Entity Linking and Coreference Resolution

CSCI 699
Instructor: Xiang Ren
USC Computer Science
Entity Linking:

CSCI 699
Entity Linking: The Problem

Given a source document, identify entities mentioned in text, and find the knowledge base entities they represent.
Problem: Example

Example Query:

Northern Ireland has a population of about one and a half million people. At the time of partition in 1921 Protestants / unionists had a two-thirds majority in the region. The first Prime Minister of Northern Ireland, Sir James Craig, described the state as having ‘a Protestant Parliament for a Protestant people.’ The state effectively discriminated against Catholics in housing, jobs, and political representation.

http://cain.ulst.ac.uk/othelem/incorepaper09.htm

Search for: Northern Ireland
Northern Ireland

For the European Parliament constituency, see Northern Ireland (European Parliament constituency).

Northern Ireland (Irish: Tuaiscirt Éireann) pronounced [ˈtuːəʃəɾt əˈɾʲən̪ˠ] (listen), Ulster Scots: Norlin Airlan or Norlin Airlan) is a part of the United Kingdom in the north-east of the island of Ireland. It is variously described as a country, province or region of the UK, amongst other terms.\(^3\)\(^4\)\(^5\) Northern Ireland shares a border with the Republic of Ireland to the south and west. As of 2011, its population was 1,810,863,\(^2\) constituting about 30% of the island's total population and about 3% of the population of the United Kingdom. Since the signing of the Good Friday Agreement in 1998, Northern Ireland is largely self-governing. According to the agreement, Northern Ireland co-operates with the rest of Ireland – from which it was partitioned in 1921 – on some policy areas, while other areas are reserved for the Government of the United Kingdom, though the Republic of Ireland "may put forward views and proposals".\(^8\)

Northern Ireland was for many years the site of a violent and bitter inter-communal conflict – the Troubles – which was caused by divisions between nationalists, who see themselves as Irish and are predominantly Roman Catholic, and unionists, who see themselves as British and are predominantly Protestant. (Additionally, people from both sides of the community may describe themselves as Northern Irish.)\(^7\)

Unionists want Northern Ireland to remain as a part of the United Kingdom,\(^8\) while nationalists want reunification with the rest of Ireland, independent of British rule.\(^9\)\(^10\)\(^11\)\(^12\) Since 1998, most of the paramilitary groups involved in the Troubles have ceased their armed campaigns.

Northern Ireland has traditionally been the most industrialised region of the island. After declining as a result of political and social unrest for much of the 20th century, the economy has started to grow in the 1990s and 2000s. In 2017, the gross domestic product (GDP) of Northern Ireland was £52.8 billion, or £30,098 per head, making it the fourth largest region in the UK, behind the London, South-East and Wales regions.
Problem: Example

Example Query:

Northern Ireland has a population of about one and a half million people. At the time of partition in 1921, Protestants / unionists had a two-thirds majority in the region. The first Prime Minister of Northern Ireland, Sir James Craig, described the state as having ‘a Protestant Parliament for a Protestant people.’ The state effectively discriminated against Catholics in housing, jobs, and political representation.

http://cain.ulst.ac.uk/othelem/incorepaper09.htm

Search for: James Craig
near miss! :(
Human immunodeficiency virus (HIV) is the primary etiologic agent responsible for the AIDS pandemic. We constructed a fusion of the gp41 membrane-proximal external region (MPER) peptide along with a variable-length (Gly4Ser)x linker (where x is 4 or 8) between the C terminus of the former and N terminus of the latter. The His-tagged recombinant proteins, expressed in BL21(DE3)pLysS cells and purified by immobilized metal affinity chromatography followed by gel filtration, were found to display a nanomolar efficacy in blocking BaL-pseudotyped HIV-1 infection of HOS.T4.R5 cells. This antiviral activity was HIV-1 specific, since it did not inhibit cell infection by vesicular stomatitis virus (VSV). The chimeric proteins were found to release intraviral p24 protein from both BaL-pseudotyped HIV-1 and fully infectious BaL HIV-1 in a dose-dependent manner in the absence of host cells. The addition of either MPER or CVN was found to outcompete this virolytic effect, indicating that both components of the chimera are required for virolysis. The finding that engaging the Env protein spike and membrane using a chimeric ligand can destabilize the virus and lead to inactivation opens up a means to investigate virus particle metastability and to evaluate this approach for inactivation at the earliest stages of exposure to virus and before host cell encounter.
Application: Navigating Unfamiliar Domains

Educational Applications: Unfamiliar domains may contain terms unknown to a reader. The Wikifier can supply the necessary background knowledge even when the relevant article titles are not identical to what appears in the text, dealing with both ambiguity and variability.
| It’s a version of Chicago – the standard classic Macintosh menu font, with that distinctive thick diagonal in the ”N”. | Chicago was used by default for Mac menus through MacOS 7.6, and OS 8 was released mid-1997.. | Chicago VIII was one of the early 70s-era Chicago albums to catch my ear, along with Chicago II. |
Application: Organizing knowledge

<table>
<thead>
<tr>
<th>Chicago</th>
<th>Chicago was used by default for Mac menus through MacOS 7.6, and OS 8 was released mid-1997.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chicago VIII</td>
<td>Chicago VIII was one of the early 70s-era Chicago albums to catch my ear, along with Chicago II.</td>
</tr>
</tbody>
</table>

It’s a version of *Chicago* – the standard classic *Macintosh* menu font, with that distinctive thick diagonal in the ”N”.

*Chicago* was used by default for *Mac* menus through *MacOS 7.6*, and *OS 8* was released mid-1997.
Application: Organizing knowledge

It’s a version of **Chicago** – the standard classic **Macintosh** menu font, with that distinctive thick diagonal in the “N”.

**Chicago** was used by default for **Mac** menus through **MacOS 7.6**, and **OS 8** was released mid-1997.

**Chicago VIII** was one of the early 70s-era **Chicago** albums to catch my ear, along with **Chicago II**.
**Background Knowledge**

<table>
<thead>
<tr>
<th>Chicago</th>
<th>Chicago was used by default for Mac menus through MacOS 7.6, and OS 8 was released mid-1997.</th>
<th>Chicago VIII was one of the early 70s-era Chicago albums to catch my ear, along with Chicago II.</th>
</tr>
</thead>
</table>

It’s a version of **Chicago** – the standard classic **Macintosh** menu font, with that distinctive thick diagonal in the ”N”.

**Chicago**

**Macintosh**

**MacOS 7.6**

**OS 8**

**Chicago VIII**

**Chicago II**
It’s a version of **Chicago** – the standard classic **Macintosh** menu font, with that distinctive thick diagonal in the ”N”.

