AUTOMATIC ACQUISITION OF A THESAURUS OF INTEROPERABILITY TERMS

Roberto Navigli, Paola Velardi

Dipartimento di Informatica
Università di Roma "La Sapienza"
(navigli,velardi)@di.uniroma1.it

Abstract: Enterprise Interoperability is the ability of different organizations to work together and exchange documents, data, services, best practices, etc. A preliminary step towards achieving this goal is the assessment of a common vocabulary of relevant domain concepts, a sort of initial, semi-formal, representation of shared domain knowledge. In this paper we describe an experiment, conducted within the INTEROP Network of Excellence (NoE), aimed at automatically extracting a thesaurus of interoperability terms from the web and a corpus of available documents. Statistical and text mining techniques have been used to extract a glossary of relevant terms and term definitions, as well as a preliminary taxonomic structure of the glossary. The result has been evaluated by a team of partners belonging to the NoE.

Keywords: knowledge acquisition, semantic networks, artificial intelligence, domain analysis, search engines, computer-integrated enterprises, documents.

1. INTRODUCTION

INTEROP is a Network of Excellence whose primary goal is to sustain European research on interoperability for enterprise applications and software. The originality of the project lies in the multidisciplinary approach merging different research areas which support the development of interoperable systems: architectures and platforms, enterprise modelling, and ontology.

One of the most common goals of ontologies is sharing a common understanding of the structure of information among people or software agents (Musen 1992; Gruber 1993). For example, suppose several different web sites contain tourism information or provide tourism e-commerce services. If these sites share and publish the same ontology of the terms they all use, or, more realistically, they have different ontologies but provide mappings to a common upper level ontology, then computer agents can extract and aggregate information from these different resources. The agents can use this semantic information to answer user queries or as input data to other applications (Noy and Mc Guiness, 2001), e.g. automatic reservation services or trip planners.

A useful and commonly adopted initial step in ontology building is to obtain a list of relevant domain terms and term definitions, i.e. a glossary.

Often, natural language definitions introduce, first, the class an object belongs to (the *genus*) and then, the properties that characterize that object (the *differentia*) with respect to its class. Consider for example the first sentence of our abstract: "*Enterprise Interoperability is the ability of different organizations to work together and exchange documents, data, services, best practices, etc.*". Here, *ability* is the genus, or hyperonym (enterprise interoperabiliy is a kind of *ability*, the capacity to do something) and the description of this ability, i.e. the rest of the sentence, is the *differentia*. Notice furthermore that we said that ability is a capacity, and therefore the definition suggests a hierarchical chain:

enterprise_interoperability→ability→capacity.

In subsequent steps of the ontology building process, this hierarchy of terms can be used to populate the lower levels of a software artifact, the ontology, in which *terms* are replaced by *concepts* and informal descriptions by *formal specification*s expressed in some knowledge representation language, e.g. OWL, a W3C[1] ontology language standard.

In line with this strategy, during the first months of the INTEROP project it was decided to develop a hierarchically structured glossary of interoperability terms. The glossary is meant to provide common meta-data to annotate structured (databases) and unstructured (papers and deliverables) data produced by project work-packages dealing with the INTEROP knowledge map, researchers mobility, educational objectives, and state of the art. These data are progressively made available on the INTEROP collaborative workspace[2]. The glossary will eventually evolve towards an ontology in a later stage of the project.

To speed up the glossary definition, the Department of Computer Science of the University of Roma made available a battery of ontology building algorithms and tools, the OntoLearn system (Navigli and Velardi, 2004; Navigli et al., 2003). The OntoLearn system builds a domain ontology relating domain terms to the concepts and conceptual relations of the WordNet lexicalised ontology. The final ontology is therefore an extended and trimmed version of WordNet. In OntoLearn, WordNet acts as a “general purpose” upper ontology, but other more specialised upper ontologies can be used, if available.

Since the use of WordNet as a reference ontology is not a current choice of the INTEROP project, and since for the glossary acquisition task some additional feature was foreseen, we conceived a partly new experiment, using some of the tools and algorithms provided by the OntoLearn system, and some new feature that we developed for the purpose of the task at hand.

In this paper we describe the steps and the results of this experiment, that led to the acquisition of a hierarchically structured glossary of about 380 interoperability terms, subsequently evaluated by a team of 6 domain experts selected from INTEROP partners.

