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

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

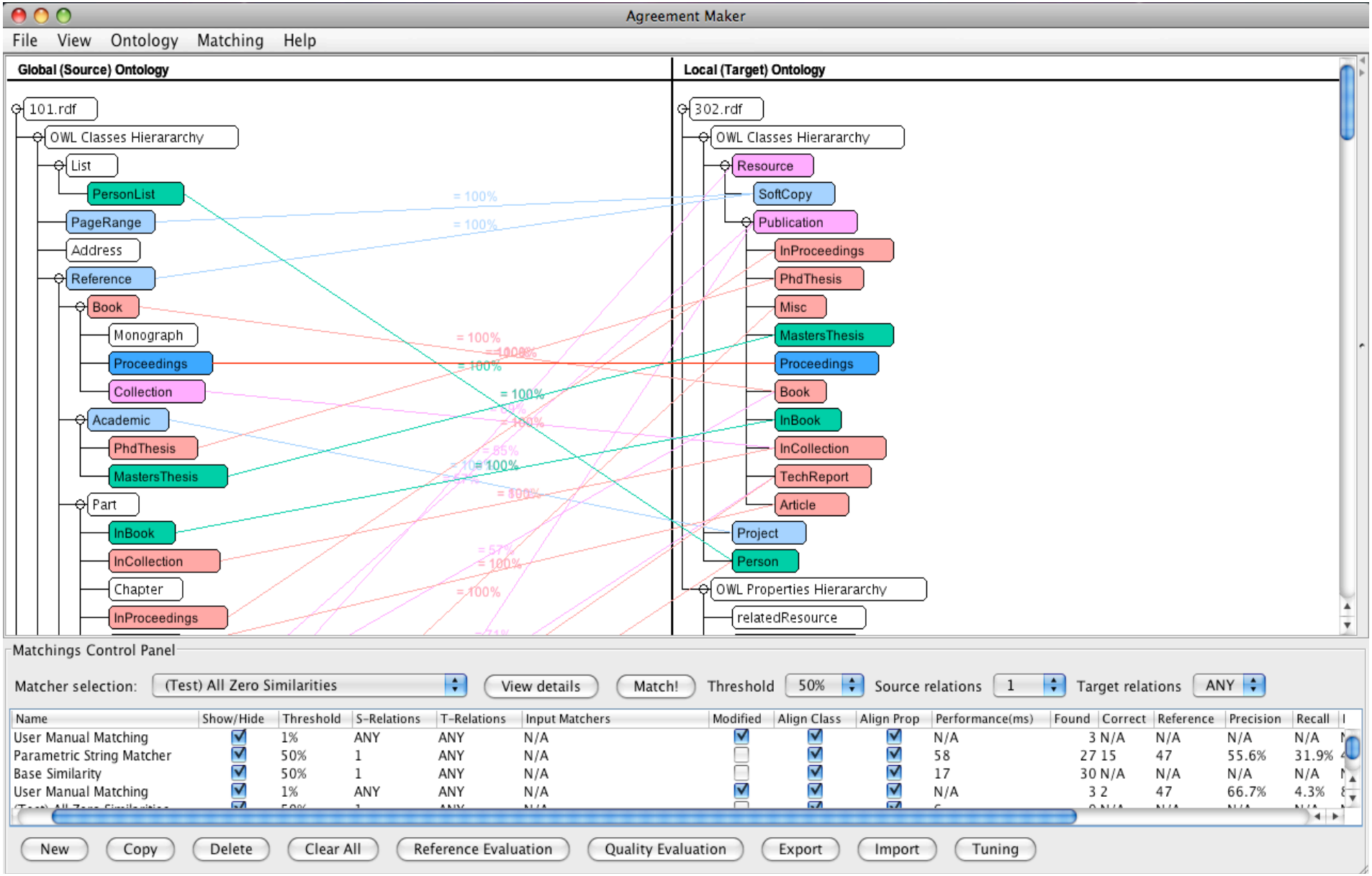
Outline

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

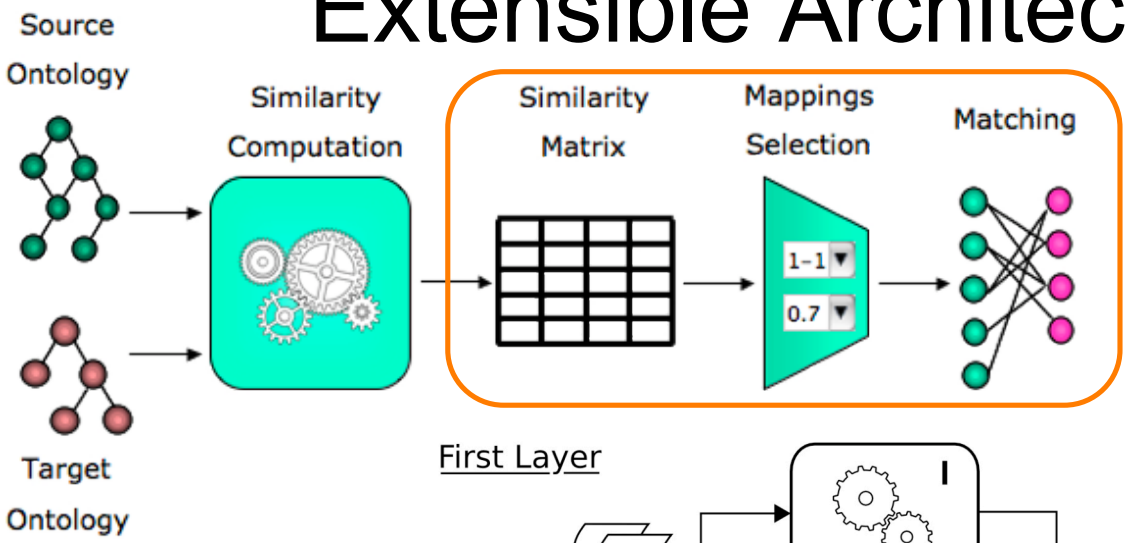
AgreementMaker
Ontology Matching System

- 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

