Gold Standard Evaluation of Ontology Learning Methods Through Ontology Transformation and Alignment

Elias Zavitsanos, Georgios Paliouras, and George A. Vouros

Abstract—This paper presents a method along with a set of measures for evaluating learned ontologies against gold ontologies. The proposed method transforms the ontology concepts and their properties into a vector space representation to avoid the common string matching of concepts and properties at the lexical layer. The proposed evaluation measures exploit the vector space representation and calculate the similarity of the two ontologies (learned and gold) at the lexical and relational levels. Extensive evaluation experiments are provided, which show that these measures capture accurately the deviations from the gold ontology. The proposed method is tested using the Genia and the Lonely Planet gold ontologies, as well as the ontologies in the benchmark series of the Ontology Alignment Evaluation Initiative.

Index Terms—Knowledge Valuation, Machine Learning, Concept Learning, Ontology Design

1 Introduction

In the context of this paper, ontology evaluation concerns the assessment of an ontology that is produced by an ontology learning method. Ontology evaluation is performed in order to ensure that the learned ontology adheres to some predefined standards, represents accurately the domain that it covers, and in general, fulfills the requirements of its deployment. Regarding the automated learning of ontologies, evaluation methods are very much needed in order to decide which learning method produces the most suitable ontology in terms of the concepts and properties learned and in terms of their relations, with respect to a given domain. Beyond the evaluation of ontology learning methods, the automated ontology evaluation is crucial for the engineering of ontologies, since developers need to decide which existing ontologies to re-use. However, evaluation techniques are not useful only during the engineering process of the ontology: they are also useful to an end-user who is looking for an ontology that is suitable for her application domain. Thus, although there is a clear need for methods for evaluating and comparing ontologies, it is most probable that an evaluation method will not be suitable for all the tasks. This paper proposes a method for the evaluation of ontology learning methods. This method compares learned ontologies with gold standard ontologies by means of ontology alignment techniques.

Regarding the evaluation of ontology learning methods, four major approaches are often adopted: (a) those comparing the learned ontology to a predefined gold standard ontology, which is usually hand-crafted by domain experts, (b) those embedding the learned ontology in a complete system and evaluating the performance of the system [1], (c) those relying on a data-driven evaluation by assessing the ontology on existing data from the domain of the ontology ([2], [3]), and (d) those in which the evaluation is performed purely by human experts ([4], [5], [6]). Many approaches fall into the first category, i.e. evaluation using a gold standard ontology (e.g. [7], [8], [9], [10]).

Methods that rely on gold standard ontologies have strong points, as well as weaknesses. On the one hand, they support the evaluation of the learned ontology at several levels, such as the lexical and the relational one. In addition, they support the automated evaluation of learned ontologies using relevant tools and standard metrics from the field of information retrieval, since there exists the "ground truth" for comparison. Finally, through a gold standard-based evaluation method, qualitative, as well as quantitative results may be derived. On the other hand, this type of ontology evaluation assumes that the gold ontology represents well and accurately the significant knowledge of the domain. This assumption may be faulty in many cases, since the gold standard is usually created by human experts. Therefore, it may be incomplete or developed in a biased way. Finally, to a large extent, gold standard evaluation depends heavily

- *E. Zavitsanos is with the Institute of Informatics and Telecommunications, NCSR "Demokritos", Athens, Greece, 15310 and with the Department of Information and Communication Systems Engineering, University of the Aegean, Samos, Greece, 83200. E-mail: izavits@iit.demokritos.gr*
- *G. Paliouras is with the Institute of Informatics and Telecommunications, NCSR "Demokritos", Athens, Greece, 15310. Email: paliourg@iit.demokritos.gr*
- *G.A. Vouros is with the Department of Information and Communication Systems Engineering, University of the Aegean, Samos, Greece, 83200. Email: georgev@aegean.gr*

on the matching[1] between the learned and the gold ontology elements, as well as on the similarity measures that are used to compare the two ontologies. This is still an open research issue.

In this paper, a new method for the automated evaluation of learned ontologies against gold standards is proposed, avoiding common pitfalls of ontology comparison methods that apply string matching techniques on concept names. For instance, let us assume the ontologies in Figure 1. Using a string-matching technique, such as the edit-distance, the concept *RNA* may be matched with *DNA*, *RNA_mol* with *DNA_mol*, *RNA_domain* with *DNA_domain* and *Nucleid_acid* with *Nucleid_acid*, unless a domain-specific matching method is used. Obviously, the overall performance of this matching would be poor, since the majority of the concepts that are matched have completely different meaning and thus instances. On the other hand, comparing similar or even identical concepts that are lexicalized with very different terms, such as "car" and "automobile", would possibly never lead to a match, unless a lexicon of synonyms is used.

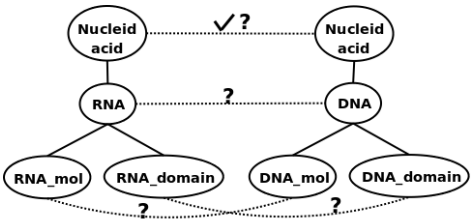

Fig. 1. Example of possible wrong matches between a gold ontology (left) and a learned ontology (right). The matching between the concepts named "Nucleid_acid" is correct, but the rest of the matchings are not.

In contrast to this superficial string matching of concepts, the proposed method transforms the concepts and properties of the gold standard and the learned ontology into probability distributions over the term[2] space of the dataset from which the ontology has been learned. In this way, the superficial string matching is avoided, since probability distributions are compared, rather than strings. Additionally, it becomes possible to choose among a variety of measures in order to assess the similarity of two concepts or two properties. Another major advantage is that the learning method is not required to label the identified concepts, since the concept name is either not used, or it simply participates as one term in the distributional representation.

Besides the simple and generic transformation method of concepts and properties into probability distributions, this article contributes the following: (a) a novel set of evaluation measures for automatically assessing the quality of the learned ontology, taking into account the degree of similarity between the elements of the two ontologies, as well as their positions in the ontologies, (b) an extension of our recently proposed method [11], in order to evaluate fully-fledged ontologies, rather than simple hierarchies, and deal with cases where the dataset is not available, (c) a flexible mechanism that allows different methods to be used for the alignment of the learned with the gold ontology, as well as different similarity measures to be used for the comparison of ontology elements, (d) a thorough investigation of the importance of computing matches between the gold and the learned ontology elements with precision in ontology evaluation, through an extensive evaluation, and (e) a novel evaluation methodology for assessing evaluation methods, such as the one proposed in this article.

In the remaining sections, we start by studying related work concerning the gold standard evaluation of ontologies (Section 2). In Section 3, the proposed evaluation method and the evaluation measures are presented. Finally, Section 4 presents extensive experimental results and discussion regarding the behavior of the proposed method using different similarity measures, while Section 5 summarizes the main contributions of the paper and presents future directions.

2 Related Work

In a gold-standard evaluation of ontology learning methods, it is assumed that the gold ontology is the "correct" answer to the specific ontology learning problem. In general, when this evaluation approach is selected, the strategy is as follows. First, domain experts manually construct the domain ontology that serves as the gold standard. Then, ontology pruning is performed. Ontology pruning is the process of hiding some components of the gold ontology, in order to assess the degree to which the learning mechanism is able to reconstruct the hidden components. Then, the ontology produced by the learning method is compared to the gold standard.

The comparison between the learned and the gold ontology can be done at various layers of the ontology elements. Based on the lexicalization of the concepts, one can compare them using the edit distance [12] and obtain an ontology matching at the lexical level (e.g. [10], [13], [14]). Quantitative results using such methods can be derived by the measures of Term/Lexical Precision and Recall, which have been introduced in [15]. In [17] and [18], where the learned ontologies are compared to known semantic nets, such as WordNet, Precision and the percentage of the correctly matched ontology elements are measured respectively.

Regarding the evaluation of concept hierarchies, the work in [16] evaluates learned taxonomies using the measures of Precision and Recall, assuming that the correct subsumption relations are those between the correctly matched concepts. However, the position of the concepts that are matched is not taken into account in this evaluation process.

1. The terms "matching" and "match" refer to equivalence mappings between ontology elements.

2. "Terms" does not necessarily denote domain terms, but words that constitute the vocabulary over which concepts are specified. In the following, "terms" and "words" are used interchangeably.

On the other hand, Augmented Precision and Recall [9] evaluate and penalize the learned hierarchy according to the position of the concepts in the hierarchy by measuring their distances from the root concept and their most common abstraction. Similar ideas have been used in [13], where the measure of Taxonomic Similarity is introduced, based on the length of the shortest path between the matched concepts in the concept hierarchies. However, although such measures focus on penalizing relational differences, they do not take into consideration the degree to which the concepts of the learned ontology differ from the concepts of the gold standard.

The position of the concepts in the hierarchy and the concepts in their vicinity is of paramount importance, regarding the taxonomic evaluation of ontologies. The method in [10] introduces 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. Techniques like [8] 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 and penalize the learned ontology according to the position of the learned concepts in the hierarchy. Thus, they are limited with respect to the alignment of the ontologies, as they require the concepts in the gold standard to match exactly the concepts of the learned ontology.

Another perspective of viewing the hierarchy is that of a partitioning of a set of instances. Based on this idea, the OntoRand index [7] measures the similarity between gold and learned concepts, based either on their common ancestors, or on their distances in the hierarchy, taking also into account the overlap of their instance sets. Although this method treats concepts as clusters of instances, going beyond their lexical representation, it requires that both hierarchies contain exactly the same set of instances, which limits the applicability of the method in the case of having a learned ontology without instances.

According to the criteria for good evaluation measures presented in [8], our aim is to evaluate two ontologies by measuring their similarity, avoiding common problems introduced by matching only concept lexicalizations. This is particularly important for many ontology learning methods, which are unable to label the identified concepts. Moreover, in contrast to most of the related efforts towards the evaluation of ontology learning methods, which focus on the evaluation of concept hierarchies (e.g. our method in [11]), the proposed method is suitable for evaluating concept hierarchies, as well as ontologies enriched with other semantic relations and properties.

3 The DMA Method

The proposed method, called Distributional Method for Alignment (DMA), assumes three main steps for evaluating the learned ontology against the gold standard: (a) transformation, (b) matching, and (c) evaluation. The first step is required in order to transform the ontology elements into probability distributions over terms. The matching step matches the learned ontology elements to the gold ones. The evaluation is based on this set of matches. The final step is the actual evaluation that penalizes the learned ontology according to its deviation from the gold ontology.

Regarding the transformation of the ontology elements, two cases are foreseen. In the first case, the dataset that was used for ontology learning is available, while in the second it is not. Subsection 3.2 presents the ontology matching process. Subsection 3.3 introduces the similarity measures and discusses their properties. Finally, the last subsection presents how the overall method works by emphasizing on its generality, its ability to be independent of the dataset, its flexibility regarding the choice of different matching methods for the matching step, as well as its ability to choose different metrics for measuring the dissimilarity between the matched ontology elements.

