Measuring Similarity between Ontologies

Alexander Maedche[1] and Steffen Staab[2,3]

[1] Forschungszentrum Informatik at the Univ. Karlsruhe,
D-76131 Karlsruhe, Germany
http://www.fzi.de/WIM
[2] Institute AIFB, Univ. Karlsruhe,
D-76128 Karlsruhe, Germany
http://www.aifb.uni-karlsruhe.de/WBS
[3] Ontoprise GmbH,
76131 Karlsruhe, Germany
http://www.ontoprise.de

Abstract. Ontologies now play an important role for many knowledge-intensive applications for which they provide a source of precisely defined terms. However, with their wide-spread usage there come problems concerning their proliferation. Ontology engineers or users frequently have a core ontology that they use, e.g., for browsing or querying data, but they need to extend it with, adapt it to, or compare it with the large set of other ontologies. For the task of detecting and retrieving relevant ontologies, one needs means for measuring the similarity between ontologies. We present a set of ontology similarity measures and a multiple-phase empirical evaluation.

1 Introduction

A core purpose for the use of ontologies is the exchange of data not only at a common syntactic, but also at a shared semantic level. Especially on the WWW more and more ontologies are constructed and used, beginning to replace the old-fashioned ways of exchanging business data via standardized comma-separated formats by standards that adhere to semantic specifications given through ontologies. Thus, in the near future more and more ontologies will be made available on the WWW. With this upswing and beginning widespread usage of ontologies, however, new problems are incurred. Ontology engineers or users frequently have a core ontology that they use, e.g., for browsing or querying data, but they need to extend it with, adapt it to, or compare it with the large set of other ontologies. For the task of detecting and retrieving relevant ontologies, one needs means for measuring the similarity between ontologies on a canonical scale (e.g., the reals in $[0, 1]$).

So, how may we measure the similarity of ontologies or of ontology parts? One could make use of the formal structures of ontologies and try at the unification of ontologies or ontology parts (which is essentially subgraph matching). The drawback here would be that all real-world ontologies that we know of do not only specify its conceptualization by logical structures, but to a large extent also by reference to terms that are grounded through human natural language use. For instance, modeling that MAN

and WOMAN are subordinates of PERSON suffices for many purposes even without any further differentiae. Two ontologies that contain these parts agree on their semantics only to a small extent by formal means, but to a larger extent by reference to common terminology. Furthermore, missing structures need not be problematic. For instance, if one ontology comes with concepts referred to by VEHICLE, CAR, SPORTSWAGON and the other with VEHICLE and SPORTSWAGON only, the semantic exchange of data may still be rather easy, even though the second ontology lacks the two taxonomic links from VEHICLE to CAR and to SPORTSWAGON.

Looking at these requirements, we have found a lack of comprehensive methodological inventory to measure similarity between real-world ontologies, as well as practical, reproducable experiences with measuring similarity between ontologies. Firstly, this paper is about introducing the necessary inventory. We break down the overall task and propose a set of measures that capture the similarity of ontologies at two different levels, the lexical and the conceptual. In general our similarity measures describe the extent to which one ontology specification is covered by the other — and *vice versa*. Secondly, this paper is about providing some practical experiences and results with the proposed measures. Five subjects, four novices and one ontology engineering expert, have modeled ontologies in three different phases about a commonly well-known domain given some additional background knowledge in form of domain texts. The ontologies generated by the different subjects then served as input to an empirical evaluation study of our similarity measuring framework.

In the following, we first prepare the ground for our proposal and our empirical evaluation study by formally specifying the ontology structure and its semantic we refer to subsequently. In the two sections thereafter, we propose measures for describing the similarity of different ontology parts at the lexical and conceptual level. In Section 5, we describe the empirical evaluation study and the results we achieved there, before we relate to other research and conclude the paper with an outlook on future challenges.

2 A Two-Layer View of Ontologies

In order to compare two ontologies and measure similarity between them (or between parts of them), one may consider different semiotic levels. The two levels that we can focus on (abstracting from an actual application) are: First, at the lexical level we may investigate how terms are used to convey meanings. Second, at the conceptual level we may investigate what conceptual relations exist between the terms.[4] For this investigation we define a simple notion of ontology and some auxiliary functions in six steps.

Definition 1 (Concept Language). *Our simple concept language is defined starting from* atomic *concepts and roles.* Concepts *are unary predicates and* roles *are binary predicates over a domain $\mathcal{U}$, with* individuals *being the elements of $\mathcal{U}$. Correspondingly, an interpretation $\mathcal{I}$ of the language is a function that assigns to each concept symbol (taken from the set $\mathcal{A}$) a subset of the domain $\mathcal{U}$, $\mathcal{I} : \mathcal{A} \mapsto 2^{\mathcal{U}}$, to each role symbol (taken from the set $\mathcal{P}$) a binary relation of $\mathcal{U}$, $\mathcal{I} : \mathcal{P} \mapsto 2^{\mathcal{U}\times\mathcal{U}}$. Concept terms and role*

[4] Further studies could look at the pragmatic and the social level and try find out about the application of terms in concrete applications and social contexts.

terms are defined inductively with terminological axioms and using operators. C and D denote concept terms, R and S denote roles.