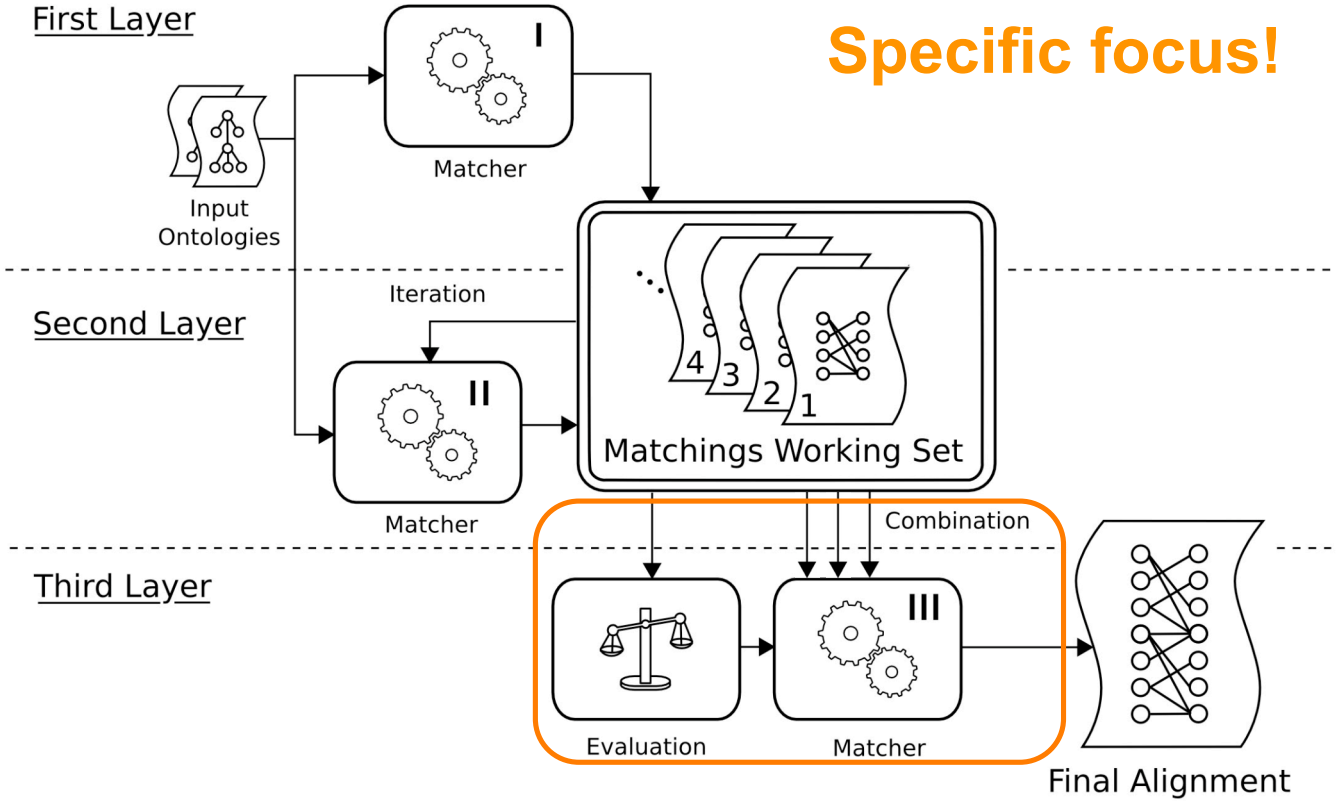


Extensible Architecture



Matcher's structure

Overall architecture



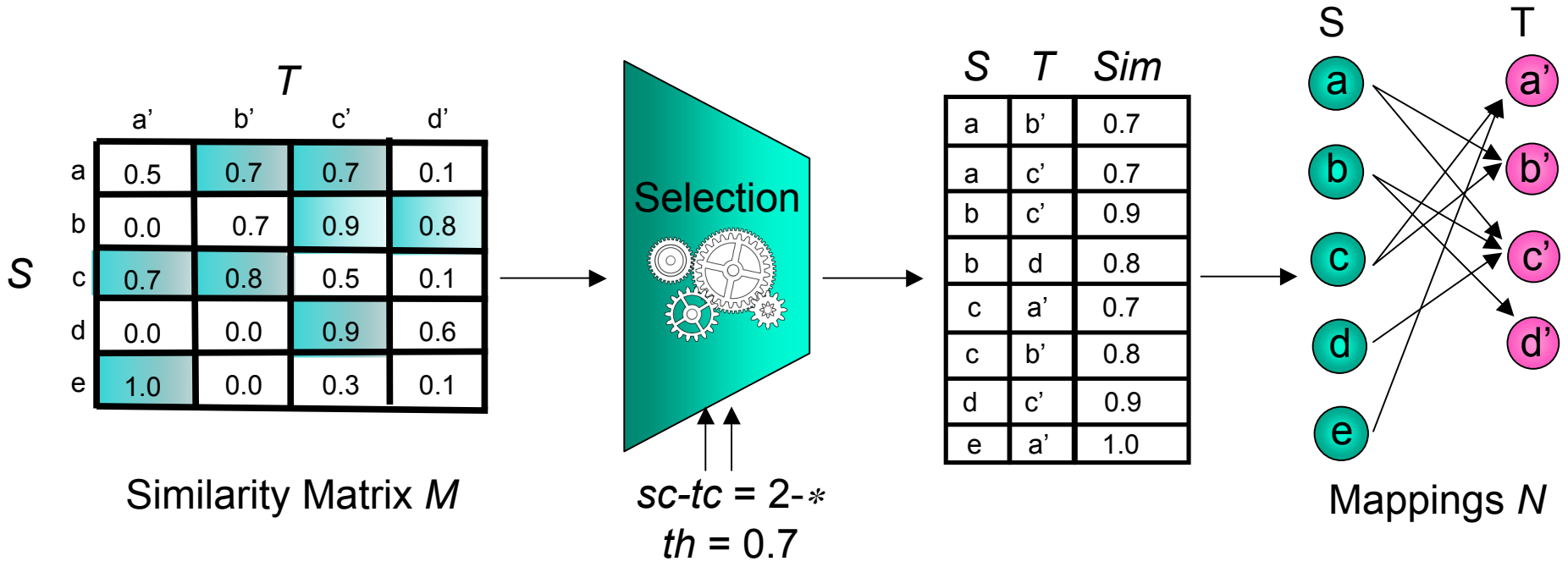
Specific focus!

Outline

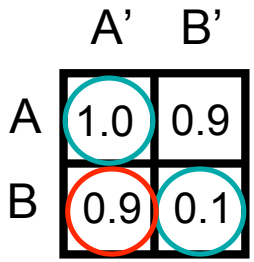
- The AgreementMaker
- Efficient Selection of Mappings
- Automatic Quality-driven Combination of Matching Methods
