A SURVEY OF ONTOLOGY EVALUATION TECHNIQUES

Janez Brank, Marko Grobelnik, Dunja Mladenić
Department of Knowledge Technologies
Jozef Stefan Institute
Jamova 39, 1000 Ljubljana, Slovenia
Tel: +386 1 4773778; fax: +386 1 4251038
e-mail: janez.brank@ijs.si, marko.grobelnik@ijs.si, dunja.mladenic@ijs.si

ABSTRACT

An ontology is an explicit formal conceptualization of some domain of interest. Ontologies are increasingly used in various fields such as knowledge management, information extraction, and the semantic web. Ontology evaluation is the problem of assessing a given ontology from the point of view of a particular criterion of application, typically in order to determine which of several ontologies would best suit a particular purpose. This paper presents a survey of the state of the art in ontology evaluation.

1 INTRODUCTION

The focus of modern information systems is moving from "data processing" towards "concept processing", meaning that the basic unit of processing is less and less an atomic piece of data and is becoming more a semantic concept which caries an interpretation and exists in a context with other concepts. Ontology is commonly used as a structure capturing knowledge about a certain area via providing relevant concepts and relations between them.

A key factor which makes a particular discipline or approach scientific is the ability to evaluate and compare the ideas within the area. The same holds also for Semantic Web research area when dealing with abstractions in the form of ontologies. Ontologies are a fundamental data structure for conceptualizing knowledge, but we are generally able to build many different ontologies conceptualizing the same body of knowledge and we should be able to say which of them best suits some predefined criterion.

Thus, ontology evaluation is an important issue that must be addressed if ontologies are to be widely adopted in the semantic web and other semantics-aware applications. Users facing a multitude of ontologies need to have a way of assessing them and deciding which one best fits their requirements the best. Likewise, people constructing an ontology need a way to evaluate the resulting ontology and possibly to guide the construction process and any refinement steps. Automated or semi-automated ontology learning techniques also require effective evaluation measures, which can be used to select the "best" ontology out of many candidates, to select values of tunable parameters of the learning algorithm, or to direct the learning process itself (if the latter is formulated as a path through a search space).

2 A CLASSIFICATION OF ONTOLOGY EVALUATION APPROACHES

Various approaches to the evaluation of ontologies have been considered in the literature, depending on what kind of ontologies are being evaluated and for what purpose. Broadly speaking, most evaluation approaches fall into one of the following categories:

- those based on comparing the ontology to a "golden standard" (which may itself be an ontology; e.g. MAEDCHE AND STAAB, 2002);
- those based on using the ontology in an application and evaluating the results (e.g. PORZEL & MALAKA, 2004);
- those involving comparisons with a source of data (e.g. a collection of documents) about the domain to be covered by the ontology (e.g. BREWSTER *et al.*, 2004);
- those where evaluation is done by humans who try to assess how well the ontology meets a set of predefined criteria, standards, requirements, etc. (e.g. LOZANO-TELLO AND GÓMEZ-PÉREZ, 2004).

In addition to the above categories of evaluation, we can group the ontology evaluation approaches based on the level of evaluation, as described below.

An ontology is a fairly complex structure and it is often more practical to focus on the evaluation of different levels of the ontology separately rather than trying to directly evaluate the ontology as a whole. This is particularly true if we want a predominantly automated evaluation rather than entirely carried out by human users/experts. Another reason for the level-based approach is that when automatic learning techniques have been used in the construction of the ontology, the techniques involved are substantially different for the different levels. The individual levels have been defined variously by different authors, but these various definitions tend to be broadly similar and usually involve the following levels:

Lexical, vocabulary, or data layer. Here the focus is on which concepts, instances, facts, etc. have been included in the ontology, and the vocabulary used to represent or identify these concepts. Evaluation on this level tends to involve comparisons with various sources of data concerning the problem domain (e.g. domain-specific text corpora), as well as techniques such as string similarity measures (e.g. edit distance).