**Chicago** was used by default for **Mac** menus through **MacOS 7.6**, and **OS 8** was released mid-1997. **Chicago VIII** was one of the early 70s-era **Chicago** albums to catch my ear, along with **Chicago II**.
Task Definition

• A formal definition of the task consists of:

1. A definition of the mentions (concepts, entities) to highlight

2. Determining the target encyclopedic resource (KB)

3. Defining what to point to in the KB (title)
1. Mentions

• A mention: a phrase used to refer to something in the world
  • Named entity (person, organization), object, substance, event, philosophy, mental state, rule ...

• Task definitions vary across the definition of mentions
  • All N-grams (up to a certain size); Dictionary-based selection; Data-driven controlled vocabulary (e.g., all Wikipedia titles); only named entities (by NER).

• Ideally, one would like to have a mention definition that adapts to the application/user
Blumenthal (D) is a candidate for the U.S. Senate seat now held by Christopher Dodd (D), and he has held a commanding lead in the race since he entered it. But the Times report has the potential to fundamentally reshape the contest in the Nutmeg State.

Some task definitions insist on dealing only with mentions that are named entities

How about: *Hosni Mubarak’s wife?*
Both entities have a Wikipedia page

Blumenthal (D) is a candidate for the **U.S. Senate** seat now held by Christopher Dodd (D), and he has held a commanding lead in the race since he entered it. But the *Times* report has the potential to fundamentally reshape the contest in the *Nutmeg State*. 
Examples of Mentions

Alex Smith
offseason
turnover
feet
Examples of Mentions

Perhaps the definition of which mentions to highlight should depend on the expertise and interests of the users?
2. Concept Inventory (KB)

• Multiple KBs can be used, in principle, as the target KB.
2. Concept Inventory (KB)

• Multiple KBs can be used, in principle, as the target KB.

• Wikipedia has the advantage of a broad coverage, regularly maintained KB, with significant amount of text associated with each title.
  • All type of pages?
    • Content pages
    • Disambiguation pages
    • List pages
3. What to Link to? (Disambiguation)

The veteran tight end suffered a wrist injury in the third quarter during the regular season finale against Baltimore. Bengals head coach Marvin Lewis described the injury as a "wrist dislocation".

**Baltimore Raven:** Should the link be any different? **Both?**

**Baltimore:** The city? Baltimore Raven, the Football team? **Both?**

**Atmosphere:** The general term? **Or the most specific** one "Earth Atmosphere?"
3. Dealing with *Null* Links

- Often, there are multiple sensible links.

*Dorothy Byrne*, a state coordinator for the *Florida Green Party*,…

- How to capture the fact that *Dorothy Byrne* does not refer to any concept in Wikipedia?
3. Dealing with *Null* Links

- Often, there are multiple sensible links.

*Dorothy Byrne*, a state coordinator for the *Florida Green Party*,…

- How to capture the fact that *Dorothy Byrne* does not refer to any concept in Wikipedia?

- **Current practice**: If multiple mentions in the given document(s) correspond to the same concept, which is outside KB
  - First cluster relevant mentions as representing a single concept
  - Map the cluster to *Null*
Why EL is Challenging?
General Challenges

Blumenthal (D) is a candidate for the U.S. Senate seat now held by Christopher Dodd (D), and he has held a commanding lead in the race since he entered it. But the Times report has the potential to fundamentally reshape the contest in the Nutmeg State.

• Ambiguity
  Times ⇔ The New York Times
  The Times

• Variability
  CT ⇔ Connecticut
  The Nutmeg State

• Concepts outside of KB (NIL)
  • Blumenthal ?

• Scale
  • Millions of labels
Language Variability

**ambiguity**

/ˌambiˈɡyoo-ɪtē/ (noun)

1. uncertainty or inexactness of meaning in language.
   "we can detect no ambiguity in this section of the Act"
   synonyms: vagueness, obscurity, abstruseness, doubtfulness, uncertainty
   More

**synonym**

/ˈsɪnəˌnɪm/ (noun)

1. a word or phrase that means exactly or nearly the same as another word or phrase in the same language, for example shut is a synonym of close.
   synonyms: alternate, substitute, alternative, equivalent, euphemism
   More
Language Variability

**ambiguity**

/ˌambiˈɡjuətɪ/  

noun

1. uncertainty or inexactness of meaning in language.
   "we can detect no ambiguity in this section of the Act"
   synonyms: vagueness, obscurity, abstruseness, doubtfulness, uncertainty;  More

**synonym**

/ˈsɪnəˌnɪm/  

noun

1. a word or phrase that means exactly or nearly the same as another word or phrase in the same language, for example shut is a synonym of close.
   synonyms: alternate, substitute, alternative, equivalent, euphemism;  More
Example Query:

Northern Ireland has a population of about one and a half million people. At the time of partition in 1921, Protestants/unionists had a two-thirds majority in the region. The first Prime Minister of Northern Ireland, Sir James Craig, described the state as having ‘a Protestant Parliament for a Protestant people.’ The state effectively discriminated against Catholics in housing, jobs, and political representation.

http://cain.ulst.ac.uk/othelem/incorepaper09.htm

Search for:

James Craig
near miss! :(
**Synonym**: One concept/entity can have many reference names.
Other Challenges

• Dealing with Popularity Bias

• Recovering from gaps in background knowledge
  • Mostly when dealing with short texts and social media

• Exploiting common sense knowledge
Popular Bias: If you search for “Michael Jordan”
Evaluation of Entity Linking
Step-wise Evaluation Metrics

• Detection of mentions in text
  • Are the detected concepts/entities accurate?
  • Same as NER: Precision, Recall, F-measure
Step-wise Evaluation Metrics

• Detection of concepts / entities in text
  • Same as NER: Precision, Recall, F-measure

• Disambiguation accuracy
  • Evaluation quality of links per mention
  • Ranking-based metrics: Mean average precision (MAP), NDCG, MRR, ...
  • Accuracy @ K (K=1, 5, 10...) – this includes NIL label
Step-wise Evaluation Metrics

• Detection of concepts / entities in text
  • Same as NER: Precision, Recall, F-measure

• Disambiguation accuracy
  • Evaluation quality of links per mention
  • Ranking-based metrics: Mean average precision (MAP), NDCG, MRR, ...
  • Accuracy @ K (K=1, 5, 10...) – this includes NIL label

• NIL clustering
  • Grouping of out-of-KB mentions into coherent clusters
End-to-end Evaluation Metrics

• End-to-end mention detection + mention disambiguation + NIL Clustering
  • CEAF
  • B-cubed
  • Graph Edit Distance
Entity Linking: Subtasks

- Entity Linking requires addressing several sub-tasks:
  - Identifying Target Mentions
    - Mentions in the input text that should be linked to KB
  - Identifying Candidate KB entities
    - Candidate KB entities that could correspond to each mention
  - Candidate Entity Ranking
    - Rank the candidate entities for a given mention
  - NIL Detection and Clustering
    - Identify mentions that do not correspond to a KB entity
    - (optional) cluster NIL mentions that represent the same entity.
Entity Linking: Subtasks

• Entity Linking requires addressing several sub-tasks:
  • Identifying Target Mentions
    • Mentions in the input text that should be linked to KB
  • Identifying Candidate KB entities
    • Candidate KB entities that could correspond to each mention
  • Candidate Entity Ranking
    • Rank the candidate entities for a given mention
  • NIL Detection and Clustering
    • Identify mentions that do not correspond to a KB entity
    • (optional) cluster NIL mentions that represent the same entity.
Mention Identification

• Highest recall: Each n-gram is a potential concept mention
  • Intractable for larger documents

• Surface form based filtering
  • Shallow parsing (especially NP chunks), NP’s augmented with surrounding tokens, capitalized words
  • Remove: single characters, “stop words”, punctuation, etc.