The paper is structured as follows: in Section 2 we give an overview of the adopted methods. In section 3 we present the results of the glossary acquisition experiment. Section 4 describes the hierarchical structuring of the glossary. Section 5 presents related work. Finally, Section 6 discusses foreseen developments in INTEROP.

2. OVERVIEW OF THE GLOSSARY ACQUISITION METHODOLOGY

Figure 1 provides a snapshot of the OntoLearn ontology learning methodology. The following steps are performed by the system:

1. *Extract pertinent domain terminology*

Simple and multi-word expressions are automatically extracted from domain-related corpora, like enterprise interoperability (e.g. *collaborative work*), hotel descriptions (e.g. *room reservation*), computer network (e.g. *packet switching network*), art techniques (e.g. *chiaroscuro*). Statistical and natural language processing (NLP) tools are used for automatic extraction of terms (Navigli and Velardi, 2004).

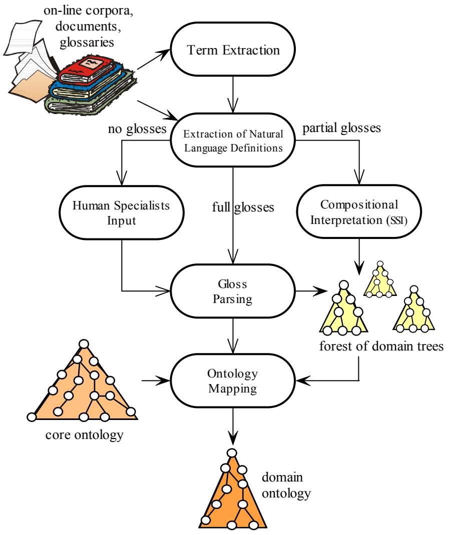

Fig. 1. An outline of the ontology learning phases in the OntoLearn system.

Statistical techniques are specifically aimed at simulating *human consensus* in accepting new domain terms. Only terms uniquely and consistently found in domain-related documents, and not found in other domains used for contrast, are selected as candidates for the domain terminology.

2. *Search on web available natural language definitions from glossaries or documents*

Available natural language definitions are searched on the web using on-line glossaries or extracting definitory sentences in available documents. A context free (CF) grammar is used to extract definitory sentences. An excerpt is:

S → PP ‘,’ NP SEP
NP → N1 KIND1
KIND1 → MOD1 NOUN1
MOD1 → Verb | Adj | Verb ‘,’ MOD1 | Adj ‘,’ MOD1
NOUN1 → Noun
N1 → Art | Adj
SEP → ‘,’ | ‘.’ | Prep | Verb | Wh
PP → Prep NP

In this example, *S*, *NP* and *PP* stand for sentence, noun phrase and prepositional phrase, respectively. KIND1 captures the portion of the sentence that identifies the kind, or genus, information in the definition.

This grammar fragment identifies (and analyses) definitory sentences like e.g.: “[In a programming language]$_{PP}$, [an *aggregate*]$_{NP}$ [that consists of data objects with identical attributes, each of which may be uniquely referenced by subscription]$_{SEP}$”, which is a definition of *array* in a computer network domain.

[1] http://www.w3.org/TR/2003/CR-owl-features-20030818/

[2] http://interop.aquitaine-valley.fr

The grammar is tuned for high precision, low recall. In fact, certain expressions (e.g. *X is an Y*) are overly general and produce mostly noise when used for sentence extraction.

3. *IF definitions are found*:

3.1 Filter out non relevant definitions

Multiple definitions may be found on the internet, some of which may be not pertinent to the selected domain (e.g. in the interoperability domain *federation* as "the forming of a nation" rather than "a common object model, and supporting Runtime Infrastructure."). A similarity-based filtering algorithm is used to prune out "noisy" definitions, with reference to a domain. Furthermore, an extension of the CF grammar of step 2 is used to select[3], when possible, "well formed" definitions. For example, definitions with *genus (kind-of)* and *differentia (modifier)*, like the *array* example in step 2, are preferred to definitions by example, like: *Bon a Tirer* "When the artist is satisfied with the graphic from the finished plate, he works with his printer to pull one perfect graphic and it is marked "Bon a Tirer," meaning "good to pull". These definitions can be pruned out since they usually do not match any of the CF grammar rules.

