Semantic Indexing of a Competence Map to support Scientific Collaboration in a Research Community

Paola Velardi, Roberto Navigli
Università di Roma "La Sapienza"
Via Salaria, 113 – 00198 Roma Italy
{velardi,navigli}@di.uniroma1.it

Michaël Petit
University of Namur (FUNDP)
Rue Grandgagnage, 21 – B-5000 Namur Belgium
mpe@info.fundp.ac.be

Abstract

This paper describes a methodology to semi-automatically acquire a taxonomy of terms and term definitions in a specific research domain. The taxonomy is then used for semantic search and indexing of a knowledge base of scientific competences, called Knowledge Map. The KMap is a system to support research collaborations and sharing of results within and beyond a European Network of Excellence. The methodology is general and can be applied to model any web community - starting from the documents shared and exchanged among the community members - and to use this model for improving accessibility of data and knowledge repositories.

1 Introduction

The NoE (Network of Excellence) INTEROP[1] is an instrument for strengthening excellence of European research in interoperability of enterprise applications, by bringing together the complementary competences needed to develop interoperability in a more global and innovative way. One of the main objectives of INTEROP has been to build a so-called "*Knowledge Map*" (KMap) of partner competences, to perform a periodic diagnostics of the extent of research collaboration and coordination among the NoE members. The aim is to monitor the status of research in the field of interoperability through a web-based platform that allows the user to retrieve information according to his/her actual need in a specific situation.

The main benefits of the KMap (Figure 1) for its users are:

- To be able to diagnose current interoperability research inside INTEROP and in Europe;
- To receive an overview of all European research activities on interoperability and subordinated topics;
- To receive an overview of organisations and experts as well as research results;
- To find relevant information for specific needs quickly;
- To find potential partners for collaborating in research activities.

The target groups of the INTEROP KMap system are:

- **Members** of the KMap management team, who are in charge of producing a periodic diagnostics of current research in interoperability performed by INTEROP partners in the first place and in Europe in the second place;
- **INTEROP partners** who contribute with information about their research and that of other researchers in the domain of interoperability, and retrieve knowledge about the current status of interoperability research;
- The **scientific community** in the field of interoperability, including universities, research institutes, researchers, companies, etc.

These objectives and targets can be considered relevant for any scientific web community in any research field.

Figure 1. The INTEROP KMap.

The KMap is a Knowledge Management application, exploiting recent research results in the area of Semantic Web, Text Mining, Information Retrieval and Ontology Enrichment. These techniques have been put in place to create a *semantically indexed* information repository, storing data on active collaborations, projects, research results, and organizations. A query interface allows users to retrieve information about partner collaborations, research results and available (or missing) competences, as well as to obtain summarized information (presented in a graphical or tabular format) on the overall degree of collaboration and

[1] http://www.interop-noe.org (2003-2007), NoE-IST 508011.

overlapping competence, based on a measure of *semantic similarity* between pieces of information.

The paper is organized as follows: first, we provide a general picture of the knowledge acquisition value chain, and the motivations behind the adopted approach. Then, we summarize the learning techniques used to bootstrap the creation of a domain taxonomy (currently evolving towards an ontology). Finally, we describe the implementation and preliminary results of the semantically indexed KMap. Related research and future activities are dealt with in the Concluding Remarks section.

2 The Knowledge Acquisition Value Chain

Figure 2 schematizes the Knowledge Acquisition Value Chain adopted in INTEROP. Progressively richer knowledge structures (Lexicon, Glossary, Taxonomy, Ontology) are first bootstrapped through automatic text mining techniques, and then refined through manual validation and enrichment, supported by appropriate tools and collaborative web interfaces. Each knowledge structure builds on previously acquired knowledge, e.g. automatic glossary extraction exploits knowledge on domain terminology (the lexicon), automatic identification of taxonomic relations is based on glossary parsing, etc.

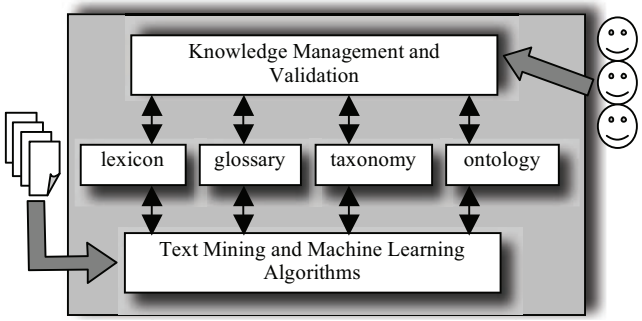

Figure 2. The Knowledge Acquisition Value Chain.

In short, the steps of the knowledge acquisition chain are the following[2] (steps marked A are automatic, steps marked M are manual, supported by web applications):

1. (A) Text and documents exchanged by the members of the research community are parsed to extract a set of domain terms, constituting the *domain lexicon L*;
2. (A) For each term t in *L*, one or more *definitions* are automatically extracted from the available documents and from the web, constituting the *glossary G*;
3. (M) The lexicon and glossary are validated through a collaborative web application by all the members of the community, who are also asked to express a fine-grained evaluation of the definition *quality*;
4. (A) Definitions in the validated glossary *G* are parsed to extract hypernymy (kind-of) relations. Additional hypernymy relations are extracted from a general-purpose lexicalized taxonomy, WordNet [Fellbaum, 1998], and tailored to the domain using a word sense disambiguation algorithm. The validated set of hypernymy relations is used to automatically structure the terms in *L* into a forest *F* of taxonomically ordered sub-trees;
5. (M) A taxonomy management and validation web application is used to: i) manually create a taxonomy *C* of "core" domain terms ii) enrich *C* with the automatically created sub-trees *F*, and iii) allow a collaborative validation of the resulting taxonomy, *T*;
6. (M+A) The same application is being used (this is an in-progress activity) to let the taxonomy evolve towards the full power of an ontology. Again, automatic techniques are used to start the ontology enrichment process, followed by a validation and refinement task.

The idea behind this approach is that, despite many progresses in the area of ontology building and knowledge acquisition, automated techniques cannot fully replace the human experts and the stakeholders of a semantic web application. On the other side, manual ontology building methods as for example METHONTOLOGY [Fernández et al. 1997] or the NASA taxonomy development framework [Dutra and Busch, 2003] are very costly and require an effort in terms of time and competences, not affordable by loosely structured web communities. In our view, automated procedures are useful to achieve a significant speed-up factor in the development of semantic resources, but human validation and refinement is unavoidable when resources are to be used in real environments and applications.

In the rest of this paper, we overview the methodologies used for bootstrapping the knowledge acquisition process. Because of space restrictions, and because some of the methods have been already described in literature (see [Velardi et al., 2007] for steps 1 and 2 of the procedure outlined above), we provide details only on the taxonomy ordering algorithm (step 4). As far as the validation tasks are concerned, (steps 3 and 5), only the final results of the evaluation are presented and discussed; to obtain details, the interested reader is invited to access the deliverables of the project[3].

2.2 Learning a domain terminology and glossary

Web communities (groups of interest, web enterprises, research communities) share information in an implicit way through the exchange of mail, best practices, white papers, publications, announcements, etc. In INTEROP, many documents (state of arts, deliverables, workshop proceedings, etc.) have been stored on the network collaborative platform, like state of arts, deliverables, workshop proceedings, etc.. We applied to these documents a terminology extraction algorithm based on four measures: *Lexical Cohesion* [Park et al., 2002], *Domain Relevance* and *Domain Consensus* [Navigli and Velardi, 2004] and *Text Layout*. The algorithm puts together among the best available term extraction techniques in the literature, and

[2] The examples throughout this paper are in the domain of enterprise interoperability, but the reader can easily convince him/herself that the outlined procedure is fully general.

[3] http://interop-noe.org/backoffice/deliv/dg-1-interoperability-glossary

proved to have a very high precision[4] in different domains and applications [Navigli and Velardi, 2004]. The output of this phase is a *domain lexicon L*.

For each term t in *L*, candidate definitions are then searched in the document repository and on the web. Automatic extraction of definitions relies on an incremental filtering process:

1. (A) Definitions are firstly searched in existing web glossaries. If not found, simple patterns at the lexical level (e.g. "*t is a Y"*, "*t is defined as Y",* etc.) are used to extensively search an initial set of candidate definitions from web documents. Let D_t be the set of candidate definitions for each t in *L*.
2. (A) On the set D_t a first statistical filtering is applied, to verify *domain pertinence*. A statistical indicator of pertinence is computed for each definition $d_t \in D_t$, based on the number and statistical relevance of domain words (e.g. those in *L*) occurring in d_t;
3. (A) A subsequent *stylistic* filtering is applied, based on fine-grained regular expressions at the *lexical, part-of-speech* and *syntactic* level. The objective is to select "*well-formed*" definitions, i.e. definitions expressed in terms of *genus* (the *kind* a concept belongs to) and *differentia* (what specializes the concept with respect to its kind).

There are three advantages in applying the stylistic filtering criterion: i) To prefer definitions adhering to a uniform style, commonly adopted by professional lexicographers. For example, the following definition is not well-formed in the stated sense: "*component integration is obtained by composing the component's refinement structures together, resulting in (larger) refinement structures which can be further used as components*", ii) To be able to distinguish definitions from non-definitions (especially when candidate definitions are extracted from free texts, rather than glossaries). For example, "*component integration has been recently proposed to provide a solution for those issues*" is not a definition; iii) To be able to extract from definitions the *kind-of* information, subsequently used to help taxonomic ordering of terms. For example: "*In the traditional software engineering perspective, domain model is a precise* ***representation*** *of specification and implementation concepts that define a class of existing systems*" is well-formed, and its parsing returns the hypernym: *representation*.

The regular expressions used for stylistic filtering are domain-general, and the patterns are learned from definitions in professional glossaries on the web.

2.2.1 Collaborative Lexicon and Glossary Validation

During the subsequent evaluation phase, all INTEROP partners were requested, through a collaborative voting interface[5], first to validate the lexicon, rejecting inappropriate terms in *L*, and then to express their judgment on the definitions of the survived terms. The actual decision to reject or accept a term or definition was based on the sum of all expressed votes (*cumulated* vote). As far as definitions are concerned, the request was for a fine-grained voting of definition's *quality*. Votes were ranging from +1(adequate) to -3 (totally wrong). Partners were also requested to add missing definitions (the coverage of the automated gloss extraction procedure was about 70%) and to manually adjust some near-good definition. Table I summarizes the results of this phase. The performance is comparable with published results (e.g. [Park et al. 2002]) but the "real life" value of the experiment increases its relevance. In the literature, evaluation is mostly performed by two or three domain experts with adjudication, or by the authors themselves. In our case, the validation was performed by an entire research community with rather variable expertise (mainly: enterprise modeling, architectures and platforms, knowledge management) and different views of the interoperability domain.

Number of partners who voted the **lexicon**	35
Total expressed votes	2453
Accepted terms	1120 (59%)
Number of partners who voted the **glossary**	15
Total expressed votes	2164
Analysed definitions	1030
Accepted definitions	595 (57.76%)
Reviewed definitions[6] (cumulated vote =-1)	260 (25.24%)
Rejected definitions (cumulated vote <-1)	175 (17%)
New definitions added (terms without definition)	108

Table I. Result of collaborative lexicon and glossary evaluation.

2.2.2 Computing the Speed-up factor for the Glossary

The need to use automated glossary learning techniques in INTEROP was motivated by the absence of skilled personnel to create a high quality glossary, but most of all, by the fact that the knowledge domain of the INTEROP community was vaguely defined (as it often happens in emerging communities), making it particularly difficult to identify the truly pertinent domain concepts. However, as already remarked, the aim of the learning procedure described so far is not to replace humans, but to significantly reduce the time needed to build lexico-semantic resources.

To our knowledge, and after a careful study of the relevant literature, no precise data are available on glossary development costs, except for [Kon and Hoey, 2005] in which a cost of 200-300 dollars per term is estimated, but no details are given to motivate the estimate. We consulted several sources, finally obtaining the opinion of an experienced professional lexicographer[7] who has worked for many important publishers. The lexicographer outlined a three-step procedure for glossary acquisition including: i) internet search of terms ii) production of definitions iii)

[4] The interested reader can experiment through a web application, TermExtractor, available from http://lcl.di.uniroma1.it, which allows users to upload a document archive, obtain a domain terminology, and validate it.

[5] http://interop-noe.org/backoffice/workspaces/wpg/ps_interop

[6] All the definitions with a cumulated vote "-1" (i.e. only one partner expressed the opinion "not fully adequate") where manually adjusted.

[7] We thank Orin Hargraves for his very valuable comments.

harmonization of definitions style. The lexicographer evaluated the average time spent in each step in terms of 6 minutes, 10 min. and 6 min. per definition, respectively. Notice that the creation of a list of relevant terms (lexicon) is not included in this computation. The lexicographer also pointed out that conducting this process with a team of experts could be rather risky in terms of time, however he admits that in very new fields the support of experts is necessary, and this could significantly increase the above figures (however he did not provide an estimate of this increase). Starting from the professional lexicographer's figures, that clearly represent a sort of "best case" performance, we attempted an evaluation of the obtained speed-up. The glossary acquisition procedure has three phases in which man-power is requested: lexicon and glossary validation, and manual refinement of definitions. Each of these phases require from few seconds to few minutes, but actions are performed both on "wrong" and "good" data (with respect to the results of Table I, to obtain 83 "good" definitions, 100 must be inspected: 58 of them just accepted, 25 to be manually adjusted, etc.). We omit for the sake of space the details of the computation that led to over 50% speed up with respect to the lexicographer's estimate. In this comparison we exclude the stylistic harmonization (step (iii) of the lexicographer's procedure), which is indeed necessary to obtain a good quality glossary. However, since this phase would be necessarily manual in both cases, it does not influence the computation of the speed-up factor.

2.3 Learning taxonomic relations

The application of the *well-formedess* criterion discussed in section 2.2 (implemented with regular expressions), allows to extract from definitions the *kind-of* information, as defined by the author of a definition. This information may help structuring the terms of L in taxonomic order. However, ordering terms according to the hypernyms extracted from definitions has well-known drawbacks [Ide and Véronis, 1994]. Typical problems found when attempting to extract (manually or automatically) hypernymy relations from natural language definitions, are: over-generality of the provided hypernym (e.g. "Constraint checking is one of many *techniques*…"), unclear choices for more general terms, or-conjoined hypernyms (e.g. "Non-functional aspects define the overall *qualities* or *attributes* of a system"), absence of hypernym (e.g. "Ontological analysis is accomplished by examining the vocabulary that…"), circularity of definitions, etc. These problems – especially over-generality – are more or less evident when analysing the hypernyms learned through glossary parsing. To reduce these problems, we defined the following procedure:

1. (A) First, terms in the lexicon L are ordered according to simple *string inclusion*. String inclusion is a very reliable indicator of a taxonomic relation, though it does not capture all possible relations. This step produces a forest F of sub-trees. Let ST_{int} be one of such trees, for example:

integration
- representation integration
- model integration
  - enterprise model integration
- schema integration
- ontology integration
- knowledge integration
- data integration
- information integration

This "string inclusion" heuristics created a forest of 621 isolated trees out of the 1120 validated terms in L (cf. Table I).

2. (M) The trees in F are manually connected to a *core taxonomy*[8] C of high-level concepts, defined by a team of experts who basically reused WordNet and previous available work on enterprise ontologies[9]. Let $T_0=C \cup F$ be the resulting, fully connected, taxonomy.

In the INTEROP domain, C includes 286 concepts and T_0 includes 1406 nodes in total.

3. (A) The set of multi-word concept names in T_0 is decomposed in a list L' of singleton words, to which we added also the hypernyms automatically extracted from definitions. For example, if T_0 is the sub-tree ST_{int}, L' is: *representation, integration, model, data, ontology, specification, information*, etc. Terms in L' are used to search *hypernymy relations* in the WordNet sense inventory. For example:

   representation#n#2 → knowledge#n#1
   scheme#n#1 → representation#n#2
   data#n#1 → information#n#2
   workflow#n#1 → development#n#1

   All the WordNet word senses in the above example have a lexical counterpart in L'. Let R_{Wn} be the set of extracted hypernymy relations.

Some of the senses in R_{Wn} are not appropriate in the interoperability domain, e.g.: architecture#n#3 → activity#n#1, which refers to the "profession" sense of *architecture* rather than to *computer architecture* (sense #4 in WordNet). However, the objective is to apply these relations in a restrictive way, i.e. only to *sibling* terms in T_0. For example, the first rule of the above list can be used to move a term starting with "*representation*" below a term starting with "*knowledge*" iff if these two terms are siblings in some sub-tree of T_0 (e.g. in ST_{int}). The number of "applicable" rules is therefore reduced to a subset $R_{Wn}^{T_0} \subseteq R_{Wn}$.

In our domain, L' includes 607 different words, (since certain words occur many times in terminological strings), R_{Wn} includes 4015 *kind-of* relations, but $R^{T0}{}_{Wn}$ includes only 67 relations.

[8] "core" is a very basic and minimal taxonomy consisting only of the minimal concepts required to understand the other concepts.

[9] We omit details of this work, for the sake of space and because it is not central to the purpose of this paper.

4. (A) An on-line word sense disambiguation algorithm, SSI [Navigli and Velardi, 2005], is used to detect wrong senses[10] in $R^{T0}{}_{Wn}$, with respect to the domain. We use SSI to disambiguate each word in *L'* that appears in at least one of the kind-of relations in $R^{T0}{}_{Wn}$. The context for disambiguation is provided by co-occurring words in each sub-tree, e.g. in ST_{int}: *representation, integration, model*, etc. Let R_{SSI} be the relations in $R^{T0}{}_{Wn}$ survived after this step.

Step 4 returned 196 sense selections, which have been manually validated by two judges. 158 sense selections (80.62%) were judged as correct, given the domain.

5. (A) Relations in R_{SSI} are used to restructure T_0. For example, according to the relations available in R_{SSI} (e.g. those in the example of step 3), ST_{int} becomes:

knowledge integration
  representation integration
    schema integration
  model integration
    enterprise model integration
  information integration
    data integration
  ontology integration

Let T_1 be the resulting taxonomy after step 5. Following the learn-and-validate methodology adopted throughout the project, a web interface[11] has been developed to allow a collaborative validation of T_1. Table II provides a summary of the validation task.

Number of partners who voted the **taxonomy**		11
Total number of activated polls		21
Total number of performed actions		34
Of which:	Movement of single terms or term sub-trees	25
	Deleted core nodes	3
	Created core nodes	6

Table II. Results of collaborative taxonomy validation.

In Table II, "activated polls" refers to the fact that before making a change, partners need to activate a poll and receive consensus. The table shows that only 25 moves have been approved. A comparison between the number of actions performed by partners in Table I and Table II suggests that domain specialists can easily perform certain tasks (i.e. lexicon pruning) but are less confident when asked to contribute in creating progressively more "abstract" representations of their domain of expertise (from glossary to taxonomy and, eventually, to an ontology). This seems to further support the use of automated techniques.

3 Semantic Indexing and Semantic Search

The taxonomy created through the procedure illustrated so far has been used to semantically index the INTEROP KMap. Figure 3 shows the screen dump of a possible query type (*"find all the results – papers and projects – dealing with a subset of concepts in the taxonomy"*). The user can select concepts (referred to as *knowledge domains*, or simply *domains*, in the query interface) by "string search" in the taxonomy (as in the example of Figure 3), they can arrange concepts in boolean expressions, and perform query expansion (including in the query all or some of the concept's hyponyms).

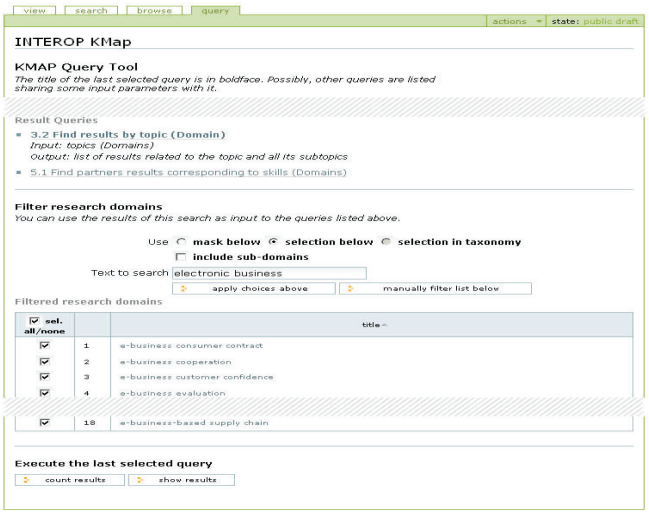

Figure 3. Taxonomy-based search of INTEROP Research Results.

