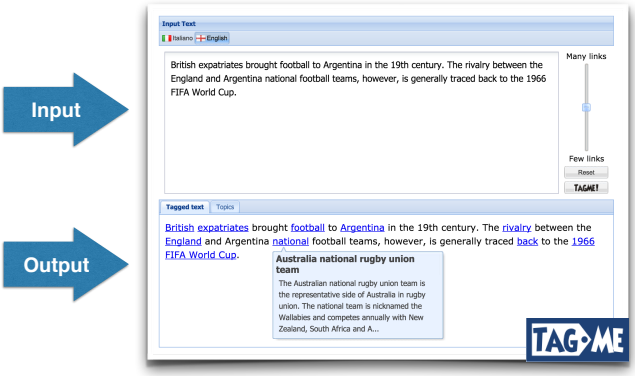
DAT630
Entity linking I.

08/11/2016

Faegheh Hasibi | University of Stavanger

What is entity linking?



What is entity linking?

British expatriates brought football to Argentina in the 19th century.
The rivalry between the England and Argentina national football teams, however, is generally traced back to the 1966 FIFA World Cup.

1966 FIFA World Cup | England national football team | Argentina national football team

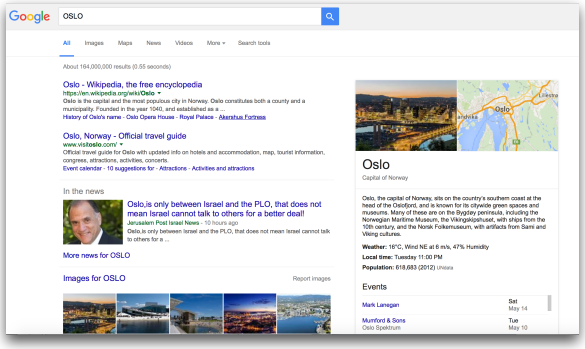
What is entity linking?

Linking **free text** to **entities**

- **Text:** any piece of text
  - documents (news, blog post, etc.)
  - tweets
  - queries
  - ….
- **Entities:** typically taken from a knowledge graph
  - Wikipedia
  - DBpedia
  - Freebase
  - …

Why entity linking?

Why entity linking?



Why entity linking?

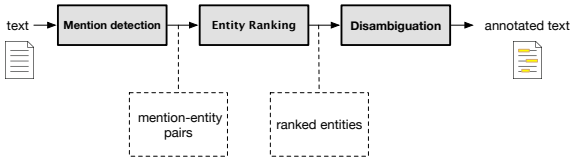
Enables:

- Semantic search
- Automatic document enrichment; go-read-here
- Ontology learning, KB population

"Used as feature" to improve:

- Classification
- Retrieval

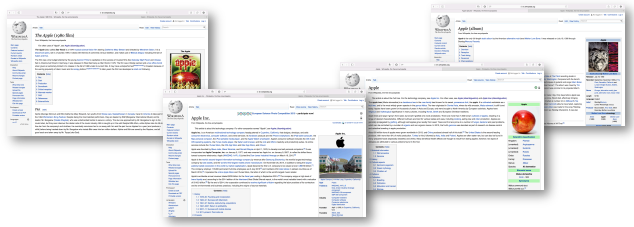
Approach



Step1: Mention detection

Determine “linkable” phrases

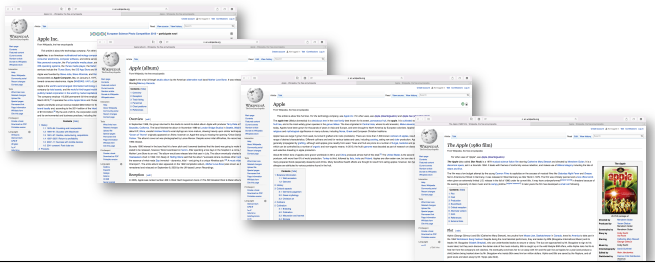
Apple is the second largest company ….



Step 2: Entity ranking

Rank candidate entities

Apple is the second largest company ….

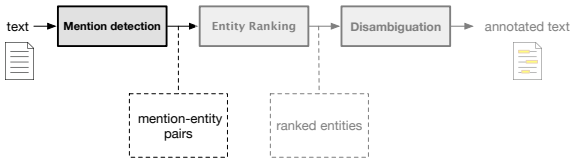


Step 3: Disambiguation

Disambiguate (filter or select)

Apple is the second largest company ….