Concept Forming Operator	
Syntax	*Semantics*
C_{atom}	$\{d \in \mathcal{U}^{\mathcal{I}} \mid C_{atom}$ *atomic*, $d \in \mathcal{I}(C_{atom})\}$
$C \sqcap D$	$C^{\mathcal{I}} \cap D^{\mathcal{I}}$
$\forall R.C$	$\{d \in \mathcal{U}^{\mathcal{I}} \mid \forall e(d,e) \in R^{\mathcal{I}} \Rightarrow e \in C^{\mathcal{I}}\}$

Role Forming Operators	
Syntax	*Semantics*
R_{atom}	$\{(d,e) \in \mathcal{U}^{\mathcal{I}} \times \mathcal{U}^{\mathcal{I}} \mid R_{atom}$ *atomic*, $(d,e) \in \mathcal{I}(R_{atom})\}$
$R \sqcap S$	$R^{\mathcal{I}} \cap S^{\mathcal{I}}$
$C \times D$	$\{(d,e) \in C^{\mathcal{I}} \times D^{\mathcal{I}}\}$

Terminological Axioms			
Axiom	*Semantics*	*Axiom*	*Semantics*
$D \doteq C$	$D^{\mathcal{I}} = C^{\mathcal{I}}$	$D \sqsubseteq C$	$D^{\mathcal{I}} \subseteq C^{\mathcal{I}}$
$S \doteq R$	$S^{\mathcal{I}} = R^{\mathcal{I}}$	$S \sqsubseteq R$	$S^{\mathcal{I}} \subseteq R^{\mathcal{I}}$

Definition 2 (Lexicon). *The lexicon consists of a set of terms (lexical entries) for concepts, $\mathcal{L}^c$, and a set of terms for relations, $\mathcal{L}^r$. Their union is the lexicon $\mathcal{L} := \mathcal{L}^c \cup \mathcal{L}^r$.*

Definition 3 (Reference Function). *The **reference functions** $\mathcal{F}$, $\mathcal{G}$, with $\mathcal{F} : 2^{\mathcal{L}^c} \mapsto 2^{\mathcal{A}}$ and $\mathcal{G} : 2^{\mathcal{L}^s} \mapsto 2^{\mathcal{P}}$. $\mathcal{F}$ and $\mathcal{G}$ link sets of lexical entries*[5] *$\{L_i\} \subset \mathcal{L}$ to the set of concepts and relations they refer to, respectively. In general, one lexical entry may refer to several concepts or relations and one concept or relation may be refered to by several lexical entries. Their inverses are $\mathcal{F}^{-1}$ and $\mathcal{G}^{-1}$.*

We distinguish between terms and concept/relation symbols, because we want to allow for the explicit expression of ambiguities. For instance, one term like "bank" may refer to two concept symbols, viz. BANK-1 being a subconcept of FURNITURE and BANK-2 being a subconcept of COMPANY. Expressing this by disjunction (e.g., BANK $\doteq$ BANK-1 $\sqcup$ BANK-2) would be logically equivalent, but it would conflate two ontological states, viz. "bank" being an ambiguous natural language term and BANK-1 being a construed symbol for precise logical denotation.

Definition 4 (Core Ontology). *A core ontology $\mathcal{O}$ is a tuple $(\mathcal{A}, \mathcal{P}, \mathcal{D}, \mathcal{L}, \mathcal{F}, \mathcal{G})$, which consists of a set of concept symbols $\mathcal{A}$, a set of relation symbols $\mathcal{P}$, a set of statements $\mathcal{D}$ in the concept language defined above, a lexicon $\mathcal{L}$ and two reference functions $\mathcal{F}, \mathcal{G}$.*

Definition 5 (Concept Hierarchy). *The concept hierarchy $\mathcal{H}$ is defined by* $\mathcal{H} := \{(C,D) \mid C, D \in \mathcal{A} \wedge C^{\mathcal{I}} \subseteq D^{\mathcal{I}}\}$

Definition 6 (Domain/Range). *Domain ($d(R)$) and range ($r(R)$) of a relation R are defined by $\{d \mid \exists e(d,e) \in R^{\mathcal{I}}\}$ and $\{e \mid \exists d(d,e) \in R^{\mathcal{I}}\}$, respectively.*

In the following sections we propose and use methods for measuring similarity of ontologies based on the lexical and the conceptual level of ontologies.

[5] The reference functions are defined on sets of lexical entries (instead of single entities) in order to allow for a more compact description of formulae later on.

3 Lexical Comparison Level

The *edit distance* formulated by Levenshtein [5] is a well-established method for weighting the difference between two strings. It measures the minimum number of token insertions, deletions, and substitutions required to transform one string into another using a dynamic programming algorithm. For example, the edit distance, ed, between the two lexical entries "TopHotel" and "Top_Hotel" equals 1, ed("TopHotel", "Top_Hotel") $= 1$, because one insertion operation changes the string "TopHotel" into "Top_Hotel".