- A Practical Example: OAEI 2009
- Demonstration

Mappings Selection Module

- Input:
  - the similarity matrix *M*
  - a threshold value *th* ∈ [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

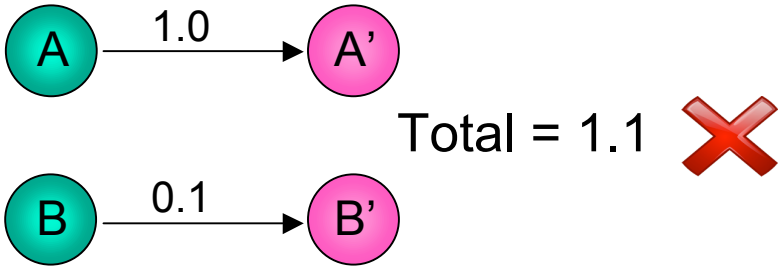


1-1 Matching Optimization Problem

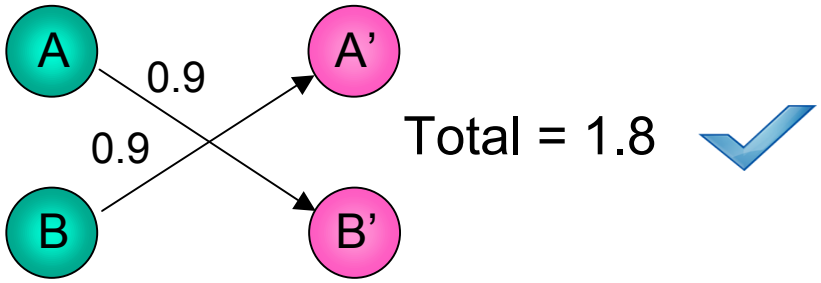


Even a simple scenario can be tricky!

