Evaluation Framework for Local Ontologies Interoperability

Paolo Avesani
ITC-Irst, Via Sommarive – Località Povo, 38050 Trento, Italy
E-mail: avesani@itc.it

Introduction

The latest trends of knowledge management is to enable the interoperability among distributed and autonomous sources of knowledge (Bonifacio, Bouquet, & Traverso 2002). The innovative claim is that a centralized encoding of the knowledge is not suitable in the real world. The social process that underlies the knowledge creation and distribution is too complex to be forced into a centralized model.

If we assume that two different sources of knowledge could be developed autonomously a new issue rises: how to support their interoperability without a common interpretation agreement in advance? The open challenge is to support a process of meaning negotiation that considers a shared understanding the outcome and not the premise of the interaction between two knowledge sources.

Although with different background many scientific communities are looking at this problem. Distributed databases and semantic web are two major examples where the match between two different schemas and ontologies, respectively, is becoming one of the main target of the research effort. As a consequence the solutions to map two sources of knowledge is growing. At the same time there is a lack of methodology to evaluate alternative solutions.

The objective of this work is to propose a framework to evaluate new competing solutions without any assumption on the matching techniques. The main goal is to provide an answer to the big issue related to the matching problem: how to assess the accuracy of the mapping between two concepts? This question is really hard because we need to know in advance the answer we are looking for, but if we know in advance the criteria to assess the equivalence of two concepts representation we could use such criteria as a method itself.

I will argue that a viable way to proceed is to acquire an annotated collections of knowledge representations where the annotation can be conceived as an association between an object and a concept. Such a kind of annotation will play the role of an approximation of the knowledge interpretation.

In this paper we will focus our attention on a specific knowledge representation suitable to encode local ontologies, namely *context* (Benerecetti, Bouquet, & Ghidini 2000). In the next section I will briefly summarize the notion of context as a concept hierarchy and the related definition of context mapping. After that I will restrict the generic interoperability between two sources of knowledge to a specific task of information retrieval. Finally I will introduce the notion of annotated collections and their use to support the computation of evaluation measures.

Mapping as Context Matching

The representation of knowledge is commonly formulated through ontology languages. Up to now there isn't a common standard and the choice usually depends on the the purpose. Let imagine to address the categorization task. In this case the representation of knowledge can be conceived as a taxonomy of concepts.

Definition 1 (Concept hierarchy)
A concept hierarchy*, namely a context, is a graph* $C = \langle N, E \rangle$ *where* N *is a finite set of nodes and* E *a finite set of directed edges between nodes; all the nodes have a label, chosen in a set* L *of labels, while all the edges take value in* {*is-a, part-of, instance-of*} *such that the edges labelled with hierarchical labels forms an acyclic graph.*

If we have two different contexts, that play the role of taxonomies, we could be interested to know how to map one concept defined in the *source* context into an other concept defined in the *target* context. Such a kind of contexts overlapping is represented by a structure called *context mapping* (Bouquet *et al.* 2002).

Definition 2 (Context Mapping)
The context mapping is a 4-tuple $\langle m, c_s, \mathcal{M}, c_t \rangle$*, where*

1. m *a unique identifier associated with a mapping;*
2. c_s *and* c_t *are distinct context identifiers, called* source *and* target *context;*
3. $\mathcal{M}$ *is a* concept mapping *from the content of the source context to the content of the target context. Where a context mapping from a concept hierarchies*

$C_s = \langle N_s, E_s \rangle$ to $C_t = \langle N_t, E_t \rangle$, is a tuple of relations each of which is a subset of $C_s \times C_t$.

The mapping between two contexts can be considered a way to assess the similarity among the concepts defined autonomously by two sources of knowledge.

Information Retrieval

The notion of similarity is always related to a given target or goal. For this reason we have to specify in advance what are the tasks related to the context similarity assessment. We could identify at least two:

- **Concept-based Retrieval.** We can design a scenario where given a context C_1 a concept $s \in C_1$ is selected as representative of the meaning that a seeker is looking for; the goal in this scenario is to detect in a context C_2 a concept $p \in C_2$ that has the same meaning of s, i.e. the document classified under the node p would the same classified under the node s.
- **Context-based Retrieval.** A different scenario can be conceived where a knowledge engineer is going to design a context to be deployed as categorization taxonomy; given a sketch of such a context the knowledge engineer could be interested to find a past developed context for revision or completion purposes.