Based on Levenshtein's edit distance we propose a *lexical similarity measure* for strings, the String Matching (SM), which compares two lexical entries L_i, L_j:

$$\text{SM}(L_i, L_j) := \max\left(0, \frac{\min(|L_i|, |L_j|) - \text{ed}(L_i, L_j)}{\min(|L_i|, |L_j|)}\right) \in [0, 1].$$

SM returns a degree of similarity between 0 and 1, where 1 stands for perfect match and zero for bad match. It considers the number of changes that must be made to change one string into the other and weighs the number of these changes against the length of the shortest string of these two. In our example from above, we compute SM("TopHotel", "Top_Hotel") $= \frac{7}{8}$. In order to provide a summarizing figure for the lexical level of two sign systems, e.g. for the lexica referring to concepts $\mathcal{L}_1^c, \mathcal{L}_2^c$ of two ontologies $\mathcal{O}_1, \mathcal{O}_2$, we compare two sets $\mathcal{L}_1, \mathcal{L}_2$ returning the averaged String Matching $\overline{\text{SM}}(\mathcal{L}_1, \mathcal{L}_2)$:

$$\overline{\text{SM}}(\mathcal{L}_1, \mathcal{L}_2) := \frac{1}{|\mathcal{L}_1|} \sum_{L_i \in \mathcal{L}_1} \max_{L_j \in \mathcal{L}_2} \text{SM}(L_i, L_j).$$

$\overline{\text{SM}}(\mathcal{L}_1, \mathcal{L}_2)$ is an asymmetric measure that determines the extent to which the lexical level of a sign system $\mathcal{L}_1$ (the target) is covered by the one of a second sign system $\mathcal{L}_2$ (the source). Obviously, $\overline{\text{SM}}(\mathcal{L}_1, \mathcal{L}_2)$ may be quite different from $\overline{\text{SM}}(\mathcal{L}_2, \mathcal{L}_1)$. E.g., when $\mathcal{L}_2$ contains all the strings of $\mathcal{L}_1$, but also plenty of others, then $\overline{\text{SM}}(\mathcal{L}_1, \mathcal{L}_2) = 1$, but $\overline{\text{SM}}(\mathcal{L}_2, \mathcal{L}_1)$ may approach zero. Compared to the relative number of hits,

$$\text{RelHit}(\mathcal{L}_1, \mathcal{L}_2) := \frac{|\mathcal{L}_1 \cap \mathcal{L}_2|}{|\mathcal{L}_1|},$$

$\overline{\text{SM}}$ diminishes the influence of string pseudo-differences in different ontologies, such as use vs. not-use of underscores or hyphens, use of singular vs. plural, or use of additional markup characters. Of course, SM may sometimes be deceptive, when two strings resemble each other though they there is no meaningful relationship between them, e.g. "power" and "tower". In our case study, however, we have found that in spite of this added "noise" SM may be very helpful for proposing good matches of strings.

4 Conceptual Comparison Level

At the conceptual level we may compare semantic structures of ontologies $\mathcal{O}_1, \mathcal{O}_2$, that vary for concepts $\mathcal{A}_1, \mathcal{A}_2$. In our model the conceptual structures are constituted by $\mathcal{H}_1, \mathcal{H}_2$ and $\mathcal{P}_1, \mathcal{P}_2$.

4.1 Comparing taxonomies $\mathcal{H}_1, \mathcal{H}_2$

Though there has been a long discussion in the literature about comparing the similarity of two concepts in a common taxonomy (cf. Section 6), we have not found any discussion about *comparing two taxonomies*.

We start by determining the extent to which two taxonomies as seen from two particularly identified concepts compare. More precisely, we assume that we have one lexical entry $L \in \mathcal{L}_1^c \cap \mathcal{L}_2^c$ that refers via $\mathcal{F}_1$ and $\mathcal{F}_2$ to two concepts C_1, C_2 from two different taxonomies $\mathcal{H}_1, \mathcal{H}_2$. The intensional semantics of C_1 (C_2) may be seen to be constituted by the *semantic cotopy* (SC) of C_1 (C_2), i.e. all its super- and subconcepts:

$$\text{SC}(C_i, \mathcal{H}) := \{C_j \in \mathcal{A} | \mathcal{H}(C_i, C_j) \vee \mathcal{H}(C_j, C_i)\}.$$

SC is overloaded to process sets of concepts, too.

$$\text{SC}(\{C_1, \ldots, C_n\}, \mathcal{H}) := \bigcup_{i:=1\ldots n} \text{SC}(C_i, \mathcal{H}).$$

The taxonomic overlap (TO) between $\mathcal{H}_1$ and $\mathcal{H}_2$ as seen from the concepts referred to by L may then be computed by following $\mathcal{F}_1^{-1}$ and $\mathcal{F}_2^{-1}$ back to the common lexicon.

$$\text{TO}'(L, \mathcal{O}_1, \mathcal{O}_2) := \frac{|\mathcal{F}_1^{-1}(\text{SC}(\mathcal{F}(\{L\}), \mathcal{H}_1)) \cap \mathcal{F}_2^{-1}(\text{SC}(\mathcal{F}(\{L\}), \mathcal{H}_2))|}{|\mathcal{F}_1^{-1}(\text{SC}(\mathcal{F}(\{L\}), \mathcal{H}_1)) \cup \mathcal{F}_2^{-1}(\text{SC}(\mathcal{F}(\{L\}), \mathcal{H}_2))|}$$

Averaging over all lexical entries we may thus compute a semantic similarity for the two given hierarchies.

