Cross-language Retrieval of E-gov Services

Fedelucio Narducci, Matteo Palmonari, Giovanni Semeraro
• Motivations and Problem Definition
• Cross-Language Semantic Matching by CroSeR
• Semantic Annotation Techniques
• Experiments
• Conclusions
Linking Open Government Data

• Open Government Data (OGD)
  o Many governments decided to make public their data about spending, service provision, economic indicators, and so on

• A large amount of data in many languages*:
  o 1,000,000+ datasets published online (October 2015)
  o 43 different countries
  o 24 different languages

• Standardization and Sharing: Linked Open Government Data
  o Roadmap for Linked Open Government Data [Ding et al.2012]
  o Open stage / link stage / reuse stage

*http://logd.tw.rpi.edu/iogds_data_analytics
eGov Services Data ESD & LGLS

- 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, Sweden, and Norway

- European Local Government Service List (LGSL) as part of the Electronic Service Delivery (ESD)-toolkit website
  - Defines the semantics of public sector services
  - 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
Cross-language Matching of eGov Services

- ≈ sameAs links
- semantic heterogeneities
  - not a mere “translation” problem (cultural bias)
- ultra-short descriptions

1. Primary school places
2. Education - grants - student awards
3. Nursery education grant
4. Adult residential care

---

1. Basisonderwijs
   - *Basic Education*
2. Inschrijvingsgeld – vrijstelling
   - *Registration fee-exemption*
3. Kinderopvangtoeslag
   - *Childcare Allowance*
4. Rust- en verzorgingstehuizen
   - *Retirement and nursing homes*

---

1. Leerplicht
   - *Compulsory Schooling*
2. Studiefonds Provincie groningen
   - *Study Fund province of groningen*
3. Subsidie Kinderopvang
   - *Grant Babysitting/Child Services*
4. Opname Verzorgingshuis
   - *Recording Care Home*
**CroSeR**

- Cross-language Service Retrieval (CroSeR)
  - a tool to support the linkage of a source eGov service catalog represented in any language to a target catalog represented in English
  - based on Cross-language Semantic Matching

- Given a source service \( s \) described in a language \( L \), CroSeR retrieves \( k \) target services described in the LGSL (English), which are best candidate to be linked to \( s \)
  - Based on automatic extraction of Wikipedia-based service annotations
  - Unsupervised
  - Minimal descriptions (average number of terms \( \approx 3 \))
  - Language independent
CroSeR: Architecture
Wikipedia-based Annotation Techniques

• Feature Generation
  o Translation-based Explicit Semantic Analysis (TR-ESA)
  o E.g., Home Schooling $\rightarrow$ Home (0.67), School (0.55), Education (0.48), Family (0.35)

• Wikipedia-based Tagging
  o Tag.me
  o Wikifi
  o Dbpedia Spotlight
  o E.g., Home Schooling $\rightarrow$ Homeschooling (1)
Experiment: Design

• Datasets
  o LGSL 1435 services
  o 225 Dutch services
  o 190 German services
  o 341 Belgian services
  o 165 Norwegian services
  o 66 Swedish services

• Methodology
  o Gold standard represented by links defined by human experts

• Configurations
  o Keyword (baseline)
  o Tagme
  o Tagme + keyword
  o Wikiminer
  o Wikiminer + keyword
  o Dbpedia
  o Dbpedia + keyword
  o TR-Esa
  o TR-Esa + keyword
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**
    
    \[
    MRR = \frac{\sum_{i=1}^{N} \frac{1}{\text{rank}_i}}{N}
    \]

    \( \text{rank}_i \) is the rank of the correctly retrieved services in the ranked list, and \( N \) is the number of the services correctly retrieved **with the configuration**.
## Experiment: Results

<table>
<thead>
<tr>
<th>Representation</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>
<th>MRR</th>
</tr>
</thead>
<tbody>
<tr>
<td>keyword (baseline)</td>
<td>0.26</td>
<td>0.37</td>
<td>0.43</td>
<td>0.46</td>
<td>0.48</td>
<td>0.48</td>
<td>0.33</td>
</tr>
<tr>
<td>tagme</td>
<td>0.11</td>
<td>0.15</td>
<td>0.16</td>
<td>0.16</td>
<td>0.16</td>
<td>0.16</td>
<td>0.42</td>
</tr>
<tr>
<td>tagme+keyword</td>
<td>0.27</td>
<td>0.39</td>
<td>0.44</td>
<td>0.48</td>
<td>0.50</td>
<td>0.50</td>
<td>0.33</td>
</tr>
<tr>
<td>wikiminer</td>
<td>0.10</td>
<td>0.12</td>
<td>0.12</td>
<td>0.13</td>
<td>0.13</td>
<td>0.13</td>
<td>0.45</td>
</tr>
<tr>
<td>wikiminer+keyword</td>
<td>0.27</td>
<td>0.38</td>
<td>0.43</td>
<td>0.47</td>
<td>0.49</td>
<td>0.49</td>
<td>0.33</td>
</tr>
<tr>
<td>tr-esa</td>
<td>0.28</td>
<td>0.41</td>
<td>0.47</td>
<td>0.56</td>
<td>0.64</td>
<td>0.68</td>
<td>0.22</td>
</tr>
<tr>
<td>tr-esa+keyword</td>
<td>0.28</td>
<td>0.41</td>
<td>0.47</td>
<td>0.56</td>
<td>0.64</td>
<td>0.68</td>
<td>0.22</td>
</tr>
<tr>
<td>dbpedia</td>
<td>0.16</td>
<td>0.21</td>
<td>0.22</td>
<td>0.23</td>
<td>0.24</td>
<td>0.24</td>
<td>0.43</td>
</tr>
<tr>
<td>dbpedia+keyword</td>
<td>0.28</td>
<td>0.39</td>
<td>0.44</td>
<td>0.47</td>
<td>0.49</td>
<td>0.50</td>
<td>0.33</td>
</tr>
</tbody>
</table>
Conclusions

- We proposed a cross-language matching method to support domain experts in linking eGov services
- Promising experimental results despite the minimal descriptions available
- Trade-off between coverage and quality of ranking
  - The best configuration in this domain is obtained using TR-ESA
- Adding external knowledge for representing a very short textual description is an effective solution in this specific domain
Thanks for your attention

cedelucio.narducci@uniba.it