The Workshop Programme

9.00-9.20 **Welcome**

9.20 - 11.00 **Session 1**

9.20-9.40: Ontology-based question analysis in a multilingual environment: The MOSES project (Paggio, Hansen, Basili, Pazienza, Zanzotto)

9.40-10.00: A Content Creation Process for the Semantic Web (Hyvonen, Salminen, Junnila, Kettula)

10.00-10.20: Proper Nouns Ontology for Document Retrieval and Question Answering (de Loupy, Crestan, Lemaire)

10.20-10.40: A logic-based system for textual document classification (Cumbo, Ettorre, Iiritano, Mastratisi, Ruffolo, Rullo)

10.40-11.00: Ontological Types of Associative Relations in Information Retrieval Thesauri and Automatic Query Expansion (Loukachevitch, Dobrov)

11.00 - 11.30 **Coffee Break**

11.30 - 13.00 **Panel**

The vision of the Semantic Web is to turn the Internet into a machine understandable resource so that digital information can be accessed in a more flexible, content-based fashion. Annotation of texts, data and services with ontological information constitutes one of the backbones of the Semantic Web technology and to this end, lexical resources - preferably multilingual – can provide the basic substratum. In this context, a central issue concerns the trade off between representing ontological information through formal languages and maintaining the richness of natural language to characterize ontological categories and relations. Therefore, the HLT and Semantic Web communities could profit from joining forces in order to guide each other into the next generation of content-based web-technology.

Panel Members: Maria Teresa Pazienza, Aldo Gangemi/Alessandro Oltramari, Kiril Simov, Alessandro Lenci, Bernardo Magnini

LUNCH

14.30 - 15.30 **Posters and Demos**

Posters:

- Building Business Ontologies with Language Technology Techniques – The VID Project (Pedersen, Navarretta, Henriksen)
- Natural Language Expression of User Policies in Pervasive Computing Environments (Weeds, Keller, Weir)
- Reconstructing the Ontology of the Tang Dynasty – a pilot study of the Shakespearean-garden Approach (Chu-Ren Huang, Feng-ju Lo, Ru-Yng Chang, Sueming Chang)

Demo:

- OntoTag's Linguistic Ontologies as a Reference for Semantic Web

Annotations (de Cea, de Mon, Pareja-Lora)

- Sinica BOW: Integrating bilingual WordNet and SUMO Ontology (Chu-Ren Huang)
- Searching and Browsing Collections of Finnish Museums on the Semantic Web (Hyvonen, Junnila, Kettula, Makela, Saarela, Salminen, Syreeni, Valo, Viljanen)

15.30-16.30 **Session 2/a**

15.30-15.50: Context Related Derivation of Word Senses (Kunze, Rosner)

15.50-16.10: Automatic Thai Ontology Construction and Maintenance System (Kawtrakul, Suktarachan, Imsombut)

16.10-16.30: An Ontology for Multilingual Treatment of Proper Names (Tran, Grass, Maurel)

16.30-17.00 **Coffee Break**

17.00-18.00 **Session 2/b**

17.00-17.20: Integrating Semantic Lexicons and Domain Ontologies (Basili, Vindigni, Zanzotto)

17.20-17.40: Ontological Knowledge and language in modelling classical architectonic structures (Cappelli, Giovannetti, Michelassi)

17.40-18.00: Comparison of Principles Applying to Domain Specific versus General Ontologies (Madsen, Thomsen, Vikner)

18.00-18.15 **Conclusion**

Workshop Organisers

Alessandro Oltramari
Laboratory for Applied Ontology, ISTC-CNR; Department of Cognition and Education Sciences, Trento University, oltramari@loa-cnr.it
Patrizia Paggio
Center for Sprogteknologi, University of Copenhagen, patrizia@cst.dk
Aldo Gangemi
Laboratory for Applied Ontology, ISTC-CNR Rome, a.gangemi@istc.cnr.it
Maria Teresa Pazienza
Roma Tor Vergata University, pazienza@info.uniroma2.it
Nicoletta Calzolari
Istituto di Linguistica Computazionale del CNR, glottolo@ilc.cnr.it
Bolette Sandford Pedersen
Center for Sprogteknologi, University of Copenhagen, bolette@cst.dk
Kiril Simov, Bulgarian
Academy of Sciences, kivs@bultreebank.org

Workshop Programme Committee

Roberto Basili (Roma Tor Vergata University)
Werner Ceusters (Language & Computing)
Nicoletta Calzolari (Istituto di Linguistica Computazionale del CNR)
Aldo Gangemi (Laboratory for Applied Ontology, ISTC-CNR, Rome)
Eric Gaussier (Xerox Research Centre Europe, Grenoble Laboratory)
Maria Toporowska Gronostaj (Språkdata, University of Gothenburg)
Nicola Guarino (Laboratory for Applied Ontology, ISTC-CNR, Trento)
Arne Jönsson (Linköping Universitet)
Dimitrios Kokkinakis (Språkdata, University of Gothenburg)
Alessandro Lenci (Universitá di Pisa)
Claude de Loupy (Sinequa and University of Paris 10)
Bernardo Magnini (ITC-IRST, Trento)
Jørgen Fischer Nilsson (Technical University of Denmark)
Alessandro Oltramari (Laboratory for Applied Ontology, ISTC-CNR, Trento)
Patrizia Paggio (Center for Sprogteknologi)
Maria Teresa Pazienza (Roma Tor Vergata University)
Bolette Sandford Pedersen (Center for Sprogteknologi)
Guus Schreiber (Vrije Universiteit Amsterdam)
Kiril Simov (Bulgarian Academy of Sciences)
Paola Velardi (Università “La Sapienza”, Rome)

Table of Contents

Author Index

Ontology-based question analysis in a multilingual environment: the MOSES case study

Patrizia Paggio (**), Dorte H. Hansen (**),

Roberto Basili (*), Maria Teresa Pazienza (*), Fabio Massimo Zanzotto (*)

(*) Dip. di Informatica Sistemi e Produzione
University of Rome "Tor Vergata"
{basili,pazienza,zanzotto}@info.uniroma2.it

(**) Centre for Language Technology
University of Copenhagen
{patrizia,dorte}@cst.dk

Introduction

Question Answering (QA) systems (as QA track of the Text Retrieval conference (TREC-QA) competitions (Voorhees 2001)), are able both to understand questions in natural language and to produce answers in the form of selected paragraphs extracted from very large collections of text. Generally, they are open-domain systems, and do not rely on specialised conceptual knowledge, while using a mixture of statistical techniques and shallow linguistic analysis. Ontological Question Answering systems, e.g. (Woods et al. 1972, Zajac 2000) attack the problem by using an internal unambiguous knowledge representation. As any knowledge intensive application, ontological QA systems have as intrinsic limitation the small scale of the underlying syntactic-semantic models of natural language. While limitations are well-known, we are still questioning if any improvement has occurred since the development of LUNAR, the first ontological QA system. Several important facts have emerged that could influence related research approaches:

- a growing availability of lexical knowledge bases that model and structure words: WordNet (Miller 1995) and EuroWordNet (Vossen 1998) among others; some open-domain QA systems have proven the usefulness of these resources (e.g. Harabagiu et al. 2001);
- the vision of a Web populated by "ontologically" tagged documents which the semantic Web initiative has promoted; this would require a world-wide collaborative work for building interrelated "conceptualisations" of domain specific knowledge;
- the trend in building shallow, modular, and robust natural language processing systems (Abney 1996, Hobbs et al. 1996, Ait-Moktar&Chanod 1997, Basili&Zanzotto 2002) which is making them appealing in the context of ontological QA systems, both for text interpretation (Andreasen et al. 2002) and for database access (Popescu et al. 2003).

In such a new fascinating context, we are investigating a novel approach to ontology-based QA in which users ask questions in natural language to knowledge bases of facts extracted from a federation of Web sites and organised in topic map repositories (Garshol 2003). Our approach is investigated in the context of the EU project MOSES[1], with the explicit objective of developing an ontology-based methodology to search, create, maintain and adapt semantically structured Web contents according to the vision of the Semantic Web. The test-bed chosen in the project is related to the development of an ontology-based knowledge management system and an ontology-based search engine that will both accept questions and produce answers in natural language for the Web sites of two European universities. Challenges of the project are:

- developing an ontological QA system;
- supporting a multilingual environment which implies the ability to treat several languages, and, crucially, several conceptualisations.

In this paper, after briefly describing how the project is trying to comply with the semantic Web vision, we will focus on question processing, and in particular on the way in which NLP techniques and ontological knowledge interact in order to support questions to specific sites or to site federations.

An ontology-based approach to question answering

In our ontological QA system, both questions and domain knowledge are represented through the same ontological language. QA system will be developed in two steps: firstly a prototypical implementation is planned to answer questions related to the current "state-of-affairs" of the site to which the question is posed; secondly step, given a "federation" of sites within the same domain, we will investigate how to support QA across the sites. Answering a question can then be seen as a collaborative task among ontological nodes belonging to the same QA system. Since each node has its own version of the domain ontology, the task of passing a question from node to node may be reduced to a mapping task between (similar) conceptual representations. To make such an approach feasible, a number of difficult problems must still be solved. In this paper, we will provide details on how:

- to build on existing ontologies by interfacing between them and language resources;
- to interpret questions wrt the ontological language;
- to model the mapping task for federated questions.

[1] MOSES is a cooperative project under the 5th Framework Programme. The project partners are FINSA Consulting, MONDECA, Centre for Language Technology, University of Copenhagen, University of Roma Tre, University of Roma Tor Vergata and ParaBotS.

Building on off-the-shelf semantic Web ontologies

One of the results of the Semantic Web initiative will be the production of several interrelated domain-specific ontologies that provide the formal language for describing the content of Web documents. In spite of the freedom allowed in the production of new conceptualisations, it is reasonable to expect that a first knowledge representation jungle will leave room to a more orderly place where only the most widely shared conceptualisations have survived. This is a prerequisite for achieving interoperability among software agents. In view of this, and since publicly available non-toy ontology examples are already available, the effort of adapting an existing ontology to a specific application is both useful and possible. This experiment is being conducted in MOSES to treat the university domain.

The conceptualisation of the university world provided in the DAML+OIL ontology library is an interesting representation for the application scenarios targeted in MOSES (i.e. *People/Course/Research*), and has therefore been used as starting point to develop the project's ontologies. Described classes and relations cover in fact, at least at a high level, most of the relevant concepts of the analysed scenarios. The ontology has been adapted to develop conceptualisations for each of the two national university sub-systems (i.e. Italian and Danish) while providing additional information required for answering the input questions. To give an idea of the coverage achieved, the Danish ontology contains about 200 classes and 50 relations. Instances of the classes are being added by the project's user groups by downloading them from the respective sites' databases as well as by manually extracting data from the Web pages. This manually acquired knowledge will be used to develop machine learning algorithms that will allow a semi-manual construction of the domain knowledge.

The first challenge deriving from having two separate ontologies for the same domain is the language. Whereas concept and relation labels in the Italian ontology are expressed either in English (for concepts directly taken from the original source) or in Italian, in the Danish counterpart all labels are in Danish. This means that a mapping algorithm making use of string similarity measures applied to concept labels will have to work with translation, either directly between the two languages involved, or via a pivot language like English. The goal would be to establish correspondences such as 'Lektor' ↔ ('AssociateProfessor') ↔ 'ProfessoreAssociato'.

Another challenge comes from the structural differences: not all the nodes in one ontology are represented also in the other and vice-versa; moreover, nodes modelling concepts that seem intentionally "equivalent", may have different structural placements. This is the case for the 'Lektor'/'ProfessoreAssociato' pair just mentioned: in the Danish system, 'Lektor' is not a subclass of 'Professor', although *associate professor* is considered a correct translation.

Finally, domain relations are treated somewhat differently in the two ontologies. In the Italian one, all relations are binary in keeping with the original DAML-OIL model, whereas the Danish ontology makes use of n-nary relations in the spirit of the Topic Maps (Garshol. 2003) formalism.

Linguistic interfaces to ontologies

Ontologies for the Semantic Web are written in formal languages (OWL, DAML+OIL, SHOE) that are generalisations/restrictions of Description Logics (Baader et al. 2003). TBox assertions describe concepts and relations. A typical entry for a concept is:

ID	Course
Label	Course
Subclassof	Work

Table 1 A concept

where ***ID*** is the concept unique identifier, ***label*** is the readable name of the concept, ***subclassof*** indicates the relation to another class. As the label has the only purpose of highlighting the concept to human readers, alternative linguistic expressions are not represented. On the contrary, this piece of information is recorded in a lexical data base like WordNet. The problem is even more obvious when considering relationships.

ID	teacherOf
Label	Teaches
Domain	#Faculty
Range	#Course

Table 2 A relationship

In Table 2, *Domain* and *Range* contain the two concepts related to the described binary relation. The label *teacherOf* does not mention alternative linguistic expressions like: *#Faculty* ***gives*** *#Course* or *#Faculty* ***delivers*** *#Course*, etc.

For the ontology producers, only one concept or relation name is sufficient. Synonymy is not a relevant phenomenon in ontological representations. In fact, it is considered a possible generator of unnecessary concept name clashes, i.e. concept name ambiguity. Conceptualisations (as in tables 1,2) are, however, inherently weak whenever used to define linguistic models for NLP applications. Interpreting questions like:

(1) Who **gives**/**teaches** the database **class**/**course** this year?

with respect to a university domain ontology means in fact mapping all four questions onto the concepts and relations in Table 2. There is a gap to be filled between linguistic and ontological ways of expressing the domain knowledge.

In developing an ontological QA system, the main problem is then to build what we call the "linguistic interface" to the ontology, which consists of an explicit mapping between linguistic expressions and the concepts

and relationships they convey. To make this attempt viable, we are currently studying methods to automatically relate lexical knowledge bases like WordNet (Miller 1995) to domain ontologies (Basili et al 2003a) and to induce syntactic-semantic patterns for relationships (Basili et al 2003b). In the current phase of the project, however, the linguistic interface is being created manually. The approaches used to treat the two languages differ with respect to both formalism and tools, as well as in the way in which syntactic and semantic analyses are combined. In both cases, however, the central aspect is the mapping between conceptual knowledge and alternative linguistic expressions in which this knowledge can be conveyed.

Classifying questions

To facilitate recognition of what are the relevant expressions to be encoded in the linguistic interface, the project's user groups[2] identified a corpus of possible questions in each of the two languages supported by the system, and classified them in co-operation with the system's developers. A classification often quoted is that in Lauer et al. (1992), which mainly builds on speech act theory. Another influential, more syntactically-oriented approach is that in Harabagiu et al. (2001) where to each syntactic category correspond one or several possible answer types, or focuses (a person, a date, a name, etc.).
Several dimensions have been identified as relevant for MOSES and explored in "question cards" by the user groups:

1. the number of sites and pages in which the answer is to be found. Thus, a first distinction is done between site-specific and federated questions. In the first case, analysis involves only one language and one knowledge domain. In the second, the interpretation of a question produced by a local linguistic analyser is matched against the knowledge domain of other sites;
2. sub-domain coverage (e.g. people, courses, research).
3. format of the answer: in MOSES the answer consists not only of a text paragraph as in standard QA, but could also be composed of one or more instances of semantic concepts (professors, courses) or relations (courses being taught by specific professors), whole Web pages, tables, etc. due to the heterogeneity of information sources

From the point of view of the linguistic analysis, however, syntactic category and content are the central dimensions of sentence classification. Syntactic categories are e.g. *yes/no question, what-question, who-question, etc.* Subtypes relate to the position inside the question where the focus is expressed, e.g. depending on whether the wh-pronoun is a determiner, or the main verb is a copula. The content consists of concepts and relations from the ontology, the focus constraint[3] (the ontological type being questioned), and a count feature indicating the number of instances to be retrieved. Table 3 shows an example of linguistic classification. For each sentence type, several paraphrases are described.

FORM 1	
Input	Hvem underviser i filmhistorie (*Who teaches film history)*
Syntactic type	Who (Hvem)
Syntactic subtype	V ≠ copula
CONTENT	
Focus constraint	Teacher
Concepts	Faculty Course.Name: *history of film*
Relations	TeacherOf(Faculty, Course)
Answer count	List

Table 3: Example of question classification

Question analysis

Question analysis is carried out in the MOSES linguistic module associated with each system node. To adhere to the semantic Web approach, MOSES poses no specific constraints on how the conceptual representation should be produced, nor on the format of the output of each linguistic module. The agent that passes this output to the content matcher (an ontology-based search engine) maps the linguistic representation onto a common MOSES interchange formalism (still in an early development phase). Two independent modules have been developed for Danish and Italian language analysis. They have a similar architecture (both use preprocessing, i.e. POS-tagging and lemmatising, prior to syntactic and semantic analyses), but specific parsers. Whereas the Danish parser, an adapted version of PET (Callmeier 2000) produces typed feature structures (Copestake 2002), the Italian one outputs quasi-logical forms. Both representation types have proven adequate to express the desired conceptual content.
The question analysis components have not yet been fully integrated within the overall system. Therefore, at the current stage the only possible evaluation, is an account of the syntactic and semantic coverage of each system in isolation, where coverage is defined in terms of a subset of the questions constructed by the user groups.
The Italian system has been tested on a test set consisting of 83 questions yielding 58 (70%) semantically correct analyses and 21 partial analyses. Sources of complexity are mainly due to the presence of temporal pronouns not correctly handled.
The Danish system has been tested on 85 questions, out of which 65 (= 76%) are correctly analysed, 3 get incorrect analyses and 17 get no analyses. The fact that 17 questions yield no analysis is due to lack of grammatical coverage concerning genitives, complex nominal heads, split NPs and relative clauses without relative marker. Unlike the Italian system, the Danish one doesn't have a robustness mechanism outputting partial analyses at the moment. The plan is to use POS-tags along with dictionary and thesaurus look-up in order to produce fragmented output for further use in the search process.

[2] The University of Roma III and the Faculty of Humanities at the University of Copenhagen.
[3] In the sense of Rooth (1992).

Analysis of Italian questions

Analysis of Italian questions is carried out by using two different linguistic interpretation levels. The syntactic interpretation is built by a general purpose robust syntactic analyser, i.e. Chaos (Basili&Zanzotto 2002). This will produce a Question Quasi-Logical Form (Q-QLF) of an input question based on the extended dependency graph formalism (XDG) introduced in (Basili&Zanzotto 2002). In this formalism, the syntactic model of the sentence is represented via a planar graph where nodes represent constituents and arcs the relationships between them. Constituents produced are chunks, i.e. kernels of verb phrases (VPK), noun phrases (NPK), prepositional phrases (PPK) and adjectival phrases (ADJK). Relations among the constituents represent their grammatical functions: logical subjects (lsubj), logical objects (lobj), and prepositional modifiers. For example, the Q-QLF of the question

(2) Chi insegna il corso di Database?
(Who teaches the database course?)

is shown in Figure 1.

Figure 1 A Q-QLF within the XDG formalism

Then a robust semantic analyser, namely the Discourse Interpreter from LaSIE (Humphreys et al. 1996) is applied.

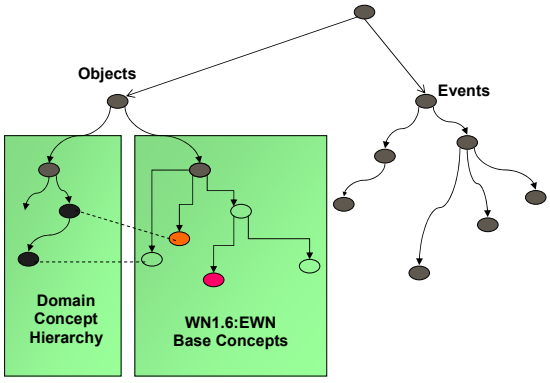


Figure 2 The world model taxonomy

An internal world model has been used to represent the way in which the relevant concepts (i.e. objects) and relationships (i.e. events) are associated with linguistic forms (see Figure 2). Under the Objects node, concepts from the domain concept hierarchy are mapped onto synsets (sets of synonyms) in the linguistic hierarchy EWN (i.e. the EuroWordNet.base concepts). This is to guarantee that linguistic analysis is carried out using already existing general lexical knowledge.

The association of objects and events with linguistic forms is used in matching rules as shown in Figure 3. The rule expresses the fact that, if any of the words *tenere*, *insegnare* or *fare* is encountered in relation with a *human_1* (represented by the base concept *ewn4123*) and the word *education_1* (*ewn567704*), the relation ***teacherOf*** can be induced.

```
TEACH_EVENT ==> teach_course.
        teach_course ==> tenere v insegnare v fare.

props(teach_course(E),[
        (consequence(E,
        [relation(E,teacherOf),r_arg1(E,X),r_arg2(E,Z)]     ):-
                nodeprop(E,lsubj(E,X)),
                X <- ewn4123(_), /* human_1 */
                nodeprop(E,lobj(E,Z)),
                Z <- ewn567704(_)              /* education_1 */
        )]).
```

Figure 3 Example of syntactic-semantic interpretation rule

The analysis resulting for sentence (2) is then:

```
focus(e2),
relation(e1,teacherOf),
r_arg1(e1, person_dch(e2)),
r_arg2(e1,course_dch(e3)),
relation(e4,hasSubject),
r_arg1(e4, course_dch(e3)),
r_arg2(e4,topic_dch("Database")).
```

This means that the user is interested in a person, the entity *e2* of the class person_dch, that is in a relation teacherOf with the entity *e4* (instance of the class course_dch), that is in turn related by hasSubject with the topic (i.e. topic_dch) "Database". This result can be passed on to the content matcher.

Analysis of Danish questions

Danish linguistic analysis consists of a preprocessing and a parsing step. The job of preprocessing is to prepare the linguistic input for semantic parsing by applying a number of shallow NLP techniques: tokenisation, named entity recognition, part-of-speech tagging and lemmatisation. The parser builds a semantic representation of the input in terms of semantic relations among domain concepts.

The parser uses a typed-feature structure formalism very similar to that of Head-Driven Phrase Structure Grammar (Pollard & Sag, 1994) which has the central characteristic of being able to conflate syntactic and semantic information in one and the same feature structure representation. In other words, rather than producing a syntactic structure (a parse tree) first, and then the semantic representation corresponding to this structure as done in many other systems (e.g. in the Italian analysis), here syntax and semantic analysis proceed in an integrated fashion, and yield a unified output result. This strategy ensures that the parser does not produce syntactic structures that are not semantically well-formed. Only the semantic part of the result is sent on to the content matcher. An example are the semantics produced for the

Danish equivalent of "Who teaches database theory at the faculty of humanities?" (Danish conceptual labels have been replaced by English ones throughout this section):

```
(3)  [FOCUS-CONST #1: Faculty
      COUNT: all
      LOA      <   CourseOffer
                   [COURSE   #2: Course
                    TEACHER  #1: Faculty
                    PROVIDER: FacultyOrg
                        [NAME "Humanities"]]
                   CourseSubject
                   [SUBJECT: Subject
                        [NAME
                          "databaseTheory"]
                    WORK    #2: Course]  > ]
```

The value of the attribute FOCUS-CONST indicates the class to which the instances we are looking for belong, here the class of university teachers (Faculty). The index "#1" shows that the teacher in focus plays the TEACHER role in the relation CourseOffer. The value of COUNT indicates that we want all relevant instances. LOA (List Of Associations) is a list of relations expressing constraints on the instances to be found. In this case, the teacher must be related to a course as well as the Faculty of Humanities via a CourseOffer relation, and the subject of this same course (Course #2) must be "Database Theory" as expressed by the relation CourseSubject.

The domain concepts and relations are part of the type system available to the parser, which also contains syntactic types (such as 'verb', 'head', 'interrogative clause'), grammar rules ('head-complement-rule'), lexical types ('active-verb') and lexical entries.

The mechanism that drives the construction of the desired semantic representation relies on a combination of syntactic and semantic constraints. Some of these constraints – both syntactic and semantic ones – are quite general and therefore domain independent. For example, in a Danish yes-no question the finite verb always precede the subject, and an interrogative pronoun constrains the number and type of the instances in focus. Other are more tightly associated with specific words, and are expressed in the system's lexicon. A straightforward example is *universitet* (university), the content of which is a nominal object restricted to being of type 'University', which is a concept in the concept hierarchy. (Syntactic features are largely omitted here).

```
(4)  universitet_1 := lex-phrase &
      [ORTH  "universitet",
      SYNSEM.LOC
      [CAT.HEAD noun, CONT nom-obj &
        [RESTR University] ] ].
```

A more complex example is that of a deverbal noun like *undervisning* (teaching), which contains a mapping between the two syntactic arguments and semantic roles in the appropriate relations (#arg1, #arg2 and #course are variable names):

```
(5)  undervisning_1 := lex-phrase &
      [ORTH  "undervisning" ,
       SYNSEM.LOC
          [ARG-ST <
            [LOC.CONT #arg1],
            [LOC.CONT #arg2] >,
           CONT verb-obj &
            [LOA < CourseOffer &
                   [ TEACHER #arg1,
                     COURSE #course ],
                      CourseSubject &
                   [ WORK #course,
                     SUBJECT #arg2 ]>] ] ].
```

The lexicon entry of the verb *undervise* (teach) contains very similar information to account for different versions of questions revolving around the same content, modelled by the CourseOffer relation:

(6) Hvem har undervisning i filmhistorie?
(Who does teaching in film history?)
Hvem underviser i filmhistorie?
(Who teaches film history?)