It is also possible to obtain "global" information, e.g. a map of member's competence *similarity*, or an analysis of research results similarity. Figure 4 shows the screen dump of a graph in which nodes represent INTEROP organizations and the similarity value is highlighted by the thickness of edges. The number shown on each edge is the result of a *semantic similarity* computation (see [Velardi et al., 2007] for details). In short, the information (text or data) concerning each organization and its affiliated partners, is automatically parsed, and a weighted vector P_M of taxonomy concepts is associated to each member M. The well-known *cosine-similarity* measure is computed between vector pairs, but rather than considering only direct matches between terms, we also consider *indirect matches*, i.e. term pairs $t_x \in P_{M1}$ and $t_y \in P_{M2}$ related by direct ($t_x \rightarrow t_y$) or semi-direct hypernymy relations ($t_x \rightarrow t \leftarrow t_y$).
In the current version of the KMap (to be enhanced in the last year of the project) indirect matches represent a 38.41% of the total matches used to compute partner similarity.

4 Related Work and Concluding Remarks

This paper illustrated (in a forcefully sketchy way) a complete application of Semantic Web techniques to the task of modeling the competences of a web-based research community, INTEROP. We are not aware of any example of fully implemented knowledge acquisition value chain, where the acquired knowledge is first, extensively validated through the cooperative effort of an entire web community, and then, put in operation, to improve accessibility of web

[10] In principle, the domain appropriateness of the 67 hypernymy relations could be verified manually, given the limited dimension of the task, but we used a WSD algorithm for the sake of generality.

[11] Available online from http://lcl.di.uniroma1.it/tav

resources. The adopted techniques are fully general and the tools and interfaces developed within INTEROP can be applied to any other domain. For example, in the last year of the project the glossary learning procedure will be available as a web application and will be experimented by industrial partners to build glossaries in different business domains.

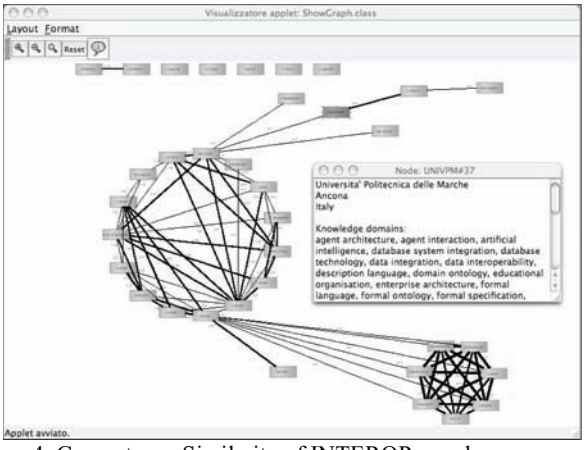

Figure 4. Competence Similarity of INTEROP members.

Given the wide spectrum of methodologies used (text mining, glossary and taxonomy enrichment, semantic indexing) a complete analysis of related work is impossible for space restrictions. We concentrate on what we consider the most original part of this work, taxonomy learning. Taxonomy learning is a three stage process: *terminology extraction* (e.g. [Park et al., 2002]), *glossary extraction* (like [Klavans and Muresan, 2001]), and finally, *extraction of hypernymy relations* between terms (among the others, the surveys in [Maedche et al., 2002] and [Cimiano et al., 2004]). While a variety of methods address specific phases of taxonomy learning, no published work addresses the complete process in all its aspects, like we do.

Another difference is the predominant use, in the literature, of trained machine learning methods [Miliaraki and Androutsopoulos, 2004]: the availability of training sets cannot be assumed in general, and, furthermore, preparing a training set requires professional annotators, like e.g. in TREC[12] contests.