Hierarchy or taxonomy. An ontology typically includes a hierarchical *is-a* relation between concepts. Although various other relations between concepts may be also defined, the is-a relationship is often particularly important and may be the focus of specific evaluation efforts.

Other semantic relations. The ontology may contain other relations besides is-a, and these relations may be evaluated separately. This typically includes measures such as precision and recall.

Context or application level. An ontology may be part of

a larger collection of ontologies, and may reference or be referenced by various definitions in these other ontologies. In this case it may be important to take this context into account when evaluating it. Another form of context is the application where the ontology is to be used; evaluation looks at how the results of the application are affected by the use of the ontology.

Syntactic level. Evaluation on this level may be of particular interest for ontologies that have been mostly constructed manually. The ontology is usually described in a particular formal language and must match the syntactic requirements of that language. Various other syntactic considerations, such as the presence of natural-language documentation, avoiding loops between definitions, etc., may also be considered (GÓMEZ-PÉREZ, 1994).

Structure, architecture, design. This is primarily of interest in manually constructed ontologies. We want the ontology to meet certain pre-defined design principles or criteria; structural concerns involve the organization of the ontology and its suitability for further development (GÓMEZ-PÉREZ, 1994, 1996). This sort of evaluation usually proceeds entirely manually.

The following table summarizes which approaches from the list at the beginning of this section are commonly used for which of these levels.

Table 1. An overview of approaches to ontology evaluation.

	Approach to evaluation			
Level	Golden standard	Application-based	Data-driven	Assessment by humans
Lexical, vocabulary, concept, data	x	x	x	x
Hierarchy, taxonomy	x	x	x	x
Other semantic relations	x	x	x	x
Context, application		x		x
Syntactic	x[1]			x
Structure, architecture, design				x

[1] "Golden standard" in the sense of comparing the syntax in the ontology definition with the syntax specification of the formal language in which the ontology is written (e.g. RDF, OWL, etc.).

The next sections will present more details about the various approaches and the levels of evaluation.

3 EVALUATION ON THE LEXICAL/VOCABULARY AND CONCEPT/DATA LEVEL

An example of an approach that can be used for the evaluation of a lexical/vocabulary level of an ontology is the one proposed by MAEDCHE AND STAAB (2002). Similarity between two strings is measured based on the Levenshtein edit distance, normalized to produce scores in the range [0, 1]. A *string matching* measure between two sets of strings is then defined by taking each string of the first set, finding its similarity to the most similar string in the second set, and averaging this over all strings of the first set. One may take the set of all strings used as concept identifiers in the ontology being evaluated, and compare it to a "golden standard" set of strings that are considered a good representation of the concepts of the problem domain under consideration. The golden standard could be in fact another ontology (as in Maedche and Staab's work), or it could be taken statistically from a corpus of documents (see sec. 7), or prepared by domain experts.

The lexical content of an ontology can also be evaluated using the concepts of precision and recall, as known in information retrieval. In this context, *precision* would be the percentage of the ontology lexical entries (strings used as concept identifiers) that also appear in the golden standard, relative to the total number of ontology words. *Recall* is the percentage of the golden standard lexical entries that also appear as concept identifiers in the ontology, relative to the total number of golden standard lexical entries. A way to achieve more tolerant matching criteria (allowing synonyms, etc.) is to augment each lexical entry with its hypernyms from WordNet or some similar resource (BREWSTER *et al.*, 2004); then, instead of testing for equality of two lexical entries, one can test for overlap between their corresponding sets of words (each set containing an entry with its hypernyms).

The same approaches could also be used to evaluate the lexical content of an ontology on other levels, e.g. the strings used to identify relations, instances, etc.

VELARDI *et al.* (2005) describe an approach for the evaluation of an ontology learning system which takes a body of natural-language text and tries to extract from it relevant domain-specific concepts (terms and phrases), and then find definitions for them (using web searches and WordNet entries) and connect some of the concepts by is-a relations. Part of their evaluation approach is to generate natural-language glosses for multiple-word terms. These glosses can then be evaluated by domain experts, who therefore do not have to be familiar with formal languages in which ontologies are commonly described.

