Aligner automatiquement des ontologies avec MAP

Raphaël Troncy

Tuesday 23rd of January, 2007
Motivation

- Various SW repositories, using different vocabularies, distributed on the web
- Already large amounts of data out there
  - Swoogle hits $10^7 - 10^9$ unique Semantic Web documents (16/11/2006)

Problem:
- How to search and retrieve information in such an environment?
- How to map the various vocabularies used?
oMAP: Ontology Alignment Tool

- Formal and open framework
- Classifiers customization (parameter, chaining)
oMAP: A Formal Framework

Sources of inspiration:
- Formal work in data exchange [Fagin et al., 2003]
- GLUE: combining several specialized components for finding the best set of mappings [Doan et al., 2003]

Notation:
- A mapping is a triple: \( M = (T, S, \Sigma) \)
  - \( S \) and \( T \) are the source and target ontologies
  - \( S_i \) is an OWL entity (class, datatype property, object property) of the ontology
  - \( \Sigma \) is a set of mapping rules: \( \alpha_{ij} T_j \leftarrow S_i \)
oMAP: Combining Classifiers

- Weight of a mapping rule:
  \[ \alpha_{ij} = w(S_i, T_j, \sum) \]

- Using different classifiers:
  \[ w(S_i, T_j, CL_k) \text{ is the classifier's approximation of the rule } T_j \leftarrow S_i \]

- Combining the approximations:
  - Use of a priority list: \( CL_1 \prec CL_2 \prec \ldots \prec CL_n \)
  - Weighted average of the classifiers prediction
Terminological Classifiers

- **Same entity names (or URI)**

  \[ w(S_i, T_j, CL_N) = \begin{cases} 
  1 & \text{if } S_i, T_j \text{ have same name,} \\
  0 & \text{otherwise} 
\end{cases} \]

- **Same entity name stems**

  \[ w(S_i, T_j, CL_S) = \begin{cases} 
  1 & \text{if } S_i, T_j \text{ have same stem,} \\
  0 & \text{otherwise} 
\end{cases} \]
Terminological Classifiers

- **String distance name**

\[
w(S_i, T_j, CL_{LD}) = \frac{\text{dist}_{\text{Levenshtein}}(S_i, T_j)}{\max(\text{length}(S_i), \text{length}(T_j))}\]

- **Iterative substring matching**

\[
w(S_i, T_j, CL_{IS}) = \text{Comm}(S_i, T_j) - \text{Diff}(S_i, T_j) + \text{winkler}(S_i, T_j)\]

- See [Stoilos et al., ISWC'05]
Terminological Classifiers

WordNet distance name

\[
w(S_i, T_j, CL_{WN}) = \begin{cases} 
1 & \text{if } S_i, T_j \text{ are synonyms,} \\
\max \left( \frac{2 \cdot lcs}{\text{length}(S_i) + \text{length}(T_j)} \right) & \text{otherwise}
\end{cases}
\]

- \(lcs\) is the longest common substring between \(S_i\) and \(T_j\)
- \(sim = \frac{|\text{synonym}(S_i) \cap \text{synonym}(T_j)|}{|\text{synonym}(S_i) \cup \text{synonym}(T_j)|}\)
Collecting bag of words:
- *label* for the named individuals
- *data value* for the datatype properties
- *type* for the anonymous individuals and the range of object properties
- ...

Recursion on the OWL definition:
- *depth* parameter

Use statistical methods on the collected bag of words
Machine Learning-Based Classifiers

Example

Individual ($x_1$ type (Workshop))
value (label "DECOR") value (location $x_2$)

Individual ($x_2$ type (Address))
value (city "Namur") value (country "Belgium")

$u_1 = ("DECOR", "Address")$

$u_2 = ("Address", "Namur", "Belgium")$

Naïve Bayes text classifier

$$w(S_i, T_j, CL_{NB}) = Pr(S_i) \cdot \sum_{(x,u) \in T_j} \prod_{m \in u} Pr(m | S_i)$$

kNN text classifier
Structural and Semantics-Based Classifier

- $\sum$ is a set of mapping rules: $\alpha_{ij} T_j \leftarrow S_i$
- $\sum$ sets are computed by taking the OWL definition of the entities to align
  - recursively in the OWL structure
  - ... without looping thanks to cycles detection
Structural and Semantics-Based Classifier

If $S_i$ and $T_j$ are property names:

$$w(S_i, T_j, \Sigma) = \begin{cases} 0 & \text{if } T_j \leftarrow S_i \notin \Sigma \\ w'(S_i, T_j, \Sigma) & \text{otherwise} \end{cases}$$

If $S_i$ and $T_j$ are concept names$^1$:

$$w(S_i, T_j, \Sigma) = \begin{cases} 0 & \text{if } T_j \leftarrow S_i \notin \Sigma \\ w'(S_i, T_j, \Sigma) & \text{if } |D| = 0 \text{ and } T_j \leftarrow S_i \in \Sigma \\ \frac{1}{(|\text{Set}|+1)} \left( w'(S_i, T_j, \Sigma) + \max_{set} \left( \sum_{(C_i, D_j) \in \text{set}} w(C_i, D_j, \Sigma) \right) \right) & \text{otherwise} \end{cases}$$

$^1$Where $D = D(S_i) \times D(T_j)$; $D(S_i)$ represents the set of concepts directly parent of $S_i$. 


Structural and Semantics-Based Classifier

Let $C_S = (QR.C)$ and $D_T = (Q'R'.D)$, then\(^1\):

\[
  w(C_S, D_T, \Sigma) = w_Q(Q, Q') \cdot w(R, R', \Sigma) \cdot w(C, D, \Sigma)
\]

Let $C_S = (op C_1 \ldots C_m)$ and $D_T = (op' D_1 \ldots D_m)$, then\(^2\):

\[
  w(C_S, D_T, \Sigma) = w_{op}(op, op') \cdot \frac{\max_{set} \left( \sum_{(C_i, D_j) \in set} w(C_i, D_j, \Sigma) \right)}{\min(m, n)}
\]

\(^1\) Where $Q, Q'$ are quantifiers, $R, R'$ are property names and $C, D$ concept expressions

\(^2\) Where $op, op'$ are concept constructors and $n, m \geq 1$
Evaluation

  http://oaei.ontologymatching.org/
  - Systematic benchmark tests on bibliographic data
    - Tests 2xx: aligning an ontology with variations of itself where each OWL constructs are discarded or modified one per one
    - Tests 3xx: four real bibliographic ontologies
  - Web categories alignment
## Benchmark Tests (2005)

<table>
<thead>
<tr>
<th>Precision</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>dublin20</td>
<td>0.92</td>
</tr>
<tr>
<td>Falcon</td>
<td>0.91</td>
</tr>
<tr>
<td>FOAM</td>
<td>0.90</td>
</tr>
<tr>
<td>oMAP</td>
<td>0.85</td>
</tr>
<tr>
<td>CMS</td>
<td>0.81</td>
</tr>
<tr>
<td>OLA</td>
<td>0.80</td>
</tr>
<tr>
<td>ctxMatch</td>
<td>0.72</td>
</tr>
<tr>
<td>edna</td>
<td>0.45</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Recall</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Falcon</td>
<td>0.89</td>
</tr>
<tr>
<td>OLA</td>
<td>0.74</td>
</tr>
<tr>
<td>dublin20</td>
<td>0.72</td>
</tr>
<tr>
<td>FOAM</td>
<td>0.69</td>
</tr>
<tr>
<td>oMAP</td>
<td>0.68</td>
</tr>
<tr>
<td>edna</td>
<td>0.61</td>
</tr>
<tr>
<td>ctxMatch</td>
<td>0.20</td>
</tr>
<tr>
<td>CMS</td>
<td>0.18</td>
</tr>
</tbody>
</table>

- **oMAP:**
  - 4th with the global F-Measure
  - 1st on 3xx tests (*real* ontologies to align)
Aligning Web Categories (2005)

- Aligning Google, Loksmart and Yahoo web categories [Avesani et al., ISWC'05]
- Blind tests: only recall results are available

<table>
<thead>
<tr>
<th>ctxMatch</th>
<th>FOAM</th>
<th>CMS</th>
<th>Dublin20</th>
<th>Falcon</th>
<th>OLA</th>
<th>oMAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>9.4%</td>
<td>11.9%</td>
<td>14.1%</td>
<td>26.5%</td>
<td>31.2%</td>
<td>32.0%</td>
<td>34.4%</td>
</tr>
</tbody>
</table>
Conclusion

- oMAP: a formal framework for aligning automatically OWL ontologies
- Combining several specific classifiers
  - Terminological classifiers
  - Machine learning-based classifiers
  - Structural and semantics-based classifier
Future Work

- oMAP
  - Using additional classifiers:
    - KL-distance, other resources, background K, etc.
    - Straightforward theoretically but practically difficult!
  - Finding complex alignment
    - name = firstName + lastName
  - OWL and rule-based languages:
    - Take into account this additional expressivity
  - Other KR languages: e.g. SKOS
http://www.cwi.nl/~troncy/oMAP/

Any questions?
Structural and Semantics-Based Classifier

Possible values for $w_{op}$ and $w_Q$ weights

<table>
<thead>
<tr>
<th>$\cap$</th>
<th>$\cup$</th>
<th>$\neg$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\cap$</td>
<td>1</td>
<td>$1/4$</td>
</tr>
<tr>
<td>$\cup$</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$\neg$</td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$\exists$</th>
<th>$\forall$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\exists$</td>
<td>1</td>
</tr>
<tr>
<td>$\forall$</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$\leq n$</th>
<th>$\geq n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\leq m$</td>
<td>1</td>
</tr>
<tr>
<td>$\geq m$</td>
<td>1</td>
</tr>
</tbody>
</table>