Greedy approach



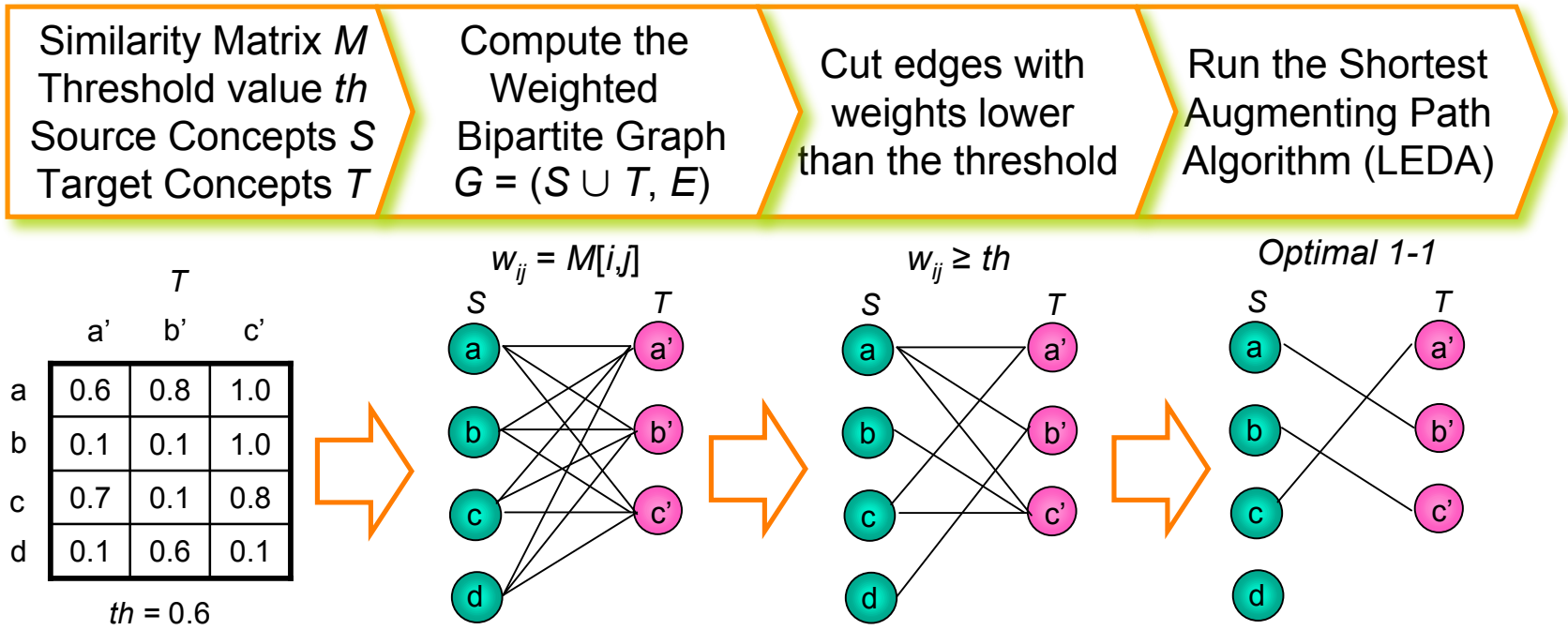
Optimal approach



- 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

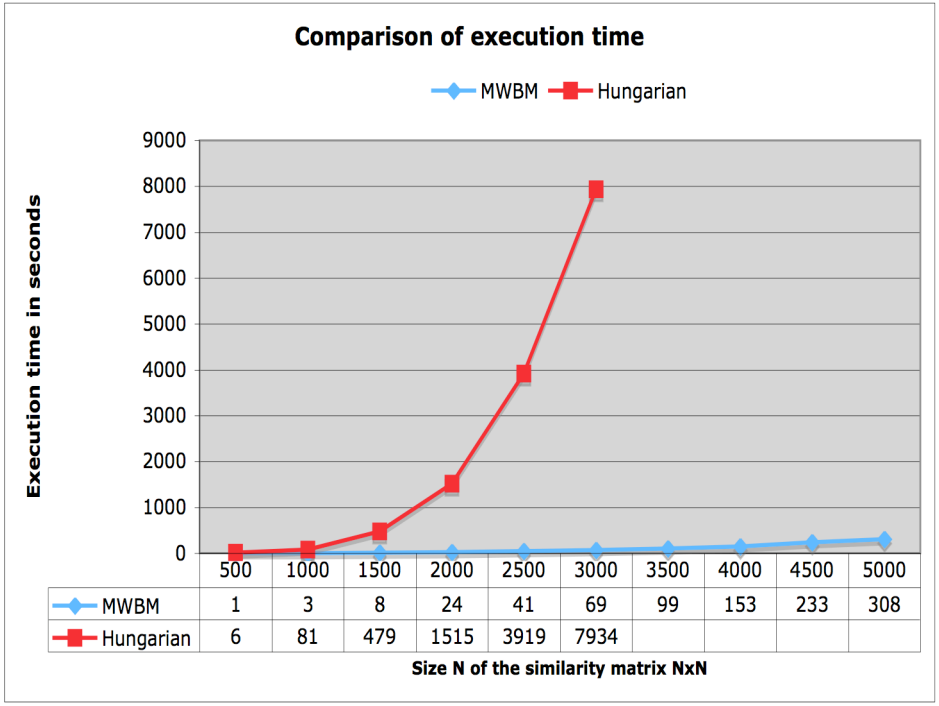
- Reduce the 1-1 Mappings Selection problem to the Maximum Matching in a Weighted Bipartite Graph