4 EVALUATION OF TAXONOMIC AND OTHER SEMANTIC RELATIONS

BREWSTER *et al.* (2004) suggested using a data-driven approach to evaluate the degree of structural fit between an ontology and a corpus of documents. (1) Given a corpus of documents from the domain of interest, a clustering algorithm based on EM is used to determine, in an unsupervised way, a probabilistic mixture model of hidden "topics" such that each document can be modeled as having been generated by a mixture of topics. (2) Each concept *c* of the ontology is represented by a set of terms including its name in the ontology and the hypernyms of this name, taken from WordNet. (3) The probabilistic models obtained during clustering can be used to measure, for each topic identified by the clustering algorithm, how well the concept *c* fits that topic. (4) At this point, if we require that each concept fits at least some topic reasonably well, we obtain a technique for lexical-level evaluation of the ontology. Alternatively, we may require that concepts associated with

the same topic should be closely related in the ontology (via is-a and possibly other relations). This would indicate that the structure of the ontology is reasonably well aligned with the hidden structure of topics in the domain-specific corpus of documents. A drawback of this method as an approach for evaluating relations is that it is difficult to take the directionality of relations into account (e.g. we may know that concepts c_1 and c_2 should be related, but we cannot really infer whether c_1 is-a c_2, or c_2 is-a c_1, or if some completely different relation should be used).

Given a golden standard, evaluation of an ontology on the relational level can also be based on precision and recall measures, comparing the ontology either with a human-provided golden standard, or with a list of statistically relevant terms. This was used by SPYNS (2005) to evaluate an approach for automatically extracting a set of lexons, i.e. triples of the form $\langle \text{term}_1, \text{role}, \text{term}_2 \rangle$, from natural-language text. Unfortunately preparing the golden standard requires a lot of manual human work.

A somewhat different aspect of ontology evaluation has been discussed by GUARINO AND WELTY (2002). They point out several philosophical notions (essentiality, rigidity, unity, etc.) that can be used to better understand the nature of various kinds of semantic relationships that commonly appear in ontologies, and to discover possible problematic decisions in the structure of an ontology (for example, is-a is sometimes used to express meta-level characteristics of some class, or is used instead of is-a-part-of, or is used to indicate that a term may have multiple meanings). A downside of this approach is that it requires manual intervention by a trained human expert familiar with the above-mentioned notions such as rigidity; the expert should annotate the concepts of the ontology with appropriate metadata tags, whereupon checks for certain kinds of errors can be made automatically.

MAEDCHE AND STAAB (2002) propose several measures for comparing the relational aspects of two ontologies. Although this is in a way a drawback of this method, an important positive aspect is that once the golden standard is defined, comparison of two ontologies can proceed entirely automatically. The *semantic cotopy* of a term c in a given hierarchy is the set of all its super- and sub-concepts. Given two hierarchies H_1, H_2, a term t might represent some concept c_1 in H_1 and a concept c_2 in H_2. One can then compute the set of terms which represent concepts from the cotopy of c_1 in H_2, and the set of terms representing concepts from the cotopy of c_2; the overlap of these two sets can be used as a measure of how similar a role the term t has in the two hierarchies H_1 and H_2. An average of this may then be computed over all the terms occurring in the two hierarchies; this is a measure of similarity between H_1 and H_2. Similar ideas can also be used to compare other relations besides is-a.

5 CONTEXT-LEVEL EVALUATION

Sometimes the ontology is a part of a larger collection of ontologies that may reference one another (e.g. one ontology may use a class or concept declared in another ontology), for example on the web or within some institutional library of ontologies. This context can be used for evaluation of an ontology in various ways. For example, the Swoogle search engine of DING *et al.* (2004) uses cross-references between semantic-web documents to define a graph and then compute a score for each ontology in a manner analogous to PageRank used by the Google web search engine. A similar approach has been used in the OntoKhoj portal of PATEL *et al.* (2003). Not all "links" or references between ontologies are treated the same. For example, if one ontology defines a subclass of a class from another ontology, this reference might be considered more important than if one ontology only uses a class from another as the domain or range of some relation.