3.1 Ontology Transformation

Towards the objective of representing each ontology element as a probability distribution over terms, two cases are foreseen: (a) creating the term space from the dataset used to learn the ontology, and (b) creating the term space from the terms occurring in the concept specifications, in case the dataset is not available.

A point to be made here concerns the case of having the learned and the gold ontologies alone, without the dataset. Since the majority of ontology learning methods rely on a training dataset of text documents, in the majority of the cases, such a dataset will be available. However, in order to illustrate the applicability of the method in cases where the data are not available, the handling of this case is presented here. This allows also to evaluate the method on the very valuable benchmark of the Ontology Alignment Evaluation Initiative (OAEI).

Regarding the first case, assuming that the concept instances are annotated in the text documents, he frequencies of the terms that appear in the *context* of each concept instance are measured. The context of the concept instance in this case is the surrounding text or the complete document, in which the instance appears. By concept instances we mean instantiations of concepts that appear in the text. For example, the word “Hania” is annotated as an instance of the concept “City”, and thus, the context of this occurrence of “Hania” provides terms for the representation of the concept “City”.

As Figure 2 illustrates, having the instances annotated in the corpus, it is possible to associate each document to the concept(s) that it refers to. In cases where the concept instances are themselves documents, e.g. in a document indexing task, the mapping between concepts and documents is directly provided and the population process of Figure 2 can be skipped. The distributional

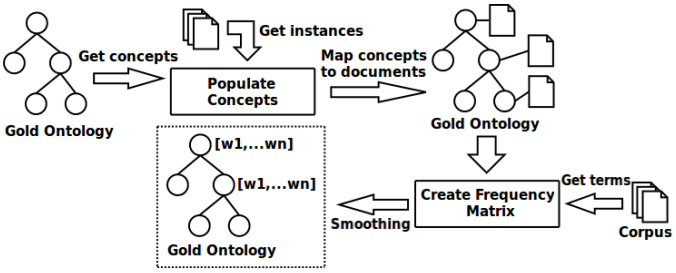

Fig. 2. The transformation of the ontology elements into probability distributions.

representation of a concept records the frequency of each term in the context of all instances of that concept in the dataset.

At the end of the process, for each concept, the frequencies are normalized to obtain a probability distribution over the term space of the dataset. At this final step, there is the option of performing Laplace smoothing (Equation (1)) of the probability distributions to eliminate possible zero values of unseen terms. Algorithm 1 describes the way the distribution vectors are created.

$$\hat{P}_L(w_i) \doteq \frac{\hat{P}(w_i)+1}{N+1}, \forall i, \quad (w_i: \ word, \ N: \ term \ space \ size). \tag{1}$$

Both ontologies, i.e. the learned, and the gold standard one, are transformed to a common representation following Algorithm 1. The representation used for concept properties is analogous to that of concepts, as long as the context of the properties can be located. In this case, the context comprises the document(s) where the values, labels, domain and range of the properties appear. In particular, the context of the domain and range of a property, which usually correspond to concepts, is the document(s) where instances of those concepts appear.

The transformation step relies heavily on the context of ontology elements and provides useful information for the matching task. It must be noted that, via this context-oriented ontology transformation process, ambiguity and polysemy are addressed to some extent, since a term may be annotated as an instance of different concepts. The context in which this instance appears provides useful information for deciding the exact concept that it instantiates. In addition, this transformation is particularly suitable for comparing topic ontologies [16], where concepts are already represented as multinomial distributions over terms, as well as in cases where the learning method is not able to provide labels for the learned concepts.

According to Algorithm 1, when concepts are mapped to documents (Create DocumentConceptMatrix block), the *degree of participation* of a concept c_j in a document d_i is calculated as the number of instances of concept c_j in document d_i. Thus, different concepts participate with a different degree in the same context (document

Data: Documents and concept instances
Result: Distributional representation of concepts
$DTArray[][]$=createDocumentTermMatrixOfFrequencies()
// **Create DocumentConceptMatrix**
$DCArray[][]$=null
for *all documents* $d_i \in 1,..,D$ **do**
  for *all concepts* $c_j \in 1,..C$ **do**
    for *all instances of* c_j, i_k **do**
      if i_k *exists in* d_i **then**
        DCArray[d_i][c_j]++
      end
    end
  end
end
// **Create ConceptTermMatrix**
$CTArray[][]$=null
for *all concepts* $c_i \in 1,..,C$ **do**
  for *all terms* $t_j \in 1,..,T$ **do**
    for *all documents* $d_k \in 1,..,D$ **do**
      $percent =$
      $DCArray[d_k][c_i] * DTArray[d_k][t_j]$
    end
    $CTArray[c_i][t_j] = percent$
  end
end

Algorithm 1: Transformation of concepts to distributional representations.

d_i), leading to different distributional representations.

In the extreme case, where two concepts have the same number of instances, identically distributed in the same context, then they share the same distributional representation. Let us consider the following document:

> *Although the island is formally divided into four prefectures (Hania, Rethimnon, Heraklion and Lassithi), it is more readily divided into east, west and central Crete.*

Two concepts appear in this context. "City" and "Island". However, "City" participates "more" than "Island", since four instances of "City" appear in that context (Hania, Rethimnon, Heraklion, Lassithi), while only one instance of "Island" appears in the same context (Crete). Thus, the distributional representations of "City" and "Island", will be different. Figure 3 presents the two vectors.

	island	Hania	Rethimnon	Heraklion	Lassithi	Crete	devided	east	west	central	...
City	4	4	4	4	4	4	8	4	4	4	...
Island	1	1	1	1	1	1	2	1	1	1	...

Fig. 3. Distributional representation of concepts "City" and "Island" before normalization. Concept "City" more than "Island" in the same context, since fourfold instances of the former appear in that context.

When comparing an ontology against a gold standard one, without knowing the dataset from which this ontology has been constructed/learned, the distributional

representation is obtained, using vocabulary terms. As Figure 4 shows, a common term space for the gold and the learned ontology is created, by extracting terms that appear in the labels, comments and descriptions of concepts and properties, as well as in concept instances, domain and range of properties. For a specific property, the terms that appear in its label, comments and so forth, define the context for that property.

Following a similar procedure as the one described in Figure 2, the distributional representation of each ontology element is created, based on its context.

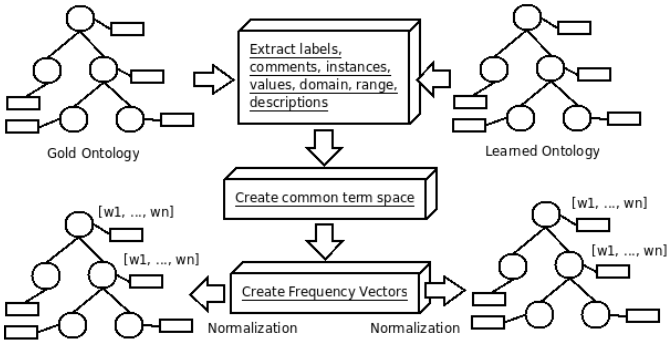

Fig. 4. The representation of ontology elements in the case where an annotated dataset is not available.

The transformation method proposed here, bares some similarities with Formal Concept Analysis [19], where a concept is defined as the set of its attributes, and for the whole ontology a formal context is defined, including the set of concepts and their attributes. However, the method proposed here aims to go one step further by defining concepts as multinomial probability distributions over the term space of the dataset or the ontologies. This way, a variety of probability distribution measures and metrics can be used to decide how "close" two concepts (or properties) are.

3.2 Ontology Matching

Using the vector representation of ontology elements, it is possible to measure the dissimilarity between the concepts and properties of the gold and the learned ontology. Since both representations are based on probability distributions, an appropriate probability metric can be used to measure how "close" two concepts or two properties are. In this paper, three probability metrics are used in order to measure the dissimilarity between concepts or properties: (a) the *Total Variational Distance* (TVD), (b) the *Kolmogorov Distance* (d_K), and (c) the *Separation Distance* (S).

In order to measure the dissimilarity ($SimDist$) of two probability distributions $p(\cdot)$ and $q(\cdot)$ over a countable state space Ω, such as the term space of the corpus here, the TVD is defined according to Equation (2).

$$TVD = \frac{1}{2}\sum_i \mid p(i) - q(i) \mid . \quad (2)$$

In Equation (2), $p(\cdot)$ and $q(\cdot)$ represent ontology elements in the learned and gold ontologies respectively. TVD is one of the most commonly used probability metrics, because it admits natural interpretations, as well as useful bounding techniques. Given two distributions as input, TVD measures their average distance. For individual terms, TVD measures the largest possible difference between the probabilities that the two distributions assign to the same term ($p(i), q(i)$).

For the same countable space Ω, the Kolmogorov Distance and the Separation Distance are given by Equations (3) and (4) respectively. However, the Kolmogorov distance can be also defined over a state space $\Omega = \Re$, while the Separation distance is not a metric, due to symmetry reasons. Finally, one may choose any other probability measure (e.g. see [20]) to measure similarity.

$$d_K = max_i \mid p(i) - q(i) \mid . \quad (3)$$

$$S = max_i(1 - \frac{p(i)}{q(i)}). \quad (4)$$

DMA determines a one-to-one matching between the gold and the learned concepts using the TVD metric, or any other metric. Therefore, the set of matching pairs, includes as many pairs as the number of concepts in the smaller of the two ontologies. Respectively, the matching of properties includes as many pairs as the number of properties of the ontology that has the fewest properties.

Assuming any of the above probability measures of dissimilarity, among the possible matches, the best set of matches is determined, by minimizing the aggregate $SimDist$. According to Equation (5), among all the possible matches N (of concepts and properties), the one that minimizes the sum of $SimDist$ over all matching pairs M is chosen.

$$argmin_N\{\sum_i^M SimDist_i\}. \quad (5)$$

It should be stressed that the learned ontology is evaluated over the complete set of matches, i.e. the "quality" of each individual match affects the estimated deviation of the learned ontology from the gold standard. The impact of each deviation on the evaluation measure depends on the probability measure that is chosen and the way these measures are aggregated. For instance, d_K may be more strict than S in some situations. Irrespective however, of the probability measure, one difference in the distributional representation of a concept (i.e. the value of one term) suffices to distinguish it from another concept.

Instead of one-to-one matching, one could choose to perform (a) one-to-many matches, by matching a single ontology element in the gold ontology to many elements in the learned ontology, (b) many-to-one matching, by finding many elements in the gold ontology that probably match to a single learned ontology element, or

(c) many-to-many matching. However, when a gold-standard evaluation is performed, the assumption is made that the gold ontology is the best among all the possible ontologies in a domain, given a particular source of information. Thus, a one-to-one matching is the most appropriate choice, as it imposes a more strict evaluation of the learned ontologies. Following this approach, each ontology element is judged separately and an overall score for the learned ontology is provided.

