Cross Document Person Name Disambiguation Using Entity Profiles

Harish Srinivasan
John Chen and Rohini Srihari

Janya Inc.

IIWEB July 11, 2009
Overview

Introduction

Motivation

Previous Work

Entity Profiles

Features

Model and Approach

Experiments and Results

Conclusion
Outline

Introduction
Motivation
Previous Work
Entity Profiles
Features
Model and Approach
Experiments and Results
Conclusion
Introduction

- Cross document entity coreference: Linking entities in one document to the “same” entity in other documents
- Different entities often have the same name (name mention) - Disambiguation
- Same entities go by different names - Alias detection

Problem Formulation: Given a corpus and an ambiguous name, cluster the documents such that each cluster contains all and only those documents that mention the same person.

Salisbury is getting ready to celebrate the 400th anniversary of Captain John Smith’s voyage through the Chesapeake. His explorations of the Delmarva Peninsula were a vital part of his journey. That’s why the Salisbury Zoo is making plans for the Delmarva Trail, which will include an endangered red wolf exhibit. This comes at the same time many are pushing for a national water trail tracing Smith’s steps across nearly 3,000 miles.

“This water trail will commemorate that, but also provide an avenue for nature and heritage tourism for the region,” said Joel Dunn, Conservation Fund.

Congress has authorized a feasibility study to establish the national water trail commemorating Smith’s travels in 1608.

COLUMBIA, S.C. - Great Falls coach John Smith has become the winningest boy’s basketball coach in state history after the Red Devils beat Blacksburg 74-67 in the Class A playoffs Saturday night.


Smith has been with Great Falls for 37 years and won six Class A titles in 15 tries. His overall record is 743-225 and he has just one losing season.

Smith treated win No. 743 just like any other, The (Rock Hill) Herald reported on its Web site.

Smith shook hands with Blacksburg’s players and coaches, then headed to the locker room to talk to his team. There was no mention of the record.

Chris Huhe’s audacious bid to lead the Liberal Democrats resulted in a strong showing but no victory.

He has only been an MP for nine months, a period of time many party members appear to believe may have been too short to have proven his ability as a leader. Many of the party’s younger MPs had given their backing to Sir Menzies Campbell.

But Huhe, not unlike Tony Blair in the wake of John Smith’s death, saw an opportunity and decided to go for it.

The fourth contender to declare his candidacy, he was virtually unknown at the outset of the campaign.

At first written off as a no-hoper, the Treasury spokesman benefited from Mark Oaten’s decision to quit the race, revelations about Simon Hughes’s private life and an efficiently run campaign.
Outline

Introduction

Motivation

Previous Work

Entity Profiles

Features

Model and Approach

Experiments and Results

Conclusion
Motivation

- 10% of news text consists of proper names
- Growing size of corpus and generality - A challenge for NLP
- Consolidating entity information from a large corpus.
  - Question answering systems
- Problem by nature is interesting !!
  - Solution solely relies on understanding context around entity mention.
- Increased attention in the last few years
  - SemEval 2007 - Disambiguating names from workshop
  - TAC 2009 (Current) - Entity disambiguating and linking to knowledge base
Related Work

- Bagga and Baldwin (1998) - First to address this problem
  - Bag of words approach
  - B-Cubed F-Measure scoring metric
- Mann and Yarowsky(2003), Chen and Martin (2007)
  - Unsupervised learning approaches
- Our Work
  - Builds on features on top of Chen and Martin (2007)
  - Utilizing topic model features, entity profiles.
Outline

Introduction
Motivation
Previous Work
Entity Profiles
Features
Model and Approach
Experiments and Results
Conclusion
Entity Profile: A summary of the entity that combines in one place features of the entity, attributes of the entity, relations (to or from another entity), and events that this entity is involved in as a participant.

According to the Boston Globe, the al-Barakaat network of Dubai was founded by Osama bin Laden in the late eighties... Mohamed Barre has been the money transfer agency’s broker...
Entity Profiles cont...

- Profile Name (longest name mention in the document)
  - *John Smith, Smith, He* ——> *John Smith*

- Attributes - *President, Captain*

- Links to other entities - Persons, Organization, Location etc.