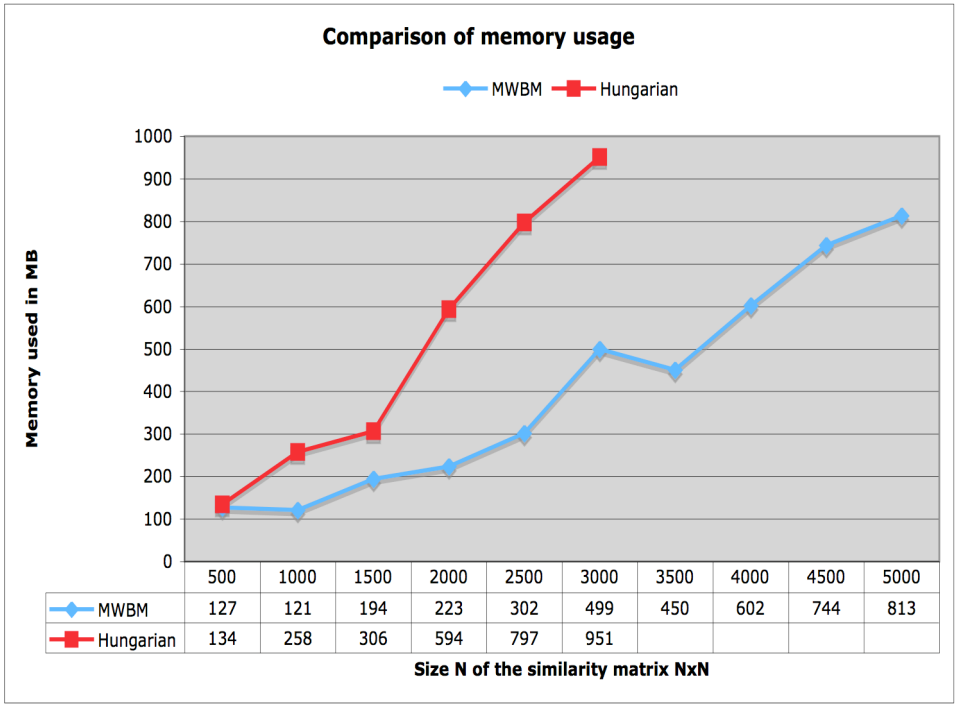
- Worst case $O(n(m + n \log n))$
- Experimentally shown to be better

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



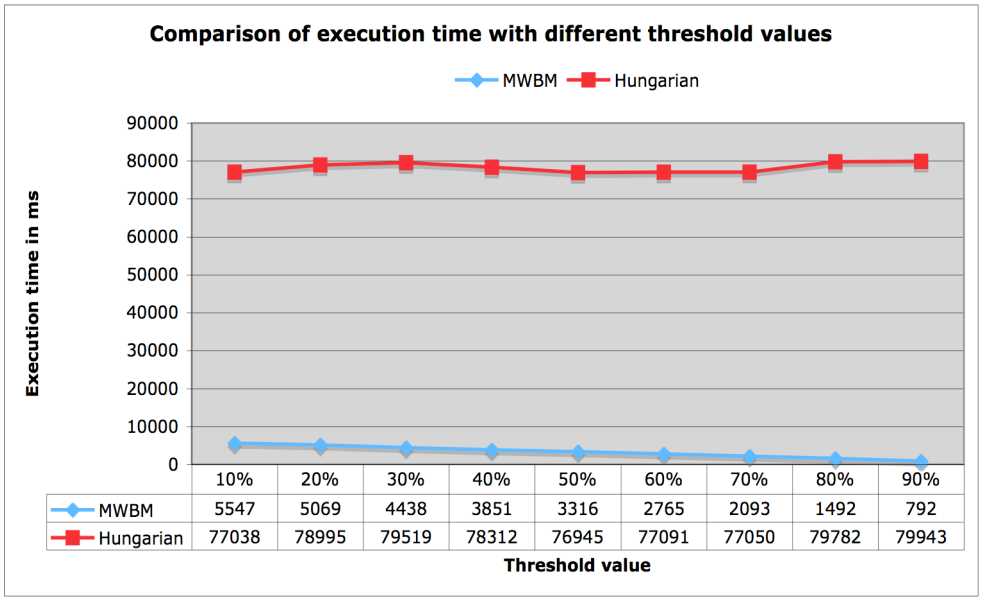
	500	1000	1500	2000	2500	3000	3500	4000	4500	5000
MWBM	1	3	8	24	41	69	99	153	233	308
Hungarian	6	81	479	1515	3919	7934				



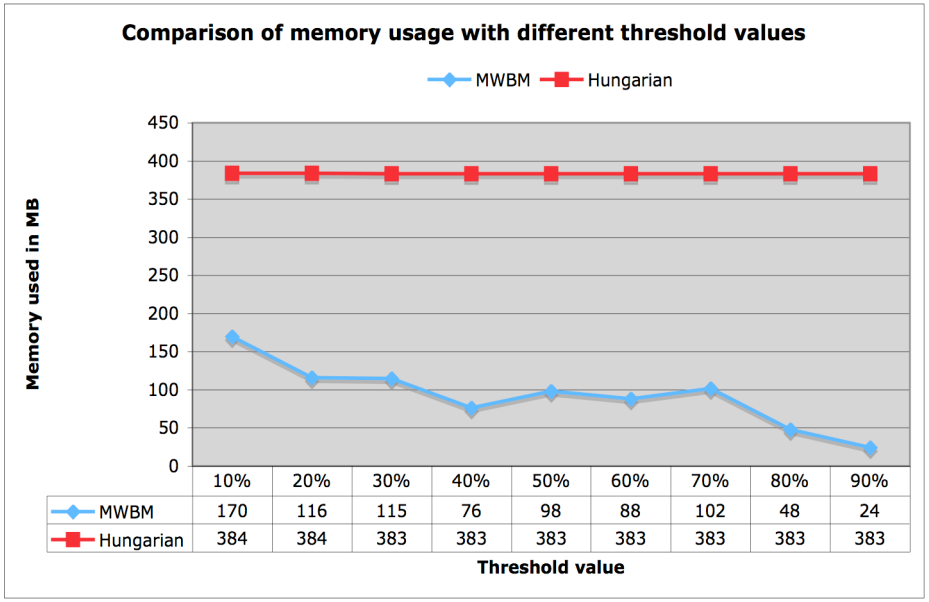
	500	1000	1500	2000	2500	3000	3500	4000	4500	5000
MWBM	127	121	194	223	302	499	450	602	744	813
Hungarian	134	258	306	594	797	951				

