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

Elias Zavitsanos[1,2] and **Georgios Paliouras**[1] and **George A. Vouros**[2]

Abstract. This paper presents a method for the evaluation of learned ontologies against gold standards. The proposed method transforms the ontology concepts to a vector space representation to avoid the common string matching of concepts at the lexical layer. We propose a set of evaluation measures that exploit the concepts' representations and calculate the similarity of the two hierarchies. Experiments show that these measures scale gradually in the closed interval of $[0, 1]$ as learned ontologies deviate increasingly from the gold standard. The proposed method is tested using the Genia and the Lonely Planet gold standard ontologies.

1 INTRODUCTION

In the context of our work, ontology evaluation concerns the assessment of a learned ontology. This is done to ensure that it fulfills some predefined standards, and that it fulfills the requirements of its deployment. We may distinguish four major categories of ontology evaluation approaches: (a) those comparing the learned ontology to a predefined gold standard, which is usually a hand-crafted ontology, (b) those using the ontology in a system and evaluating the performance of the system, (c) those relying on a data-driven evaluation by comparing the ontology to existing data from the domain to which the ontology refers, and (d) those in which the evaluation is performed purely by human experts. Many approaches fall into the first category, i.e. evaluation using a gold standard ontology ([1], [2], [10], [9]).

A method that relies on a gold standard ontology allows easy evaluation of several levels of the learned ontology specifications (e.g. lexical, taxonomic, relational). On the other hand, this type of evaluation assumes that the gold standard represents well and captures all the significant knowledge of the domain, an assumption that in many cases may be faulty, since the gold standard is created by human experts and in many cases may be incomplete or developed in a biased way. To a large extent, this type of evaluation depends on the similarity measures that are used to compare the learned ontologies with the gold standard. This is a field on which little work has been done so far.

In this paper we describe a new method for automated evaluation of learned ontologies using a gold standard, avoiding common pitfalls of comparison at the lexical layer. For instance, assume the two ontologies illustrated in Figure 1. According to a string-matching technique, e.g. edit distance, the concept RNA would be matched with DNA, RNA_mol with DNA_mol and RNA_domain with DNA_domain, since they only differ in one letter. Obviously, this case would lead to matching completely different concepts, which have completely different instances (and thus meaning). Furthermore, comparing concepts lexicalized with very different terms, such as "car" or "automobile", would possibly never lead to a match.

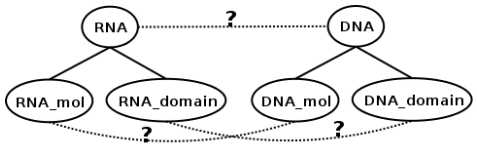

Figure 1. Example of possible mismatches between a gold ontology (left) and a learned ontology (right).

In contrast to this superficial lexical matching of concepts, the proposed method transforms the concepts of the gold standard and the learned ontology into distributions over the term[3] space of the dataset from which the ontology has been learned. A major advantage is that the learning method is not required to label the identified concepts. Moreover, we introduce a set of measures for automatically assessing the quality of the learned ontology by means of measuring its similarity with the gold standard.

In what follows we start by studying related work concerning the gold standard evaluation of ontologies (Section 2), while Section 3 presents the proposed method. In Section 4, we perform experiments and discuss the empirical evaluation results, and finally, Section 5 concludes the paper sketching plans for future work.

2 RELATED WORK

While a gold standard evaluation method, i.e. a method relying on the use of a gold standard ontology, may evaluate a learned ontology at the lexical, the taxonomic and the non-taxonomic levels, in this paper we concentrate on the evaluation of concept hierarchies. Therefore, we are not considering methods that evaluate ontologies at the non-taxonomic levels.