The algorithms used to learn taxonomic relations are mostly based on the analysis and comparison of contextual features of terms, extracted from their occurrences in texts (see [Cimiano et al., 2004] for a comparison of different vector-based hierarchical clustering algorithms). Instead, we use a knowledge-based method that orders terms using kind-of relations extracted from their definitions and from a general-purpose semantic lexicon. The main advantage is that the principles that justify the resulting term ordering are *clear, consistently applied, easier to evaluate and modify*. On the contrary, evaluation of clustering algorithms is difficult and the results are hardly comparable: usually, the error rate of hypernymy extraction is over 40% [Caraballo, 1999; Widdows, 2003; Cimiano et al., 2004]. Furthermore, performance is evaluated with reference to the judgment of two-three human evaluators, often the authors themselves, rather than submitted to the user community, as we do.

In summary, the comparison with existing literature shows that the work presented in this paper promotes some progress in the automatic enrichment and use of semantic resources for knowledge management in real-world applications.

[12] The TREC (http://trec.nist.gov /data/qa.html) task relevant to our application is Question Answering (answering "what is" questions , i.e. to find definitions).

Acknowledgments

This work is partially funded by the Interop NoE (508011), 6th European Union FP.

References

[Caraballo, 1999] S. A. Caraballo. Automatic construction of a hypernym-labeled noun hierarchy from text. In *37th Annual Meeting of the Association for Computational Linguistics*, pages 120-126,1999.

[Cimiano et al., 2004] P. Cimiano, A. Hotho and S. Staab. Comparing Conceptual, Divisive and Agglomerative Clustering for learning Taxonomies from Text, In *16th ECAI 2004*, 22-27 August 2004, Valencia, Spain.

[Dutra and Busch, 2003] J. Dutra and J. Busch. Enabling Knowledge Discovery: Taxonomy Development for NASA. White Paper, 2003.

[Fellbaum, 1998] C. Fellbaum (ed.). *WordNet: an Electronic Lexical Database*, MIT Press, 1998.

[Férnandez et al., 1997] M. Fernández, A. Gómez-Pérez, N. Juristo. METHONTOLOGY: From Ontological Art Towards Ontological Engineering. In *Spring Symposium Series*. Stanford, pp. 33-40, 1997.

[Ide and Véronis, 1994] N. Ide and J. Véronis. Refining Taxonomies extracted from machine readable Dictionaries. *Research in Humanities Computing 2*, Oxford University Press, pp. 145-59, 1994.

[Klavans and Muresan, 2001] J. Klavans and S. Muresan. Text Mining Techniques for fully Automatic Glossary Construction, In *Proceedings of the HTL Conference*, San Diego (CA), March, 2001.

[Kon and Hoey, 2005] H. Kon and M. Hoey. Leveraging Collective Knowledge, In *Proc. of CIKM 2005*.

[Maedche et al., 2002] A. Maedche V. Pekar and S. Staab. Ontology learning part One: On Discovering Taxonomic Relations from the Web. In *Web Intelligence*, Springer, Chapter 1, 2002

[Miliaraki and Androutsopoulos, 2004] S. Miliaraki and I. Androutsopoulos. Learning to identify single-snippet answers to definition questions. In *Proceedings of COLING-2004*, pages 1360–1366, 2004.

[Navigli and Velardi, 2004] R. Navigli, P. Velardi. Learning Domain Ontologies from Document Warehouses and Dedicated Websites. *Computational Linguistics* (30-2), MIT Press, 2004.

[Navigli and Velardi, 2005] R. Navigli, P. Velardi. Structural Semantic Interconnections: a Knowledge-Based Approach to Word Sense Disambiguation, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 27(7), July, 2005

[Park et al., 2002] Y. Park, R. Byrd and B. Boguraev. Automatic glossary extraction: beyond terminology identification. In *19th Int. Conf. on Computational Linguistics*, Taipei, Taiwan, 2002.

[Velardi et al., 2007] P. Velardi , A. Cucchiarelli and M. Pétit. A Taxonomy Learning Method and its Application to Characterize a Scientific Web Community. To appear on IEEE Transactions on Knowledge and Data Engineering, 2007

[Widdows, 2003] D. Widdows Unsupervised methods for developing taxonomies by combining syntactic and statistical information. In *Proc. of HLT-NAACL 2003*, Edmonton, 2003.