Named Entity Disambiguation in Digital Libraries

Le Dieu Thu

Supervisors: Raffaella Bernardi
Massimo Poesio
Patrick Blackburn
1. Introduction

2. Author Name Disambiguation: State of the art

3. Our Disambiguation Framework
   - Blocking module
   - Metadata Enriching module
   - Feature representation & cosine similarity
   - Sparsity problem & Dimensionality Reduction
   - Clustering module

4. Experiments & Discussion

5. Conclusions
Introduction

- **Ambiguous author name:** Different authors having the same name
- **Aim:** Disambiguate those authors

R Smith  Author of  Books

- High finance in the Euro zone: competing in the new European capital market
- Governing the modern corporation: capital markets, corporate control and economic performance
- The money wars: the rise and fall of the great buyout boom of the 1980s
- Ecology and field biology
- Elements of ecology
Bolzano Library Catalogue Searching

- **Classification Numbers**: encodings (QK 620, WI 2000), to organize books on shelves
- **Subject Headings**: keywords

- **High finance in the Euro zone**: competing in the new European capital market
  - QK 620: Finance, Economics
  - WI 2000: Ecology, Biology

- **Ecology and field biology**
  - QK 620: Finance, Economics
  - WI 2000: Ecology, Biology

**Searching by CN & SH:**

**Goal**: Develop a disambiguation method that can be applied in any catalogue (even without manual annotations, e.g. CN & SH)

*Not of support in Digital Libraries (CiteSeer, DBLP)*
Metadata Enriching Module

Wikipedia Dataset -> Model Estimation -> Estimated Topic Model

Topic Models:

Ambiguous Candidates

Enriched Ambiguous Candidates
Hidden Topic Discovery from Documents

Latent Dirichlet Allocation (LDA) [Blei et al. 2003]

Document

For decades, German software giant SAP (SAP) has been
relying on organic growth. During the
decade, the company has spent a relatively modest $1
billion on acquisitions. During this period, rival
Oracle (ORCL) has announced plans to spend billions
of dollars, according to research analysts at CitiGroup (C). But all of
that changed on Oct. 7 when SAP said it would make a
$3.5 billion to $4 billion bid for
SAP AG (SAP), a business intelligence and data mining company based in France.

It's a sign of how mergers and acquisitions will reshape the
software sector in the months and years ahead. Growth in
software is to
Now that they're multi-billion dollar deals, and as investors ask
how much they're being paid to acquire the
editors in the

Hidden Topic Analysis/Discovery

Topics \{1, 2, …, K\} are unknown (i.e., hidden and need to be discovered)
Wikipedia data preprocessing

Wikipedia → Dumps in SQL & XML → HTML removal → Punctuation Stop-word Removal → Stop-word List → 266M data

80,000 documents

Computer science or computing science (sometimes abbreviated CS) is the study of the theoretical foundations of information and computation, and of practical techniques for their implementation and application.
**Model Estimation using LDA & Gibbs Sampling**