3.2 Parse definitions to extract kind-of information

The CF grammar of step 3.1 is again used to extract kind-of relations from natural language definitions. For example, in the *array* example reported in step 2, the same grammar rule can be used to extract the information (corresponding to the KIND1 segment in the grammar excerpt):

$$\text{array} \xrightarrow{kind-of} \text{aggregate}$$

4. *ELSE IF definitions are not found*

4.1 *IF definitions are available for term components* (e.g. no definition is found for the compound *integration strategy* but *integration* and *strategy* have individual definitions)

4.1.1 Solve ambiguity problems

In technical domains, specific unambiguous definitions are available for the component terms, e.g.: *strategy*: "a series of planned and sequenced tasks to achieve a goal" and *integration*: "the ability of applications to share information or to process independently by requesting services and satisfying service requests" (interoperability domain). In other domains, like tourism, definitions of component terms are often extracted from general purpose dictionaries (e.g. for *housing list*, no definitions for *list* are found in tourism glossaries, and in generic glossaries the word *list* is highly ambiguous). In these cases, a word sense disambiguation algorithm, called SSI (Navigli and Velardi, 2004 and 2004b), is used to select the appropriate meaning for the component terms.

4.1.2 Create definition compositionally

Once the appropriate meaning components have been identified for a multi-word expression, a generative grammar is used to create definitions. The grammar is based on the presumption (not always verified, see (Navigli et al., 2004) for a discussion) that the meaning of a multi-word expression can be generated compositionally from its parts. According to this compositional view, the syntactic head of a multi-word expression represents the *genus* (kind-of), and the other words the *differentia* (modifier). For example, *integration strategy* is a *strategy* for *integration*.

Generating a definition implies, first, to identify the *conceptual relations* that hold between the complex term components[4], and then, to compose a definition using segments of the components' definitions. For example, given the term *integration strategy*, the selected underlying conceptual relation is *purpose*:

$$\text{strategy} \xrightarrow{purpose} \text{integratim}$$

and the grammar rule for generating a definition in this case is:

(1) <MWE>:: = **a kind of** <H>, <HDEF>**, for** <M>, <MDEF>

where MWE is the complex term, H is the syntactic head, HDEF is the main sentence of the selected definition for H, M is the modifier of the complex term, and MDEF is the main sentence of the selected definition for M.

For example, given the previous definitions for *strategy* and *integration,* the following definition is generated by the rule (1):

integration strategy: a **kind of** *strategy*, a series of planned and sequenced tasks to achieve a goal, **for** *integration*, the ability of applications to share information or to process independently by requesting services and satisfying service requests.

As better discussed in (Navigli et al., 2004) this definition is quite verbose, but has the advantage of showing explicitly the sense choices operated by the sense disambiguation algorithm. A human supervisor can easily verify sense choices and reformulate the definition in a more compact way.

4.2 ELSE ask expert

If it is impossible to find even partial definitions for a multi-word expression, the term is submitted to human specialists, who are in charge of producing an appropriate and agreed definition.

5. *Arrange terms in hierarchical trees*

Terms are arranged in forests of trees, according to the information extracted in steps 3.2 and 4.1.1. Figure 2 shows examples from a computer domain.

[3] The grammar used for analysing definitions is a superset of the grammar used to extract definitions from texts. The analysed sentences are extracted both from texts and glossaries, therefore expressions like X is an Y must now be considered.

[4] Machine learning techniques are used to assign appropriate conceptual relations, see referenced papers for details.

6. *Link sub-hierarchies to the concepts of a Core Ontology*.

The semantic disambiguation algorithm SSI (mentioned in step 4.1.1) is used to append sub-trees under the appropriate node of a Core Ontology. In our work, we use a general purpose wide-coverage ontology, WordNet. This is motivated by the fact that sufficiently rich domain ontologies are currently available only in few domains (e.g. medicine). With reference to Figure 2, the root *artificial language* has a monosemous correspondent in WordNet, but *temporary or permanent termination* has no direct correspondent. The node is then linked to *termination*, but first, a disambiguation problem must be solved, since *termination* in WordNet has two senses: “end of a time span”, and “expiration of a contract”, therefore disambiguation is necessary.

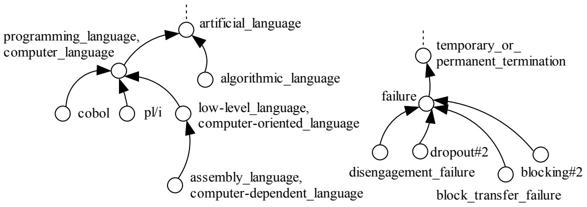

Fig. 2. Examples of taxonomic trees of terms (computer domain).

