A Distributional Approach to Evaluating Ontology Learning Methods Using a Gold Standard

E. Zavitsanos[1,2], G. Paliouras[1], G.A. Vouros[2]
izavits@iit.demokritos.gr, paliourg@iit.demokritos.gr, georgev@aegean.gr

[1]Institute of Informatics and Telecommunications, NCSR "Demokritos", Greece
[2]Department of Information and Communication Systems Engineering, AI-Lab, University of Aegean, Greece

3rd Ontology Learning and Population Workshop (OLP3), ECAI 2008

Outline

Why evaluate learned ontologies?

To ensure that:

1. The ontology fulfills some predefined standards
2. The ontology fulfills the requirements of its deployment
3. The learning method works well

Evaluation: How to..

Ontology evaluation categories:

1. Gold-standard-based evaluation
2. Application-based evaluation
3. Data-driven evaluation
4. Human evaluation

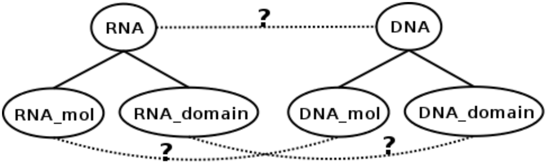
Evaluation using a Gold Standard

- Many approaches fall into this category
- Allows easy evaluation of several levels of the ontology
- Assumes that the gold ontology is the best one for the domain
- Depends on similarity measures used to compare the learned and the gold ontology

In a perfect world..

.. a perfect evaluation method

- would avoid common pitfalls of superficial string matching
- would recognize synonyms (car ?= automobile)



Our aim

We want to:

- Introduce an automated gold-standard-based evaluation method
- Avoid mismatches due to superficial string matching
- Deal with ontologies where the learning method does not label the identified concepts
- Introduce measures for automatically assessing the quality of the learned ontology

How do we evaluate so far?

- We need to match learned concepts to gold concepts. Use string matching and evaluate at the lexical layer (Ehrig et al. 2005) using Term/Lexical Precision and Term/Lexical Recall (Sabou et al. 2005 - Zavitsanos et al. 2007).
- We need to match learned subsumption relations to gold subsumption relations. Assume that the correct subsumption relations are those between the correctly identified concepts (Zavitsanos et al. 2007) and evaluate using Precision and Recall.

More fine-grained methods..

Need to take into account the position of the concepts in the hierarchy.

- Use Augmented Precision and Recall (Maynard et al. 2006). Take into account the distance from root and from common abstractions.
- Use OntoRand index (Brank et al. 2006). Take into account common ancestors for each concept, distances from other concepts in the hierarchy and the overlap of their sets of instances.
- Take into account the concepts in the vicinity of a concept. Semantic Cotopy; evaluate with Taxonomic Overlap (Maedche, Staab 2002).
- Common Semantic Cotopy; use the vicinity of concepts that belong in both ontologies (Dellschaft, Staab 2006).

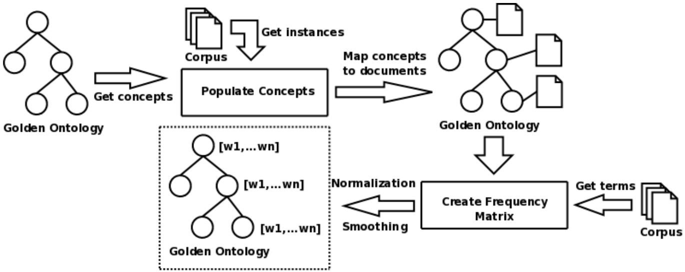
What we propose

- A method that evaluates learned hierarchies
- A distributional perspective to avoid common string matching
- Evaluation measures assessing the matching of concepts and hierarchical properties.

Ontology Transformation

- Represent each concept as multinomial probability distribution over the term space of the dataset.
- We assume that concept instances are known in texts.
- The context of a concept is the document(s) in which concept instances appear.
- Create a vector of frequencies of terms appearing in the context of each concept.
- Normalize and smooth (Laplace smoothing) to produce probability distributions for each concept.

