On How to Perform a Gold Standard Based Evaluation of Ontology Learning

Klaas Dellschaft and Steffen Staab

Universität Koblenz-Landau, ISWeb Working Group
Universitätsstr. 1, 56070 Koblenz, Germany
{klaasd, staab}@uni-koblenz.de,
WWW home page: http://isweb.uni-koblenz.de

Abstract. In recent years several measures for the gold standard based evaluation of ontology learning were proposed. They can be distinguished by the layers of an ontology (e.g. lexical term layer and concept hierarchy) they evaluate. Judging those measures with a list of criteria we show that there exist some measures sufficient for evaluating the lexical term layer. However, existing measures for the evaluation of concept hierarchies fail to meet basic criteria. This paper presents a new taxonomic measure which overcomes the problems of current approaches.

1 Introduction

The capabilities of ontology learning approaches may be tested by (i) evaluation in a running application, (ii) a posteriori evaluation by experts, or (iii) evaluation by comparison of learned results against a pre-defined ”gold standard”. Though approaches (i) and (ii) exhibit some considerable advantages over approach (iii), when it comes to frequent and large-scale evaluations and comparisons of multiple ontology learning approaches, only approach (iii) is feasible in practice. Since such – comparably – easily repeatable evaluation schemes contributed heavily to the overwhelming success of disciplines like information retrieval, machine learning or speech recognition, we conjecture that a similar success of ontology learning requires an analogous scheme for evaluation with gold standards, too.

Examples of gold standard-based evaluations of ontology learning can be found in [1], [2] and [3] – to name but a few. However, it is apparent that there does not exist a canonical way of performing gold-standard based evaluations of ontology learning. Moreover, we argue in this paper that existing gold-standard based evaluations are faulty and that a well-founded evaluation model is largely missing. Therefore, we describe here a new framework for gold standard-based evaluation of ontology learning that avoids common mistakes and we show by some analytical considerations and by some experiments that the new framework fulfills crucial evaluation criteria that other frameworks do not meet.

2 Related Work

There exist many measures for the reference-based evaluation of ontologies. One may distinguish between measures which only evaluate the lexical term layer of an ontology, those which also take the concept hierarchy into account and the ones which evaluate the non-taxonomic relations contained in an ontology. In this paper we will concentrate on the measures for evaluating concept hierarchies and the lexical term layer.

On the lexical term layer ”binary” measures are often used that compare the terms from the reference and the learned ontology based on an exact match of strings. Examples for this kind of measure are the *Term Precision and Term Recall* as they are presented in [3]. There exist also several other names for these measures like *Lexical Precision and Recall* or simply *precision and recall* (see [4] and [5]). Another example of a term level evaluation measure is the *String Matching* measure presented in [6] and [7]. This measure is based on the edit distance between two strings. It is therefore more robust with regard to slightly different spellings and typing errors (e.g. ”center” and ”centre”).

The comparison of concept hierarchies is more complicated than the comparison of the lexical term layer of ontologies. Such concept hierarchy measures are often divided into kinds of local and global measures. The local measure compares the similarity of the positions of two concepts in the learned and the reference hierarchy. The global measure is then computed by averaging the results of the local measure for concept pairs from the reference and the learned ontology.

One of the first examples of a concept hierarchy evaluation measure is the *Taxonomic Overlap* (TO) presented in [6] and [7]. The local taxonomic overlap compares two concepts based on the set of all their super- and sub concepts. In opposite to the local overlap, which is a symmetric measure, this is not the case for the global taxonomic overlap measures proposed in [6], [7] and [8], i.e. they can be computed into two directions. In [8] this asymmetry is interpreted as a kind of precision and recall. But in section 4.5 we will show that this is a misinterpretation of the asymmetry, as local taxonomic overlap already constitutes a kind of combination of precision and recall.

Another example is the *Augmented Precision and Recall* (AP & AR) presented in [9]. It is also divided into a global and a local part of the measure. For the local part two alternatives may be used: The *Learning Accuracy* (LA) and the *Balanced Distance Metric* (BDM). LA was proposed by [10]. It compares two concepts based on their distance in the tree (e.g. the length of the shortest path between the root and their most specific common abstraction). BDM further develops the idea of LA by taking further types of paths and a branching factor of the concepts into account (see [9]).

The latest measure for comparing concept hierarchies is the *OntoRand* index proposed in [11]. It is a symmetric measure which extends techniques used in the clustering community for comparing two partitions of the same set of instances. A concept hierarchy is seen as a hierarchical partitioning of instances. For OntoRand two alternatives exist to measure the similarity of concepts. The first alternative is based on the set of common ancestors. The second alternative is based on the distance between two concepts in the tree (like LA and BDM). An important constraint imposed on the concept hierarchy is that both compared hierarchies must contain the same set of instances.

3 Criteria for Good Evaluation Measures

Given this variety of evaluation measures for concept hierarchies it is now the question what is a ”good” measure and can we give some criteria according to which to evaluate the different measures. Measures fulfilling the following criteria will help to avoid the misinterpretation of evaluation results and ease drawing the right conclusions for the improvement of the evaluated ontology learning procedure.