The similarity between two strings can be measured using the edit distance [7]. Based on the edit distance, a string-matching measure between two sets of strings can be defined by taking each string of the first set, finding its similarity to the "closest" string in the second set and averaging this over all strings of the first set. Using concept

[1] Institute of Informatics and Telecommunications, NCSR "Demokritos", Greece, email: {izavits | paliourg}@iit.demokritos.gr
[2] University of Aegean, Department of Information and Communication Systems Engineering, AI-Lab, Greece, email: georgev@aegean.gr

[3] "Terms" does not necessarily denote domain terms, but words that will constitute the vocabulary over which concepts will be specified. In the following, we use "terms" and "words" interchangeably.

names as strings we can obtain a comparison of two ontologies at the lexical level ([10], [3], [8]). The measures Term/Lexical Precision and Term/Lexical Recall have been introduced in [12].

Concerning the evaluation of concept hierarchies, the work in [13] evaluates the learned taxonomy using the measures of Precision and Recall, assuming that the correct subsumption relations are those between the correctly matched concepts. Furthermore, the measures of Augmented Precision and Recall presented in [9] can be used in the evaluation of taxonomies taking into account the position of the concepts in the hierarchy, as well as their distances from the root concept and their most specific common abstraction. Similar ideas have been followed in [3], where the measure of Taxonomic Similarity is introduced, given by the length of the shortest path between the matching concepts in the concept hierarchies.

Since the position of concepts in the hierarchy and the concepts in their vicinity play an important role on the taxonomic evaluation of ontologies, the method in [10] introduces the Taxonomic Overlap to compare two concepts in different hierarchies based on their Semantic Cotopies. The Semantic Cotopy of a concept is defined to be the set of all its super and sub concepts.

Finally, the OntoRand index [1] has also been proposed for comparing concept hierarchies. The hierarchy is viewed as a means for partitioning the set of instances. The comparison can be performed by measuring the similarity between concepts of different hierarchies based either on their common ancestors, or on their distances in the hierarchy, taking also into account the overlap of their sets of instances. Although this method treats concepts as clusters of instances, going beyond their lexical representation, it demands that both hierarchies contain exactly the same set of instances. Finally, techniques like [2], that use the notion of *Common Semantic Cotopy*, take into account only concepts that appear in both the learned and the gold ontologies with the same name.

According to the criteria for good evaluation measures presented in [2], we aim to evaluate two concept hierarchies by measuring their similarity, avoiding common problems introduced by matching only concept lexicalizations. The proposed method transforms the concepts into distributions over the term space of the dataset used for learning the ontology (in our case a set of text documents). An additional contribution of this paper is the introduction of a set of evaluation measures, inspired by information retrieval. These measures exploit the similarity between concept representations as vectors of distributions and their position in the hierarchies. The proposed measures scale gradually in the closed interval $[0, 1]$, according to the number of "errors" introduced in the learned ontology, compared to the gold standard.

3 THE PROPOSED EVALUATION METHOD

3.1 Ontology Transformation

The first step of the proposed method concerns the transformation of ontologies, so as to represent each concept as a probability distribution over the term space of the dataset that it covers. Towards this objective, assuming that the concept instances are annotated in the text documents that are the data sources, we measure the frequency of the terms that appear in the context of each concept instance. The context of a concept instance is assumed to be the document where the concept instance appears. As Figure 2 illustrates, having annotated the instances of concepts, it is possible to associate each document to the concept(s) that it refers to, by counting the concept instances that appear in the document[4]. Feature vectors corresponding to concepts record the frequency of each term in the context of each concept's instance in the documents.

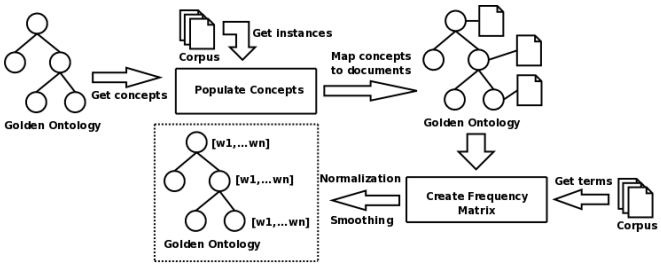

Figure 2. The transformation of the ontology concepts into probability distributions. Concepts are first populated in order to locate their contexts in the data set (corpus). Then, vectors of term frequencies are created based on the context of each concept. Finally, normalization and smoothing is performed to transform the vectors into probability distributions.