Although expressed in a different formalism and directly integrated in the lexicon, the kind of mapping described here is essentially the same done in the Italian system by means of mapping rules.

Treating federated questions

A further research step is the extension of this approach to question analysis in order to manage federated questions. A possible solution would be sending the natural language question to several nodes and let each node interpret it against its own domain knowledge. This is unfeasible in a multilingual environment. The solution we are investigating is based on the notion of ontology mapping.

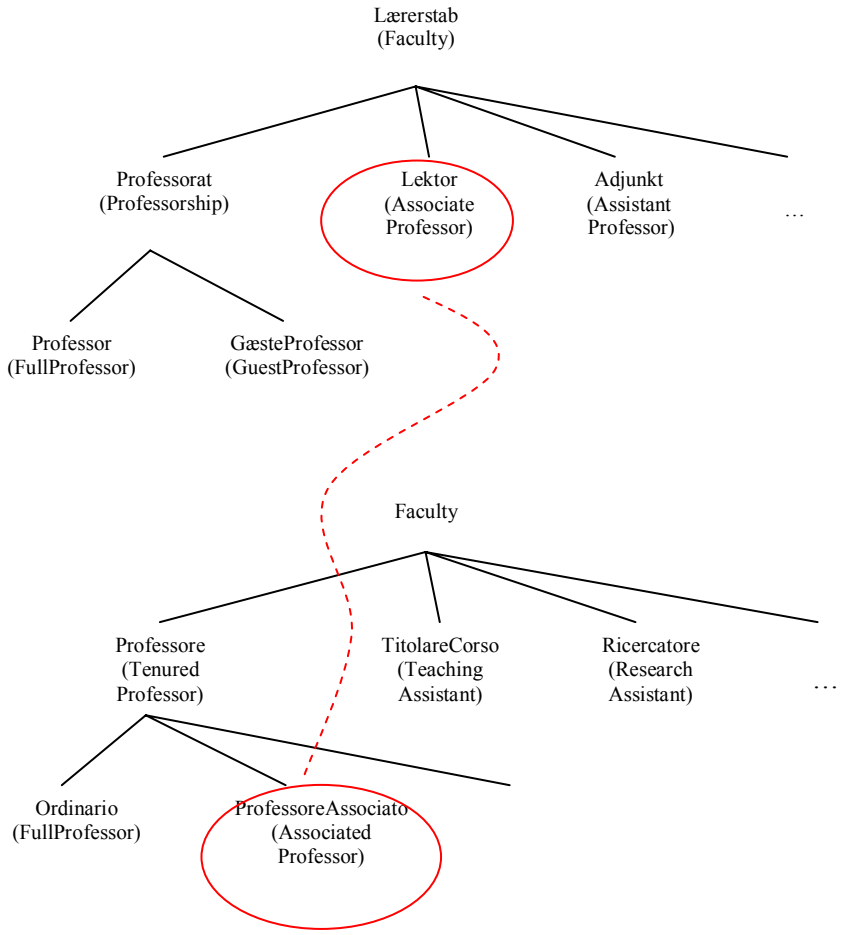


Figure 4: The "Faculty" Danish and Italian sub-ontologies

Let us consider the case of a student questioning not only the Danish but also the Italian site (by selecting specific modalities for entering questions):

(7) Hvem er lektor i fransk?
(Who is associate professor of French?)

As the question is in Danish, it has to be analysed by the Danish analysis component, which will produce a semantic interpretation roughly corresponding to the following term:

(8) all(x) (lektor(x) & CourseOffer(x,y) & Course(y) & Name(y, French))[4]

Since all concepts and relations come from the Danish ontology, it is not a problem to query the Danish knowledge base for all relevant examples. In order to query the Italian knowledge base, however, equivalent concepts and relations must be substituted for those in the "Danish" interpretation. The corresponding Italian representation is:

(9) all(x) (ProfessoreAssociato(x) & TeacherOf(x,y) & Course(y) & Subject(y, French))

The first problem is establishing a correspondence between 'lektor' and 'ProfessoreAssociato', which are not structurally equivalent (Fig. 4). As suggested in (Pazienza&Vindigni 2003, Medche&Staab 2001), equivalence relations must be established by considering *is-a* structures and lexical concept labels together. In the example under discussion, an initial equivalence can be posited between the top nodes of the two ontology fragments, since they both refer explicitly to the original DAML+OIL ontology via a *sameAs* relation. However, none of the concept labels under 'Faculty' in the Italian ontology are acceptable translations of 'Lektor', nor do any of the nodes refer to common nodes in a common reference ontology. Thus, the matching algorithm must search further for en equivalent concept by considering possible translations of concept labels and testing the relations that equivalence candidates participate in. Distance from a common starting node, lexical equivalence and occurrence in similar relations are all constraints to be considered.
The same problem of finding a correct mapping exists for the relations. In this case, we must be able to discover that CourseOffer and TeacherOf represent the same relation.

[4] All concepts and relations will in fact be expressed in Danish. Here, to facilitate non-Danish readers, we are using English equivalents with the exception of the concept 'Lektor' under discussion.

For instance, we can rely on the fact that they have both two roles, and the concepts filling these roles, Faculty and Course (or rather the Danish and Italian equivalent concepts) correspond. Discovering similarities between relations, however, may be a much more complex task than shown in this example. In general, it presupposes the ability to map between concepts.

Conclusion

Our focus in this paper has been, in the context of ontology-based QA, to discuss how to interface between ontology and linguistic resources on the one hand, and ontology and natural language questions on the other while remaining within a unique framework. An interesting issue in a multilingual environment is how to support questions to federation of sites organised around local ontologies. We have begun to address this issue in terms of ontology mapping. Specific algorithms for machine learning and information extraction have also been identified and are under development.

References

Steven Abney (1996) *Part-of-speech tagging and partial parsing*. In G.Bloothooft K.Church, S.Young, editor, Corpus-based methods in language and speech. Kluwer academic publishers, Dordrecht.

Salah Ait-Mokhtar and Jean-Pierre Chanod. (1997) *Incremental Finite-state parsing*. In Proceedings of ANLP97, Washington.

Andreasen, Troels, Per Anker Jensen, Jørgen F. Nilsson, Patrizia Paggio, Bolette Sandford Pedersen and Hanne Erdman Thomsen (2002) Ontological Extraction of Content for Text Querying, in *Natural Language Processing and Information Systems*, Revised Papers of NLDB 2002. Springer-Verlag, pp. 123–136.

Baader, F., D. Calvanese, D. McGuinness, D. Nardi, P.F. Patel-Schneinder, eds. (2003) *The Description Logics Handbook: Theory, Implementation, and Applications*, Cambridge University Press

Basili, Roberto, Michele Vindigni, Fabio Massimo Zanzotto (2003a) *Integrating ontological and linguistic knowledge for Conceptual Information Extraction*, Web Intelligence Conference, Halifax, Canada, September 2003

Basili, Roberto, Maria Teresa Pazienza, and Fabio Massimo Zanzotto (2003b) *Exploiting the feature vector model for learning linguistic representations of relational concepts* Workshop on Adaptive Text Extraction and Mining (ATEM 2003) held in conjuction with Europena Conference on Machine Learning (ECML 2003) Cavtat (Croatia), September 2003

Basili, Roberto and Fabio Massimo Zanzotto (2002) *Parsing Engineering and Empirical Robustness* Journal of Natural Language Engineering 8/2-3 June 2002

Burger, John *et al* (2002) *Issues, tasks and program structures to roadmap research in question & answering* (Q&A). NIST DUC Vision and Roadmap Documents, http://www-nlpir.nist.gov /projects/duc/roadmapping.html.

Callmeier, Ulrich (2000) PET – a platform for experimentation with efficient HPSG processing techniques. In Flickinger, D., Oepen, S., Tsujii, J. and Uszkoreit, H. (eds.) *Natural Language Engineering. Special Issue on Efficient Processing with HPSG*. Vol. 6, Part 1, March 2000, 99–107.

Copestake, Ann (2002) *Implementing Typed Feature Structure Grammars*. CSLI Publications. Stanford University.

Garshol, Lars Marius (2003) Living with Topic Maps and RDF. Technical report. http://www.ontopia.net/topicmaps/materials/tmrdf.html.

Harabagiu, Sanda, Dan Moldovan, Marius Paca, Rada Mihalcea, Mihai Surdeanu, Rhzvan Bunescu, Roxana Girju, Vasile Rus, and Paul Morrescu (2001) *The role of lexico-semantic feedback in open-domain textual question-answering*. In Proceedings of the Association for Computational Linguistics, July 2001.

Hobbs, Jerry R., Douglas E. Appelt, John Bear, David Israel, Megumi Kameyama, Mark Stickel, and Mabry Tyson (1996). *FASTUS: A cascaded finite-state transducer for extracting information from natural-language text*. In Finite State Devices for Natural Language Processing. MIT Press, Cambridge, MA.

Humphreys, K., R. Gaizauskas, S. Azzam, C. Huyck, B. Mitchell, H. Cunningham, and Y. Wilks (1998) *University of sheffield: Description of the LASIE-II system as used for MUC-7*. In Proceedings of the Seventh Message Understanding Conferences (MUC-7). Morgan Kaufman, 1998.

Lauer, Thomas W., Eileen Peacock and Arthur C. Graesser (eds.) (1992) *Questions and Information Systems*. Hillsdale, NJ: Lawrence Erlbaum.

Meadche, Alexander and Steffen Staab (2001) *Comparing Ontologies-Similarity Measures and Comparison Study*, Internal Report No. 408, Institute AIFB, University of Karlsruhe, Germany, 2001

Miller, George A. (1995) WordNet: A lexical database for English. Communications of the ACM, 38(11):39--41, 1995.

Pazienza, Maria Teresa and Michele Vindigni (2003) *Agent-based Ontological Mediation in IE systems* in M.T. Pazienza ed. Information Extraction in the Web Era, LNAI 2700, Sprinter Berlin 2003.

Pollard, Carl and Ivan Sag (1994) *Head-Driven Phrase Structure Grammar* The University of Chicago Press, Chicago.

Rooth, M. (1992) A Theory of Focus Interpretation. In *Natural Language Semantics*, Vol. 1, No. 1, pp. 75-116.

Voorhees, Ellen M. (2001) The TREC question answering track. *Natural Language Engineering* 7(4), pp. 361–378.

Vossen, Piek (1998) *EuroWordNet: A Multilingual Database with Lexical Semantic Networks* Kluwer Academic Publishers, Dordrecht, October 1998

Woods, W., R. Kaplan, and B. Nash-Weber (1972) *The Lunar Sciences Natural Language Information System: Final Report*. Technical Report, Bolt Beranek and Newman, Number 2378, June 1972.

Zajac, Remi (2001) *Towards Ontological Question Answering*, ACL-2001 Workshop on Open-Domain Question Answering, Toulose, France, 2001

A Content Creation Process for the Semantic Web

Eero Hyvönen, Mirva Salminen, Miikka Junnila, Suvi Kettula

Helsinki Institute for Information Technology (HIIT), University of Helsinki
P.O. Box 26, 00014 UNIV. OF HELSINKI, FINLAND
{firstname.lastname}@cs.helsinki.fi
http://www.cs.helsinki.fi/group/seco/

Abstract

This paper discusses the creation of terminologies, ontologies, and annotations when publishing semantic web content. The problem is approached by presenting the content creation processes of the semantic portal MUSEUMFINLAND that is intended for publishing collections of Finnish museums on the web.

1. Introduction

The key idea of the Semantic Web (Berners-Lee et al., 2001) is to annotate web resources with machine interpretable metadata. Based on the metadata, intelligent applications such as semantic portals (Maedche et al., 2001) can be created. Metadata creation includes two major parts. First, the ontologies (Fensel, 2004) and vocabularies used as the basis in metadata descriptions are defined. Second, the web resources are annotated with metadata conforming to the definitions.

A crucial question for the breakthrough of the Semantic Web approach is how easily the needed metadata can be created. Annotating data by hand is laborious and resource-consuming and usually economically infeasible with larger datasets. Automation of the annotation process is therefore needed. This paper addresses the problem of metadata creation for the Semantic Web through a real life case study. We describe the content creation process developed for the MUSEUMFINLAND[1] (Hyvönen et al., 2004a) semantic portal. This application publishes cultural collection data from several heterogeneous distributed museum databases in Finland. We describe what kind of data is needed in bringing the heterogeneous cultural collections into one uniform semantically linked WWW space and focus on how this process can be done with minimal human intervention.

2. Specification for Content Need

MUSEUMFINLAND provides the user with two services: 1) a multi-facet (Pollitt, 1998; Hearst et al., 2002) search engine based on ontologies and 2) a recommendation system for semantic browsing[2].

In order to provide the semantically interlinked and machine understandable inter-museum exhibition and the facets underlying the services, four kinds of content creation processes are needed:

1. **Ontology Creation.** The core of the system is the set of seven domain ontologies listed in table 1.

2. **Terminology Creation.** The museums have heterogeneous contents and use different vocabularies, so a term ontology is needed to define linguistic words and expressions and their relation to ontological concepts. A separate term ontology makes MUSEUMFINLAND flexible with respect to variance in terminologies used at different museums and by different catalogers. The museums can keep their local terminological conventions as long as they tell the meaning of their own terms by a (URI) reference to the ontologies.

3. **Annotation Creation.** During the annotation creation process the data from the museum databases is annotated semantically. The process makes the heterogeneous collection data syntactically and semantically interoperable.

4. **Recommendation Creation.** Rules that define more associative relations between different metadata items need to be created. These rules are based on the domain ontologies, the collection item annotations, and expert knowledge.

Figure 1 depicts the corresponding content creation processes in MUSEUMFINLAND. The final result of the process is the MUSEUMFINLAND RDF(S)[3] Knowledge Base. It consists of the ontologies, the annotated collection data, and an additional Rule Base that is used for enriching the metadata. With the rules new implicit relations are inferred from the explicit metadata.

In the following the sub-processes of figure 1 are explained in more detail.

3. Ontology Creation

In the ontology creation process, three main methods were needed: *manual editing*, *thesaurus transformation*, and *ontology population*. These methods are discussed next.

3.1. Manual Editing

Ontologies are typically created or enhanced by hand using an ontology editor. This is feasible, e.g., with small ontologies, semantically complex ontologies, or if there are no thesauri or other data repositories available for

[1] http://museosuomi.cs.helsinki.fi

[2] The idea of these services is explained in (Hyvönen et al., 2004b).

[3] http://www.w3.org/RDF/, http://www.w3.org/2000/01/rdf-schema

Ontology	Content	Classes	Instances
Artifacts	Classes for tangible collection objects	3227	0
Materials	Substances that the artifacts are made of	364	0
Situations	Situations, events, and processes in the society	992	0
Actors	Persons, companies, organization, and other active agents	26	1715
Locations	Continents, countries, cities, villages, farms etc.	33	864
Times	Eras, centuries, etc. as time intervals	57	0
Collections	Museum collections included in the system	22	24

Table 1: Ontologies in the MUSEUMFINLAND portal.

View category	View	Underlying ontology
Object	Object type	Artifacts
	Material	Materials
Creation	Creator	Actors
	Location of creation	Locations
	Time of creation	Times
Usage	User	Actors
	Location of usage	Locations
	Situation of usage	Situations
Museum	Collection	Collections

Table 2: View facets in the MUSEUMFINLAND portal.

computer-based ontology creation. In our case, the Collections ontology classifying the collections in MUSEUMFINLAND and the Times ontology that represents a taxonomy of different time eras and periods by time intervals were created in this way. All ontologies have been enhanced manually to some extent even if much of the creation work could be automated. In this work the Protégé-2000[4] editor with its RDF plug-in was mostly used.

3.2. Thesaurus Transformation

Controlled vocabularies and thesauri are usually used when indexing collection items in a database. A thesaurus employs a small number of relationships to organize the terms, such as those listed in table 3 (Foskett, 1980). Also references to synonyms, antonyms, and homonyms may be explicitly presented.

In Finland, the most notable and widely used thesaurus for cultural content in Finnish is MASA (Leskinen, 1997) maintained by the National Board of Antiquities[5]. MASA consists of over 6000 terms and employs the relational structure of table 3. This repository was available as a database and its terms could be used as a basis for creating ontologies.

When transforming a thesaurus into an ontology, the NT/BT relations can be used as a first approximation for the subsumption taxonomy. However, lots of manual corrections are needed for several reason. First, the semantics of the NT/BT relation typically includes different forms of both hyponymy and meronymy, which may not be desirable. Second, the relations are often defined locally without considering a larger global context. For example, the entry Make-up mirror can be a narrower term (NT) of Mirror and the entry Mirror can be a narrower term of Furniture. However, one should not infer from this transitively that a make-up mirror is a piece of furniture like one could with a proper subsumption (subClassOf) hierarchy. Third, the NT/BT relations are not systematically developed in thesauri. For example, in the case of MASA it turned out that there were about 2600 roots that had no broader term among the 6000 terms. The thesauri may also contain some errors that have not been detected by the term bank system used for editing the thesaurus. In our case, some missing reciprocal links and even circularity in the NT/BT relation was detected.

MASA thesaurus was transformed into a new taxonomic ontology called MAO in three steps:

1. A meta-level for MAO-ontology was created using Protégé-2000. This meta-level consists of meta-classes that describe the properties of the ontological classes to be created as MAO-classes. The meta-properties fall into two categories: 1) Semantic relations of the thesaurus as they are, such as BT, NT, etc. 2) Metadata documenting the meaning and creation history of the classes, such as creator, date-of-creation, etc.

2. An RDF Schema structure conforming to the RDFS representation conventions of Protégé-2000 was created automatically from the database. This structure represented the entries of the thesaurus as classes organized into an initial subClassOf taxonomy corresponding to the NT/BT relation.

3. A human editor, museum curator, edited the hierarchy further with Protégé-2000 into a proper taxonomy by introducing new concepts and by re-organizing the classes. Some 600 new classes were created during this phase.

[4] http://protege.stanford.edu

[5] http://www.nba.fi

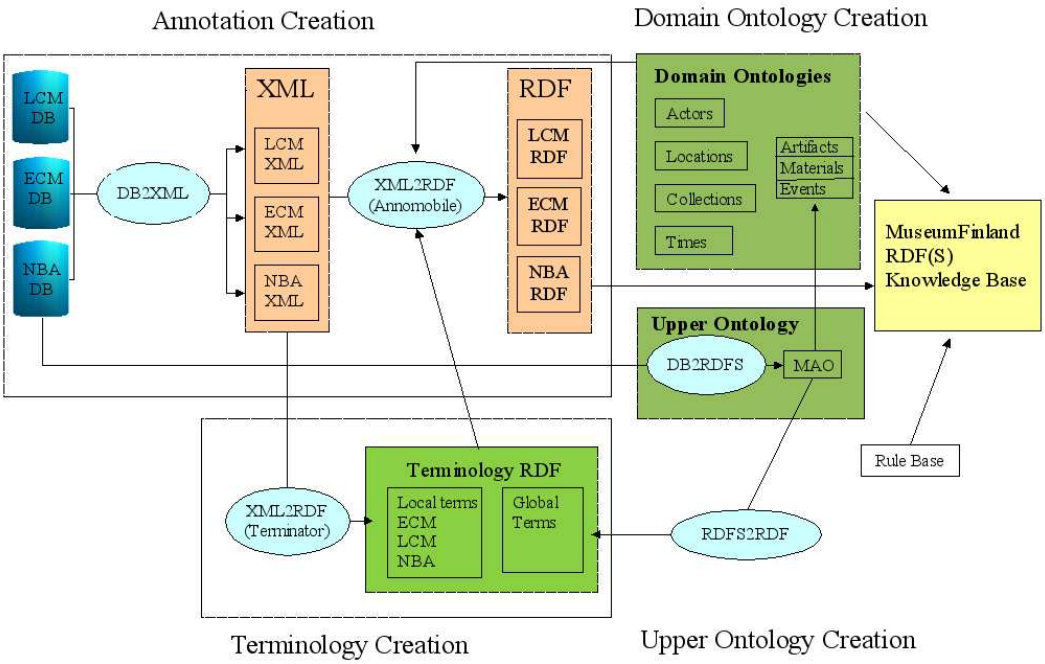


Figure 1: Content creation process in MUSEUMFINLAND.

Symbol	Relationship
USE	Equivalent to ”see” reference
UF	Use for, reciprocal of USE
SN	Scope note
BT	Broader term, in a hierarchical array
NT	Narrower term, in a hierarchical array; the reciprocal of BT
RT	Related term, expressing any useful relation other than BT/NT

Table 3: Typical relationships and their symbols used in thesauri (Foskett, 1980).

The transformation in step (2) can be done easily by an algorithm that created RDF(S) classes for thesaurus entries and an initial subsumption hierarchy. For each entry a *term card* mapping the term to a class URI on the ontology was created. Obsolete terms identified by the USE property were omitted from the taxonomy in order to prevent creation of multiple classes for a single concept. However, term cards were created for these entries since obsolete terms are encountered in databases that have evolved during long time periods, and thus need to be mapped on ontology concepts.

In this way, three domain ontologies, Artifacts, Materials, and Events in table 1 emerged as sub-ontologies of MAO. These ontologies were later on extended based on collection item data from the collections of the National Museum[6], Espoo City Museum[7], and Lahti City Museum[8].

3.3. Ontology Population

By ontology population we refer to a process, where the class structure of the ontology already exists and is extended with instance data (individuals). This can be done either by a computer or by a human editor. In our case, the Actors and Locations ontologies in table 1 were created in this way by a semi-automatic process.

The class structure of the Locations ontology is small and could be created by hand. The main content in the ontology is its individual location instances (e.g., Helsinki or Finland) and their mutual meronymy relations (e.g., Helsinki is a part of Finland). An initial set of individual countries and cities (a couple hundred individuals) was generated automatically from official data sources, such as the list of Finnish cities and counties. However, most of the instance data had to be populated from the collection databases, since the museum databases include specific location information — for example specific estates or historic locations — that were not available in the official data sources. For these locations some meronymy relations could be identified automatically. This is because many collection data entries contain both a general and a more particular location term (e.g., Paris in Texas or Paris in France), from which the meronymy relation could be deducted. For ambiguous location names, the rdf:type and part-of properties had to be edited by a human editor.

As in Locations, the class structure of the Actors ontology is small (Person, Company, etc.) and could be cre-

[6] http://www.nba.fi
[7] http://www.espoo.fi /museo/
[8] http://www.lahti.fi /museot/

ated by hand. Most of the resources in the ontology are instances, such as particular persons. The individuals were populated from the databases. In some cases, the class of the instance could be deduced from the original data. If not, the computer made a guess and let the human editor check the result. For example, it may be known that a certain string, say "John Doe", is a person's name but the sex has not been represented explicitly. The computer can then create an instance of class Person and let the editor change the class to either Woman or Man.

4. Terminology Creation

A thesaurus organizes words. This is in contrast with conceptual ontologies that organize concepts underlying the words. For example, a single conceptual ontology can manifest itself as a set of thesauri in different languages. An ontology is — in principle — language independent in nature, but in practice many concepts are language dependent. The distinction between terms and concepts has many practical consequences also within one language. It is possible to define and use different terminologies as long as a mapping from the terms to concepts is provided. In this way, for example, old collection metadata containing obsolete terms can be used and different terminologies of different museums and of different persons can be made interoperable.

In MUSEUMFINLAND a terminology is represented by a term ontology, where the notion of the term is defined by the class Term. The class Term has the properties of table 4. They are inherited by the term instances, term cards. A term card associates a term as a string with an URI in an ontology represented as the value of the property `concept`. Both `singular` and `plural` forms are stored explicitly for two reasons. First, this eliminates the need for Finnish morphological analysis that is complex even when making the singular/plural distinction. Second, singular and plural forms are used with different meaning in Finnish thesauri. For example, the plural term "operas" would typically refer to different compositions and the singular "opera" to the abstract art form. To make the semantic distinction at the term card level, the former term can be represented by a term card with missing singular form and the latter term with missing plural form. Property `definition` is a string representing the definition of the term. Property `usage` is used to indicate obsolete terms in the same way as the USE attribute is used in thesauri. Finally, the `comment` property can be filled to store any other useful information concerning the term, like context information, or the history of the term card.

