Engaging Content
Engaging People

The Cultured Machine

Gary Munnelly

The ADAPT Centre is funded under the SFI Research Centres Programme(Grant 13/RC/2106)and is co-funded under the European Regional Development Fund.

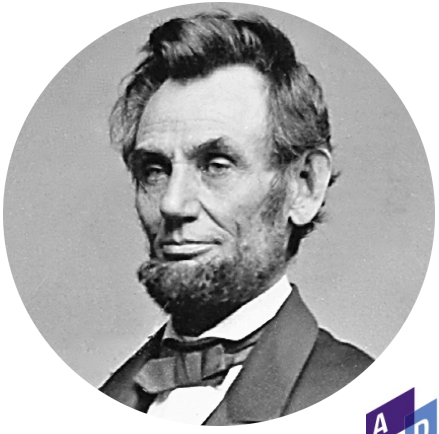
1. Digital Humanities

2. Entity Disambiguation

Digital Humanities

Accessibility

Accessibility



Digitisation

1. Mitigates risk of damage to artifacts
2. Facilitate parallel research
3. Reduces risk of “losing” artifacts
4. Makes the archive available to all

Towards a Generous Interface

"... open the doors, tear down the drab lobby; instead of demanding a query it would offer multiple ways in, and support exploration as well as the focused enquiry where search excels. In revealing the complexity of digital collections, a generous interface would also enrich interpretation by revealing relationships and structures within a collection."

1641 Depositions

Brennan, first [struk] att the said Richard Barnard being then young about 10 years of age with his sword drawne, & cutt him first a deep wound vpon his head & presently after ouer his Nose & face whervpon the said Richard fell to the ground and the said Lewis Brennon not being therwith satisfied in pursuance of his bloody & murderout disposacion took e of a collar hempen Cord from a grey hownds neck ther present, & therwith (putt vp about the said Richard neck, he draggd the said Richard to his fathers tenter hooks & ther the said Lewes hanged the said Richard

Brennan, first [struck] at the said Richard Barnabas being then young about 10 years of age with his sword drawn, & cut him first a deep wound upon his head & presently after over his Nose & face whereupon the said Richard fell to the ground and the said Lewis Brennon not being therewith satisfied in pursuance of his bloody & murderout disposation took of a collar hempen Cord from a grey hounds neck there present, & therewith (put up about the said Richard neck, he dragged the said Richard to his fathers tenter hooks & there the said Lewis hanged the said Richard

Entity Disambiguation

Definition

Entity Disambiguation: The problem of establishing a real world referent for a given mention of an entity.

Entity Disambiguation

Definition

Entity Disambiguation: The problem of establishing a real world referent for a given mention of an entity.

Not to be confused with coreference resolution or entity recognition

Entity Recognition

Brennan, first struk att the said Richard Barnard being then young about 10 years of age with his sword drawne, & cutt him first a deep wound vpon his head & presently after ouer his Nose & face whervpon the said Richard fell to the ground and the said Lewis Brennon not being therwith satisfied in pursuance of his bloody & murderout disposacion took e of a collar hempen Cord from a grey hownds neck ther present, & therwith (putt vp about the said Richard neck, he draggd the said Richard to his fathers tenter hooks & ther the said Lewes hanged the said Richard

Brennan, first struk att the said Richard Barnard being then young about 10 years of age with his sword drawne, & cutt him first a deep wound vpon his head & presently after ouer his Nose & face whervpon the said Richard fell to the ground and the said Lewis Brennon not being therwith satisfied in pursuance of his bloody & murderout disposacion took e of a collar hempen Cord from a grey hownds neck ther present, & therwith (putt vp about the said Richard neck, he draggd the said Richard to his fathers tenter hooks & ther the said Lewes hanged the said Richard

Entity Disambiguation

... amongst many books brought into the City of Limerick from foreign parts, & seized upon by the reverend Bishop of that Sea as prohibited ... one had a written addition to the first part which was printed, containing a discourse of the friars of the Augustine order, sometimes seated in the town of Armagh in Ulster

Entity Disambiguation

... amongst many books brought into the City of Limerick from foreign parts, & seized upon by the reverend Bishop of that Sea as prohibited ... one had a written addition to the first part which was printed, containing a discourse of the friars of the Augustine order, sometimes seated in the town of Armagh in Ulster

