Linking Multilingual eGov Services to the LOD Cloud

Fedelucio Narducci*, Matteo Palmonari*, Giovanni Semeraro
*DISCO, University of Milan-Bicocca, Italy
°Department of Computer Science, University of Bari Aldo Moro, Italy

AI for Smart Cities Workshop
AI*IA 2013 - 25th Year Anniversary
XIII Conference
Turin (Italy), December 5, 2013
EC 6 axes
to evaluate ‘city smartness’

http://www.smart-cities.eu/model.html
Cross-language Linking of Open Government Data

• A large amount of Open Government Data in many languages*:
  o 1,000,000+ datasets published online (February 2013)
  o 40 different countries
  o 24 different languages
Cross-language Linking of eGov Service Descriptions

- Government service catalogs are part of the LOD cloud
  - Effective Service Delivery (ESD)-toolkit
  - European Local Government Service List (LGSL)
    - 2000+ interlinked public services in 6 languages
eGov LOD Services  
ESD & LGSL

• Effective Service Delivery (ESD)-toolkit
  o defines the semantics of public sector services
  o The Smart Cities Project (2009-2011)
    • innovation network in the domain of the development and uptake of e-services in the whole North Sea region
    • England, Netherlands, Belgium, Germany, Scotland, Sweden, and Norway

• European Local Government Service List (LGSL)
  o each country responsible to build and maintain its list of public services
  o all services interlinked to the services delivered by other countries
  o linked to the LOD cloud
Cross-language Linking of eGov Services

Why is it useful?

• Advantages for PAs
  o Compare local service offerings with best practices in other countries
  o Support interoperability among PAs of different countries and other service providers
  o Enrich service descriptions with additional information via links to LGSL (e.g., link to life event ontologies)

• Advantages for citizens
  o Find eGov services when in a foreign country
  o Towards cross-language service access

Costly and Error Prone Activity
Catalogs of several hundreds of services
Cross-language Linking of eGov Services

Why is it challenging?

≈ sameAs links

Semantic heterogeneity
  o not a mere “translation” problem
  o cultural bias

Ultra-short descriptions

• Challenging cross-language matching problem
• Most of the approaches:
  • or report problems when automatic translation returns descriptions with heterogeneous vocabulary [Hertling & Paulheim 2012]
CroSeR
Cross-language Service Retriever

CROSER: Cross-language Service Retriever
connects your service catalogs

Web tool to support the linkage of a source eGov service catalog represented in any language to a target catalog represented in English

Based on Machine Translation and Explicit Semantic Analysis (ESA)
Web tool to **support the linkage** of a source eGov service catalog represented in **any language** to a target catalog represented in **English**.

Based on **Machine Translation** and **Explicit Semantic Analysis (ESA)**.
CroSeR
Cross-language Service Retriever

CROSER: Cross-language Service Retriever
connects your service catalogs

Italian  Dutch  Belgian  German  Swedish  Norwegian

• Load a catalog

Service Catalog

Service Matched from ESD
Search

Service Info

Web tool to support the linkage of a source eGov service catalog represented in any language to a target catalog represented in English

Based on Machine Translation and Explicit Semantic Analysis (ESA)
Web tool to support the linkage of a source eGov service catalog represented in any language to a target catalog represented in English.

Based on Machine Translation and Explicit Semantic Analysis (ESA).
CroSeR
Cross-language Service Retriever

CROSER: Cross-language Service Retriever
connects your service catalogs

- Load a catalog

- Select a source service

- Look at the retrieved services (link recommendations)

Web tool to support the linkage of a source eGov service catalog represented in any language to a target catalog represented in English.

Based on Machine Translation and Explicit Semantic Analysis (ESA)
CroSeR
Cross-language Service Retriever

CROSER: Cross-language Service Retriever
connects your service catalogs

- Load a catalog
- Select a source service
- Look at the retrieved services (link recommendations)
- Link SKOS broader / exact / narrower match

Web tool to **support the linkage** of a source eGov service catalog represented in **any language** to a target catalog represented in **English**

Based on **Machine Translation** and **Explicit Semantic Analysis (ESA)**
Web tool to **support the linkage** of a source eGov service catalog represented in **any language** to a target catalog represented in **English**.