The most important criterion is that a measure allows to evaluate an ontology along multiple dimensions. This criterion is formulated in several papers like [9] and [12]. Thus a user can weight different kinds of errors based on his own preferences. This enables to better analyze the strengths and weaknesses of a learned ontology.

If a multi dimensional evaluation is performed, each measure should be influenced just by one dimension, i.e. by one type of error only. For example, if one uses measures for evaluating the lexical term layer of an ontology (e.g the lexical precision and recall) and one also wants to evaluate the quality of the learned concept hierarchy (e.g. with the taxonomic overlap), then a dependency between those measures should be avoided.

The second criterion is that the effect of an error onto the measure should be proportional to the distance between the correct and the given result. For example, an error near the root of a concept hierarchy should have a stronger effect on the evaluation measure than an error nearer to the leafs (see also [12]).

The third criterion is closely related to the previous one. For measures with a closed scale interval (e.g. $[0..1]$), a gradual increase in the error rate should also lead to a gradual decrease in the evaluation results. For example, if a measure has the interval $[0..1]$ as its scale but already slight errors lead to a decrease of the returned results from 1 to 0.2 then it is difficult to distinguish between slight and severe errors (see [11]).

In Tab. 1 it is shown in how far the measures described in section 2 meet the criteria listed in this section. The rating is based on the descriptions in [7], [9] and [11]. Additionally, the new findings from section 4.5 were used for rating the taxonomic overlap. A measure can improve its multi dimensionality by two factors: either by removing the influence of the lexical term layer on the evaluation of the concept hierarchy or by separately measuring different aspects of the hierarchy (e.g. precision and recall). None of the measures removes the influence of the lexical term layer and only the augmented precision and recall distinguishes between two aspects of the hierarchy. The Learning Accuracy does not achieve the best score for the proportional error effect because it

Table 1. Rating of concept hierarchy measures

	multi dimensionality	proportional error effect	usage of interval
TO	$-$	$+$	?
AP & AR	$\circ$	$+$	?
LA	$-$	$\circ$	?
OntoRand[1]	$-$	$+$ / $-$	$+$ / $-$
TP_{csc} (cf. section 4.3)	$+$	$+$	$+$

considers the distance between the correct and the given answer only to some small extent (see [9]). In the following a truly multi dimensional approach for evaluating an ontology will be presented, thus overcoming the problems of the current measures.

4 Comparing Learned Ontologies with Gold Standards

In this section measures will be presented which can be used for an evaluation of the lexical term layer and the concept hierarchy of an ontology. The measures extend the idea of precision and recall to the gold standard based evaluation of ontologies. The lexical term layer of an ontology will be evaluated with lexical precision and recall (see section 4.2). For the concept hierarchy a framework of building blocks will be defined in section 4.3. This framework defines a family of measures and it will be used for systematically constructing a measure which fulfills the criteria from section 3.

In the following the simplified definition of a core ontology will be used. This definition of an ontology only contains the lexical term layer and the concept hierarchy. Similarly to [8], we define a core ontology as follows:

Definition 1. *The structure* $\mathcal{O} := (\mathcal{C}, root, \leq_{\mathcal{C}})$ *is called a core ontology.* $\mathcal{C}$ *is a set of concept identifiers and* $root$ *is a designated root concept for the partial order* $\leq_{\mathcal{C}}$ *on* $\mathcal{C}$*. This partial order is called concept hierarchy or taxonomy. The equation* $\forall c \in \mathcal{C} : c \leq_{\mathcal{C}} root$ *holds for this concept hierarchy.*

In this definition of a core ontology the relation between lexical terms and their associated concept is a bijection, i.e. each term is associated with exactly one concept and each concept with exactly one term. Thus it is possible to use the a lexical term as the identifier of a concept. This restriction simplifies the following formulas. Nevertheless it would be possible to generalize them to the case where an $n : m$ relation between concepts and lexical terms exists (in analogy to [6] and [7]).

4.1 Precision & Recall

This section gives a short overview of precision, recall and F-measure, as they are known from information retrieval (see [13]). They are used for comparing a reference retrieval (*Ref*) with a computed retrieval ($Comp$) returned by a system. Precision and recall are defined as follows:

$$P(\mathit{Ref}, Comp) = \frac{|Comp \cap \mathit{Ref}|}{|Comp|} \qquad R(\mathit{Ref}, Comp) = \frac{|Comp \cap \mathit{Ref}|}{|\mathit{Ref}|} \tag{1}$$

It is interesting that precision and recall are the inverse of each other:

$$P(\mathit{Ref}, Comp) = \frac{|Comp \cap \mathit{Ref}|}{|Comp|} = R(Comp, \mathit{Ref}) \tag{2}$$

[1] It is shown in [11] that the measures based on tree distance in some cases do not show an proportional error effect and that they do not use the complete interval. These problems do not exist for the OntoRand measure based on common ancestors.