• Classification and statistics based filtering
  • Name tagging (Finkel et al., 2005; Ratinov and Roth, 2009; Li et al., 2012)
  • Mention extraction (Florian et al., 2006, Li and Ji, 2014)
  • Key phrase extraction, independence tests (Mihalcea and Csomai, 2007), common word removal (Mendes et al., 2012; )
Mention Identification

• Multiple input sources are being used
  • Some build on the given text only, some use external resources.

• Methods used by some popular systems
  • Illinois Wikifier (Ratinov et al., 2011; Cheng and Roth, 2013)
    • NP chunks and substrings, NER (+nesting), prior anchor text
  • TAGME (Ferragina and Scaiella, 2010)
    • Prior anchor text
  • DBPedia Spotlight (Mendes et al., 2011)
    • Dictionary-based chunking with string matching (via DBpedia lexicalization dataset)
  • AIDA (Finkel et al., 2005; Hoffart et al., 2011)
    • Name Tagging
  • RPI Wikifier (Chen and Ji, 2011; Cassidy et al., 2012; Huang et al., 2014)
    • Mention Extraction (Li and Ji, 2014)
Mention Identification (Mendes et al., 2012)

<table>
<thead>
<tr>
<th>Method</th>
<th>P</th>
<th>R</th>
<th>Avg Time per mention</th>
</tr>
</thead>
<tbody>
<tr>
<td>L&gt;3</td>
<td>4.89</td>
<td>68.20</td>
<td>.0279</td>
</tr>
<tr>
<td>L&gt;10</td>
<td>5.05</td>
<td>66.53</td>
<td>.0246</td>
</tr>
<tr>
<td>L&gt;75</td>
<td>5.06</td>
<td>58.00</td>
<td>.0286</td>
</tr>
<tr>
<td>LNP*</td>
<td>5.52</td>
<td>57.04</td>
<td>.0331</td>
</tr>
<tr>
<td>NPL*&gt;3</td>
<td>6.12</td>
<td>45.40</td>
<td>1.1807</td>
</tr>
<tr>
<td>NPL*&gt;10</td>
<td>6.19</td>
<td>44.48</td>
<td>1.1408</td>
</tr>
<tr>
<td>NPL*&gt;75</td>
<td>6.17</td>
<td>38.65</td>
<td>1.2969</td>
</tr>
<tr>
<td>CW</td>
<td>6.15</td>
<td>42.53</td>
<td>.2516</td>
</tr>
<tr>
<td>Kea</td>
<td>1.90</td>
<td>61.53</td>
<td>.0505</td>
</tr>
<tr>
<td>NER</td>
<td>4.57</td>
<td>7.03</td>
<td>2.9239</td>
</tr>
<tr>
<td>NER U NP</td>
<td>1.99</td>
<td>68.30</td>
<td>3.1701</td>
</tr>
</tbody>
</table>

Dictionary-Based chunking (LingPipe) using DBPedia Lexicalization Dataset (Mendes et al., 2011)

LNP Extends L with simple heuristic to isolate NP’s

NPL>ₖ Same as LNP but with Statistical NP Chunker

CW Extends L by filtering out common words (Daiber, 2011)

Kea Uses supervised key phrase extraction (Frank et al., 1999)

NER Based on OpenNLP 1.5.1

NER U NP Augments NER with NPL
Entity Linking: Subtasks

• Entity Linking requires addressing several sub-tasks:
  • Identifying Target Mentions
    • Mentions in the input text that should be linked to KB
  • Identifying Candidate KB entities
    • Candidate KB entities that could correspond to each mention
  • Candidate Entity Ranking
    • Rank the candidate entities for a given mention
  • NIL Detection and Clustering
    • Identify mentions that do not correspond to a KB entity
    • (optional) cluster NIL mentions that represent the same entity.
Generating Candidate Entities

• 1. Based on canonical names (e.g. Wikipedia page title)
  • Titles that are a super or substring of the mention
    • Michael Jordan is a candidate for “Jordan”
  • Titles that overlap with the mention
    • “William Jefferson Clinton” → Bill Clinton;
    • “non-alcoholic drink” → Soft Drink
Candidate entities by names

James Craig

James Craig

James Craig (actor)

title: James Craig (actor) anchor text: James Craig
    James Craig
    in
disambiguation: James Craig
freebase name: James Craig (actor)

James Craig

JC, 1st Viscount Craigavon

title: James Craig, 1st Viscount Craigavon
anchor text: Sir James Craig's Craig Administration
disambiguation: James Craig
freebase name: Lord Craigavon
Generating Candidate Entities

• 1. Based on canonical names (e.g. Wikipedia page title)
  • Titles that are a super or substring of the mention
    • Michael Jordan is a candidate for “Jordan”
  • Titles that overlap with the mention
    • “William Jefferson Clinton” → Bill Clinton;
    • “non-alcoholic drink” → Soft Drink

• 2. Based on previously attested references
  • All Titles ever referred to by a given string in training data
    • Using, e.g., Wikipedia-internal hyperlink index
    • More Comprehensive Cross-lingual resource (Spitkovsky & Chang, 2012)
Candidate entities by attested references

In-content interlinking
Entity Linking: Subtasks

- Entity Linking requires addressing several sub-tasks:
  - Identifying Target Mentions
    - Mentions in the input text that should be linked to KB
  - Identifying Candidate KB entities
    - Candidate KB entities that could correspond to each mention
  - Candidate Entity Ranking
    - Rank the candidate entities for a given mention
  - NIL Detection and Clustering
    - Identify mentions that do not correspond to a KB entity
    - (optional) cluster NIL mentions that represent the same entity.
Entity Linking Solution Overview

• Identify mentions \( m_i \) in document \( d \)

• (1) Local Inference
  • For each \( m_i \) in \( d \):
    • Identify a set of relevant KB entities \( T(m_i) \)
    • Rank entities \( t_i \in T(m_i) \)
      [E.g., consider local statistics of edges \((m_i, t_i)\), \((m_i, *)\), and \((*, t_i)\)]
      occurrences in the Wikipedia graph]
Simple heuristics for initial ranking

• Initially rank titles according to…
  • Wikipedia article length
  • Incoming Wikipedia Links (from other titles) or incoming link to the KB entity
  • Number of inhabitants or the largest area (for geo-location titles)
Simple heuristics for initial ranking