- Events involved - *Attack, Kill*

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRF_NAM</td>
<td>John Smith</td>
</tr>
<tr>
<td>CE_MODIFIERS</td>
<td>Still alive</td>
</tr>
<tr>
<td>EVENTS_INVOLVED</td>
<td>Ran into</td>
</tr>
<tr>
<td>CE_PER_TITLE</td>
<td>Captain</td>
</tr>
<tr>
<td>Ce_Association_Entity</td>
<td>Joe Grahame</td>
</tr>
</tbody>
</table>
Cross Document Entity Profiles

**Mohammad El Baradei** (Strength: 722) View documents | events

196 mentions in 72 documents

**ALIASES**

Displaying 1-5 of 11

- ElBaradei
- Mohamed El Baradei
- Mohammad Al-Baradi'i
- Mohamed ElBaradei

**ASSOCIATIONS**

**People**

7 of 7

- George W. Bush
- Ali Larijani
- Jacques Chirac
- Konstantin Kosachyov
- Ali Asghar Soltaniyeh
- Mohamed
- Sabah Zanganeh

**Organizations**

1-10 of 12

- IAEA
- UN Security Council
- Agence France-Presse
- IAEA Board of Governors
- IAEA Meeting
- IAEA Resolution
- al-Baradi'i

**Places**

4 of 4

- Iran
- United States
- Russia
- Tehran

**RELATED INFORMATION**

**Affiliations**

Displaying 1-5 of 6

- Agency
- IAEA
- International Atomic Energy Agency
- al-Baradi'i
- during his meeting with IAEA

**Description**

Displaying 1-5 of 14

- Director General
- chief
- head

**Management affiliations**

Displaying 2 of 2

- Agency
- IAEA

**Positions**

Displaying 1-5 of 10

- Director General
- Dr
- chief
- director
- director general

**QUOTES**

Displaying 5 of 5

"Due process, therefore, must take its course before the Secretariat (the IAEA administration) is able to submit a detailed report,"

"They have not said anything"
Outline

Introduction
Motivation
Previous Work
Entity Profiles
Features
Model and Approach
Experiments and Results
Conclusion
Features

1. Summary (S)
   ▶ All sentences which contain the ambiguous person’s full name, last name.
   ▶ Include coreference sentences (*Nominal* and *Pro-Nominal*).

2. Base Noun Phrases (BNP)
   ▶ Noun recursive noun phrases in sentence of mention

3. Document Entities (DE)
   ▶ Named entities and nominals of organization, vehicle, weapon, location etc.
   ▶ Persons other than ambiguous name.

4. Profile Features (PF)
   ▶ All attributes and relational features (forward and backward)
     - Two tuple Attribute-Value pairs.

5. Topic Model Features (TM)
Profile Features

- Reinstate critical information about entity on top of summary term features
  - Doc 1: “Captain John Smith first beheld American strawberries in Virginia”
  - Doc 2: “John Smith had strawberries”
  - Doc 3: “… discovered by Captain John Smith”

- Feature Population: Non stop words in their morphological form
Topic Model Features

- Bag of words approach fails to capture abstract similarity
  - island, bay, water, ship vs.
  - founder, voyage, captain

- Every document assigned with a probability distribution over set of topics

- Every topic associated with a probability distribution over words
Topic model learnt from all nouns in the corpus (extracted for each document)

Total number of topics - 50

Top 10 words with highest joint probability are chosen.

\[ P(w, t|D) = P(w|t, D) \times P(t|D) = P(w|t) \times P(t|D) \]

where \( w, t \) and \( D \) are word, topic and document respectively.

Feature population: Morphological forms of these 10 words appended to features.
Features - Example

Document
Salisbury is getting ready to celebrate the 400th anniversary of Captain John Smith's voyage through the Chesapeake. His explorations of the Delmarva Peninsula were a vital part of his journey. That's why the Salisbury Zoo is making plans for the Delmarva Trail, which will include an endangered red wolf exhibit. This comes at the same time many are pushing for a national water trail tracing Smith's steps across nearly 3,000 miles.

"This water trail will commemorate that, but also provide an avenue for nature and heritage tourism for the region," said Joel Dunn, Conservation Fund.