Ontology Transformation (cont'd)



Benefits?

Through the distributional representation of each concept:

- Avoid typical string matching
- Various similarity measures can be used for matching the learned concepts to the gold concepts.
- Polysemy is addressed. Instances of the same term may be assigned to different concepts, according to the context(s) in which they appear.

Matching the Ontologies

- Use the distributional representation of a concept to perform one-to-one matching
- Total Variational Distance (TVD) used as distance measure between concepts.
  $TVD = \frac{1}{2} \sum_i \mid P(i) - Q(i) \mid$
- Number of mappings = Number of concepts in the smallest ontology
- Best mapping configuration: minimize $\sum_i^M TVD_i$, for all possible mappings M

Remarks

- In a perfect matching configuration the sum over all TVDs is zero
- One could choose to perform one-to-many matching
- A one-to-one matching is more demanding
- We have the flexibility to choose practically any matching method

Similarity Measures

We need similarity measures that follow well-defined criteria (Dellschaft, Staab 2006).

- We need measures that support automatic evaluation.
- Such measures must support ontology evaluation along multiple dimensions.
- An error must cause a change to the measure proportional to the distance between the gold and the learned ontology
- For measures with a range in a closed interval [0,1], gradual increase in error rate should lead to gradual decrease in the value of the function.

Proposed Measures

- $P_{value} = \frac{1}{M}\sum_{i=1}^{M}(1 - SD_i)PCP_i$
- $R_{value} = \frac{1}{M}\sum_{i=1}^{M}(1 - SD_i)PCR_i$
- $F_{value} = \frac{(\beta^2 + 1)P_{value} * R_{value}}{(\beta^2 R_{value}) + P_{value}}$

M: number of mapping pairs
SD: a distance measure between concepts ranging in [0,1]
The intensity of an error should depend on the position at which the error occurred in the taxonomy.
PCP: Probabilistic Cotopy Precision
PCR: Probabilistic Cotopy Recall

SD: which distance measure?

- We choose SD = TVD
- However, there exist many alternatives (Gibbs, Su 2002):
  - Kolmogorov Metric (KM)
  - Separation Distance (S)
  - Levy Metric (LM)
  - Prokhorov Metric (PM)

We have the flexibility to choose a distance measure

PCP ?? PCR ?? - Definitions

PCP and PCR are influenced by the notion of Semantic Cotopy.

For a matching i between a learned concept C_L and a gold concept C_G:

- PCP_i is the number of concepts in the cotopy set of C_L matched to concepts in the cotopy set of C_G, divided by number of concepts participating in the cotopy set of C_L.
- PCR_i is the number of concepts in the cotopy set of C_L matched to concepts in the cotopy set of C_G, divided by number of concepts participating in the cotopy set of C_G.

Cotopy Set of a concept C: the set of all its super and sub-concepts, including itself.

Remarks

- P_{value} reflects the similarity of two ontologies in the spirit of Precision
- R_{value} reflects the similarity of two ontologies in the spirit of Recall
- F_{value} combined measure governed by parameter β
- SD reflects matching differences between concepts
- PCP and PCR reflect differences at the taxonomical level
- P_{value}, R_{value}, and F_{value} incorporate all differences giving values in [0,1]

Setting the Experiments

We use two gold ontologies and the corresponding datasets:

1. Genia ontology: comprises 45 concepts from the domain of molecular biology
2. LonelyPlanet ontology: comprises 60 concepts from the tourism domain

We want to measure the scaling of the measures by introducing errors in the gold ontologies and by comparing the resulting ontologies to the original ones.