The F_1-measure is used for giving a summarizing overview and for balancing the precision and recall values. The F_1-measure is the harmonic mean of P and R.

$$F_1(\mathit{Ref}, \mathit{Comp}) = \frac{2 \cdot P(\mathit{Ref}, \mathit{Comp}) \cdot R(\mathit{Ref}, \mathit{Comp})}{P(\mathit{Ref}, \mathit{Comp}) + R(\mathit{Ref}, \mathit{Comp})} \tag{3}$$

4.2 Lexical Precision & Recall

There exist several measures sufficient for evaluating the lexical term layer of an ontology (see section 2). In this subsection the lexical precision and recall measures, as they are described in [4], will be explained in a bit more detail. Later on they will be used in conjunction with the measures for evaluating concept hierarchies, as they are presented in section 4.3. Given a computed core ontology $\mathcal{O}_C$ and a reference ontology $\mathcal{O}_R$, the lexical precision (LP) and lexical recall (LR) are defined as follows:

$$LP(\mathcal{O}_C, \mathcal{O}_R) = \frac{|\mathcal{C}_C \cap \mathcal{C}_R|}{|\mathcal{C}_C|} \qquad LR(\mathcal{O}_C, \mathcal{O}_R) = \frac{|\mathcal{C}_C \cap \mathcal{C}_R|}{|\mathcal{C}_R|} \tag{4}$$

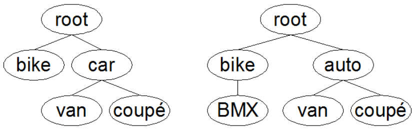

Fig. 1. Example reference ontology ($\mathcal{O}_{R1}$, left) and computed ontology ($\mathcal{O}_{C1}$, right)

The lexical precision and recall reflect how good the learned lexical terms cover the target domain. For example, if one compares $\mathcal{O}_{C1}$ and $\mathcal{O}_{R1}$ in Fig. 1 with each other, one gets $LP(\mathcal{O}_{C1}, \mathcal{O}_{R1}) = \frac{4}{6} = 0.67$ and $LR(\mathcal{O}_{C1}, \mathcal{O}_{R1}) = \frac{4}{5} = 0.8$.

4.3 Taxonomic Precision & Recall

In this subsection a framework of building blocks is described. It defines a family of taxonomic precision and recall measures from which two concrete measures will be selected afterward. Only the equations for the taxonomic precision measures will be presented. The corresponding equations for the taxonomic recall measures can be easily derived from them because of equation (2). This framework extends and improves the framework used for the taxonomic overlap measures in [7]. It especially replaces the previously used equation for comparing the position of two concepts with each other *leading to a completely different behavior of the measure* (see also section 4.5).

Comparing Concepts As mentioned before, measures for comparing two concept hierarchies with each other are usually divided into a kind of local and a global measure (cf. section 2). The local measure compares the positions of two concepts and the global measure is used for comparing two whole concept hierarchies. We start describing the framework's local measure. It is then used in the definition of the global measure.

For the local taxonomic precision the similarity of two concepts will be computed based on characteristic extracts from the concept hierarchy. Such an extract should characterize the position of a concept in the hierarchy, i.e. two extracts should contain many common objects if the characterized objects are at similar positions in the hierarchy. The proportion of common objects in the extracts should decrease with increasing dissimilarity of the characterized concepts. Given such an characteristic extract ce, the local taxonomic precision tp_{ce} of two concepts $c_1 \in \mathcal{O}_C$ and $c_2 \in \mathcal{O}_R$ is defined as

$$tp_{ce}(c_1, c_2, \mathcal{O}_C, \mathcal{O}_R) := \frac{|ce(c_1, \mathcal{O}_C) \cap ce(c_2, \mathcal{O}_R)|}{|ce(c_1, \mathcal{O}_C)|} \tag{5}$$

The characteristic extract from the concept hierarchy is an important building block of the local taxonomic measure and several alternative instantiations exist. As we will see below, they have a major influence on the properties of the corresponding global measure. For the taxonomic overlap measure described in [7] it was suggested to characterize a concept by its semantic cotopy, i.e. all its super- and subconcepts. Given the concept $c \in \mathcal{C}$ and the ontology $\mathcal{O}$, the semantic cotopy sc is defined as follows:

$$sc(c, \mathcal{O}) := \{c_i | c_i \in \mathcal{C} \wedge (c_i \leq c \vee c \leq c_i)\} \tag{6}$$

If one uses the semantic cotopy for defining the local taxonomic precision measure tp_{sc}, the results will be heavily influenced by the lexical precision of $\mathcal{O}_C$ because with decreasing lexical precision more and more concepts of $sc(c, \mathcal{O}_C)$ are not contained in $\mathcal{O}_R$ and $sc(c, \mathcal{O}_R)$. This increases the probability that $sc(c, \mathcal{O}_C)$ contains such concepts, leading to a direct dependency between the lexical and the taxonomic precision. But according to section 3, evaluation measures should be judged by whether the different measures are independent of each other. So taxonomic measures based on the semantic cotopy shouldn't be used in conjunction with the lexical precision and recall.