Sample topics extracted from the estimated model

<table>
<thead>
<tr>
<th>Topic 0</th>
<th>Topic 8</th>
<th>Topic 23</th>
<th>Topic 39</th>
<th>Topic 68</th>
<th>Topic 86</th>
<th>Topic 96</th>
</tr>
</thead>
<tbody>
<tr>
<td>company</td>
<td>album</td>
<td>cells</td>
<td>law</td>
<td>storm</td>
<td>war</td>
<td>school</td>
</tr>
<tr>
<td>business</td>
<td>music</td>
<td>disease</td>
<td>court</td>
<td>tropical</td>
<td>army</td>
<td>university</td>
</tr>
<tr>
<td>services</td>
<td>band</td>
<td>medical</td>
<td>police</td>
<td>damage</td>
<td>force</td>
<td>college</td>
</tr>
<tr>
<td>market</td>
<td>song</td>
<td>patients</td>
<td>legal</td>
<td>winds</td>
<td>battle</td>
<td>high</td>
</tr>
<tr>
<td>companies</td>
<td>released</td>
<td>treatment</td>
<td>rights</td>
<td>typhoon</td>
<td>military</td>
<td>students</td>
</tr>
<tr>
<td>million</td>
<td>singer</td>
<td>cell</td>
<td>public</td>
<td>cyclone</td>
<td>air</td>
<td>schools</td>
</tr>
<tr>
<td>bank</td>
<td>rock</td>
<td>blood</td>
<td>justice</td>
<td>storms</td>
<td>navy</td>
<td>education</td>
</tr>
<tr>
<td>service</td>
<td>guitar</td>
<td>health</td>
<td>laws</td>
<td>caused</td>
<td>ship</td>
<td>institute</td>
</tr>
<tr>
<td>industry</td>
<td>live</td>
<td>medicine</td>
<td>judge</td>
<td>landfall</td>
<td>command</td>
<td>year</td>
</tr>
<tr>
<td>financial</td>
<td>records</td>
<td>brain</td>
<td>criminal</td>
<td>season</td>
<td>attack</td>
<td>program</td>
</tr>
<tr>
<td>tax</td>
<td>vocals</td>
<td>protein</td>
<td>supreme</td>
<td>pacific</td>
<td>fire</td>
<td>campus</td>
</tr>
</tbody>
</table>

Toolkit: GibbsLDA++; 1000 iterations; 2.8GHz computer; Heap size: 3G; took 14 hours
Hidden Topic Inference for Metadata

Record

“Introduction to futures and options market”

Record

“Innovations in investments and corporate finance”

Estimated Topic Model

Topic inference

**Topic 0:**
company business services market companies million bank corporation service management industry financial new production products development tax trade sold price ...

\[
Frequency_{m,k} = \begin{cases} 
\text{round}(\text{scale} \times \vartheta_{m,k}), & \text{if } \vartheta_{m,k} \geq \text{cut-off} \\
0, & \text{if } \vartheta_{m,k} < \text{cut-off}
\end{cases}
\]

probability of topic \( k \) in record \( m \)
Feature Representation

- Features: (co-author names, title, publishers)
- Feature Representation: Vector Space Model
- Record similarity: Cosine similarity

\[
\text{cosin}_{\text{sim}}(r_i, r_j) = \frac{r_i \cdot r_j}{|r_i| |r_j|} = \frac{\sum_{t \in V} w_{ti} \cdot w_{tj}}{\sqrt{\sum_{t \in V} w_{ti}^2} \cdot \sqrt{\sum_{t \in V} w_{tj}^2}}
\]

High Dimensional Space

Sparseness
Dimensionality Reduction with PCA

**Principle Component Analysis (PCA):** reduce each vector to few dimensions while keeping as much of the variance as possible.

- Less sparse
- More understandable model (visualization for better quantity analyses)
- Reduce speed & complexity of the Clustering process
Proposed Disambiguation Framework

Library Catalogue

Metadata Enriching Module

Ambiguous Candidates

PCA Module

Clustering Module

Ambiguous Candidates

A unique Author

A unique Author

A unique Author

A unique Author

A unique Author
• **Cluster books:** each cluster corresponds to a unique author

• Which clustering algorithm?
  – Number of clusters?
  – Distance metric?
- Clustering algorithm: Hierarchical Agglomerative Clustering (HAC)
- Distance between Clusters (A & B): Average Linkage

\[ d(A, B) = \frac{1}{|A| \cdot |B|} \sum_{x \in A} \sum_{y \in B} d(x, y) \]
## Experimental Settings

<table>
<thead>
<tr>
<th>Settings</th>
<th>Records</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>$c \cup t \cup p$</td>
</tr>
<tr>
<td>CNSH</td>
<td>$(c \cup t \cup p) \oplus (CN \cup SH)$</td>
</tr>
<tr>
<td>CNSH-Enriched</td>
<td>$(c \cup t \cup p) \oplus (CN \cup SH) \oplus SH$-enriched</td>
</tr>
<tr>
<td>CNSH-PCA</td>
<td>$[(c \cup t \cup p) \oplus (CN \cup SH)]_{PCA}$</td>
</tr>
<tr>
<td>HT</td>
<td>$(c \cup t \cup p) \oplus HT$</td>
</tr>
</tbody>
</table>