In addition, however, we must consider the case where a lexical entry L is in $\mathcal{L}_1^c$, but not in $\mathcal{L}_2^c$. Then, the simplest assumption is that the L is simply missing from $\mathcal{L}_2^c$, but when comparing the two hierarchies the optimistic taxonomic approximation is the one that searches for the maximum overlap given a fictive membership of L to $\mathcal{L}_2^c$ by

$$\text{TO}''(L, \mathcal{O}_1, \mathcal{O}_2) := \max_{C \in \mathcal{C}_2} \left\{ \frac{|\mathcal{F}_1^{-1}(\text{SC}(\mathcal{F}(\{L\}), \mathcal{H}_1)) \cap \mathcal{F}_2^{-1}(\text{SC}(C), \mathcal{H}_2)|}{|\mathcal{F}_1^{-1}(\text{SC}(\mathcal{F}(\{L\}), \mathcal{H}_1)) \cup \mathcal{F}_2^{-1}(\text{SC}(C), \mathcal{H}_2)|} \right\}$$

Given these premises the averaged similarity $\overline{\text{TO}}$ between two taxonomies $(\mathcal{H}_1, \mathcal{H}_2)$ of ontologies $(\mathcal{O}_1, \mathcal{O}_2)$ may then be defined by:

$$\overline{\text{TO}}(\mathcal{O}_1, \mathcal{O}_2) := \frac{1}{|\mathcal{L}_1^c|} \sum_{L \in \mathcal{L}_1^c} \text{TO}(L, \mathcal{O}_1, \mathcal{O}_2), \text{ with}$$

$$\text{TO}(L, \mathcal{O}_1, \mathcal{O}_2) := \begin{cases} \text{TO}'(L, \mathcal{O}_1, \mathcal{O}_2) & \text{if } L \in \mathcal{L}_2^c \\ \text{TO}''(L, \mathcal{O}_1, \mathcal{O}_2) & \text{if } L \notin \mathcal{L}_2^c \end{cases}$$

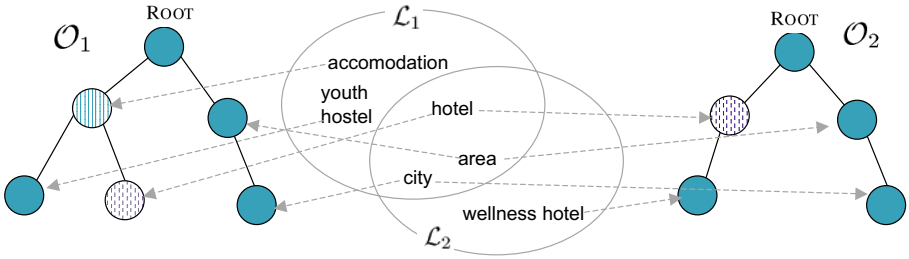

Fig. 1. Two Example Ontologies $\mathcal{O}_1, \mathcal{O}_2$

Example: A partial example for comparing taxonomies is given in Figure 1: The taxonomic overlap TO′(“hotel”, $\mathcal{H}_1, \mathcal{H}_2$) is determined by $\mathcal{F}_1^{-1}(\text{SC}(\mathcal{F}(\{\text{“hotel”}\}), \mathcal{H}_1)) = \{\text{“hotel”, “accomodation”}\}$ and $\mathcal{F}_2^{-1}(\text{SC}(\mathcal{F}(\{\text{“hotel”}\}), \mathcal{H}_2)) = \{\text{“wellness hotel”, “hotel”}\}$ resulting in TO′(“hotel”, $\mathcal{H}_1, \mathcal{H}_2) = \frac{1}{3}$ as input to $\overline{\text{TO}}$.

When we consider the lexical entry “accomodation”, which is only in $\mathcal{L}_1^c$, we compute the taxonomic overlap as follows: We compute for the lexical entry “accomodation” $\mathcal{F}_1^{-1}(\text{SC}(\mathcal{F}(\{\text{“accomodation”}\}), \mathcal{H}_1))$ = {“youth hostel”, “accomodation”, “hotel”}. The concept referred to by “hotel” in $\mathcal{A}_2$ yields the best match resulting in $\mathcal{F}_2^{-1}(\text{SC}(\mathcal{F}(\{\text{“hotel”}\}))) = \{\text{“wellness hotel”, “hotel”}\}$ and, thus, TO″(“accomodation”, $\mathcal{H}_1, \mathcal{H}_2) = \frac{1}{4}$.

The reader may note several properties of $\overline{\text{TO}}$: First, $\overline{\text{TO}}$ is asymmetric. While TO′ is a symmetrical measure, TO″ is asymmetric, because depending on coverage it may be very easy to integrate one taxonomy into another one, but it may be very difficult to do it the other way around. Second, for ease of presentation of the basic principles we have given here a shortened definition. The longer version specially considers the (minority of) cases, where one lexical entry refers to several concepts. The longer version does not consider the semantic cotopies of all referred concepts for computing TO, but only those that eventually optimize TO. Third, obviously $\overline{\text{TO}}$ becomes meaningless when $\mathcal{L}_1^c$ and $\mathcal{L}_2^c$ are disjoint. The more $\mathcal{L}_1^c$ and $\mathcal{L}_2^c$ overlap (or are made to overlap, e.g. through a syntactic merge), the better $\overline{\text{TO}}$ may focus on existing hierarchical structures and not on optimistic estimations of adding a new lexical entry to $\mathcal{L}_2^c$.