This influence of lexical precision and recall on the taxonomic measures can be avoided if one uses the common semantic cotopy csc as the characteristic extract. The common semantic cotopy excludes all concepts which are not also available in the other ontology's set of concepts:

$$csc(c, \mathcal{O}_1, \mathcal{O}_2) := \{c_i | c_i \in \mathcal{C}_1 \cap \mathcal{C}_2 \wedge (c_i <_1 c \vee c <_1 c_i)\} \tag{7}$$

In Tab. 2 and 3 one can see the influence of inserting and replacing concepts in a hierarchy. The tables contain the sets sc and csc for the ontologies $\mathcal{O}_{R1}$ and $\mathcal{O}_{C1}$ which were already used as an example for lexical precision and recall (see Fig. 1). One can see that inserting and replacing concepts without actually changing the hierarchy has no effect on the common semantic cotopy while the semantic cotopy is heavily influenced by these changes on the lexical term layer of an ontology.

Besides the previously described extracts of the concept hierarchy, further extracts are imaginable. For example, the upwards cotopy (see [7]) or the set of all direct subconcepts might be used. In [14] also measures based on the direct subconcepts were evaluated. But [14] shows also that measures based on the semantic cotopy meet more of the criteria from section 3.

Table 2. Semantic cotopies for the ontologies in Fig. 1.

c	$sc(c, \mathcal{O}_{R1})$	$sc(c, \mathcal{O}_{C1})$
root	{root, bike, car, van, coupé}	{root, bike, BMX, auto, van, coupé}
car	{root, car, van, coupé}	–
auto	–	{root, auto, van, coupé}
van	{root, car, van}	{root, auto, van}
coupé	{root, car, coupé}	{root, auto, coupé}
bike	{root, bike}	{root, bike, BMX}
BMX	–	{root, bike, BMX}

Table 3. Common semantic cotopies for the ontologies in Fig. 1.

c	$csc(c, \mathcal{O}_{R1}, \mathcal{O}_{C1})$	$csc(c, \mathcal{O}_{C1}, \mathcal{O}_{R1})$
root	{bike, van, coupé}	{bike, van, coupé}
car	{root, van, coupé}	–
auto	–	{root, van, coupé}
van	{root}	{root}
coupé	{root}	{root}
bike	{root}	{root}
BMX	–	{root, bike}

Comparing Concept Hierarchies It is now possible to define a framework for constructing a global taxonomic precision measure. Fig. 2 shows the building blocks used in this framework for a global taxonomic precision measure.

$$TP(\mathcal{O}_C, \mathcal{O}_R) := \frac{1}{|\mathcal{C}_C|} \underbrace{\sum_{\boxed{c \in \mathcal{C}_C}}}_{\text{concept set}} \begin{cases} \overbrace{\boxed{tp(c, c, \mathcal{O}_C, \mathcal{O}_R) \qquad \text{if } c \in \mathcal{C}_R}}^{\text{local taxonomic precision}} \\ \underbrace{\boxed{\max_{c' \notin \mathcal{C}_R} tp(c, c', \mathcal{O}_C, \mathcal{O}_R) \quad \text{if } c \notin \mathcal{C}_R}}_{\text{estimation}} \end{cases}$$

Fig. 2. Building blocks of the global taxonomic precision measure

The *set of concepts* whose local taxonomic precision values are summed up is the first building block. Two alternatives may be used. The first alternative is to use the set of concepts $\mathcal{C}_C$ from the learned ontology. If one chooses this alternative, the global taxonomic precision is influenced by the lexical precision. For example, if the lexical precision of a learned ontology is approximately 5% (like in the empirical evaluation in section 5.2) then for 95% of the concepts a local taxonomic precision value has to be estimated because there doesn't exist a corresponding concept in the reference ontology (see below). If such an influence of the lexical precision should be avoided then the set of common concepts $\mathcal{C}_C \cap \mathcal{C}_R$ should be preferred. It especially makes sense if one also uses a local taxonomic precision value based on the common semantic cotopy.

The *local taxonomic precision* is the next building block. It is used for comparing the position of a concept in the learned hierarchy with the position of the same concept in the reference hierarchy. Thus the current concept has to exist in both hierarchies.

An *estimation* of a local taxonomic precision value is the last building block. It is only used if the current concept isn't contained in both ontologies. Its usage is therefore influenced by the chosen set of concepts (see above). In [7] it is suggested to make an optimistic estimation by comparing the current concept with all concepts from the reference ontology and choose the highest local taxonomic precision value. This ensures that concepts which do not match on the lexical term layer (e.g. "auto" and "car" in Fig. 1) will nonetheless match in the concept hierarchy and thus return a high local taxonomic precision value. The optimistic estimation reduces the influence of lexical precision but it may also cause misleading results.