Regarding alternative matching methods, the ASMOV [24] and Lily [23] approaches are included in the experiments. These methods have been chosen due to their superior performance in the OAEI 2008 matching contest.

ASMOV uses an algorithm that automates the ontology alignment process, optionally incorporating feedback from a user. It uses a weighted average of measurements of similarity along four different features (lexical description, external structure, internal structure, and individual similarity) of the ontologies, and performs semantic validation of the resulting alignments.

Lily, on the other hand, uses a hybrid strategy to perform the alignment between two ontologies. It comprises four main components that are responsible for: (a) the generic alignment between ontologies, (b) the large-scale ontology alignment, (c) the semantic matching between ontology elements and (d) alignment debugging.

Both Lily and ASMOV provide a confidence degree that can be considered to be a degree of similarity for each matching pair of elements between the two ontologies. This is important for using them as alternatives to DMA in the proposed evaluation method. In these cases, $SimDist$ is equal to $(1-\gamma_x)$, where γ_x is the confidence degree provided by a matching method X. For instance, for Lily, the $SimDist$ for each pair of ontology elements is $(1-\gamma_{Lily})$. Subsequently, we refer to the dissimilarity $SimDist$ provided by a matching method, rather than the similarity/confidence degree γ provided by it for each matching pair of elements.

3.3 Ontology Evaluation Measures

Moving from matched concepts and properties to an overall evaluation measure is a non-trivial issue. According to [8], a measure must evaluate an ontology along multiple dimensions (e.g. lexical and relational levels). Furthermore, in ontology learning, an error, or deviation of the learned ontology from the gold one, must cause a change to the measure proportional to the dissimilarity 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 should lead to a gradual decrease in the value of the evaluation measure.

In this work, a set of similarity measures is also proposed, which is presented in Equations (6), (7) and (8). Details about these measures are presented in the paragraphs that follow.

$$P = \frac{1}{M}\sum_{i=1}^{M}(1-SimDist_i)PCP_i. \quad (6)$$

$$R = \frac{1}{M}\sum_{i=1}^{M}(1-SimDist_i)PCR_i. \quad (7)$$

$$F = \frac{(\beta^2+1)P*R}{\beta^2 R+P}. \quad (8)$$

In the above equations, M is the number of matching pairs between the learned and gold ontology, while $SimDist$ is a measure of dissimilarity between the matched concepts and properties ranging in $[0,1]$.

It is already stressed the importance of the correct matching as the basis for the evaluation of the learned ontology. The matching of concepts and properties of the two ontologies can be performed with any ontology alignment method ([21], [22]). In the case where an ontology alignment method is used, the *SimDist* factor can be replaced by the degrees of matching that this method provides, such as the ones discussed in subsection 3.2.

For the evaluation of the structure of the learned ontology (relational level), the degree of the deviation from the gold ontology should also depend on the position at which this deviation occurred in the learned ontology. For instance, missing a leaf concept that participates in a single subsumption relation and has no properties should cause a smaller penalty than missing a central concept that is subsumed by some concepts and has a number of children and properties. The *PCP* and *PCR* factors in Equations (6) and (7) measure the impact of an error to the relational structure of the ontology.

PCP and *PCR* stand for *Probabilistic Cotopy Precision* and *Probabilistic Cotopy Recall* respectively. They are both influenced by the notion of Semantic Cotopy [10]. For a matching i between a concept C_L in the learned ontology and a concept C_G in the gold ontology, PCP_i and PCR_i are defined based on the *Cotopy Set* of the concepts.

Definition 1 (Cotopy Set): The cotopy set of a concept C ($CS(C)$) is the set of all its direct and indirect super and sub-concepts and its direct properties, including the concept C itself. For instance, Figure 5 illustrates the cotopy set of concept RNA, which includes the shaded concepts (depicted as circles), as well as its direct property (depicted as a rectangle).

Definition 2 (PCP): 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 the number of concepts in the cotopy set of C_L (Equation (9)).

$$PCP_i = \frac{\|CS(C_L)\cap CS(C_G)\|}{\|CS(C_L)\|}. \quad (9)$$

Definition 3 (PCR): 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 the number of concepts participating in the cotopy set of C_G (Equation (10)).

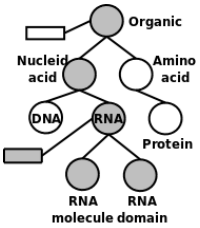

Fig. 5. The cotopy set of concept RNA comprises all the direct and indirect super and sub-concepts of RNA, its direct properties, as well as RNA itself (all the shaded elements).

$$PCR_i = \frac{\|CS(C_L) \cap CS(C_G)\|}{\|CS(C_G)\|}. \quad (10)$$

For instance, in Figure 6, depicting a specific match between the concepts named *RNA*, the PCP_{RNA} is equal to $\frac{3}{4}$, while the PCR_{RNA} is equal to $\frac{3}{5}$.

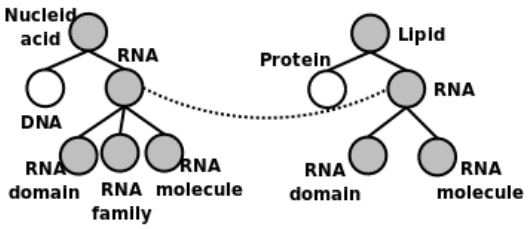

Fig. 6. An example regarding a specific match.

Therefore, the P measure (Equation (6)) reflects the similarity of the two ontologies in the spirit of Precision, penalizing learned elements that do not appear in the gold standard ontology. On the other hand, the R measure (Equation 7), similar to Recall, penalizes the learned ontology in cases where it does not include elements that appear in the gold ontology. F is a combined measure of P and R (Equation 8). The mismatches between properties of the two ontologies penalize the learned ontology through the PCP and PCR factors, since concept properties participate in the cotopy sets of the concepts.

Finally, it should be noted that the gold ontology, being hand-crafted by human experts, may be biased, incomplete, or inaccurate. Therefore, one may be interested more in the precision of the learning method, or in recall. Thus, one could adjust the F measure of Equation (8) to focus more on the impact of P or R, by adjusting the parameter β. The gold ontologies used in section 4.1 came with the datasets. Although they are hand-crafted by humans, the assumption is made that they are accurate conceptualizations of the available data we study and thus, in our evaluation settings we choose $\beta = 1$, reflecting the harmonic mean of P and R.

3.4 Overview of the Evaluation Method

Figure 7 summarizes the proposed evaluation method, consisting of an inventory of matching methods, including the DMA method that was proposed in subsection 3.2, an inventory of dissimilarity measures, including TVD, d_K and S, presented in 3.2, and the evaluation measures proposed in subsection 3.3.

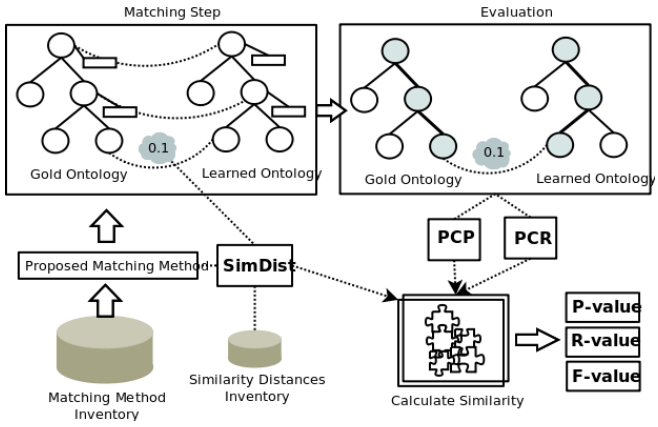

Fig. 7. The proposed evaluation method. An inventory of matching methods and similarity measures can be used to perform the matching between the learned and the gold ontology. For each matching pair, the PCP and PCR factors are calculated, which along with the *SimDist* factor provide the final evaluation results in terms of P and R.

The proposed evaluation method can be used with a variety of methods from the field of Ontology Alignment, as long as they provide a "degree" of similarity for the matching pairs between the elements of the learned and the gold ontology. If, on the other hand, the DMA matching method is chosen, one may choose any dissimilarity measure from the corresponding inventory, such as the TVD. Once a set of matches has been established, the evaluation of the learned ontology is performed, taking into account the dissimilarity ($SimDist$) of each matching pair, and (b) PCP and PCR for each matching pair.

As an example, Figure 7 depicts the case where a leaf concept from the learned ontology matches to a concept in the gold ontology with $SimDist$ equal to 0.1. For this specific matching pair, the cotopy sets are determined, which include the shaded concepts in Figure 7. The final step is the calculation of PCP and PCR for this matching. Following the same procedure for all matching pairs the P, R and F measures are calculated according to Equations (6), (7) and (8).

4 Experimental Evaluation

In this section, the DMA is assessed through a set of experiments, examining the robustness of the evaluation measures when "errors" are introduced to the gold standard ontology: Therefore, different versions of the gold standard ontology are assumed to be the "ontologies learned", and are compared to the gold standard.

Two sets of tests have been performed, presented in the following subsections. First, two real datasets are used, comprising documents that contain annotated instances of gold ontology concepts. In this case, we

study the behavior of DMA and the robustness of the evaluation measures when introducing errors in the corresponding gold standard ontologies. In the second subsection, the set of ontologies provided by the Ontology Alignment Evaluation Initiative (OAEI) [21] contest are used, comprising a reference (gold) ontology and variants of it (considered to be learned ontologies). In this case, we observe the behavior of the measures as the variants of the reference ontology deviate in various ways from it.

The aim of these experiments is to assess how small or large deviations from the gold ontology affect the F value. We also observe the impact of different types of error on the evaluation measures. Finally, the behavior of the method is studied, when the $SimDist$ metric is calculated using the Kolmogorov or the Separation Distance, besides the TVD, or when DMA is substituted by highly accurate ontology matching methods.

4.1 Experimental Assessment with Real Datasets

In this task, two gold ontologies are used with their corresponding datasets: the Genia[3], comprising 43 concepts from the domain of molecular biology, and the Lonely Planet[4] ontology, comprising 60 concepts from the tourism domain. As mentioned above, the "learned" ontologies are generated, by introducing errors (deviations) to the gold ones.

The intuition behind this approach is that a learned ontology will approximate the gold standard but will not be identical to it. This means that it may be incomplete, by missing some concepts, or it may have more concepts than the gold one. In addition, it may comprise more or fewer relations between concepts, or it may contain reversed taxonomic relations, i.e. a concept A may subsume a concept B, while the opposite holds for the gold standard. On this basis, we define six elementary "damage" operators that capture the majority of the errors in a learned ontology.

This approach is particularly suitable for the evaluation of evaluation methods. It provides a controlled experimental setting, avoiding the bias introduced by the use of an ontology learning method. In this context, the following "damage" operators are used: (1) Swap Concepts, (2) Remove Concepts, (3) Add Subconcepts, (4) Add Superconcepts, (5) Add Taxonomic Relations, and (6) Change Concept Representation.

Each of these operators takes as input a number that indicates the extent of the damage to the gold ontology. This degree of "damage" has the following effect on each of the "damage" operators: *Swap Concepts*:

- *Swap Concepts*: the number of subsumption relations to be swapped.
- *Remove Concepts*: the number of concepts to be removed.
- *Add Subconcepts*: the number of concepts to be added as subsumees of existing concepts.
- *Add Superconcepts*: the number of concepts to be added as subsumers of existing concepts.
- *Add Taxonomic Relations*: the number of relations to be added among existing concepts.
- *Change Concept Representation*: the number of concepts, whose representation as distributions of terms will be changed.

With the exception of the last operator (Change Concept Representation), the results are obtained using only the TVD as the *SimDist* measure. This is because *SimDist* is equal to zero for all pairs of matching concepts in the first five cases, irrespective of the measure that is used (TVD, d_K, S). In the last "damage" operator, where the representation of the concepts changes, the three different *SimDist* measures are compared.

More complicated operators can be built on the basis of the proposed elementary ones. For instance, a "Merge Concepts" or a "Split Concepts" operator would consist of a sequence of "Remove Concepts" or "Add Concepts" operations. The impact of such composite operators would be an aggregate of the impact observed for the elementary operations. Thus, the elementary operators provide a finer evaluation of the methods.

For each ontology and for each "damage" operator, 50 different tests are run for each of 10 different "damage" degrees, thus performing 500 tests per "damage" operator and a total of 3000 tests per ontology. In each test the similarity of the resulting ontology to the original one is measured.

In the following, average values over the 50 tests are presented for each damage degree of each "damage" operator. Figures 8 and 9 present the results that were obtained for the Genia and the Lonely Planet ontologies respectively, while Figures 11 and 12 provide a different view of the results, according to the level of the subsumption hierarchy at which the "damage" was done. The results in Figures 11 and 12 focus on a specific "damage" degree for each "damage" operator. The rest of this section summarizes the main observations per "damage" operator.

Swap Concepts

This operator picks randomly a predefined number of concept pairs and swaps them, introducing in this way invalid subsumption relations to the ontologies. The number of concepts, as well as the "nature" of the concepts, i.e. their representations as distributions of terms, remain the same. Thus, only the cotopy sets of the concepts are affected, resulting in different PCP and PCR values.

Swapping a single pair of concepts leads to a small taxonomic difference (Figs. 8 and 9), especially when this operation is performed on the leaf concepts of the hierarchy (Figs. 11 and 12). For the two reference ontologies, the 50 different experiments of swapping one pair of concepts result to an F between 0.81 and 0.99 (Figs. 11

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

4. The Lonely Planet travel advise and information, http://www.lonelyplanet.com

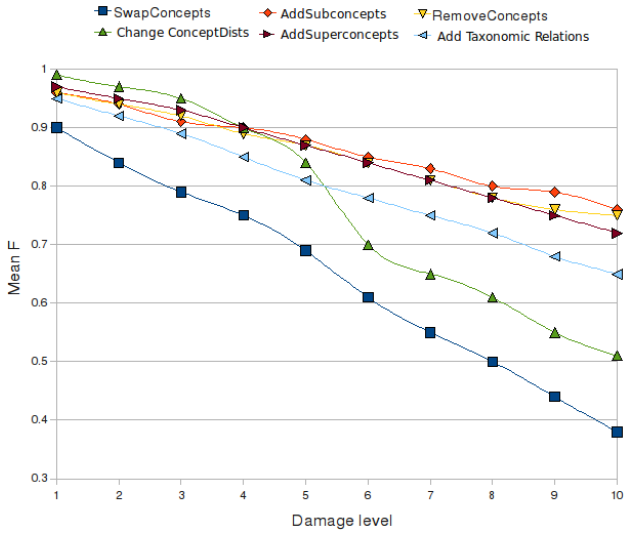

Fig. 8. Combined diagram for all "damage" operators in the case of the Genia ontology.

and 12), depending on the position of the concepts in the hierarchy. As the "damage" increases, i.e. as the number of concepts that are swapped gradually increases, the F measure decreases almost linearly (Figs. 8 and 9), reaching a situation where 10 pairs of concepts are swapped, changing more than half of the subsumption relations of the ontologies.

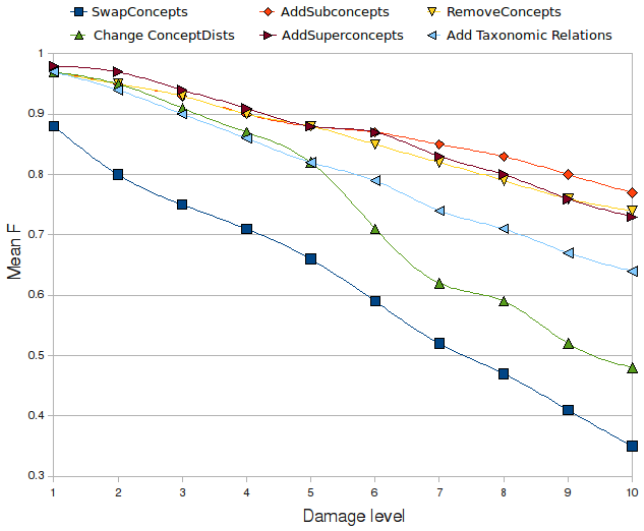

Fig. 9. Combined diagram for all "damage" operators in the case of the Lonely Planet ontology.

Swapping concepts affects heavily the evaluation results because of the significant change of the cotopy sets. A swap between two concepts may affect their cotopy sets, as well as the cotopy sets of other concepts, by introducing or removing a large number of neighboring concepts. This effect depends on the new position of the swapped concepts. Figure 10 illustrates such an example. Before the swap operator, the cotopy set of concept *Nucleid_acid* is the whole ontology (the shaded concepts), while the cotopy set of concept *DNA* comprises *DNA* itself and *Nucleid_acid*. Recall that the cotopy set of a concept includes all its super and subconcepts. When these concepts are swapped, the cotopy sets of *Nucleid_acid* and *DNA* change significantly. Now, the cotopy set of *DNA* is the whole ontology, while the cotopy set of *Nucled_acid* consists only of itself and *DNA*.

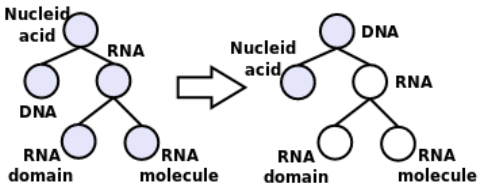

Fig. 10. Example of swapping two concepts that participate in a subsumption relation, showing the effect of the Swap Operator on their cotopy sets.

Remove Concepts

In this case, a predefined number of randomly chosen concepts is removed from the ontologies. The remaining concepts stay intact, i.e. their representation is not altered. Again, only the cotopy sets in which the removed concepts participate change, influencing the PCP and PCR factors.

The errors that this operator introduces affect the hierarchical structure of the ontologies, as some concepts of the gold standard disappear. Removing only one concept from the hierarchy leads - quantitatively - to a small relational difference. Especially, if this is a leaf concept, the penalty is small. For the two ontologies, the 50 different tests of removing a single concept result in F between 0.90 and 0.99 (Figs. 11 and 12), depending again on the position of the concept, i.e. whether it is a leaf or a "central" concept.

As shown in Figures 8 and 9, the decrease of the F measure is linearly related to the extent of the damage and it is milder than for concept swapping. This is because the removal of a concept A affects only the cotopy set of the concepts in the vicinity of A by decreasing their size by one. Thus, the PCP and PCR factors are affected, but not substantially.

Add Subconcepts

This operator adds a predefined number of new concepts randomly to the ontologies, as children of existing concepts, maintaining the tree-like structure of the hierarchies. Since the added concepts are assigned random distributions of terms, it is almost impossible to be matched to concepts in the gold ontology. In addition, the fact that the rest of the concepts remain intact, results in a perfect match between them. Thus, the P value is only affected, while R remains equal to 1.

The behavior of this operator is similar to the one that removes concepts. In this case, the cotopy set of some concepts is affected by increasing their size by one. The

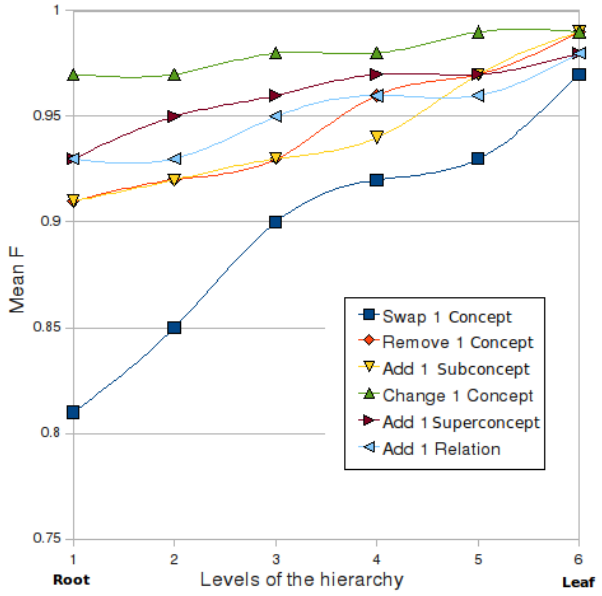

Fig. 11. Combined diagram for all "damage" operators in the case of the Genia ontology.

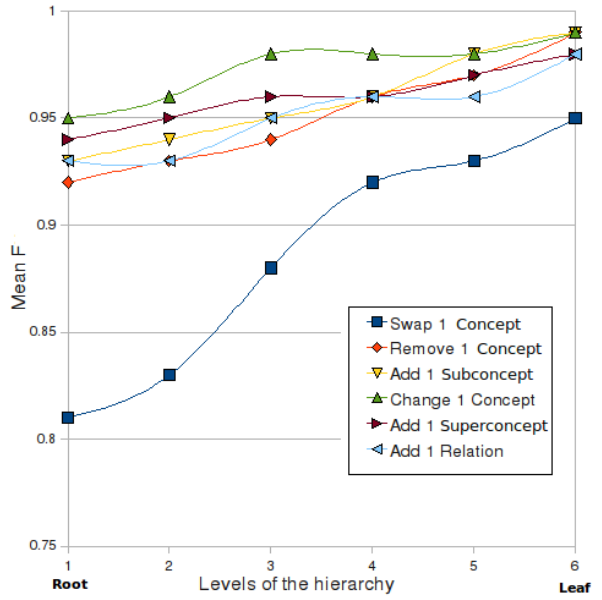

Fig. 12. Combined diagram for all "damage" operators in the case of the Lonely Planet ontology.

concepts that are affected through their cotopy sets are those which are parents and children (direct or indirect) of the newly added concept. As in the case of removing concepts, the effect on the PCP and PCR factors is not substantial.

