Named Entity Disambiguation And Linking

Entity Based Cross-Document Coreferencing using the Vector Space Model

Andreas Schwarte

Text Mining Seminar – WS 09/10

Outline

- Motivation
- Problem Statement and Implementation Idea
- Architecture and Methodology
- Evaluation
- Conclusion

- Paper for this presentation
  - Bagga, & Baldwin. (1998). *Entity-Based Cross-Document Coreferencing Using the Vector Space Model*.

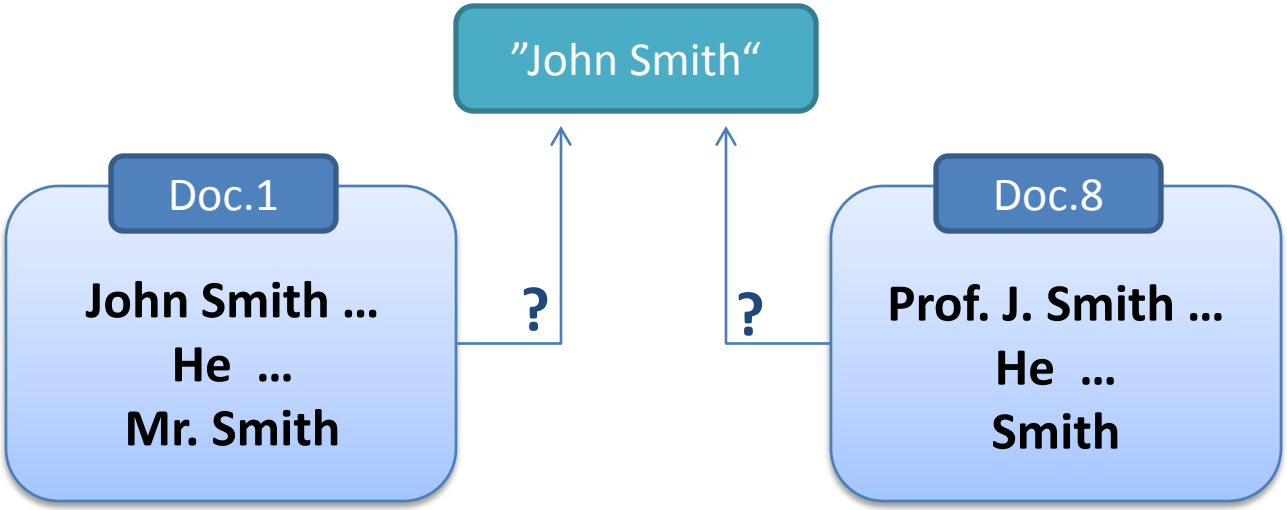
Motivation

- Disambiguation of Named Entities across documents

"Is John Smith mentioned in doc1 the same person as Prof. J. Smith mentioned in doc8?"

"How can cross document coreference of entities be evaluated by a computer system?"

"How can a scoring model be implemented which reflects similarity of entities?"



Problem Statement and Implementation Idea

The Problem:

Multiple text sources mention the same name, place or concept, possibly in slight variations . How can a computer system "decide" if the instances reflect the same entity?

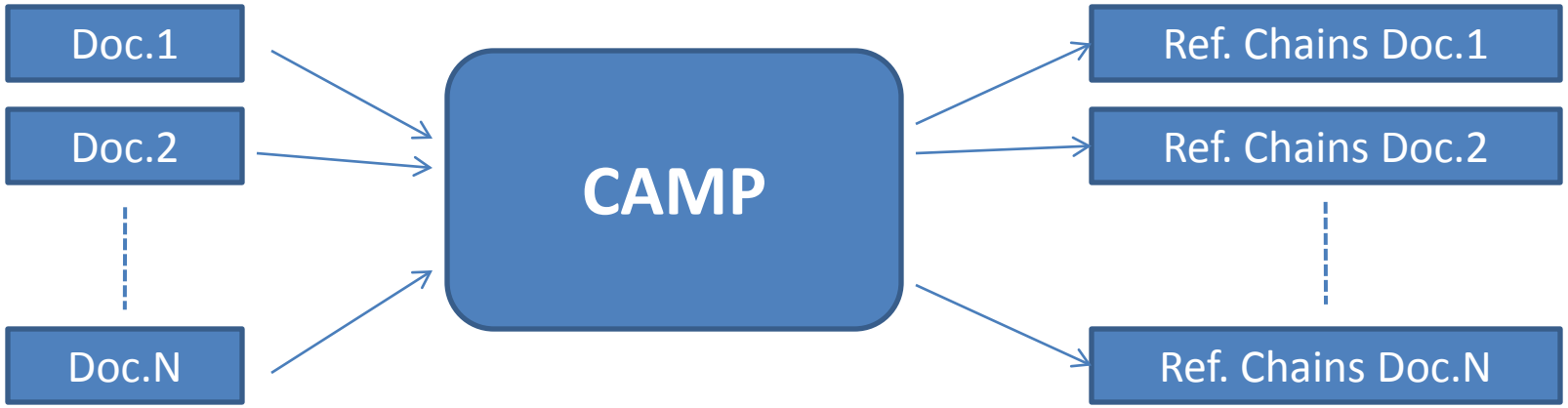
Bagga's & Baldwin's Solution Idea:

Mapping of the within-document coreference problem to a cross-document co-reference problem and evaluation of similarity by means of the vector space model.

Architecture and Methodology (1)

Step 1:

Building intra document co-reference chains



CAMP: University of Pennsylvania‘s Pennlight Coreference System

Example →

Architecture and Methodology (2)

Step 1 – Example

John Perry, of Weston Golf Club, announced his resignation yesterday. He was the President of the Massachusetts Golf Association. During his two years in office, Perry guided the MGA into a closer relationship with the Women's Golf Association of Massachusetts.

Figure 2: Extract from doc.36

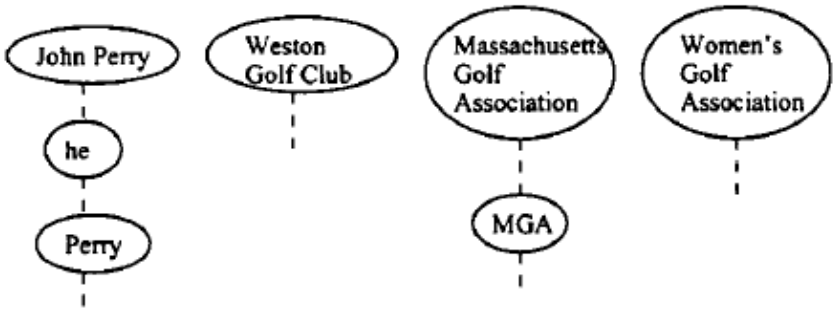


Figure 3: Coreference Chains for doc.36

Oliver "Biff" Kelly of Weymouth succeeds John Perry as president of the Massachusetts Golf Association. "We will have continued growth in the future," said Kelly, who will serve for two years. "There's been a lot of changes and there will be continued changes as we head into the year 2000."

Figure 4: Extract from doc.38

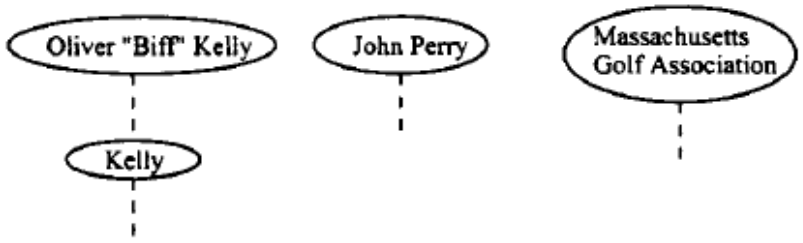


Figure 5: Coreference Chains for doc.38

Architecture and Methodology (3)

Step 2:

Producing "Summaries" using the *SentenceExtractor*

→ Take all sentences from the input document in which the respective entity names occurs (e.g. for Doc.36 this is the complete document, for Doc.38 the first sentence)

These summaries serve as input for the next step

Architecture and Methodology (4)

Step 3:

Disambiguating entities using vector space similarity

- Summaries are mapped to N-dim. feature vectors
  - Each vector component corresponds to a distinct term
  - Weighting is done by means of *tf*idf* of the term in the document (tf ~ term frequency; idf ~ inversed doc. freq.)
- Decision based on some similarity metric: sim(S1,S2)
  - Commonly cosine similarity of summary feature vectors

Illustration →

Architecture and Methodology (5)

Step 3 – Illustration:

- Doc.1: {“John Smith is a president of the Massachusetts Golf Association. Smith is member of this association since 1999.”}
- Doc.2: {“John Perry Smith became member of the Golf Association in 1999. Since 2005 he is the president.”}

