Data Driven Ontology Evaluation

Christopher Brewster*, Harith Alani†, Srinandan Dasmahapatra†, Yorick Wilks*

*Department of Computer Science, University of Sheffield,
211 Portebello Street, Sheffield, S1 4DP, U.K.
{C.Brewster|Y.Wilks}@dcs.shef.ac.uk

† School of Electronics and Computer Science,
University of Southampton, SO17 1BJ, United Kingdom
{ha|sd}@ecs.soton.ac.uk

Abstract

The evaluation of ontologies is vital for the growth of the Semantic Web. We consider a number of problems in evaluating a knowledge artifact like an ontology. We propose in this paper that one approach to ontology evaluation should be corpus or data driven. A corpus is the most accessible form of knowledge and its use allows a measure to be derived of the 'fit' between an ontology and a domain of knowledge. We consider a number of methods for measuring this 'fit' and propose a measure to evaluate structural fit, and a probabilistic approach to identifying the best ontology.

1. Introduction

Current work in Knowledge Management, the Semantic Web, and a variety of Semantic Web Services depends on ontologies as the backbone to application development (Fensel et al., 2003). This paper argues for the need to develop a clear set of evaluation methodologies and makes the case for the importance of an approach based on a data-driven evaluation methodology. We believe that for significant progress to be achieved in the development and deployment of ontologies, evaluation metrics must be available similar to those used in TREC, MUC or Senseval. The usage of measures like precision and recall, used in such evaluations has been successful in providing a set of performance benchmarks. We address some of the issues that arise in the choice and usage of empirical methods in addressing the task of evaluating ontologies.

Good ontologies are the ones that serve their purpose. Complete ontologies are probably more than what most knowledge services require to function properly. The biggest impediment to ontology use is the cost of building them, and deploying "scruffy" ontologies that are cheap to build and easy to maintain might be a more practical and economical option. Equally there has been much focus on the potential of ontology re-use, which would also lower the entry cost. In both cases, the existence of appropriate evaluation methodologies is essential.

The rest of this paper is organised as follows. In Section 2, we discuss the problem of evaluating a knowledge representation, in Section 3 we enumerate different types of evaluation, and in Section 4 we present our ideas on data-driven evaluation. This is followed by a brief account of related work and a conclusion.

2. The Evaluation of a Representation of Knowledge

There are inherent problems in trying to evaluate an ontology as it is not clear what exactly one is trying to evaluate. An ontology is a representation or model of knowledge, a "formal, explicit specification of a shared conceptualisation" according to (Gruber, 1993), and this means that however 'shared' it may be it is still extremely subjective, representing the time, place and cultural environment in which it is created. A particular ontology reflects the interests of the knowledge users, which must be captured in the design criteria for ontology construction. The labels of the concepts picked out to describe the concepts of interest to a user or application context, is an act of interpretation over the information available.

Ontology evaluation cannot be compared to the evaluation tasks in Information Retrieval or classic Natural Language Processing tasks such as POS tagging, because the notion of precision and recall cannot be easily used. One would like *precision* to reflect the amount of knowledge correctly identified (in the ontology) with respect to the whole knowledge available in the ontology. One would like to define *recall* to reflect the amount of knowledge correctly identified with respect to all the knowledge that it should identify. But we have yet to develop any clear notion of what this means. *Precision* and *recall* depend on a clear set of items concerned, for example Parts of Speech. There is no clear set of "knowledge to be acquired" because the same set of facts can give rise to very different interpretations and therefore different kinds of "knowledge".

One way of approaching the problem might be to decompose it into its constituent parts. An ontology, at its simplest, is composed of concepts and relations, some of which are explicitly defined, others which follow from a set of axioms. We may view these constructs as abstractions from a set of natural language texts describing the domain. Our approach involves reversing this (tacit) process of abstraction and we will propose to find signatures in natural language texts of the relevant concepts from the ontology. It is hoped that the variability in the definition of concepts of interest within an ontology and the varied

ways in which constructions in natural language express the shared nature of our understanding of a particular domain will exhibit visible correlations. Thus we need to identify word (co-)occurrences that stand in for concepts and relations in the ontology. Given a corpus of texts representative of a given domain of knowledge, it is relatively easy to identify the salient terms used in the texts using current technology (Maynard and Ananiadou, 2000). Determining the set of relations immediately raises problems. On the one hand there are common relations such IS-A (hyponomy) and part-of (meronymy) but it is not clear that these are of the right granularity to represent knowledge in the ontology. Finally, to identify the appropriate relationship between each concept is currently the greatest challenge and one where automatic approaches are generally unsatisfactory.