The two tasks illustrated above address different notions of similarity. For example the former could promote as most similar a mapping where all the concepts of a context C_1 are projected on a single concept of a context C_2. The same mapping could not be considered a good approximation of the context similarity because in this case a non injective mapping should be preferred.

In the following we will focus on the concept-based retrieval task.

Approaching Context Matching

Before to introduce a methodology of evaluation we need to detail more the interoperability model of the concept-based retrieval scenario. The interaction between a seeker and a provider implements a meaning negotiation process that can be of two types:

- **Supervised.** In this case the similarity assessment can refer both the concepts defined by the contexts and the document classified under the given contexts.
- **Unsupervised.** In this case the similarity assessment can refer only the concepts as defined by the contexts without taking advantage of the information that could derive from the documents classified under a given context.

In the following we will focus only on the unsupervised scenario.

Annotated Context Collection

Although we have excluded the possibility to refer the documents classified under the given contexts it doesn't mean that we give up to exploit them. As mentioned in the introduction the big issue is to acquire in advance the right mapping between two concepts that belong to different contexts.

The trivial solution could be to acquire in advance the optimal mapping taking advantage of an expert that interprets both the contexts. But this approach suffers of at least a couple of drawbacks. The former is concerned with the scalability: what about the expert when the size of the contexts increases? what about the combinatorial explosion when the number of contexts increases? The latter is even worse. The idea that an expert can provide the optimal mapping between two contexts is a little contradictory. We have made the assumption that the two contexts are autonomously defined then two interpretation schemas apply not necessarily covered by a single user.

It should be evident that the elicitation of the optimal mapping can be achieved only as a result of a collaborative effort of many experts that through a process of negotiation achieve a common agreement. This kind of approach has been applied by the community of the natural language processing to address the evaluation of the message understanding task (Hirschman 1998).

Here I would like to propose an approach that preserves the initial assumption of autonomy and at the same time doesn't require a cognitive overload not sustainable by a real expert.

The basic idea is like in other scientific community (Hirschman 1998) to build annotated collections where usually the target function is known. In our case the annotation shouldn't record directly the optimal mapping between contexts, i.e. the relation that occur when between two interpretations of two contexts the same meaning occurs. On the contrary , the goal of the annotation will be the elicitation of the interpretation of the single context. As it is well known the representation of the interpretation is not trivial but taking into account the specific task of concept-based retrieval we could proceed as follows:

Definition 3 (Classification Function)
Given a context c, a classification function f_c projects a given document d in the subset of concept $\{n_k\} \in c$:

$$f_c : D \times C \longrightarrow 2^N$$

The definition above enable the modelling of the interpretation for a given context. If we look at a subset of documents we could have an approximation based on examples: the pairs $(document, concept)$ produced by the classification function f_c.

A kind of inverse function can be defined that allows us to complete the model for the concept-based retrieval:

Definition 4 (Retrieval Function)
Given a annotated context c, a retrieval function f_r returns the set of documents $\{d_i\}$ classified under a given concept $n_k \in c$:

$$f_r : C \longrightarrow 2^D$$

Let imagine to get available the funtions above. We could claim that given a concept $s \in C_1$ and a concept $p \in C_2$ they have the same meaning if $f_r(s) = f_r(p)$. Of course it is true when the set of documents is close to the infinity.

If we accept to have a partial definition of the above function the condition $f_r(s) = f_r(p)$ will be only an approximation. The estimate $\hat{f}_r$ will be easily obtained through a process of annotation that starting from a corpus of documents D derives the association with the concepts of a given collection of contexts.

But at this stage it is possible to assess something more than the simple equality relation, for example taking advantage of the precision and recall measures, typical of information retrieval (Baeza-Yates & Ribeiro-Neto 1999), a similarity between concepts can be formulated as

$$Sim(s,p) = 2|\hat{f}_r(s) \cap \hat{f}_r(p)| \cdot |\hat{f}_r(s)|^{-1} \cdot |\hat{f}_r(p)|^{-1}$$

Differently from information retrieval, where no relations hold among the keywords, in the concept-based retrieval the context provides the auxiliary information on the concepts relationships. How to exploit this further information to obtain a more accurate assessment of concept similarity is matter of the next section.