• Initially rank titles according to…
  • Wikipedia article length
  • Incoming Wikipedia Links (from other titles) or incoming link to the KB entity
  • Number of inhabitants or the largest area (for geo-location titles)

• More sophisticated measures of prominence
  • Prior link probability
  • Centrality on graph
P(tlm): “Commonness”

\[
\text{Commonness}(m \Rightarrow t) = \frac{\text{count}(m \rightarrow t)}{\sum_{t' \in W} \text{count}(m \rightarrow t')}
\]

Typography

By default, a font called Charcoal is used to replace the similar Chicago typeface. Additional system fonts are also provided including Capitals, Gadget, Sand, Te operating system need to be provided, such as the Command key symbol, ††. |

Airlines and destinations

Although the population of Iceland is only about 300,000, there are scheduled flights to and from seven locations in the United States (Boston, Chicago, Minneapolis, New York, Orlando, Seattle, and Washington), three in Canada (Halifax, Toronto and Winnipeg) and 30 cities across Europe. The largest carriers at Keflavik are Icelandair and Iceland Express.

The Greatest Show on Earth were a British rock band, who recorded two albums for Harvest Records in 1970. The band had been conceived by Harvest Records in an attempt to create a horn-based rock combo, such as Blood Sweat & Tears or Chicago.\[1\]
P(t|m): “Commonness”

| Rank | t                        | P(t | ”Chicago”) |
|------|--------------------------|----------------|
| 1    | Chicago                  | .76            |
| 2    | Chicago (band)           | .041           |
| 3    | Chicago (2002_film)      | .022           |
| 20   | Chicago Maroons Football | .00186         |
| 100  | 1985 Chicago Whitesox Season | .00023448    |
| 505  | Chicago Cougars          | .0000528       |
| 999  | Kimbell Art Museum       | .00000586      |

- First used by Medelyan et al. (2008)
- Most popular method for initial candidate ranking
Note on Domain Dependence

• “Commonness” Not robust across domains

<table>
<thead>
<tr>
<th>Formal Genre</th>
<th>Corpus</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACE</td>
<td>86.85%</td>
<td></td>
</tr>
<tr>
<td>MSNBC</td>
<td>88.67%</td>
<td></td>
</tr>
<tr>
<td>AQUAINT</td>
<td>97.83%</td>
<td></td>
</tr>
<tr>
<td>Wiki</td>
<td>98.59%</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Tweets</th>
<th>Metric</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>60.21%</td>
<td></td>
</tr>
<tr>
<td>R-Prec</td>
<td>52.71%</td>
<td></td>
</tr>
<tr>
<td>Recall</td>
<td>77.75%</td>
<td></td>
</tr>
<tr>
<td>MRR</td>
<td>70.80%</td>
<td></td>
</tr>
<tr>
<td>MAP</td>
<td>58.53%</td>
<td></td>
</tr>
</tbody>
</table>

Ratinov et al. (2011)

Meij et al. (2012)
Graph-based Initial Ranking

- Centrality (Hachey et al., 2011; Hakimov et al., 2012)

\[ \text{Centrality}(a) = \sum_{b \in W} \frac{\partial_a}{s(a,b)} \ast \text{in\_links}(a) \ast \text{out\_links}(a) \ast k \]

- \( \partial_a \): the number of all reachable nodes from \( a \)
- \( s(a,b) \): the distance between \( a \) and \( b \)

- Importance of the title with respect to Wikipedia - Similar to PageRank (Brin & Page, 1998)
  - Hachey et al. (2011) showed the centrality works slightly better than PageRank
Local Ranking: How to?
Local Ranking: Basic Idea

• Use **similarity measure** to compare the **context of the mention** with the **text or structural info** associated with a candidate entity entity in KB (e.g., entity description in the corresponding KB page)

• “Similarity” can be (1) manually specified a-priori, or (2) machine-learned (w/ training examples)
Local Ranking: Basic Idea

• Use **similarity measure** to compare the **context of the mention** with the **text or structural info** associated with a candidate entity in KB (e.g., entity description in the corresponding KB page)

• “Similarity” can be (1) manually-specified, or (2) machine-learned

• **Mention-entity similarity** can be further combined with **entity-wise metrics** (e.g., entity popularity)
Context Similarity Measures

\[
\Gamma^* = \arg\max_{\Gamma} \sum_{i} \phi(m_i, t_i)
\]

Feature vector to capture degree of contextual similarity

Mapping from mentions to entities

Mention-concept assignment

Determine assignment that maximizes pairwise similarity
Context Similarity Measures: Context Source

- Varying notion of distance between mention and context tokens
  - Token-level, discourse-level
- Varying granularity of concept description
  - Synopsis, entire document

Text document containing mention

mention’s immediate context

all document text

Compact summary of concept

Text associated with KB concept

\[ \phi \]

The Chicago Bulls are a professional basketball team...

Chicago won six championships...
Context Similarity Measures: *Context Analysis*

- Context is processed and represented in a variety of ways

TF-IDF; Entropy based representation (Mendes et al., 2011)

Topic model representation

Facts about concept (e.g. <Jerry Reinsdorf, owner of, Chicago Bulls> in Wikipedia Info box)

1993 NBA all document playoffs

1990’s NBA Derrick Rose

Structured text representations such as chunks, dependency paths

Automatically extracted Keyphrases, named entities, etc.

The Chicago Bulls are a professional basketball team...
Typical Features for Candidate Ranking

<table>
<thead>
<tr>
<th>Mention/Concept Attribute</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td></td>
</tr>
<tr>
<td>Spelling match</td>
<td>Exact string match, acronym match, alias match, string matching...</td>
</tr>
<tr>
<td>KB link mining</td>
<td>Name pairs mined from KB text redirect and disambiguation pages</td>
</tr>
<tr>
<td>Name Gazetteer</td>
<td>Organization and geo-political entity abbreviation gazetteers</td>
</tr>
<tr>
<td>Document surface</td>
<td></td>
</tr>
<tr>
<td>Lexical</td>
<td>Words in KB facts, KB text, mention name, mention text.</td>
</tr>
<tr>
<td></td>
<td>Tf.idf of words and ngrams</td>
</tr>
<tr>
<td>Position</td>
<td>Mention name appears early in KB text</td>
</tr>
<tr>
<td>Genre</td>
<td>Genre of the mention text (newswire, blog, ... )</td>
</tr>
<tr>
<td>Local Context</td>
<td>Lexical and part-of-speech tags of context words</td>
</tr>
<tr>
<td>Entity Context</td>
<td></td>
</tr>
<tr>
<td>Type</td>
<td>Mention concept type, subtype</td>
</tr>
<tr>
<td>Relation/Event</td>
<td>Concepts co-occurred, attributes/relations/events with mention</td>
</tr>
<tr>
<td>Coreference</td>
<td>Co-reference links between the source document and the KB text</td>
</tr>
<tr>
<td>Profiling</td>
<td>Slot fills of the mention, concept attributes stored in KB infobox</td>
</tr>
<tr>
<td>Concept</td>
<td>Ontology extracted from KB text</td>
</tr>
<tr>
<td>Topic</td>
<td>Topics (identity and lexical similarity) for the mention text and KB text</td>
</tr>
<tr>
<td>KB Link Mining</td>
<td>Attributes extracted from hyperlink graphs of the KB text</td>
</tr>
<tr>
<td>Popularity</td>
<td></td>
</tr>
<tr>
<td>Web</td>
<td>Top KB text ranked by search engine and its length</td>
</tr>
<tr>
<td>Frequency</td>
<td>Frequency in KB texts</td>
</tr>
</tbody>
</table>