Based on **Machine Translation** and **Explicit Semantic Analysis (ESA)**.
CroSeR: Matching Approach

Machine Translation of Service Descriptions

Extraction of ESA-based representations and indexing (Vector Space Model)

Top-k Service Retrieval by Cosine Similarity
Explicit Semantic Analysis (ESA)

Technique able to provide a fine-grained semantic representation of natural language texts in a high-dimensional space of comprehensible concepts derived from Wikipedia [GM06]

Wikipedia viewed as an ontology = a collection of ~1M concepts

Explicit Semantic Analysis (ESA)

- Sparse matrix where rows are terms in the whole vocabulary and columns are the Wikipedia concepts (articles)
- TF-IDF score for representing the semantic relatedness between a term and a Wikipedia concept

<table>
<thead>
<tr>
<th>ESA</th>
<th>Concept 1</th>
<th>...</th>
<th>Concept n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Term 1</td>
<td>TF-IDF</td>
<td></td>
<td>TF-IDF</td>
</tr>
<tr>
<td>...</td>
<td>TF-IDF</td>
<td></td>
<td>TF-IDF</td>
</tr>
<tr>
<td>Term k</td>
<td>TF-IDF</td>
<td></td>
<td>TF-IDF</td>
</tr>
</tbody>
</table>
Every Wikipedia article represents a **concept**.

Article **words** are associated with the **concept** (TF-IDF).
Every Wikipedia article represents a concept

Panthera

Article words are associated with the concept (TF-IDF)
Every Wikipedia article represents a **concept**
Article words are associated with the **concept** (TF-IDF)

The **semantics** of a word is the **vector** of its **associations** with Wikipedia concepts
The semantics of a text fragment is the average vector (centroid) of the semantics of its words.

Explicit Semantic Analysis (ESA)

In practice – Disambiguation
ESA effectively used for

- **Text Categorization** [Gabri09]
  - Experiments on diverse collection of datasets
  - Significant improvement using Support Vector Machines
- **Semantic relatedness** of words and texts [Gabri09]
  - Cosine similarity between vectors of ESA concepts
  - Significant improvement for both the tasks
- **Information Retrieval** [Egozi08, Egozi11]
  - Morag: an ESA-based IR algorithm enriching documents and queries
  - The generated features improve the performance of BOW-based systems
- **Information Filtering** [NMSLdG13]
  - ESA-based enrichment of user profiles in content-based RecSys
  - Improved transparency and serendipity: more comprehensible concepts to represent user profiles and justify recommendations

---


Explicit Semantic Analysis (ESA) in CroSeR

Wikipedia-based representation of natural language expressions with the ESA matrix

A set of terms is represented by the centroid of the vectors associated with the individual terms

- E.g.: “Unemployment Support” → Job Interview (0.42), Employment Agency (0.55), …, Unemployment benefits (0.62)

- Feature generation/semantic enrichment + lightweight disambiguation
Experimental Evaluation: Design

• Dataset
  - Any language LGSL vs. English LGSL
    - Dutch (#225)
    - German (#190)
    - Flemish (#341)
    - Norwegian (#165)
    - Swedish (#66)
    - TOT = #997 vs. #1425

• Methodology
  - Gold standard: ≈sameAs links defined by human experts in the LGSL dataset
  - Accuracy @1 ... @30
  - MRR (Mean Reciprocal Rank)

• Comparative evaluation against baseline and other techniques based on similar principles*
  - ESA
  - ESA + keyword vs
  - Keyword (baseline)
  - Tagme
  - Tagme+Keyword
  - WikiMiner
  - WikiMiner+Keyword
  - DBpedia Spotlight
  - DBpedia Spotlight+Keyword

*Experiments with CL-ESA [Sorg et al. 2012] have also been carried out
Experiment: Design

- **Metrics**
  - **Accuracy@n**: is calculated considering only the first $n$ retrieved services. If the correct service occurs in the top-$n$ items, the service is marked as correctly retrieved ($n = 1; 3; 5; 10; 20; 30$)
  
  - **MRR**
    
    $$MRR = \frac{1}{N} \sum_{i=1}^{N} rank_i$$
    
    $rank_i$ is the rank of the correctly retrieved service in the ranked list, and $N$ is the number of services correctly retrieved **with the** configuration.
Experiments: Accuracy
Norwegian catalog (best average accuracy)

- Best approach in terms of Accuracy@n (n ≥ 5) is ESA
- ESA becomes more effective when more services are returned
- Merging the Wikipedia-based representation with keywords does not improve accuracy in ESA
Experiments: Accuracy
Swedish catalog (worst average accuracy)

- Low accuracy for every approach
- Best approach in terms of Accuracy@n is ESA
- ESA relative performance is more evident
  - ESA more effective for any \( n \)
Wilcoxon test for Keyword vs. ESA