In opposite to that, assuming a local taxonomic precision value of 0% if no match on the lexical term layer can be found maximizes the influence of the lexical precision. But if one wants to completely eliminate the influence of lexical precision one should avoid this estimation building block anyway. This is done by only averaging the local taxonomic precision values of the common concepts.

Concrete Measures In the following the previously presented building blocks will be combined to concrete measures fulfilling the criteria from section 3. The measures will be evaluated in section 5. In [14] further measures are described and evaluated. This paper only contains the best two pairs of measures.

The first pair of measures consists of TP_{sc} and TR_{sc}. They are based on the semantic cotopy and are thus influenced by the lexical term layer. In the evaluation in section 5 they will be used for demonstrating the disadvantages of mixing the evaluation of lexical term layer and concept hierarchy. The other building blocks are selected so that they further increase this influence. This is achieved by computing the local taxonomic precision for all learned concepts and by estimating the local taxonomic precision as 0 if the current concept isn't also contained in the reference ontology.

$$TP_{sc}(\mathcal{O}_C, \mathcal{O}_R) := \frac{1}{|\mathcal{C}_C|} \sum_{c \in \mathcal{C}_C} \begin{cases} tp_{sc}(c, c, \mathcal{O}_C, \mathcal{O}_R) & \text{if } c \in \mathcal{C}_R \\ 0 & \text{if } c \notin \mathcal{C}_R \end{cases} \tag{8}$$

$$TR_{sc}(\mathcal{O}_C, \mathcal{O}_R) := TP_{sc}(\mathcal{O}_R, \mathcal{O}_C) \tag{9}$$

All in all, the measures TP_{sc} and TR_{sc} do not allow a separate evaluation of lexical term layer and concept hierarchy. For evaluation scenarios where a thorough analysis of the learned ontologies is needed the measures TP_{csc} and TR_{csc} are better suited. Here the building blocks will be selected so that the influence of the lexical term layer is minimized. This is achieved by using the common semantic cotopy and by computing the taxonomic precision values only for the common concepts of both ontologies. The latter makes the estimation of local taxonomic precision values unnecessary.

$$TP_{csc}(\mathcal{O}_C, \mathcal{O}_R) := \frac{1}{|\mathcal{C}_C \cap \mathcal{C}_R|} \sum_{c \in \mathcal{C}_C \cap \mathcal{C}_R} tp_{csc}(c, c, \mathcal{O}_C, \mathcal{O}_R) \tag{10}$$

$$TR_{csc}(\mathcal{O}_C, \mathcal{O}_R) := TP_{csc}(\mathcal{O}_R, \mathcal{O}_C) \tag{11}$$

4.4 Taxonomic F- and F'-Measure

Like it is the case for precision and recall in information retrieval, also the taxonomic precision and recall have to be balanced if one wants to output a combined measure. Therefore the taxonomic F-measure is introduced, which is the harmonic mean of the global taxonomic precision and recall.

$$TF(\mathcal{O}_C, \mathcal{O}_R) := \frac{2 \cdot TP(\mathcal{O}_C, \mathcal{O}_R) \cdot TR(\mathcal{O}_C, \mathcal{O}_R)}{TP(\mathcal{O}_C, \mathcal{O}_R) + TR(\mathcal{O}_C, \mathcal{O}_R)} \tag{12}$$

A higher taxonomic F-measure corresponds to a better quality of the concept hierarchy. The meaningfulness with regard to the overall quality of the ontology (lexical level + taxxonomy) depends on the chosen building blocks. If TF is not influenced by the lexical level then the taxonomic F'-measure (see [8]) may additionally be computed. It is the harmonic mean of LR and TF:

$$TF'(\mathcal{O}_C, \mathcal{O}_R) := \frac{2 \cdot LR(\mathcal{O}_C, \mathcal{O}_R) \cdot TF(\mathcal{O}_C, \mathcal{O}_R)}{LR(\mathcal{O}_C, \mathcal{O}_R) + TF(\mathcal{O}_C, \mathcal{O}_R)} \tag{13}$$

4.5 Taxonomic Overlap

In [6] and [8] the taxonomic overlap measure is defined. It is also divided into a global and a local part of the measure. The global taxonomic overlap TO has the same building blocks like TP but instead of the local taxonomic precision it uses the local overlap to:

$$to_{sc}(c_1, c_2, \mathcal{O}_1, \mathcal{O}_2) := \frac{|sc(c_1, \mathcal{O}_1) \cap sc(c_2, \mathcal{O}_2)|}{|sc(c_1, \mathcal{O}_1) \cup sc(c_2, \mathcal{O}_2)|} \tag{14}$$

Because to is a symmetric measure, it depends on the other building blocks (concept set and estimation component) whether the global taxonomic overlap is symmetric or asymmetric. We have shown the following lemma (cf. [14] for its proof):

Lemma 1. *Symmetric global taxonomic overlap measures can be solely derived from taxonomic F-measures. The equation $TO = TF/(2 - TF)$ holds.*

This lemma implies that symmetric TO measures behave like TF measures (see [14] for a symmetric TO measure). In [6] and [8] an asymmetric overlap measure is defined. There, this asymmetry is interpreted like a kind of precision and recall. But in [14] it was shown that no strictly monotonic dependency exists between that asymmetric TO measure and corresponding TP and TR measures. Thus the asymmetry can not be interpreted like precision and recall. It should be avoided to use asymmetric TO measures until the unclarity with regard to their interpretation is resolved. Instead corresponding taxonomic precision and recall measures should be used.