3. Types of Evaluation

A major distinction needs to be made between a qualitative and a quantitative approach to evaluation. A qualitative approach might present users with an ontology, or subparts of an ontology, and ask them to rate it. The problem here is that it is quite hard to determine who the right users are, and what criteria to propose they use for their evaluation. Should the domain experts be considered the users, or the knowledge engineers, or even the end users? Should they evaluate an ontology more highly because it is 'sensible', 'coherent', 'complete' or 'correct', and what do we mean by these terms? Furthermore most users could not evaluate the logical correctness of an ontology.

A closely related qualitative approach would be to evaluate an ontology from the perspective of the principles used in its construction. Such an approach has been especially espoused by (Guarino, 1998) and (Gómez-Pérez, 1999). While some of these design principles are valid in theory, it is extremely difficult to construct automated tests which will comparatively evaluate two or more ontologies as to their consistent use of "identity criteria" or their taxonomic rigour. This is because such principles depend on an external semantics to perform that evaluation, which currently only human beings are capable of providing. Furthermore, there is a significant danger that in applying a principles-based approach to ontology construction the result could be vacuous and of no practical use, as argued in (Wilks, 2002).

Another approach would be to evaluate how effective a particular ontology is in the context of an application. Currently, no work has been done to take a given application environment and test a number of similar but different ontologies in order to evaluate which is most appropriate for the application concerned and determine why that is so. The establishment of a clear set of simple application suites which would allow a number of different ontologies to be 'slotted in' in order to evaluate the ontologies would be an important research step. Another way of achieving the same result would be to set up the detailed tasks presented in TREC or MUC. From the perspective of the machine-readability vision of the Semantic Web, where ontologies are an enabling technologies for interoperability of processes, it may even be entirely inappropriate for humans to read and assess the ontology, only the effects of the ontology are to be judged.

The focus of this paper is on a third approach which concerns the congruence or 'fit' between an ontology and a domain of knowledge. It is impossible to automatically evaluate directly the fit between a knowledge artefact such as an ontology and a person's knowledge of a domain, let alone the knowledge of a group. One standard approach would be to compare a new ontology with an existing 'gold standard' one. Such has been the approach espoused by authors such as (Grefenstette, 1994). The problem here is that if the results differ from the gold standard, it is hard to determine whether that is because the corpus is inappropriate, the methodology is flawed or there is a real difference in the knowledge present in the corpus and the gold standard. In any case, this approach is more applicable when one is trying to evaluate ontology *learning* methodologies. In the Semantic Web scenario, it is likely that one has to choose from a range of existing ontologies the most appropriate for a particular domain, or the most appropriate to adapt to the specific needs of the domain/application.

Elsewhere it has been argued that a corpus of texts might be the most effective source of information for the construction of a large proportion of ontologies (Brewster et al., 2001). The traditional methods of ontology construction of protocol analysis or introspection are extremely time-consuming, laborious, and expensive. For the evaluation of an ontology (however built) a major part of the evaluation should be to identify the 'fit' of the ontology with a domain specific corpus. We purposefully use the word 'fit' because there are a number of ways in which this congruence can be assessed.

4. Data-driven Evaluation

4.1. The Scenario

Let us imagine a situation where the knowledge engineer has an application in mind (a specific Semantic Web Service, let us say) and can identify the knowledge area needed. This can be represented initially as a corpus of texts concerning the domain. From a number of existing ontologies, they must select the most appropriate for the application, and determine if it needs significant revision for the intended application. We consider below a number of methods which could facilitate this selection process.

In our case, we chose the art and artists domain for which we had developed the ARTEQUAKT application (Alani et al., 2003). Using this we collected 41 arbitrary texts from the Internet on a number of artists. The ARTEQUAKT ontology was compared with four others: The Ontology of Science (Ontology of Science, n.d.) was a revised version of the KA2 ontology, the AKT Reference Ontology (AKT, 2003) concerns the academic domain, the CIDOC Conceptual Reference Model (CRM) (CIDOC, 2003) is an ontology representing cultural heritage, and SUMO is the Suggested Upper Merged Ontology (IEEE P1600.1 Standard Upper Ontology Working Group, 2003).

4.2. Basic Comparison of Ontologies with Texts

Rather than compare one ontology with another existing one as undertaken by (Maedche and Staab, 2002), we propose to compare one or more ontologies with a corpus.

To achieve this, one could, for example, perform automated term extraction on the corpus and simply count the number of terms that overlap between the ontology and the corpus. The ontology can be penalised for terms present in the corpus and absent in ontology, and for terms present in the ontology but absent in the corpus.

Another approach is to use a vector space representation of the terms in both the corpus and the ontologies under evaluation. This permits an overall measure of the 'fit' between the ontologies and the corpus. Thus for example, when comparing the five ontologies mentioned above with our corpus of artist related texts, we obtained the figures shown in Table 1.

Ontology	Similarity
Artequakt	0.0518
CRM	0.0460
AKT	0.0259
Science	0.0221
SUMO	0.0355

Table 1: Vector comparison of five ontologies

This fits with our intuitive and objective understanding of the Artequakt ontology as having the closest fit with the selected corpus.

4.3. An Architecture for Ontology-Corpus Evaluation

We propose a somewhat more sophisticated architecture for ontology evaluation in view of a corpus, and thus obtaining a measure of the overall fit. There are three steps each of which can be undertaken using a variety of methodologies:

1. **Identifying keywords/terms.** This is essentially a form of automated term recognition, and thus the whole panoply of techniques existing can be applied (Maynard and Ananiadou, 2000). In our simple test case we applied Latent Semantic Analysis (Hofmann, 1999) and used a clustering method.

2. **Query expansion.** Because a concept in the ontology is a compact representation of a number of different lexical realisations in a number of ways, it is important to perform some form of query expansion of the concept terms. In our test case, we used WordNet to add two levels of hypernyms to each term in a cluster. There are other ways to expand the a term using (for example) IR techniques.

3. **Ontology Mapping.** Finally, the set of terms identified in the corpus need to be mapped to the ontology.

Given a corpus appropriately annotated against an ontology, we could count how many concept terms in the ontology match those lexical items that have been marked up. This would yield initial (crude) measures of lexical keyword coverage by ontology labels (precision and recall). This provides figures which reflect the coverage of the ontology of the corpus. The most common scenario is one where there are items absent as well as items unneeded.

The advantage of using a cluster analysis approach is that it permits the creation of a measure of *structural fit*. We can imagine two ontologies with identical concept sets which, however, have the concepts differently organised and thus concepts are at a different distance from each other. Thus we propose a 'tennis measure' (cf. (Stevenson, 2002)) for an ontology which evaluates the extent to which items in the same cluster are closer together in the ontology than those in different clusters. What is determined as close is dependent on the probability model used to derive the clusters.

4.4. A Probabilistic Approach

Within a probabilistic setting, we express the evaluation of the "best fit" between a corpus and one among a set of ontologies as the requirement of finding the conditional probability of the ontologies given the corpus. The ontology that maximises the conditional probability of the ontology O given a corpus C is then the best fit ontology O^*:

$$O^* = \text{argmax}_O P(O|C) = \text{argmax}_O \frac{P(C|O)P(O)}{P(C)}$$

If ontology-tagged corpora were available, it would be possible to estimate $P(C|O)$ in Bayes' theorem above. Otherwise, we take recourse to a variety of ways in which one could attempt to extract the information content of the corpus in order to correlate that with the ontology. The identification of words that relate best to the concepts and relations in the ontology may be extracted by assembling all the concept labels of each ontology and collecting terms from the corresponding hypernym tree in WordNet for each concept. The match between these terms and words from the corpus provide a measure of the aptness of the association between ontologies and corpora.

In order to assess the closeness of concepts in the ontology as compared with some unsupervised measure of relatedness or clustering of terms within the corpus, we need to find a way of extracting clusters. The approach we have taken as a first step is along the lines of an aspect model (Pereira et al., 1993; Hofmann, 1999).

$$P(d, w) = \sum_{t \in T} P(w|t)P(d|t)P(t),$$

where the word-document (w, d) co-occurrences in the corpus is modelled by mutinomial distributions indexed by a set T of hidden "topic" variables $t \in T$. We trained these distributions using expectation-maximisation (EM). For each cluster variable, we estimate the conditional probability of concept labels ℓ:

$$\begin{aligned} P(\ell|t \in T) &= \alpha \sum_{w_t} P(\ell|w_t)P(w_t) \\ &+ (1-\alpha) \sum_{w_t} \sum_{c_i} P(\ell|c_i)P(c_i|w_t)P(w_t). \end{aligned} \quad (1)$$

The two terms in the convex combination denote the direct word-concept matches and those mediated by query expansion, here implemented by traversing the hypernym tree of WordNet. The constant α can be set by an EM algorithm

in the presence of training data, or else set by heuristic fiat. While this gives a measure of coherence of concepts within an ontology, an accumulation of these probability values in (1), say by taking a product over all concept labels, gives a measure of fit between the corpus and an ontology.

The use of clustering to find the degree of fit allows us the flexibility of discovering varying degrees of fitness across fragments of a corpus. If there is a variability in the highest ranks among ontologies across the corpus, we could identify the documents corresponding to different clusters from high values of $P(d|t)$.