Finally, for each concept, the frequencies are normalized giving a probability distribution over the term space. In addition, we perform Laplace smoothing (Equation (1)) of the probability distributions to eliminate possible zero values of unseen terms. Both ontologies, i.e. the learned as well as the gold-standard one, are transformed to a common representation following the aforementioned procedure.

$$P(w_i) = \frac{P(w_i) + 1}{N + 1}, \forall i, \ (N : \ size\ of\ term\ space). \quad (1)$$

Via the ontology transformation process, ambiguity and polysemy are addressed, since instances of the same (ambiguous or polysemous) term may be assigned to different concepts, according to the context(s) in which they appear.

3.2 Matching the Ontologies

Using the representation of concepts, we perform a one-to-one matching between the gold and the learned concepts. Since both representations are based on probability distributions, an appropriate probability metric can be used to measure how "close" two concepts are. In this paper, we use the *Total Variational Distance* (TVD) [11], [4] to measure the similarity of two probability distributions $P(\cdot)$ and $Q(\cdot)$, which is defined as follows for a countable state space Ω (in our case, this is the countable term space of the corpus):

$$TVD = \frac{1}{2} \sum_i \mid P(i) - Q(i) \mid \quad (2)$$

TVD is one of the most commonly used probability metrics, because it admits natural interpretations, as well as useful bounding techniques. TVD takes into account each element of the two distributions, that is each term, in order to measure their average distance. At the level of individual terms, TVD reflects the largest possible difference between the probabilities that the two distributions can assign to the same term.

[4] In cases where the concept instances are documents (e.g. in a document indexing task), the mapping between concepts and documents is directly provided. Therefore, the concept population step can be skipped.

The mapping configuration, i.e. the set of matching pairs, determined by TVD, includes as many pairs as the number of concepts in the smaller ontology. Among all the possible mappings, the best configuration is determined, as the one that minimizes the sum of TVD over all mappings. In a perfect matching the TVD is equal to zero, and thus, in a perfect matching configuration, the sum over all mappings is equal to zero. Thus, according to Equation (3), among all the possible matching configurations N, we choose the one that minimizes the sum of TVD over all matching pairs M.

$$argmin_N\{\sum_i^M TVD_i\} \quad (3)$$

Besides the one-to-one matching, one could choose to perform a one-to-many matching, by matching one gold concept to many concepts in the learned ontology, or vice versa. However, assuming that the gold ontology is the best among all the possible ontologies representing concepts in a domain, a one-to-one matching is the most suitable, in the sense that it imposes a more strict evaluation.

3.3 Similarity Measures

Measuring the similarity between two ontologies automatically with a standard measure is an open research issue. According to [2], a measure must support evaluating an ontology along multiple dimensions (lexical, relational levels). Second, an error must cause a change to the measure proportional to the distance between the correct and the given result. Finally, for measures with a range in a closed interval, e.g. $[0, 1]$, a gradual increase in the error rate should also lead to a gradual decrease in the value of the evaluation function.

The proposed set of similarity measures is given in Equations (4), (5) and (6).

$$P_{value} = \frac{1}{M}\sum_{i=1}^{M}(1 - SD_i)PCP_i \quad (4)$$

$$R_{value} = \frac{1}{M}\sum_{i=1}^{M}(1 - SD_i)PCR_i \quad (5)$$

$$F_{value} = \frac{(\beta^2 + 1)P_{value} * R_{value}}{(\beta^2 R_{value}) + P_{value}} \quad (6)$$

In the above equations, M is the number of mapping pairs. SD is a distance measure between concepts, ranging in $[0, 1]$.