4.2 Comparing relations $\mathcal{P}_1, \mathcal{P}_2$

At the lexical level a relation R_1 is referred to by a lexical entry L_1. At the conceptual level it specifies a pair $(C_1, D_1), C_1, D_1 \in \mathcal{C}$ describing the concept C_1 that the relation belongs to and its range restriction D_1.

We determine the accuracy that two relations match, RO (relation overlap), based on the geometric mean value of how similar their domain and range concepts are. The geometric mean reflects the intuition that if either domain or range concepts utterly fail to match, the matching accuracy converges against 0, whereas the arithmetic mean value might still turn out a value of 0.5.

The similarity between two concepts (the concept match CM) may be computed by considering their semantic cotopy. However, the measures derived from complete co-

topies underestimate the place of concepts in the taxonomy. For instance, the semantic cotopy of the concept corresponding to "hotel" in $\mathcal{L}_2$ (Figure 1) is identical to the semantic cotopy of the one corresponding to "wellness hotel". Hence, for the purpose of similarity of concepts (rather than taxonomies), we define the upwards cotopy (UC) as follows:

$$\text{UC}(C_i, \mathcal{H}) := \{C_j \in \mathcal{A} | \mathcal{H}(C_i, C_j)\}.$$

Based on the definition of the upwards cotopy (UC) the concept match (CM) is then defined in analogy to TO':

$$\text{CM}(C_1, \mathcal{O}_1, C_2, \mathcal{O}_2) := \frac{|\mathcal{F}_1^{-1}(\text{UC}(C_1, \mathcal{H}_1)) \cap \mathcal{F}_2^{-1}(\text{UC}(C_2, \mathcal{H}_2))|}{|\mathcal{F}_1^{-1}(\text{UC}(C_1, \mathcal{H}_1)) \cup \mathcal{F}_2^{-1}(\text{UC}(C_2, \mathcal{H}_2))|}.$$

Then RO$'$ of relations R_1, R_2 may be defined by:

$$\text{RO}'(R_1, \mathcal{O}_1, R_2, \mathcal{O}_2) := \sqrt{\text{CM}(\text{d}(R_1), \mathcal{O}_1, \text{d}(R_2), \mathcal{O}_2) \cdot \text{CM}(\text{r}(R_1), \mathcal{O}_1, \text{r}(R_2), \mathcal{O}_1)}.$$

In order to take reference by $L \in \mathcal{L}_1^r, L \in \mathcal{L}_2^r$ into account:

$$\text{RO}''(L, \mathcal{O}_1, \mathcal{O}_2) := \frac{1}{|\mathcal{G}_1(\{L\})|} \sum_{R_1 \in \mathcal{G}_1(\{L\})} \max_{R_2 \in \mathcal{G}_2(\{L\})} \{\text{RO}'(R_1, \mathcal{O}_1, R_2, \mathcal{O}_2)\}$$

Some lexical entries only refer to relations in $\mathcal{P}_1$:

$$\text{RO}'''(L, \mathcal{O}_1, \mathcal{O}_2) := \frac{1}{|\mathcal{G}_1(\{L\})|} \sum_{R_1 \in \mathcal{G}_1(\{L\})} \max_{R_2 \in \mathcal{P}_2} \{\text{RO}'(R_1, \mathcal{O}_1, R_2, \mathcal{O}_2)\}$$

Combined we have for $L \in \mathcal{L}_1^r$:

$$\text{RO}(L, \mathcal{O}_1, \mathcal{O}_2) := \begin{cases} \text{RO}''(L, \mathcal{O}_1, \mathcal{O}_2) & \text{if } L \in \mathcal{L}_2^r \\ \text{RO}'''(L, \mathcal{O}_1, \mathcal{O}_2) & \text{if } L \notin \mathcal{L}_2^r \end{cases}$$

The averaged relation overlap $\overline{\text{RO}}$ is then defined by:

$$\overline{\text{RO}}(\mathcal{O}_1, \mathcal{O}_2) := \frac{1}{|\mathcal{L}_1^r|} \sum_{L \in \mathcal{L}_1^r} \text{RO}(L, \mathcal{O}_1, \mathcal{O}_2).$$

Example. We take Figure 1 as an example setting for computing RO. We assume one relation R_1 in $\mathcal{O}_1$, referenced by "located at" and specifying the domain and range corresponding to ("hotel", "area"). In $\mathcal{O}_2$, the same lexical entry may refer to R_2, with domain and range corresponding to ("hotel", "city"). Computing CM for the concepts referred to by "hotel" in $\mathcal{O}_1$ and $\mathcal{O}_2$ results in $\frac{1}{2}$. The CM between the concepts referred

to by "area" in $\mathcal{O}_1$ and "city" in $\mathcal{O}_2$ also returns $\frac{1}{2}$. Thus, the RO$'$ for the lexical entry "located at" boils down to $\sqrt{\frac{1}{2} \cdot \frac{1}{2}} = 0.5$ as input to $\overline{\text{RO}}$.