MWBM vs. Hungarian Method 2/2

- Performances improve when the threshold value increases



	10%	20%	30%	40%	50%	60%	70%	80%	90%
MWBM	5547	5069	4438	3851	3316	2765	2093	1492	792
Hungarian	77038	78995	79519	78312	76945	77091	77050	79782	79943

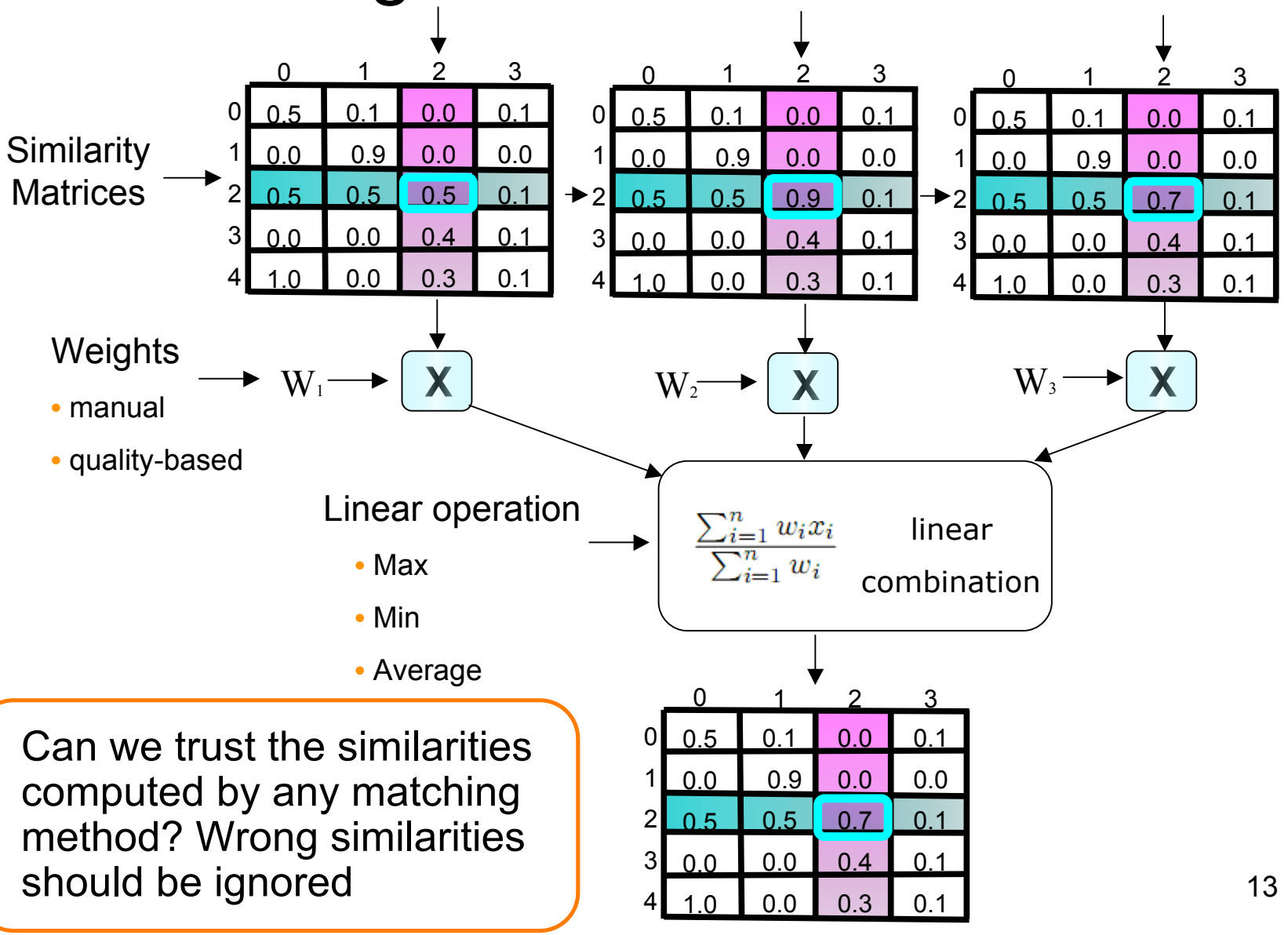


	10%	20%	30%	40%	50%	60%	70%	80%	90%
MWBM	170	116	115	76	98	88	102	48	24
Hungarian	384	384	383	383	383	383	383	383	383