Congress has authorized a feasibility study to establish the national water trail commemorating Smith's travels in 1608.

Summary
Salisbury is getting ready to celebrate the 400th anniversary of Captain John Smith's voyage through the Chesapeake.

His explorations of the Delmarva Peninsula were a vital part of his journey.

Terms
salisbury, readi, celebr, anniversari, captain, voyag, chesapeake, explor, delmarva, peninsula, vital, part, journei

Profile Features

Base Noun Phrases

Document Entities

Entities: chesapeake, delmarva peninsula, salisbari zoo, delmarva trail, region, conserv, congress
Outline

Introduction

Motivation

Previous Work

Entity Profiles

Features

Model and Approach

Experiments and Results

Conclusion
Model and Approach

IE pipeline (Semantex)

Disambiguation Model

- Bag of words and phrases obtained from features
- Compute TF, IDF
  - Prefix match is used for term frequency.
  - ‘Capt’ vs ‘Captain’
  - Ambiguous name used as stop word
- Vector Space Model populated with features.
- Hierarchical agglomerative clustering with single linkage
Similarity and Clustering

- Cosine similarity (1)
- Log-transformed measure (2).

\[
\text{Sim}(S_1, S_2) = \sum_{\text{common terms } t_j} w_{1j} \times w_{2j},
\]

where \( w_{ij} = \frac{\text{tf} \times \ln \frac{N}{df}}{\sqrt{s_{i1}^2 + s_{i2}^2 + \ldots + s_{in}^2}} \) (1)

where \( S_1 \) and \( S_2 \) are the term vectors for which the similarity is to be computed. \( \text{tf} \) is the frequency of the term \( t_j \) in the vector. \( N \) is the total number of documents. \( df \) is the number of documents in the collection that the term \( t_j \) occurs in. The denominator is the cosine normalization.

\[
\text{Sim}(S_1, S_2) = \sum_{\text{common terms } t_j} w_{1j} \times w_{2j},
\]

where \( w_{ij} = \frac{\ln \left( \frac{\text{tf} \times \ln \frac{N}{df}}{\sqrt{s_{i1}^2 + s_{i2}^2 + \ldots + s_{in}^2}} \right)}{\sqrt{s_{i1}^2 + s_{i2}^2 + \ldots + s_{in}^2}} \) (2)

- Hierarchical agglomerative clustering
- Single linkage (Best amongst others)
Outline

Introduction

Motivation

Previous Work

Entity Profiles

Features

Model and Approach

Experiments and Results

Conclusion
<table>
<thead>
<tr>
<th>Ambiguous Name</th>
<th>John Smith</th>
<th>James Jones</th>
<th>John Smith</th>
<th>Michael Johnson</th>
<th>Robert Smith</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corpus</td>
<td>Bagga Baldwin</td>
<td>English Boulder</td>
<td>English Boulder</td>
<td>English Boulder</td>
<td>English Boulder</td>
</tr>
<tr>
<td>Total No of Documents</td>
<td>197</td>
<td>104</td>
<td>112</td>
<td>101</td>
<td>100</td>
</tr>
<tr>
<td>No of Clusters (Unique Names)</td>
<td>35</td>
<td>24</td>
<td>54</td>
<td>52</td>
<td>65</td>
</tr>
</tbody>
</table>

- Clusters are highly skewed
- Several clusters with a single document
### Results