A terminology ontology is represented by a Protégé-2000 project that consists of the Term class as an RDF Schema, term instances in RDF, and the referenced ontology represented as an included project. Three different methods were used in terminology creation:

1. **Manual development**

   The terminology ontology can be enhanced and new individual terms created by hand with the ontology editor.

2. **Thesaurus to taxonomy transformation**

   New term instances can be created when transforming a thesaurus into an ontology. Here a term card for each thesaurus entry is created and associated with the ontology class corresponding to the entry. For obsolete terms, the associated ontology resource can be found by the USE attribute value. For entries in singular form (e.g., abstract concepts such as "opera" and materials) the plural form is empty. For those entries in plural form whose singular form represents some other concept, the singular form should be empty. For other entries, both singular and plural forms are created. The morphological tool MachineSyntax[9] was used for creating the missing plural or singular forms for the term cards.

3. **New term generation**

   New term cards are created automatically for unknown terms that are found in artifact record data. The created term cards are automatically filled with contextual information concerning the meaning of the term. This information helps the human editor to fill the `concept` property. For example, assume that one has an ontology M of materials and a related terminology T. To enhance the terminology, the material property values of a collection database can be read. If a material term not present in T is encountered, a term card with the new term but without a reference to an ontological concept can be created. A human editor can then define the meaning by making the reference to the ontology.

Figure 2 depicts the general term extraction process in MUSEUMFINLAND. The process involves a local process at the museum and a global process at MUSEUMFINLAND. There are four different term ontologies: one for terms related to MAO concepts, one for Locations, one for Actors, and one for Collections. For the museum side, we created a tool called Terminator. It extracts individual term candidates from the collection data records. A human editor annotates ambiguous terms or terms not known by the system. The result is a set of new term cards. This set is included in the museum's local terminology and terms of global interest can be included in the global terminology of the whole system for other museums to use.

The global terminology consists of terms that are used in all the museums. It reduces the workload of individual museums, since these terms need not be included in local terminologies. The local term base is important because it makes it possible for individual museums to use and maintain their own terminologies.

The global term base can be extended when needed. For example, when creating new terms, it may occur that there is no appropriate concept in the ontologies that a new term can be associated with. In this case, the term is associated with a more general concept and a suggestion is made to MUSEUMFINLAND for extending the ontology later on with a more accurate concept.

Property	Meaning
`singular`	Singular form of the term as a string
`plural`	Plural form of the term
`concept`	URI of the concept in an ontology
`definition`	Definition of the term or info from a data source
`usage`	Value that tells whether the term is obsolete or in use
`comment`	Any additional information concerning the term

Table 4: Term card properties.

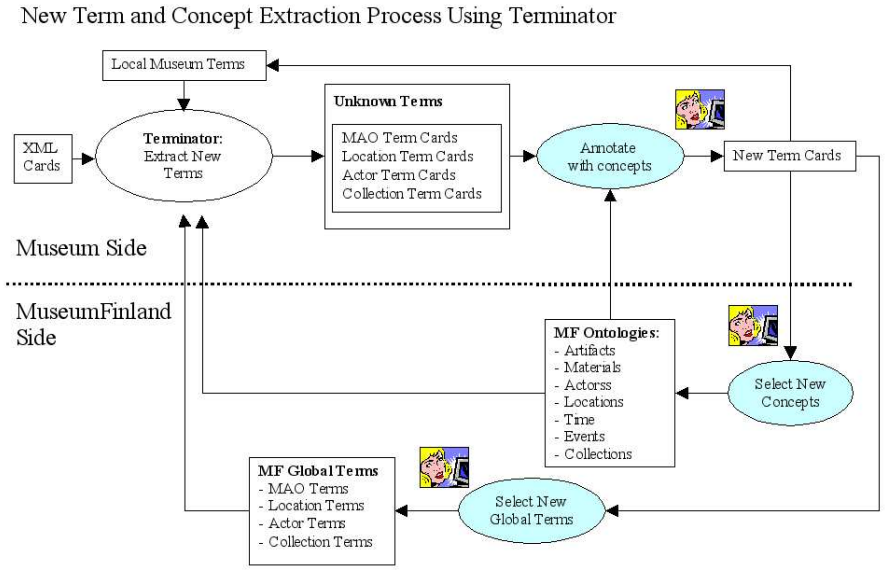


Figure 2: Creating new term cards in MUSEUMFINLAND.

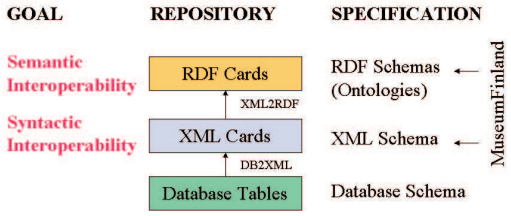


Figure 3: Transforming museum collection data from database into RDF.

5. Annotation Creation

Figure 3 depicts the process of transforming collection data records into RDF format in MUSEUMFINLAND. The first step towards semantic inter-linkage is to attain syntactic interoperability among all the collections. This is done by transforming the collections into XML that is shared by the co-operating museums. As the database schemas of museums are not conforming, the XML card lets every museum to decide which of their database fields to use in filling the XML cards.

Next, the XML is transformed into the final RDF metadata form used by the portal. The RDF conforms to the RDF Schema ontologies of table 1, which guarantees semantic interoperability. The XML to RDF transformation is essentially based on the terms cards by which string values at the XML level, such as “Finland”, are transformed into corresponding concept URIs of the ontologies, such as `http://www.fms.fi/locations#Finland`. A semi-automatic tool called Annomobile has been implemented to perform the transformation. The XML to RDF process is discussed and its algorithm is described in more detail in (Hyvönen et al., 2003).

The XML to RDF transformation cannot be done fully automatically due to unknown and homonymous terms. The problem of unknown terms can, in principle, be solved by generating all needed term cards before running the XML2RDF transformation. The problem of homonymous terms occurs when there are homonyms within the context of a data field (e.g., material, location, etc.) each of which refers to one domain ontology (Material, Location, etc.). Homonymous terms that belong to different domains (e.g., term “Malmi” that refers to both a material and a location concept) can be distinguished without human intervention. Our first experiments indicate that, at least in Finnish, homonymy typically occurs between terms referring to different domain ontologies, and the problem of semantic disambiguation is smaller than initially expected. For example, there are only 29 homonyms in the MAO ontology corresponding to 0,8% of the total number of classes in MAO.

[9] http://www.conexor.fi /m_ syntax.html

6. Discussion

6.1. Contributions

This paper presented an overview of the content creation process for the Semantic Web portal MUSEUMFINLAND. The process was evaluated through a real life case and was found to be useful in many ways:

Terminological interoperability. The terms used in different organizations by different catalogers can be made semantically interoperable mapping the terms onto common shared ontologies.

Terminology sharing. Terms that are commonly used in all the museums can be shared by all the museums, which lowers the number of local terms needed.

Ontology sharing. Ontologies provide means to make exact references to the external world. For example, the Locations ontology and Actors ontology are shared by the museums in order to make correct and interoperable references.

Automatic content enrichment. Artifact descriptions can be annotated semi-automatically based on term ontologies. In addition, ontological class definitions, rules, and consolidated metadata enrich collection data semantically.

As far as we know, MUSEUMFINLAND is the first system to provide semantical interoperability and enrichment across several heterogeneous museum collections.

6.2. Related Work

The idea of annotating cultural contents in terms of multiple ontologies has been explored, e.g. in (Hollink et al., 2003). Other ontology-related approaches used for indexing cultural content include Iconclass[10] (van den Berg, 1995) and Art and Architecture Thesaurus[11] (Peterson, 1994).

Computer-based ontology creation and ontology population can be done using domain texts as discussed, e.g., in (Velardi et al., 2001). Mining of taxonomic relations and instances from text is more error prone but obviously feasible if no other data is available. Our approach of using data-to-be-annotated as the source for ontology population ensures that we create only the instances actually needed. The transformation process of thesauri into semantic web ontologies has been discussed also in (Wielinga et al., 2004).

6.3. Further Research

Practical problems were encountered when transforming the database contents into RDF. For example, the museum collection data used as the input for Annomobile includes not only terms, but also complex phrases and free text, such as the value "*case: case for a prize spoon, competition at Salpausselka, 1924, 10 km skiing*". To handle such descriptions, the free text and complex phrases were tokenized into words and phrases which were then interpreted as keywords. This approach works, when term cards with ontological links are created from these keywords, and the idea was adopted in both Terminator and Annomobile. The drawback here is, that if the vocabulary used in the free texts is large, also the number of new term cards and thus also the manual workload in their annotation will be high. In MUSEUMFINLAND the keyword approach was feasible, since the number of new terms created decreased considerably after the initial term creation phase.

The annotation process cannot be fully automated due to homonymy. This problem is most severe in free text fields, since they are most prone to consist of conceptually general data where disambiguation cannot be based on the facet/ontology to which the text field is related. To solve this problem completely, museum cataloging systems should be enhanced with support for ontology-based indexing.

In the near future we plan to extend the collections of MUSEUMFINLAND with new kinds of ontologies and content, such as paintings and graphics from the Finnish National Gallery. Our goal is to show that RDF can be used as the basis for making very different kind of contents semantically interoperable.

[10] http://www.inconclass.nl

[11] http://www.getty.edu/research/conducting_research/vocabularies/aat/

Acknowledgments

Our work is funded mainly by the National Technology Agency Tekes, Nokia Corp., TietoEnator Corp., the Espoo City Museum, the Foundation of the Helsinki University Museum, the National Board of Antiquities, and the Antikvaria Group consisting of some 20 Finnish museums.

7. References

Berners-Lee, T., J. Hendler, and O. Lassila, 2001. The semantic web. *Scientific American*, 284(5):34–43.

Fensel, D., 2004. *Ontologies: Silver bullet for knowledge management and electronic commerce (2nd Edition)*. Springer-Verlag.

Foskett, D. J., 1980. Thesaurus. In *Encyclopaedia of Library and Information Science, Volume 30*. Marcel Dekker, New York, pages 416–462.

Hearst, M., A. Elliott, J. English, R. Sinha, K. Swearingen, and K.-P. Lee, 2002. Finding the flow in web site search. *CACM*, 45(9):42–49.

Hollink, L., A. Th. Schereiber, J. Wielemaker, and B.J. Wielinga, 2003. Semantic annotations of image collections. In *Proceedings KCAP'03, Florida*.

Hyvönen, E., M. Junnila, S. Kettula, E. Mäkelä, S. Saarela, M. Salminen, A. Syreeni, A. Valo, and K. Viljanen, 2004a. Finnish Museums on the Semantic Web. User's perspective on MuseumFinland. In *Selected Papers from an International Conference Museums and the Web 2004 (MW2004), Arlington, Virginia, USA*. http://www.archimuse.com/mw2004/papers/hyvonen/hyvonen.html.

Hyvönen, E., M. Junnila, S. Kettula, S. Saarela, M. Salminen, A. Syreeni, A. Valo, and K. Viljanen, 2003. Publishing collections in the Finnish Museums on the Semantic Web portal – first results. In *Proceedings of the XML Finland 2003 conference. Kuopio, Finland*. http://www.cs.helsinki.fi/u/eahyvone/publications/xmlfinland2003/FMSOverview.pdf.

Hyvönen, E., S. Saarela, and K. Viljanen, 2004b. Application of ontology based techniques to view-based semantic search and browsing. In *Proceedings of the First European Semantic Web Symposium, May 10-12, 2004, Heraklion, Greece,(forthcoming)*. Springer–Verlag, Berlin.

Leskinen, R. L. (ed.), 1997. *Museoalan asiasanasto*. Museovirasto, Helsinki, Finland.

Maedche, A., S. Staab, N. Stojanovic, R. Struder, and Y. Sure, 2001. Semantic portal — the SEAL approach. Technical report, Institute AIFB, University of Karlsruhe, Germany.

Peterson, T., 1994. Introduction to the Art and Architechure thesaurus. http://shiva.pub.getty.edu.

Pollitt, A. S., 1998. The key role of classification and indexing in view-based searching. Technical report, University of Huddersfield, UK. http://www.ifla.org/IV/ifla63/63polst.pdf.

van den Berg, J., 1995. Subject retrieval in pictorial information systems. In *Proceedings of the 18th international congress of historical sciences, Montreal, Canada*. http://www.iconclass.nl/texts/history05.html.

Velardi, P., P. Fabriani, and M. Missikoff, 2001. Using text processing techniques to automatically enrich a domain ontology. In *Proceedings of the international conference on Formal Ontology in Information Systems, 2001, Ogunquit, Maine, USA*.

Wielinga, B., J. Wielemaker, G. Schreiber, and M.van Assem, 2004. Methods for porting resources to the semantic web. In *Proceedings of the First European Semantic Web Symposium, May 10-12, 2004, Heraklion, Greece,(forthcoming)*.

Proper Nouns Thesaurus for Document Retrieval and Question Answering

Claude de Loupy(1,2), Eric Crestan(1,3), Elise Lemaire(1,2)

(1) Sinequa – SinequaLabs
51, rue Ledru-Rollin
94200 Ivry-sur-Seine – France
http://www.sinequa.com
{crestan, loupy}@sinequa.com

(2) Laboratoire Informatique d'Avignon
B.P. 1228 Agroparc
339 Chemin des Meinajaries
84911 Avignon Cedex 9, France
http://www.lia.univ-avignon.fr

(3) Laboratoire MoDyCo - UMR 7114
Université Paris 10, Bâtiment L
200, avenue de la République
92001 Nanterre Cedex, France
http://infolang.u-paris10.fr/modyco/

Abstract

In this paper, we address the problem of proper noun thesaurus as resource for document retrieval and question-answering tasks. Firstly, as an example of the usefulness of a knowledge on proper nouns, we show that using geographical named entity detection helps to improve document retrieval. Then, we analyze the benefit that could be drawn form using thesaurus as source of answer and as help in order to find answers. Finally, we tackle some problems related to the use of a thesaurus such as *WordNet* and what are the drawbacks when trying to augment it.

1. Introduction

From the beginning of automatic Document Retrieval (DR), researchers have tried to use thesaurus. But results were often disappointing, from Salton (1968) to Voorhees (1994). Many criticisms have been made of WordNet for instance. Even papers reporting better results using thesaurus (Loupy & El-Bèze, 2002) have only slight improvements. These one not really justify the cost of a thesaurus for general language. On the other hand, knowledge on proper nouns can be very effective for document retrieval (Frid *et al.*, 1997). More precisely, geographical information can improve performances. This kind of information brought an improvement of more than 11 points for a TREC query (Loupy & El-Bèze, 2002). We evaluated more precisely the contribution of proper noun knowledge on Amaryllis, a TREC-like evaluation for French (see Section 2).

On an other level, proper name thesaurus can be used for question-answering systems (Mann, 2002). It can contribute at different steps while searching the answers, especially in the case of factual questions. This point of view and the addition of 130.000 proper nouns in the *WordNet* hierarchy are presented in section 3.

2. Thesaurus for document retrieval

Classical TF-IDF term weighting method has been widely used in the last decade for document indexing. The term weight is defined according to its frequency within a document (TF) and according to the number of documents in which it appears (IDF). When querying a search engine, the same term weighting is applied to the query. However, the TF is usually equal to one in queries. Then, the term weighting is only based on IDF. Although two query terms can have the same IDF, they might not necessarily have the same weight in a semantic point of view.

The following query is an example extracted from the Amaryllis corpus, the French evaluation campaign on document retrieval systems.

"Les troubles politiques et civils au Sénégal en 1993"
(Political and civil unrest in Senegal in 1993)

Three different parts can be distinguished in this sentence that can be associated to three "*types of semantic*". The first part is composed from the terms "*troubles politiques et civils"*. It can be classified as "pragmatic semantic" (based only on word meaning) because of the nature of the words. The second part, *Senegal*, has a geographical semantic. The last part, composed by date *1993*, can be classified as temporal semantic. The query can be interpreted as an *event* occurring in a *place* at a given *time*. As for the event, it can have many variations (synonym and structure). For its part, the term *Senegal* has more importance in the query and is more likely to be present under this form in the documents. The date also has a great importance, because it restricts documents on a temporal window. In order to answer the query, the documents must contain the date *1993* or have to be published at this date. The date of publication is usually difficult to catch for Internet based documents, but easily extractable from meta-data when dealing with newspaper

articles (like within the Amaryllis campaign (Coret *et al.*, 1997)).
The tests made with Sinequa's search engine on the Amaryllis corpus (about 11 000 newspaper articles from *Le Monde*) show benefits from distinguishing geographical and temporal semantics in queries. Actually, when we force the presence of place and time entities in the answering documents, an average precision gain of 20% can be observed at 5 documents (precision for the top 5 documents returned). Figure 1 shows precision according to the number of document retrieved when using (or not) thesaurus for geography. At 5 documents, we observe a 20 points gain using geographical thesaurus (*Geography*) over baseline (*Base*). For this test, every time a location occurred in a query, this one, or one of its hyponyms, was required to be present in the returned documents.

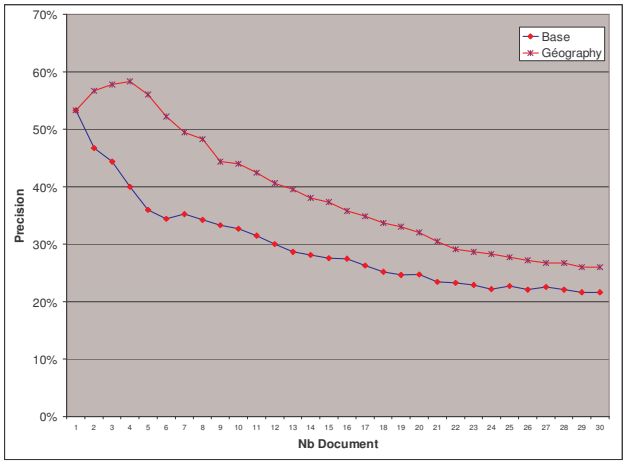


Figure 1: Improvement due to geographical information detection

However, it is sometime necessary to take care of the syntactic structure of query. Although the sequences "*in France*" and "*of France*" contain the same geographical term, they are semantically distinct. The first sequence relates an event occurring in France. But the term can be replaced by one of its hyponyms in documents answering the question (*Paris*, *Marseille*, *Normandy*…). As for the second sequence, the term *France* is more likely to appear in answers because it does not refer to a geographical place, but to a political group. As an example, a query "*population of Germany*" will require the presence of *Germany* or one of its related terms (e.g. *German population*) because we are searching how many people live there. On the other hand, a query "*population in Germany*" will probably refer to the type of population the country is composed of. Then, it is necessary to use hyponyms of the country name *Germany* in order to retrieve documents talking about the population composition in different *länders* (states).
In most case, proper nouns are detected using capital letters (Liddy, 1998). Nevertheless, case is not always respected, especially in queries. Relying on thesaurus can significantly help in detecting proper names and thus, can be used to modified the term weighting.

3. Thesaurus for Question Answering systems

If it is not clear whether a thesaurus can help document retrieval system or not, most of the Question-Answering (Q&A) systems use them for TREC evaluation. Firstly, thesaurus like knowledge is necessary in order to recognize named entities. Secondly, a thesaurus is a specific knowledge base that can be very useful in a Q&A system.

3.1 Named entities recognition

Most of the questions in the past TREC campaigns were factual ones, requiring a named entity (person, country, city, etc.) or a numerical entity (date, measure, etc.) for answer. For this purpose, we developed an entity recognition system in English. It is based on a series of transducer using lexical information as well as semantic clues. In the framework of EQueR[1] (a TREC-like environment for Q&A evaluation), we had to redevelop those transducers for French. This last is more complete than the English one, because it enables the recognition of 83 kinds of entities versus 32 for English language.
A lot of transducers need lists of words with certain properties. For instance, you need a list of first names in order to recognize names of persons. A list of titles and particles can also be used. In the system we developed, these information are placed in a thesaurus-like hierarchy. About 30 lists were added for that purpose.

3.2 Thesaurus as knowledge base

There are two main uses of a thesaurus as a knowledge base. Firstly, a thesaurus can be used in order to verify if a potential answer corresponds to what is needed. For instance, if the question is "what is the fastest animal?", a system could hesitate between "cheetah" and "Ferrari" (an automatic system can be mistaken). If the thesaurus confirms that a "cheetah" is an animal and "Ferrari" a car, the wrong answer can be eliminated.
Secondly, it can be used to answer a question (Harabagiu *et al.*, 2000). For example, a user can ask "what is an armadillo?". One of the answers is in the thesaurus if we use the hyperonymy relation: an "armadillo" is an animal.
In order to determine the importance of proper noun thesaurus, we analyzed a corpus of question. Approximately 11.000 questions in French were collected from different sources, particularly from a list of questions submitted to *Infoclic*[2], a French equivalent of AskJeeves. For each question, we determined if:

- The answer is a proper noun
- The question contains at least one proper noun
- A thesaurus could answer the question
- A thesaurus could help to answer the question

For instance, if the question is "*who was Sherley Temple?*", a thesaurus can answer "*an actress*" if *Sherley Temple* is a hyponym of actress. The example of "cheetah" and "Ferrari" concerns a thesaurus as help.
The study of the corpus of questions shows that 29.16% of them contain a proper noun. The question can be about the proper nouns ("What was the nickname of Elvis

[1] EQueR (Évaluation en Question et Réponse) is a project financed by the French government in the framework of Technolangue (http://www.technolangue.net/)
[2] Infoclic does not exist anymore but the Infoclic team kindly gives us a great amount of the questions submitted to their engine.

Presley?") or the proper noun is used to precise the questions ("What is the biggest town crossed by the Rhône river?"). The answer is a proper name for 13.7% of the questions.
A proper noun thesaurus as a knowledge base is not very advantageous because it gives the answer for only 2.5% of the questions. But this is not insignificant. On the contrary, the thesaurus can help the system to treat 12.3% of the questions (hyperonym verification). So, thesaurus of proper nouns can be a great help for a Q&A system.

3.3 Using *WordNet*

The most widely used otology is *WordNet*. The flaws of *WordNet* have been pointed out in many papers:

- Too fine grained senses (Gonzalo *et al.*, 1998; Palmer, 1998)
- No thematic links (Leacock *et al.*, 1996; Fellbaum et al. 1996). This problem is partially addressed in *WordNet* 2.0
- No inter-POS link (Gonzalo *et al.*, 1998). This problem is addressed in *WordNet* 2.0.
- Some senses have been forgotten (Schütze & Pedersen, 1995) like "derby", which is only a hat and cannot be a horse race. This problem is not *WordNet* specific.

We also note some inconsistency in *WordNet* hierarchy. For instance, the first link between "*king*" and "*queen*" is "*person*" (see Figure 2). Nevertheless, it is interesting to note that in *WordNet* 2.0 many problems have

King : Sense 1
king, male monarch
=> sovereign, crowned head, monarch
=> ruler, swayer
=> person, individual, someone, somebody, mortal, human, soul
…
=> head of state, chief of state
=> representative
=> negotiator, negotiant, treater
=> communicator
=> person, individual, someone, somebody, mortal, human, soul
…

Queen : Sense 2
queen, queen regnant, female monarch
=> female aristocrat
=> aristocrat, blue blood, patrician
=> leader
=> person, individual, someone, somebody, mortal, human, soul
….

Figure 2: Hyperonyms for *King* and *Queen* in the sense on monarchy

disappeared. Moreover, even if there are still some problems, this resource is freely available for research purpose. So we decided to use it for our prototype of Q&A system in English. We enrich the thesaurus with 130.675 proper nouns founded on several knowledge bases and on Internet. These entries were added as hyponyms of existing entries. For instance, 7.089 airports were placed under the synset *02288554* (airport). When there is no direct possible link, we created a new synset, linked to existing entries. The Table 1 gives the different types of entries, their number and where they were placed in the *WordNet* hierarchy.

