Linked Enterprise Data for Fine Grained Named Entity Linking and Web Intelligence

Albert Weichselbraun
Agenda

1. Introduction
2. Datasets: Linked Enterprise Data
3. Method
   - Major challenges
   - Pre-processing
   - Disambiguation
4. Evaluation
5. Outlook and conclusions
Introduction

• Named Entity Linking

• Key issues
  • extraction of company names and useful context and structural information from linked data sources
  • mention generation and assessment
  • disambiguation – locate mentions in text documents and ground them to the corresponding entities in the linked data source
What If China Does Land Hard? will likely affect everyone and every market," UBS said in its "How Might a China Hard Landing... blogs.wsj.com

Stocks End Higher Despite Mixed Data said David Lefkowitz, equity strategist with UBS Wealth Management. "These are not the sect... online.wsj.com

Fall in Centrica pegs back UK's FTSE 100 FTSE 100 down 0.2 pct. * Centrica hit by UBS downgrade. * FTSE down 0.4 pct start of 2... reuters.com

Manulife subsidiary among institutional investors buying U.S. farmland: Report subsidiary of Manulife Financial Corp.; and UBS Agrinvest, also known as UBS Global Real Estat... bnn.ca

Fed Chief's Spouse Advises Center Funded by UBS on the board of an academic center funded by UBS AG, the giant Swiss bank, which is posed t... online.wsj.com

U.S. rules show limits of global bank resolution Barclays, Credit Suisse, Deutsche Bank and UBS: Deutsche could face as much as a 1 billion eu... reuters.com

UPDATE 2-Detroit bankruptcy judge urges settlement in bond dispute settlement proposals with swap counterparties UBS AG and Merrill Lynch Capital Services, cal... reuters.com

UK fraud agency charges three ex-Barclays bankers over Libor against three former employees of Swiss bank UBS and UK brokerage RP Martin, the first peo... reuters.com

Italy yields hit 8-year lows as Renzi reforms plan cheers investors reforms," said Justin Knight, a strategist at UBS. "We are bullish on the periphery and we expe... reuters.com
Linked Enterprise Data

• Orell Füssli Business Information AG → Switzerland's largest provider of business inf.

• Linked Enterprise Repository
  • based on a number of business databases
  • comprises 2.9 million companies and background information (names, key people, products, contact information, brand names, turnover, ...)
  • removal of duplicates and inactive companies
  • conversion to linked data using well known name spaces → 570,000 organizations; 9 million triples
# linked data database

teledata-company:775 rdf:type owl:Company.
teledata-company:775 rdfs:label "American Optical Company International AG".
teledata-company:775 rdfs:label "Carl Zeiss Vision AG".
teledata-company:775 dbpedia-owl:numberOfEmployees "35".
teledata-company:775 dbprop-de:umsatz "4183400.0".
teledata-company:775 ofwi:company-status "active".
teledata-company:775 dbpedia-owl:industry ofwi-industry:8962, ofwi-industry:7752.
teledata-company:037041 schema-org:address ofwi-address:037041.

# kompass database

ofwi-company:037041 owl:sameAs teledata-company:775.
# contact and legal information
ofwi-company:037041 dbprop-de:unternehmensform  
dbpedia-de:Aktiengesellschaft.

ofwi-company:037041 schema-org:email "office@zeis.com".

ofwi-company:037041 schema-org:faxNumber "05525473730".

ofwi-company:037041 schema-org:telephone "0552547373".

ofwi-company:037041 schema-org:email "info.swiss@vision.zeiss.com".


# keywords regarding the company's products and services
ofwi-company:037041 dbprop:products ofwi-productgroup:38371, ...

ofwi-company:037041 dbprop:distributor "Teflon easycare",
    "i.Profiler", "Carl Zeiss".

# key people
ofwi-company:037041 dbprop:keyPeople ofwi-person:Peter_Däpp_(0432);
    dbprop:keyPeople ofwi-person:Sven_Hermann_(0341).
Linked Enterprise Data | Example

# address information
ofwi-address:037041 rdf:type schema-org:PostalAddress
ofwi-address:037041 schema-org:addressCountry "CH".
ofwi-address:037041 schema-org:addressRegion "ZH".
ofwi-address:037041 schema-org:postalCode "8714".
ofwi-address:037041 schema-org:addressLocality "Feldbach".
ofwi-address:037041 schema-org:streetAddress "Feldbacherstrasse 81".

# product groups
ofwi-productgroup:38371 rdfs:label "Optische Linsen", "Gläser", "Spiegel".
ofwi-productgroup:3837122 rdfs:label "Brillenlärser".