- $c$ = co-author names
- $t$ = book’s title
- $p$ = book’s publisher
- $CN$ = book’s Classification Numbers
- $SH$ = book’s Subject Headings
- $SH$-Enriched: Set of Enriched Subject Headings
- PCA: applying PCA to reduce dimensions
- HT: Set of most likely hidden topics inferred from the estimated topic model
**Goal:** exploit as much as possible the manual annotation information (i.e., $CN$, $SH$)

- Extract all English books (29,000 books): group $SH$ by $CN$
<table>
<thead>
<tr>
<th>Subject Headings</th>
<th>CC</th>
<th>DK</th>
<th>ET</th>
<th>ST</th>
<th>WF</th>
</tr>
</thead>
<tbody>
<tr>
<td>religion</td>
<td>observation</td>
<td>phonetics</td>
<td>metadata</td>
<td>microbiology</td>
<td></td>
</tr>
<tr>
<td>addresses</td>
<td>research</td>
<td>grammar</td>
<td>PHP</td>
<td>organisms</td>
<td></td>
</tr>
<tr>
<td>essays</td>
<td>project</td>
<td>lexicography</td>
<td>XSLT</td>
<td>soil</td>
<td></td>
</tr>
<tr>
<td>lectures</td>
<td>education</td>
<td>philosophy</td>
<td>computer</td>
<td>food industry</td>
<td></td>
</tr>
<tr>
<td>civilization</td>
<td>school</td>
<td>semantics</td>
<td>program</td>
<td>crops</td>
<td></td>
</tr>
<tr>
<td>philosophy</td>
<td>program</td>
<td>cognition</td>
<td>language</td>
<td>nitrogen</td>
<td></td>
</tr>
<tr>
<td>ethics</td>
<td>learning</td>
<td>phonology</td>
<td>software</td>
<td>microbial</td>
<td></td>
</tr>
<tr>
<td>cognition</td>
<td>daycare</td>
<td>typology</td>
<td>Microsoft</td>
<td>innovations</td>
<td></td>
</tr>
<tr>
<td>evolution</td>
<td>child</td>
<td>linguistics</td>
<td>design</td>
<td>government</td>
<td>policy</td>
</tr>
<tr>
<td>development</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Each record is enriched with the **first 20** SHs in the corresponding **CN**
Authors sharing the first initial name & last name (e.g., R Smith) [Torvik 05, 09; Huang 06]

- Data for testing from Bolziano Library catalogue
  - 28 groups of authors (each group contains 2 – 30 distinguished authors)
  - books in English, having full information

→ Preprocessing: normalization, segmentation, stop-word removal
Evaluation Metrics

\[ p_{Pre} = \frac{\text{number of correct pairs in the output clusters}}{\text{number of total pairs in the output clusters}} \]

\[ p_{Re} = \frac{\text{number of correct pairs in the output clusters}}{\text{number of total pairs in the truth clusters}} \]

\[ pF1 = 2 \cdot \frac{p_{Pre} \cdot p_{Re}}{p_{Pre} + p_{Re}} \]
Results

Baseline

CNSH

CNSH-Enriched

CNSH-PCA
Hidden Topic Enriching performance
Conclusions

• Proposed a framework for author name disambiguation:
  – exploit manual annotations \((CN, SH)\)
  – use Hidden Topics estimated from Wikipedia to automatically enrich record’s information
    • can be applied to federated libraries, digital libraries like CiteSeer, DBLP, PubMed
    • take advantage of available large-scale knowledge-base dataset, Wikipedia
    • can be used for different languages
  – use PCA to represent data in a more compact way
    • reduce number of dimensions \(\rightarrow\) reduce speed & complexity, reduce noises
    • achieve satisfactory results
    • can be used for visualization for better quantity clustering analyses in the future
Future works

- Contribution of different features
- Experiment in a multilingual environment
- Optimize cutting points for HAC
Acknowledgement

Raffaella Bernardi
Massimo Poesio
Patrick Blackburn
Luigi Siciliano
CACAO project
Marco Baroni
Cristiano Cumer
Manuel Kirschner
European Commission & LCT program