Entrie	Example	Nb.	Hyperonym
airports	*Shiphol*	7.089	2288554
base-ball players	*Jeff Abbott*	1.845	8520545
base-ball teams	*Boston Red Sox*	30	6562404
capitals	*Amsterdam*	198	6855757
astronomical objects	*Lune*	1	18241
American counties	*Abbeville*	1.688	6877736
car manufacters	*Alfa Romeo*	60	6551769
buildings	*Notre-Dame de Paris*	1.917	3709177
continents	*Africa*	9	7433085
deserts	*Arabian Desert*	34	6846427
baies	*Hudson Bay*	9	7403637
bights	*Great Australian Bight*	1	11461437
canals	*Panama Canal*	2	2508163
capes	*Cape of Good Hope*	1	7394165
channels	*English Channel*	3	7423612
falls	*Iguazu Falls*	2	7604618
gulfs	*Gulf of Aden*	35	7466538
lakes	*Great Lakes*	73	7491139
oceans	*Arctic Ocean*	8	11384719
passages	*Drake Passage*	1	-
rios	*Rio Negro*	5	7555949
rivers	*Amazon River*	141	7555949
seas	*Adriatic Sea*	54	11384719
straits	*Strait of Gibraltar*	14	7582914
enterprises	*3Com*	3.376	6543284
famous persons	*A. Conan Doyle*	4.518	-
inhabitants	*Albanian*	122	5145
human beings	*Alfred Nobel*	2.094	5145
inventors	*Alfred Nobel*	249	8193474
islands	*Aegina*	706	7481328
newspapers	*Arizona Republic*	133	5179803
languages	*Abasakur*	5.987	5191436
locations	*Acatenango*	3.140	18241
measures	*Celsius*	294	11218300
professions	*Academic teacher*	11.073	7737402
monarchs	*Baudouin I*	82	8205809
currencies	*Algerian dinar*	407	11235674
mountains	*Annapurna*	63	7514464
mythological characters	*Abaangui*	2.866	7627143
Nobel winers	*14th Dalai Lama*	722	10969039
organisations	*Bundestag*	240	6535161
political groups	*ANC*	1.174	6682947
countries	*Afghanistan*	511	6876468
planets	*Jupiter*	11	-
politicians	*Aage Brusgaard*	6.652	8374033
ports	*Aden*	1160	6952625
presidents of the USA	*Abraham Lincoln*	42	8386224
areas	*Abaiang*	4.217	6881757
artificial satellites	*Spoutnik*	1.531	3644977
political titles	*1st secretary*	136	7729804
natural locations	*archipel*	796	6976689
touristic locations	*Abu Simbel*	124	6976689
towns	*Abainville*	48.354	-
towns of the USA	*Abercrombie*	15.235	-
volcanos	*Acatenango*	1440	7602352
TOTAL		**130.675**	

Table 1: Proper nouns added into *WordNet*

Some problems appeared when adding a so great amount of entries. It is very useful to have a list of first names (for named entity recognition or in order to find the first name of someone). But many first names are ambiguous with frequent words like "*Who*". So it is important to add new words very carefully in order not to introduce too much ambiguity. It is important to know that *Who* can be a first

name. But this fact is very rare in general corpus. So special analysis must be made for this kind of rare ambiguity.

4. Conclusion

The idea of this paper is that proper noun thesaurus can improve the performances of document retrieval and Q&A systems. The experiment presented in the first part shows the improvements for document retrieval. The second part shows it is possible to build such thesaurus and to plug it in *WordNet*. The analysis of almost 11 000 questions shows that such a thesaurus could be a great help for Q&A systems.

In the near future, we plan to evaluate our proper nouns thesaurus with a real evaluation of Q&A systems. Moreover, a comparison with the contribution of common name thesaurus has also to be done.

References

Coret, A.; Kremer, P.; Landi, B.; Schibler, D.; Schmitt, L. & Viscogliosi, N. (1997). Accès à l'information textuelle en français : le cycle exploratoire Amaryllis, Actes des Premières Journées Scientifiques et Techniques de l'AUPELF-UREF, pp. 5-8, Avignon, France.

Fellbaum, C.; Grabowski, J., Landes, S. & Baumann, A. (1996). Matching words to senses in WordNet: naive vs expert differentiation of senses. WordNet: An electronic lexical database and some of its applications, (editor C. Fellbaum), MIT Press, Cambridge, USA.

Frid, B.; Logounova, L.; Michailov, A., Nusinzon, O. & Zeltser, L. (1997). High precision information retrieval with natural language processing techniques.

Gonzalo, J.; Verdejo, F.; Peters, C. & Calzolari, N. (1998). Applying EuroWordNet to Cross-Language Text Retrieval. Computers and the humanities, Special Issue on EuroWordNet.

Harabagiu, S.; Moldovan, D.; Pasca, M.; Mihalcea, R.; Surdeanu, M.; Bunescu, R.; Girju, R.; Rus, V.; Morarescu, P. (2000). FALCON: Boosting Knowledge for Answer Engines, in Proceedings of the Text Retrieval Conference (TREC-9).

Mann, G. (2002). Building proper noun ontologies for question answering. Proceedings of the Coling 2002 Workshop "SemaNet'02: Building and Using Semantic Networks", Taipei.

Palmer, M. (1998). Are WordNet sense distinctions appropriate for computational lexicons? in Proceedings of SENSEVAL Workshop. Herstmonceux Castle, England.

Leacock, C.; Towell, G. & Voorhees, E.M. (1996) Towards building contextual representations of word senses using statistical models; in B. Boguraev & J. Pustejovsky (Editeurs), Corpus Processing for Lexical Acquisition, MIT Press ; pp. 97-113 ; Cambridge, MA, USA.

Liddy, E. (1998). Enhanced text retrieval using natural language processing. ASIS Bulletin.

Loupy, C de & El-Bèze, M. (2002) "Managing Synonymy and Polysemy in a Document Retrieval System Using WordNet", Actes de l'atelier Creating and Using Semantics in Information Retrieval and Filtering. LREC 2002.

Salton, G. (1968). Automatic information organization and retrieval. McGraw-Hill Book Company.

Schütze, H. & Pedersen, J. (1995). Information retrieval based on word senses. in Proceedings of the 4th Annual Symposium on Document Analysis and Information Retrieval, pp. 161-175.

Voorhees, E.M. (1994). Query expansion using lexical-semantic relations. in Proceedings of the 17th annual international ACM-SIGIR Conference on Research and Development in Information Retrieval (SIGIR'94), pp. 61-69.

A logic-based system for textual document classification

Chiara Cumbo[2], Mario Ettorre[1], Salvatore Iiritano[1,3], Marco Antonio Mastratisi[1], Massimo Ruffolo[1,4], Pasquale Rullo[1,2]

[1]Exeura s.r.l.
{ettorre, iiritano, mastratisi, ruffolo }@exeura.it
[2]Dipartimento di Matematica - Università della Calabria, 87036 Rende (CS), Italy
{cumbo,rullo}@mat.unical.it
[3]DEIS - Università della Calabria, 87036 Rende (CS), Italy
[4]ICAR-CNR - Istituto di CAlcolo e Reti ad alte prestazioni

Abstract

This paper describes a prototypical system supporting the entire classification process: document storage and organization, pre-processing, ontology construction and classification. Document classification relies on two basic ideas: first, using ontologies for the formal representation of the domain knowledge; second, using a logic language (an extension of Datalog by aggregate functions that we call Datalog^f) as the categorization rule language. Classifying a document w.r.t. an ontology means associating it with one or more concepts of the ontology. Using Datalog^f provides the system with a natural and powerful tool for capturing the semantics provided by the ontology and describing complex patterns that are to be satisfied by (pre-processed) documents. The combined use of ontologies and Datalog^f allows us to perform a high-precision document classification.

1. Introduction

Managing the huge amount of textual documents available on the web and the intranets has become an important problem of knowledge management. For this reason, modern Knowledge Management Systems need for effective mechanisms to classify information and knowledge embedded in textual documents (Ciravegna, 2001; Riloff, 2001).
A number of classification approaches have been so far proposed, such as those based on machine learning (Cohen, 1995) and those based on clustering techniques using the vector space model (Díaz, 1998; Hsu, 1999; Brank, 2002).
In this paper we describe a prototypical classification system which relies on two basic ideas: first, using ontologies for the formal representation of the domain knowledge and, second, using a logic language as the categorization rule language.
An ontology is a formal representation of an application domain (Decker, 1999; Fensel, 2001). In the context of a classification process, an ontology is intended to provide the specific knowledge concerning the universe of discourse (categorization based on the domain context). Classifying a document w.r.t. a given ontology means associating it with one or more concepts of the ontology. To this end, each concept is equipped with a set of logic rules that describe features of a document that may relate to the given concept. The logic language we use in our system is an extension of Datalog (Ullman, 1988) with aggregate functions (Dell'Armi, 2003). Throughout this paper we refer to this language as Datalog^f. The advantage of using Datalog^f as the categorization rule language is twofold: first, we can exploit its expressive power to capture the domain semantics provided by the ontology and describe complex patterns that are to be satisfied by documents; second, the encoding of such patterns is very concise, simple, and elegant. We notice that others rule-based techniques have been proposed by several authors, but they are mainly devoted to the resolution of linguistic problems, such as the disambiguation of terms for the reduction of the vector dimensions (Paliouras, 1999), or for the improvement of the results of the classification task (Cohen, 1995).
The execution of Datalog^f programs is carried out by the DLV system (Faber, since 1996), which is part of our categorization engine. DLV is a well-known reasoning system which supports a completely declarative style of programming based on a bottom-up evaluation of the stable model semantics of disjunctive logic programs.

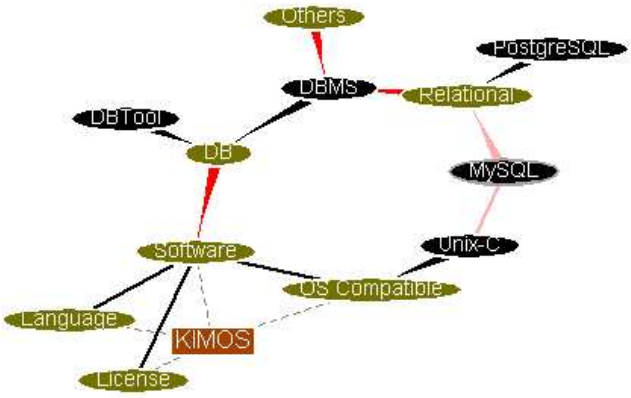
2. A system overview

The prototype is intended as a corporate classification system supporting the entire process life-cycle: document storage and organization, ontology construction, pre-processing and classification. It has been developed as a Web-Application. In the following sections we shall focus our attention on ontology management, pre-processing and classification.

3. Ontology Management

Ontologies in our system provide the knowledge needed for a high-precision classification. The ontology specification language supports the following basic constructs: Concepts, Attributes, Properties (attribute values), Taxonomic (is-a and part-of) and Non-Taxonomic binary associations, Association cardinality constraints, Concept Instances, Links (association instances), Synonyms.

Example 3..1 KIMOS is an ontology developed within Exeura with the purpose of classifying all company's software resources and the respective documentation. A fragment of KIMOS is given in figure 1. Here, the central concept is "Software" which is related to the other concepts by both taxonomic and non-taxonomic relations. For an instance, the edge connecting "Software" with "Language" represents the (many-to-many) relation "developed-in", while the one between "Software" to "OS Compatible"



..

Figure 1: The KIMOS Ontology

represents the relation "runs-on"; the concept "Software" is subdivided into a number of sub-concepts that group the different instances of "Software" into the appropriate categories. In figure we have reported only the concept "DB" that represents the class of softwares for databases. This concept is related to "Software" by an is-a relation and it is classified into "DBMS" and "DB Tool". In turn, "DBMS" is classified as either "Relational DBMS" or "Others" (i.e., DBMS of different types). An instance of "Relational DBMS" is "MySQL" which is related to "Unix-C" (an instance of "OS Compatible") by the link "runs-on". □

Internally, an ontology is stored as a set of facts. As we will see in section 5., these facts represent an input to categorization programs.

Example 3..2 The internal representation of KIMOS consists of facts representing: (1) concepts, e.g., *concept(DB)*; (2) attributes, each identified by an Id, a Data-Type and the Id of the concept which belongs to; for instance, *attribute(size-MB,real,Software)* represents the attribute "size-MB", of type real, of the concept "Software"; (3) Properties (i.e., attribute instances) each characterized by an attribute name and a value; for instance, *property(size-MB,1.35)*; (4) Taxonomic relationships of the form *is-a(DB,Software)*; (5) Non-taxonomic relationships such as *association(runs-on,Software,OS Compatible)* which represents the relation "runs-on" between "Software" and "OS Compatible"; we represent also the inverse *inverse-of(runs-on, supports)*; (6) association cardinality constraints, e.g., *cardinality(runs-on, "⩾ 1")* and *cardinality-inverse(runs-on, "⩾ 1")*; (7) Link associations (i.e., binary association between instances), e.g., *link(runs-on, MySql ,Unix-C)*; (8) Concept Instances such as *instance-of(Relational DBMS,MySql)*; (9) Synonyms such as *synonym(Database,DB)*. □

The creation of an ontology is supported by the Ontology Editor which provides a powerful visual interface based on a graph representation. During the editing of an ontology, the system guides the user to obtain a consistent ontology representation. Once created, the user can navigate the ontology using the Ontology Browser.

4. Pre-processing

The aim of the Pre-Processing is to obtain a machine-readable representation of textual documents (Yang, 1997). This is done by *annotating* documents with meta-textual information obtained by a linguistic and structural analysis.
The Pre-Processor module supports the following two tasks: (*a*) Pre-Analysis based on three main activities: Document Normalization, Structural Analysis and Tokenization, and (*b*) Linguistic Analysis, which in turn consists of a Lexical Analysis (where the PoS Tagger is a variant of the Brill Tagger (Brill, 1995)) and a Quantitative Analysis.
The output of the Pre-Processing phase is a set of facts representing the relevant information about the processed document. As we shall see in section 5., these facts represent an input to our categorization programs.

Example 4..1 Consider, for example, a textual document about databases, with 247 different tokens, and suppose that the third paragraph of this document contains the following fragment of text: "... A **database** is a structured.... ". The representation of this paragraph is like this:

```
...
word(57,'a','a','at').
word(58,'database','databas','nn').
word(59,'is','is','bez').
word(60,'a','a','at').
word(61,'structured','structur','vbn').
bold(58).
par(3,57,148).
tokenFrequency('database',13).
stemFrequency('databas',16).
numberOfTokens(247).
numberOfStems(218).
```

□

5. Document Classification

The basic idea is that of using logic programs to recognize concepts within texts. Logic rules, indeed, provide a natural and powerful way to describe features of document contents that may relate to concepts. To this end, we use the logic language Datalog^f (Dell'Armi, 2003), an extension of Datalog by aggregate functions. The system supporting the efficient bottom-up evaluation of Datalog^f programs is DLV (Faber, since 1996).

5.1. The Datalog^f language

We call Datalog^f the logic language obtained by extending Datalog (Ullman, 1988) by aggregate functions. A *function* has the form $f(Vars : Conj)$ where f is the name (count, sum, min, max, sum) and $Vars$ a set of variables occurring in the conjunction $Conj$. Intuitively the expression $Vars : Conj$ represents the set of values assumed by the variables in $Vars$ making $Conj$ true. An *aggregate atom* is an expression of the type $Lg \leq f(Vars : Conj) \leq Ug$ where Lg and Ug are positive integer constants or variables called *guards*. For instance, $count\{V : a(V)\} < value$ is an aggregate atom whose informal meaning is: the number of ground instances of $a(V)$ must be less than

$value$. A *Datalog^f program* is a logic program in which aggregate literals can occur in the body of rules. Rules with aggregate atoms are required to be $safe$ (Dell'Armi, 2003). It is worth noticing that the result of an aggregate function can be saved by an assignment. For instance in the following rule $h(X) : -X = \#count\{V : a(V)\}$, all the ground instances of $a(V)$ are counted up and the value of count is assigned to X.

5.2. Categorization programs

By combining the expressive power of Datalog with that of aggregate functions, Datalog^f provides a natural and powerful tool for describing categorization rules within our system. A categorization program relies on a number of predefined predicates, that are of two types:

1. *Pre-processing predicates* representing information generated by the pre-processing phase; examples of such predicates are: $word(Id, Token, Stem, PoS)$ and $title(Id, Token, Stem, PoS)$ where *Id* represents the position of *Token* within the text, *Stem* is the stem of the token and *PoS* its Part-of-Speech; $tokenFrequency(Token, Number)$ which represents the number of times Token occurs in the text.

2. *Ontology predicates* representing the domain ontology; examples of this kind of predicates are the following: $instance\_of(I, C)$ (I is instance of the concept C), $synonym(C1, C2)$, $isa(C1, C2)$, $part\_of(C1, C2)$, $association(A, C1, C2)$, etc..

In addition, we use the predicate $relevant(D, C)$ to state that document D is relevant for concept C.
Now, we equip each concept C of a given ontology with a set of Datalog^f rules, the *categorization program* P_C of C, used to recognize C within a given document D. The set of facts of P_C consists of the facts representing the domain ontology (see Section 3.) as well as those representing the pre-processed document (see Section 4.). The rules of P_C represent conditions that are to be satisfied in order D be considered relevant for C.

Example 5..1 We next provide an incremental construction of a categorization program associated with the concept "DB" of the KIMOS ontology (see 3..1).
Rules looking for keyword. We start with the following simple rules looking for the keyword "DB":
r_0: $t0 : -title(\_, "DB", \_, \_)$.
r_1: $t1 : -tokenFrequency("DB", F), F > a$.
In rule r_0 above, the predicate $t0$ is true if "DB" occurs in the title, while t_1 in r_1 is true if the frequency F of the token "DB" is greater than a given constant a.
We can now refine our keyword search by exploiting synonyms; for instance, we can restate r_0 as
r_0: $t0 : -title(\_, X, \_, \_), synonym(X, "DB")$.
and replace r_1 by the following two rules:
r_2: $t2(X, F) : -synonym(X, "DB"), word(\_, X, \_, \_)$, $tokenFrequency(X, F)$.
r_3: $t3$:- $F1 = \#sum\{F, X : t1(F, X)\}, F1 > a$.
Rule r_2 above "evaluates", for the concept "DB" and each of its synonyms, the respective frequency F; rule r_3, in turn, determines the total number $F1$ of times the concept "DB" and each of its synonyms appears in the text (this is performed by the aggregate function sum).
Rules looking for terms. Using the next rules we look for the term "structured data" within the document:
r_4 : $t4(I)$:- $word(I, "structured", \_, \_)$, $word(J, "data", \_, \_), J = I + 1$.
r_5 : $t5(F)$:- $F = \#count\{I : t5(I)\}$.
We may relax the above condition, requiring the words "structured" and "data" to be found, in the specified order, within a distance of at most 5 words inside the same paragraph:
r_6 : $t6(I)$:- $word(I, "structured", \_, \_)$, $word(J, "data", \_, \_), J > I$, $L = J - I, L <= 5, sameParagraph(I, J)$.
r_7: $sameParagraph(I, J)$:- $par(Id, Init, Fin), I >= Init, J <= Fin$.
r_8 : $t8(F)$:- $F = \#count\{I : t7(I)\}$.
Rule r_8 above counts the number of times the searched term occurs in the same paragraph.
Rules matching expressions. Next we write rules to recognize, within a paragraph, an expression of the following type: a verb with stem "store", followed by a name having "tabl" or "relat" as its stem (i.e., we are trying to recognize sentences such as "data are stored within tables...").
r_9 : $t9(I)$:- $word(I, \_, "store", "vb")$, $word(J, \_, "tabl", \_), sameParagraph(I, J)$.
r_{10} : $t10(I)$:- $word(I, \_, "store", "vb")$, $word(J, \_, "relat", \_), sameParagraph(I, J)$.
r_{11} : $t11(F)$:- $F = \#count\{I : t9(I)\}$.
Rules exploiting the ontology knowledge. We can improve precision of the classification process by using the underlying domain ontology. For instance, if a document talks about some specific instances of the concept "db", such as Oracle, Access, etc. (note that an instance of "relational DBMS", which is a sub-concept of "db", is also an instance of "DB"), it is quite obvious considering the document as pertinent to the concept "db". So, we write the following rules:
r_{12} : $t12(I, F)$:- $instance\_of("DB", I)$, $tokenFrequency(I, F)$.
r_{13}: $t13(N)$:- $N = \#count\{I : t11(I, \_)\}$.
r_{14} : $t14(F)$:- $F = \#sum\{F1, I : t11(I, F1)\}$.
r_{15} : $t15(T)$:-
$T = \#count\{I : instance\_of(I, "DB")\}$.
where: r_{12} provides the number of occurrences of each instance of "db" in the document; r_{13} counts the number of distinct instances of "db"; r_{14} provides the total number of instances (duplicated included) of "db" and r_{15} gives the number of instances of "db" in the ontology. Finally, the rule
r_{16} : $t16(K, L)$:- $t13(N), t14(F), t15(T)$, $K = N/T, L = F/N$.
expresses a measure, in terms of K (the fraction of the instances of "db" that are cited within the document) and L (which takes into account the fact that each instance might be cited several times), of the presence into the document of words representing instances of the concept "db". □

As we have mentioned before, we use DLV as the categorization engine in our system. DLV is a very powerful system for the bottom-up evaluation of disjunctive logic programs extended by a number of constructs (Datalog^f is a subset of the DLV language). It is used in many real applications where efficiency is a strong constraint.

The evaluation strategy of categorization programs is based on the following two observations. First, there are documents that are straightforward to classify, i.e., for which simple keyword-based rules (like $r_1 - r_2$ above) are enough; suppose, for instance, that the word "db" is contained in the title or it occurs frequently throughout the text; in such cases we can confidently classify the document at hand as relevant for the given concept only by using few simple rules (like r_1 and r_2) and forgetting of the remaining ones occurring in the rest of categorization program. Second, a deeper semantic analysis is needed only in case of documents that are difficult to classify because concepts do not appear explicitly; to this end, the execution of more complex rules (for instance, rules trying to match complex expressions) is required.

Having this in mind, the implementation of the above evaluation strategy proceeds, roughly speaking, as follows: we structure a categorization program P_C, associated to the concept C, into a number of components, say, $c_1, ..., c_n$. Each component groups rules performing some specific retrieval task, such as word-based search, term matching, etc., of increasing semantic complexity – that is, each component is capable to recognize texts that are possibly inaccessible to the "previous" ones. Given a document D, the evaluation of P_C (w.r.t. D) starts from c_1 (the "lowest" component) and, as soon as a component c_i, $1 \leq i \leq n$, is "satisfied" (by D), the process stops successfully (i.e., D is recognized to be relevant for C); if no such a component is found, the classification task fails.

5.3. Ontology-driven Classification Strategy

Let D be a document that has to be classified w.r.t. an ontology O. As we have seen in the previous subsection, each concept C of O is equipped with a suitable categorization program P_C whose evaluation determines whether D is relevant for C or not. An exhaustive approach would require to "prove" D w.r.t. the categorization program of each concept of O, and this could result in a rather heavy computation. However, we can drastically reduce the "search space" if we adopt an ontology-driven classification technique which exploits the presence of taxonomic hierarchies. This technique is based on the principle that if a document is relevant for a concept then it is so for all of its ancestors within an is-a taxonomy (unless the contrary is explicitly stated). This principle is expressed by the following recursive rule:

$relevant(D, X) : -relevant(D, Y), isa(Y, X)$

As an example, if a document is relevant for the concept "Relational DBMS" of the KIMOS ontology, then it is so for the concepts "DBMS", "DB" and "Software". If we want to exclude the latter, we simply write:

$relevant(D, X) : -relevant(D, Y), isa(Y, X),$
$X <>$ "$Software$".

The above inheritance principle suggests us a classification strategy where concepts within a sub-class hierarchy are processed in a bottom-up fashion. As soon as D is found to be relevant for a concept C in the hierarchy H, it is not any more processed w.r.t. any of the ancestors of C in H. The relevance association of D to the ancestors of C is automatically performed by the above recursive rule.

6. References

Brank, et al., 2002. Feature selection using support vector machines. In *Proc. of the 3rd International Conference on Data Mining Methods and Databases for Engineering, Finance, and Other Fields*.

Brill, 1995. Tranformation-based error-driven learning and natural language processing: A case study in part of speech tagging. In *Computational Linguistics*.

Ciravegna, 2001. (LP)2, an Adaptive Algorithm for Information Extraction from Web-related Texts. In *Proc. IJCAI-2001 Work. on Adaptive Text Extraction and Mining*.

Cohen, 1995. Text categorization and relational learning. In *Proc. of ICML-95, 12th Int. Conference on Machine Learning*.

Decker, et Al., 1999. *Ontobroker: Ontology Based Access to Distributed and Semi-Structured Information*. Proc. of DS-8. Kluwer Academic Publ, pages 351–369.

Dell'Armi, 2003. Aggregate Functions in Disjunctive Logic Programming: Semantics, Complexity, and Implementation in DLV. In *Proc. IJCAI 2003*. Acapulco, Mexico: Morgan Kaufmann Publishers.

Díaz, et Al., 1998. Integrating linguistic resources in an uniform way for text classification tasks. In *Proc. of LREC-98, 1st Int. Conference on Language Resources and Evaluation*.

Faber, since 1996. `DLV` homepage. `http://www.dlvsystem.com/`.

Fensel, 2001. OIL: An ontology infrastructure for the semantic web. *IEIS*, 16(2):38–45.

Hsu, 1999. Classification algorithms for NETNEWS articles. In *Proc. of CIKM-99, 8th ACM Int. Conference on Information and Knowledge Management*.

Paliouras, 1999. Learning rules for large vocabulary word sense disambiguation. In *Proc. of IJCAI-99*.

