Cross-language Semantic Retrieval and Linking of eGov Services

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

*http://logd.tw.rpi.edu/iogds data analytics
Cross-language Linking of eGov Service Descriptions

- Government service catalogs are part of the LOD cloud
  - Electronic Service Delivery (ESD)-toolkit
  - European Local Government Service List (LGSL)
    - 2000+ interlinked public services in 6 languages
    - Each country maintains its list of public services
Cross-language Linking of eGov Services

Why it is 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 it is Challenging

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

connects your service catalogs

<table>
<thead>
<tr>
<th>Language</th>
</tr>
</thead>
<tbody>
<tr>
<td>Italian</td>
</tr>
<tr>
<td>Dutch</td>
</tr>
<tr>
<td>Belgian</td>
</tr>
<tr>
<td>German</td>
</tr>
<tr>
<td>Swedish</td>
</tr>
<tr>
<td>Norwegian</td>
</tr>
</tbody>
</table>

**Service Catalog**

**Service Matched from ESD**

**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)**
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).
CroSeR
Cross-language Service Retriever

CROSER: Cross-language Service Retriever
connects your service catalogs

- 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)
CroSeR

Cross-language Service Retriever

CROSER: Cross-language Service Retriever
connects your service catalogs

- Load a catalog

Service Catalog

- Select a source service

Service Matched from ESD

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)
CroSeR
Cross-language Service Retriever

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

Service Catalog
- Select a source service

Service Matched from ESD
- Look at the retrieved services
  (link recommendations)

Service Info
- 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)
CroSeR
Cross-language Service Retriever

CROSER: Cross-language Service Retriever
connects your service catalogs

- Load a catalog
- Scan the list or search for a specific service
- Look at target service description (if available)
- Look at the 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)
CroSeR
Cross-language Service Retriever

CROSER: Cross-language Service Retriever
connects your service catalogs

<table>
<thead>
<tr>
<th>Italian</th>
<th>Dutch</th>
<th>Belgian</th>
<th>German</th>
<th>Swedish</th>
<th>Norwegian</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Service Catalog

Service Matched from ESD

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)
CroSeR: Matching Approcah

Machine Translation of Service Descriptions

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

Top-k Service Retrieval by Cosine Similarity
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 + light-weight disambiguation
Experimental Evaluation: Design

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

• Methodology
  o **Gold standard:** `sameAs` links defined by human experts in the LGSL
  o **Accuracy** @1 … @30
  o **MRR** (Mean Reciprocal Rank)

• Comparative evaluation against baseline and other techniques based on similar principles*
  o **Esa**
  o **Esa + keyword** vs
  o Keyword (baseline)
  o Tagme
  o Tagme + keyword
  o Wikiminer
  o Wikiminer + keyword
  o Dbpedia Spotlight
  o Dbpedia Sp + keyword

*Experiments with CL-ESA [Sorg et al. 2’12] have also been carried out
Experiments: Accuracy
Dataset in which Best Average Accuracy

- Best approach in terms of Accuracy@n 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
Dataset in which 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 $n$
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>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>0.336</td>
<td>0.258</td>
<td>0.291</td>
<td>0.167</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 reduce 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

"Absentee Ballot"
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 similar approach to other cross-language matching problems
Thanks, questions?

Demo at http://siti-rack.siti.disco.unimib.it:8080/croser/
eGov Services LOD
ESD & LGSL

- Electronic Service Delivery (ESD)-toolkit
  - Define the semantics of public sector services
  - The SmartCities project
    - 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)
  - Each country responsible to build and maintain its list of public services
  - All of those services are interlinked to the services delivered by other countries
  - Linked to the LOD cloud
CroSeR
Cross-language Service Retriever

CROSER: Cross-language Service Retriever
connects your service catalogs

Load a catalog

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)
• The semantic relatedness between a term and a Wikipedia concept (article) is in terms of **TF-IDF** score.

• Several heuristics are applied in order to reduce the number of concepts and terms.
ESA: Wikipedia Concepts and Terms

Every Wikipedia article represents a concept.

Article words (terms) are associated with the concept (TF-IDF).

Panthera

Cat [0.92]
Leopard [0.84]
Roar [0.77]
The **semantic interpretation vector** of a text fragment is the **centroid** vector of the terms occurring in

**Centroid vector**

- **mouse**
  - Mouse 
  - Rodent [0.91]
  - Computing [0.89]
  - Mickey Mouse [0.81]
  - John Steinbeck [0.17]

- **button**
  - Button [0.93]
  - Dick Button [0.84]
  - Computing Mouse [0.81]
  - Game Controller [0.32]

- **mouse**
  - Computing Mouse [0.85]
  - Rodent [0.46]
  - IBM PS/2 [0.35]
  - Drag-and-drop [0.32]
• **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**
    
    \[
    MRR = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{\text{rank}_i}
    \]

    \(\text{rank}_i\) is the rank of the correctly retrieved servicei in the ranked list, and N is the number of the services correctly retrieved **with the configuration**.
• Motivation
• CroSeR
• Experimental Evaluation
• Conclusions
We can certainly state that esa is the most effective representation in terms of accuracy. In order to statistically validate our experiment we compared results obtained by the keyword-based representation with results obtained by the esa-based one. Table 1 reports levels of significance (p-value) obtained by performing the Wilcoxon Matched Pairs Test. More specifically, the number reported in each cell shows the statistic significance (p-value) of the differences between keywords and esa for each n value of a@n. Empty cells show statistically significant difference. We can observe that for the Belgian catalog (that is also the richest one) the improvement of esa is statistically significant for each n value. Conversely, other catalogs show statistically significant differences from n = 10 onwards. These results can be considered actually satisfying, since starting from 10 retrieved items CroSeR becomes significantly better than a keyword-based model. The second analysis focuses the attention on the capability of CroSeR to boost relevant services in the first positions of the retrieved list. Results in terms of mrr for each representation are reported in Table 2. For Belgian, German, and Swedish catalogs the representation based on esa shows the highest values, but differences with other representations (keyword and wikifi+keyword) are really slight. Accordingly, there is not a representation that decisively outperform the baseline for this metric. However, since CroSeR is a retriever system, and not, for example, a question-answering engine (for which to have the correct answer in the first position plays a crucial role), we can consider these results good. Indeed, the average rank of the correct service for the esa representation is between the

<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></td>
<td></td>
<td></td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>Belgian</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>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>