Although in this paper we use the TVD to measure the distance between concept representations, other metrics can also be used: (a) the Kolmogorov metric $KM = sup_{\forall i} \mid P(i) - Q(i) \mid, i \in \Re$, (b) the Separation distance $S = max_i(1 - \frac{P(i)}{Q(i)})$, (c) the Lévy metric $LM = inf\{\epsilon > 0 : P(x-\epsilon)-\epsilon \leq Q(x) \leq P(x+\epsilon)+\epsilon, \forall x \in \Re\}$, and (d) the analogue of the Lévy metric for more general spaces, called Prokhorov metric $PM = inf\{\epsilon > 0 : P(B) \leq Q(B^\epsilon)+\epsilon\}$, where $B^\epsilon = \{x : inf_{y \in B} d(x, y) \leq \epsilon\}$. The interested reader is referred to [4]. In addition, taking into account that the matching between the concepts of the two hierarchies can also be performed by applying any ontology alignment method ([6], [5]), one could use metrics introduced by these methods as the SD distance measure.

Concerning the taxonomic evaluation of an ontology, the intensity of an error should also depend on the position at which the error occurred in the taxonomy. For instance, removing a leaf concept that participates in a single subsumption relation should impose a smaller penalty than removing a central concept that is subsumed by some concepts and that also has a number of children.

The PCP and PCR (*Probabilistic Cotopy Precision and Recall*) factors in Equations (4) and (5) respectively, are influenced by the notion of Semantic Cotopy. For a matching i, where a concept C_L in the learned ontology and a concept C_G in the gold ontology are matched, PCP_i is defined as the number of concepts in the cotopy set of C_L matched to concepts in the cotopy set of C_G, divided by the number of concepts participating in the cotopy set of C_L. For the same matching i, PCR_i is defined as the number of concepts in the cotopy set of C_L matched to concepts in the cotopy set of C_G, divided by the number of concepts participating in the cotopy set of C_G. The cotopy set of a concept C is the set of all its super and sub-concepts including also the concept C.

Therefore, P_{value}, reflects the similarity of two ontologies in the spirit of Precision, penalizing learned concepts that do not appear in the gold standard ontology. On the other hand, the R_{value}, similar to Recall, reflects the similarity of the two ontologies, penalizing the learned ontology in cases where it does not include concepts that appear in the gold ontology. The F_{value} is a combined measure of the P_{value} and the R_{value}. It must be pointed that through the SD measure, differences at the lexical layer are captured, since a change in the lexicalization of a concept could lead the ontology transformation process to represent this concept via a different distribution of terms, and on the other hand, changes to the distributional representation of a concept actually result in describing a different concept. These differences are incorporated in the similarity measures through the SD factor, while the PCP and PCR factors are responsible for penalizing differences at the taxonomical level given the two ontologies.

In the general case, one could argue that the gold ontology, being hand-crafted by human experts, may have been developed in a biased manner or it may be incomplete for a specific domain. Therefore, one may be possibly interested only in the precision of the learning method, or only in recall. Thus, one could adjust the F_{value} of Equation (6) to focus more on the impact of the P_{value} or the R_{value} by adjusting the parameter β. In this paper the gold ontologies came with the datasets. Although they are hand-crafted by humans, we assume that they are accurate conceptualizations of the data that we study and thus, in our evaluation settings we choose $\beta = 1$ for the calculation of the F_{value}, which reflects the harmonic mean of P_{value} and R_{value}.

4 EXPERIMENTAL ASSESSMENT

In this section, we assess the method presented in section 3, by providing experimental results, checking the scaling of the measures according to errors introduced in a gold standard ontology.

The set of measures introduced in section 3.3 measure the similarity of the learned ontology to the gold one in a way that takes into account the differences between the gold and the learned concepts through the distributional distance metric SD, and the taxonomic differences of the hierarchies through the Probabilistic Cotopy Precision and Recall.