7. *Provide output to domain specialists for evaluation and refinement.*

The outcome of the ontology learning process is then submitted to experts for corrections, extensions, and refinement.

In the current version of OntoLearn, the output of the system is a taxonomy, not an ontology, since the only information provided is the kind-of relation. However, extensions are in progress, aimed at extracting other types of relations from definitions and on-line lexical resources.

3. THE INTEROP EXPERIMENT

For the purpose of the INTEROP glossary acquisition task, step 6 above has been omitted, since an interoperability Core Ontology was not available, and the adoption of an available reference ontology (like WordNet) is not agreed in the project.

The preliminary objective in this phase of INTEROP was to obtain a sort of partially structured glossary, rather than an ontology, i.e. a forest of term trees, where, for each term, the following information has to be provided: *definition* of the term, *source* of the definition (domain specialist or web site), *kind-of* relation, e.g.:

interoperability: The ability of information systems to operate in conjunction with each other encompassing communication protocols, hardware software, application, and data compatibility layers.
source: www.ichnet.org/glossary.htm
kind-of: ability

3.1 Term extraction

The first step of the INTEROP glossary procedure was to derive an initial list of terms using the evidence provided by interoperability-related documents. The INTEROP collaborative workspace was used to collect from all partners the relevant documents, among which, the proceedings of INTEROP workshops and deliverables, highly referenced scientific papers, partners’ papers, tutorials, etc. The OntoLearn TermExtractor module (Navigli and Velardi, 2004) extracted from these documents 517 terms. A generic computer science glossary was used to remove overly general technical terms (e.g. *computer network*), and the list was then quickly reviewed manually to delete clearly identifiable spurious terms. The final list included 376 terms.

3.2 Generation of definitions

Given the list of terms, we activated step 2 of the automatic glossary acquisition procedure. During this step, 28 definitions were not found, 22 were generated compositionally, and the remaining have been extracted either from glossaries or from available documents. For each definition, we kept track of the source (URL of the web page). For some term, more than one definition survived the well-formedness and domain similarity criteria (step 3.1 of section 2), therefore the total number of definitions submitted to the experts for revision was 358.

3.3 Evaluation by experts

Six domain experts[5] in INTEROP were asked to review and refine the glossary. Each expert could review (*rev*), reject (*rej*), accept (*ok*) or ignore (*blank*) a definition, acting on a shared database. The experts added new definitions for brand-new terms, but they also added new definitions for terms that may have more than one sense in the domain. There have been a total of 67 added definitions, 33 substantial reviews, and 26 small reviews (only few words changed or added). Some term (especially the more generic ones, e.g. *business domain, agent, data model*) was reviewed by more than one expert who proposed different judgements (e.g. ok and rev) or different revised definitions. In order to harmonise the results, a first pass was conducted automatically, according to the following strategy:

- If a judgement is shared by the majority of voters, then select that judgement and ignore the others (e.g. if a definition receives two *ok* and one *rev*, then, ignore *rev* and accept the definition as it is).
- If the only judgement is *rej*(ect), then delete the definition
- If a definition has a *rej* and one (or more) reviewed versions, then, ignore the reject and keep the reviews.

This step led to a final glossary including 425 definitions, 23 of which with a surviving ambiguity that could not be automatically conciliated. Therefore

[5] The experts have been chosen according to their expertise in the three INTEROP domains: ontology, architecture and enterprise modelling, but also to include representatives or leaders of the project work-packages that will actually *use* the glossary.

a second, short manual pass was necessary, involving this time only three reviewers.

3.4 Speed-up factors

The objective of the procedure describe in section 2 is to *speed-up* the task of building a glossary by a team of experts. Evaluating whether this objective has been met is difficult, since no studies are available for a comparison. We consulted several sources, finally obtaining the opinion of a very experienced professional lexicographer[6] who has worked for many important publishers. He outlined a three-steps procedure for glossary acquisition including: i) internet search of terms, ii) production of definitions, and iii) harmonization of definitions style. He evaluated the average time spent in each step in terms of 6 minutes, 10 min. and 6 min. per definition, respectively. He also pointed out that conducting this process with a team of experts could be rather risky in terms of time[7], however he admits that in very new fields the support of experts could be necessary.