The reader may note two major characteristics of $\overline{\text{RO}}$. First, it depends on the agreement of the lexica and the taxonomies of $\mathcal{O}_1$ and $\mathcal{O}_2$. Without reasonable agreement, $\overline{\text{RO}}$ may not reach high values of similarity. Second, $\overline{\text{RO}}$ is also asymmetric reflecting the coverage of relations of the first by the second ontology.

5 Empirical Evaluation

In this section we present a case study that has been carried out in a seminar on ontology engineering at our institute. We have pursued two main objectives with our evaluation study: *(i)* we wanted to determine the quality of our measures and evaluate them on actual data, and, *(ii)*, we wanted to investigate and get an intuition about how similar ontologies about the same domain are that have been modeled by different persons.

5.1 Evaluation Study

The experiment was carried out with four subjects, viz. undergraduates in industrial engineering. The modeling expertise of the subjects was limited. Before actual modeling, they received 3 hours training in ontology engineering in general and 3 hours in using our ontology engineering workbench. Furthermore, they were acquainted with the purpose of the ontology, viz. as an ontology for information extraction and semantic search. Our study required from each of them the building of ontologies in the tourism domain using their background knowledge and using web pages from a WWW site about touristic offers, e.g. hotels with various attractions or cultural events. Our objective was an overall cross-comparison of ontologies, but we also wanted to test the appropriateness of single measures, To avoid error chaining, we therefore performed the evaluation in three phases (resulting in $4 \cdot 3 = 12$ ontologies). Furthermore, an expert ontology engineer (subject 0) modeled a "gold standard" for the task (a 13th ontology).

Phase I: A small top level structure was given to the subjects.[6] Based on this top level and the available knowledge sources, the subjects had to model a *complete* tourism domain ontology. To keep the ontologies within comparable ranges, the students were required to model around 300 concepts and 80 relations.

Phase II: The second phase was geared to produce results for $\overline{\text{TO}}$, while avoiding the uncertainties of lexical disagreement. Therefore, the subjects were given 310 lexical entries (for concepts) from the gold standard and the top level structure described before. Then everyone of them had to, first, model the taxonomy for concepts referred to by the 310 lexical entries and, second, model about 80 relations.

Phase III: The last phase was defined to control $\overline{\text{RO}}$ in absence of "noise" from different taxonomies and lexica. There the taxonomy (from the gold standard) was given. It consisted of 310 lexical entries, $\mathcal{L}^c$, and a set of 310 corresponding concepts, $\mathcal{A}$, taxonomically related by $\mathcal{H}$. The subjects had to model about 80 relations.

[6] It contained four concepts referred to by "thing", "material", "intangible", and "situation".

5.2 Lexical Comparison Level

The phase I-ontologies described above are used for general cross-comparison, including the lexical level. The pairwise string matching ($\overline{\text{SM}}$, cf. Section 3) of the five lexica referring to concepts and relations, respectively, returned the results depicted in Table 1.

Results: The results for computing $\overline{\text{SM}}(\mathcal{L}_1^c, \mathcal{L}_2^c)$ of matching lexical entries referring to concepts vary between 0.38 and 0.65 with an average of 0.45. Comparing lexical entries referring to relations $\overline{\text{SM}}(\mathcal{L}_1^s, \mathcal{L}_2^s)$ results in values between 0.16 and 0.53 with an average of 0.36. Several typical, though not necessarily good, pairs for which high string match values were computed are shown in Table 2. RelHit$(\mathcal{L}_1^c, \mathcal{L}_2^c)$ ranged between 20 to 25%, *i.e.* this percentage of lexical entries referring to concepts matched exactly. For lexical entries referring to relations the results were much worse, *viz.* between 10 to 15%.

	Subject				
$i \backslash j$	0	1	2	3	4
0	-	0.51,0.35	0.53,0.21	0.46,0.39	0.5,0.29
1	0.43,0.52	-	0.65,0.43	0.43,0.53	0.39,0.41
2	0.42,0.24	0.54,0.37	-	0.36,0.24	0.4,0.2
3	0.38,0.47	0.43,0.45	0.38,0.28	-	0.38,0.36
4	0.46,0.38	0.41,0.5	0.48,0.16	0.43,0.39	-

Table 1. $\overline{\text{SM}}(\mathcal{L}_i^c, \mathcal{L}_j^c)$, $\overline{\text{SM}}(\mathcal{L}_i^s, \mathcal{L}_j^s)$ for phase I-ontologies.

Interpretation: Analysing the figures we find that human subjects have a considerable higher agreement on lexical entries referring to concepts than on ones referring to relations. Investigating the auxiliary measures we have found that SM values above 0.75 in general retrieve meaningful matches — in spite of few pitfalls (cf. Table 2).

