Named Entity Disambiguation and Linking
Unlocking the Secrets of the Past: Text Mining for Historical Documents (WS 2008/09)

Tassilo Barth
Saarland University

18. February 2009
Quiz

Who is Michael Jackson?

- a) A British beer guru
- b) A pop star
- c) None of these


Problem: One name, multiple persons
Quiz

**Who is Michael Jackson?**

- a) A British beer guru
- b) A pop star
- c) None of these

Who is Michael Jackson?

- a) A British beer guru
- b) A pop star
- c) None of these


Problem: One name, multiple persons
Overview

1. Introduction

2. Methods
   - Bagga/Baldwin 1998: Entity-Based Cross-Document Coreferencing Using the Vector Space Model
   - Mann/Yarowsky 2003: Unsupervised Personal Name Disambiguation
   - Fleischman/Hovy 2004: Multi-Document Person Name Resolution

3. Conclusion
Definition

**Named Entity Disambiguation** is the task of exploring which real person (e.g. pop star, beer guru), place, event.. is referred to by a certain instance of a name (e.g. Michael Jackson in a certain context).

**Named Entity Linking** means to connect all references to the correct entity.
What is it good for?

- Basically for all tasks related to Text Mining (Text Summarization, Question Answering..)
- In the historical domain: For example extracting biographical information about one person from a couple of sources
- Named Entity Disambiguation especially necessary when there is more than one source.
Basic Terminology

Michael Jackson released a new album

Michael Jackson (1942-2007) was a beer lover
Basic Terminology

Michael Jackson released a new album

Michael Jackson (1942-2007) was a beer lover

Referent / Entity

Reference

Reference

Reference
Basic Terminology

Michael Jackson released a new album

Michael Jackson (1942-2007) was a beer lover
Basic Terminology

Reference

Instance

Reference

Referent / Entity

Michael Jackson released a new album

Features:

{beer, lover}

{brthyr: 1942}

{...}
Given: A set of references together with context (usually the sentence in which they occur. e.g.: „Michael Jackson, the famous British beer guru, ...“, „M.J. has finished a new album..“)
Basic Approach

**Given:** A set of references together with context (usually the sentence in which they occur. e.g.: „Michael Jackson, the famous British beer guru, ...“, „M.J. has finished a new album..“)

1. Find a suitable representation for the reference’s context (usually, a set of characteristic elements distinguishing the reference: features)
Basic Approach

**Given:** A set of references together with context (usually the sentence in which they occur. e.g.: „Michael Jackson, the famous British beer guru, ...“, „M.J. has finished a new album..“)

1. Find a suitable representation for the reference’s context (usually, a set of characteristic elements distinguishing the reference: features)
2. Calculate similarity between two references
Given: A set of references together with context (usually the sentence in which they occur. e.g.: „Michael Jackson, the famous British beer guru, ...“, „M.J. has finished a new album."

1. Find a suitable representation for the reference’s context (usually, a set of characteristic elements distinguishing the reference: features)

2. Calculate similarity between two references

3. Create clusters of references that belong to one entity
Overview

1. Introduction

2. Methods
   - Bagga/Baldwin 1998: Entity-Based Cross-Document Coreferencing Using the Vector Space Model
   - Mann/Yarowsky 2003: Unsupervised Personal Name Disambiguation
   - Fleischman/Hovy 2004: Multi-Document Person Name Resolution

3. Conclusion
Motivation + Source data

- **Purpose**: **Multi-Document Summarization**: Automatically create a summary of two or more documents
- 173 articles from NYT
- 11 different „John Smiths“ mentioned
- For a single document, create a list of terms that refer to a certain person (e.g. „John Smith“, „he“..) = summary
Example: Two documents

Looking for „John Perry“

**Document 1**

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.

**Document 2**

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.‟
Looking for „John Perry“

Document 1
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.

Document 2
Oliver „Biff“ Kelly of Weymouth succeeds John Perry as president of the Massachusetts Golf Association.
Features

- Idea: Assign every word in summaries a weight
- If word occurs often in current summary: Increase weight
- If word occurs seldom in other documents: Increase weight (since it seems to be a specific cue for this summary)
- Example: „president“ $\Rightarrow$ 0.6, „Massachusetts Golf Association“ $\Rightarrow$ 0.8
- Measure called „Term Frequency / Inverse Document Frequency“ (TF/IDF)
Two summaries are similar if:

- They share many words
- Shared words have a high weight
- Example: "president" and "Massachusetts Golf Association" are terms that occur in both summaries
- Their weights are combined (by using "Cosine Distance") and result in a single output value
- The higher this value, the more similar the summaries are
Finally: Clustering

Calculate the similarity of any two summaries
If value is above a certain threshold $\Rightarrow$ Assume they are referring to the same entity
Overview

1. Introduction

2. Methods
   - Bagga/Baldwin 1998: Entity-Based Cross-Document Coreferencing Using the Vector Space Model
   - Mann/Yarowsky 2003: Unsupervised Personal Name Disambiguation
   - Fleischman/Hovy 2004: Multi-Document Person Name Resolution

3. Conclusion
Source Data

Sentences extracted from web, containing references. For example:

- „Early in his career, Jim Clark was involved in one of the worst accidents in the history of Formula 1 racing“
- „There’s quite an important debate raging on the Cypherpunks list these days over Netscape CEO Jim Clark […]“

Convert these extracts to a bundle of features!
Features

1. Default: A list of words / nouns in the context (e.g., "career", "accident")
2. Plus: Apply term weights (as seen before) and use only most relevant words (e.g., "accident")
4. Use these to give higher weight to terms looking like biographical features (e.g., "CEO" for occupation)
Mann/Yarowsky 2003: Unsupervised Personal Name Disambiguation

**Similarity + Clustering**

- **Idea**: Instances with identical biographical features share probably one referent. Merge them first.
- **Next**, apply the similarity measure seen before to form bigger clusters.
- **In later stages of clustering**: Few "outlaw" single instances left, would be merged with random bigger clusters.
- "Cluster Refactoring": After the first few clustering steps, save the clusters formed that far (high quality / accuracy) as "seeds".
- **Later on**: Assign all remaining instances to the most similar seed.
Example: Clustering for name „Jim Clark“

- driver, racing \( \text{occ: driver} \)
- championship \( \text{occ: driver} \)
- rally
- Netscape \( \text{occ: founder} \)
- Software
- browser, graphics
- prevention
Example: Clustering for name „Jim Clark“

- driver, racing
- championship
- rally
- Netscape
- Software
- browser, graphics
- prevention
Example: Clustering for name „Jim Clark“
Example: Clustering for name „Jim Clark“

- Seed 1:
  - Driver, racing: occ: driver
  - Championship: occ: driver
  - Rally

- Cluster 1

- Seed 2:
  - Netscape: occ: founder
  - Software
  - Browser, graphics
  - Prevention

- Cluster 2
Overview

1. Introduction

2. Methods
   - Bagga/Baldwin 1998: Entity-Based Cross-Document Coreferencing Using the Vector Space Model
   - Mann/Yarowsky 2003: Unsupervised Personal Name Disambiguation
   - Fleischman/Hovy 2004: Multi-Document Person Name Resolution

3. Conclusion
Fleischman/Hovy: Source data

- Set of 2 million pairs: <Concept [e.g. „beer expert“], Name [e.g. „Michael Jackson“]>
- Originally extracted from a newspaper corpus
- 2675 pairs selected, each one randomly connected to another pair with same name but different concept (e.g. <„beer expert“, „Michael Jackson“> connected to <“King of Pop“, „Michael Jackson“>)
Features

- **Name features:** commonality, fame
- **Overlap (of sentential context):** Ratio of words shared between two contexts
- **Semantic features:** Different semantic measures to grasp the semantic similarity of both concepts (e.g. have “King of Pop“ and “beer guru“ a similar meaning?)
- **Statistics features:** Several probabilites (likelihood that two instances share the same referent given their concepts, etc.)
- **Web features:** Web search hits for several queries (e.g. “Michael Jackson +beer expert“)
Similarity measure

- A statistical model (Maximum Entropy) is used
- Training process: model should arrange weight of features such that evidence can be explained
- Usage: Feed with created instance pairs, get a probability that both references point to same person
Imagine: Each instance is connected to each other by edges specifying their similarity.

As long as maximum similarity above a threshold (e.g. 70%):
Merge two instances with max. similarity into one.

Result: one or more clusters with max. inter-cluster similarity less than threshold.
Example: Clustering for name „Michael Jackson“

- <Michael Jackson, beer expert>
- <Michael Jackson, pop star>
- <Michael Jackson, beer guru>
- <Michael Jackson, King of Pop>
Example: Clustering for name „Michael Jackson“
Example: Clustering for name „Michael Jackson“

\[
\begin{align*}
\text{< Michael Jackson, beer expert>} & \quad 0.80 \\
\text{< Michael Jackson, pop star, King of Pop>} & \quad 0.55 \\
\text{< Michael Jackson, beer guru>} & \quad 0.50 \\
\end{align*}
\]

0.80 > 0.7: CLUSTER
Example: Clustering for name „Michael Jackson“

< Michael Jackson, pop star, King of Pop >

< Michael Jackson, beer guru, beer expert >

0.52 < 0.7: STOP
Overview

1. Introduction

2. Methods
   - Bagga/Baldwin 1998: Entity-Based Cross-Document Coreferencing Using the Vector Space Model
   - Mann/Yarowsky 2003: Unsupervised Personal Name Disambiguation
   - Fleischman/Hovy 2004: Multi-Document Person Name Resolution

3. Conclusion
Conclusion

Comparison between methods is hard. Differences in:

- ... used data (newspaper articles, hand-annotated and preprocessed data set, web pages...)
- ... evaluation methods („Pseudonames“, match against hand-annotated „gold standard“..)
- ... evaluation measures (General „Accuracy“, F-Measure, example clusterings..)
Several features have been proven useful:

- Biographical features
- Statistics features
- Web search hits (e.g. Google page count)
- Term weight
- etc. pp.

Text mining in historical domain: Biographical features of special importance?
References


Gideon Mann; David Yarowsky. Unsupervised Personal Name Disambiguation. CoNLL-03. 2003

Thanks for your attention!