Though the comparison is not fully possible, the procedure described in this paper has three phases in which man-power is requested:

- After term extraction (step 1), to prune non-terminological and non-domain relevant strings. This requires 0.5 minutes per term.
- After the extraction of definitions (step 2), to evaluate and refine definitions. We asked each expert to declare the time spent on this task, and we came out with an average of 4 minutes per definition. Since some definition was examined by more than one expert, this amount must be increased to 6 minutes approximately.
- In a second-pass review, to agree on the conflicting judgements. This depends on the number of conflicts, that in our case was less than 10%, mostly solved automatically (section 3.3). Overestimating, we may still add 1 minute per definition.

The total time is then 7.5 minutes per definition, against the 16 declared by the lexicographer for steps 1 and 2 of his procedure. In this comparison we exclude the stylistic harmonisation (step 3 of the lexicographer), which is indeed necessary to obtain a good quality glossary, but has not been conducted in the case of the INTEROP experiments.

However, since this phase would be necessarily manual in both cases, it does not influence the computation of the speed-up factor.

The above evaluation is admittedly very questionable, because on one side we have an experienced lexicographer, on the other side we have a team of people that are certainly experts of a very specific domain, but have no lexicographic skills. Our intention here was only to provide a very rough estimate of the manpower involved, given that no better data are available in literature. Apparently, a significant speed-up is indeed obtained by our procedure.

4. GENERATION OF DOMAIN SUB-TREES

As remarked in the introduction, the glossary terms must have some kind of hierarchical ordering, leading eventually to a formal ontology. A hierarchical structure simplifies the task of document annotation, and is a basis for further developments such as automatic clustering of data (e.g. for document classification), identification of similarities (e.g. for researchers mobility), etc. In other words, it is a first step towards semantic annotation.

To arrange terms in term trees, we used the procedure described in steps 3.2 and 4.1.1 of section 2. The definitions have been parsed and the word, or complex term, representing the hyperonym (genus) has been identified. Given the limited number of definitions, we verified this task manually, obtaining a figure of 91.76% precision, in line with previous evaluations that we did on other domains (computer networks, tourism, economy). Contrary to the standard OntoLearn algorithm, we did not attached sub-trees to WordNet, as motivated in previous sections.

Overall, the definitions were grouped in 125 sub-trees, of which 39 including only 2 nodes, 43 with 3 nodes, and the other with >3 nodes. Examples of two term trees are shown in figure 3.

In the figure (tree on top), the collocation of the term *system* might seem inappropriate, since this term has a very generic meaning. However, the definition of *system* in the interoperability glossary is quite specific: "*a set of interacting components for achieving common objectives*", which justifies its collocation in the tree. A similar consideration applies to *service* in the bottom tree.

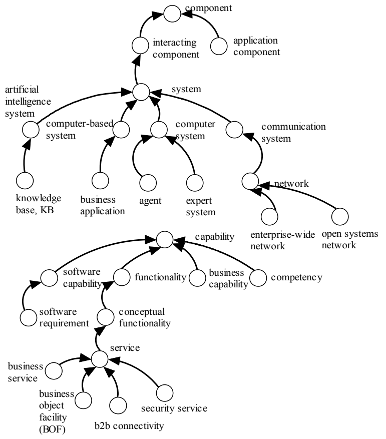

Fig. 3. Sub-trees extracted from the Interoperability domain.

An interesting paper (Ide and Véronis, 1993) provides an analysis of typical problems found when attempting to extract (manually or automatically) hyperonymy relations from natural language definitions, e.g. attachments too high in the

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

[7] To cite his words: "no commercial publisher would subject definitions to a committee for fear of never seeing them in recognizable form again"

hierarchy, unclear choices for more general terms, or-conjoined heads, absence of hyperonym, circularity, etc. These problems are more or less evident – especially over-generality – when analysing the term trees forest extracted from the glossary. However, our purpose here is not to overcome problems that are inherent with the task of building a domain concept hierarchy: rather, we wish to automatically *extract, with high precision, hyperonymy relations* embedded in glossary definitions, just as they are: possibly over-general, circular, or-conjoined. The target is, again, to speed up the task of ontology building and population, extracting and formalizing domain knowledge expressed by human specialists in an unstructured way. Discrepancies and inconsistencies can be corrected later by the human specialists, who will verify and rearrange the nodes of the forest. The study of automatic methods to rearrange trees and reduce these discrepancies is one of our on-going research streams.

