Strategies for the Evaluation of Ontology Learning

Klaas DELLSCHAFT [a] and Steffen STAAB [a]
[a] *Universität Koblenz-Landau, Koblenz, Germany*

Abstract. An important aspect of ontology learning is a proper evaluation. Generally, one can distinguish between two scenarios: (i) quality assurance during an ontology engineering project in which also ontology learning techniques may be used and (ii) evaluating and comparing ontology learning algorithms in the laboratory during their development. This paper gives an overview of different evaluation approaches and matches them against the requirements of the scenarios. It will be shown that different evaluation approaches have to be applied depending on the scenario. Special attention will be paid to the second scenario and the gold standard based evaluation of ontology learning for which concrete measures for the lexical and taxonomic layer will be presented.

Keywords. ontology learning, evaluation

1. Introduction

When it comes to the evaluation of ontology learning one has to distinguish between two different scenarios: On the one hand there is the scenario where an ontology learning algorithm is used in the context of an automatic or semi-automatic approach to ontology engineering (cf. [1] and [2]). On the other hand there is the evaluation of the ontology learning algorithm itself (cf. [3] and [4]). Both scenarios differ in their requirements with regard to their evaluation thus leading to different evaluation approaches.

In Fig. 1 one can see the general approach of ontology learning: It takes as input a domain centered corpus and tries to learn an ontology which conceptualizes the information implicitly available in the corpus. Depending on the scenario, different aspects of ontology learning have to be evaluated. In the first scenario not only the learning algorithm influences the results but another important aspect is the choice of the correct corpus which has to contain information relevant for the task. In the second scenario one is only interested in the quality of the learning algorithm itself.

We argue in this paper that in the first scenario the functional dimension of an ontology should be evaluated by means of an extrinsic, task-based evaluation, i. e. in the running application for which the ontology is engineered. This kind of evaluation ensures that the objective of using an ontology, improving a certain task, is really achieved. But this approach is not feasible in the second scenario where the aim is to compare ontology learning algorithms. Its objective is the assessment of the quality of different learning algorithms. In this scenario, an intrinsic or task-neutral evaluation by means of a gold-standard based evaluation is usually the better choice.

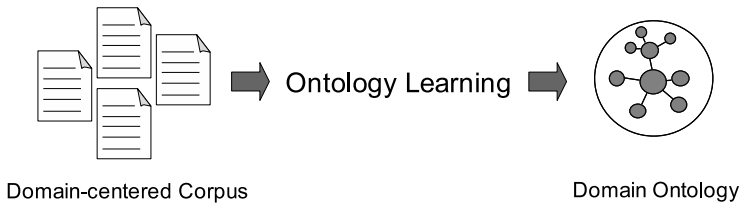

Figure 1. General Approach of Ontology Learning

After describing the two scenarios and their requirements with regard to the evaluation in more detail we will focus in the remainder of this paper on the second scenario and especially on the gold standard based evaluation of the lexical and taxonomic layer of ontologies. It will be shown that existing measures have been faulty and that a well-founded evaluation model is largely missing. Therefore, we describe a new framework for gold standard-based evaluation of ontology learning which includes measures for the lexical and the taxonomic layer of an ontology. The framework avoids common mistakes and we show by analytical considerations and by some experiments that it fulfills crucial evaluation criteria that other frameworks do not meet.

2. Evaluation Scenarios and Approaches

In the following, we will present the two scenarios in more detail and list several applicable evaluation approaches. We will distinguish between approaches which try to measure the functional dimension of a learned ontology and the ones which measure the structural dimension (cf. [5]). The functional dimension of an ontology is related to its conceptualization while the structural dimension is related to the representation of an ontology as a graph. It will be shown that, depending on the scenario, different evaluation approaches for the functional dimension should be used.

2.1. Scenario 1: Quality Assurance During Ontology Engineering

In this scenario ontologies are evaluated during the ontology engineering process as part of the quality assurance. Typical questions for evaluating an ontology are whether it is consistent, complete, concise and expandable (see [6]). For this purpose, in [5] and [7] it is proposed to measure the structural and functional dimension of ontologies and their usability profile. The requirements which should be fulfilled by an ontology with regard to the dimensions will be usually defined during the start phase of an ontology engineering project.

For example, one may check whether the target domain of the ontology is sufficiently modeled to fulfill the functional requirements and/or whether the ontology helps to improve the performance in the task for which it is designed. As a consequence of such an evaluation one may e. g. decide to further extend some aspects of the ontology. During the structural evaluation it is checked whether certain criteria are fulfilled which are related to the design principles of good ontologies and which help to improve an ontology's overall quality. In [5] it is additionally proposed to evaluate the usability pro-

file of an ontology. During such an evaluation the quantity and quality of the ontology's metadata is checked which address the communication context of an ontology.