Adding a single concept to the ontology introduces a small error to the hierarchical structure. For the 50 different tests of adding only one concept, the F measure is between 0.91 and 0.99 (Figs. 11 and 12). As the number of added concepts increases, the F measure is affected similarly to concept removal.

Add Superconcepts

This operator introduces new taxonomic relations in the ontology by adding new concepts as parents to randomly chosen existing concepts. This process is similar to that of adding subconcepts. However, in this case, the error introduced has an impact on the relational layer of the ontologies, in the sense that multiple inheritance between concepts may be introduced. This means that one concept may now be subsumed by more than one concept. Thus, P is affected, while R remains equal to 1. Again, the PCP and PCR factors are affected, resulting in the F values of Figures 8 and 9.

Figures 11 and 12 show that the addition of a superconcept has a similar effect to that of adding a subconcept, which is expected, as the cotopy sets of the concepts in the vicinity of the newly added concept increase by one.

One may argue at this point that when adding a concept A, either as a subsumee or a subsumer, there is a new cotopy set introduced, i.e. that of concept A. However, this does not affect the evaluation results, because matching is performed prior to the evaluation. The resulting set of matches comprises as many matches as the number of concepts in the smaller ontology. As already said, the new concept A is not matched to any of the gold concepts. Thus, its cotopy set does not affect the evaluation results. The fact that A has been introduced in the learned ontology is reflected only through the PCP factor for the concepts in the vicinity of A, the cotopy sets of which are affected. Figure 13 illustrates an example of adding a concept to the learned ontology.

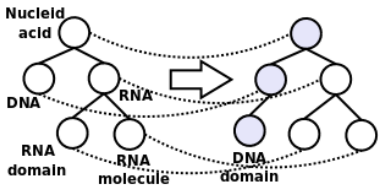

Fig. 13. Example of adding the concept *DNA_domain* in the ontology. This concept is not matched to any of the gold ones. Thus, PCP and PCR are not calculated for this concept. Its presence affects the PCP factor of the root concept and its left child, the cotopy sets of which increase by one.

Add Taxonomic Relations

This operator introduces new subsumption relations among the existing concepts of the ontologies. Like the Add Superconcepts/Subconcepts operator, the impact is on the relational layer of the ontologies. The number of concepts remains the same, as well as their representations, but the number of relations increases. Therefore, P is affected, while R remains equal to 1. The mean F value is depicted in Figures 8 and 9. As expected, the more relations added, the larger the impact on F.

Figures 11 and 12 illustrate that the impact of adding a new relation between existing concepts is a little higher than adding a new relation through the introduction of a new superconcept. Although at first this may seem counter-intuitive, it is due to the fact that the addition of a single new subsumption relation may change the hierarchy significantly, through the growth of the cotopy sets of the concepts that participate in this relation. Figure 14 provides such an example.

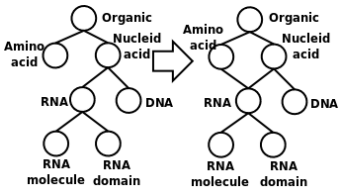

Fig. 14. Example of adding a relation between the concepts *Amino_acid* and *RNA*. The cotopy set of *Amino_acid* is now affected and includes almost the whole ontology. Furthermore, the cotopy sets of *RNA*, *RNA_molecule* and *RNA_domain* are increased by one (which is irrelevant, since it concerns proteins).

Another side effect of this operator is that the addition of new subsumption relations may introduce cycles to the resulting ontologies. In terms of the evaluation measures, cycles affect the PCP and PCR, since they change the cotopy sets of the concepts that participate in the cycle by increasing their size.

Change Concept Representation

In this last case, the number of concepts remains intact. The error is introduced in the distributional representation of randomly picked concepts. The changes affect the frequency of the terms that appear in the context of the concepts.

When changing a single randomly chosen concept, the F measure ranges between 0.97 and 0.99 (Figs. 11 and 12). However, changing the representation of more concepts, F decreases more steeply. It should be pointed out that this operator can lead to the extreme situation where a concept is changed completely, leading to $SimDist$ equal to 1. In reality, this can happen by (a) changing the probabilities of terms in the distribution, as this operator does, (b) changing the concept instances and thus changing the context of the concept and its representation, (c) removing the concept instances, and thus the context of the concept.

For this operator, an additional set of tests has been performed, using the Kolmogorov distance (Equation (3)) and the Separation distance (Equation (4)) measures. These additional tests were meaningful only for this operator, as it affects the representation of the concepts, affecting also the *SimDist* factor between pairs of concepts. Figure 15 depicts the behavior of the evaluation measures in terms of the F measure, for different "damage" degrees.

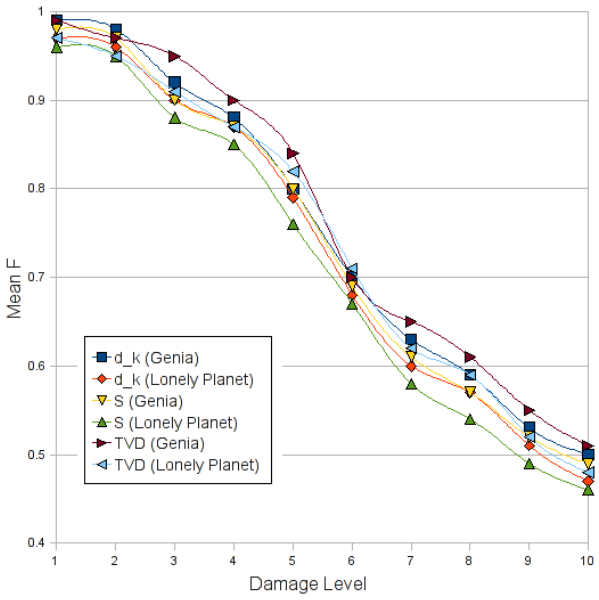

Fig. 15. Combined diagram for the Change Concept Representation operator.

The behavior of the three measures is very similar. DMA determined the same matches using the three different dissimilarity measures. In the general case, this behavior is expected when changing a single concept, since the rest of the concepts that remain intact would be matched correctly to the gold ones. When changing more than one concept, different similarity measures should provide different sets of matches. However, in the specific experiment this was not the case.

Summarizing the results of these experiments, we observe that the gradual increase of the "damage" leads to a gradual decrease in the F value in the closed interval $[0, 1]$. All the "damage" operators lead to a near-linear degradation of the F measure. Cases where the F measure does not exhibit a linear behavior are due to the fact that the learned ontology maintains some part in common to the gold one, even when the "damage" degree is high. If for instance all the subsumption relations are swapped, the F measure will still not be equal to zero, due to the remaining similarities between the two ontologies. The general observation is that different errors affect differently the evaluation measures, which are able to capture types of errors in the lexical and the relational layer of the ontologies.

The similar behavior of the method in the two different datasets is an initial indication of its invariance to the properties of the dataset and the ontology. Considering the slope of the F curves for the two ontologies, their position is also very similar. This can be attributed to the fact that the two ontologies have the same depth. Similarity in the depth of the two ontologies is important since PCP and PCR use the cotopy set of concepts, which depends on the number of parents and children. Assuming that no properties exist, as it is the case for

the ontologies used here, the cotopy sets comprise only super and subconcepts. Thus, the similarity of different ontologies tends to increase when the branching factor of the two hierarchies is also similar. However, this is not the case in the two ontologies used in our experiments, and thus, the evaluation results are not identical.

4.2 Experimental Assessment with the OAEI Ontologies

In this experiment, the benchmark series of the OAEI 2008[5] [21] is used. These consist of pairs of ontologies, comprising a reference ontology and various modifications of it. Specifically, in this experiment our aim is to study (a) the behavior of the evaluation measures when comparing the reference ontology to its variants, (b) the flexibility of the evaluation measures by substituting the matching measures with two methods from the field of ontology alignment[6], and (c) the effect that the choice of a matching method can have on ontology evaluation.

The benchmark test set that was used can be divided into five groups: $101-104$, $201-210$, $221-247$, $248-266$, and $301-304$. Each of these numbers indicates a pair of ontologies: the reference ontology and a variant of it (or in some cases an irrelevant ontology). For instance, the set 101 of the first group consists of the reference ontology compared to itself, the set 102 compares the reference ontology to an irrelevant one, and so forth.

For each of these five groups experimental results are provided in terms of F value using the method proposed in Section 3.2 (DMA), in conjunction with the TVD (DMA.TVD), the Kolmogorov (DMA.d_K), or the Separation distance (DMA.S) measures. Two state-of-the-art ontology matching methods are also included in the experiments: Lily [23] and ASMOV [24].

In this task, there are only pairs of ontologies without any text collection. Thus, as already stated, each pair comprises the reference ontology (the gold one), and its variant, which is assumed to be the "learned" ontology. In this case, the ontology elements are represented as multinomial probability distributions over a common term space created by the terms in the names, comments, instances, properties, domain, range, labels, values and descriptions of the ontology elements, as explained in subsection 3.1.

The first group (101-104) of pairs compares the reference ontology with (a) itself, (b) an irrelevant ontology, (c) a variant containing language generalizations in OWL Lite, where constraints and property types (such as the transitive property) are replaced with more general ones, and (d) a variant containing language restrictions in OWL Lite, where constraints have been discarded. Therefore, using this group, conclusions are mainly derived regarding the effect of modifications at the lexical level of the ontology. Figure 16 illustrates the behavior of the evaluation measures in terms of F value.

In the first case (101), all methods give (correctly) F values equal to one, since they provide perfect matches between the elements of the reference ontology and itself. Besides 102 (the irrelevant ontology) where all methods give (correctly) F values equal to zero, in the other two cases, all methods find the correct matches between the elements of the gold and the learned ontology. The highest F values are obtained using the d_K and S similarity measures, as they are more tolerant to language generalizations and restrictions. They give a smaller penalty to ontology elements that differ in their representation. The same holds for Lily and ASMOV that replace DMA. Between TVD, d_K and S, TVD is the most strict similarity measure, as it is more sensitive to changes in the distributional representation of the elements. This is due to the nature of d_K and S that focus on the maximum difference between the elements of two distributions (Equations 3 and 4), while TVD measures the average distance between two distributions.