5. Related Work

Research has been undertaken on applying qualitative principle-based approaches to evaluating ontologies as mentioned above (Guarino, 1998; Gómez-Pérez, 1999). Hovy presents an extensive set of parameters by which to compare one ontology with another but does not present any quantitative methods (Hovy, 2001). The standard approach has always been to compare an ontology to a gold standard. Thus Grefenstette used Roget, and even the recent work of Maedche takes for granted that the evaluation of the ontology learning methodology must be done with reference to a hand-built reference ontology, and discusses the evaluation of one ontology with respect to another (Maedche and Staab, 2002). Researchers have not previously considered comparing different ontologies with a given set of texts.

6. Conclusions and Future Work

In this paper, we have argued for the need to establish objective measures for ontology creation. We have proposed a number of methods to evaluate the congruence of an ontology (or set of ontologies) with a given corpus in order to determine how appropriate it is for the representation of the knowledge of the domain represented by the texts. Future research in this area should seek to develop further techniques for evaluating how appropriate a given ontology is for a domain. Furthermore we might envisage a Semantic Web Service for automated ontology evaluation. This would recommend the most appropriate ontology for a given set of documents to the knowledge engineer and eventually for automated Semantic Web annotation agents.

7. References

AKT, 2003. AKT reference ontology. Available at www.aktors.org/publications/ontology. An ontology of the academic domain.

Alani, Harith, S. Kim, D.E. Millard, M.J. Weal, W. Hall, P.H. Lewis, and Nigel R. Shadbolt, 2003. Automatic ontology-based knowledge extraction and tailored biography generation from the web. *IEEE Intelligent Systems*, 18(1).

Brewster, Christopher, Fabio Ciravegna, and Yorick Wilks, 2001. Knowledge acquisition for knowledge management: Position paper. In *Proceeding of the IJCAI-2001 Workshop on Ontology Learning*. Seattle, WA: IJCAI.

CIDOC, 2003. The CIDOC conceptual reference model. Available at http://cidoc.ics.forth.gr/. A cultural heritage ontology.

Fensel, Deiter, James Hendler, Henry Lieberman, and Wolfgang Wahlster, 2003. *Spinning the Semantic Web*. Cambridge, MA: MIT Press.

Gómez-Pérez, Asuncion, 1999. Evaluation of taxonomic knowledge in ontologies and knowledge bases. In *Proceedings of the 12th Banff Knowledge Acquisition for Knowledge-Based Systems Workshop, Banff, Alberta, Canada*.

Grefenstette, Gregory, 1994. *Explorations in Automatic Thesaurus Discovery*. Amsterdam: Kluwer.

Gruber, T. R., 1993. A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition*, 6(2):199–221.

Guarino, Nicola, 1998. Some ontological principles for designing upper level lexical resources. In *Proceedings of the First International Conference on Language Resources and Evaluation, LREC 98*.

Hofmann, T., 1999. Probabilistic latent semantic indexing. In *Proceedings of the 22nd ACM-SIGIR International Conference on Research and Development in Information Retrieval*. ACM.

Hovy, Eduard, 2001. Comparing sets of semantic relations in ontologies. In Rebecca Green, Carol A Bean, and Sung Hyon Myaeng (eds.), *Semantics of Relationships*, chapter 6. Dordrecht, NL: Kluwer.

IEEE P1600.1 Standard Upper Ontology Working Group, 2003. Suggested upper merged ontology. http://ontology.teknowledge.com/.

Maedche, Alexander and Steffen Staab, 2002. Measuring similarity between ontologies. In *Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web, Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management, EKAW02*, volume 2473 of *LNAI*. Berlin: Springer Verlag.

Maynard, Diana and Sophia Ananiadou, 2000. TRUCKS: a model for automatic term recognition. *Journal of Natural Language Processing*.

Ontology of Science, n.d. Available at http://protege.-stanford.edu/ontologies/ontologyOfScience/Science.zip. A modified version of the KA2 ontology.

Pereira, Fernando C. N., Naftali Z. Tishby, and Lillian Lee, 1993. Distributional clustering of english words. In *Proceedings of 30th Annual Meeting of the Association for Computational Linguistics*. Ohio State University, Association for Computational Linguistics.

Stevenson, Mark, 2002. Combining disambiguation techniques to enrich an ontology. In *Proceedings of the Fifteenth European Conference on Artificial Intelligence (ECAI-02) workshop on "Machine Learning and Natural Language Processing for Ontology Engineering", Lyon, France*.

Wilks, Yorick, 2002. Ontotherapy: or how to stop worrying about what there is. Invited presentation, Ontolex 2002, Workshop on Ontologies and Lexical Knowledge Bases, 27th May. Held in conjunction with the Third International Conference on Language Resources and Evaluation - LREC02, 29-31 May, Las Palmas, Canary Islands.