In this scenario, the evaluation of ontologies is seen as an important part of the quality assurance process. In many cases the ontologies will be manually engineered. But in case of semi-automatic approaches to ontology engineering (cf. [1] and [2]) also the output of ontology learning algorithms is contained in the engineered ontology and it is thus also evaluated during the quality assurance process.

The following approaches to a functional and structural evaluation can be identified in this scenario:

Task-based Approaches Task-based approaches try to measure in how far an ontology helps to improve the results of a certain task. They will usually measure the functional dimension of an ontology but also the structural dimension may influence the outcome of task-based evaluations. For example, if one designs an ontology for improving the performance of a web search engine (cf. [8]) one may collect several example queries and compare whether the search results contain more relevant documents if a certain ontology is used. A task-based evaluation is influenced by many aspects which have to be kept constant during all evaluations so that changes in the results can be put down to the changes in the used ontologies. The choice of concrete measures for such an evaluation is dependent on the task, e. g. for the web search engine example one may adapt measures known from information retrieval but also other success criteria may be defined.

Because every task-based evaluation is individual, no finite set of well-suited measures can be defined. Nevertheless, some principles can be identified: Usually, it is not enough to know whether an ontology is better or worse than another but one wants to conclude on concrete shortcomings in its conceptualization. Thus, in [9] it is demanded that a task-based evaluation allows for concluding on insertion, deletion and substitution errors in the ontology, i. e. whether there are superfluous, missing or off-target concepts and/or relations. But again, there is no universally valid way how the principles can be realized in a concrete task-based evaluation. In [9] it is only demonstrated for one example task.

Corpus-based Approaches Corpus-based approaches are used for checking in how far an ontology sufficiently covers a given domain. They address the functional dimension of an ontology. For this purpose, the ontology is compared with the content of a text corpus which is representative for the domain. The content of the corpus is analyzed with natural language techniques, e. g. in [10] Latent Semantic Analysis and a clustering method were applied for identifying terms in the corpus. The list of identified terms was then compared with the terms in the evaluated ontology. Similar approaches for evaluating the lexical layer of an ontology are described in [11] and [12] while [13] contains a preliminary method applicable for evaluating triples in ontologies.

All the corpus-based approaches have in common that they involve information extraction and/or ontology learning techniques in the evaluation. Thus they are only partially suitable for evaluating ontologies which were learned with other ontology learning algorithms because the information extraction and/or ontology learning techniques from the corpus-based approaches are like a benchmark which can not be outperformed by other ontology learning algorithms. In [11] it is proposed to

evaluate and extend ontologies at the same time with such an approach, e. g. by suggesting terms which are currently missing in the ontology and which would improve the evaluation results.

Criteria-based Approaches In this category fall a wide variety of evaluation measures which all have in common that they measure in how far an ontology or taxonomy adheres to certain desirable criteria. One can distinguish between measures related to the structure of an ontology, e. g. if it is represented as a graph, and more sophisticated measures which e. g. evaluate a taxonomy based on philosophical notions.

Structural measures are quite straightforward and easy to understand: For example, one may measure the average depth of paths from root to leaf nodes in a directed graph, how many nodes have more than one ingoing arc (i. e. multi-hierarchical nodes) or whether there are cycles in the directed graph (cf. [5] and [7]). But also for ontologies based on frame logic or description logic one may define structural measures, e. g. for detecting potential inconsistencies in the partitioning of a taxonomy (cf. [6]). Such a partitioning error measure may for example find instances belonging to more than one class where two or more of the classes are defined as disjoint.

For the structural measures it is usually no problem to have a fully automatic evaluation. This is not the case for the more sophisticated measures like OntoClean [14] which evaluates taxonomies based on philosophical notions like the essence, identity and unity which should be taken into account during modeling an ontology in order to avoid common pitfalls. For example, a property is essential for an entity if it holds for that entity in every possible world. Furthermore, a property is rigid if it is essential for all its possible instances. In [14], this is explained with the example relations *having a brain* and *being a student*. In this example, the *having a brain* relation is essential for all human beings thus it is a rigid property. In contrast the *being a student* relation, which is not essential for any human being as everyone can become a student or cease to be a student at any time. Thus it would be an anti-rigid property. (For more examples and explanations see [14].) Because of this high complexity, OntoClean is designed for manually analyzing ontologies although an approach for partially automating this process was recently proposed (see [15]).