# industry mapping
ofwi-industry:8962 rdf:label "Wholesale of photographic and .."@en;
rdf:label "Commercia all'ingrosso di..."@it;
rdf:label "Commerce de gros d'appareils ...."@fr;
rdf:label "Grosshandel mit Foto ..."@de.
# Challenges

<table>
<thead>
<tr>
<th>ID</th>
<th>Description</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Data quality</td>
<td></td>
</tr>
<tr>
<td>1.1</td>
<td>ambiguous short names</td>
<td>Aktien (shares), Hell (bright), ...</td>
</tr>
<tr>
<td>1.2</td>
<td>uppercase only company names</td>
<td>DER SA, DER HEIZER</td>
</tr>
<tr>
<td>2</td>
<td>ambiguities</td>
<td></td>
</tr>
<tr>
<td>2.1</td>
<td>many very small companies</td>
<td>1300 x Meyer, 1018 x Personalfürsorgestiftung</td>
</tr>
<tr>
<td>2.2</td>
<td>legally related companies</td>
<td>83 x Credit Suisse, 92 x UBS</td>
</tr>
<tr>
<td>2.3</td>
<td>Similar company names with no or little metadata</td>
<td>ABSOLUT, ABSOLUT SA, ABSOLUT COSMETICS</td>
</tr>
</tbody>
</table>
## Challenges

<table>
<thead>
<tr>
<th>ID</th>
<th>Description</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>Smaller data granularity</td>
<td></td>
</tr>
<tr>
<td>3.1</td>
<td>Ambiguous company names</td>
<td>13 x IST (be); 0 x @DBpedia WEG (way)</td>
</tr>
<tr>
<td>3.2</td>
<td>Ambiguous person names</td>
<td>Robert Frey vs. Rober Frey Consulting</td>
</tr>
<tr>
<td>4</td>
<td>Use of casual name forms</td>
<td></td>
</tr>
<tr>
<td>4.1</td>
<td>Short names</td>
<td>IST AG rather than Innovative Sensort Technology AG</td>
</tr>
<tr>
<td>4.2</td>
<td>Use of „insider“ casual names</td>
<td>Sonova (Phonak Sounds AG) CS (Credit Suisse)</td>
</tr>
</tbody>
</table>
Method | System Architecture

- **Method**
  - **System Architecture**

- **Components**
  - **SPARQL Query & Object Mapping**
  - **GeoNames**
  - **Linked Enterprise Data Repository**
  - **Stopword Lists**
  - **EntityLyzard Configuration**
  - **Dictionaries**
  - **Name and Business Objectives Database**
  - **Name and Business Objectives Database**
  - **Servlets**
  - **Context Analyzer**
  - **Structure Analyzer**
  - **Name Analyzer**
  - **Disambiguation**
    - **Person Disambiguation Profile**
    - **Location Disambiguation Profile**
    - **Company Disambiguation Profile**

- **Keywords**
  - **Named entity linking and disambiguation Profile**
  - **Raw names**
  - **Ambiguous and unambiguous search needles**
  - **Context terms**

- **Algorithm**
  - **Lucene-based Disambiguation Algorithm**
  - **Amitay-based Disambiguation Algorithm**

- **Process**
  - **Document**
  - **Recognize**
  - **Named Entity Annotations**
Method | Name Analyzer

- Extract potential mentions from linked data
  - Entity: UBS Financial Service Basel AG
  - Possible mention Strings: UBS, UBS Financial Service, ...

- Entropy-based metric
  \[ H(m) = H_{\text{init}}(m) + H_C(C_{ij}) + \sum_{t_k \in T_{ij}} H_{t,t_{caseSensitive}}(m) + H_{\text{compl}} \]

- Entropy threshold determines minimum mention length and ensures “complete” names

- Entries below the threshold are disambiguated using prefixes and/or suffixes
Method | Context & structure analyzer

- structural information
  - related companies and subsidiaries
  - company's management
  - address information

- context information
  - products and services offered by the company
  - industry
  - revenue and number of employees
Method | Recognize Profiles

<table>
<thead>
<tr>
<th>field</th>
<th>value</th>
</tr>
</thead>
<tbody>
<tr>
<td>label</td>
<td>eval.5.context</td>
</tr>
<tr>
<td>source</td>
<td>http://../r/de.dbpedia.org</td>
</tr>
<tr>
<td></td>
<td>WHERE {?s rdf:type dbpedia-owl:Company .}</td>
</tr>
<tr>
<td></td>
<td>{?s rdfs:label ?companyName .} OPTIONAL {</td>
</tr>
<tr>
<td></td>
<td>{?s prop-de:sitz ?city .} } FILTER (LANG(?companyName) = 'de') ...</td>
</tr>
</tbody>
</table>