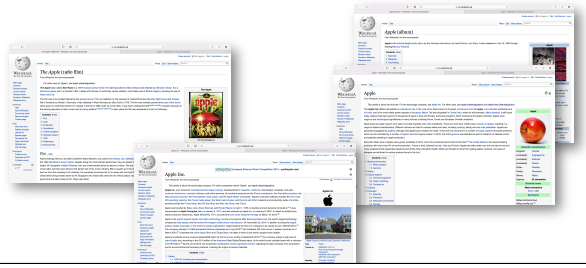
Approach



Step1: Mention detection

Determine “linkable” phrases

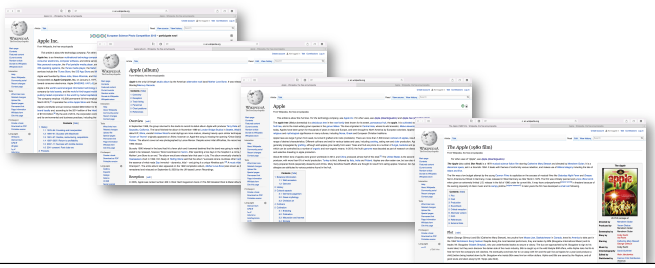
Apple is the second largest company ….



Step 2: Entity ranking

Rank candidate entities

Apple is the second largest company ….



Step 3: Disambiguation

Disambiguate (filter or select)

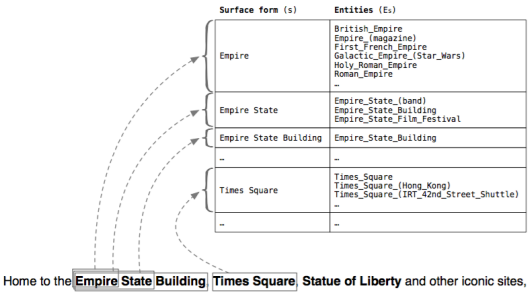
Apple is the second largest company ….

Mention detection

Detecting all “linkable phrases” (mentions) of the text, with their corresponding entities.

- 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; all are not interesting

Mention detection-Example



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.

Image taken from Balog. (2016). Entity Linking. In10th Russian Summer School in Information Retrieval.

Mention detection

1. Build a dictionary of entity surface forms
   - contains a mapping from entity name variants to entities
2. Find all document n-grams (substrings) against the dictionary
   - The length of n-gram is typically between 6 and 8
3. Filter out undesired entities

Mention detection

Key questions:

- What is the data source for entity name variants?
  - Wikipedia
- How to filter out inappropriate entities?
  - Statistical hints
  - Mention length

Building the dictionary

- **Page title**
  - the most common name of the entity

Building the dictionary

- Page title
  - the most common name of the entity
- **Redirect pages**
  - alternative name for referring to the entity

Building the dictionary

- Page title
  - the most common name of the entity
- Redirect pages
  - alternative name for referring to the entity
- **Disambiguation pages**
  - entities that share the same name

Building the dictionary

- Page title
  - the most common name of the entity
- Redirect pages
  - alternative name for referring to the entity
- Disambiguation pages
  - entities that share the same name
- **Anchor texts**
  - Wikipedia hyper links

Building the dictionary

- Page title
  - the most common name of the entity
- Redirect pages
  - alternative name for referring to the entity
- Disambiguation pages
  - entities that share the same name
- Anchor texts
  - Wikipedia hyper links
- **Bold texts from the first paragraph**
  - denotes other name variants of the entity

Filtering mentions

- Surface form dictionaries are rich and large
- A mention can be associated to too many entities
  - esp. the very commons names (e.g. 'new york', 'us')
- Some mentions are unlikely to be linked any entity
  - 'the' -> [The The]
  - 'b' -> [B (I Am Kloot album)]

Keyphraseness

Probability of a word being linked

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

- $|D_{link}(m)|$: number of Wikipedia articles where ***m*** **appears as link**
- $|D(m)|$: number of Wikipedia articles that **contain *m***

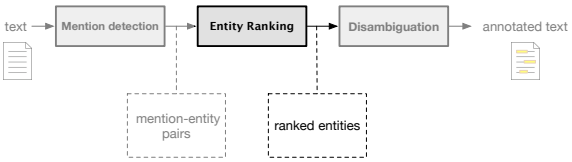
Commonness

Probability of a word referring to the entity

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

- $n(m,e)$: number of times entity *e* is the link target of mention *m*
- $\sum_{e'} n(m,e')$: total number of times mention *m* appears as link

Approach



Entity ranking

Ranking entities and narrowing down the space of disambiguation possibilities.