The most important success criterion for an ontology engineering project is whether the final ontology helps to improve the task for which it was engineered. Thus, improving the results during a task-based evaluation can be seen as the most important goal. Corpus-based and criteria-based evaluation approaches only help to pinpoint the remaining problems which should be addressed in an improved version of the ontology. The main assumption behind corpus-based and criteria-based evaluation measures is that an improvement with regard to the measures correlates with an improvement in the task-based evaluation (see [8] where the correlation was shown for OntoClean [14]).

2.2. Scenario 2: Comparing Ontology Learning Algorithms

In this scenario one tries to assess and compare ontology learning algorithms with each other. It can be used by researchers to improve an existing learning algorithm or to find out how changing the values of input parameters affects the results. An example how

such an evaluation may look like is available in [4]. It is the goal in this scenario to measure the quality of an ontology learning algorithm. This should ideally be done by looking at the output (i. e. the learned ontology) and comparing it with the input (i. e. the content of the corpus). As we will see below, there basically exist two approaches how one can approximate the comparison with the input by either making a manual evaluation by human experts or a gold-standard based evaluation where the gold-standard covers the content of the corpus.

With regard to the functional dimension of the learned ontology one is interested in measuring in how far the learning algorithm is able to conceptualize the information from a given corpus (e. g. whether it extracts isA-relations between relevant concepts) and which fraction is found. This corresponds to measuring the precision and recall (see 4.1). But also from evaluating the structural dimension of the learned ontology one may draw interesting conclusions on the qualities of a learning algorithm.

In the following, two approaches to measuring the functional dimension will be presented which are specific for the needs in this scenario and which are different to the approaches from the first scenario. In contrast, it is possible to re-use a subset of the structural measures described in 2.1. Thus, we will concentrate here on the evaluation of the functional dimension.

In the previous scenario, a task-based evaluation was considered as ideal for evaluating the functional dimension of an ontology. This is not the case for the evaluation and comparison of ontology learning algorithms. Here, it would be necessary to filter out the influence of the task on the evaluation results in order to make valid conclusions on the strength and weaknesses of the learning algorithm itself. For example, the results of a task-based evaluation would be influenced by many other factors like the choice of the corpus, the task itself or the algorithm used for performing the task. Additionally, it is very difficult to conclude from the results of a task-based evaluation on the concrete precision and recall values achieved by the learning algorithm. Instead, it would be valuable to have a more direct approach to measuring those dimensions of interest. All in all, the following list of criteria should be fulfilled by an evaluation in this scenario:

- The evaluation should be task neutral and allow developers to easily pinpoint the advantages and disadvantages of a learning algorithm. Weighing the different advantages and disadvantages of a learning algorithm is then up to the ontology engineer who has a concrete task in mind. This weighing can be based on his experience or even on a task-based evaluation where it was shown that certain aspects of an ontology are more important than others.
- All influencing factors of the evaluation have to be sufficiently described so that its results can be reproduced at another time and place. This is important for having a proper scientific evaluation in general and also applies for other approaches and scenarios like task-based evaluation approaches.
- It should be possible to do additional evaluation runs at low cost because frequent and large-scale evaluations are required during developing ontology learning algorithms, e. g. in order to find the parameter values of the learning algorithm for which the best results are achieved. It has to be ensured that all evaluation runs are performed under the same conditions in order to have comparable results.

By looking at the literature, one can identify the following two approaches for measuring the functional dimension in this scenario:

Manual Evaluation by Human Experts This evaluation approach can be found in several papers about ontology learning algorithms like in [16] and [17] where the learned ontology is presented to one or more human experts which have to judge in how far the extracted information is correct (i. e. the precision is measured). But the approach has several downsides: First of all, the extracted information is not compared with the information found in the corpus but with the knowledge of the human expert. While this is not so problematic for measuring the precision of the learning algorithm it makes a reliable measurement of the recall nearly impossible. Furthermore, the most important influencing factor of the evaluation is the choice of the human experts. Because they may not be available at another time and place the last two criteria outlined before are not fulfilled. This problem can only be avoided by asking a sufficiently large number of experts. Additionally, every evaluation run comes with the same high costs as the first run thus making frequent and large-scale evaluations unfeasible.

Gold Standard Based Approaches Gold standard based approaches compare the learned ontology with a previously created gold standard which represents an idealized outcome of the learning algorithm. A learning algorithm is considered to be better when the learned ontology has a high similarity with the gold standard. Examples for this kind of evaluation can be found in papers like [3], [12] and [18]. The gold standard based evaluation fulfills all the criteria from above: It can be used for directly measuring the precision and recall of the learned ontology compared to the gold standard. Furthermore, the evaluation results can be reproduced and are comparable if the same corpus, learning algorithm and gold standard are used. Additionally, only for the first run of the evaluation the high costs of creating the gold standard exist. Subsequent runs of the evaluation are then fully automatic.