L_1	L_2	SM$(L1, L2)$
Sehenswuerdigkeit [seesight]	Sehenswürdigkeit [seesight]	0.875
Verkehrsmittel [vehicle]	Luftverkehrsmittel [air vehicle]	0.71
Zelt [tent]	Zeit [time]	0.75
Anzahl_Betten [number_beds]	hat_Anzahl_Betten [has_number_beds]	0.77

Table 2. Typical string matches

5.3 Conceptual Comparison Level

At the conceptual level we may compare semantic structures of ontologies $\mathcal{O}_1, \mathcal{O}_2$, that vary for concepts $\mathcal{A}_1, \mathcal{A}_2$. We use the ontologies of phase I, II, and III for evaluating our measures introduced in Section 4.

Results: Table 3 presents the results we have obtained for the phase I-ontologies using the similarity measures taxonomy overlap ($\overline{\text{TO}}$) and relation overlap ($\overline{\text{RO}}$). The reader may note that these ontologies have been built without any previous assumptions about the lexica $\mathcal{L}_1$ and $\mathcal{L}_2$, thus their similarity values are well below those of later phases where the lexica for concepts were predefined.

	Subject				
$i \backslash j$	0	1	2	3	4
0	-	0.33,0.35	0.31,0.25	0.32,0.5	0.29,0.28
1	0.35,0.15	-	0.4,0.41	0.34,0.03	0.28,0.15
2	0.28,0.12	0.36,0.25	-	0.25,0.04	0.24,0.15
3	0.36,0.4	0.31,0.32	0.24,0.04	-	0.26,0.03
4	0.38,0.29	0.31,0.21	0.32,0.2	0.32,0.26	-

Table 3. $\overline{\text{TO}}(\mathcal{O}_i, \mathcal{O}_j)$, $\overline{\text{RO}}(\mathcal{O}_i, \mathcal{O}_j)$ for phase I-ontologies.

Table 4 depicts the similarity measures computed for phase II-ontologies. Values for $\overline{\text{TO}}$ range between 0.47 and 0.87, the average $\overline{\text{TO}}$ over all 20 cross-comparisons results in 0.56. $\overline{\text{RO}}$ yields values from 0.34 to 0.82 with an average of 0.47.

	Subject				
$i \backslash j$	0	1	2	3	4
0	-	0.57,0.5	0.54,0.47	0.54,0.48	0.59,0.39
1	0.57,0.44	-	0.86,0.78	0.48,0.45	0.55,0.35
2	0.54,0.46	0.87,0.82	-	0.46,0.46	0.58,0.35
3	0.54,0.44	0.48,0.5	0.46,0.47	-	0.47,0.34
4	0.58,0.4	0.55,0.45	0.57,0.45	0.47,0.35	-

Table 4. $\overline{\text{TO}}(\mathcal{O}_i, \mathcal{O}_j)$, $\overline{\text{RO}}(\mathcal{O}_i, \mathcal{O}_j)$ for phase II-ontologies.

Interpretation: The figures indicate that subjects tend to agree or disagree on taxonomies irrespective of the amount of material being predefined. In fact, correlation between $\overline{\text{TO}}$ values of phase I- and phase II- ontologies support this indication, because correlation is 0.58 — distinctly positive — for the ontologies with and without predefined lexica. Furthermore, we may conjecture that comparison between $\overline{\text{TO}}$ values (in order to select the best) remains meaningful even with a restricted overlap of lexica.

Results: Table 5 depicts the similarity measures computed for phase III-ontologies, where only $\overline{\text{RO}}$ has been computed, because the taxonomy was predefined. $\overline{\text{RO}}$ here ranges between 0.23 and 0.71, the average $\overline{\text{RO}}$ over all 20 cross-comparisons achieving 0.5.

	Subject				
$i \backslash j$	0	1	2	3	4
0	-	0.61	0.38	0.51	0.54
1	0.69	-	0.56	0.57	0.55
2	0.4	0.49	-	0.35	0.23
3	0.67	0.71	0.5	-	0.57
4	0.45	0.44	0.3	0.41	-

Table 5. $\overline{\text{RO}}(\mathcal{O}_i, \mathcal{O}_j)$ for phase III-ontologies.

Interpretation: The correlation of $\overline{\text{RO}}$ values between phases I and II computes to 0.34, between phases I and III to 0.27, and between phases II and III to 0.16. In general, higher $\overline{\text{RO}}$ values are reached without a predefined taxonomy — this reflects the observation that subjects found it easy to use a predefined lexicon, but extremely difficult to continue modeling given a predefined taxonomy.

Overall, we may conjecture that the engineers' use of their lexicon correlates rather strongly with their conceptual model and *vice versa*: The similarity measures for subject 3 ontologies with subject 4 ontologies result in very low values at the lexical and at the conceptual level. In contrast, subject 1 ontologies reach high similarity values with subject 2 ontologies at all levels.

6 Related Work

Similarity measures for ontological structures have been widely researched, e.g. in cognitive science, databases [9], software engineering[11], and AI (e.g., [8, 1, 4, 3]). Though this research covers many wide areas and application possibilities, most of it has restricted its attention to the determination of similarity of lexicon, concepts, and relations *within one ontology*.