Riloff, 2001. A Case Study in Using Linguistic Phrases for Text Categorization on the WWW. In *AAAI/ICML Work.Learning for Text Categorization*.

Ullman, 1988. *Principles of Database and Knowledge-Base Systems*. Rockville (Md.).

Yang, 1997. A comparative study on feature selection in text categorization. In *International Conference on Machine Learning*. ACL.

Ontological Types of Associative Relations in Information-Retrieval Thesauri and Automatic Query Expansion

Natalia V. Loukachevitch and Boris V. Dobrov

Research Computing Center of M.V.Lomonosov Moscow State University;
NCO Center for Information Research
339, Research Computing Center of M.V.Lomonosov Moscow State University,
Leninskie Gory, Moscow, 119992, Russia
{louk, dobroff}@mail.cir.ru

Abstract

Traditional information-retrieval thesauri describe a lot of knowledge about various domains. Therefore it is very important to use the thesaurus knowledge in automatic regimes of query expansion. In the paper we will show that analysis of associative relations from the point of view of their ontological properties allows us to understand, what types of such relations can lead to more effective automatic expansion of queries in information-retrieval systems.

1. Introduction

For many years information-retrieval thesauri were used as important tools for document search. Hundreds of thesauri were created in various domains. However traditional information-retrieval thesauri were intended for manual indexing and information search. Contemporary large electronic text collections need automatic means for text processing and information retrieval. Therefore it is very important to know how it is possible to use the knowledge described in traditional information-retrieval thesauri, in automatic regimes.

The use of thesaurus relations for effective automatic query expansion is a serious problem. Most information-retrieval thesauri have two types of relations between concepts (descriptors): Broader-Narrower Terms and Related Terms (associative relations). The present state of affairs here is as follows: various techniques of use of thesaurus relations for automatic query expansion usually give more recall, however precision of the retrieval becomes considerably less, and the overall efficiency of the retrieval decreases (Voorhees, 1999).

Especially a lot of problems are related to associative relations because they comprise a broad range of semantically different relations between concepts. Therefore, it is important to understand, how associative relations or maybe what associative relations can be used in automatic query expansion for recall increasing without decrease of integral characteristics of retrieval.

The problem of the associative relations use is studied usually from two points of view. Tudhope and Taylor (1997), Chen et. al. (1993) consider what weights can be assigned to associative relations in automatic query expansion, in other papers (Tudhope, Alani & Jones, 2001; Rada et. al., 1991) problems of additional semantic classification of associative relations are studied. Also problems of dependence of the weights from semantic types of associative relations are discussed in (Jones, 1993).

In the paper we will show, that analysis of associative relations from the point of view of their ontological properties allows us to understand,

- what relations can be used only in a manual mode of search,
- what relations are in practice non-symmetric and can be effectively used for automatic query expansion only in one direction,
- what associative relations have to be added to descriptor's articles.

In this analysis we will use a notion of "relation of conceptual dependency" from the formal ontology theory.

We will illustrate our analysis on the basis of the information-retrieval multilingual thesaurus of European community EUROVOC (1995), which currently used for manual indexing and retrieval of European documents. In this paper we do not study specific features of EUROVOC in comparison to other similar resources. The thesaurus EUROVOC is a typical example of information-retrieval thesauri (UNBIS, 1976; LIV, 1984), which have similar structures and similar problems. The thesaurus was translated into Russian language therefore we can demonstrate potential capability of associative relations for query expansion through retrieval of documents in University Information System RUSSIA (Russian inter-University Social Science Information and Analytical consortium www.cir.ru/eng), containing more than 800 thousand contemporary Russian documents: official and legislative documents, international treaties, analytical and research papers on social sciences, newspaper articles.

2. Thesaurus Relations and Simple Queries

Queries in an information-retrieval system can consist of different numbers of terms and words. From the thesaurus point of view the simplest query is a query consisting of a single term T of a thesaurus. All other queries, including several terms, words and terms have to be processed as a function from elementary queries.

We hypothesize that potential quality of query expansion based on thesaurus relations can be studied using the simplest queries. If search characteristics of expansion of

elementary queries are low, then processing quality of complicated queries can not be better. If thesaurus relations allow for effective query expansion in simple cases then it is an important step to study techniques for expansion of a complex query. The meaning of such the simplest query is "all about T" and we will denote it as SQ(T).

From this point of view we can study potential search characteristics of every thesaurus relation. Let us see two concepts C1 and C2, between which relation R is established. We consider a simple query consisting of a single term corresponding to concept C1 – SQ(C1), and we would like to know how relation R between C1 and C2 can be used for expansion of this query. In this process documents containing terms of C2 have to be joined to the retrieved set of documents, maybe with certain weights. Hence without any real query expansion we can take documents, containing C2, and try to determine how many of these documents can be relevant to the query SQ(C1).

We will use this method of research for analysis of potential search usefulness of associative relations described in EUROVOC.

3. Associative Relations of EUROVOC and Information Retrieval

As an example we took concept *LAND REGISTER* of the thesaurus EUROVOC. This concept has the following associative relations:

LAND REGISTER
RT *BUILDING PERMIT*
RT *LOCAL TAX*
RT *PROPERTY TAX*
RT *TOWN-PLANNING REGULATIONS*

Associative relations are symmetric therefore we can fulfil a search using query "land register" and evaluate a set of retrieved documents from the point of view of their relevance to every of these four associated concepts. We searched documents using the text collection of UIS RUSSIA. For the retrieval a statistically-based technique similar to (Callan et.al, 1992) was used.

We analysed the contents of 50 first documents in the retrieved set mentioning *land register*. The set included documents by the President of the Russian Federation, the Russian government, daily records of the State Duma of the Federal Assembly of Russia, newspaper articles.

Among 50 documents 41 documents were relevant to the query "land register", other documents discussed specific problems of a state authority (Committee on land register) We received that among these 41 documents:

- 11 documents were also relevant to the simple query "property tax";
- 9 documents were relevant to the simple query "local tax";
- 9 documents were relevant to the simple query "town-planning regulations";
- 3 documents were relevant to the simple query "building permit".

So we can see that query expansion of four different simple queries based on associative relations with concept *LAND REGISTER* will lead to the same result: very low potential precision of retrieval. The reverse relations behave in the same way, so among 50 documents discussing *property tax* only 5 ones could be considered as relevant to the "land register " simple query.

4. Analysis of Problem

The considered relations were not erroneous, they describe important relations, that a land register can be a source of information for building permits, town planning and tax collection, however majority of the texts containing term *land register* were not relevant to the corresponding simple queries. The retrieved texts discussed such topics as creation of a land register, registration of rights to real property, price of land, sale of land and other important problems.

So the considered relations and situations are only several ones among many others where land registers can participate. In fact we can state that

- land register's information can be needed for different goals and sometimes for receiving building permits, sometimes for local tax or property tax calculation etc.,
- the receiving of building permits requires different documents, and one of them is an extract from the land register and so on.

The main problem here is that the contemporary level of automatic text processing does not allow qualitative automatic recognition of such situations in texts of large and heterogeneous text collections. Therefore if we really want to use non-taxonomic relations of concept C1 in automatic query expansion, we have to use (and describe in information-retrieval thesauri) such relations that do not loss their relevance in majority situations of C1.

For concept *LAND REGISTER* such a persistent relation is a relation to concept *LAND*. The land register is a register of land lots, their borders and quality of land. Therefore a lot of situations and actions with *LAND REGISTER* concern lands. If we consider the same 41 texts we can see that 33 (85%) texts were about lands, were relevant to simple query containing concept *LAND* (8 texts were about results of voting for a draft law on land register*)*.

Another important feature of the relation of concept *LAND REGISTER* to concept *LAND* is that concept *LAND REGISTER* can not come to existence without the existence of concept *LAND*. This fact makes it important to consider problems of associative relations from the point of view of the philosophical theory of formal ontology, which studies existence of various entities in the world.

5. Retrieval of Documents and Relations of Ontological Dependence

Basic notions of philosophical formal ontology applied to contemporary conceptual research are philosophical notions of rigidity, identity, unity and dependence (Guarino, 1998). The experimental results very correlate with such a notion of philosophical theory of formal ontology (Smith, 1998) as relation of ontological dependence, which studies the various forms of existential dependence involving special individuals that belong to different classes.

There are three main types of this relation:

- rigid dependence, when the actual existence of an individual necessarily implies the actual existence of another specific individual, so
  - a specific example of *summit* implies existence of specific examples of *heads of states*;
  - a specific example of *forest* is impossible without specific *trees.*
- generic dependence, when the actual existence of an individual necessarily implies the actual existence of some individual belonging to another class, so
  - a specific example of *garage* implies existence of any example from class *cars;*
  - a specific example of *pianist* implies existence of any example from class *pianos;*
  - a specific example of *grocery* implies existence of examples from class *food supplies;*
- historical dependence, when the existence of an entity in moment T presumes the existence of another entity in moment T1 before T, so
  - concept *car* depends from concept *car plant*, because cars are produced in car plants;
  - concept *straw* historically depends from *threshing process*, because it can not appear without threshing, but it can exist for a long time after threshing has finished.

It is easy to see that in case of the rigid dependence the existence of a dependent concept is very tightly connected with the existence of a main concept. It is difficult to imagine a situation (and a text) where a dependent concept participates and this situation has no relation to a main concept.

In case of the generic dependence examples of a dependent concept usually participate in situations related to a main concept, however sometimes situations, not relevant to a main concept, can arise (for example, a crime in a garage can have no relation to automobiles).

At last the historical type of dependence is the weakest type among existential situations. A main concept is necessary for appearance of a dependent concept, but then a dependent concept can exist for a long time and participate in various situations not relevant to the main concept.

Let us study potential retrieval efficiency of simple queries, equal to main concepts M, expanded by text with ontologically dependent concept D. We will analyse 50 best texts from retrieval set for simple query SQ(D). The search was implemented on the full Russian collection of University Information System RUSSIA. Results for several mentioned examples are presented in Table 1.

Dependent concept D	Type of dependence	Main concept M	*nD*	*nM*
FOREST	Rigid	*TREE*	49	12
SUMMIT	Rigid	*HEAD OF STATES*	49	20
PIANIST	Generic	*PIANO*	44	16
GARAGE	Generic	*CAR*	43	1
CAR	Historical	*CAR PLANT*	18	44

Table 1.

In Table 1 *nD* - number of texts containing D, relevant to D and relevant to SQ(M), *nM* - number of texts containing M, relevant to M and relevant to SQ(D).

The table demonstrates the correlation between a type of dependence and search characteristics of simple queries. In case of the rigid dependence for almost all texts if a text is relevant to a dependent concept, it is relevant to a main concept also. In case of the generic dependence the ratio is less but high enough. In case of the historical dependence ratio much decreases. Search characteristics of reverse simple queries are low (that is there are a lot of texts, which are relevant to a main concept and are not relevant to a dependent concept), and this corresponds to absence of dependence. In the fifth pair a lot of texts about car plants are texts about cars at the same time, because concept *car plant* also depends on concept *car*. Car plants can not exist without existence of the class of cars therefore this is the generic type of dependence and again we can see correlation of search characteristics.

6. Ontological Dependence Relations as a Basis for Description of Associative Relations

So in our opinion search characteristics of information-retrieval thesauri created for manual indexing can be improved only if their associative relations are marked from the point of view of rigid and generic ontological dependence:

1) associative relations that are not relations of rigid or generic ontological dependence are marked for use only in manual expansion of a query in a user interface;

2) associative relations describing such relations of ontological dependence become non-symmetric – the direction of query expansion from a main concept to a dependent concept;

3) relations of ontological dependence which are not represented in a thesaurus are added as non-symmetric associations;

4) in several cases associations connect very semantically related concepts. In these cases associative relations preserve their symmetry and can be used for automatic query expansion in both directions.

Analysis of 100 first associative relations (in alphabetic order of concepts) in the Russian version of the thesaurus EUROVOC showed that:

1) 33 associative relations are in fact taxonomic relations, for example, *bus – means of transport.* They are presented as RT relations because EUROVOC does not allow description of more than one Broader Term relations. Therefore these relations are non-symmetric and can be used in query expansion after their marking;

2) 27 associative relations can be used only in manual retrieval, because two concepts have such a relation that appears only in several of all possible situations for every concept, for example,
   - *AIRCRAFT INDUSTRY – NEW TECHNOLOGY,*
   - *AGRICULTURAL EDUCATION – ON-THE-JOB TRAINING,*
   - *AGRARIAN REFORM – REDIRECTION OF PRODUCTION,*
   - *AQUACULTURE – SOLAR ENERGY END-USE APPLICATIONS*;
   - *ALCOHOLISM – ROAD SAFETY;*

3) 41 associations are relations of ontological dependence and can be used in automatic expansion of a query containing a main concept with texts containing a dependent concept for example,
   - *MOTOR CAR – PARKING AREA,*
   - *MOTOR CAR – MOTOR INDUSTRY,*
   - *SHARE – SHAREHOLDER,*
   - *ALCOHOLIC BEVERAGES – ALCOHOLISM;*

4) 3 associations describe symmetric relations between very semantically close concepts, for example, *FARMING SECTOR – AGRICULTURE.*

7. Semantic Names of Ontological Dependence Relations

Let us see what a relationship between the notion of ontological dependence and semantic names of relations. Table2 presents some of associative relations of the thesaurus EUROVOC that are relations of rigid or generic ontological dependence. Every relation is characterized from the semantic point of view – we describe a suppositional semantic name of the relation from the main (ontologically) concept to the dependent concept:

Main concept (M)	Dependent Concept (D)	Semantic relation R: M is R for D
CHILD	*LARGE FAMILY*	*Part*
CHILD	*ADOPTION OF A CHILD*	*Object*
ILLNESS	*PREVENTION OF ILLNESS*	*Counter agent*
PRODUCT QUALITY	*QUALITY LABEL*	*Content*
HEAD OF STATE	*SUMMIT MEETING*	*Agent*
PARLIAMENT	*PARLIAMENTARY SESSION*	*Agent*
MEMBER OF PARLIAMENT	*PARLIAMENTARY IMMUNITY*	*Bearer of property*
TREE	*FOREST*	*Part*

Table 2.

We can see a variety of semantic names of relations describing the ontological dependence. Therefore the ontological dependence represents another point of view to conceptual relations in comparison with the semantic point of view. And from our experiments it is possible to see that the query expansion process (for the existing level of automatic text analysis) is determined by dependence characteristics, not by semantic names of conceptual relations. Therefore any semantic subdivision of associative relations, inclusion additional semantic relations to a set of conceptual relations of information-retrieval thesauri will not give additional retrieval effectiveness.

8. Relations of Ontological Dependence in Thesaurus for Automatic Conceptual Indexing

Nowadays there is a very important question what a structure and a set of relations can be described in domain-specific linguistic resources created specially for automatic text processing of large text collections in information retrieval applications. From our analysis only one conclusion can follow: the contemporary level of automatic processing of large and heterogeneous text collections allows working mainly with a set of relations based on properties of taxonomy and ontological dependence.

We tried to implement idea of descripiton of ontological dependence relations in a linguistic resource specially created as a tool for automatic text processing - Thesaurus on Sociopolitical Life and its applications (Loukachevitch & Dobrov, 2002).

The Thesaurus on Sociopolitical Life (below CIR*Thesaurus – the Thesaurus of Center for Information Research) is a hierarchical net of concepts. It contains a

lot of terms from economical, financial, political, military, social, legislative, cultural and other spheres and is used for automatic text processing the following types of texts: official documents, legislative documents, international treaties, news stories and newspaper articles. Now the CIR*Thesaurus includes more than 70 thousand terms, words and proper names, more than 30 thousand concepts and more than 107 thousand conceptual relations.

Since 1996 the CIR*Thesaurus is used in automatic processing applications such as conceptual indexing, automatic text categorization, automatic text summarization. The CIR*Thesaurus is a searching tool in University Information System RUSSIA

The relations of the CIR*Thesaurus are intensively used in automatic text processing for lexical disambiguation, computation of term weights, query expansion. The system of conceptual relations grew from experiments, was experimentally supported. Now the system of non-taxonomic relations in the CIR*Thesaurus is based on relations of ontological dependence.

Now besides taxonomic relations the thesaurus includes the following types of relations:

- Whole–Part for description of physical parts, relations "situations – their participants", objects – their properties. Generic or rigid ontological dependence of "parts" (participants, properties) from their "wholes" is required.
- Non-symmetric association RT1-RT2 (RT1 = "ontologically dependent of", RT2 = "ontologically main for"), used for other relations of ontological dependence;
- Symmetric association for description of semantically close concepts.

9. Evaluation of Thesaurus in Information Retrieval Applications

To evaluate CIR*Thesaurus-based information retrieval in University information system "Russia" we took 20 topics from list of "Subject Headings for Legislative Acts" adopted as an official system of subject headings in the Russian Federation. The system has 1168 subject headings and 20 main thematic subdivisions. The topics for evaluation were extracted from every subdivision of the system (Loukachevitch & Dobrov 2002).

Topics were usually short and consisted of 1-4 words. Examples of chosen subject headings were as follows: "Water supply", "Use of nuclear energy", "Migration of population". Documents were searched in the subcollection of legislative acts of the Russian Federation (50 thousand documents).

Every search was implemented twice. The first search was implemented using statistical retrieval model similar to (Callan et.al., 1992). In the second search we manually represented a topic as a Boolean expression of concepts from the CIR-Thesaurus and words which are absent in the thesaurus. Such translation was literal without any additions or deletions. For example, subject heading "Use of nuclear energy" was represented as

/Word = '*use*'
AND
/Concept(with Tree) = '*NUCLEAR ENERGY*'

During search every term was automatically expanded using its full thesaurus tree including all described lower concepts and dependent concepts (Parts and RT2 relations). Properties of transitivity of taxonomic relations and dependent parts were used.

Most of the queries resulted in several hundreds documents from the full subcollection. To economize time of evaluation without losses in quality of evaluation we reduced time interval of documents publication to receive 30-40 documents.

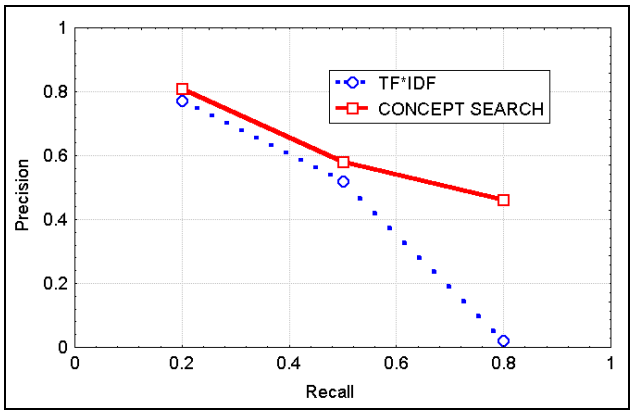


Figure 1. 3-point Recall-Precision Data
for Vector Search (tf*idf) and
Thesaurus-Based Search (Concept Search)

We used "3-point" evaluation to calculate average precision for 0.2, 0.5, 0.8 recall values (Voorhees, 1999), as a measure of retrieval effectiveness. Results of evaluation are presented in Table 3 and on Figure 1.

Type of search	0.2	0.5	0.8	Average
Vector Search	0.77	0.52	0.02	0.44
Thesaurus Search	0.81	0.58	0.46	0.62

Table 3.

Conclusion

Contemporary level of automatic text processing of large heterogeneous text collections does not allow stable identification of semantic relations described in linguistic resources in texts.

Taking into consideration this fact, we proposed to analyse conceptual relations in traditional information-retrieval thesauri for manual indexing from the point of view of the notion of ontological dependence. We showed that if associative relations in such thesauri describe rigid or generic relations of ontological dependence, then these relations can be used for effective automatic query

expansion in the direction from a main concept to an ontologically dependent concept.

We proposed to mark relations of ontological dependence in existing information-retrieval thesauri and this makes it possible more effective usage of knowledge of numerous thesauri developed in a lot of domains.

Also in our opinion if development of a linguistic resource for automatic information retrieval applications in any domain begins nowadays, then description of relations of ontological dependence is an important condition of further effective use of this linguistic resource.

Acknowledgements

Partial support for this work is provided by the Russian Foundation for Basic Research through grant # 03-01-00472.

References

Callan, J.P., Croft, W.B. & Harding, S.M. (1992) *The INQUERY Retrieval System*. In A.M. Tjoa & I. Ramos (eds.), Database and Expert System Applications. Springer Verlag, New York.

Chen, H.; Lynch, K.J., Basu, K. & Ng, T.D. (1993) *Generating, integrating, and activating thesauri for concept—based document retrieval*. IEEE Expert, pp. 25—34.

Guarino, N. (1998) *Some Ontological Principles for Designing Upper Level Lexical Resources*. In "Proceedings of First International Conference on Language Resources and Evaluation".

Jones, S. (1993) *A Thesaurus Data Model for an Intelligent Retrieval System*. Journal of Information Science, 19, pp. 167—178

LIV (1994) *Legislative Indexing Vocabulary*. Congressional Research Service. The Library of Congress. Twenty first Edition.

Loukachevitch, N. & Dobrov, B. (2002) *Evaluation of Thesaurus on Sociopolitical Life as Information Retrieval Tool.* In: "Proceedings of Third International Conference on Language Resources and Evaluation (LREC2002)", M.Gonzalez Rodriguez, C. Paz Suarez Araujo, eds. – Vol.1 – Gran Canaria, Spain, pp.115—121.

Rada, R., Barlow, J., Potharst, J., Zanstra, P. & Bijstra, D. (1991) *Document ranking using an enriched thesaurus*. Journal of Documentation, 47(3), pp. 240—253.

Smith, B. (1998) *Basic tools of formal ontology*. In "Formal Ontology in Information Systems", N. Guarino, ed.

Thesaurus EUROVOC (1995) Vol. 1—3 / European Communities. – Luxembourg:Office for Official Publications of the European Communities, – Ed.3. – English Language.

Tudhope, D., Alani, H. & Jones C. (2001) Augmenting Thesaurus Relationships: Possibilities for Retrieval. – Journal of Digital Libraries. Volume 1, Issue 8.

Tudhope, D. & Taylor, C. (1997) Navigation via Similarity: automatic linking based on semantic closeness. *Information Processing and Management*, 33(2), pp. 233—242.

UNBIS Thesaurus *(*1976) English Edition, Dag Hammarskjold Library of United Nations, New York.

Voorhees, E.M. (1999) Natural Language Processing and Information Retrieval. In M.T. Pazienza (ed.), *Information Extraction: Towards Scalable, Adaptable Systems*. New York: Springer, 32-48.

Building Business Ontologies with Language Technology Techniques

-The VID project

Bolette Sandford Pedersen, Costanza Navaretta, Lina Henriksen

Center for Sprogteknologi, University of Copenhagen
Njalsgade 80, 2300 Copenhagen S - DK
{bolette,costanza,lina}@cst.dk

Abstract

This paper presents work from the Danish VID project and focuses primarily on HLT-assisted ontology construction. Firstly, we investigate how we can combine and refine existing Danish language technology tools and resources for term extraction, and we discuss some of the practical problems encountered when facing this challenge on the basis of company-specific text material originating from a specific business work flow. Secondly, we present the ongoing work of ontology building, where we combine bottom-up and top-down strategies in an attempt to build what we define as *linguistic ontologies* for the domains. We discuss the problematic middle layer of the ontology where term experts and term definitions are heavily drawn upon, and we sketch out how the SIMPLE core ontology is applied as a top layer where classes are characterised primarily by means of linguistic tests.

1. Introduction to VID

Several Danish business companies are beginning to realise the need for knowledge organisation systems that can apply and combine human language technology (HLT) and domain ontologies with emerging technologies in the field of the semantic web in terms of semantically oriented metadata. This paper presents work from the Danish VID project (VIden og Dokumenthåndtering med sprogteknologi – Knowledge and Document Handling with Language Technology) and focuses primarily on HLT-assisted ontology construction, the ontologies being a main component of a knowledge organisation system.

VID comprises the research institution Center for Sprogteknologi (CST), three large Scandinavian companies with high demands for the quality and efficiency regarding document production, as well as two Danish technology companies specialised in search and knowledge organisation participating as technology providers. The aim of the project is to examine, develop and/or refine HLT techniques for acquiring and representing relevant parts of domain knowledge and corporate language in the participating companies.

In this paper we first present the environment of the knowledge base; then we investigate how we can combine and refine existing Danish language technology tools and resources for term extraction, and we discuss some of the practical problems encountered when facing company-specific text material originating from a specific business work flow. After this, we present the ongoing work of ontology building, where we combine bottom-up and top-down strategies in an attempt to build what we define as *linguistic ontologies* for the domains. In this respect, the middle layer constitutes the biggest challenge since it requires a deeper understanding of the domain than what can be found in the text; a knowledge mainly possessed by the term experts or rooted in term dictionaries with term definitions. As top layer, we apply the SIMPLE core ontology where classes are characterised by means of linguistic tests, and not as in traditional formal ontologies on the basis of axiomatic characterisations.