Although the gold standard based evaluation seems to be ideal in this scenario there remains one big issue: Where to get or how to create such a gold standard? On the one hand, one may ask a human expert to create a gold standard based on the information in the used corpus. Depending on the size of the corpus, this can constitute a very work intensive approach. Another approach might be to take an already existing ontology and choose the corpus accordingly so that it can be assumed that most of the information of the gold standard is available in the corpus. An example of the latter approach is available in [4].

Independent from this decision, the term “gold standard” may be misleading as there exists not only one gold standard but, depending on who is asked, one may get several gold standards which differ in their details. This is due to the different conceptualization humans may have of a domain (cf. [5]). The same problem exists for the manual evaluation by human experts. There it is typically addressed by measuring the consensus between several experts (cf. [19]). A similar way may be used for the creation of the gold standard. For example, one may involve several experts in the creation of the gold standard and measure their consensus or one may compare with several gold standards (and measuring the agreement between those gold standards). But regardless of this decision, the main advantage of gold standard based evaluation remains that the conceptualizations of the experts become explicitly available in form of the gold standard. This ensures that every learning algorithm is compared against the same standard and that everyone can control how thoroughly the gold standard was created.

There exist many measures for the gold standard based evaluation of ontologies. They can be distinguished between measures which only evaluate the lexical layer of an ontology, the ones which also take the concept hierarchy or taxonomic layer 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 the lexical and the taxonomic layer.

On the lexical 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 [18]. There exist several other names for these measures like *Lexical Precision and Recall* or simply *precision and recall* (see [20] and [21]). Another example of a lexical evaluation measure is the *String Matching* measure presented in [22] and [19]. 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 or taxonomies is more complicated than the comparison of the lexical 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.

Furthermore, we have to distinguish between different learning approaches. An example for such an approach is the *General Named Entity Identification* (GNE) where the algorithm has to find for previously unknown concepts their maximally specific generalization from a given ontology, i. e. it adds them as leaf nodes to the ontology. Examples of measures suitable for evaluating GNE algorithms are available in [23]. They partially depend on the assumption that the compared ontologies only differ in their leaf nodes.

But in this paper we will concentrate on another, more general approach where ontologies are learned from scratch, i. e. without a seed ontology which is extended. In the case of concept hierarchies, it leads to the fact that not only the positions of leaf nodes may differ between the learned and the reference hierarchy but also the position of inner concepts. Thus, the evaluation measures can not depend on the assumption that large portions of the two compared hierarchies (i. e. the seed hierarchy) match exactly.

One of the first examples of such a concept hierarchy evaluation measure is the *Taxonomic Overlap* (TO) presented in [22] and [19]. 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 [22], [19] and [4], i. e. they can be computed into two directions. In [4] 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 [24] and [25]. 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 [26]. 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 [24]).

Table 1. Rating of concept hierarchy measures

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

The latest measure for comparing concept hierarchies is the *OntoRand* index proposed in [27]. 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 doing a gold standard based evaluation of 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 misinterpreting evaluation results and ease drawing the right conclusions for the improvement of the evaluated ontology learning algorithm.

The most important criterion is that an ontology is evaluated along **multiple dimensions**. This criterion is formulated in several papers like [24] and [28]. But instead of having a measure which aggregates the evaluation of all those dimensions into a single value one should use separate measures for each of the dimensions. 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.

As we will show in 5, it is very important that a measure is only influenced by exactly one dimension and/or type of error. For example, if one uses measures for evaluating the lexical 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 [28]).

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 [27]).

In Tab. 1 it is shown in how far the measures for the functional dimension described in section 2.2 meet the criteria listed in this section. The rating is based on the descriptions in [19], [24] and [27]. 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 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 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 considers the distance between the correct and the given answer only to some small extent (see [24]). 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 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 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 layer and the concept hierarchy. Similarly to [4], 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 terms on the lexical layer 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 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 terms exists (in analogy to [22] and [19]).

4.1. Precision & Recall

This section gives a short overview of precision, recall and F-measure, as they are known from information retrieval (see [29]). They are used for comparing a reference retrieval

[1] In [27] two different variants of OntoRand are presented. One variant is based on a tree distance while the other is based on finding the common ancestor in the concept hierarchy. For the comparison in Tab. 1 the measure based on the common ancestors was used because it was shown in [27] that it is superior to the tree distance based measure.

(Ref) with a computed retrieval ($Comp$) returned by a system. Precision and recall are defined as follows:

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

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

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

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.