Various types of features can be used:

- Context independent
- Context dependent
- Entity relatedness

Context independent features

Neither the text nor other mentions in the document are taken into account

- Keyphraseness
- Commonness
- Link prior
  - Probability of the entity measured in terms of incoming links
- Page views
  - Probability of the entity measured in terms of traffic volume

Context dependent features

Compare the surrounding **context of a mention** with the textual representation of the entity

- **Context of a mention**
  - Window of text (sentence, paragraph) around the mention
  - Entire document
- **Similarity function**
  - Cosine similarity $\text{sim}_{cos}(m,e) = \frac{\vec{d}_m \cdot \vec{d}_e}{\| \vec{d}_m \| \| \vec{d}_e \|}$

Entity relatedness

Captures coherence between entity linking decisions in the text

- **Assumption:** a document focuses on one or at most a few topics
- Entities mentioned in a document should be topically related to each other
- Relatedness can be captured between two entities

Entity relatedness

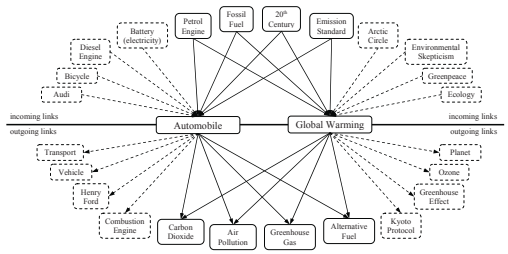


Image taken from Milne and Witten (2008a). An Effective, Low-Cost Measure of Semantic Relatedness Obtained from Wikipedia Links. In AAAI WikiAI Workshop.

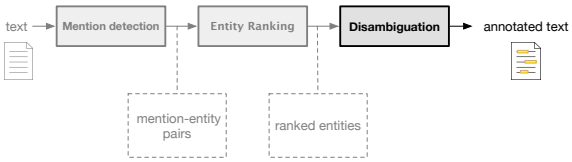
Wikipedia Link-based Measure (WLM)

Two entities are related if there is a large overlap between their incoming links

$$\mathrm{WLM}(e, e') = 1 - \frac{\log(\max(|L_e|, |L_{e'}|)) - \log(|L_e \cap L_{e'}|)}{\log(|E|) - \log(\min(|L_e|, |L_{e'}|))}$$

$|E|$: total number of entities; $|L_e|$, $|L_{e'}|$: set of entities that link to e

Approach



Disambiguation

Selecting single entity or none for each mention

Approaches

- Pruning based on score threshold
- Classification algorithms
- Graph-based approaches

Graph based approach

- **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
- The graph with the highest density is kept as the solution

Graph based approach

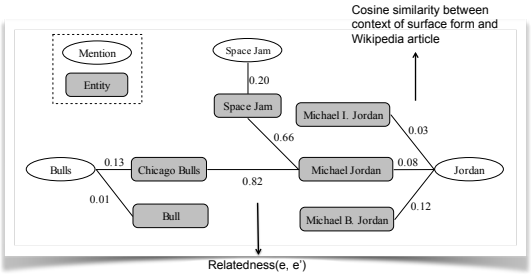


Image taken from Han et. al. (2011). Collective Entity Linking in Web Text: A Graph-based Method. In SIGIR

Exercise

Example of an entity linking system

Sample text:

"... Angola changed from a one-party Marxist-Leninist system ruled by the MPLA to a formal multiparty democracy following the 1992 elections ..."

Table 1: An excerpt from the surface form dictionary.

Mention	Entity	Count
1992 elections	<wikipedia:Philippine_general_election,_1992>	9
1992 elections	<wikipedia:Angolan_presidential_election,_1992>	1
1992 elections	*_total*	98
angola	<wikipedia:Angola>	4026
angola	<wikipedia:Angola_(Portugal)>	6
angola	<wikipedia:Angola_national_football_team>	120
angola	*_total*	4298
democracy	<wikipedia:Democracy>	108
democracy	<wikipedia:Democracy_(album)>	3
democracy	*_total*	2162
multiparty democracy	<wikipedia:multiparty_democracy>	11
multiparty democracy	*_total*	11
one party	<wikipedia:Non-possessors>	1
one party	<wikipedia:Single-party_state>	5
one party	*_total*	983

Mention detection

Question: Considering Table 1, what is the output of the mention detection step for the given sample text?

Answer:

All mention-entity pairs of Table 1 are considered, except the ones related to the mention “democracy“. We ignore this mention, because the longer mention “multiparty democracy“ is considered.

Entity ranking

Question: Compute the commonness for all mention-entity pairs, where mention is “1992 elections”.

Mention	Entity	Commonness
1992 elections	<wikipedia:Philippine_general_election,_1992>	$9/98 = 0.09$
1992 elections	<wikipedia:Angolan_presidential_election,_1992>	$1/98 = 0.01$
angola	<wikipedia:Angola>	$4026/4298 = 0.93$
angola	<wikipedia:Angola_national_football_team>	$120/4298 = 0.03$
angola	<wikipedia:Angola_(Portugal)>	$6/4298 = 0.001$
multiparty democracy	<wikipedia:multiparty_democracy>	$11/11 = 1$
one party	<wikipedia:Single-party_state>	$5/983 = 0.005$
one party	<wikipedia:Non-possessors>	$1/983 = 0.001$

Disambiguation

Question: Considering $\tau_s = 0.01$, what is the output of the CMNS approach?

Mention	Entity
1992 elections	<wikipedia:Philippine_general_election,_1992>
angola	<wikipedia:Angola>
multiparty democracy	<wikipedia:multiparty_democracy>