2. Supporting document production with a knowledge organisation system

One of the participating companies is a consultancy company with offices in several of the Nordic countries. The corporate language is English, but the company produces a large quantity of standard documents and correspondences in the Scandinavian languages. Maintaining and updating the standard documents requires a lot of work, together with detailed knowledge about the company working processes, the relevant domain(s), and the legislation in the relevant countries. The company wants to systematise and automate their document production and has therefore acquired a system for semiautomatic saving and production of standard documents. This system is currently being tuned to the needs of the company.

In order to use the system in an optimal way, the company is systematically storing knowledge about the content of their documents. The aim of constructing such a knowledge system is not only to make the document production and maintenance more effective, but also to increase the quality of the documents as well as the knowledge-sharing inside and in-between the different departments of the company. Because the quantity of standard documents is very large, it is important to be able to find relevant documents and/or text chunks in an easy and flexible way, preferably by natural language queries.

All the involved Nordic languages are to be supported by the system, and therefore a system covering all the involved languages as well as English constitutes the final goal of the initiative.

HLT is applied in the project as a facility to semi-automate the *building* of the knowledge organisation system on the basis of existent documentation as well as a facility to be applied in content-based *search* and *document production.*

3. Methodology: linguistic ontology applying HLT techniques

3.1 Danish as a starting point

In this paper we mainly describe methodologies for building the ontologies of the relevant domains, and we primarily focus on the building of the *Danish* sub-ontology. There are several reasons for this: first of all, we have chosen to apply a linguistic, textual basis for ontology building; that is to take the documents of the company as our starting point. In this context, the Danish documents were the ones easiest at hand for the Danish VID project, and furthermore, the Danish branch office are the initiators of the knowledge organisation project and have therefore – as a preparatory action - invested considerable time in structuring and streamlining their standard documents during the last few years. Last but not least, the Danish material gave us a possibility to combine, adjust and validate - on a realistic case - the Danish HLT components that have been developed at CST during the last few years, one of the primary ones being the recently finalised Danish computational lexicon, STO (cf. Braasch & Pedersen 2002 and Braasch & Olsen 2004).

.

3.2 Building term lists

In order to automatically generate term lists as a backbone for the ontology, corpora from two different domains have been constructed in the project, encompassing standard documents and instructions about how to use these standards. The corpora are relatively small; the first corpus containing approx. 87,000 running words, the second approx. 23,000.

For the term list generation, we apply and refine a methodology proposed by Jørgensen et al. (2003). The refined methodology consists in the following steps:

- tokenising,
- POS tagging,
- lemmatising using the morphology encoding in the Danish computational lexicon, STO,
- extracting nominals, verbs and adjectives,
- comparing these with a list of approx. 65,000 general language lemmas in the STO lexicon,
- proposing lemmas that do not occur in STO as term candidates, and finally
- extending the term candidate lists by automatically finding additional term candidates by looking at words in the original list which are encoded as general language words in STO, but which are part of extracted compound terms, such as *gebyr* (fee) which is a component of the term *ekstensionsgebyr* (extension fee).

Examples from the automatically generated term lists are seen in Figure 1 and read as follows: total number of occurrences of the lemma in the corpus, the lemma itself, the POS tag (N for noun, EGEN for proper noun), and the number of occurrences of each inflected form in the corpus.

142 nyhedsundersøgelse N (93 nyhedsundersøgelse/N -, 45 nyhedsundersøgelsen/N -, 4 nyhedsundersøgelsens/N_GEN -)
111 epo EGEN (111 EPO/EGEN -)
101 patenterbarhedsprøvning N (85 patenterbarhedsprøvning/N -, 16 patenterbarhedsprøvningen/N -)
99 prioritetsdato N (43 prioritetsdato/N -, 56 prioritetsdatoen/N -)
88 pct-ansøgning N (72 pct-ansøgning/N -, 14 pct-ansøgningen/N -, 2 pct-ansøgningens/N_GEN -)
84 ansøgningstekst N (41 ansøgningstekst/N -, 43 ansøgningsteksten/N -)
62 præliminær ADJ (54 præliminær/ADJ -, 8 præliminære/ADJ -)
50 nyhedsrapport N (23 nyhedsrapport/N -, 27 nyhedsrapporten/N -)
48 designering N (14 designering/N -, 2 designeringen/N -, 32 designeringer/N -)
47 indleveringsdato N (27 indleveringsdato/N -, 20 indleveringsdatoen/N -)
38 epc EGEN (38 epc/EGEN -)
37 grundansøgning N (33 grundansøgning/N -, 2 grundansøgningen/N -, 2 grundansøgningens/N_GEN -)

Figure 1: Extract of automatically generated term candidate list

We have also extracted multiword term candidates and collocation candidates automatically using pointwise mutual information of the tagged bigrams and trigrams in our corpus (Church and Hanks, 1989). Mutual information was calculated with the CMU-Cambridge Statistical Language Tool (Clarkson and Rosenfeld, 1997). Furthermore we reduced our tag-set to exclusively indicate word-class information. Thus tags such as N_GEN (nominal-genitive) and V_PRES (verb in present form) were changed to N and V respectively. In the resulting analysis we focused particularly on bigrams and trigrams with high mutual information and consisting of subsequent nominals (proper nouns and/or common nouns) and on nominals followed by a preposition and a nominal. Many of the phrases extracted were English company names or organisations, countries, and names of patent-related standards. Examples of these phrases are in figure 2.

saudi EGEN arabien EGEN
eurasian ADJ patent N office N
information N disclosure N document N
det PRON_DEMO ikke-registrerede ADJ design N
(the unregistered design)
ef EGEN design N (EC design)
skånefrist N for PRÆP design N (protective time-limit for design).

Figure 2: Extract of automatically generated multiword terms

When identifying terms as the basis for the ontology, we distinguish between terms and the concepts these terms represent. Terms and term synonyms may change rapidly and though the semantic meaning of a term's

abstract concept is not independent or everlasting as e.g. numbers, concepts of this domain are at least more stable than their linguistic expressions. In the presented approach, the term representations of the concepts are considered lexical entries belonging to a terminology database that is developed to interact with the ontology[1].

One interesting aspect that deserves mentioning, is the fact that most of the texts consist of standard documents with open slots meant for instanciation. This characteristicum has required some tuning of the input texts[2] and of the language tools (particularly tokeniser and tagger) in order to obtain the desired precision in the preprocessing procedures. At the same time, however, it has the clear advantage that instances like proper names (of i.e. appliers and lawyers) and dates are not very frequent in the texts. This means that most of the terms on the automatically generated lists actually refer to ontological concepts relevant for the ontology building. However, some proper names *are* of terminological relevance; and these are – not surprisingly - present in the standard documents. Examples of such are country names related to different IPR legislations as well as relevant institutions like *Patent- og Varemærkestyrelsen*. Such proper names are encompassed by the ontology and referred to as instances (cf. section 3.3).

Evaluation of the automatically extracted wordlists

The lists of term candidates including collocations and multiword terms and the list of relevant general word candidates have subsequently been evaluated by the term experts in the company by means of a two-step procedure; first obvious mistakes have been discarded. For instance, some of the found collocations were not relevant to the specific domain (but were interesting from a linguistic point of view), such as *af hensyn til* (out of consideration for). Secondly, the experts have considered possible extensions of the lists on the basis of additional knowledge. The experts added to the list a number of terms (approx. 16%) which were not contained in the text corpora. These are excluded from the evaluation of the automatically extracted term identification, but they indicate that the corpora we received did not cover all the terminology used in the two departments.
Precision, i.e. the proportion of identified terms that are relevant, is 71,14%. Recall, i.e. the proportion of relevant terms that were identified, is 77,24 %.
The low precision was partly due to the fact that some words had been wrongly marked as content words by the tagger. This was especially the case in sentences containing foreign words. Other errors were due to the fact that some of the term candidates are terms which were used by the company at the time the documents were written, but which have in the meantime been substituted by new synonym terms. Finally a group of words were general language words, but were not encoded in our lexicon. Some terms were not recognised because they were encoded in our lexicon as general words. This was especially the case for juridical words such as *ret* (law) and *domstol* (court of justice).

We are currently investigating, in the line of Jaquemin (2001), whether using the terms in a patent term dictionary, in addition to the STO lexicon, can improve the process of extracting terms from the text corpora.

3.3 OWL in Protégé

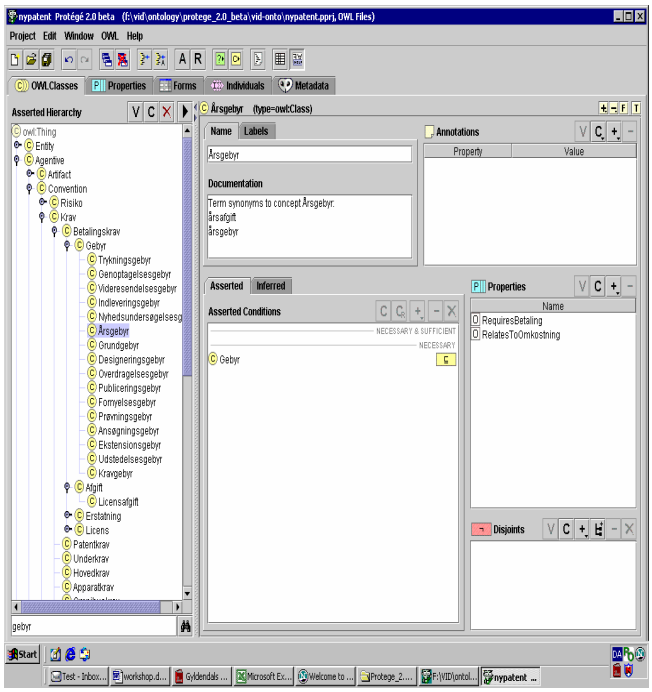
One of the aims of the project is to share research results with the participating companies in the project, and our focus is on the (re)use of data on intranet and internet as well as on the integration of ontological information with other metadata types such as those defined in Dublin Cor). Therefore we decided to use the standard W3C Ontology Web Language (OWL) (http://www.w3.org/TR/owl-ref/) as ontology encoding language and exchange format in the project. As encoding tool we used Protégé-2000 (beta-version) and the corresponding owl plugin, both developed at Stanford University (http://protege.stanford.edu/). Protégé-2000 was chosen because it is freely available, could be run by all the project participants and is already used by some of the partners. For further argumentation, see Pedersen, Navarretta & Haltrup (2003).

3.4 Building ontologies: The bottom layer

We refer to the ontologies that we develop in VID as *linguistic ontologies* for three reasons: (i) they are linguistically 'anchored' being produced primarily on the basis of text corpora, (ii) they are language specific at the lower levels, in this case Danish although mapped into a language-independent upper level-ontology (iii), they address linguistic problems like synonymy, synonymous expressions and polysemy. The ontologies are constructed with a combination of bottom-up and top-down strategies.

The lower nodes are established bottom-up on the basis of the term lists and the generated corpora. In Figure 3 are given extracts from the bottom layer concerning administrative procedures in the patent domain. The class Gebyr (fee) regarding patent applications has 16 different subclasses relating to all the different fees relevant for the domain. Relations are encoded by means of properties and are primarily established between two classes; in this case relations are established between fees and payment as well as between fees and expenses.

[1] Note that for clarification purposes we begin concepts with a capital letter, whereas terms are in italics.
[2] Open slots in the standard documents have been replaced with dummies with appropriate word tags.



Figur 3: The concept Årsgebyr (Annual fee) and its term synonyms *årsgebyr* and *årsafgift* as well as its relations to Betaling (Payment) and Omkostning (Expenses)

As mentioned previously, the ontology also includes instances in terms of proper names of terminological relevance. Figure 4 illustrates the concept Styrelse (Management/Administration) and one of its instances: *Patent- og Varemærkestyrelsen.*

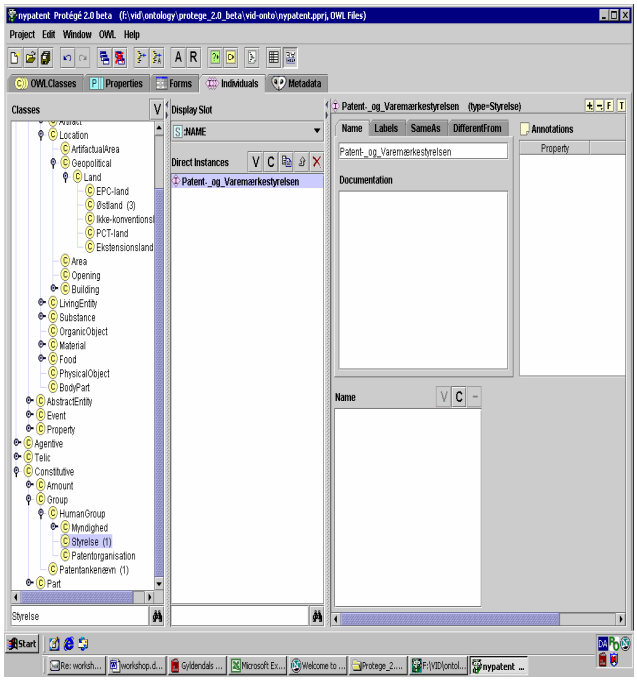


Figure 4: Instance of Styrelse (Government/Administration)

3.5 The middle layer

Building the middle layer has proven to be the most difficult part of the task; and this part of the ontology is still under development. The company is currently in the phase where the actual needs and demands of the ontology are getting more and more specific during the development of the system (see also Navaretta et al. 2004 for considerations on the system architecture).

Apart from the term lists, two sources are heavily drawn upon for the construction of the middle layer. Firstly, the term experts play a central role in this phase since the concepts of the middle layer are partly metaconcepts referred to in the texts (such as Ansøgning, 'application'), but also concepts introduced by the experts themselves as a structuring device (such as Officielt_dokument, 'official document').

A second source is a company specific patent dictionary, which has been scanned in order to construct an electronic version. This source proved to be of considerable relevance and help for the ontology structuring of the middle concepts. Term definitions from the patent dictionary include a genus proximum (closest superconcept) - sometimes already present in the term list; sometimes not - and these function as middle layer concepts providing the basic structuring of the bottom layer. To give an example, *justifikationssag* (justification trial) is defined as a *retssag* (trial) during which it is tested if a prohibity injunction is issued correctly.

3.6 The top layer

The top-down strategy consists of organising the higher levels of the ontology adapting the top categories from the SIMPLE (Semantic Information for Plurilingual, Multifunctional LExica) Core Ontology which contains approx. 135 upper level concept categories (cf. Lenci et al. 2001, Pedersen & Paggio (2004)).

The SIMPLE ontology is multidimensional applying orthogonal inheritance in that it applies the four-dimensional qualia structure proposed by Pustejovsky (1995) and consequently organised according to the four qualia roles Formal, Constitutive, Agentive and Telic. It is originally built as an organisational tool for the encoding of concept lexicons in 12 different European languages and therefore meant to facilitate multilingual mapping. Like in formal ontologies (SUMO, DOLCE, BFO and others[3]), inclusion defines the basic skeleton of the ontology, however, in contrast to these, the characterisation of the categories relies on linguistic tests and not on a formal characterisation based on axioms.

One of the fundamental assumptions behind the SIMPLE ontology is that concepts vary in their internal complexity. Simple types are applied to basic categories

[3]For SUMO (Suggested Upper Merged Ontology), see Sevcenko (2000), DOLCE (A Descriptive Ontology for Linguistic and Cognitive Engineering), see Masolo et al. 2003, and for BFO (Basic Formal Ontology) see http://ontology.buffalo.edu/bfo/BFO.htm.

and to concepts with *rigid properties* (as presented in Guarino 2000:Sec.3.1) such as *himmel* 'sky', *blomst* 'flower', and *bakke* 'hill'. Basic categories are considered to be monodimensional and thus only inherit from the formal role. They can be defined in terms of a monodimensional hierarchy by which we mean that they are organised uniquely by means of hyponymy relations. In contrast, unified types are multidimensional with multiple coordinates although also *grounded* on a simple type.

We have encoded the SIMPLE Ontology in OWL; an extract can be found in figure 5. Examples of multidimensional concepts inheriting from Constitive, Telic and Agentive respectively, are such as BodyPart, Artifact and MovementofThought, whereas examples of simple types are such as OrganicObject and Location.

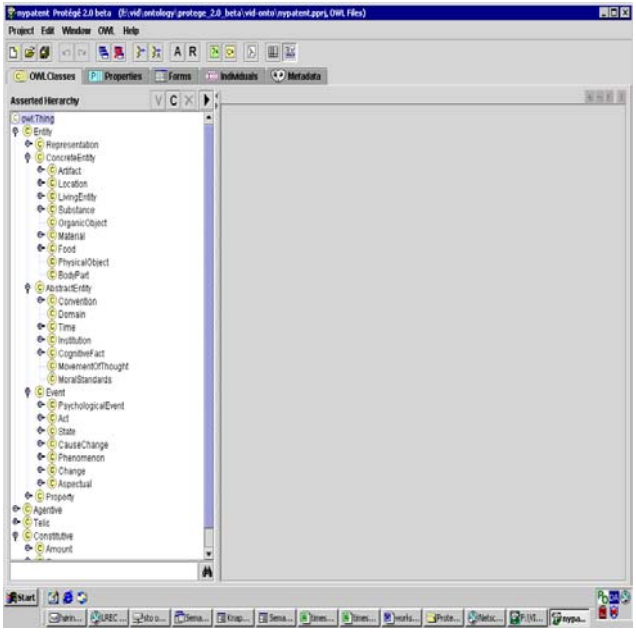


Figure 5: Sample of SIMPLE classes

The lower and middle nodes are linked to the SIMPLE categories manually, applying the linguistic tests provided with the SIMPLE Guidelines (Lenci et al. 2000a). Examples of linguistic tests as a characterisation of a class is given below for the class AgentOfPersistentActivity (Lenci et al. 2000a:207):

1. Do not allow locative modifiers to make reference to the event (*John is a pedestrian in France / *John is a linguist in France*).
2. With the exception of items denoting habits, they do not allow temporal modifiers (*Rush-hour pedestrian / *Rush-hour smokers; frequent customers / * frequent violinist*).
3. With numeral modifiers only individuals can be *counted (five doctors* = 5 different individuals; *five customers* = 5 or less individuals, namely you can also count the events irrespectively of the individuals).
4. If the defining event is negated the result is not a contradiction - with the exception of habits -, unlike with temporary nouns (e.g., *John is a violinist but does not play the violin anymore / *John is a passenger but does not travel anymore*).

4. On-going and future work

4.1 Supporting ontology-building with clustering methods

At present we are investigating to which extent clustering methods can be used to support the process of constructing ontologies and to validate the manual categorisation. We have started collecting a large domain-specific corpus which we will use in clustering. In the first phase of the project we have mainly used POS-tagging and morphologic information in STO in order to lemmatise our corpora. We are now investigating to which extent syntactic and semantic information encoded in the STO lexicon can support the construction of the ontology, especially the encoding of relations between concepts and constraints on the encoded concepts.

4.2 Moving from monolingual to multilingual ontologies

The above ontology is language dependent and reflects aspects of the Danish patent world. However, this is as mentioned a multilingual environment requiring identification of relationships and differences between concepts, also across languages.

The proposed approach involves the creation of an upper-level domain-specific language-independent ontology combined with a number of language dependent ontologies representing the Scandinavian languages. We propose to apply an approach similar to the one applied in Vatant (2003) by linking the language independent ontology to the language dependent ontologies via SAMEAS relations - and similarly to link identical concepts in a subset of the language dependent ontologies with the same relationship. An example demonstrating ontology linking between the Danish ontology and a language-independent ontology is seen in figure 6.

```
<owl:Class rdf:ID="Gebyr">
<owl:sameAs
rdf:resource=http://www.cst.dk/ontology/otm.xml#fee
</owl:sameAs>
 </owl:Class>
```

Figure 6: SameAs relations in OWL

5. Concluding remarks

In this paper we have presented ongoing work in the Danish VID project. The results achieved until now can be summarized as follows: The practical exercise of combining and refining existing Danish HLT tools for

term extraction (in particular the Danish tagger, lemmatiser and STO lexicon) on real business text material has proven feasible and efficient although several practical problems have been encountered. One of these problems relates to the Danish STO lexicon, where we can conclude that part of the so-called 'grey zone' vocabulary in STO overlaps with what the domain experts consider to be relevant terms in the specific domain. This is particularly the case for juridical terms. As indicated above, we are therefore investigating whether using a patent term dictionary in addition to the STO lexicon, can improve the process of extracting terms from the text corpora. The coverage and the quality of the data produced semi-automatically from the corpora are continuously evaluated by the company experts. In spite of the reservations mentioned, the results of these evaluations are promising and indicate that HLT is useful as a substantial support to the construction of knowledge organisation systems.

As regards ontology building, we have again focused on the anchoring in actual text. We claim to build linguistic ontologies, and this is one of the reasons for applying the SIMPLE top ontology which uses linguistic tests as a characterisation of its classes. However, the middle layer requires inclusion of extra-textual knowledge, and here the term experts and term definitions play the central role. An evaluation of the constructed ontologies as backbone of a semi-automatic document production system will be carried out when the ontologies are fully integrated in the document production system.

Acknowledgements

The VID-project is funded by the Danish Research Councils. We would like to thank Dorte Haltrup Hansen, Bart Jongejan and Bente Maegaard (CST) for their useful feedback and the companies participating in the project for a fruitful cooperation.

References

Braasch, A. & B. S. Pedersen. 2002. Recent Work in the Danish Computational Lexicon Project "STO", in EURALEX Proceedings 2002, Center for Sprogteknologi, Copenhagen.

Braasch, A. & S.Olsen. 2004. STO: A Danish Lexicon Resource - Ready for Applications. Proceedings from LREC 2004, Lissabon.

Church, K.W. and P.Hanks. 1989. Word association norms, mutual information and lexicography. In: Proceedings of ACL 27, pp.76-83.

Clarkson, P. and R. Rosenfeld. 1997. Statistical Language Modeling Using the CMU-Cambridge Toolkit. In Proceedings of ESCA Eurospeech 1997.

Guarino, N. & Welty, C. 2000. Ontological Analysis of Taxonomic Relationships, in: A. Laender V. Storey (eds.) Proceedings of ER-2000. The International Conference of Conceptual Modeling. Springer Verlag. (available (http://citeseer.nj.nec.com/correct/309633)

Jacquemin, C. 2001. Spotting and Discovering Terms thorugh National Language Processing. MIT Press. Cambridge, Massachusetss.

Jørgensen, S.W., Hansen, C., Drost, J., Haltrup, D., Braasch, A., Olsen, S.: Domain specific corpus building and lemma selection in a computational lexicon, 2003 In: Corpus Linguistics 2003 Proceedings, Lancaster.

Lenci, A., Bel, N., Busa, F., Calzolari, N., Gola, E., Monachini, M., Ogonowski, A., Peters, I., Peters, W., Ruimy, N., Villages, M. , Zampolli, A. 2000. "SIMPLE – A General Framework for the Development of Multilingual Lexicons", in: T. Fontenelle (ed.) International Journal of Lexicography Vol 13. pp. 249-263. Oxford University Press.

Lenci, A. F. Busa, N. Ruimy, E. Gola, M. Monachini, N. Calzolari, A. Zampolli, J. Pustejovsky, E. Guimier, G. Recourcé, L. Humphreys, U. Von Rekovsky, A. Ogonowski, C. McCauley, W. Peters, I. Peters, R. Gaizauskas, M. Villegas, O. Norling-Christensen. 2000a. SIMPLE Linguistic Specifications, University of Pisa.

Masolo, C. S. Borgo, A. Gangemi, N. Guarino, A. Oltramari, L. Schneider. 2003. WonderWeb Deliverable D17, Preliminary Report, Padova, Italy.

Navaretta, C., B.S. Pedersen, D. H.Hansen. 2004. Human Language Techbology Elements in a Knowledge Organisation system. LREC Proceedings 2004, Lissabon, Portugal.

Pedersen, B., Paggio, P.2004. The Danish SIMPLE Lexicon and its Application in Content-based Querying, Nordic Journal of Linguistics Vol 27:1.

Pedersen, B., C. Navarretta, D. Haltrup. 2003. 'Ontologier og metadata i relation til søgning i tekster', VID Technical Report, Center for Sprogteknologi. Available at http://cst.dk/vid/public/index.html

Pustejovsky, J. 1995. The Generative Lexicon, Cambridge, MA. The MIT Press.

Sevcenko, M. 2003. Online Presentation of an Upper Ontology. In Proceedings of Znalosti 2003, Ostrava, Czech Republic, February 19-21, 2003.

Vatant, B. 2003. Semantic Structure for an ontology-based Knowledge Management System, IST-2001-37244. Unpublished Technical Report.