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

4.2. Lexical Precision & Recall

There exist several measures sufficient for evaluating the lexical layer of an ontology (see section 2.2). In this subsection the lexical precision and recall measures, as they are described in [20], 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:

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

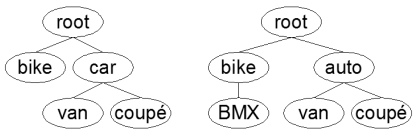

Figure 2. Example reference ontology ($\mathcal{O}_{R1}$, left) and computed ontology ($\mathcal{O}_{C1}$, right)

The lexical precision and recall reflect how good the learned terms cover the target domain. For example, if one compares $\mathcal{O}_{C1}$ and $\mathcal{O}_{R1}$ in Fig. 2 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 [19]. 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).

4.3.1. 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 2.2). The local measure compares the positions of two concepts and the global measure is used for comparing two whole concept hierarchies. We start with 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 extracts from the concept hierarchy, which are characteristic for the position of a concept in the hierarchy. That is,. 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 a 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 [19] 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. 2). 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 layer of an ontology.

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

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. 2.

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}

Besides the previously described extracts of the concept hierarchy, further extracts are imaginable. For example, the upwards cotopy (see [19]) or the set of all direct subconcepts might be used. In [30] also measures based on the direct subconcepts were evaluated. But [30] shows also that measures based on the semantic cotopy meet more of the criteria from section 3.

4.3.2. Comparing Concept Hierarchies

It is now possible to define a framework for constructing a global taxonomic precision measure. Fig. 3 shows the building blocks used in this framework for a global taxonomic precision measure.

$$TP(\mathcal{O}_C, \mathcal{O}_R) := \frac{1}{|\mathcal{C}_C|} \sum_{c \in \mathcal{C}_C} \begin{cases} tp(c, c, \mathcal{O}_C, \mathcal{O}_R) & \text{if } c \in \mathcal{C}_R \\ \max_{c' \notin \mathcal{C}_R} tp(c, c', \mathcal{O}_C, \mathcal{O}_R) & \text{if } c \notin \mathcal{C}_R \end{cases}$$

local taxonomic precision

concept set

estimation

Figure 3. 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 [19] 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 layer (e. g. "auto" and "car" in Fig. 2) 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 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.

4.3.3. 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 [30] 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 layer. In the evaluation in section 5 they will be used for demonstrating the disadvantages of mixing the evaluation of lexical 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 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 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 [4]) 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 [22] and [4] 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. [30] 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 [30] for a symmetric TO measure). In [22] and [4] an asymmetric overlap measure is defined. There, this asymmetry is interpreted like a kind of precision and recall. But in [30] 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.

Table 4. Evaluation of the ontologies in Fig. 4 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%

Table 5. Evaluation of the ontologies in Fig. 4 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%

5. Evaluation

In this section the measures presented in 4.3.3 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 layer. This corresponds to the first criterion that a good set of measures allows for evaluating along multiple dimensions. Closely related to this criterion is the objective that each measure is independent of the other measures. The ontologies in Fig. 4 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 layer. This has to be reflected by taxonomy measures which are not influenced by errors on the lexical 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 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 layer.

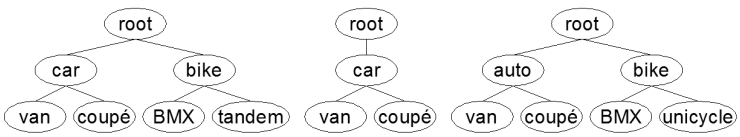

Figure 4. Reference ontology ($\mathcal{O}_{R2}$, left) and two learned ontologies ($\mathcal{O}_{C2}$, middle; $\mathcal{O}_{C3}$, right)

Table 6. Evaluation of the ontologies in Fig. 5 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%

Table 7. Evaluation of the ontologies in Fig. 5 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%

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. 5. 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.

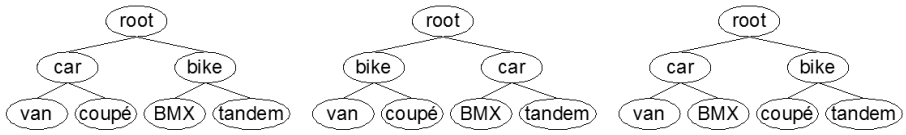

Figure 5. Reference ontology ($\mathcal{O}_{R3}$, left) and two learned ontologies ($\mathcal{O}_{C4}$, middle; $\mathcal{O}_{C5}$, right)

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. 4 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 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 more strongly influenced by errors in the lexical 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.

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. [31], [3]). In this evaluation it will be shown in how far the choice of the measure influences the nature of the results and subsequently the conclusions which are drawn from the evaluation of a learning algorithm. For the 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 [32] and Tab. 8 for more details about the ontology).