City of Limerick	`http://dbpedia.org/resource/Limerick`
Bishop of that Sea	`NIL`
Augustine order	`http://dbpedia.org/resource/Order_of_Saint_Augustine`
Armagh	`http://dbpedia.org/resource/County_Armagh`
Ulster	`http://dbpedia.org/resource/Ulster`

Entity Disambiguation

1. Choose Knowledge Base
   - What information does the disambiguation system have about the world?
   - Often use Wikipedia or DBpedia, but these aren't always good for cultural heritage.

2. Identify Candidates
   - For a single recognised entity, who or what might it be referring to.

3. Select Referents
   - Which of the candidates is the "right" candidate.

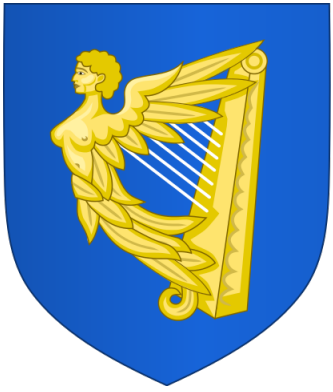
Choosing a Knowledge Base - Entity Representation

King of Spain (Felipe VI)

King of Spain (Philip IV)

Choosing a Knowledge Base - Entity Representation

Ireland (Republic)



Ireland (Kingdom)

Choosing a Knowledge Base - Entity Representation

William Alrich

Entity Disambiguation

1. Choose Knowledge Base
   - What information does the disambiguation system have about the world?
   - Often use Wikipedia or DBpedia, but these aren't always good for cultural heritage.

2. Identify Candidates
   - For a single recognised entity, who or what might it be referring to.

3. Select Referents
   - Which of the candidates is the "right" candidate.

Candidate Selection - Casting the Net

James I & VI

Candidate Selection - Casting the Net

James I & VI

Maiesties/Maiesty/Majesty

Candidate Selection - Casting the Net

James I & VI

Maiesties/Maiesty/Majesty

Rebel Bastard

Entity Disambiguation

1. Choose Knowledge Base
   - What information does the disambiguation system have about the world?
   - Often use Wikipedia or DBpedia, but these aren't always good for cultural heritage.

2. Identify Candidates
   - For a single recognised entity, who or what might it be referring to.

3. Select Referents
   - Which of the candidates is the "right" candidate.

Disambiguation

What information do we have at our disposal?

Disambiguation

What information do we have at our disposal?

Depends on the knowledge base:

- String Similarity
- Attributes of entity – age, date of birth etc.
- Contextual Similarity – word embeddings are great for this
- Popularity
- Relationships between entities

Disambiguation

Given these features, how do we solve the problem of choosing a referent?

Disambiguation

Given these features, how do we solve the problem of choosing a referent?

Typically treat the problem as a Learning to Rank task comprised of two parts:

- Local Similarity
  - The direct similarity bewteen a mention of an entity and a candidate referent
  - Features include string similarity, contextual similarity, popularity, and attributes.
- Global Coherence
  - Entities mentioned in the same context are probably linked by some topic. So, the correct referents likely have some relationship in the knowledge base.
  - Usually a graph problem.

Highest ranked candidate for each entity is chosen as the referent.

Disambiguation

For learning feature weights and candidate ranks, Support Vector Machines and Conditional Random Fields are popular, but require labeled training data.

In the absence of training data, ranking can be based on raw similarity metrics computed for global coherence and local similarity

Conclusion

- There are many opportunities for Machine Learning in Digital Humanities
- I think Entity Disambiguation has great potential in this field
- Disambiguation in a nutshell
  1. Pick your knowledge base
     - Make sure your entities are appropriately represented
  2. Identify your candidates
     - Focus on recall over precision. There is a balance to be struck, but we're going to filter candidates later anyway
  3. Disambiguate
     - Solve in two parts - local and global
     - Learning to Rank view is popular: SVMs and CRFs are common
     - Innovate! Know what you have available to you and exploit

Thank You

Gary.Munnelly@adaptcentre.ie

github.com/munnellg