Natural Language Expression of User Policies in Pervasive Computing Environments

Julie Weeds*, Bill Keller*, David Weir*, Ian Wakeman†, Jon Rimmer† and Tim Owen†

*Natural Language Processing Group
†Networks Laboratory
Department of Informatics, School of Science and Technology
University of Sussex, Brighton, BN1 9QH, UK
{juliewe,billk,davidw,ianw,jonr,timo}@sussex.ac.uk

Abstract

The pervasive computing environments of tomorrow will typically consist of a large heterogeneous collection of networked services. In an ongoing research project, we are exploring ways to enable non-technical users to configure their environment. Our architecture includes an ontology that precisely describes the available services, a formal language for defining user policies and a middleware implementation of formal policies. In this paper, we analyse how existing natural language technologies can be applied to bridge the gap from a natural language description of user policies to a formal representation.

1. Introduction

The pervasive computing environments of tomorrow will typically consist of a large heterogeneous collection of networked services. These could include services associated with physical devices such as printers, mobile phones, motion detectors, and household heating systems, services associated with non-physical resources such as personal calendars or pdf to postscript converters, and services associated with intelligent agents that could, for example, determine your current location, the location of your smart trousers, or identify items that you need to include in your next Internet shop.

We assume a model where service provision is mediated through a **broker** which implements a collection of **policies** that configure the actual services in the environment in a variety of ways. Policies can be used to automate routine tasks and may involve instantiating underspecified descriptions of events and/or combining different services. For example, a user may have a policy to always use the printer nearest to their current location. With knowledge of this policy, we would expect a broker to route an underspecified print request to the most appropriate printer.

Given the current state of the art, there are severe limits on the kinds of policies that it is currently realistic to expect a broker to implement, but there is clearly substantial scope for configuring the basic set of actual services. To some extent, it will be possible to preconfigure the services provided by a broker (for example, by providing defaults that can be used to instantiate underspecified service requests). However, there is clearly a need to allow users to add personalised policies. For example, you might want your phone calls forwarded to your answer machine when your calendar indicates that you are busy, unless the caller is a family member, and you aren't in a meeting with your boss.

Making it possible for non-technical users to configure their environment through a broker represents a major technological challenge. Users cannot be assumed to have significant technological expertise, indeed, they may not even be aware of the existence of some of the more *invisible* services in their environment. Furthermore, there is likely to be a significant gap between the system of rules and constraints that determine the behaviour of the broker and a user's conceptualisation of their environment. This paper arises from an ongoing research project in which we are exploring the role that natural language (NL) descriptions can play in bridging this gap. Our architecture includes an ontology that precisely describes the actual services, a formal language for defining policies, and a middleware implementation of formal policies[1]. In the remainder of the paper we present an analysis of how existing NL technologies can be applied to this problem.

2. An Abstract Policy Representation Language

In order to modularise the problem, we consider an architecture which has an (unambiguous) abstract policy representation language at its centre (see Figure 1). This language is "spoken" (or manipulated) by all of the agents in our system, which we envisage including the policy broker, a natural language user interface (NLI) and a graphical user interface (GUI). Given some unambiguous formal language, we are left with the smaller problems of developing the individual agents that can manipulate it. For example, the NLI, which is the focus of this paper, is required to "translate" from user policies expressed in natural language to user policies expressed in the formal language.

The design of the formal language is the subject of ongoing research within our project. Here, we will just give a flavour of what might be required. We will start by discussing the domain ontology, which is used to establish a shared terminology between the agents. We will then discuss how policies might be expressed with reference to concepts in the ontology.

[1]The latter is also being developed within our project (Owen et al., 2003; Robinson and Wakeman, 2003) but is outside the scope of this paper

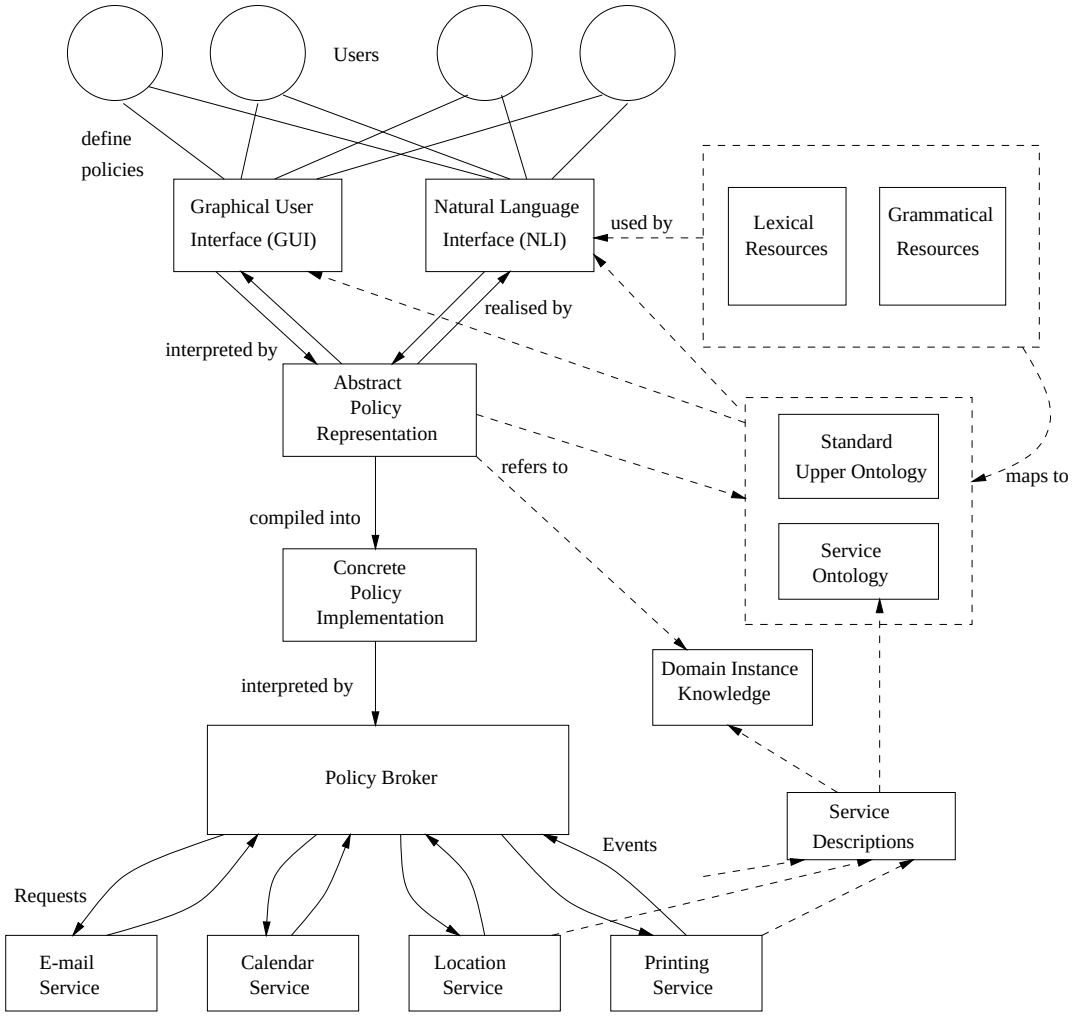


Figure 1: System Architecture

2.1. Domain Ontology

Throughout this paper we will use an office scenario to explore the issues in generating interpretations of NL policy descriptions. In this scenario there are classes of **objects**, corresponding to entities such as printers, scanners, documents and users and classes of **events**, which correspond to different types of services within the environment. For example, events may be user **requests**, e.g., "print document A". Other events may be system **notifications** of state changes, e.g., "user B is in the office".

The ontology is a shared terminology which captures the properties of and relationships between objects and events. For example, printers are devices with attributes including name, location and status; and print events are user requests that involve a user, a file and a printer.

Recent interest in the Semantic Web (Berners-Lee et al., 2001) has led to a proliferation of potential ontology representation formalisms. Baader et al. (2003) note that description logics (DLs) are ideal candidates for ontology languages since they provide a well-defined semantics and powerful reasoning ability. Further, as noted by Stevens et al. (2001), they are the underlying logical formalism of the web ontology languages OIL (Fensel et al., 2001) and DAML+OIL (Horrocks and Patel-Schneider, 2001), whilst being more intuitive to the human user.

Using DL, descriptions of concepts within the domain are built from atomic concepts (unary predicates) and atomic roles (binary predicates). Figure 2 shows a fragment of the concept hierarchy for our office scenario expressed in a typical DL. For example, the concept printer is declared as a subconcept of device and printers have attributes (functional roles) such as *colourness* and *status*. *colourness* is declared as a total attribute of the concept printer since every printer is related to exactly one ColourValue. Following Borgida and Brachman , we use the concept constructor **the** to declare total attributes.

Other roles, however, are relational rather than functional i.e., they do not define a one-to-one mapping. For example, the expression Email $\sqsubseteq \forall recipient.$User $\sqcap \geq 1\ recipient$ states that every recipient of an Email event is a User and that there is at least one recipient.

The relatively intuitive syntax of this DL makes it ideal for generating from natural language. For example, we can represent the natural language expression "an online colour printer" using the description:

$x \in$ Printer $\sqcap \exists colourness.$COLOUR $\sqcap \exists status.$ONLINE

In order to be able to distinguish between individual objects (e.g., a particular printer) and values (e.g. an integer

Object $\sqsubseteq$ **TOP** $\sqcap$ **the** $name$.**String**
PhysicalObject $\sqsubseteq$ Object
$\sqcap$ **the** $location$.PhysicalLocation
Device $\sqsubseteq$ PhysicalObject
Printer $\sqsubseteq$ Device
$\sqcap$ **the** $colourness$.ColourValue
$\sqcap$ **the** $status$.StatusValue
$\sqcap$ **the** $duplicity$.DuplexValue
$\sqcap$ **the** $idletime$.Duration
Scanner $\sqsubseteq$ Device
$\sqcap$ **the** $colourness$.ColourValue
$\sqcap$ **the** $status$.StatusValue
$\sqcap$ ....
User $\sqsubseteq$ PhysicalObject
ElectronicObject $\sqsubseteq$ Object
File $\sqsubseteq$ ElectronicObject
$\sqcap$ **the** $colourness$.ColourValue
$\sqcap$ **the** $security$.SecurityValue
$\sqcap$ **the** $owner$.User
$\sqcap$ **the** $format$.FormatValue
$\sqcap$ **the** $size$.Size
Document $\sqsubseteq$ File
$\sqcap$ **the** $version$.VersionValue
$\sqcap$ **the** $length$.Length
Event $\sqsubseteq$ **TOP** $\sqcap$ **the** id.**String**
Request $\sqsubseteq$ Event
Print $\sqsubseteq$ Request
$\sqcap$ **the** $agent$.User
$\sqcap$ $\forall patient$.File
$\sqcap$ $\geq 1\ patient$
$\sqcap$ $\forall target$.Printer
TurnOn $\sqsubseteq$ Request
$\sqcap$ $\forall patient$.Device
$\sqcap$ $\geq 1\ patient$
Email $\sqsubseteq$ Request
$\sqcap$ **the** $sender$.User
$\sqcap$ $\forall recipient$.User
$\sqcap$ $\geq 1\ recipient$
$\sqcap$ **the** $message$.String
$\sqcap$ $\forall attachment$.File
ClassObject $\sqsubseteq$ **TOP** $\sqcap$ **the** $name$.**String**
Relation $\sqsubseteq$ **TOP**
Nearest $\sqsubseteq$ Relation
$\sqcap$ **the** $objectclass$.ClassObject
$\sqcap$ **the** $location$.PhysicalLocation
$\sqcap$ **the** $individual$.Object

Figure 2: Fragment of the office scenario concept hierarchy

or the value "COLOUR"), we allow attributes to have values from concrete domains (Horrocks and Patel-Schneider, 2001). For example, the *length* of a Document is an **INTEGER** and is indistinct from another **INTEGER** with the same value. Similarly, the concept ColourValue is not defined in terms of other concepts and roles but in terms of the individual values:

$$ColourValue \equiv \{COLOUR, MONO\}$$

Having declared the primitive concepts and roles in our scenario, it is also possible to define new ones. For example, we might want to define a long document as a document with more than 10 pages:

Document $\sqcap$ **the** $long$.**TRUE** $\equiv$ Document $\sqcap \exists length > 10$

We also want to model superlative concepts e.g. "the longest document", "the busiest printer" and "the nearest printer to me". This type of information is most naturally associated with the entire concept rather than with each of its individual instances (Borgida and Brachman, 2003). However, as noted by Borgida and Brachman, DLs do not currently have the ability to be able to treat concepts as objects (as might be possible in some object-oriented systems). Thus it is necessary to create a meta-individual that is related to the concept by a naming convention. For example, we might create the individual PRINTER-CLASS-OBJECT (as an instance of the ClassObject concept) and attach the information regarding *busiest* and *nearest* as roles of this individual. However, certain relationships, such as *nearest*, will have to be reified (i.e. represented as a concept rather than a role) as they involve more than two objects. For example, Nearest is a ternary relation between a concept or class of objects, a physical location and an individual object.

Our discussion so far has focussed on the concept hierarchy. Most DLs also support the notion of a role hierarchy. In our example, there is a similarity between the *patient* role of a Print event (what is being printed) and the *message* role of an e-mail event (what is being e-mailed); and also between the *target* role of a Print event (where it is being printed) and the *recipient* role of an e-mail event (where it is being e-mailed). We will model these similarities by turning each pair of roles into sub-roles of a common super-role. Davis and Barrett (2002) note that very general roles prove useful for stating linguistic regularities in the linking between semantic roles and syntactic arguments, which is something we aim to exploit in our mapping between natural language and logical descriptions.

Finally, we note that our domain ontology is intended to encode only the concepts, objects and services specific to the user's environment. When new services are added, a description must be included by the service provider which will allow it to be incorporated into the ontology. Our ontology will be integrated with existing high-level ontologies such as the Suggested Upper Merged Ontology (SUMO) (Pease et al., 2002) to provide definitions of non-domain specific concepts such as time. Current research (Pease and Fellbaum, 2004) on integrating the machine readable dictionary WordNet (Fellbaum, 1998) with SUMO increases

User	Policy Set
1	*Normally, I use lja.* *If lja is off-line, I use ljx.* *If the document consists of more than 15 or 20 pages then I use ljb.* *If the document is in colour then I might use cclj.*
2	*I primarily use printer cork, because its near my desk.* *If the document is long and only a draft then I will use ljb.*
3	*If I am printing the final copy of a document which is in colour, then I will print it on cclj.* *Otherwise, if the document is long I will print it on ljb.* *Otherwise, I use the closest printer.*

Figure 3: Excerpts from user statements describing how they select what printer to use

its appeal for use with natural language. In order to integrate our ontology with SUMO, it will be necessary to provide a mapping between the DL representation of our ontology and a KIF or DAML representation.

2.2. User Policies

There is a range of different types of policies that a user might want to express:

default rules: filling in the gaps in an underspecified user request, possibly based on the characteristics of other objects involved in the event.

ontological definitions: defining a new concept or role in terms of other concepts and roles.

rewrite rules: changing a user request under some specified condition.

blocking rules: blocking a user request under some specified condition.

event generation: generating a new event on the basis of a trigger event.

In this paper, we focus on default rule policies, which have a first order logic (FOL) interpretation. Ontological definitions can also be expressed in FOL since they are assertions in English. We expect to be able to apply much of our work on default rule policies to ontological definitions. We also note that there is previous work (Pease and Murray, 2003) on the translation of ontological definitions from controlled English to logic. Rewrite and blocking rules present problems for any monotonic logic formalism, since the consequents of these rules may contradict other facts in the database (including the conditions of the rule). Event generation policies are similar in form to the complex sentences used in task-oriented dialogues (Balkanski, 1992) and their logical expression is the subject of ongoing research.

With a view to collecting real-life policies, we performed an initial study of how users manage a collection of available printers, and how they might express their policies using natural language. In the study, twenty-four users within a university department were asked to write down how they decide what printer to use. Figure 3 shows some typical examples extracted from their statements. In many cases, a default rule or policy could be applied which would, say, route a colour document to a colour printer if the target of the print event is unspecified. In principle, this rule could also be interpreted as a rewrite rule and apply when the broker receives a print request in which a specific printer is specified, but whether or not this is desirable, and how problems relating to this can be resolved is a complex issue that is outside the scope of this paper.

Another issue that is clear from our study is that policies cannot be interpreted in isolation. In addition to a policy about colour documents, a user may also have a policy to print long documents on a double-sided printer. One complication here is that *long* does not have a precise definition and thus a policy is required stating how it should be interpreted in this context. A second complication is the possible interaction or conflict between these two policies. What happens if the user attempts to print a document which is colour and in its final version and long? In this case, it may be possible to print the document to a duplex, colour printer. However, this may not be what the user wants and even if it is, it may not be possible given the actual services available i.e., there may be no double-sided, colour printer in the user's domain.

One possible solution is to interpret the semantics of user policies as preferences or soft constraints (e.g., Bartak (2002)) on the broker's reasoning. In the above example, a preference for a colour printer and a preference for a double-sided printer will be generated. We can also model a preference for a certain default printer, or an on-line printer, or the nearest printer to my current location in terms of constraints. The problem of finding an appropriate printer then becomes one of (possibly weighted) constraint satisfaction. Figure 4 shows constraints that we might wish to generate for a set of policies.

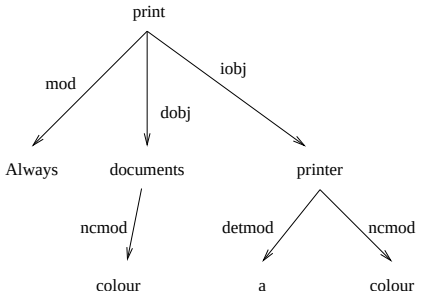
Another approach, which could be used alongside or instead of constraints, is to provide a multimodal interface which will allow the user to simulate and debug their own policies. For example, using the interface, a user could select which policy has highest priority. Using this type of model, the system can be thought of a tool for aiding the translation between different policy representations: NL, logical, graphical and software implementation.

3. From NL Descriptions to Constraints

Our general approach to obtaining constraints from NL policy statements is fairly standard (Allen, 1984) and involves mapping syntactic structure in the natural language to the semantic representation provided by the ontology. This section describes some of the more interesting details. In particular, we will discuss the recovery of syntactic dependencies using a shallow parser, extension of the lexicon using pre-existing lexical resources and distributional similarity methods, word sense disambiguation using knowledge of the semantic argument types from the ontology and recovery of implicit event participants using the ontology.

No.	NL Description	Policy Constraint	Strength
1	*Always print colour documents on a colour printer.*	$x \in$ Print $\sqcap\ patient.$(Document $\sqcap\ colourness.$COLOUR) $\rightarrow$ $x \in target.$(Printer $\sqcap\ colourness.$COLOUR)	Strong
2	*I usually print draft copies double-sided.*	$x \in$ Print $\sqcap\ agent.name.$\$Username $\sqcap$ $patient.$(Document $\sqcap\ version.$DRAFT) $\rightarrow$ $x \in target.$(Printer $\sqcap\ duplicity.$DRAFT)	Weak
3	*I never print confidential documents on lja.*	$x \in$ Print $\sqcap\ agent.name.$\$Username $\sqcap\ patient.$(Document $\sqcap\ security.$CONFIDENTIAL) $\rightarrow$ $x \notin\ target.$(Printer $\sqcap\ name.'lja'$)	Strong
4	*Never send documents to an off-line printer.*	$x \in$ Print $\sqcap\ patient.$Document $\rightarrow$ $x \notin\ target.$(Printer $\sqcap\ status.$OFFLINE)	Strong
5	*By default, I print documents on the closest printer.*	$(x \in$ Print $\sqcap\ patient.$Document $\sqcap\ agent.(name.$\$Username $\sqcap\ location.y))$ $\wedge\ w \in$ (Nearest $\sqcap\ objectclass.$PRINTER-CLASS-OBJECT $\sqcap\ location.y \sqcap\ individual.z) \rightarrow\ x \in target.z$	Weak

Figure 4: Examples of NL policies and corresponding constraints




```
(dobj, print:2_VV0, document+s:4_NN2)
(iobj, on:5_II, print:2_VV0, printer:8_NN1)
(ncmod, document+s:4_NN2, colour:3_NN1)
(ncmod, printer:8_NN1, colour:7_NN1)
(detmod, printer:8_NN1, a:6_AT1)
(mod, print:2_VV0, Always:1_RR)
```

Figure 5: Dependency analysis for the user policy "Always print colour documents on a colour printer"

3.1. Shallow Parsing

Shallow, dependency-based parsing can be used to determine the local, grammatical relations between the words in a sentence. These grammatical dependencies are closely related to the logical dependencies that hold between objects and events in our ontology. A key advantage of shallow parsing over deep syntactic analysis is its robustness. We aim to show that combining dependency-based parsing with the domain ontology will provide a robust and accurate approach to the interpretation of NL policy statements.

Our approach makes use of the RASP toolkit (Briscoe and Carroll, 2002), a pipelined, modular parsing system comprising separate processing stages for: tokenisation, part-of-speech and punctuation tagging, lemmatisation and shallow parsing. The output of the RASP parser is a dependency analysis of the input sentence, represented as a set of grammatical relations between lexical heads. An example of a dependency parse for the user policy *"Always print colour documents on a colour printer"* is shown in Figure 5, together with the corresponding set of grammatical relations that is output by the parser. The dependency structure shows that *"print"* is the lexical head of the whole sentence, and that it has a direct object (dobj) *"documents"*, an indirect object (iobj) *"printer"* and a modifier (mod) *"always"*. Further, the direct object *"documents"* is the lexical head of the sub-phrase *"colour documents"*, where *"colour"* is a (non-clausal) modifier (ncmod) of the head; and similarly for the indirect object *"printer"*.

3.2. The Lexicon

The words used by the user need to be mapped onto concepts in the ontology. We assume the existence of a **core lexicon** that associates a small number of words or phrases with each concept or class in the ontology. For example, the core lexicon might well assign the word *printer* to the concept or class `Printer` in the ontology.

There are, of course, many alternative ways that the same concept can be expressed, for example, by the use of synonyms or hypernyms/hyponyms. Rather than trying to include all of these directly to our lexicon, we are investigating how machine readable dictionaries (such as WordNet) and distributional similarity techniques can be used to overcome sparseness of the core lexicon.

This approach is bound to introduce a certain amount of noise, but the domain ontology can be used to resolve some of the potential uncertainty in how to map from lexical items into the ontology. For each word w in the dependency tree, a set of possible concepts in the ontology will be generated. Each possible concept c will have a **similarity score** c_w associated with it which indicates the similarity between the lexical item w and the core lexicon's entry for the concept c. For example, the lexical items associated with the `ColourValue` concept in the ontology are *colour* and *monochrome*. However, if a user uses the term *black-and-white*, we identify that the user means *monochrome* without explicitly stating this in our domain ontology since *black-and-white* is synonymous with *monochrome* in WordNet and can therefore be given a high similarity score. Similarly, we can determine that a user who refers to a *copy*, as in policy example 2, might be referring to a *document* since these are related in WordNet via their common hypernym *Writing, written material*. As will be discussed in Section 3.3., where the user refers to *draft copies*, there is further evidence for the document interpretation, since *draft* is a

possible value of an attribute of a document.

There are at least two problems with using WordNet to augment our domain ontology. First, a word may not exist in WordNet. For example, there is no entry for the word *double-sided* (in WordNet 1.6). Second, words tend to have multiple senses in WordNet, many of which are unlikely interpretations given the domain. To some extent, the domain ontology can be used directly for disambiguation. For example, we know that the most likely sense of *printer* is the *device* sense, rather than the *person* sense, since the entry for *printer* in the core lexicon maps it to the `printer` concept in the ontology, which is a subconcept of `device`. This idea can also be extended to words outside of the core lexicon. For example, there are four senses of the word *copy* in WordNet 1.6. In our office scenario, the highest similarity between concepts in WordNet occurs between the written material sense of *copy* and document, and so disambiguation is comparatively straightforward. However, as the scenario and the ontology are scaled-up, the process becomes more problematic.

These problems can be tackled using lexical distributional similarity methods (e.g., Weeds (2003)) to automatically generate thesauruses from domain-specific corpora. Such techniques can be used to find semantically similar neighbours of words not in WordNet. Further, it has been shown (McCarthy et al., 2004) that the most distributionally similar neighbours of a word can be used to select the most likely sense of a word in WordNet given the domain. There is also related work (Buitelaar, 2001) which uses a relative term frequency score to compute the domain relevance of a term and thus of a concept in a semantic lexicon such as WordNet. In both approaches, a domain specific corpus is required (either to derive reliable similarity scores or to compute domain relevance scores) and to this end, we are in the process of creating large domain-specific corpora consisting of text retrieved from the Internet using search engines given words in the core lexicon as queries.