In order to study the behavior of the measures, we used two gold ontologies with their corresponding datasets: the Genia[5] ontology, comprising 45 concepts from the domain of molecular biology, and the Lonely Planet[6] ontology, comprising 60 concepts from the tourism domain. In order to measure the scaling of the measures, we

[5] The Genia project, http://www.tsujii.is.s.u-tokyo.ac.jp/GENIA

[6] The Lonely Planet travel advise and information, http://www.lonelyplanet.com

experimented by introducing errors in the gold ontologies, comparing the resulting ontologies to the original ones.

More specifically, we define the following "damage" operators: (a) Swap Concepts, (b) Remove Concepts, (c) Add Concepts, (d) Change Concept Distribution, and (e) Introduce taxonomic relations with existing or new concepts. Each "damage" operator, takes as input a number indicating the extent of the damage, i.e. how many concept pairs to be swapped, how many concepts to be removed/added, etc. Thus, for each ontology and for each "damage" operator we have run 50 experiments for 10 different values of its parameter, measuring the similarity of the resulting ontology with the original. We provide mean values that result from the averaging of the 50 experimental results for each of the values of the "damage" operators' parameters. Figures 3 and 4 depict this situation for all the experiments, while Figures 5 and 6 depict the mean value of the F_{value} in conjunction with the levels of the hierarchies for a specific parameter of all the "damage" operators. The resulting mean values reflect the behavior of the measures for various cases of the aforementioned operators:

(a) Swap Concepts: This operator picks randomly a predefined number of concept pairs and swaps them, introducing in this way invalid subsumption relations to the hierarchy. Since the number of concepts remains the same, the P_{value} and the R_{value} are affected in the same way. Figure 3 depicts the mean F_{value} obtained from these experiments in the Genia ontology, while the mean F_{value} obtained in the Lonely Planet ontology is depicted in Figure 4.

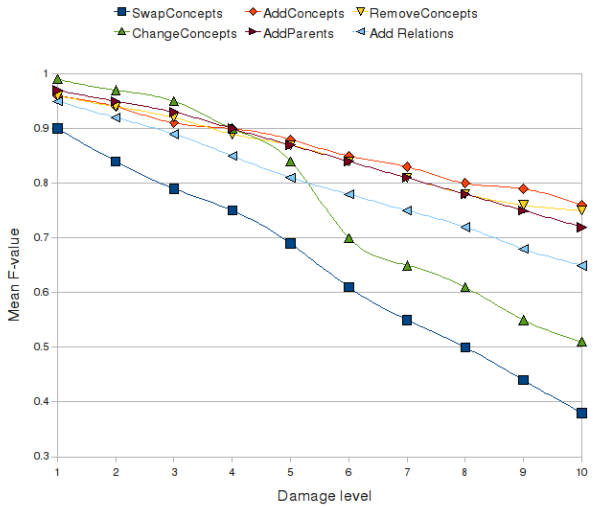

Figure 3. Combined diagram for all "damage" operators in the case of the Genia ontology. The mean F_{value} for "damage" level ranging from 1 to 10.

Swapping only one pair of concepts leads to a small taxonomic difference, especially when this operation is performed in the leaf concepts of the hierarchy. For the two ontologies, the 50 different experiments of swapping one pair of concepts result to an F_{value} between 0.81 and 0.99 (Figures 5, 6), depending on the position of the concepts that are swapped. As the number of concepts that are swapped gradually increases, the F_{value} decreases almost linearly, reaching a situation of swapping 10 pairs of concepts that changes over half of the subsumption relations of the hierarchy.