...
# Method | Recognize Profiles

<table>
<thead>
<tr>
<th>field</th>
<th>value</th>
</tr>
</thead>
<tbody>
<tr>
<td>entity type</td>
<td>Recognize.OrganizationEntity</td>
</tr>
<tr>
<td>disambiguation algorithm</td>
<td>Lucene similarity</td>
</tr>
</tbody>
</table>
| pre-processor               | binding: ?companyName
  handler: OrganizationNameHandler                                   |
| filter                      | scope: name
  dict=dict.C, dict.de_CH, dict_de, dict.en                           |
| filter                      | scope: context
  dict=stopwords.C, stopwords.de, stopwords.en, ...                   |
| affix filter (disambiguation)| Recognize.OrganizationAffix                                        |
Method | Disambiguation

- Geo → Amitay
- Organizations → modified Lucene similarity
- ambiguous mentions are disambiguated using prefix and suffix terms
  \[
  s(e_i, d) = f_c(mentions_e, d) \cdot |mentions_e| \sum_{t \in mentions_e} \left[ idf(t^2) \cdot boost(t) \right]
  \]
- ranking is refined by using weights obtained from context information (number of employees and turnover)
Evaluation | Corpora

- extended AWP.ch news dataset
  - 320,000 manually annotated news messages
  - 150 randomly selected German-speaking news messages
  - annotations of all covered companies which have been manually confirmed by domain experts

- NZZ (Neue Zürcher Zeitung) news dataset
  - 150 randomly selected NZZ business news articles manually annotated by domain experts
Evaluation | Setting

- raw names
  → extracted names “as is”
- simple
  → tokenize names and generate standardized alternative names (e.g. I.B.M. > IBM, ...)
- advanced
  → full Recognyze name pre-processing
# Evaluation | Estimated Coverage

<table>
<thead>
<tr>
<th>Setting</th>
<th>Rescore</th>
<th>AWP messages R</th>
<th>NZZ articles R</th>
</tr>
</thead>
<tbody>
<tr>
<td>raw names</td>
<td></td>
<td>0.52</td>
<td>0.13</td>
</tr>
<tr>
<td></td>
<td>√</td>
<td>0.52</td>
<td>0.13</td>
</tr>
<tr>
<td>simple</td>
<td></td>
<td>0.95</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td>√</td>
<td>0.81</td>
<td>0.66</td>
</tr>
<tr>
<td>advanced</td>
<td></td>
<td>0.87</td>
<td>0.81</td>
</tr>
<tr>
<td></td>
<td>√</td>
<td>0.83</td>
<td>0.76</td>
</tr>
</tbody>
</table>
## Evaluation | Linking Performance

<table>
<thead>
<tr>
<th>Setting</th>
<th>Rescore</th>
<th>AWP messages P</th>
<th>R</th>
<th>F1</th>
<th>NZZ articles P</th>
<th>R</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>raw names</td>
<td></td>
<td>0.44 0.52 0.44</td>
<td></td>
<td></td>
<td>0.14 0.13 0.11</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>√</td>
<td>0.49 0.52 0.47</td>
<td></td>
<td></td>
<td>0.16 0.13 0.13</td>
<td></td>
<td></td>
</tr>
<tr>
<td>simple</td>
<td></td>
<td>0.07 0.52 0.10</td>
<td></td>
<td></td>
<td>0.03 0.45 0.06</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>√</td>
<td>0.09 0.61 0.14</td>
<td></td>
<td></td>
<td>0.04 0.55 0.07</td>
<td></td>
<td></td>
</tr>
<tr>
<td>advanced</td>
<td></td>
<td>0.36 0.71 0.41</td>
<td></td>
<td></td>
<td>0.38 0.75 0.46</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>√</td>
<td>0.50 0.80 0.59</td>
<td></td>
<td></td>
<td>0.60 0.74 0.63</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Outlook and conclusions

- Recognyze draws upon linked data sources → no learning step involved → tested with OFWI linked enterprise data and Dbpedia

- Data pre-processing considerably improves the component's performance

- Future work will focus on
  - adding support for additional named entity types (people and events)
  - improved extraction of contextual information (e.g. obtain abbreviations from Dbpedia abstracts)
  - create easy ways to create and share Recognyze profiles