"Damage" Operators

Define some "damage" operators that introduce errors in the ontologies:

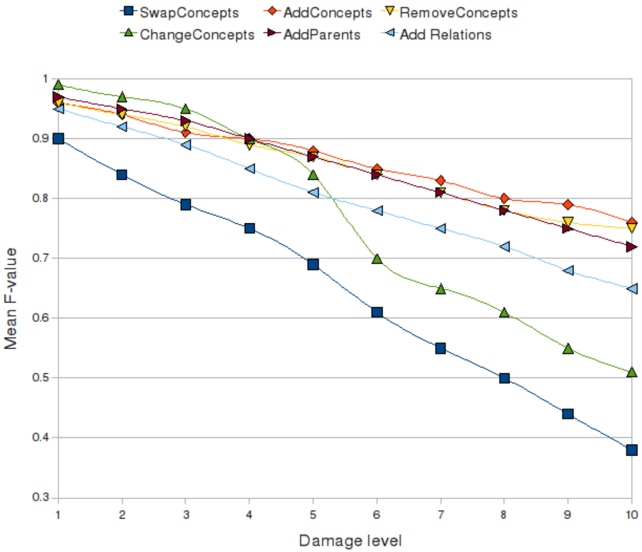
1. Swap Concepts
2. Remove Concepts
3. Add Concepts
4. Change Concept Distributions
5. Add Parents
6. Add Taxonomic Relations

A parameter indicates the extent of the damage (e.g. how many concepts to remove)

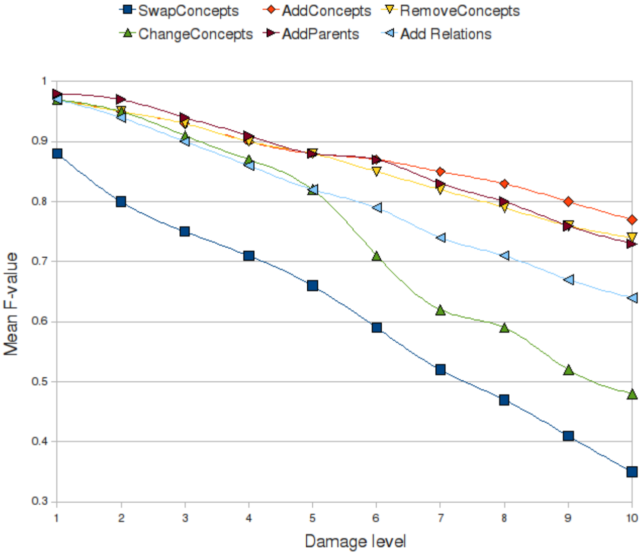
Setting the Experiments (cont'd)

- For each ontology, for each "damage" operator and for each of 10 "damage levels" we run 50 experiments and calculate the mean values.
- We present the F_{value} with $\beta = 1$, reflecting the harmonic mean of P_{value} and R_{value}.

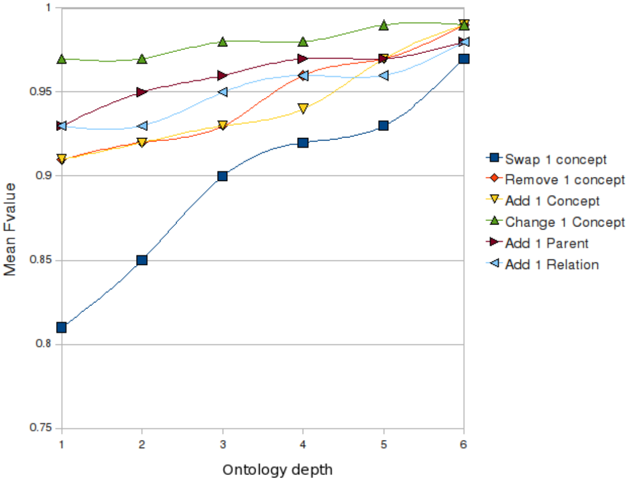
All "damage" operators in the Genia case



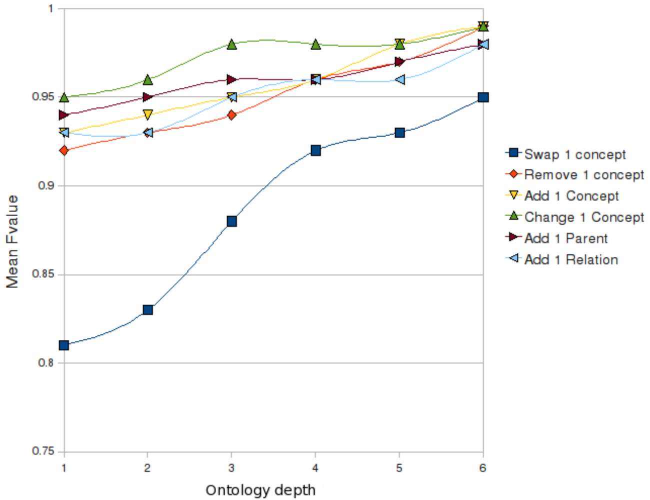
All "damage" operators in the LonelyPlanet case