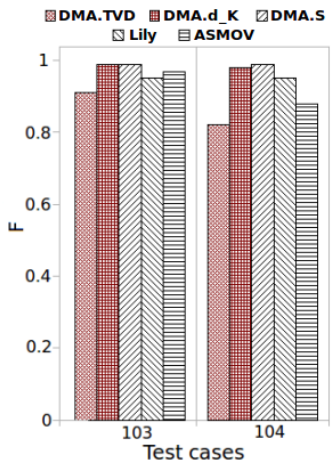

Fig. 16. Behavior of the evaluation measures in the test cases of the first group of the OAEI set. All methods determined the same set of matches.

In the second group ($201-210$), the reference ontology is compared with (a) variants without the names of ontology elements (201), (b) variants without names and comments or with misspelled comments ($202-203$), (c) variants with naming conventions, synonyms or translations of the ontology elements ($204-207$), and (d) variants combining naming conventions, synonyms and translations ($208-210$).

As Figure 17 shows, for the first case, where names are missing, DMA gives high values since it is less dependent on the names of the ontology elements, and relies more on the context of these elements. This is true for all three measures (TVD, d_K, and S). Therefore, the produced matches are correct and carry small penalties. Lily and ASMOV on the other hand, being dependent on names, impose a larger penalty and lead to lower F values, managing at the same time to determine the

5. Benchmark series: http://oaei.ontologymatching.org/versions/bench50.zip

6. We have used the results for these methods as provided on the OAEI web site [21].

correct matches between the elements of the ontologies.

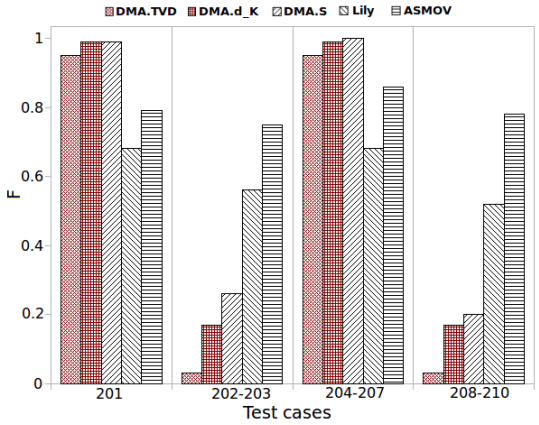

Fig. 17. Behavior of the evaluation measures in the test cases of the second group of the OAEI set.

In the second case though, where comments are also missing $(202-203)$, the common term space created by DMA is reduced significantly. Thus, the method faces difficulties in locating the context of each element and determining the matches. As a result, PCP and PCR impose a heavy penalty on the learned ontology. The fact that TVD is more strict than d_K and S is due to the way dissimilarity is calculated by its formula. On the other hand, Lily and ASMOV provide a set of matches that is close to the correct one, with fewer mistakes than DMA. Hence PCP and PCR penalize less the learned ontologies. This case $(202-203)$ indicates that DMA should not be used for evaluating learned ontologies when the term space is very small, i.e. it comprises very few terms.

When synonyms or translations are introduced $(204-207)$, DMA becomes more tolerant than Lily and ASMOV. This is mainly attributed to the fact that DMA allows a learned concept to have a different name from a gold concept. Thus, if this concept is a synonym or a translation of the gold one, it is assumed correct and shall not be penalized heavily. Recall that the name of a concept is just a single feature in its representation. Thus, a change in a single feature (translation or synonym) does not lead to a heavy penalty. On the other hand, Lily and ASMOV, which are more dependent on concept names make some mistakes in matching concepts.

The final set of the second group $(208-210)$ combines synonyms, translations and other name conventions all together. The effect of this additional deviation of the ontology on Lily and ASMOV is relatively small and it is attributed to a penalty through *SimDist* and PCR for some matching mistakes. On the other hand, DMA produces very low F values, due to its inability to match the context of each learned element to the context of the corresponding gold element. The corresponding representations differ significantly, because of the concurrent introduction of synonyms, translations, conventions, acronyms, and so forth, that change the majority of the features in the vector representations. As a general conclusion for this group, DMA seems sensitive to drastic changes of the term space, while it is suitable for cases where concept names change.

Regarding the third test group $(221-247)$, the interesting cases are (a) those introducing changes at the relational level of the ontology $(221-223)$, (b) those removing instances or local property restrictions $(224-225)$, and (c) those substituting concepts with concept sets $(230-231)$.

As Figure 18 illustrates, for the first case $(221-223)$, the changes at the relational layer affect heavily all methods, which is expected, since the resulting ontology differs significantly from the reference one. The main penalty here is introduced by the PCP and PCR factors that are sensitive to changes in the relations of the ontology.

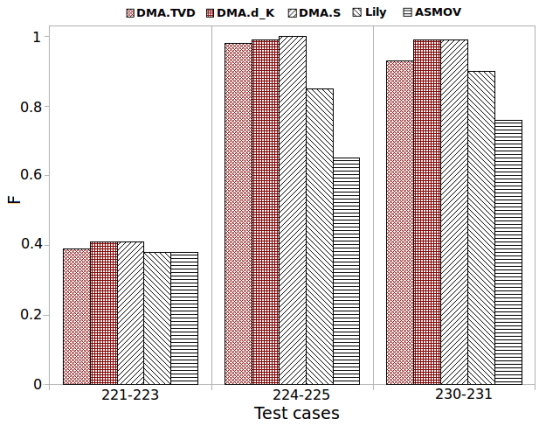

Fig. 18. Behavior of the evaluation measures in the test cases of the third group of the OAEI set.

In the second case $(224-225)$, the fact that instances and restrictions are missing affects the lexical layer of the ontology. DMA is not affected significantly and is able to locate the correct context of each ontology element, thus creating the corresponding representation. This is because the term space incorporates terms from labels, lexicalizations, domain, range, etc. that do not change. On the other hand, Lily and ASMOV are more sensitive and impose a larger penalty to the learned ontology, leading to lower F values.

Regarding the last case $(230-231)$, almost all methods behave similarly to the second case. The observed effect is mainly due to the PCP factor that penalizes extra concepts that do not appear in the gold ontology. The R value is equal to 1 in the majority of the cases for DMA, pushing the F value also very close to 1. The tolerance of the proposed method in this case is justified by the fact that existing concepts are left intact and only some extra concepts are introduced as specializations of leaf concepts. This affects only the PCP factor, and does not result in significant changes in the cotopy sets of the learned concepts.

The fourth group $(248-266)$ is the most difficult one for matching, as the "learned" ontology is very different from the reference one. The variants of the

reference ontology comprise errors in both the lexical and the relational layer. Concept names and labels may be scrambled, in the sense that they are replaced by random strings, comments are missing and in many cases, instances and properties are missing also. Regarding the hierarchical layer, it may be expanded or flattened.

As Figure 19 shows, the evaluation results for all methods are rather low. This is particularly true for DMA.TVD, as it penalizes heavier the large differences in the term distributions. On the other hand, when using Lily and ASMOV, the F is near 0.2. These methods manage to determine a small set of good matches, carrying a smaller penalty through the PCP and PCR factors.

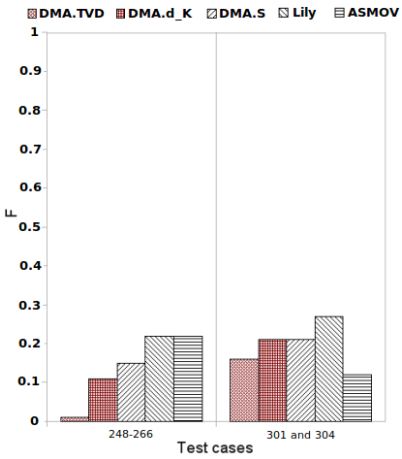

Fig. 19. Behavior of the evaluation measures in the test cases of the fourth and fifth group of the OAEI set.

Finally, the last group $(301 - 304)$ compares the reference ontology against four real ontologies. Only 301 and 304 have some elements in common with the reference ontology. For the other two, no matching was possible. Figure 19 shows that for the cases 301 and 304 where the "learned" ontology has some elements in common with the gold standard, all methods result in low F values. In the more extreme cases 302 and 303 (not depicted in the figure), zero F values are obtained, as expected.

Table 1 summarizes the basic conclusions from this experiment. A threshold to the F value has been introduced, in order to illustrate the tolerance of each method. From figures 16, 17, 18 and 19, three major cases can be distinguished: (a) that where the majority of the methods achieves F values near 1.0, (b) that where the highest F value is around 0.8, and (c) that where all methods score below 0.5. In the first case, values near 1.0 indicate that the method is tolerant to a particular situation, values near 0.9 indicate relative tolerance, while values below 0.9 indicate no tolerance. In the second case, values around 0.8 indicate relative tolerance, while lower values indicate no tolerance. Finally, regarding the last case, values below 0.5 indicate no tolerance.

Based on this qualitative interpretation of the results, DMA (with d_K and S) and ASMOV seem to be tolerant to generalizations and restrictions, regarding the representation language, e.g. moving from OWL Full to OWL Lite. Lily is relatively tolerant to restrictions but insensitive to generalizations. When names are removed, DMA is the only method that is able to provide a correct set of matches and evaluate objectively the learned ontology, but when comments are also removed, DMA becomes very sensitive, while ASMOV and Lily are more tolerant. DMA is also tolerant to synonyms and translations. Replacing concept names with synonyms has a small effect, since the concept label changes, but the reference concept remains the same (recall Fig. 1 in Section 1). In contrast, state-of-the-art ontology matching methods, such as Lily and ASMOV, impose higher penalties in such cases. When extreme lexical changes or big relational changes occur, all methods become very sensitive, as is to be expected. Finally, DMA is also tolerant to the removal of instances and restrictions. Therefore, in cases where names are missing, concepts are replaced by synonyms or translations, or instances and restrictions are missing, it is safer to use DMA, since it is able to judge more objectively by a relevant tolerance the learned ontology with respect to the gold one.

Thus, the experiments have emphasized that the proposed method (DMA) goes beyond the superficial matching of the ontologies. The results of the proposed evaluation method depend mainly on the effectiveness of the matching between the ontologies, which in turns affects the penalty through PCP and PCR, and the strictness of the method in penalizing lexical differences through $SimDist$.

What needs to be stressed however is that the proposed evaluation framework supports a variety of different matching methods. Thus, in cases where sophisticated methods, like ASMOV and Lily are more suitable than DMA, then they should be preferred. Otherwise, the DMA provides a good evaluation approach. The ultimate goal is to evaluate the learned ontology as objectively as possible.

5 Conclusions

An automated ontology evaluation method needs to be flexible and to support the use of measures that take into account various aspects of the ontology. Furthermore, it is important to use effective ontology matching methods that go beyond superficial string matching. In this paper, a novel method for evaluating learned ontologies against gold ones was presented, as well as a new set of evaluation measures that rely on a distributional representation of the ontology elements, based on their contexts. The proposed method provides a flexible framework for evaluation, enabling the use of various methods, such as the ones of the field of ontology matching that avoid common problems of evaluating ontologies, such as the superficial string matching between concepts.