If the Hearst patterns are applied on a collection of texts, it is very likely that the same relation is extracted more than once. This information can be used for defining a confidence value in the extracted relation. The confidence is increased, with the number of occurrences. The most often extracted relation gets a confidence value of 1.0. It drops to 0.0 with descending occurrences. Four different thresholds θ were applied to the confidence value for filtering the taxonomic relations. For more details on the experiment and further results for other learning algorithms and document corpora see [30].

In Fig. 6, 7 and 8 one can see the evaluation results for the taxonomic and the lexical layer of the learned ontologies. These raw evaluation results should now be used for deciding for which threshold the best results were achieved. Fig. 7 and 8 contain the evaluation of the taxonomic layer of the same ontologies but evaluated with the two different measures from section 4.3.3.

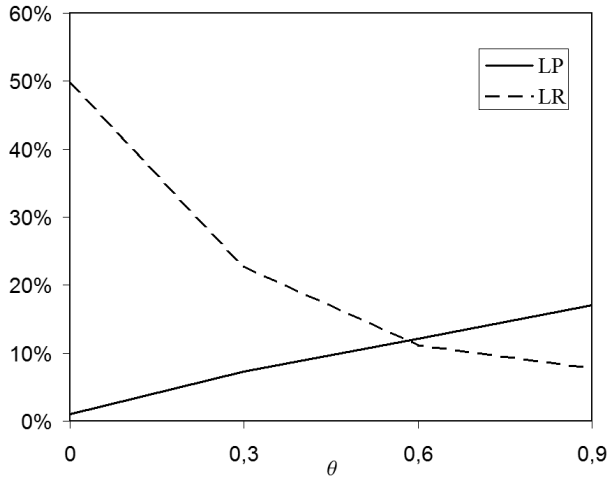

Figure 6. Evaluation of the lexical layer depending on threshold θ

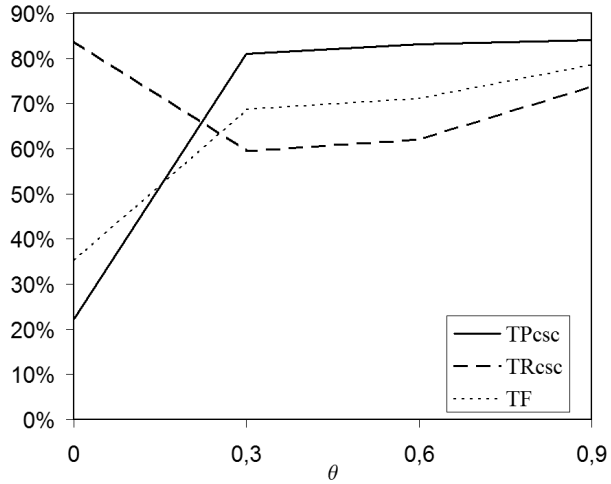

Figure 7. Evaluation of learned ontologies with TP_{csc} depending on threshold θ

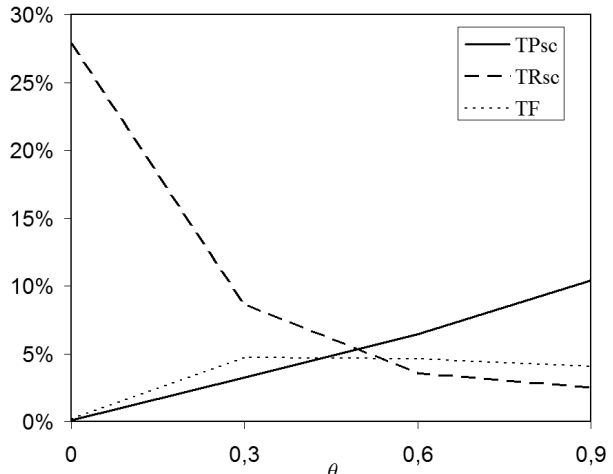

Figure 8. Evaluation of learned ontologies with TP_{sc} depending on threshold θ

Table 8. Structural evaluation of the reference ontology and the learned ontologies

θ	concepts	circles	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

Looking at the results in Fig. 7 one can see that there is a major improvement of the quality on the taxonomic layer if θ is increased from 0.0 to 0.3. But this improvement on the taxonomic layer is accompanied by a decrease of the lexical recall (see Fig. 6) thus it isn't so clear whether the ontologies with $\theta = 0.0$ or 0.3 are better. But from the low lexical and taxonomic precision for a $\theta = 0.0$ one may also conclude that this ontology more or less "accidentally" contains correct terms and taxonomic relations (which lead to the high recall values). So after a deeper analysis of the evaluation results one may come to the conclusion that learning taxonomic relations with Hearst patterns works best if the output ontology is moderately filtered based on the threshold values.