The nearest to our comparison *between two ontologies* come [2, 3] and [13]. [2] introduces several similarity measures in order to locate a new complex concept into an existing ontology by similarity rather than by logic subsumption. Bisson restricts the attention to the conceptual comparison level. In contrast to our work the new concept is described in terms of the existing ontology. Furthermore, he does not distinguish relations into taxonomic relations and other ones, thus ignoring the semantics of inheritance. [13] compute description compatibility in order to answer queries that are formulated with a conceptual structure that is different from the one of the information system. In contrast to our approach their measures depend to a very large extent on a shared ontology that mediates between locally extended ontologies. Their algorithm

also seems less suited to evaluate similarities of sets of lexical entries, taxonomies, and other relations.

Dieng & Hug [3] compare concept lattices in order to find out about the common location of two concepts in a merged ontology using several measures taking also advantage of the lattice. Again, however, their concerns are different from ours as they do not determine similarities of ontologies.

Research in the area of schema integration has been carried out since the beginning of the 1980s. Schema comparison analyzes and compares schema in order to determine correspondences and comes therefore near to our approach. However, their purpose is the alignment of pairs of tables or concepts [9] and often restricted to string and data type similarities.

Finally, so-called pathfinder networks [10] began in 1981 as an attempt to develop a network model for proximity data. They use use multidimensional scaling techniques. This statistical techniques transforms the concept network relationships into inter-point distances in a space of minimal dimensionality. In this space different similarity operations are performed. In contrast to our work, however, pathfinder networks do not focus on "real-world ontologies" including a lexical layer.

7 Conclusion

We have considered ontologies as two-layered systems, consisting of a lexical and a conceptual layer. Based on this core ontology model a methodological inventory to measure similarity between ontologies with each other based on the notions of lexicon $\mathcal{L}$, reference functions $\mathcal{F}, \mathcal{G}$ and semantic cotopy (SC, UC) has been described. Then, we have performed a three-phase empirical evaluation study to see how our measures perform in isolation and in combination. With our investigation we have created a methodological baseline and collected some empirical experiences.

Our measures may be applied in different application fields. First, we are currently working on an "ontology search engine" that will use the proposed measures as a basis retrieving ontologies based a user-defined core ontology that matches against available ontologies. Classical evaluation measures like precision and recall from the information retrieval community will serve as input for a quality-based evaluation of the proposed measures. Second, in [7] we describe how the measures presented in this paper may be extended for the instance level. Based on these instance-based similarity measures we provide means for computing a hierarchical clustering of ontology-based instances. Preliminary evaluation studies of applying the instance-based similarity measures within a clustering algorithm have shown promising results. Third, the measures proposed within this paper have shown to be very useful for supporting the discovery of mappings between two ontologies (see [6]). Fourth, such applications scenarios will become important for integrating existing ontologies into an ontology engineering process or for facilitating collaborative ontology engineering (cf. [12]).

Acknowledgements. Research for this paper was partially funded by the EU IST projects Bizon (IST-2001-33506) and SWAP (IST-2001-34103).

References

1. E. Agirre and G. Rigau. Word sense disambiguation using conceptual density. In *Proc. of COLING-96*, 1996.
2. G. Bisson. Learning in FOL with a similarity measure. In *Proc. of AAAI-1992*, pages 82–87, 1992.
3. R. Dieng and S. Hug. Comparison of personal ontologies represented through conceptual graphs. In *Proceedings of ECAI 1998*, pages 341–345, 1998.
4. E. Hovy. Combining and standardizing large-scale, practical ontologies for machine translation and other uses. In *Proc. of the First Int. Conf. on Language Resources and Evaluation (LREC)*, 1998.
5. I. V. Levenshtein. Binary Codes capable of correcting deletions, insertions, and reversals. *Cybernetics and Control Theory*, 10(8):707–710, 1966.
6. A. Maedche, B. Motik, N. Silva, and R. Volz. MAFRA – A MApping FRamework for Distributed Ontologies. In *Proceedings of the 13th European Conference on Knowledge Engineering and Knowledge Management EKAW-2002, Madrid, Spain*, 2002.
7. A. Maedche and V. Zacharias. Clustering Ontology-based Metadata in the Semantic Web. In *Proceedings of the Joint Conferences 13th European Conference on Machine Learning (ECML'02) and 6th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'02), Springer, LNAI, Finland, Helsinki*, 2002.
8. R. Rada, H. Mili, E. Bicknell, and M. Blettner. Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man, and Cybernetics*, 19(1), 1989.
9. E. Rahm and P. Bernstein. A survey of approaches to automatic schema matching. *VLDB Journal*, 10(4):334–350, 2001.
10. R. W. Schvanefeldt. *Pathfinder Associative Networks: Studies in Knowledge Organization*. Ablex Publishing Corporation, Norwood, New Jersey, 1989.
11. G. Spanoudakis and P. Constantopoulos. Similarity for analogical software reuse: A computational model. In *Proc. of ECAI-1994*, pages 18–22, 1994.
12. Y. Sure, M. Erdmann, J. Angele, S. Staab, R. Studer, and D. Wenke. Ontoedit: Collaborative ontology development for the semantic web. In *Proceedings of the 1st International Semantic Web Conference (ISWC2002), June 9-12th, 2002, Sardinia, Italia*, LNCS 2342, pages 221–235. Springer, 2002.
13. P. Weinstein and W. Birmingham. Comparing concepts in differentiated ontologies. In *Proc. of KAW-99*, 1999.