	D1	D2
John	1	1
Smith	2	1
President	1	1
Golf	1	1
…	…	…

$$Sim(D1, D2) = \sum w_{1i} * w_{2i}$$

$$Sim(D1, D2) > threshold \Rightarrow same$$

- Very Simplified Illustration
  - N-Dimensional vector (stop words like ‘a’ are eliminated)
  - Plain Term Frequency is used (idf is omitted)
  - Similarity Based on Scalar Product (Cosine Normalization omitted)

Evaluation (1)

- Experiments based on highly ambiguous dataset
  - 197 NY Times articles, 35 different John Smiths
- MUC Coreference Scoring algorithm
  - Based of Precision and Recall estimates, roughly

    Precision: how precise, i.e. how many "correct" items

    Recall: how many of the correct links are found
  - Hand Marked "Truth" set, versus System Output
  - F-measure: Weighted Harmonic Mean of Precision and Recall

Details →

Evaluation (2)

- Recall and Precision estimation based on sets

  S ~ equivalence set generated by the “Truth”

  R ~ equivalence classes generated by the “Response”

  p(s) ~ partitioning of S relative to the “Response”

  * For precision the role of “Truth” and “Response” is reversed

$$R = \frac{\sum |S_i| - |p(S_i)|}{\sum |S_i| - 1} \qquad P = \frac{\sum |S'_i| - |p'(S'_i)|}{\sum |S'_i| - 1}$$

Example →

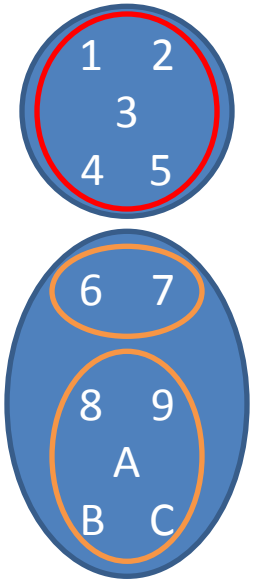
Evaluation (3)

"Truth": S1={1,2,3,4,5} S2={6,7} S3={8,9,A,B,C}

"Response": R1={1,2,3,4,5} R2={6,7,8,9,A,B,C}

$$R = \frac{(5-1)+(2-1)+(5-1)}{(5-1)+(2-1)+(5-1)} = \frac{9}{9}$$

$$P = \frac{(5-1)+(7-2)}{(5-1)+(7-1)} = \frac{9}{10}$$



Partition w.r.t. response

Partition w.r.t. key

Evaluation (4)

- Bagga & Baldwin developed B-Cubed Scoring Algorithm
  - Overcome shortcomings of MUC-6 (“errors are equal”)
  - Model accuracy on a per-document basis (“weighting”)

Output	MUC Algorithm	B-CUBED Algorithm (equal weights for every entity)
Example 1	P: $\frac{9}{10}$ (90%)	P: $\frac{1}{12} * [\frac{5}{5} + \frac{5}{5} + \frac{5}{5} + \frac{5}{5} + \frac{5}{5} + \frac{2}{7} + \frac{2}{7} + \frac{5}{7} + \frac{5}{7} + \frac{5}{7} + \frac{5}{7}] = 76\%$
	R: $\frac{9}{9}$ (100%)	R: $\frac{1}{12} * [\frac{5}{5} + \frac{5}{5} + \frac{5}{5} + \frac{5}{5} + \frac{5}{5} + \frac{2}{2} + \frac{2}{2} + \frac{5}{5} + \frac{5}{5} + \frac{5}{5} + \frac{5}{5}] = 100\%$

$$\text{Precision}_i = \frac{\textit{number of correct elements in the output chain containing entity}_i}{\textit{number of elements in the output chain containing entity}_i} \quad (1)$$

$$\text{Recall}_i = \frac{\textit{number of correct elements in the output chain containing entity}_i}{\textit{number of elements in the truth chain containing entity}_i} \quad (2)$$

- The values refer to the example on the previous slides.
- Weighting in this example is done uniformly, i.e. w=1/N

Conclusion

- Disambiguation method results are quite promising
  - F-Measure Performance: 84.6%
  - Previous methods (NetOwl/Textract) could not distinguish John Smiths of this data set -> poor results
- Other methodologies and ideas
  - Unsupervised Clustering techniques (Patterns)
  - Classification based on a Maximum Entropy Model

Thanks for Your Attention!

References

Bagga, & Baldwin. (1998). *Entity-Based Cross-Document Coreferencing Using the Vector Space Model*.

Bagga, & Baldwin, et al. (1998). *Description of the UPENN CAMP System as used for Coreference*.

Vilain, Marc et al. (1995) *A Model-Theoretic Coreference Scoring Scheme*, Proceedings of the MUC-6 Conference.