• (Ji et al., 2011; Zheng et al., 2010; Dredze et al., 2010; Anastacio et al., 2011)
Entity Profiling Feature Examples

Disambiguation

Name Variant Clustering
Topical features or topic based document clustering for context expansion (Milne and Witten, 2008; Syed et al., 2008; Srinivasan et al., 2009; Kozareva and Ravi, 2011; Zhang et al., 2011; Anastacio et al., 2011; Cassidy et al., 2011; Pink et al., 2013)
Context Similarity Measures: *Context Expansion*

- Obtain additional documents related to mention
  - Consider mention as information retrieval query
- KB may link to additional, more detailed information

"collaborator" mentions in other documents

The Chicago Bulls are a professional basketball team...

Additional info about entity

related documents, e.g. “External Links” in Wikipedia
Context Similarity Measures: *Computation*

\[ \Phi \]

- Cosine similarity (via TF-IDF)
- Other distance metrics (e.g. Jaccard)
- 2\textsuperscript{nd} order vector composition (Hoffart et al., EMNLP2011)
- Mutual Information
Entity Linking Solution Overview

• Identify mentions $m_i$ in document $d$

• (1) Local Inference
  • For each $m_i$ in $d$:
    • Identify a set of relevant KB entities $T(m_i)$
    • Rank entities $t_i \in T(m_i)$
      [E.g., consider local statistics of edges $[(m_i, t_i), (m_i, *), \text{ and } (*, t_i)]$ occurrences in the Wikipedia graph]
How these features weigh in the model? – Machine-learned ranking functions

NIL classification: Is it similar enough to be a match?
Putting it All Together

<table>
<thead>
<tr>
<th></th>
<th>Score Baseline</th>
<th>Score Context</th>
<th>Score Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chicago_city</td>
<td>0.99</td>
<td>0.01</td>
<td>0.03</td>
</tr>
<tr>
<td>Chicago_font</td>
<td>0.0001</td>
<td>0.2</td>
<td>0.01</td>
</tr>
<tr>
<td>Chicago_band</td>
<td>0.001</td>
<td>0.001</td>
<td>0.02</td>
</tr>
</tbody>
</table>

• **Learning to Rank** [Ratinov et. al. 2011]
  • Consider all pairs of title candidates
    • Supervision is provided by Wikipedia
  • Train a ranker on the pairs (learn to prefer the correct solution)
  • A Collaborative Ranking approach: outperforms many other learning approaches (Chen and Ji, 2011)
Ranking Approach Comparison

• Unsupervised or weakly-supervised learning (Ferragina and Scaiella, 2010)
  • Annotated data is minimally used to tune thresholds and parameters
  • The similarity measure is largely based on the unlabeled contexts

• Supervised learning (Bunescu and Pasca, 2006; Mihalcea and Csomai, 2007; Milne and Witten, 2008, Lehmann et al., 2010; McNamee, 2010; Chang et al., 2010; Zhang et al., 2010; Pablo-Sanchez et al., 2010, Han and Sun, 2011, Chen and Ji, 2011; Meij et al., 2012)
  • Each <mention, title> pair is a classification instance
  • Learn from annotated training data based on a variety of features
  • ListNet performs the best using the same feature set (Chen and Ji, 2011)

• Graph-based ranking (Gonzalez et al., 2012)
  • context entities are taken into account in order to reach a global optimized solution together with the query entity

• IR approach (Nemeskey et al., 2010)
  • the entire source document is considered as a single query to retrieve the most relevant Wikipedia article
Entity Linking Solution Overview

- Identify mentions \( m_i \) in document \( d \)

(1) Local Inference
  - For each \( m_i \) in \( d \):
    - Identify a set of relevant KB entities \( T(m_i) \)
    - Rank entities \( t_i \in T(m_i) \)
      [E.g., consider local statistics of edges \([(m_i, t_i), (m_i,*), and (*, t_i)]\) occurrences in the Wikipedia graph]

(2) Global Inference
  - For each document \( d \):
    - Consider all \( m_i \in d \); and all \( t_i \in T(m_i) \)
    - Re-rank entities \( t_i \in T(m_i) \)
      [E.g., if \( m, m' \) are related by virtue of being in \( d \), their corresponding entities \( t, t' \) may also be related]
Global Inference: Illustration

Northern Ireland

James Craig

Catholics

James Craig
James Craig (actor)

James Craig
JC, 1st Viscount Craigavon

Catholic Church
Roman Catholic Church

Catholic Church
American Catholic Church
Global Inference: Illustration

Northern Ireland

James Craig

Catholics

not compatible

Catholic Church

Roman Catholic Church

American Catholic Church

James Craig

James Craig (actor)

James Craig

JC, 1st Viscount Craigavon

Not compatible

Global Inference: Illustration
Global Inference: Illustration
Global Inference: A Combinatorial Optimization Problem
Global Inference/Ranking: Problem Formulation

\[ \Gamma^* \approx \arg \max_{\Gamma} \sum_{i=1}^{N} [\phi(m_i, t_i) + \sum_{t_i \in \Gamma, t_j \in \Gamma'} \psi(t_i, t_j)] \]

- How to define relatedness between two candidate entities? (What is \( \Psi \)?)
Conceptual Coherence

• Recall: The reference collection (might) have structure.

It’s a version of **Chicago** – the standard classic **Macintosh** menu font, with that distinctive thick diagonal in the "N".

**Chicago** was used by default for **Mac** menus through **MacOS 7.6**, and **OS 8** was released mid-1997.

**Chicago VIII** was one of the early 70s-era **Chicago** albums to catch my ear, along with **Chicago II**.

• Co-occurrence:
  • Textual co-occurrence of concepts is reflected in the KB (Wikipedia)

• In-text referencing:
  • Preferred disambiguation contains structurally coherent concepts
Co-occurrence (Entity 1, Entity 2)

Typography

By default, a font called Charcoal is used to replace the similar Chicago typeface; additional system fonts are also provided including Capitals, Gadget, Sand, Te operating system need to be provided, such as the Command key symbol, Ṣe. |

Airlines and destinations

Although the population of Iceland is only about 300,000, there are scheduled flights to and from seven locations in the United States (Boston, Chicago, Minneapolis, New York, Orlando, Seattle, and Washington), three in Canada (Halifax, Toronto and Winnipeg) and 30 cities across Europe. The largest carriers at Keflavik are Icelandair and Iceland Express.