(b) Remove Concepts: In this case, a predefined number of randomly chosen concepts is removed from the ontology. Thus the

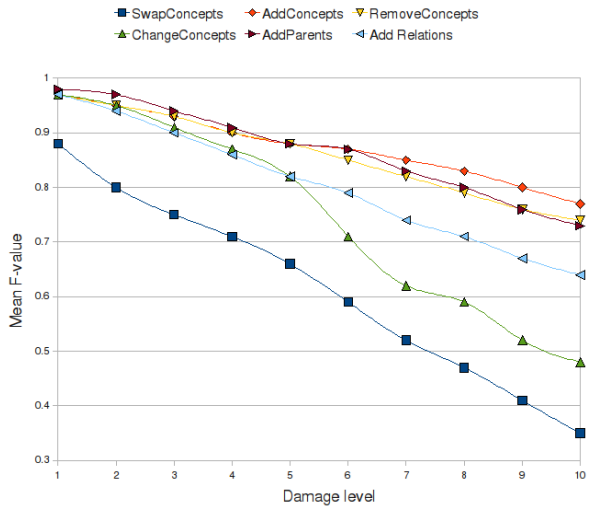

Figure 4. Combined diagram for all "damage" operators in the case of the Lonely Planet ontology. The mean F_{value} for "damage" levels ranging from 1 to 10.

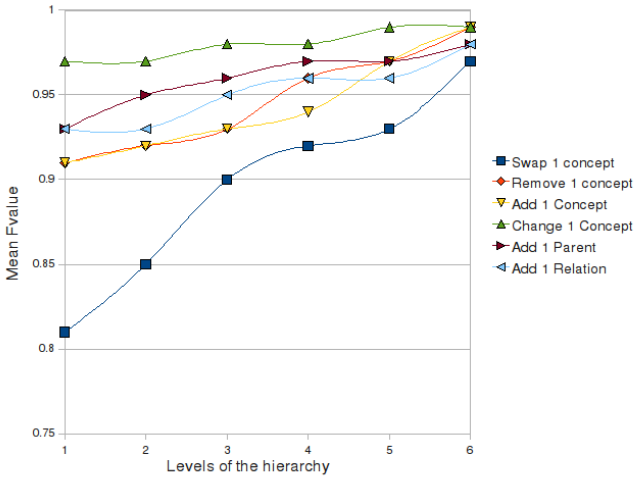

Figure 5. Combined diagram for all "damage" operators in the case of the Genia ontology. The mean F_{value} for "damage" level=1 for all levels of the hierarchy.

R_{value} is affected, while the P_{value} remains equal to 1. This operator affects the taxonomic structure of the gold ontology, as some concepts that appear in the gold standard disappear. Figures 3 and 4 present the mean F_{value} for various cases for the Genia and the Lonely Planet ontology respectively.

Removing only one concept from the hierarchy leads - quantitatively - to a small taxonomic difference. As already mentioned, if this is a leaf concept, the penalty is smaller. For the two ontologies, the 50 different experiments of removing a single concept result in F_{value} between 0.90 and 0.99 (Figures 5, 6), depending on whether this concept is a leaf or a central concept. As shown in Figures 3 and 4, the decrease of the F_{value} is linearly related to the extent of the

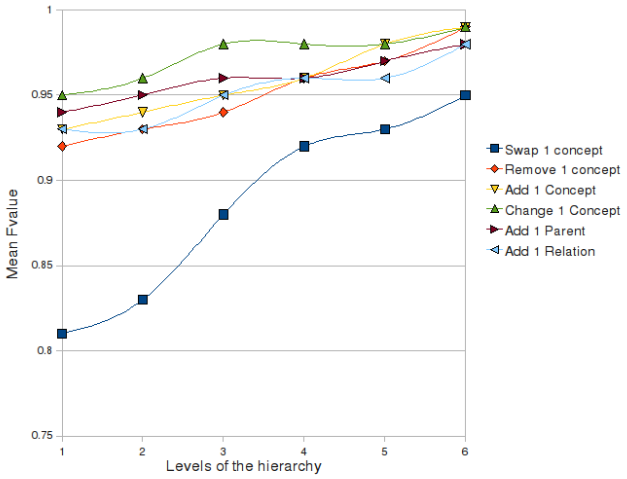

Figure 6. Combined diagram for all "damage" operators in the case of the Lonely Planet ontology. The mean F_{value} for "damage" level=1 for all levels of the hierarchy.

damage and less steep than for concept swapping. This is because the removal of a concept affects the cotopy set of a concept by decreasing its size by one. On the other hand, a swap between concepts may affect also the cotopy set of a second concept by introducing a large number of concepts to its cotopy set, depending on the new position of the swapped concepts.