Alternatively, the context for evaluation may be provided by human experts; for example, SUPEKAR (2005) proposes that an ontology be enhanced with metadata such as its design policy, how it is being used by others, as well as "peer reviews" provided by users of this ontology. A suitable search engine could then be used to perform queries on this metadata and would aid the user in deciding which of the many ontologies in a repository to use.

6 APPLICATION-BASED EVALUATION

Typically, the ontology will be used in some kind of application or task. The outputs of the application, or its performance on the given task, might be better or worse depending partly on the ontology used in it. Thus one might argue that a good ontology is one which helps the application in question produce good results on the given task. Ontologies may therefore be evaluated simply by plugging them into an application and evaluating the results of the application. This is elegant in the sense that the output of the application might be something for which a relatively straightforward and non-problematic evaluation approach already exists. For example, PORZEL AND MALAKA (2004) describe a scenario where the ontology, with its relations (both is-a and others) is used primarily to determine how closely related the meaning of two concepts is. The task is a speech recognition problem, where evaluation of the final output of the task is relatively straightforward (proposed interpretations of the sentences are compared with a gold standard provided by humans).

The application-based approach to ontology evaluation also has several drawbacks: (1) we see that an ontology is good or bad when used in a particular way for a particular task, but it's difficult to generalize this observation; (2) the ontology could be only a small component of the application and its effect on the outcome may be relatively small and indirect; (3) comparing different ontologies is only possible if they can all be plugged into the same application.

7 DATA-DRIVEN EVALUATION

An ontology may also be evaluated by comparing it to existing data (usually a collection of textual documents) about the problem domain to which the ontology refers. For

example, PATEL *et al.* (2003) show how to determine if the ontology refers to a particular topic, and to classify the ontology into a directory of topics: one extracts textual data from the ontology (such as names of concepts and relations) and uses this as the input to a text classification model (trained using standard machine learning algorithms).

Similarly, BREWSTER *et al.* (2004) extracted a set of relevant domain-specific terms from the corpus of documents, using latent semantic analysis. The amount of overlap between the domain-specific terms and the terms appearing in the ontology (e.g. as names of concepts) can then be used to measure the fit between the ontology and the corpus.

In the case of extensive ontologies incorporating a lot of factual information (such as Cyc, see e.g. *www.cyc.com*), the documents could also be used as a source of "facts" about the external world, and the evaluation examines if these facts can also be derived from the ontology.

8 MULTIPLE-CRITERIA APPROACHES

Another family of approaches to ontology evaluation deals with the problem of selecting a good ontology (or a small short-list of promising ontologies) from a given set of ontologies, and treats this problem as essentially a decision-making problem. To help us evaluate the ontologies, we can use approaches based on defining several decision criteria or attributes; for each criterion, the ontology is evaluated and given a numerical score. An overall score for the ontology is then computed as a weighted sum of its per-criterion scores. Similar strategies are used in many other contexts to select the best candidate (e.g. tenders, grant applications, etc.). A drawback is that a lot of manual involvement by human experts may be needed. In effect, the general problem of ontology evaluation has been deferred or relegated to the question of how to evaluate the ontology with respect to the individual evaluation criteria. On the positive side, these approaches allow us to combine criteria from most of the levels discussed in section 2.