Before to move forward it is worthwhile to notice that this approach could be effective with every kinds of contents (image, sounds, ...) and not only with text documents.

Measures for Context Mapping

At this stage we assume to have a collection of contexts C and a collection of documents D. Let suppose to have organized an annotation process that classified D over C. Taken two contexts we will call $S = \hat{f}_r(s)$ the set of documents classified under the concept $s \in C_1$, namely *seeker*; and $P = \hat{f}_r(p)$ the set of documents classified under the concept $p \in C_2$, namely *provider*.

Definition 5 (Ambiguity)
This notion is defined by the ratio between the marginal sets and the shared documents:

$$Ambiguity = \frac{|M_P^S| + |M_S^P|}{|O_P^S|}$$

where $M_P^S = S \setminus (S \cap P)$ is the marginal set of documents classified by S and not classified by P (similarly $M_S^P = P \setminus (P \cap S)$). The set of shared documents is defined as $O_P^S = P \cap S$ and where the following equivalence applies $O_P^S = O_S^P$.

As mentioned before in the *ambiguity* measure we didn't take into account the knowledge encoded in the contexts but only in the node labels that denotate the concepts.

Definition 6 (Generalization Ambiguity Index)

This notion refers to the ambiguity introduced by considering the concept S as a generalization of the concept P. It is defined as follows:

$$GAI = \frac{|M_P^S| + |M_S^P|}{|O_P^S| + |O_{A_S}^P| + |O_{T_P}^S|}$$

where $O_{A_S}^P$ represents the set of documents resulting from the intersection between M_S^P and the set of documents classified under the concepts in the hierarchy above S (i.e. the ancestors); similarly $O_{T_P}^S$ represents the set of documents resulting from the intersection between M_P^S and the set of documents classified under the concepts in the hierarchy below P (i.e. the children).

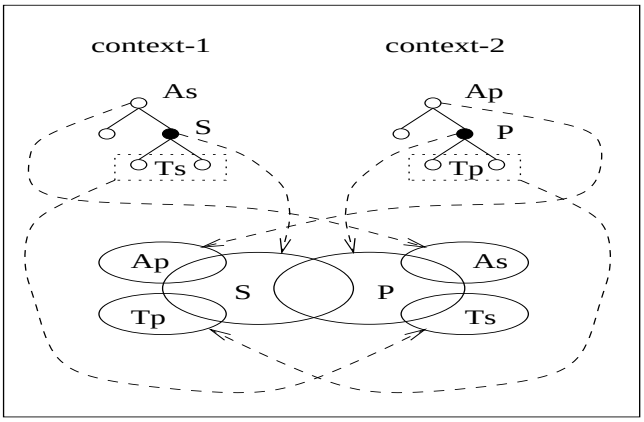

Figure 1: An example of annotated context.

Because in some way the notion of specialization is the opposite of generalization we can similarly define the corresponding measure:

Definition 7 (Specialization Ambiguity Index)
This notion refers to the ambiguity introduced by considering the concept S as a specialization of the concept P. It is defined as follows:

$$SAI = \frac{|M_P^S| + |M_S^P|}{|O_P^S| + |O_{T_S}^P| + |O_{A_P}^S|}$$

where $O_{T_S}^P$ represents the set of documents resulting from the intersection between M_S^P and the set of documents classified under the concepts in the hierarchy below S (i.e. the children); similarly $O_{A_P}^S$ represents the set of documents resulting from the intersection between M_P^S and the set of documents classified under the concepts in the hierarchy above P (i.e. the ancestors).

The computation of the ambiguity indexes, making both generalization and specialization hypothesis, introduces an error derived by the approximation of the correct interpretation of the further related concepts: the ancestors and the children.

Let define the analytical estimate of the ambiguity error associated to generalization and specialization approximation respectively.

Definition 8 (Generalization Ambiguity Error) *The error introduced by the generalization ambiguity assessment is defined as*

$$GAE = \frac{|O^P_{A_S}| + |O^S_{T_P}|}{|O^S_P|}$$

i.e. the ratio between the marginal documents classified in the generalized concepts and the shared classified documents.

In the same way the specialization hypothesis is the premise of a further ambiguity error.

Definition 9 (Specialization Ambiguity Error) *The error introduced by the specialization ambiguity assessment is defined as*

$$SAE = \frac{|O^P_{T_S}| + |O^S_{A_P}|}{|O^S_P|}$$

i.e. the ratio between the marginal documents classified in the specialized concepts and the shared classified documents.

Ambiguity index and generalization/specialization error can be combined to balance their contributions into an optimum.

Definition 10 (Generalization) *The generalization represents the trade off between the latent generalization ambiguity between two concepts (GAI) and the error introduced by the hypothesis of generalization (GAE):*

$$Generalization = \frac{1}{GAI + GAE}$$

Definition 11 (Specialization) *The specialization represents the trade off between the latent specialization ambiguity between two concepts (SAI) and the error introduced by the hypothesis of specialization (SAE):*

$$Specialization = \frac{1}{SAI + SAE}$$

The previous measures have been defined over the range $(0, \infty)$ but it is a trivial exercise to reformulate them in the range $[0, 1]$. Once moved to the unary interval we can summarize both the ambiguity indexes and errors into a more general measure.

The above definitions distinguish between generalization and specialization assessment of one concept respect with another. If we are interested to take into account the mutual influence we need a measure that will be the balance of the two alternative hypothesis: specialization and generalization. The basic intuition is that a promising evaluation of the both hypothesis seems a little contradictory. A mutual exclusion should apply between specialization and generalization.

Definition 12 (Concept Similarity) *Given two concepts and two measures respectively to assess their generalization and specialization hypothesis, two concept similarities are defined*

$$G - Similarity = Generalization \cdot (1 - Specialization)$$

$$S - Similarity = Specialization \cdot (1 - Generalization)$$

where generalization and specialization take values in $[0, 1]$.

The basic intuition underlying the previous definition is that the specialization assessment should provide the complementary result of the generalization assessment.

All these measures are based on the classification functions f^i_c that model the correct interpretation of the concepts that belong to a context representation. It is well known that to refer to a set of examples is not enough to provide a non ambiguous representation of the concept meaning. However we can model the heterogeneity related to the concept representation enabling the association of more than one classification function to the same context. The side effect of this change is that we can not assume that the mapping of two instances of the same context will be trivial.

Definition 13 (Heterogeneity Index) *Given a mapping M between two instances of the same context,* f^i_c *and* f^j_c *are two different classification functions defined on the same context*

$$H(f^i_c, f^j_c) = \sum_{v_s, v_p \in \mathcal{M}} |Similarity^{f^i_c}_{v_s, v_p} - Similarity^{f^j_c}_{v_s, v_p}|$$

Once we allow more than one user to classify a corpus of documents respect with the same context it is possible that the interpretation of the same concept hierarchy will differ. The result will be two different classifications. The computation of the *heterogeneity index* will provide the estimate of the error derived by the working hypothesis that model the interpretation through the examples.

The *heterogeneity index* could be a powerful tool to evaluate the quality (or better the complexity) of a given annotated collection of contexts.

References

Baeza-Yates, R., and Ribeiro-Neto, B. 1999. *Modern Information Retrieval*. Addison Wesley.

Benerecetti, M.; Bouquet, P.; and Ghidini, C. 2000. Contextual reasoning distilled. *JETAI* 12(3):279–305.

Bonifacio, M.; Bouquet, P.; and Traverso, P. 2002. Enabling distributed knowledge management. managerial and technological implications. *Informatik/Informatique* 3(1).

Bouquet, P.; Donà, A.; Serafini, L.; and Zenobini, S. 2002. ConTeXualized local ontology specification via ctxml. In *Submitted to AAAI Workshop on Meaning Negotiation*. Available also as IRST Technical Report TR-0204-01.

Hirschman, L. 1998. The evolution of evaluation: Lessons from the message understanding conferences. *Computer Speech and Language* 12:281–305.