(c) Add Concepts: This operator adds a predefined number of new concepts randomly to the ontology as children of already existing concepts, maintaining this way the tree-like structure of the hierarchy. Thus the P_{value} is affected, while the R_{value} remains equal to 1. Figures 3 and 4 present the results in the Genia and the Lonely Planet cases.

Adding only a single concept to the hierarchy introduces a small error to the hierarchy. For the 50 different experiments of adding only one concept, the F_{value} is between 0.91 and 0.99 (Figures 5, 6). As the number of concepts that are added increases, the F_{value} is affected similarly to concept removal.

(d) Change Concept Distribution: In this case the number of concepts remains intact. Thus the P_{value} and the R_{value} are both affected in the same way. The error is introduced in the distributional representations of randomly picked concepts. The changes affect the frequency of the terms that appear in the contexts of the concepts. Therefore this operator has an impact on the lexical layer of the ontology. Figures 3 and 4 show how the F_{value} is affected in both ontologies.

In 50 different experiments of disturbing a randomly chosen concept, the F_{value} is between 0.97 and 0.99 (Figures 5, 6). However, changing the distributional representation of more than one concept, the F_{value} decreases non-linearly. It should be pointed out that this operator can lead to the extreme situation where a concept is changed completely, setting the SD of Equations (4) and (5) equal to 1. Essentially, in this situation, the matching ceases to exist. Regarding the non-linear decrease of the F_{value}, we should stress out that changing the distribution of a concept has an impact in the evaluation measures, depending both on how much a concept's representation has been changed and on the taxonomic position of this concept. Therefore strong changes in central concepts with many relations have a larger effect in the F_{value}.

(e) Introduce Taxonomic Relations: The last operator introduces new taxonomic relations in the concepts of the ontology. Two cases are foreseen: (1) add randomly new concepts as parents to randomly chosen existing concepts, and (2) add randomly new subsumption relations between existing concepts. In both cases, the error has an impact on the taxonomic layer in the sense that the taxonomy might not be a tree-like structure anymore, and also new concepts and subsumption relations are introduced, that do not appear in the gold standard ontology. Thus, the P_{value} is affected, while the R_{value} remains equal to 1. Figures 3 and 4 depict the F_{value} for the Genia and the Lonely Planet ontology respectively. One could expect the introduction of new subsumption relations through the addition of parent concepts to have a larger impact on the measures. However, the experimental results show the opposite. This is because the addition of only one subsumption relation between existing concepts, depending on their position, may change the hierarchy significantly and introduce more than one concepts into the cotopy sets of the concepts that participate in this relation. Obviously, the more relations added, the larger the impact on the F_{value}.

We should mention at this point that the tree-like structure of the hierarchy may be damaged through the application of this operator. For instance, multiple inheritance between concepts might be introduced. Moreover, it is possible through the addition of new taxonomic relations between concepts to introduce cycles in the ontology. Due to the definition of the cotopy set of a concept, we prohibit such situations through a fail-safe mechanism in our evaluation tool. Therefore, cases where this damage operator has introduced cycles were ignored. However, this experiment gave us the opportunity to further investigate how such a situation should be addressed by the proposed method in the future.

In general, as all diagrams illustrate, gradual increases in the damage lead to gradual decreases of the F_{value} in the closed interval $[0, 1]$. Moreover, the measure seems to be sensitive enough to errors introduced both at the lexical and the taxonomic layer of the ontologies. Due to space limitations we presented here only a representative subset of the experimental results.

Concerning the two datasets, the similar behavior of the method indicates that it tends to be unbiased by different data sources, i.e. the gradual increase in the damage, leads to gradual decrease of the F_{value} irrespective of the dataset.