BURTON-JONES *et al.* (2004) propose an approach of this type, with ten simple criteria: lawfulness (i.e. frequency of syntactical errors), richness (how many of the syntactic features available in the formal language are actually used by the ontology), interpretability (do the terms used in the ontology also appear in WordNet?), consistency (how many concepts in the ontology are involved in inconsistencies), clarity (do the terms used in the ontology have many senses in WordNet?), comprehensiveness (number of concepts in the ontology, relative to the average for the entire library of ontologies), accuracy (percentage of false statements in the ontology), relevance (number of statements that involve syntactic features marked as useful or acceptable to the user/agent), authority (how many other ontologies use concepts from this ontology), history (how many accesses to this ontology have been made, relative to other ontologies in the library/repository).

FOX *et al.* (1998) propose another set of criteria, which is however geared more towards manual assessment and evaluation of ontologies. LOZANO-TELLO AND GÓMEZ-PÉREZ (2004) define an even more detailed set of 117 criteria, organized in a three-level framework.

9 CONCLUSIONS AND FUTURE WORK

Ontology evaluation remains an important open problem in the area of ontology-supported computing and the semantic web. There is no single best or preferred approach to ontology evaluation; instead, the choice of a suitable approach must depend on the purpose of evaluation, the application in which the ontology is to be used, and on what aspect of the ontology we are trying to evaluate. In our opinion, future work in this area should focus particularly on automated ontology evaluation, which is a necessary precondition for the healthy development of automated ontology processing techniques for a number of problems, such as ontology learning, population, mediation, matching, and so on.

Acknowledgments

This work was supported by the Slovenian Research Agency and the IST Programme of the European Community under SEKT Semantically Enabled Knowledge Technologies (IST-1-506826-IP) and PASCAL Network of Excellence (IST-2002-506778). This publication only reflects the authors' views.

References

1. BREWSTER, C. *et al.* Data driven ontology evaluation. Proceedings of Int. Conf. on Language Resources and Evaluation, Lisbon, 2004.
2. BURTON-JONES, A., *et al.*, A semiotic metrics suite for assessing the quality of ontologies. *Data and Knowledge Engineering* (2004).
3. DING, L., *et al.*, Swoogle: A search and metadata engine for the semantic web. *Proc. CIKM 2004*, pp. 652–659.
4. EHRIG, M., *et al.*, Similarity for ontologies — a comprehensive framework. *Proc. Eur. C. Inf. Sys.*, 2005.
5. FOX, M. S., *et al.*, An organization ontology for enterprise modelling. In: M. Prietula *et al.*, *Simulating organizations*, MIT Press, 1998.
6. GÓMEZ-PÉREZ, A. Some ideas and examples to evaluate ontologies. Knowledge Systems Laboratory, Stanford University, 1994.
7. GÓMEZ-PÉREZ, A. Towards a framework to verify knowledge sharing technology. *Expert Systems with Applications*, 11(4):519–529 (1996).
8. GUARINO, N., WELTY, C., Evaluating ontological decisions with OntoClean. *Comm. of the ACM*, 45(2):61–65, February 2002.
9. LOZANO-TELLO, A., GÓMEZ-PÉREZ, A., Ontometric: A method to choose the appropriate ontology. *J. Datab. Mgmt.*, 15(2):1–18 (2004).
10. MAEDCHE, A., STAAB, S., Measuring similarity between ontologies. *Proc. CIKM 2002*. LNAI vol. 2473.
11. PATEL, C., *et al.*, OntoKhoj: a semantic web portal for ontology searching, ranking and classification. *ACM Web Inf. & Data Mgmt.*, 2004.
12. PORZEL, R., MALAKA, R., A task-based approach for ontology evaluation. *ECAI 2004 Workshop Ont. Learning and Population*.
13. SPYNS, P., EvaLexon: Assessing triples mined from texts. Technical Report 09, STAR Lab, Brussels, Belgium, 2005.
14. SUPEKAR, K. A peer-review approach for ontology evaluation. *Proc. 8th Intl. Protégé Conference*, Madrid, Spain, July 18–21, 2005.
15. VELARDI, P., *et al.*, Evaluation of OntoLearn, a methodology for automatic learning of domain ontologies. In: *Ont. Learning from Text: Methods, Evaluation and Applications*, IOS Press, 2005.