Outline

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

Linear Weighted Combination Matcher



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

0.0	0.0	0.0
0.0	0.0	0.0
0.0	0.0	0.0
0.0	0.0	0.0

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

0.3	0.5	0.8
0.7	0.6	0.3
0.9	0.1	0.5
0.3	0.7	0.8

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

1.0	1.0	1.0
1.0	1.0	1.0
1.0	1.0	1.0
1.0	1.0	1.0

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

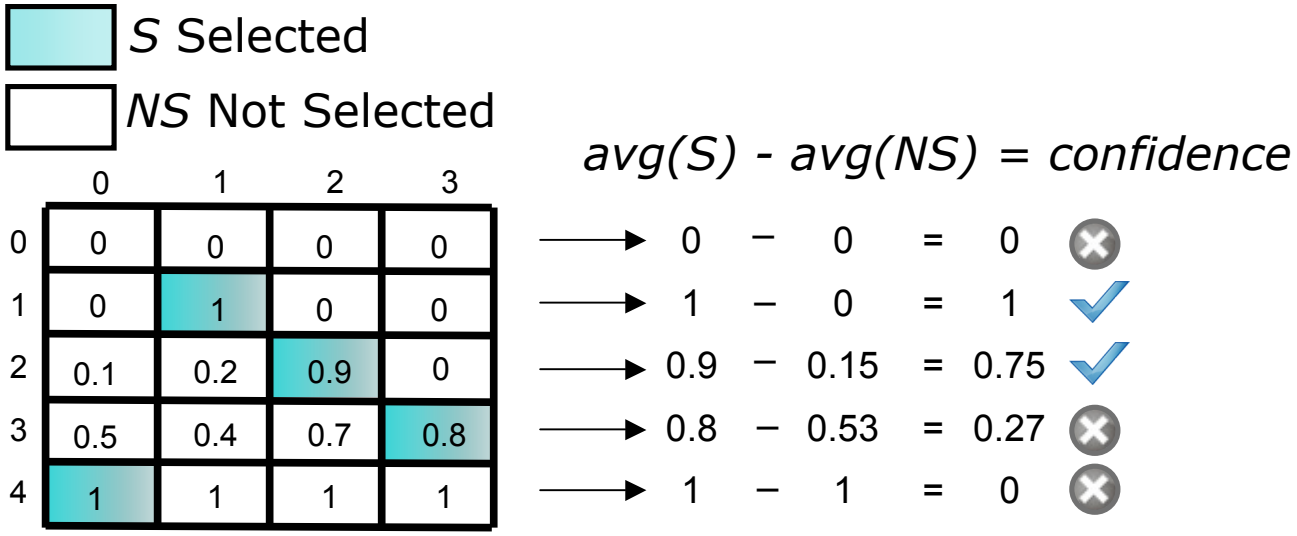
1.0	1.0	1.0
0.3	0.5	0.9
0.0	0.0	0.0
0.0	0.0	1.0

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:

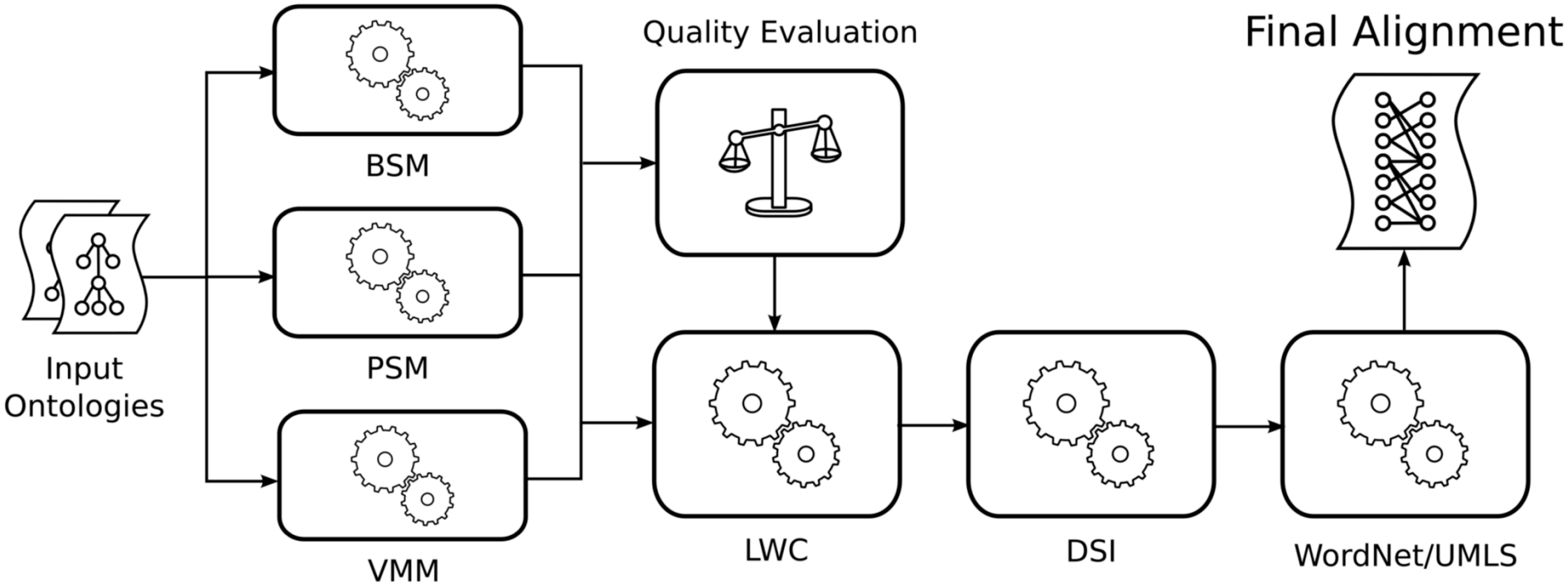
$$LC_M(c) = \frac{\sum_{c' \in m_M(c)} sim_M(c, c')}{|\, m_M(c) \,|} - \frac{\sum_{c' \in (T - m_M(c))} sim_M(c, c')}{|\, T - m_M(c) \,|}$$