All “damage” operators at different depths (Genia)



All “damage” operators at different depths (LonelyPlanet)



Remarks

- Gradual increases in the damage lead to gradual decreases of the F_{value} in [0,1]
- The measures are sensitive to errors introduced at different depths of the ontology
- Similar behavior in two different datasets and ontologies with different characteristics indicating potential robustness of the method

Conclusions

We presented:

- A method for gold-standard-based ontology evaluation
- A method that relies in the distributional representation of concepts
- A new set of similarity measures for the similarity of two hierarchies
- Experiments showing the flexibility of the method

Future Plans

1. We argue that the proposed measures can be used to evaluate ontologies containing non-taxonomic semantic relations
2. We need to provide a new definition of the Cotopy Set
3. Evaluation of ontologies learned from non-textual sources
4. Compare the behavior of F_{value} with other measures
5. Enhance the method with extra features based on lexical similarity
6. Investigate how to deal with ontologies containing cycles
7. Experiment with larger datasets and ontologies

Thank you!

Questions...?

References

- (Brank, Mladenic, Globelnik 2006) J. Brank, D. Mladenic, M. Grobelnik. "Gold standard based ontology evaluation using instance assignment", in Proceedings of the EON 2006 Workshop, 2006.
- (Dellschaft, Staab 2006) K. Dellschaft, S. Staab. "On how to perform a gold standard based evaluation of ontology learning", in Proceedings of the 5th International Conference on Semantic Web, 2006.
- (Ehrig, Haase, Stohanovic, Hefke 2005) M. Ehrig, P. Haase, N. Stohanovic, M. Hefke. "Similarity for ontologies - a comprehensive framework", In Proceedings of the European Conference in Inf. Sys., 2005.
- (Gibbs, Su 2002) A.L. Gibbs, F.E. Su. "On choosing and bounding probability metrics", International Statistical Review, 70(3), 419-435, 2002.

References (cont'd)

- (OAEI) Ontology Alignment Evaluation Initiative, http://oaei.ontologymatching.org.
- (Kalfoglou, Schorlemmer 2003) Y. Kalfoglou, M. Schorlemmer. "Ontology mapping: the state of the art", The Knowledge Engineering Review, 18(1), 2003.
- (Levenshtein 1966) I.V. Levenshtein. "Binary codes capable of correcting deletions, insertions and reversals", Cybernetics and Control Theory, 10(8), 707-710, 1966.
- (Maedche, Staab 2002) A. Maedche, S. Staab. "Ontology Learning for the Semantic Web", Kluwer 2002.

References (cont'd)

- (Maynard, Peters, Li 2006) D. Maynard, W. Peters, Y. Li. "Metrics for evaluation of ontology-based information extraction", in Proceedings of the EON 2006 Workshop, 2006.
- (Maedche, Staab 2002) A. Maedche, S. Staab. "Measuring similarity between ontologies", in Proceedings of the European Conference on Knowledge Acquisition and Management, 251-263, 2002.
- (Renner 2005) R. Renner. "On the variational distance of independently repeated experiments" 2005.
- (Sabou, Wroe, Goble, Stuckenschmidt 2005) M. Sabou, C. Wroe, C. Goble, H. Stuckenschmidt. "Learning domain ontologies for semantic web service descriptions", Journal of Web Semantics, 3(4), 2005.

References (cont'd)

- (Zavitsanos, Paliouras, Vouros, Petridis 2007) E. Zavitsanos, G. Paliouras, G.A. Vouros, S. Petridis. "Discovering subsumption hierarchies of ontology concepts from text corpora", in Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, 2007.