5 Evaluation

In this section the measures presented in 4.3 will be analytically and empirically evaluated. In the analytical evaluation it will be checked in how far they fulfill the criteria defined in section 3. Subsequently in the empirical evaluation, it will be shown in how far the choice of the measure influences the outcome of the evaluation of an ontology learning task.

5.1 Analytical Evaluation

First, it will be checked in how far the taxonomic measures are independent of the measures for the lexical term layer. This corresponds to the first criterion that a good set of measures allows an evaluation along multiple dimensions. Closely related to this criterion is the objective that each measure is independent of the other measures. The ontologies in Fig. 3 will be used for this purpose. Compared to $\mathcal{O}_{R2}$ there are three concepts missing in $\mathcal{O}_{C2}$, but the hierarchy of the remaining concepts is not changed. Also in $\mathcal{O}_{C3}$ the hierarchy is not changed but the natural language identifier of two concepts is changed (e.g. ”car” is renamed to ”auto”). Thus the hierarchy of both ontologies is perfectly learned but there are errors on the lexical term layer. This has to be reflected by taxonomy measures which are not influenced by errors on the lexical term layer.

As one can see in Tab. 4 and 5 only the measures TP_{csc} and TR_{csc} are independent of the lexical precision and recall. But this was already expected from the properties of the single building blocks of the taxonomic measures. It is more surprising to which extent the lexical precision and recall influence TP_{sc} and TR_{sc}. The errors on the lexical term layer of both learned ontologies lead to a higher decrease of the taxonomic measures than of the lexical measures. This can be seen by comparing the values of the taxonomic measures and of the lexical measures in Tab. 4. The values of the taxonomic measures are lower than the corresponding values of the lexical measures although the evaluated ontologies only contain errors on the lexical term layer.

Table 4. Evaluation of the ontologies in Fig. 3 with a semantic cotopy based measure

Compare $\mathcal{O}_{R2}$ with	LP	LR	TP_{sc}	TR_{sc}	TF_{sc}	TF'_{sc}
$\mathcal{O}_{C2}$	100.00%	57.14%	100.00%	51.02%	67.57%	61.92%
$\mathcal{O}_{C3}$	71.43%	71.43%	54.25%	54.25%	54.25%	61.67%

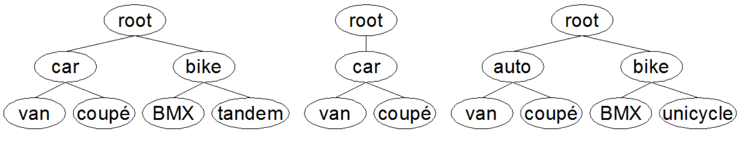

Fig. 3. Reference ontology ($\mathcal{O}_{R2}$, left) and two learned ontologies ($\mathcal{O}_{C2}$, middle; $\mathcal{O}_{C3}$, right)

Table 5. Evaluation of the ontologies in Fig. 3 with a common semantic cotopy based measure

Compare $\mathcal{O}_{R2}$ with	LP	LR	TP_{csc}	TR_{csc}	TF_{csc}	TF'_{csc}
$\mathcal{O}_{C2}$	100.00%	57.14%	100.00%	100.00%	100.00%	72.73%
$\mathcal{O}_{C3}$	71.43%	71.43%	100.00%	100.00%	100.00%	83.33%

The second criterion of good evaluation measures was that the effect of an error onto the measure should be proportional to the distance between the correct and the given result. This criterion will be checked with the ontologies in Fig. 4. There, in $\mathcal{O}_{C4}$, the two concepts "car" and "bike" are interchanged, corresponding to an error near the root of the hierarchy. In $\mathcal{O}_{C5}$ the two leaf concepts "coupé" and "BMX" are interchanged. Altogether the errors in $\mathcal{O}_{C4}$ are more serious than the errors in $\mathcal{O}_{C5}$. Thus measures which fulfill this second criterion should rate $\mathcal{O}_{C4}$ worse than $\mathcal{O}_{C5}$. In Tab. 6 and 7 one can see that both pairs of measures fulfill this criterion.