The conclusion based on the functional evaluation of the lexical and taxonomic layer is also supported by the structural evaluation in Tab. 8. The first row of the table contains the values of the reference ontology against which the learned ontologies are compared. The following rows contain the values of the learned ontologies. One can see that the unfiltered concept hierarchy contains 4,973 circularity errors in the concept hierarchy (i. e. a concept is also one of its superconcepts) and that the average cardinality of the paths from the root to the leaf nodes (i. e. the average depth of the hierarchy) is 119. Additionally, it is interesting to look at the branching factor of the hierarchy: The concepts have 3.57 direct subconcepts in average with a very high deviation of 53.2. The average number of direct superconcepts is also quite high with 1.52 and a deviation of 2.2 (i. e. there exist many multi-hierarchical concepts). All these structural measures show that the hierarchy of the unfiltered ontology is more or less degenerated while the values for the ontology with $\theta = 0.3$ are close to the values of the manually built reference ontology.

The exemplary evaluation with TP_{csc} and TR_{csc} shows that they allow for separately evaluating the taxonomic and lexical layer of an ontology. The different evaluation measures have to be weighed and prioritized thus forming an overall picture of the advantages and disadvantages of the ontologies and thus the used learning algorithm.

The separate evaluation of the functional dimension of the taxonomic and lexical layer is not possible if TP_{sc} and TR_{sc} are used instead. In constructing the measures in section 4.3 as well as in the analytical evaluation in 5.1 it was predicted that there is a strong dependency of TP_{sc} and TR_{sc} on the respective measure from the lexical layer. This dependency also becomes obvious by comparing Fig. 6 and Fig. 8. Both graphs show more or less the same information, i. e. the evaluation of the taxonomic layer is superimposed by the influence of the lexical layer. Thus drawing conclusions about the taxonomic layer and making a truly multidimensional evaluation is impossible because the used measures are not independent of each other.

6. Conclusions

In this chapter we presented an overview of several existing approaches to the evaluation of ontologies. It was shown that in the scenario of evaluating ontology learning algorithms a gold standard based evaluation approach is the best choice while for the quality assurance during an ontology engineering project a combination of task-, corpus- and criteria-based evaluation approaches should be used.

Focusing on the scenario of evaluating ontology learning algorithms, we presented a framework for gold standard based evaluations. It was used for creating a measure for the taxonomic layer. It was shown by means of an analytical and empirical evaluation that it fulfills the three basic criteria for gold standard based evaluations: (i) allowing for evaluating along multiple dimensions, (ii) taking the distance between correct and given answer into account and (iii) 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] J.-U. Kietz, A. Maedche, and R. Volz, "A method for semi-automatic ontology acquisition from a corporate intranet," in *Proc. of the EKAW'2000 Workshop "Ontologies and Texts"*, 2000.

[2] E. Simperl, C. Tempich, and D. Vrandecic, "A methodology for ontology learning," in *Bridging the Gap between Text and Knowledge – Selected Contributions to Ontology Learning and Population from Text* (P. Buitelaar and P. Cimiano, eds.), IOS Press, 2007. THIS VOLUME.

[3] P. Cimiano, A. Pivk, L. Schmidt-Thieme, and S. Staab, "Learning taxonomic relations from heterogenous sources of evidence," in *Ontology Learning from Text: Methods, Applications, Evaluation* (P. Buitelaar, P. Cimiano, and B. Magnini, eds.), Amsterdam: IOS Press, 2005.

[4] P. Cimiano, A. Hotho, and S. Staab, "Learning concept hierarchies from text corpora using formal concept analysis," *JAIR – Journal of AI Research*, vol. 24, pp. 305–339, 2005.