5 RELATED WORK

The major papers in the area of ontology and, specifically, taxonomy construction propose methods to extend an existing ontology with unknown words, e.g. (Aguirre et al., 2000), (Morin, 1999), etc. based on context similarity in running texts.

Alfonseca and Manandhar (2002) present an algorithm to enrich WordNet with unknown concepts on the basis of hyponymy patterns detected in texts. Berland and Charniak (1999) propose a method to extract whole-part relations from corpora and enrich an ontology with this information. Other papers describe methods to extensively enrich an ontology with domain terms. For example, Vossen (2001) uses statistical methods and string inclusion to create lexicalized trees. However, no semantic disambiguation of terms is performed. Very often, in fact, ontology learning algorithms regard domain terms as domain concepts.

In comparison with state-of-art literature, our method provides a richer set of methods for extensive ontology learning as well as a support for algorithmic and human evaluation.

6 CONCLUDING REMARKS AND FUTURE WORK

As already remarked, the glossary provides a first set of shared terms to be used as metadata for annotating documents and data in the INTEROP platform. Several features/improvements are foreseen to improve this initial result, both on the interface/architecture and the methodological side. For example, annotation tools must be defined and integrated in the INTEROP platform. The taxonomic structuring of the glossary must be manually reviewed in the light of a core ontology to be defined, and methods to include new terms must be provided. Finally, the use of terms (and later, of ontology concepts) for document access, clustering and retrieval must be implemented and evaluated.

ACKNOWLEDGEMENTS

This experiment was made possible thanks to the support of all INTEROP partners. Special thanks go the reviewers Raul Poler, Michael Pétit, Michele Missikoff, Valeria De Antonellis, Yves Ducq, Giuseppe Berio. Special thanks to Orin Hargraves for his valuable description of the lexicographer task.

REFERENCES

Alfonseca, E. and Manandhar, S. (2002). Improving an Ontology Refinement Method with Hyponymy Patterns. Language Resources and Evaluation (LREC-2002), Las Palmas, Spain, May 2002.

Agirre, E., Ansa, O., Hovy, E., and Martinez, D. (2000). Enriching very large ontologies using the WWW, in *ECAI Ontology Learning Workshop*.

Berland M. and Charniak, E. (1999). Finding Parts in Very Large Corpora. In *Proceedings of the the 37th Annual Meeting of the Association for Computational Linguistics* (ACL-99).

Gruber, T.R. (1993). A Translation Approach to Portable Ontology Specification. *Knowledge Acquisition* **5**, pp. 199-220.

Ide N. and Véronis J. (1993). Refining Taxonomies extracted from machine readable Dictionaries, In Hockey, S., Ide, N. *Research in Humanities Computing*, **2**, Oxford University Press.

Morin, E. (1999). Automatic Acquisition of semantic relations between terms from technical corpora, *Proc. of 5th Terminology and Knowledge Engineering* (TKE-99).

Musen, M.A. (1992). Dimensions of knowledge sharing and reuse. *Computers and Biomedical Research* **25**, pp. 435-467.

Navigli R., Velardi P and Gangemi A. (2003). Ontology Learning and its Application to Automated Terminology Translation. *IEEE Intelligent Systems*, vol. **18**, pp. 22-31

Navigli R. and Velardi P. (2004). Learning Domain Ontologies from Document Warehouses and Dedicated Web Sites. *Computational Linguistics*, vol. **50** (2).

Navigli R. and Velardi P. (2004b). Structural Semantic Interconnection: a knowledge-based approach to Word Sense Disambiguation, Proc. *3rd Workshop on Sense Evaluation*, Barcelona.

Navigli R., Velardi P., Cucchiarelli A. and Neri F. (2004). Quantitative and Qualitative Evaluation of the OntoLearn Ontology Learning System. *Proc. 20th COLING 2004*, Geneva.

Noy N. F. and D. L. McGuinness (2001). Ontology Development 101: A Guide to Creating Your First Ontology. *Stanford Knowledge Systems Laboratory Technical Report* KSL-01-05 and *Stanford Medical Informatics Technical Report* SMI-2001-0880, March 2001.

Vossen P. (2001). Extending, Trimming and Fusing WordNet for technical Documents, *NAACL 2001 workshop on WordNet and Other Lexical Resources*, Pittsburgh, July 2001.