<table>
<thead>
<tr>
<th>Ambiguous Name</th>
<th>John Smith (Bagga)</th>
<th>James Jones</th>
<th>John Smith (Boulder)</th>
<th>Michael Johnson</th>
<th>Robert Smith</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Chen and Martin - Optimal Threshold</strong> - S+BNP+DE (Separate bag of words + Soft TF-IDF)</td>
<td>92.02</td>
<td>97.10</td>
<td>91.94</td>
<td>92.55</td>
<td>93.48</td>
<td>93.41</td>
</tr>
<tr>
<td><strong>Chen and Martin - Fixed Stop Threshold</strong> - S+BNP+DE (Separate bag of words + Soft TF-IDF)</td>
<td>-</td>
<td>96.64</td>
<td>91.31</td>
<td>90.57</td>
<td>86.71</td>
<td>91.31</td>
</tr>
<tr>
<td><strong>Baseline</strong> - S+BNP+DE (Separate bag of words)</td>
<td>84.20</td>
<td>98.11</td>
<td>85.50</td>
<td>90.79</td>
<td>90.37</td>
<td>89.79</td>
</tr>
<tr>
<td><strong>Baseline + Log Transformed</strong></td>
<td>93.96</td>
<td>90.54</td>
<td>86.80</td>
<td>89.52</td>
<td>92.66</td>
<td>90.69</td>
</tr>
<tr>
<td><strong>Model (Single bag of words + Log Transformed Tf-Idf)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S+BNP+DE</td>
<td>92.28</td>
<td>95.48</td>
<td>89.50</td>
<td>91.64</td>
<td>92.42</td>
<td>92.26</td>
</tr>
<tr>
<td>S+BNP+DE + PF (A)</td>
<td>91.93</td>
<td>98.14</td>
<td>91.46</td>
<td>90.22</td>
<td>92.54</td>
<td>92.85</td>
</tr>
<tr>
<td>A + Nsw</td>
<td>92.77</td>
<td>98.14</td>
<td>90.56</td>
<td>89.85</td>
<td>93.22</td>
<td>92.90</td>
</tr>
<tr>
<td>A + Nsw + Ptf</td>
<td>92.83</td>
<td>98.14</td>
<td>91.24</td>
<td>93.27</td>
<td>94.27</td>
<td>93.95</td>
</tr>
<tr>
<td>A + Nsw + Ptf + TM (Fixed Stop Threshold)</td>
<td><strong>92.62</strong></td>
<td><strong>99.03</strong></td>
<td><strong>91.49</strong></td>
<td><strong>94.01</strong></td>
<td><strong>93.03</strong></td>
<td><strong>94.03</strong></td>
</tr>
<tr>
<td>A + Nsw + Ptf + TM</td>
<td>92.42</td>
<td>97.28</td>
<td>89.3(dev)</td>
<td>90.3(dev)</td>
<td>92.12</td>
<td>92.28</td>
</tr>
</tbody>
</table>

Results cont...

<table>
<thead>
<tr>
<th>Ambiguous Name</th>
<th>John Smith (Bagga)</th>
<th>James Jones</th>
<th>John Smith (Boulder)</th>
<th>Michael Johnson</th>
<th>Robert Smith</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>A + Nsw + Ptf + TM</td>
<td>92.62</td>
<td>99.03</td>
<td>91.49</td>
<td>94.01</td>
<td>93.03</td>
<td>94.03</td>
</tr>
</tbody>
</table>

Model (Separate bag of words with all features)

<table>
<thead>
<tr>
<th></th>
<th>Separate - Average</th>
<th>Separate - NN</th>
<th>Separate - MaxEnt</th>
</tr>
</thead>
<tbody>
<tr>
<td>A + Nsw + Ptf + TM</td>
<td>93.01</td>
<td>94.40</td>
<td>92.69</td>
</tr>
<tr>
<td>Separate - Average</td>
<td>98.37</td>
<td>99.03</td>
<td>98.14</td>
</tr>
<tr>
<td>Separate - NN</td>
<td>81.65</td>
<td>86.26</td>
<td>86.94</td>
</tr>
<tr>
<td>Separate - MaxEnt</td>
<td>87.34</td>
<td>89.19</td>
<td>88.92</td>
</tr>
<tr>
<td>Average</td>
<td>92.27</td>
<td>90.98</td>
<td>91.71</td>
</tr>
</tbody>
</table>

- Separate bag of words model - Worse
  - Restriction in matching terms across different bag of words
Outline

Introduction

Motivation

Previous Work

Entity Profiles

Features

Model and Approach

Experiments and Results

Conclusion
Conclusion

- Merging of document level entity profiles to create corpus level profiles.
- High F-Measures for person name disambiguation.
- Technique can be used for organization and location with minimal changes.
- Disambiguation along with Alias detection is a bigger challenge.
  - With aliases the search space for same entities becomes too large
  - Baseline performance can be reached with string matching variants and use of known alias lexicon.