The city senses of Boston and Chicago appear together often.

The Greatest Show on Earth were a British rock band, who recorded two albums for Harvest Records in 1970. The band had been conceived by Harvest Records in an attempt to create a horn-based rock combo, such as Blood Sweat & Tears or Chicago. [1]
Entity Coherence & Relatedness

• Let c, d be a pair of entities …
• Let C and D be their sets of incoming (or outgoing) links
  • Unlabeled, directed link structure

\[
\text{relatedness}(c,d) = \frac{\log\left(\max(|C|,|D|)\right) - \log(|C \cap D|)}{\log(W) - \log(\min(|C|,|D|))}
\]

\[
\text{PMI}(c,d) = \frac{|C \cap D| / |W|}{\left(|C| / |W|\right) \ast \left(D / |W|\right)}
\]

• Let C and D ∈\{0,1\}^K, where K is the set of all categories

\[
\text{relatedness}(c,d) = \langle C, D \rangle
\]

See García et al. (JAIR2014) for variational details

Introduced by Milne & Witten (2008)
Relatedness Outperforms Pointwise Mutual Information (Ratinov et al., 2011)

Category based similarity introduced by Cucerzan (2007)
More relatedness features (Ceccarelli et al., 2013)

<table>
<thead>
<tr>
<th>Symmetric Features</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\rho_{MW}^{MW} (a, b)$</td>
<td>co-citation based similarity [19].</td>
</tr>
<tr>
<td>$J(a, b)$</td>
<td>Jaccard similarity: $J(a, b) = \frac{\text{in}(a) \cap \text{in}(b)}{\text{in}(a) \cup \text{in}(b)}$.</td>
</tr>
<tr>
<td>$P(a, b)$</td>
<td>joint probability of entities $a$ and $b$: $P(a, b) = P(a</td>
</tr>
<tr>
<td>$\text{Link}(a \leftrightarrow b)$</td>
<td>equals 1 if $a$ links to $b$ and vice versa, 0 otherwise.</td>
</tr>
<tr>
<td>$\text{AvgFr}(a, b)$</td>
<td>average friendship: $(\text{Friend}(a, b) + \text{Friend}(b, a))/2$.</td>
</tr>
<tr>
<td>$\rho_{\text{out}}^{MW} (a, b)$</td>
<td>$\rho_{MW}^{MW}$ considering outgoing links.</td>
</tr>
<tr>
<td>$\rho_{\text{in-out}}^{MW} (a, b)$</td>
<td>$\rho_{MW}^{MW}$ considering the union of the incoming and outgoing links.</td>
</tr>
<tr>
<td>$J_{\text{out}}(a, b)$</td>
<td>Jaccard similarity considering the outgoing links.</td>
</tr>
<tr>
<td>$J_{\text{in-out}}(a, b)$</td>
<td>Jaccard similarity considering the union of the incoming and outgoing links.</td>
</tr>
<tr>
<td>$\chi^2(a, b)$</td>
<td>$\chi^2$ statistic: $\chi^2(a, b) = (</td>
</tr>
<tr>
<td>$\chi_{\text{out}}^2 (a, b)$</td>
<td>$\chi^2$ statistic considering the outgoing links.</td>
</tr>
<tr>
<td>$\chi_{\text{in-out}}^2 (a, b)$</td>
<td>$\chi^2$ statistic considering the union of the incoming and outgoing links.</td>
</tr>
<tr>
<td>$\text{PMI}(a, b)$</td>
<td>point-wise mutual information: $\log \frac{P(a</td>
</tr>
</tbody>
</table>
Entity Linking: Subtasks

- Entity Linking requires addressing several sub-tasks:
  - Identifying Target Mentions
    - Mentions in the input text that should be linked to KB
  - Identifying Candidate KB entities
    - Candidate KB entities that could correspond to each mention
  - Candidate Entity Ranking
    - Rank the candidate entities for a given mention
  - NIL Detection and Clustering
    - Identify mentions that do not correspond to a KB entity
    - (optional) cluster NIL mentions that represent the same entity.
NIL Detection

Is it in the KB?

\{ W_1, W_2, W_N, W_{NIL} \}

Jordon accepted a basketball scholarship to North Carolina, ...

In the 1980’s Jordan began developing recurrent neural networks.

Local man Michael Jordan was appointed county coroner ...

1. Augment KB with NIL entry and treat it like any other entry
2. Include general NIL-indicating features

1. Binary classification (Within KB vs. NIL)
2. Select NIL cutoff by tuning confidence threshold

Is it an entity?

• Concept Mention Identification (above)
• Not all NP’s are linkable

No spike: Not an entity

‘Prices Quoted’

Sudden Google Books frequency spike: Entity

“Soluble Fiber”
NIL Clustering

“All in one”

Simple string matching

“One in one”

Often difficult to beat!

Collaborative Clustering

Most effective when ambiguity is high
# NIL Clustering Methods Comparison
(Chen and Ji, 2011; Tamang et al., 2012)

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>B-cubed+ F-Measure</th>
<th>Complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Agglomerative clustering</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 linkage based algorithms (single linkage, complete linkage, average linkage) (Manning et al., 2008)</td>
<td>85.4%-85.8%</td>
<td>$O(n^2)$ $O(n^2 \log n)$ $n$: the number of mentions</td>
</tr>
<tr>
<td>6 algorithms optimizing internal measures cohesion and separation</td>
<td>85.6%-86.6%</td>
<td>$O(n^2 \log n)$ $O(n^3)$</td>
</tr>
<tr>
<td><strong>Partitioning Clustering</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6 repeated bisection algorithms optimizing internal measures</td>
<td>85.4%-86.1%</td>
<td>$O(\text{NNZ} \times k + m \times k)$ \text{NNZ}: the number of non-zeroes in the input matrix M: dimension of feature vector for each mention k: the number of clusters</td>
</tr>
<tr>
<td>6 direct k-way algorithms optimizing internal measures (Zhao and Karypis, 2002)</td>
<td>85.5%-86.9%</td>
<td>$O(\text{NNZ} \times \log k)$</td>
</tr>
</tbody>
</table>
Collaborative Clustering (Chen and Ji, 2011; Tamang et al., 2012)

- Consensus functions
  - Co-association matrix (Fred and Jain, 2002)
- 12% gain over the best individual clustering algorithm
New Trends

- Entity linking until now: Solving Entity Linking Problems in
  - Standard settings; Long documents
- Extending the task to new settings
  - Social media entity linking
  - Spatiotemporal entity linking
  - Handling emerging entities
  - Cross-lingual Entity Linking
  - Linking to general KB and ontologies
  - Fuzzy matching for candidates
Motivation: Short and Noisy Text

• Microblogs are data gold mines!
  • Over 400M short tweets per day

• Many applications
  • Election results [Tumasjan et al., SSCR 10]
  • Disease spreading [Paul and Dredze, ICWSM 11]
  • Tracking product feedback and sentiment [Asur and Huberman, WI-IAT 10]

• Need more research
  • Stanford NER on tweets got only 44% F1 [Ritter et. al, EMNLP 2011]
Challenges for Social Media

• Messages are short, noisy and informal
  • Lack of rich context to compute context similarity and ensure topical coherence

• Lack of Labeled Data for Supervised Model
  • Lack of Context makes annotation more challenging

who cares, nobody wanna see the spurs play. Remember they’re boring...
What approach should we use?