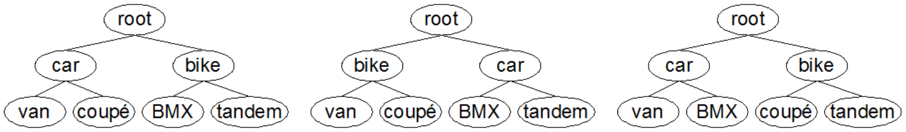

Fig. 4. Reference ontology ($\mathcal{O}_{R3}$, left) and two learned ontologies ($\mathcal{O}_{C4}$, middle; $\mathcal{O}_{C5}$, right)

Table 6. Evaluation of the ontologies in Fig. 4 with a semantic cotopy based measure

Compare $\mathcal{O}_{R3}$ with	LP	LR	TP_{sc}	TR_{sc}	TF_{sc}	TF'_{sc}
$\mathcal{O}_{C4}$	100.00%	100.00%	66.67%	66.67%	66.67%	80.00%
$\mathcal{O}_{C5}$	100.00%	100.00%	83.33%	83.33%	83.33%	90.91%

The third and last criterion of good evaluation measures was that a gradual increase in the error rate should lead to a more or less gradual decrease in the evaluation results. One can see from the previously given examples that TP_{csc} and TR_{csc} fulfill this criterion. Especially for the ontologies in Fig. 3 it returned perfect evaluation results. The opposite is true for TP_{sc} and TR_{sc}: Because these measures are influenced by errors in the lexical term layer as well as by errors in the concept hierarchy they will drop very fast if both kinds of errors occur in an ontology. Additionally it was shown that they are stronger influenced by errors in the lexical term layer than the lexical precision and recall measure itself.

TP_{csc} and TR_{csc} are all in all better suited for evaluating a concept hierarchy and drawing conclusions about the strengths and weaknesses of the used learning procedure.

Table 7. Evaluation of the ontologies in Fig. 4 with a common semantic cotopy based measure

Compare $\mathcal{O}_{R3}$ with	LP	LR	TP_{csc}	TR_{csc}	TF_{csc}	TF'_{csc}
$\mathcal{O}_{C4}$	100.00%	100.00%	52.38%	52.38%	52.38%	68.75%
$\mathcal{O}_{C5}$	100.00%	100.00%	76.19%	76.19%	76.19%	84.49%

5.2 Empirical Evaluation

In this section the previously described measures will be used in a real evaluation of concept hierarchies learned with Hearst patterns (cf. [15], [1]). In this evaluation it will be shown in how far the choice of the measure influences the lessons learned from evaluating an ontology learning task. For this evaluation, several ontologies for the tourism domain were learned from a corpus of 4596 tourism related Wikipedia articles with 6.54 million tokens. The reference ontology was created by an experienced ontology engineer within the GETESS project (see [16] and Tab. 9 for more details about the ontology). A more detailed description of the experiment and further results for ontologies learned with other learning procedures and from other document corpora are available for download [14].

Tab. 8 and 10 contain the evaluation results for the ontologies learned with the Hearst patterns applied on the Wikipedia corpus. The learned ontologies were compared with the GETESS reference ontology. These raw evaluation results should now be used for deciding for which threshold the best results were achieved. Both tables contain the results for the same ontologies but evaluated with the two different measures from section 4.3.

Table 8. Evaluation of learned ontologies with TP_{csc} depending on threshold θ

θ	LP	LR	TP_{csc}	TR_{csc}	TF_{csc}	TF'_{csc}
0.0	1.00%	49.66%	22.26%	83.81%	35.18%	41.18%
0.3	7.27%	22.79%	81.01%	59.60%	68.67%	34.22%
0.6	12.09%	11.22%	83.08%	62.11%	71.08%	19.39%
0.9	17.04%	7.82%	84.06%	73.85%	78.62%	14.23%

Looking at the results in Tab. 8 one can see that there is a major improvement of the taxonomic precision if the threshold is increased from 0.0 to 0.3. But this improvement on the taxonomic layer of the ontology is accompanied by a decrease of the lexical recall. According to the TF'_{csc} one would judge the unfiltered ontology better. But from the low lexical and taxonomic precision of the unfiltered ontology one may also conclude that it more or less ”accidentally” contains correct lexical entries and taxonomic relations. So after a deeper analysis of the evaluation results one may come to the conclusion that a moderate filtering based on the confidence value should be applied.

This conclusion based on the results in Tab. 8 are also supported by the ontology’s additional statistical values in Tab. 9. The first row of the table contains the values of

Table 9. Additional statistical values for the reference and the learned ontologies.

θ	concepts	loops	avg. depth	avg. sub	sub. dev.	avg. super	super dev.
ref	294	1	5.14	5.22	4.42	1.03	0.17
0.0	14569	4973	119.29	3.57	53.2	1.52	2.2
0.3	893	97	3.8	2.81	14.89	1.22	0.87
0.6	246	24	3.29	2.68	8.39	1.16	0.78
0.9	116	2	3.17	2.76	6.06	1.08	0.35

the reference ontology against which the learned ontologies are compared. The following rows contain the statistical values of the learned ontologies. One can see that the unfiltered concept hierarchy contains 4,973 loops (i.e. a concept is also one of its superconcepts) and that a leaf concept has 119 superconcepts in average. Additionally, it is interesting to look at the branching factor of the hierarchy. There one can see that a concept has 3.57 direct subconcepts in average with a very high deviation of 53.2. Also the average number of direct superconcepts is quite high with 1.52 and a deviation of 2.2. All these statistical values show that the unfiltered ontology is more or less degenerated. Compared to these results the statistical values of the filtered ontologies are much better.