<table>
<thead>
<tr>
<th>Catalog</th>
<th>a@1</th>
<th>a@3</th>
<th>a@5</th>
<th>a@10</th>
<th>a@20</th>
<th>a@30</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dutch</td>
<td>0.05</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>Belgian</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>German</td>
<td></td>
<td></td>
<td></td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>Norwegian</td>
<td></td>
<td></td>
<td></td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>Swedish</td>
<td></td>
<td></td>
<td></td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
</tbody>
</table>
Experiments:
Mean Reciprocal Rank

<table>
<thead>
<tr>
<th>Representation</th>
<th>Dutch</th>
<th>Belgian</th>
<th>German</th>
<th>Norwegian</th>
<th>Swedish</th>
</tr>
</thead>
<tbody>
<tr>
<td>keyword</td>
<td>0.333</td>
<td>0.320</td>
<td>0.242</td>
<td>0.273</td>
<td>0.182</td>
</tr>
<tr>
<td>tagme</td>
<td>0.120</td>
<td>0.094</td>
<td>0.147</td>
<td>0.121</td>
<td>0.091</td>
</tr>
<tr>
<td>tagme + keyword</td>
<td>0.316</td>
<td>0.334</td>
<td>0.258</td>
<td>0.273</td>
<td>0.197</td>
</tr>
<tr>
<td>wikifi</td>
<td>0.080</td>
<td>0.114</td>
<td>0.116</td>
<td>0.109</td>
<td>0.106</td>
</tr>
<tr>
<td>wikifi + keyword</td>
<td>0.324</td>
<td>0.326</td>
<td>0.258</td>
<td>0.291</td>
<td>0.167</td>
</tr>
<tr>
<td>esa</td>
<td>0.311</td>
<td>0.326</td>
<td>0.289</td>
<td>0.261</td>
<td>0.242</td>
</tr>
<tr>
<td>esa + keyword</td>
<td>0.311</td>
<td>0.328</td>
<td>0.289</td>
<td>0.261</td>
<td>0.242</td>
</tr>
<tr>
<td>dbpedia</td>
<td>0.182</td>
<td>0.202</td>
<td>0.163</td>
<td>0.164</td>
<td>0.091</td>
</tr>
<tr>
<td>dbpedia + keyword</td>
<td>0.329</td>
<td>0.334</td>
<td>0.274</td>
<td>0.285</td>
<td>0.212</td>
</tr>
</tbody>
</table>

- ESA average rank: between 3rd and 5th position
- Suboptimal MRR in two datasets: higher coverage is achieved under the condition that the list of retrieved services is extended
Experiments: Mean Reciprocal Rank

<table>
<thead>
<tr>
<th>Representation</th>
<th>Dutch</th>
<th>Belgian</th>
<th>German</th>
<th>Norwegian</th>
<th>Swedish</th>
</tr>
</thead>
<tbody>
<tr>
<td>keyword</td>
<td>0.333</td>
<td>0.320</td>
<td>0.242</td>
<td>0.273</td>
<td>0.182</td>
</tr>
<tr>
<td>tagme</td>
<td>0.120</td>
<td>0.094</td>
<td>0.147</td>
<td>0.121</td>
<td>0.091</td>
</tr>
<tr>
<td>wikihi</td>
<td>0.080</td>
<td>0.114</td>
<td>0.116</td>
<td>0.109</td>
<td>0.106</td>
</tr>
<tr>
<td>dbpedia</td>
<td>0.182</td>
<td>0.202</td>
<td>0.163</td>
<td>0.164</td>
<td>0.091</td>
</tr>
<tr>
<td>dbpedia+keyword</td>
<td>0.329</td>
<td>0.334</td>
<td>0.274</td>
<td>0.285</td>
<td>0.212</td>
</tr>
</tbody>
</table>

- ESA average rank: between 3rd and 5th position
- Suboptimal MRR in two datasets: higher coverage is achieved under the condition that the list of retrieved services is extended

Difficult matching task: user intervention is needed

Looking at a reasonable number of recommendations significantly reduces the linking effort (e.g., 30/1425)
Experiments: Discussion

- CroSeR finds matchings that cannot be discovered by machine translation + keyword comparison

- CroSeR’s recommendations can support the users to refine the links
Conclusions & Future Work

• Summary
  o CroSeR uses cross-language matching to recommend links among eGov service descriptions available in different languages
  o Good performance in semi-automatic linking settings
  o Unsupervised method based on Explicit Semantic Analysis
  o Language independent

• Future work
  o Collect additional information to improve the overall results
  o User study to further evaluate the quality of the recommendations (ongoing)
  o Application of the same approach to other similar cross-language matching problems (where similar means: little information available, only short descriptions, focus on coverage rather than precision)
Thanks, questions?


Extended versions: