Efficient Selection of Mappings and Automatic Quality-driven Combination of Matching Methods

Isabel F. Cruz, Flavio Palandri Antonelli, and Cosmin Stroe

ADVISIS Lab
University of Illinois at Chicago

With thanks to Ulas G. Keles and Angela Maduko

Supported by awards ITR IIS-0326284, IIS-0513553, and IIS-0812258
Outline

• The AgreementMaker
• Efficient Selection of Mappings
• Automatic Quality-driven Combination of Matching Methods
• A Practical Example: OAEI 2009
• Demonstration
AgreementMaker
Ontology Matching System

- Outputs *agreements* (mappings)
- Built for domain experts
- Triple focus
  - Matching methods
    - Wide range of automatic methods and their combination
  - Evaluation techniques
    - Comparison with “gold” standards and “inherent” quality measures
  - User interface
    - Supports all methods (manual, automatic, and semi-automatic) and evaluation techniques
- Extensible architecture
Multi-purpose User Interface
Extensible Architecture

Overall architecture

Matcher’s structure

First Layer

Second Layer

Third Layer

Specific focus!
Outline

• The AgreementMaker
• Efficient Selection of Mappings
• Automatic Quality-driven Combination of Matching Methods
• A Practical Example: OAEI 2009
• Demonstration
Mappings Selection Module

- **Input:**
  - the similarity matrix $M$
  - a threshold value $th \in [0,1]$ (e.g., 0.7)
  - the source and target cardinality constraints $sc-tc$ (e.g., 1-1, $n-m$, $n*$, $*-$)

- **Output:** a set of mappings $N$ that
  - maximizes the overall similarity of the selected mappings
  - satisfies threshold and cardinality constraints

\[
\begin{array}{ccccc}
S & & T & & \\
\hline
a & 0.5 & 0.7 & 0.7 & 0.1 \\
b & 0.0 & 0.7 & 0.9 & 0.8 \\
c & 0.7 & 0.8 & 0.5 & 0.1 \\
d & 0.0 & 0.0 & 0.9 & 0.6 \\
e & 1.0 & 0.0 & 0.3 & 0.1 \\
\end{array}
\]

\[
S \ T \ \text{Sim}
\begin{array}{ccc}
a & b' & 0.7 \\
a & c' & 0.7 \\
b & c' & 0.9 \\
b & d & 0.8 \\
c & a' & 0.7 \\
c & b' & 0.8 \\
d & c' & 0.9 \\
e & a' & 1.0 \\
\end{array}
\]

\[
sc-tc = 2- * \quad th = 0.7
\]

Mappings $N$
1-1 Matching Optimization Problem

Even a simple scenario can be tricky!

Greedy approach

Optimal approach

- This is an optimization problem!
  - Namely, the Assignment Problem
- Combinatorial methods are typically adopted
  - e.g., Hungarian Method $O(|M|^3)$ (too slow)
  - not feasible on large ontologies because of memory usage
Our Approach

- Reduce the 1-1 Mappings Selection problem to the Maximum Matching in a Weighted Bipartite Graph

**Similarity Matrix** \( M \)
- Threshold value \( th \)
- Source Concepts \( S \)
- Target Concepts \( T \)

**Compute the Weighted Bipartite Graph**
\( G = (S \cup T, E) \)

**Cut edges with weights lower than the threshold**

**Run the Shortest Augmenting Path Algorithm (LEDA)**

- \( w_{ij} = M[i,j] \)
- \( w_{ij} \geq th \)

**Optimal 1-1**

- Worst case \( O(n(m + n \log n)) \)
- Experimentally shown to be better
MWBM vs. Hungarian Method 1/2

- Significant improvement in execution time
- Efficient memory usage
  - with 1GB limit of memory the Hungarian Method won’t work on 3500x3500 matrices
MWBM vs. Hungarian Method 2/2

- Performances improve when the threshold value increases
Outline

• The AgreementMaker
• Efficient Selection of Mappings
• Automatic Quality-driven Combination of Matching Methods
• A Practical Example: OAEI 2009
• Demonstration
Can we trust the similarities computed by any matching method? Wrong similarities should be ignored.
Identifying Unreliable Similarity Values

1. The matching method compares features that are not available in the ontologies (e.g., label comparison of non-labeled ontologies) ⇒ all 0 similarity values

2. The matching method compares meaningless features (e.g., string comparison of numeric identifiers) ⇒ random similarity values

3. The matching method compares features that are identical for all concepts (e.g., structural comparison of non-hierarchical ontologies) ⇒ all 1 similarity values

4. All of the above: the matching method performs differently for each concept

A pretty confident mapping!
Local-Confidence Evaluation

- A local estimation of the reliability of the similarity values
- For each source concept \( c \), given the similarity matrix \( M \), the set of target concepts \( T \), and the target concepts mapped to \( c \) \( m_M(c) \), then:

\[
LC_M(c) = \frac{\sum_{c' \in m_M(c)} \text{sim}_M(c, c')}{|m_M(c)|} - \frac{\sum_{c' \in (T - m_M(c))} \text{sim}_M(c, c')}{|T - m_M(c)|}
\]

\[
S \text{ Selected} \quad \nabla \quad NS \text{ Not Selected}
\]

\[
\text{avg}(S) - \text{avg}(NS) = \text{confidence}
\]

\[
\begin{array}{cccc}
0 & 1 & 2 & 3 \\
0 & 0 & 0 & 0 \\
1 & 0 & 1 & 0 \\
2 & 0.1 & 0.2 & 0.9 & 0 \\
3 & 0.5 & 0.4 & 0.7 & 0.8 \\
4 & 1 & 1 & 1 & 1 \\
\end{array}
\]

\[
\rightarrow 0 - 0 = 0 \quad \times
\]

\[
\rightarrow 1 - 0 = 1 \quad \checkmark
\]

\[
\rightarrow 0.9 - 0.15 = 0.75 \quad \checkmark
\]

\[
\rightarrow 0.8 - 0.53 = 0.27 \quad \times
\]

\[
\rightarrow 1 - 1 = 0 \quad \times
\]
Outline

• The AgreementMaker
• Efficient Selection of Mappings
• Automatic Quality-driven Combination of Matching Methods
• A Practical Example: OAEI 2009
• Demonstration
A Practical Example: OAEI 2009

- **BSM**: Base Similarity Matcher
- **PSM**: Parametric String-based Matcher (substring + edit-dist)
- **VMM**: Vector-based Multi-term Matcher (TF-IDF + Cosine sim)
- **LWC**: Linear Weighted Combination (local-confidence weighting scheme)
- **DSI**: Descendant’s Similarity Inheritance (structural)
- **WordNet/UMLS**: Dictionaries (lexical)
Results of the Anatomy Track

- Mapping the adult mouse anatomy ontology (2744 classes) to the NCI thesaurus of the human anatomy (3304 classes)
- The AgreementMaker ranked **second** among ten systems

<table>
<thead>
<tr>
<th>Track</th>
<th>Goal</th>
<th>Rank</th>
<th>Additional Achievements</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1</td>
<td>Maximize F-measure</td>
<td>2nd</td>
<td>1st in Recall and Recall+</td>
</tr>
<tr>
<td>#2</td>
<td>Maximize Precision</td>
<td>2nd</td>
<td>0.006 distance from 1st, 1st in F-measure</td>
</tr>
<tr>
<td>#3</td>
<td>Maximize Recall</td>
<td>1st</td>
<td>1st also in Recall+</td>
</tr>
<tr>
<td>#4</td>
<td>Use a Partial Reference</td>
<td>5th</td>
<td>1st in Precision, improved execution time</td>
</tr>
</tbody>
</table>

### Additional Achievements

<table>
<thead>
<tr>
<th>System</th>
<th>Runtime</th>
<th>Task #1</th>
<th>Task #2</th>
<th>Task #3</th>
<th>Recall+</th>
</tr>
</thead>
<tbody>
<tr>
<td>AgrMaker</td>
<td>≈ 23 min</td>
<td>0.865</td>
<td>0.798</td>
<td>0.831</td>
<td>0.967</td>
</tr>
<tr>
<td>RiMOM</td>
<td>≈ 10 min</td>
<td>0.940</td>
<td>0.684</td>
<td>0.792</td>
<td>-</td>
</tr>
<tr>
<td>TaxoMap</td>
<td>≈ 12 min</td>
<td>0.870</td>
<td>0.678</td>
<td>0.762</td>
<td>0.953</td>
</tr>
<tr>
<td>DSSim</td>
<td>≈ 12 min</td>
<td>0.853</td>
<td>0.676</td>
<td>0.754</td>
<td>0.973</td>
</tr>
<tr>
<td>ASMOV</td>
<td>≈ 5 min</td>
<td>0.746</td>
<td>0.755</td>
<td>0.751</td>
<td>0.821</td>
</tr>
<tr>
<td>aflood</td>
<td>≈ 15 sec / 4 min</td>
<td>0.873</td>
<td>0.653</td>
<td>0.747</td>
<td>0.892</td>
</tr>
<tr>
<td>Lily</td>
<td>≈ 99 min</td>
<td>0.738</td>
<td>0.739</td>
<td>0.739</td>
<td>0.869</td>
</tr>
<tr>
<td>Aroma</td>
<td>≈ 1 min</td>
<td>0.775</td>
<td>0.678</td>
<td>0.723</td>
<td>-</td>
</tr>
<tr>
<td>kosimap</td>
<td>≈ 5 min</td>
<td>0.866</td>
<td>0.619</td>
<td>0.722</td>
<td>0.907</td>
</tr>
</tbody>
</table>
Results of the Conference Track

- Mapping 15 ontologies dealing with conference organization
- The AgreementMaker ranked **first** among seven systems with a threshold cutting of 75% and **second** with no threshold cutting

<table>
<thead>
<tr>
<th>matcher</th>
<th>threshold</th>
<th>P</th>
<th>R</th>
<th>F-meas</th>
</tr>
</thead>
<tbody>
<tr>
<td>afllood</td>
<td>*</td>
<td>48%</td>
<td>61%</td>
<td>52%</td>
</tr>
<tr>
<td>AgrMaker</td>
<td>0.75</td>
<td>69%</td>
<td>51%</td>
<td>57%</td>
</tr>
<tr>
<td>AMExt</td>
<td>0.75</td>
<td>54%</td>
<td>50%</td>
<td>51%</td>
</tr>
<tr>
<td>aroma</td>
<td>0.53</td>
<td>39%</td>
<td>48%</td>
<td>42%</td>
</tr>
<tr>
<td>ASMOV</td>
<td>0.23</td>
<td>68%</td>
<td>38%</td>
<td>47%</td>
</tr>
<tr>
<td>DSSim</td>
<td>*</td>
<td>15%</td>
<td>51%</td>
<td>22%</td>
</tr>
<tr>
<td>kosimap</td>
<td>0.51</td>
<td>52%</td>
<td>42%</td>
<td>45%</td>
</tr>
</tbody>
</table>
Outline

• The AgreementMaker
• Efficient Selection of Mappings
• Automatic Quality-driven Combination of Matching Methods
• A Practical Example: OAEI 2009
• Demonstration
Isabel F. Cruz, Flavio Palandri Antonelli, and Cosmin Stroe

ADVIS Lab
University of Illinois at Chicago

With thanks to Ulas G. Keles and Angela Maduko

Supported by awards ITR IIS-0326284, IIS-0513553, and IIS-0812258