In addition, the proposed similarity measures take into account the lexical and the relational dimensions of the learned ontology and penalize it in proportion

TABLE 1
Tolerance of methods in different tests. Check mark indicates tolerance, "x" indicates no tolerance and "-" indicates relative tolerance.

Test case	DMA TVD	DMA d_K	DMA S	Lily	ASMOV	Group
Language Generalizations	x	✓	✓	✓	✓	103
Language Restrictions	x	✓	✓	✓	-	104
No names	✓	✓	✓	x	x	201
No names, no comments	x	x	x	-	-	202-203
Synonyms, translations	✓	✓	✓	x	-	204-207
Synonyms, translations, conventions	x	x	x	x	-	208-210
Relational, lexical changes	x	x	x	x	x	221-223
No instances, no restrictions	✓	✓	✓	-	x	224-225

to its differences from the gold standard. Moreover, the generality of the evaluation measures allows a flexible choice of dissimilarity measure to be used. This possibility was demonstrated through experimentation with the Total Variational Distance, the Kolmogorov Distance and the Separation Distance. Additionally, the evaluation measures can be used with any method from the field of ontology alignment to determine the matching between the gold and the learned ontology. This was also tested, using the state-of-the-art ontology matching methods Lily and ASMOV. The extensive and unbiased evaluation that was performed has indicated that the proposed evaluation methodology is suitable for assessing evaluation methods and ontology learning methods as well.

From the experimental assessment, we concluded first that the method penalizes in a near-linear fashion the increasing deviation of two ontologies, taking values in the closed interval [0,1]. Second, the proposed method managed to derive a good assessment in most cases, avoiding the superficial string matching of ontology elements. Third, we have showed that it is straightforward to incorporate any ontology alignment method, and finally, the importance of choosing an appropriate matching method that provides correct sets of matches in order to obtain accurate evaluation results was also studied. Despite the simplicity of DMA, we have observed that in many cases it is preferable to use that method in conjunction with TVD, d_k, or S, rather more sophisticated and strict ones. However, the intention of the experiment is to show which alignment method must be chosen according to various cases.

Future directions include the evaluation of ontologies that have been learned from non-textual sources, where ontology elements cannot be represented as distributions over terms, but over different types of features. A final intension is to improve the proposed matching method, incorporating useful features from the literature of ontology alignment.

REFERENCES

[1] R. Porzel and R. Malaka, "A Task-based Approach for Ontology Evaluation", ECAI 2004 OLP Workshop, 2004.

[2] P. Velardi and R. Navigli and A. Cuchiarelli and F. Neri, "Evaluation of OntoLearn, a Methodology for Automatic Learning of Domain Ontologies", In Ontology Learning from Text: Methods, Evaluation and Applications, IOS Press, 2005.

[3] C. Brewster and H. Alani and S. Dasmahapatra and Y. Wilks, "Data Driven Ontology Evaluation", In Proceedings of the International Conference on Language Resources and Evaluation, Lisbon, 2004.

[4] A.B. Jones and C.V. Storey and V. Sugumaran and P. Ahluwalia, "A Semiotic Suite for Assessing the Quality of Ontologies", In Data and Knowledge Engineering, vol. 55, no. 1, pp. 84-102, 2004.

[5] A. Lozano-Tello and A. Gomez-Perez, "Ontometric: A method to choose the appropriate ontology", Journal of Database Management, vol. 15, no 2, pp. 1-18, 2004.

[6] A. Lozano-Tello and A. Gomez-Perez and E. Sosa, "Selection of Ontologies for the Semantic Web", In ICWE 2003, JM Cueva Lovelle et al. (Eds.), LNCS 2722, pages 413416, 2003.

[7] J. Brank and D. Mladenić and M. Grobelnik, "Gold Standard Based Ontology Evaluation Using Instance Assignment", In Proceedings of the EON 2006 Workshop, 2006.

[8] 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.

[9] D. Maynard and 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] E. Zavitsanos and G. Paliouras and G.A. Vouros, "A Distributional Approach to Evaluating Ontology Learning Methods Using a Gold Standard", In ECAI 2008 OLP Workshop, 2008.

[12] I.V. Levenshtein, "Binary Codes Capable of Correcting Deletions, Insertions and Reversals", In Cybernetics and Control Theory, vol. 10, no. 8, pp. 707-710, 1966.

[13] M. Ehrig and P. Haase and N. Stohanovic and M. Hefke, "Similarity for Ontologies - A Comprehensive Framework", In Proceedings of the European Conference in Information Systems, 2005.

[14] A. Meadche and S. Staab, "Ontology Learning for the Semantic Web", In IEEE Intelligent Systems, vol. 16, no. 2, pp. 72-79, 2001.

[15] M. Sabou and C. Wroe and C. Goble and H. Stuckenschmidt, "Learning Domain Ontologies for Semantic Web Service Descriptions", Journal of Web Semantics, vol. 3, no. 4, 2005.

[16] E. Zavitsanos and S. Petridis and G. Paliouras and G.A. Vouros, "Determining Automatically the Size of Learned Ontologies", In 18th European Conference on Artificial Intelligence, ECAI, 2008.

[17] H.F. Witschel, "Using Decision Trees and Text Mining Techniques for Extending Taxonomies", In Proceedings of Learning and Extending Lexical Ontologies by Using Machine Learning Methods, Workshop at ICML-05, 2005.

[18] A. Wagner, "Enriching a Lexical Semantic Net with Selectional Preferences by Means of Statistical Corpus Analysis", In Proceedings of the ECAI 2000 Workshop on Ontology Learning, Berlin, pages 3742, 2000.

[19] B. Ganter and R. Wille. "Formal Concept Analysis: Mathematical Foundations". Springer-Verlag, 1999.

[20] A.L. Gibbs and F.E. Su, "On Choosing and Bounding Probability Metrics", International Statistical Review, vol. 70, no. 3, pp. 419-435, 2002.

[21] Ontology Alignment Evaluation Initiative, http://oaei.ontologymatching.org

[22] Y. Kalfoglou and M. Schorlemmer, "Ontology Mapping: The State of the Art", The Knowledge Engineering Review, vol. 18, no. 1, 2003.

[23] P. Wang and B. Xu, "Lily: ontology alignment results for OAEI 2008", In ISWC-2008 Workshop on Ontology Matching, 2008.

[24] Y.R. Jean-Mary and M.R. Kabuka, "ASMOV: results for OAEI 2008", In ISWC-2008 Workshop on Ontology Matching, 2008.

APPENDIX
EXPERIMENTAL RESULTS

TABLE 2. Evaluation results in the OAEI benchmark series using the proposed method and the Total Variational Distance (TVD) or the Kolmogorov Distance (d_K) as dissimilarity measures.

Data	TVD		d_K	
ID	P_{value}	R_{value}	P_{value}	R_{value}
101	1.0	1.0	1.0	1.0
102	0.0	0.0	0.0	0.0
103	0.91	0.91	0.99	0.99
104	0.82	0.82	0.98	0.98
201	0.93	0.93	0.99	0.99
201-2	0.97	0.97	0.99	0.99
201-4	0.97	0.97	1.0	1.0
201-6	0.94	0.94	0.99	0.99
201-8	0.95	0.95	1.0	1.0
202	0.02	0.02	0.21	0.15
202-2	0.03	0.03	0.18	0.17
202-4	0.02	0.02	0.19	0.15
202-6	0.02	0.03	0.16	0.15
202-8	0.02	0.02	0.24	0.17
203	0.05	0.06	0.19	0.2
204	0.97	0.97	0.99	0.99
205	0.95	0.95	1.0	1.0
206	0.95	0.95	1.0	1.0
207	0.94	0.94	0.99	0.99
208	0.04	0.05	0.18	0.17
209	0.02	0.02	0.16	0.15
210	0.02	0.02	0.21	0.17
221	0.73	0.18	0.98	0.18
222	0.96	0.47	1.0	0.48
223	0.15	0.9	0.15	0.93
224	0.96	0.96	0.99	0.99
225	0.99	0.99	1.0	1.0
228	0.0	0.0	0.0	0.0
230	0.86	0.86	0.99	0.99
231	1.0	1.0	1.0	1.0
232	0.69	0.17	0.97	0.18
233	0.01	0.0	0.01	0.0
236	0.0	0.0	0.0	0.0
237	0.92	0.45	0.98	0.47
238	0.15	0.93	0.12	0.92
239	0.48	0.26	0.31	0.27
240	0.08	0.38	0.06	0.17
241	0.01	0.0	0.01	0.0
246	0.44	0.24	0.82	0.43
247	0.08	0.37	0.06	0.19
248	0.04	0.02	0.83	0.15
248-2	0.08	0.03	0.83	0.15
248-4	0.07	0.02	0.83	0.15
248-6	0.05	0.02	0.83	0.15
248-8	0.05	0.02	0.83	0.15
249	0.0004	0	0.19	0.16
249-2	0.01	0.02	0.19	0.19
249-4	0.01	0.01	0.24	0.24
249-6	0.0	0.01	0.22	0.19
249-8	0.0	0.01	0.19	0.19
250	0.0	0.0	0.0	0.0
250-2	0.0	0.0	0.0	0.0
250-4	0.0	0.0	0.0	0.0
250-6	0.0	0.0	0.0	0.0
250-8	0.0005	0.0005	0.0005	0.0005
251	0.02	0.02	0.58	0.21
251-2	0.03	0.03	0.23	0.18
251-4	0.03	0.03	0.23	0.18
251-6	0.03	0.03	0.21	0.18
251-8	0.02	0.02	0.18	0.15
252	0.0	0.01	0.08	0.24
252-2	0.01	0.03	0.08	0.22
252-4	0.01	0.03	0.0	0.16
252-6	0.01	0.03	0.05	0.24
252-8	0.01	0.03	0.05	0.24
253	0.0	0.0	0.87	0.15
253-2	0.05	0.02	0.87	0.15
253-4	0.03	0.01	0.87	0.15
253-6	0.02	0.01	0.87	0.15
253-8	0.01	0.01	0.87	0.15
254	0.0	0.0	0.0	0.0
254-2	0.01	0.0	0.01	0.0
254-4	0.0	0.0	0.0	0.0
254-6	0.0	0.0	0.0	0.0
254-8	0.0008	0.0005	0.0008	0.0005
257	0.0	0.0	0.0	0.0
257-2	0.0	0.0	0.0	0.0
257-4	0.0	0.0	0.0	0.0
257-6	0.0	0.0	0.0	0.0
257-8	0.0005	0.0005	0.0005	0.0005
258	0.0004	0.0	0.22	0.13
258-2	0.02	0.01	0.22	0.09
258-4	0.01	0.01	0.22	0.09
258-6	0.01	0.01	0.25	0.13
258-8	0.0	0.01	0.22	0.09
259	0.0	0.0	0.05	0.21
259-2	0.01	0.02	0.05	0.21
259-4	0.01	0.02	0.05	0.19
259-6	0.01	0.02	0.05	0.24
259-8	0.01	0.02	0.03	0.21
260	0.02	0.02	0.18	0.14
260-2	0.03	0.03	0.24	0.16

continued on next page

continued from previous page

continued from previous page

ID	P_{value}	R_{value}	P_{value}	R_{value}
260-4	0.04	0.03	0.14	0.08
260-6	0.02	0.03	0.21	0.12
261	0.0001	0.01	0.03	0.14
261-2	0.01	0.02	0.05	0.18
261-4	0.01	0.02	0.07	0.18
261-6	0.01	0.02	0.07	0.16
261-8	0.01	0.02	0.05	0.18
262	0.0	0.0	0.0	0.0
262-2	0.01	0.0	0.01	0.0
262-4	0.0	0.0	0.0	0.0
262-6	0.0	0.0	0.0	0.0
262-8	0.0008	0.0005	0.0008	0.0005
265	0.0	0.0	0.14	0.12
266	0.0	0.0	0.05	0.09
301	0.64	0.04	0.97	0.06
302	0.0	0.0	0.0	0.0
303	0.0	0.0	0.0	0.0
304	0.17	0.42	0.21	0.6

TABLE 3. Evaluation results in the OAEI benchmark series using the proposed method and the Separation Distance (S) as dissimilarity measure.