This exemplary evaluation with TP_{csc} and TR_{csc} shows that they allow to make conclusions about the real problems of a learned ontology and subsequently to identify the best parameters for optimizing the used learning procedure. It is now the question whether an evaluation with TP_{sc} and TR_{sc} leads to the same conclusions.

Looking at the evaluation results in Tab. 10 one may also draw the conclusion that a moderate filtering of the learned lexical entries and taxonomic relations improves the results because the best TF'_{sc} value is achieved for the ontology filtered with a threshold of 0.3. But it is not clear in how far this improvement is only caused by the changes on the lexical level (especially the improvement of the lexical precision) because the improvement of the taxonomy is superposed by the influence of lexical precision and recall on TP_{sc} and TR_{sc}. Thus, a truly multidimensional evaluation of the learned ontologies is impossible because the used measures are not independent of each other.

Table 10. Evaluation of learned ontologies with TP_{sc} depending on threshold θ

θ	LP	LR	TP_{sc}	TR_{sc}	TF_{sc}	TF'_{sc}
0.0	1.00%	49.66%	0.10%	27.84%	0.21%	0.41%
0.3	7.27%	22.79%	3.23%	8.67%	4.71%	7.80%
0.6	12.09%	11.22%	6.44%	3.61%	4.63%	6.55%
0.9	17.04%	7.82%	10.40%	2.53%	4.07%	5.35%

6 Conclusions

This paper presented a framework useful for gold standard based evaluation of ontologies. It was used for creating a new measure which allows to do a multi dimensional evaluation. Furthermore, it was ensured that errors are weighted differently based on their position in the concept hierarchy and that, compared to existing measures, the scale interval of the measure is used more evenly.

Acknowledgments

This work has been supported by the european projects Lifecycle Support for Networked Ontologies (NeOn, IST-2006-027595) and Semiotic Dynamics in Online Social Communities (Tagora, FP6-2005-34721).

References

1. Cimiano, P., Pivk, A., Schmidt-Thieme, L., Staab, S.: Learning taxonomic relations from heterogenous sources of evidence. In: Ontology Learning from Text: Methods, Applications, Evaluation. IOS Press (2005)
2. Spyns, P., Reinberger, M.L.: Lexically evaluating ontology triples generated automatically from texts. In: Proc. of the second European Conference on the Semantic Web. (2005)
3. Sabou, M., Wroe, C., Goble, C., Stuckenschmidt, H.: Learning domain ontologies for semantic web service descriptions. Journal of Web Semantics **3**(4) (2005)
4. Sabou, M., Wroe, C., Goble, C., Mishne, G.: Learning domain ontologies for web service descriptions: an experiment in bioinformatics. In: Proc. of WWW05. (2005)
5. Reinberger, M.L., Spyns, P.: Unsupervised text mining for the learning of dogma-inspired ontologies. (In: Ontology Learning from Text: Methods, Applications and Evaluation)
6. Maedche, A.: Ontology Learning for the Semantic Web. Kluwer, Boston (2002)
7. Maedche, A., Staab, S.: Measuring similarity between ontologies. In: Proc. of the European Conference on Knowledge Acquisition and Management (EKAW-2002). (2002)
8. Cimiano, P., Hotho, A., Staab, S.: Learning concept hierarchies from text corpora using formal concept analysis. JAIR – Journal of AI Research **24** (2005) 305–339
9. Maynard, D., Peters, W., Li, Y.: Metrics for evaluation of ontology-based information extraction. In: Proc. of the EON 2006 Workshop. (2006)
10. Hahn, U., Schnattinger, K.: Towards text knowledge engineering. In: Proc. of the 15th National Conference on Artificial Intelligence (AAAI-98). (1998)
11. Brank, J., Mladenic, D., Grobelnik, M.: Gold standard based ontology evaluation using instance assignment. In: Proc. of the EON 2006 Workshop. (2006)
12. Hartmann, J., Spyns, P., Maynard, D., Cuel, R., Carmen Suarez de Figueroa, M., Sure, Y.: Methods for ontology evaluation. Deliverable D1.2.3, Knowledge Web (2004)
13. van Rijsbergen, C.: Information Retrieval. Butterworths, London (1979)
14. Dellschaft, K.: Measuring the similiarity of concept hierarchies and its influence on the evaluation of learning procedures. Diploma thesis, Universität Koblenz-Landau (2005) http://www.uni-koblenz.de/FB4/Institutes/IFI/AGStaab/Theses/2005/DADellschaft.pdf.
15. Hearst, M.: Automatic aquisition of hyponyms from large text corpora. In: Proc. of the 14th International Conference on Computational Linguistics. (1992)
16. Staab, S., et al.: Getess - searching the web exploiting german texts. In: Proc. of the 3rd Workshop on Cooperative Information Agents. (1999)