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

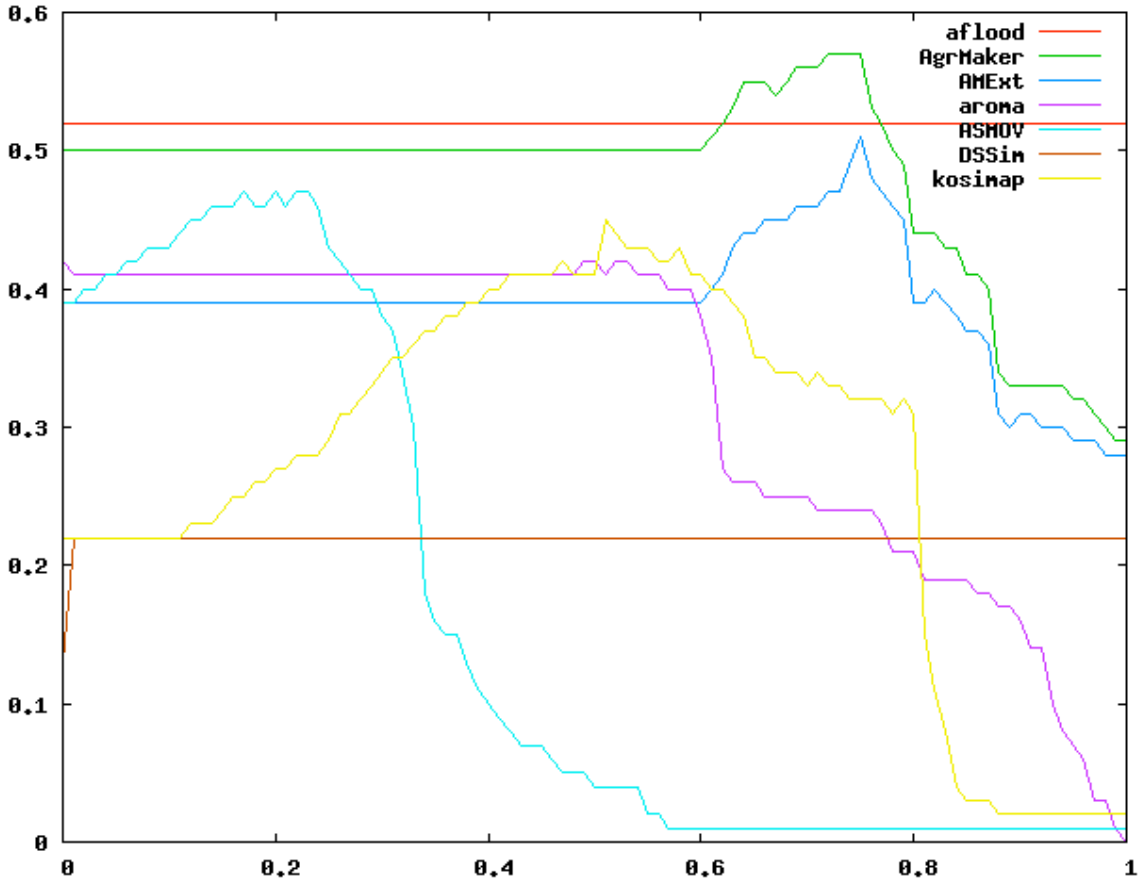
Track	Goal	Rank	Additional Achievements
#1	Maximize F-measure	2nd	1st in Recall and Recall+
#2	Maximize Precision	2nd	0.006 distance from 1st, 1st in F-measure
#3	Maximize Recall	1st	1st also in Recall+
#4	Use a Partial Reference	5th	1st in Precision, improved execution time

System	Task #1				Task #2			Task #3			Recall+	
	Runtime	Prec.	Rec.	F	Prec.	Rec.	F	Prec.	Rec.	F	#1	#3
SOBOM	≈ 19 min	0.952	0.777	0.855	-	-	-	-	-	-	0.431	-
AgrMaker	≈ 23 min	0.865	0.798	0.831	0.967	0.682	0.800	0.511	0.815	0.628	0.489	0.553
RiMOM	≈ 10 min	0.940	0.684	0.792	-	-	-	-	-	-	0.183	-
TaxoMap	≈ 12 min	0.870	0.678	0.762	0.953	0.609	0.743	0.458	0.716	0.559	0.222	0.319
DSSim	≈ 12 min	0.853	0.676	0.754	0.973	0.620	0.757	0.041	0.135	0.063	0.185	0.061
ASMOV	≈ 5 min	0.746	0.755	0.751	0.821	0.736	0.776	0.725	0.767	0.745	0.419	0.474
aflood	≈ 15 sec / 4 min	0.873	0.653	0.747	0.892	0.712	0.792	0.827	0.763	0.794	0.197	0.484
Lily	≈ 99 min	0.738	0.739	0.739	0.869	0.559	0.681	0.534	0.774	0.632	0.477	0.548
Aroma	≈ 1 min	0.775	0.678	0.723	-	-	-	-	-	-	0.368	-
kosimap	≈ 5 min	0.866	0.619	0.722	0.907	0.446	0.598	0.866	0.619	0.722	0.154	0.154

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

matcher	threshold	P	R	F-meas
aflood	*	48%	61%	52%
AgrMaker	0.75	69%	51%	57%
AMExt	0.75	54%	50%	51%
aroma	0.53	39%	48%	42%
ASMOV	0.23	68%	38%	47%
DSSim	*	15%	51%	22%
kosimap	0.51	52%	42%	45%



Outline

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

THANK YOU

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