• **Task:** Restrict mentions to **Named Entities**
  - **Named Entity Wikification**

• **Approach 1 (NER + Disambiguation):**
  - Develop a named entity recognizer for target types
  - Link to entities based on the output of the first stage

• **Approach 2 (End-to-end Wikification):**
  - Learn to jointly detect mention and disambiguate entities
  - Take advantage of Wikipedia information

---

Mature Techniques

Limited Types; Adaptation
A Simple End-to-End Linking System

- [Guo, NAACL 13, Chang et. al. #Micropost 14]

Message → Text Normalization → Candidate Generation → Joint Recognition and Disambiguation → Overlap Resolution → Entity Linking Results

Winner of the NEEL challenge; The best two systems all adopt the end-to-end approach.

There is no mention filtering stage.
Balance the Precision and Recall

In certain applications (such as optimizing F1), we need to tune precision and recall. Much easier to do in a joint model.
How Difficult is Disambiguation?

- **Commonness Baseline** [Guo et al., NAACL 13]
  - Gold mentions match the prior anchor text (e.g. the lexicon)
  - P@1 = the accuracy of the most popular entity

- The baseline for disambiguating entities is high
  - The overall entity linking performance is still low
    - Mention detection is challenging for tweets!

- The mention detection problem is even more challenging
  - The lexicon is not complete

<table>
<thead>
<tr>
<th>Data</th>
<th>#Tweets</th>
<th>#Cand</th>
<th>#Entities</th>
<th>P@1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test 2</td>
<td>488</td>
<td>7781</td>
<td>332</td>
<td>89.6%</td>
</tr>
</tbody>
</table>
Morphs in Social Media

“Conquer West King” (平西王)  =  “Bo Xilai” (薄熙来)  =  “Baby” (宝宝)  =  “Wen Jiabao” (温家宝)

Chris Christie → the Hutt
Datasets and Tools
ERD 2014

• Given a document, recognize all of the mentions and the entities;
  • No target mention is given
• An entity snapshot is given
  • Intersection of Freebase and Wikipedia

• Input: Webpages
• Output: Byte-offset based predictions

• Webservice-driven; Leaderboard
NIST TAC Knowledge Base Population (KBP)

• KBP2009-2010 Entity Linking (Ji et al., 2010)
  • Entity mentions are given, Link to KB or NIL, Mono-lingual

• KBP2011-2013 (Ji et al., 2011)
  • Added NIL clustering and cross-lingual tracks

• KBP2014 Entity Discovery and Linking (Evaluation: September)
  • Given a document source collection (from newswire, web documents and discussion forums), an EDL system is required to automatically extract (identify and classify) entity mentions ("queries"), link them to the KB, and cluster NIL mentions
  • English Mono-lingual track
  • Chinese-to-English Cross-lingual track
  • Spanish-to-English Cross-lingual track
Dataset – Long Text

• KBP Evaluations (can obtain all data sets after registration)
  • http://nlp.cs.rpi.edu/kbp/

• CoNLL Dataset

• Emerging Entity Recognition
Dataset - Short Text

• Micropost Challenge
  • http://www.scc.lancs.ac.uk/microposts2014/challenge/index.html

• Dataset for “Adding semantics to microblog posts”
  • http://edgar.meij.pro/dataset-adding-semantics-microblog-posts/

• Dataset for “Entity Linking on Microblogs with Spatial and Temporal Signals”
  • http://research.microsoft.com/en-us/downloads/84ac9d88-c353-4059-97a4-87d129db0464/

• Query Entity Linking
  • http://edgar.meij.pro/linking-queries-entities/
UIUC Wikifier

The Wikification system has identified the following entities with Wikipedia articles. Click on an entity to visit the corresponding Wikipedia page. Hover over links to view the categories associated with each entity.

**Northern Ireland** has a **population** of about one and a half million people. At the time of partition in 1921, **Protestants** / **unionists** had a **two-thirds majority** in the region. The first Prime Minister of **Northern Ireland**, **Sir James Craig**, described the state as having a **Protestant Parliament** for a **Protestant** people. The state effectively discriminated against **Catholics** in housing, **jobs**, and **political representation**.
Northern Ireland has a population of about one and a half million people. At the time of partition in 1921, Protestants/unionists had a two-thirds majority in the region. The first Prime Minister of Northern Ireland, Sir James Craig, described the state as having 'a Protestant Parliament for a Protestant people.' The state effectively discriminated against Catholics in housing, jobs, and political representation.

James Craig, 1st Viscount Craigavon
James Craig, 1st Viscount Craigavon, PC, PC (NI) (8 January 1871 – 24 November 1940), was a prominent Irish unionist politician, leader of the Ulster Unionist Party and the first Prime Minister of Nor...
Northern Ireland has a population of about one and a half million people. At the time of partition in 1921 Protestants / unionists had a two-thirds majority in the region. The first Prime Minister of Northern Ireland, Sir James Craig, described the state as having 'a Protestant Parliament for a Protestant people.' The state effectively discriminated against Catholics in housing, jobs, and political representation.
Resources

• Tool List
  • http://nlp.cs.rpi.edu/kbp/2014/tools.html

• Shared Tasks
  • KBP 2014
    • http://nlp.cs.rpi.edu/kbp/2014/
  • ERD 2014
    • http://web-ngram.research.microsoft.com/erd2014
  • #Micropost Challenge (for tweets)
    • http://www.scc.lancs.ac.uk/microposts2014/challenge/index.html
  • Chinese Entity Linking Task at NLPCC2014
Coreference Resolution

CSCI 699
Coreference Resolution

Identify the noun phrases (or *entity mentions*) that *refer* to the same real-world entity

Queen Elizabeth set about transforming her husband, King George VI, into a viable monarch. A renowned speech therapist was summoned to help the King overcome his speech impediment...
Coreference Resolution

Identify the noun phrases (or entity mentions) that refer to the same real-world entity

Queen Elizabeth set about transforming her husband, King George VI, into a viable monarch. A renowned speech therapist was summoned to help the King overcome his speech impediment...
Coreference Resolution

Identify the noun phrases (or entity mentions) that refer to the same real-world entity

Queen Elizabeth set about transforming her husband, King George VI, into a viable monarch. A renowned speech therapist was summoned to help the King overcome his speech impediment...
Coreference Resolution

Identify the noun phrases (or entity mentions) that refer to the same real-world entity

Queen Elizabeth set about transforming her husband, King George VI, into a viable monarch. A renowned speech therapist was summoned to help the King overcome his speech impediment...