The spread of the values of F_{value} in similar spectrums in both datasets is due to the fact that the Genia and the Lonely Planet ontology have similar sizes and the same depth. Particularly, similarity in the depth is important since PCP and PCR take into account concepts that are descendants and ancestors of the concepts that participate in a particular matching. Obviously, the similarity increases in cases where the branching factor of the two hierarchies is similar.

Finally, we argue that the proposed similarity measure can be applied to ontologies that contain also other semantic relations, beyond subsumption. A new definition of the *Cotopy Set* of a concept C is required, taking into account the concepts that are connected to C through various semantic relations. The evaluation measures remain intact, computing the similarity of the two ontologies over all the matching pairs of the matching configuration, irrespective of whether a matching is between concepts that participate in a subsumption or in a different semantic relation. This hypothesis requires further experimental validation.

5 CONCLUSIONS

In this paper we presented a novel method for evaluating learned ontologies against gold standard ontologies. The paper proposes a new set of evaluation measures, relying on a distributional representation of the concepts based on their instances annotated in a given corpus. The proposed approach avoids common problems of evaluating ontologies at the lexical level. The similarity measures proposed here, take into account both the lexical and the taxonomic dimensions of the ontology. The generality of the similarity measures allows a flexible choice of distance measure to be used. Experimental results showed that the method penalizes in a near-linear fashion the increasing difference of two ontologies, taking values in the closed interval $[0, 1]$.

Future plans include the use of the method in ontologies that contain relations beyond the taxonomic backbone. In addition, future experiments involve evaluation between ontologies that have been learned from non-textual sources, and thus do not have a distributional representation over terms, but over different types of feature. Finally, we also plan to compare the behavior of the F_{value} with other metrics mentioned in the literature, as well as to enhance the method with extra features based on lexical similarity.

ACKNOWLEDGEMENTS

The presented work was supported be the research and development project ONTOSUM[7], funded by the Greek General Secretariat for Research and Technology.

REFERENCES

[1] J. Brank, D. Mladenić, and M. Grobelnik, 'Gold standard based ontology evaluation using instance assignment', in *Proceedings of the EON 2006 Workshop*, (2006).
[2] K. Dellschaft and 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).
[3] M. Ehrig, P. Haase, N. Stohanovic, and M. Hefke, 'Similarity for ontologies - a comprehensive framework', in *Proceedings of the European Conference in Inf. Sys.*, (2005).
[4] A.L. Gibbs and F.E. Su, 'On choosing and bounding probability metrics', *International Statistical Review*, **70(3)**, 419–435, (2002).
[5] Ontology Alignment Evaluation Initiative, 'http://oaei.ontologymatching.org'.
[6] Y Kalfoglou and M. Schorlemmer, 'Ontology mapping: The state of the art', *The Knowledge Engineering Review*, **18(1)**, (2003).
[7] I.V. Levenshtein, 'Binary codes capable of correcting deletions, insertions and reversals', *Cybernetics and Control Theory*, **10(8)**, 707–710, (1966).
[8] A. Maedche and S. Staab, *Ontology Learning for the Semantic Web*, Kluwer, 2002.
[9] D. Maynard, W. Peters, and Y. Li, 'Metrics for evaluation of ontology-based information extraction', in *Proceedings of the EON 2006 Workshop*, (2006).
[10] A. Meadche and S. Staab, 'Measuring similarity between ontologies', in *Proceedings of the European Conference on Knowledge Acquisition and Management (EKAW)*, pp. 251–263, (2002).
[11] Renato Renner. On the variational distance of independently repeated experiments, 2005.
[12] M. Sabou, C. Wroe, C. Goble, and H. Stuckenschmidt, 'Learning domain ontologies for semantic web service descriptions', *Journal Of Web Semantics*, **3(4)**, (2005).
[13] E. Zavitsanos, G. Paliouras, G.A. Vouros, and S. Petridis, 'Discovering subsumption hierarchies of ontology concepts from text corpora', in *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence - WI '07*. Springer-Verlag, (2007).

[7] See also *http://www.ontosum.org/*