[5] A. Gangemi, C. Catenacci, M. Ciaramita, and J. Lehmann, "Modelling ontology evaluation and validation," in *Proceedings of the 3rd European Semantic Web Conference (ESWC)*, 2006.
[6] A. Gómez-Pérez, "Ontology evaluation," in *Handbook on Ontologies* (S. Staab and R. Studer, eds.), Heidelberg: Springer Verlag, 2003.
[7] A. Gangemi, C. Catenacci, M. Ciaramita, and J. Lehmann, "Ontology evaluation and validation: An integrated formal model for the quality diagnostic task," tech. rep., ISTC-CNR, 2005.
[8] C. Welty, R. Kalra, and J. Chu-Carrol, "Evaluating ontological analysis," in *Proceedings of the ISWC-03 Workshop on Semantic Integration*, 2003.
[9] R. Porzel and R. Malaka, "A task-based approach for ontology evaluation," in *Proc. of the ECAI Workshop on Ontology Learning and Population*, 2004.
[10] C. Brewster, H. Alani, S. Dasmahapatra, and Y. Wilks, "Data-driven ontology evaluation," in *Proceedings of International Conference on Language Resources and Evaluation*, 2004.
[11] W. Daelemans and M.-L. Reinberger, "Shallow text understanding for ontology content evaluation," *IEEE Intelligent Systems*, vol. 19, no. 4, pp. 74–81, 2004.
[12] P. Spyns and M.-L. Reinberger, "Lexically evaluating ontology triples generated automatically from texts," in *Proc. of the second European Conference on the Semantic Web*, 2005.
[13] P. Spyns, "EvaLexon: Assessing triples mined from texts," Tech. Rep. STAR-2005-09, STAR Lab, 2005.
[14] N. Guarino and C. Welty, "An overview of OntoClean," in *Handbook on Ontologies* (S. Staab and R. Studer, eds.), Heidelberg: Springer Verlag, 2003.
[15] J. Völker, D. Vrandecic, and Y. Sure, "Automatic evaluation of ontologies (aeon)," in *Proceedings of the 4th International Semantic Web Conference (ISWC2005)*, 2005.
[16] M. Berland and E. Charniak, "Finding parts in very large corpora," in *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, 1999.
[17] R. Girju and D. Moldovan, "Text mining for causal relations," in *Proc. of the FLAIRS Conference*, 2002.
[18] M. Sabou, C. Wroe, C. Goble, and H. Stuckenschmidt, "Learning domain ontologies for semantic web service descriptions," *Journal of Web Semantics*, vol. 3, no. 4, 2005.
[19] A. Maedche and S. Staab, "Measuring similarity between ontologies," in *Proc. of the European Conference on Knowledge Acquisition and Management (EKAW-2002)*, 2002.
[20] M. Sabou, C. Wroe, C. Goble, and G. Mishne, "Learning domain ontologies for web service descriptions: an experiment in bioinformatics," in *Proc. of WWW05*, 2005.
[21] M.-L. Reinberger and P. Spyns, "Unsupervised text mining for the learning of dogma-inspired ontologies," in *Ontology Learning from Text: Methods, Applications and Evaluation* (P. Buitelaar, P. Cimiano, and B. Magnini, eds.), IOS Press, 2005.
[22] A. Maedche, *Ontology Learning for the Semantic Web*. Boston: Kluwer, 2002.
[23] E. Alfonseca and S. Manandhar, "Proposal for evaluating ontology refinement methods," in *Proceedings of the Language Resources and Evaluation Conference (LREC-2002)*, 2002.
[24] D. Maynard, W. Peters, and Y. Li, "Metrics for evaluation of ontology-based information extraction," in *Proc. of the EON 2006 Workshop*, 2006.
[25] D. Maynard, Y. Li, and W. Peters, "Using contextual information for term extraction and ontology population," in *Bridging the Gap between Text and Knowledge - Selected Contributions to Ontology Learning and Population from Text* (P. Buitelaar and P. Cimiano, eds.), IOS Press, 2007. THIS VOLUME.
[26] U. Hahn and K. Schnattinger, "Towards text knowledge engineering," in *Proc. of the 15th National Conference on Artificial Intelligence (AAAI-98)*, 1998.
[27] J. Brank, D. Mladenic, and M. Grobelnik, "Gold standard based ontology evaluation using instance assignment," in *Proc. of the EON 2006 Workshop*, 2006.
[28] J. Hartmann, P. Spyns, D. Maynard, R. Cuel, M. Carmen Suarez de Figueroa, and Y. Sure, "Methods for ontology evaluation," Deliverable D1.2.3, Knowledge Web, 2004.
[29] C. van Rijsbergen, *Information Retrieval*. London: Butterworths, 1979.
[30] K. Dellschaft, "Measuring the similiarity of concept hierarchies and its influence on the evaluation of learning procedures," diploma thesis, Universität Koblenz-Landau, December 2005. http://www.uni-koblenz.de/FB4/Institutes/IFI/AGStaab/Theses/2005/DADellschaft.pdf.
[31] M. Hearst, "Automatic aquisition of hyponyms from large text corpora," in *Proc. of the 14th International Conference on Computational Linguistics*, 1992.
[32] S. Staab, C. Braun, I. Bruder, A. Dusterhoft, A. Heuer, M. Klettke, G. Neumann, B. Prager, J. Pretzel, H. Schnurr, R. Studer, H. Uszkoreit, and B. Wrenger, "Getess - searching the web exploiting german texts," in *Proc. of the 3rd Workshop on Cooperative Information Agents*, 1999.