Inherently a clustering task
the coreference relation is transitive
\[
\text{Coref}(A,B) \land \text{Coref}(B,C) \implies \text{Coref}(A,C)
\]
Coreference Resolution

Identify the noun phrases (or entity mentions) that refer to the same real-world entity

Queen Elizabeth set about transforming her husband, King George VI, into a viable monarch. A renowned speech therapist was summoned to help the King overcome his speech impediment...

Typically recast as the problem of selecting an antecedent for each mention, $m_j$
Coreference Resolution

Identify the noun phrases (or *entity mentions*) that refer to the same real-world entity

**Queen Elizabeth** set about transforming *her* husband, **King George VI**, into a viable monarch. A renowned speech therapist was summoned to help *the King* overcome *his* speech impediment...

Typically recast as the problem of selecting an *antecedent* for each mention, $m_j$

Does **Queen Elizabeth** have a preceding mention coreferent with it? If so, what is it?
Coreference Resolution

Identify the noun phrases (or *entity mentions*) that refer to the same real-world entity

Queen Elizabeth set about transforming her husband, King George VI, into a viable monarch. A renowned speech therapist was summoned to help the King overcome his speech impediment...

Typically recast as the problem of selecting an antecedent for each mention, $m_j$

Does her have a preceding mention coreferent with it? If so, what is it?
Why it’s challenging?

Coreference strategies differ depending on the mention type definiteness of mentions

… Then Mark saw the man walking down the street.
… Then Mark saw a man walking down the street.
Why it’s challenging?

Coreference strategies differ depending on the mention type definiteness of mentions

… Then Mark saw the man walking down the street.
… Then Mark saw a man walking down the street.

pronoun resolution alone is notoriously difficult

There are pronouns whose resolution requires world knowledge

The Winograd Schema Challenge (Levesque, 2011)
Why it’s challenging?

Coreference strategies differ depending on the mention type definiteness of mentions

… Then Mark saw the man walking down the street.
… Then Mark saw a man walking down the street.

pronoun resolution alone is notoriously difficult

There are pronouns whose resolution requires world knowledge

The Winograd Schema Challenge (Levesque, 2011)

pleonastic pronouns refer to nothing in the text

I went outside and it was snowing.
Applications: Coref in QA

Where was Mozart born?

Mozart was one of the first classical composers. He was born in Salzburg, Austria, in 27 January 1756. He wrote music of many different genres...

Haydn was a contemporary and friend of Mozart. He was born in Rohrau, Austria, in 31 March 1732. He wrote 104 symphonies...
Applications: Coref in QA

Where was Mozart born?

Mozart was one of the first classical composers. He was born in Salzburg, Austria, in 27 January 1756. He wrote music of many different genres...

Haydn was a contemporary and friend of Mozart. He was born in Rohrau, Austria, in 31 March 1732. He wrote 104 symphonies...
Applications: Coref in QA

Where was Mozart born?

Mozart was one of the first classical composers. He was born in Salzburg, Austria, in 27 January 1756. He wrote music of many different genres...

Haydn was a contemporary and friend of Mozart. He was born in Rohrau, Austria, in 31 March 1732. He wrote 104 symphonies...
Applications: Coref in QA

Where was Mozart born?

Mozart was one of the first classical composers. He was born in Salzburg, Austria, in 27 January 1756. He wrote music of many different genres...

Haydn was a contemporary and friend of Mozart. He was born in Rohrau, Austria, in 31 March 1732. He wrote 104 symphonies...
Coref: The Mention-pair model

a classifier that, given a description of two mentions, $m_i$ and $m_j$, determines whether they are coreferent or not. Coreference as a pairwise classification task.
Coref: The Mention-pair model

Training instance creation
create one training instance for each pair of mentions from texts annotated with coreference information

\[ \text{negative} \quad \text{negative} \quad \text{positive} \]

Coref: The Mention-pair model

Training instance creation

create one training instance for each pair of mentions from texts annotated with coreference information

Coref: The Mention-Entity model

a classifier that determines whether (or how likely) a mention *belongs to a preceding coreference cluster*

more expressive than the mention-pair model

an instance is composed of a *mention* and a *preceding cluster*

can employ *cluster-level* features defined over any subset of mentions in a preceding cluster

is a mention gender-compatible with *most* of the mentions in it?
Coref: The Cluster-Ranking model

Mention-ranking model

- Rank candidate antecedents

Mention-entity model

- Consider preceding clusters, not candidate antecedents
- Rank preceding clusters
Coref: The Cluster-Ranking model

<table>
<thead>
<tr>
<th></th>
<th>B³</th>
<th></th>
<th></th>
<th>CEAF</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R</td>
<td>P</td>
<td>F</td>
<td>R</td>
<td>P</td>
<td>F</td>
</tr>
<tr>
<td>Mention-Pair Baseline</td>
<td>50.8</td>
<td>57.9</td>
<td>54.1</td>
<td>56.1</td>
<td>51.0</td>
<td>53.4</td>
</tr>
<tr>
<td>Entity-Mention Baseline</td>
<td>51.2</td>
<td>57.8</td>
<td>54.3</td>
<td>56.3</td>
<td>50.2</td>
<td>53.1</td>
</tr>
<tr>
<td>Mention-Ranking Baseline (Pipeline)</td>
<td>52.3</td>
<td>61.8</td>
<td>56.6</td>
<td>51.6</td>
<td>56.7</td>
<td>54.1</td>
</tr>
<tr>
<td>Mention-Ranking Baseline (Joint)</td>
<td>50.4</td>
<td>65.5</td>
<td>56.9</td>
<td>53.0</td>
<td>58.5</td>
<td>55.6</td>
</tr>
<tr>
<td>Cluster-Ranking Model (Pipeline)</td>
<td>55.3</td>
<td>63.7</td>
<td>59.2</td>
<td>54.1</td>
<td>59.3</td>
<td>56.6</td>
</tr>
<tr>
<td>Cluster-Ranking Model (Joint)</td>
<td>54.4</td>
<td>70.5</td>
<td>61.4</td>
<td>56.7</td>
<td>62.6</td>
<td>59.5</td>
</tr>
</tbody>
</table>
Coref: Two Recent Approaches

Multi-pass sieve approach (Lee et al., 2011)
- Winner of the CoNLL-2011 shared task
- English coreference resolution

Latent tree-based approach (Fernandes et al., 2012)
- Winner of the CoNLL-2012 shared task
- Multilingual coreference resolution (English, Chinese, Arabic)