Data	S		Data	S	
ID	P_{value}	R_{value}	ID	P_{value}	R_{value}
101	1.0	1.0	250	0.09	0.09
102	0.0	0.0	250-2	0.14	0.14
103	0.99	0.99	250-4	0.12	0.12
104	0.99	0.99	250-6	0.11	0.11
201	0.99	0.99	250-8	0.11	0.11
201-2	0.99	0.99	251	0.18	0.16
201-4	1.0	1.0	251-2	0.24	0.2
201-6	0.99	0.99	251-4	0.24	0.2
201-8	1.0	1.0	251-6	0.18	0.16
202	0.19	0.2	251-8	0.25	0.2
202-2	0.16	0.2	252	0.0	0.21
202-4	0.19	0.23	252-2	0.0	0.32
202-6	0.19	0.17	252-4	0.0	0.29
202-8	0.22	0.2	252-6	0.0	0.23
203	0.62	0.62	252-8	0.0	0.32
204	0.99	0.99	253	0.89	0.16
205	1.0	1.0	253-2	0.89	0.16
206	1.0	1.0	253-4	0.89	0.16
207	0.99	0.99	253-6	0.89	0.16
208	0.19	0.26	253-8	0.89	0.16
209	0.19	0.2	254	0.5	0.09
210	0.16	0.2	254-2	0.5	0.09
221	0.98	0.18	254-4	0.5	0.09
222	1.0	0.48	254-6	0.5	0.09
223	0.15	0.94	254-8	0.5	0.09
224	1.0	1.0	257	0.11	0.09
225	1.0	1.0	257-2	0.14	0.14

continued on next page

continued from previous page

ID	P_{value}	R_{value}	ID	P_{value}	R_{value}
228	0.35	0.35	257-4	0.12	0.12
230	0.99	0.99	257-6	0.09	0.09
231	1.0	1.0	257-8	0.09	0.11
232	0.97	0.18	258	0.22	0.2
233	0.5	0.09	258-2	0.22	0.23
236	0.35	0.35	258-4	0.25	0.23
237	0.99	0.48	258-6	0.22	0.2
238	0.15	0.99	258-8	0.56	0.26
239	0.72	0.28	259	0.0	0.24
240	0.06	0.17	259-2	0.0	0.21
241	0.5	0.09	259-4	0.0	0.29
246	0.77	0.37	259-6	0.0	0.26
247	0.06	0.17	259-8	0.0	0.21
248	0.86	0.16	260	0.22	0.17
248-2	0.86	0.16	260-2	0.3	0.2
248-4	0.86	0.16	260-4	0.17	0.1
248-6	0.86	0.16	260-6	0.27	0.15
248-8	0.86	0.16	261	0.04	0.17
249	0.19	0.2	261-2	0.07	0.22
249-2	0.19	0.2	261-4	0.09	0.22
249-4	0.19	0.2	261-6	0.09	0.2
249-6	0.19	0.2	261-8	0.07	0.22
249-8	0.19	0.2	262	0.5	0.09
265	0.18	0.16	262-2	0.5	0.09
266	0.06	0.12	262-4	0.5	0.09
301	0.98	0.07	262-6	0.5	0.09
302	0.0	0.0	262-8	0.5	0.09
303	0.0	0.0	304	0.19	0.58

TABLE 4. Evaluation results in the OAEI benchmark series using Lily and ASMOV matching methods to produce the mapping.

Data	Lily		ASMOV	
ID	P_{value}	R_{value}	P_{value}	R_{value}
101	1.0	1.0	1.0	1.0
102	0.0	0.0	0.0	0.0
103	0.95	0.95	0.97	0.97
104	0.95	0.95	0.88	0.88
201	0.46	0.46	0.67	0.67
201-2	0.88	0.88	0.91	0.91
201-4	0.78	0.78	0.86	0.86
201-6	0.69	0.69	0.81	0.81
201-8	0.57	0.57	0.71	0.71
202	0.42	0.42	0.49	0.49
202-2	0.65	0.65	0.89	0.89
202-4	0.57	0.57	0.8	0.8
202-6	0.49	0.49	0.71	0.71
202-8	0.41	0.41	0.61	0.61
203	0.8	0.8	1.0	1.0
204	0.91	0.91	0.98	0.98
205	0.61	0.61	0.85	0.85
206	0.59	0.59	0.81	0.81

continued on next page

continued from previous page

ID	P_{value}	R_{value}	P_{value}	R_{value}
207	0.6	0.6	0.81	0.81
208	0.72	0.72	0.98	0.98
209	0.44	0.44	0.75	0.75
210	0.39	0.39	0.61	0.61
221	0.78	0.18	1.0	0.18
222	0.83	0.47	0.78	0.47
223	0.15	0.88	0.15	0.93
224	0.9	0.9	0.94	0.94
225	0.81	0.81	0.36	0.36
228	0.57	0.57	0.36	0.36
230	0.81	0.81	0.52	0.52
231	0.99	0.99	1.0	1.0
232	0.74	0.14	0.93	0.16
233	0.44	0.06	0.5	0.06
236	0.55	0.55	0.34	0.34
237	0.76	0.41	0.72	0.43
238	0.15	0.95	0.13	0.88
239	0.5	0.26	0.28	0.16
240	0.05	0.58	0.05	0.31
241	0.39	0.05	0.48	0.05
246	0.48	0.24	0.27	0.15
247	0.15	0.95	0.05	0.3
248	0.41	0.12	0.53	0.14
248-2	0.58	0.13	0.93	0.17
248-4	0.53	0.12	0.85	0.16
248-6	0.51	0.12	0.76	0.17
248-8	0.45	0.12	0.66	0.16
249	0.44	0.44	0.42	0.42
249-2	0.61	0.61	0.81	0.81
249-4	0.57	0.57	0.64	0.64
249-6	0.46	0.46	0.59	0.59
249-8	0.4	0.4	0.47	0.47
250	0.23	0.23	0.03	0.03
250-2	0.37	0.37	0.25	0.25
250-4	0.29	0.29	0.2	0.2
250-6	0.31	0.31	0.16	0.16
250-8	0.19	0.19	0.04	0.04
251	0.44	0.23	0.41	0.29
251-2	0.58	0.32	0.71	0.43
251-4	0.52	0.27	0.65	0.39
251-6	0.51	0.23	0.6	0.37
251-8	0.44	0.21	0.52	0.32
252	0.09	0.37	0.11	0.54
252-2	0.12	0.61	0.14	0.85
252-4	0.12	0.61	0.14	0.85
252-6	0.12	0.61	0.14	0.85
252-8	0.12	0.61	0.14	0.85
253	0.41	0.13	0.45	0.14
253-2	0.56	0.13	0.85	0.15
253-4	0.53	0.12	0.74	0.14
253-6	0.46	0.11	0.71	0.14

continued on next page

continued from previous page

ID	P_{value}	R_{value}	P_{value}	R_{value}
253-8	0.43	0.12	0.58	0.15
254	0.46	0.11	0.06	0.01
254-2	0.37	0.05	0.48	0.05
254-4	0.38	0.06	0.45	0.04
254-6	0.36	0.07	0.36	0.05
254-8	0.38	0.09	0.28	0.05
257	0.14	0.14	0.08	0.08
257-2	0.35	0.35	0.25	0.25
257-4	0.31	0.31	0.2	0.2
257-6	0.23	0.23	0.15	0.15
257-8	0.22	0.22	0.05	0.05
258	0.43	0.23	0.34	0.24
258-2	0.54	0.3	0.65	0.39
258-4	0.5	0.25	0.52	0.33
258-6	0.5	0.23	0.48	0.31
258-8	0.39	0.18	0.39	0.25
259	0.08	0.35	0.12	0.37
259-2	0.11	0.59	0.12	0.79
259-4	0.11	0.59	0.12	0.79
259-6	0.11	0.6	0.12	0.79
259-8	0.11	0.6	0.12	0.79
260	0.29	0.13	0.03	0.03
260-2	0.4	0.19	0.1	0.1
260-4	0.33	0.14	0.08	0.08
260-6	0.32	0.14	0.08	0.08
261	0.01	0.19	0.02	0.03
261-2	0.03	0.37	0.04	0.26
261-4	0.03	0.37	0.04	0.26
261-6	0.03	0.36	0.04	0.26
261-8	0.03	0.38	0.04	0.26
262	0.0	0.0	0.0	0.0
262-2	0.34	0.04	0.48	0.05
262-4	0.35	0.04	0.48	0.04
262-6	0.35	0.05	0.48	0.06
262-8	0.35	0.09	0.5	0.08
265	0.14	0.14	0.08	0.08
266	0.02	0.09	0.04	0.08
301	0.37	0.03	0.19	0.02
302	0.0	0.0	0.0	0.0
303	0.0	0.0	0.0	0.0
304	0.38	0.6	0.18	0.26

TABLE 5. Overall mean F values in the four most representative groups.

Groups	TVD	d_K	S	ASMOV	Lily
GROUP 1	0.68	0.74	0.75	0.71	0.72
GROUP 2	0.49	0.58	0.62	0.79	0.61
GROUP 3	0.39	0.41	0.47	0.36	0.47
GROUP 4	0.01	0.11	0.15	0.22	0.22