In any case, the result at this stage will be a set of possible referents within the ontology for each word in the uttered policy together with an estimate of their plausibility.

3.3. From NL Dependencies to Logical Descriptions

Having mapped each lexical item onto a set of ontological concepts, the next step is to use the output of the shallow parser and the ontology to determine the most likely combination of concepts, and how these fit together. In general, we disambiguate the referents of each local tree of the dependency parse by finding the most coherant referents in the ontology: the most tightly located collection of elements in the ontology.

In policy example 1, where each word used is mapped to a single concept in the ontology, there is also a single path through the ontology that links these concepts in the way specified by the dependency tree. In general, we would expect there to be a mapping between the grammatical dependency relation and the semantic role in the ontology. In the dependency tree, the words *documents* and *printer* are the direct object and indirect object respectively of the verb *print*. The corresponding concepts in the ontology can be the patient and target arguments respectively of a `print` event. Similarly, the adjective *colour* modifies the nouns *documents* and *printer*, which maps to the ontological fact that `colour` is a value of a role that applies to both the concepts of `document` and `printer`. Accordingly, we can generate the following expression of the type of event described by this NL description:

$$\begin{aligned}\text{Print} \quad &\sqcap \quad patient.(\text{Document} \sqcap colourness.\text{COLOUR}) \\ &\sqcap \quad target.(\text{Printer} \sqcap colourness.\text{COLOUR})\end{aligned}$$

In general, the problem is much harder since each word used will map to a number of plausible concepts in the ontology. Thus, each combination of concepts will be scored according to their syntactic dependencies and semantic coherence within the ontology.

There are three clear benefits of using an ontology to generate logical forms. First, the ontology provides a certain amount of disambiguation. In policy example 4, we can determine that *send* is referring to a print event since it is applied to a document and a printer. This approach can be used to disambiguate the word *send* between the concepts `print` and `e-mail`. If we send a document to a printer than we are referring to a print action whereas if we send a document to a person then it is likely we are referring to an e-mail action.

Second, the ontology can be used to identify parsing errors. Although RASP is designed to be robust and accurate, its precision and recall of dependency relations is unsurprisingly less than human annotators. In particular, it has low precision and recall for the indirect object relation (Carroll et al., 1999). For example, the parser may incorrectly identify two words as having an indirect object relationship when they do not. We can identify this as a parser error if the words do not map to concepts that are related by the corresponding semantic role in the ontology. Further, we might be able to use our knowledge of expected concepts in particular roles to correct the parse or hypothesise syntactic dependencies missed by the parser.

Third, the ontology can be used to discover implicit arguments of events. In policy example 2, there is no explicit mention of a printer. However, we can introduce a printer into our logical representation of the event because the only path through the ontology from print to double-sided (assuming our word similarity method has returned a high similarity between *double-sided* and *duplex*) is through the concept of printer. Thus the following logical expression can be generated:

$$\begin{aligned}\text{Print} \quad &\sqcap \quad patient.(\text{Document} \sqcap version.\text{DRAFT}) \\ &\sqcap \quad target.(\text{Printer} \sqcap duplicity.\text{DUPLEX})\end{aligned}$$

3.4. Constraint Generation

The logical expressions we have generated so far describe an event but they do not express the desired constraint on the policy broker. In order to do this we need to be able to determine which parts of the expression make up the condition and which the consequent. We also need to be able to deal with the verbal modifiers such as *always*, *usually* and *never*.

In our printing policy examples, it is always the characteristics of the printer used in a print event that are determined by the characteristics of other objects such as the

document. In general, requests have arguments that are required (and therefore their characteristics will make up part of the condition) and arguments which may be underspecified (and therefore their characteristics will make up the consequent). This information is encoded in the *qualified number restrictions* in the ontological description of the event. For example, the *patient* role of the `Print` concept has the restriction ≥ 1, whereas the *target* role does not. We are also planning investigative work to discover whether the required information can be learnt from corpus data. We expect to find that the conceptual arguments that can be underspecified will correspond to the syntactic arguments that can be omitted.

The verbal modifiers are also of key importance in deciding the overall form of the constraint. However, we note that there are a relatively small number of them and therefore it is possible to enumerate them and their effects. For example, *always* produces a strong positive constraint, "usually" produces a weak positive constraint and *never* produces a strong negative constraint.

3.5. Eliciting Further User Input

As it stands, the system we propose generates a ranked set of possible constraints for each NL policy statement. Rather than trying to resolve any remaining ambiguity (which may in any case be due to a truly globally ambiguous policy), which could result in the broker acting in undesirable or unexpected ways, we envisage presenting the user with the set of ranked alternatives. The user can then select the desired logical form, which is a much simpler task than generating it from scratch. It will also be possible to present undefined concepts (such as "long") to the user and request clarification as to the definition of a "long document". These clarifications will be in the form of *definitional* policies, which extend the ontology.

4. Conclusions and Further Work

This paper describes ongoing research on the use of natural language to express user policies in pervasive computing environments. The central issue addressed in this paper is how the ontology is being used as the basis for interpreting NL descriptions, and in particular the referents of the lexical items in the description. We have presented an approach in which grammatical dependencies generated by RASP are mapped to ontological relations expressed in DL. Throughout the paper, we have highlighted areas and issues which require further investigation. In particular, we are investigating the range of possible user policies and their characteristics, so that we can constrain the natural language interpretations.

5. References

Allen, James, 1984. *Natural Language Understanding*. Benjamin Cummings, 1st edition.

Baader, Franz, Ian Horrocks, and Ulrike Sattler, 2003. Description logics as ontology languages for the semantic web. In *Lecture Notes in AI*. Springer.

Balkanski, Cecile, 1992. Logical form of complex sentences in task-oriented dialogues. In *Proceedings of ACL-1992*.

Bartak, Roman, 2002. Modelling soft constraints: a survey. *Neural Network World*, 12(5):421–431.

Berners-Lee, T., J. Hendler, and O. Lassila, 2001. The semantic web. *Scientific American*, 284(5):34–43.

Borgida, Alex and Ronald Brachman, 2003. *Description Logic Handbook*, chapter Conceptual Modelling with Description Logics. Cambridge University Press.

Briscoe, Edward and John Carroll, 2002. Robust accurate statistical annotation of general text. In *Proceedings of the 3rd International Conference on Language Resources and Evalutation (LREC-2002)*.

Buitelaar, Paul, 2001. Semantic lexicons: between ontology and terminology. In *Proceedings of OntoLex 2001*.

Carroll, John, Guido Minnen, and Edward Briscoe, 1999. Corpus annotation for parser evaluation. In *Proceedings of EACL-99 Workshop on Linguistically Interpreted Corpora*. Bergen, Norway.

Davis, Anthony and Leslie Barrett, 2002. Relationships between roles. In *Proceedings of OntoLex 2002*.

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

Fensel, D., F. van Harmelan, I. Horrocks, D. McGuinness, and P.F. Patel-Schneider, 2001. OIL: An ontology infrastructure for the semantic web. *IEEE Intelligent Systems*, 16(2):38–45.

Horrocks, I. and P. Patel-Schneider, 2001. The generation of DAML+OIL. In *Proceedings of the 2001 Description Logic Workshop*.

McCarthy, Diana, Rob Koeling, and Julie Weeds, 2004. Ranking WordNet senses automatically. Technical Report TR 569, Department of Informatics, University of Sussex.

Owen, Tim, Julian Rathke, Ian Wakeman, and Des Watson, 2003. JPolicy: A java extension for dynamic access control. Technical Report 04-2003, University of Sussex.

Pease, Adam and Christian Fellbaum, 2004. Language to logic translation with phrasebank. In *Proceedings of the 2nd International WordNet Conference*.

Pease, Adam and William Murray, 2003. An English to logic translator for ontology-based knowledge rpresentation languages. In *Proceedings of 2003 IEEE Conference on Natural Language Processing and Knowledge Representation*.

Pease, Adam, Ian Niles, and John Li, 2002. The Suggested Upper Merged Ontology: a large ontology for the semantic web and its applications. In *Working Notes of the AAAI-2002 Workshop on Ontologies and the Semantic Web*. Edmonton, Canada.

Robinson, Jon and Ian Wakeman, 2003. The scooby event-based pervasive computing infrastructure. In *Proceedings of the 1st UK-UbiNet Workshop*. London, UK.

Stevens, R., I. Horrocks, C. Goble, and S. Bechhofer, 2001. Building a reason-able bioinformatics ontology using OIL. In *Proceedings of the IJCAI-2001 Workshop on Ontologies and Information Sharing*.

Weeds, Julie, 2003. *Measures and Applications of Lexical Distributional Similarity*. Ph.D. thesis, University of Sussex.

Reconstructing the Ontology of the Tang Dynasty
A pilot study of the Shakespearean-garden approach

Chu-Ren Huang*, Feng-ju Lo, Ru-Yng Chang*, Sueming Chang***
*Academia Sinica, **Yuan Ze University
130 SEC.2 Academia Road, Nankang, Taipei, TAIWAN 11529, R.O.C
churen@sinica.edu.tw, echinese@saturn.yzu.edu.tw, ruyng@gate.sinica.edu.tw, kati@gate.sinica.edu.tw

Abstract

We propose the Shakespearean-garden approach towards domain ontology construction in this paper. In sum, we suggest that domain lexica can be extracted and obtained for non-standard knowledge backgrounds. Once the comprehensive lexica are collected, a lexical interface between wordnet and our Sinica BOW can be applied. It will allow each lexical item to a conceptual location on Sinica BOW. With the WordNet and SUMO interface, as well as our bilingual correspondence program, each domain lexica can be mapped to a set of SUMO conceptual nodes. These nodes will each be linked to the ontology. We show that the domain ontologies can be constructed directly from synset-ontology pairs, or from the lexical information taken from Wordnet.

1 Background

1.1 Non-Standard Ontology

The construction of an ontology from a knowledge background which is substantially different from ours can be challenging yet rewarding. We will refer to this type of ontology as 'Non-Standard Ontology' for lack of better terms. Work on non-standard ontology presents a dilemma. On one hand, the structure of knowledge is often neither explicated nor represented before the non-standard ontology is constructed. On the other hand, to construct such an ontology, one needs to start with at least some pre-defined terms and conceptual taxonomy, which is in practice a small (upper) ontology. For historical ontologies, it is very rare to find a synchronous ontology from the same period, such as Wilkins (1668). In this case, the structure of the synchronous ontology can be adopted and mapped to a modern system for study. However, for the knowledge domains with no existing ontological available, the greatest challenge also underlines the greatest potential to gain new knowledge. For instance, seventh century Chinese does not have the same scientific knowledge or the philosophical tradition that the current academic world holds to be common. Hence, even though there is much knowledge to be gained, there is also very little to fall back to as the working hypothesis. We will show in this paper how such dilemma can be resolved with successful integration of lexical resources and upper ontology.

1.2Some Basic Facts

The target ontology of this study is the ontology of the Tang dynasty (618-907AD). In this pilot study, we work with the text of the collection of the Tang 300 Poems. We adopt SUMO as our upper ontology. The lexical resources used include the domain lexica extracted from the text and the English-Chinese bilingual wordnet system Sinica BOW.

2 The Shakespearean-garden Approach

We propose a Shakespearean-garden approach to the construction of non-standard ontology. This approach is both lexicon-based and domain-driven. A Shakespearean garden collects and grows all plants referred to in Shakespearean texts. The purpose of a Shakespearean garden is to replicate the botanic knowledge and flora experience of Shakespearean England. A Shakespearean garden works because we can reasonably assume that the plants we collect now are by and large identical to the Shakespearean plants and have the same functions. Similarly, when constructing a non-standard ontology, we propose to start with concrete sub-domains. A chosen domain must have two properties: that it plays roughly equivalent roles in the knowledge backgrounds of the target ontology and the reference ontology (i.e. our contemporary ontology); and that it is empirically verifiable with lexical resources supporting the target ontology. Even though the Shakespearean-garden approach does not guarantee a complete ontology, it will lead to very reliable domain ontologies. When there is sufficient data and knowledge collected, these domain ontologies can be further linked to approach a complete ontology of the target knowledge domain.

Our approach requires a shared upper ontology as the anchor for bootstrapping and for comparative studies. We assume that when two knowledge systems are studied, there will be no meaningful comparison unless both of them can be put in the same representational framework. In the current work, we adopt SUMO (Suggested Upper Merged Ontology, Niles and Pease 2003) as the framework for ontological representations. SUMO was constructed with the explicit goal to serve as the upper ontology of varying knowledge domains by the IEEE's suggested upper ontology workgroup. In other words, SUMO is supposed to be versatile and has robust coverage of general concepts used by different ontologies. Since SUMO is attested with many contemporary knowledge domains, it offers a good foundation for our comparative study of non-standard ontology. In addition, our application to a temporally and culturally far removed knowledge source offers a genuine challenge to the robustness of SUMO. Lastly, as an upper ontology, SUMO avoids elaboration of lower level nodes. Hence there is only a very low probability that it will run into contradictions with the expanded nodes of a non-standard ontology.

While an upper ontology is adopted as the anchor for domain ontology construction, such an upper ontology may not contain all the finer-grained concepts necessary to fully represent the chosen domain. Hence, we propose to use Wordnet to supplement the knowledge. Wordnet as a lexical knowledgebase provides the natural interface between the domain lexica and SUMO (Niles and Pease 2003). In addition, for concepts not explicitly represented in the upper ontology, wordnet lexical semantic relations can be used to construct a conceptual taxonomy.

All the lexical and knowledge resources required for this approach are already integrated in Sinica BOW (Academia Sinica Bilingual Ontological WordNet, Huang et al. 2004). Hence we use Sinica BOW as the primary referential knowledgebase in this study. Sinica BOW integrates three resources: WordNet, English-Chinese Translation Equivalents Database (ECTED, Huang et al. 2003), and SUMO. Referring to Sinica BOW has three advantages. First, it allows access to both lexical semantic relation in WordNet and conceptual taxonomy in SUMO. Second, it allows lexical search in either Chinese or English. Third, it allows research information to be represented in either Chinese or English.

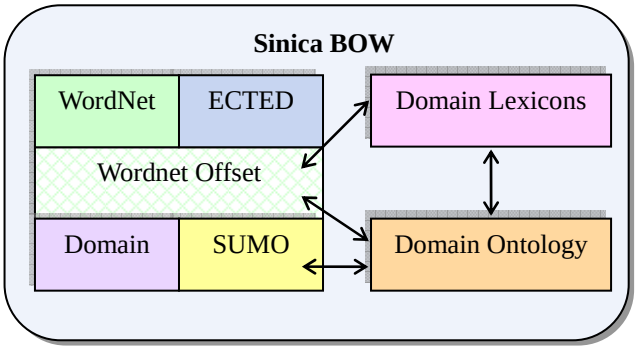


Figure 1: The resource and structure of Sinica BOW

3 Mapping Lexical Data to Ontology

3.1 Preparing the Lexical Resources

Tang civilization (618-907AD) was one of the most vibrant periods of Chinese civilization. It welcomed and integrated elements from many of the neighboring non-Han civilizations. In turn, Tang civilization was also venerated and imitated by neighboring countries. The Japanese civilization, for instance, borrowed generously from Tang, including the kanji writing system. It is not an exaggeration to claim that the classical roots of Japanese civilization are actually Tang civilization. Hence, the ontology of the Tang dynasty has far more implications than being an ontology of a long-gone historical period. It may shed light on how heterogeneous knowledge systems integrate, as well as how a borrowed knowledge system develops in the new cultural background.

As a pilot of the main study of constructing an ontology based on the more than 10 millions characters in textual archives from the Tang Dynasty, we construct an ontology based on the famous anthology of The 300 Tang Poems. The text of the 300 Tang Poems contains slightly more than 15,000 characters. This is one of the most important and popular collections of Chinese literature. Its importance far out-weights its relative small size. In addition, since it is poetry, the conceptual density, as represented by the lexical types contained, is high. In this pilot study, the words and classification of words in the text are hand-tagged. The choice of manual tagging is made because our tagger is not tested for domain classification, even though it performs the task of pos tagging very well. The relatively small size of the text also allows manual work to be done efficiently. The highly reliable result will serve as valuable training data for future automatic tagging classification. There is already a classical Chinese tokenizer combining segmentation and tagging available from Academia Sinica. This tokenization program, adopting the basic design of Chen and Liu (1992), is very robust and performed well in the first SigHAN Chinese segmentation bakeoff in 2003. It has also successfully segmented over 5 million words of classical Chinese texts for the language archives project at Academia Sinica.

Three sub-lexicons from the Tang 300 Poems were extracted for domain ontology construction: animals, plants, and artifacts. A total of 176 words were assigned to the three domain lexica: The animals lexicon contains 64 words; the plants lexicon contains 59 words; and the artifacts lexicon contains 53 words. The result from the animal and plant domains will be reported in this paper. These domains are chosen because their meanings are referential and rich. Since they are referential, it is more likely to uniquely determine the meaning of each term. On the other hand, these are familiar terms and important poetic devices used to invoke empathy or express feelings.

The second step in the preparation of the lexical resources for ontology-building is the identification of the appropriate sense of each word for the target knowledge domain. There are two issues involved here. First, as most words are assigned more than one senses in wordnet, we need to identify the correct sense. Second, as these words are used over 11 hundred years ago, some meanings may have become obscure or changed. We need to identify the intended meaning. A batch query on these 176 words was sent to Sinica BOW. Of the 176 words, only 100 words found complete matching entries in the Chinese part of the bilingual wordnet. We then expand the query to include words that share the initial or ending characters. The expanded query still left 24 words with no possible matches in the current version of BOW. These 24 words were later assigned correct translation and meaning with manual dictionary lookup. For words with direct sense assignment from WordNet, the link form BOW to SUMO ontology is utilized. When a sense does not belong to the target knowledge domain, it is discarded. The senses that belong to the target domain by SUMO assignment is kept for next step. Even though there were in average 2.18 senses assigned for each word, the domain requirement quickly reduced the number of possible senses to close to one.

It is important to notice that expertise knowledge is crucial in the identification of word senses when dealing with a non-standard knowledge domain. A good example is the word *mei2*, with grass radical found in the Tang poems. Its dominant sense in contemporary Chinese equals to berry, as in strawberry '*cao3mei2*'. However, further investigation showed that such sensed did not exist in Tang dynasty. The word refers to a kind of moss instead. In other words, although the Chinese character composition reinforces its position in the plants domain, its actual reference cannot be reliably determined by using standard lexical knowledge.

Expertise knowledge and manual editing is also crucial for the words that do not find direct match in Sinica BOW. For example, *hu2jia1* is a particular musical instrument that was first invented and played by the Tartar people and no longer commonly used. Hence its lack of an equivalent in the English language is not surprising. To solve this problem, we consult similar senses from

Wordnet. Since *hu2jia2* is a kind of tubular wind instrument, we considered it to be a kind of pipe, which does occur in WordNet and is linked to SUMO.

3.2Constructing Domain Ontology

Once each lexical item is assigned a unique correct Chinese sense and its corresponding English synset, it can be mapped through Sinica BOW to a SUMO conceptual node. When there is no exact match, lexical semantic relations from WordNet are consulted to establish relation between a lexical item and SUMO. For lexical items that are thus assigned to an appropriate SUMO node, the construction of the domain ontology is as simple as connecting two dots. This is largely the case for the animals ontology (Figure 4).

On the other hand, SUMO as an upper ontology does not necessarily offers sufficient knowledge structure for all domains. For instance, although plants can be considered to be equally salient as animals conceptually, SUMO only gives the very rough-grained classification of FloweringPlant and NonFloweringPlant. Hence we need to use the lexical semantic relations from WordNet to construct the hierarchical conceptual network, i.e. the proposed domain ontology. In this case, we cannot simply copy and connect the relations. Since WordNet's main goal is to record all cognitively relevant semantic relations, not all relations can fit in a rigorous conceptual classification and inference system. Hence, after bootstrapping with all WordNet synsets and relations marked, an important step is to prune the resultant tree for both inconsistency and redundancy. The plants ontology in Figure 5 is the wordnet-based ontology after extensive pruning.

In establishing the link between a sense and a ontology node, it is important to notice that the SUMO-WordNet link is established with the contemporary background knowledge of the English speaker world. Hence it is likely to find that a non-standard ontology based on a different system will require a totally different conceptual assignment. An instance is of such mismatches involves *mou2hu2*, which is a kind of silk flag. A flag, according to both the literary context and the assigned lexical sense, should be a piece of artifact, solid and substantial. However, the SUMO-WordNet link that Sinica Bow follows mapped it to the conceptual node of "Icon." This may be appropriate when a flag is used in signing, but not appropriate in the Chinese context. Hence we simply correct the link and assign it to artifact.

What is more interesting in terms of linguistic use involves words that seem to carry the same meaning, while involves fundamentally different conceptualization. The difference in conceptualization requires assignment to a different ontological location. One such example is *dai4mei4*, which is given the sense of 'a beaded sea turtle,' and seems to be a straightforward case of a kind of animal. However, when we refer to the context, the sentence actually refers to 'a beam inlaid with *dai4mai4* '. In other words, it refers to the materials used in decorating a building. It is the shell of the turtle that has been ground and polished like a piece of jade. It is also interesting to note the fact that these two characters used have a jade radical, rather than an animal or fish radical. Both the context and the written form suggest that the sense being used here is the material, and there in no evidence suggesting that Tang people know that the *dai4mei4* material comes from a turtle. Hence this word is not included in the animals ontology.

On the other hand, when metonymy is used, it is often possible to argue that the original sense is invoked. An example in our study is *shuang1li2*, double-carp, which refer to a letter since letters are traditionally sent in a word box with two carps carved on top. In this case, even though the actual reference is not the animal, but the lexical metonymy necessarily involve the image of the fish. Hence we consider the concept of carp is used, and hence justifying our including carp as an attested case for the animals ontology for Tang.

4 Result and Discussions

The result of this pilot study will include three semi-automatically constructed sub-ontologies: animal, plant, and artifact. The first two are completed and will be discussed here. The top part of each ontology is mapped to SUMO. The lower part of each ontology is extended using WordNet relations. These ontologies as well as the attached lexical terms will have Chinese-English bilingual representation.

The first generalizations that can be obtained are from the distribution of these domain terms in the texts. The total frequency of these three domains ranges from 1.65% to 1.89%. These are relatively high compared to a balanced corpus. In a balanced corpus, the top 20 animal or plant domain terms comprise of less than 1%.

The second generalizations can be made from the distribution among the different terms within the domain. Among animal concepts, the total frequency of birds is over 38%, and hoofed mammals over 30%. These two kinds each far exceed all the other eight kinds of animals combined. This fact should have implications on either the fauna of Tang, or the poetic choice of images. Even more striking is the fact that of all plants, flowering plants consist of over 95% of the instances in the texts. This fact should not be surprising because of the strong poetic image that a flower presents.

After the sub-ontologies are constructed, comparative studies of the Tang ontological structure with our contemporary ontology (based on SUMO) will be conducted. For instance, we found that among the order of mammals, the families of marsupials and marine mammals are missing. The absence of marsupials is expected since it is a fact of science history that they were discovered much later. The absence of marine mammals may point to the fact that the Tang civilization is mainly land-based. In addition, we also found two interesting facts in other branches. First, almost all invertebrates that are documented are (winged) insects. And among the non-mammal vertebrates, with only less than 5 exceptions, all documented lexical items refer to bird. A possible explanation of the idiosyncrasy is the Tang civilization's fascination with flying. We know as a fact that flying is a recurring theme in paintings from this period, and occur in poetry too.

The plants ontology of Tang offers a good test case of how to bootstrap an ontology with lexical knowledgebases such as wordnets. We showed that when the lexical resource contains sense and lexical semantic relations information, it is possible to use the information to bootstrap a domain ontology. The crucial challenge here is how to turn the set of pair-wise and

lexicon-driven relations to a taxonomical hierarchy. An issue that will recur is how to deal with same level nodes that are classified and assigned with diagonal criteria. One such example is the classification of plants in Figure 5. FloweringPlants and HerbaceousPlants and AcquaticPlants create partially overlapping classes. These are all linguistically and cognitively motivated and cannot subsume each other. Given the fact that even an upper ontology like SUMO acknowledes such human cognitive facts and allows multiple inheritance, there is still reservations that an ontology can quickly become non-trackable if no constraints are put on such cross-classification. This is an issue that merits in-depth formal and theoretical deliberation.

5 Conclusion

In this current study, we propose the Shakespearean-garden approach to the construction of non-standard ontology. We showed with a pilot study that such an approach is feasible, especially when supported by the right combination of lexical knowledge sources and upper ontology. In addition, we showed that the constructed sub-ontology allows us to have a comprehensive view of the knowledge system of a civilization that no longer exists. Such a representation will offer a unique opportunity to study how their world differs from ours and how they view the world differently from us.

A natural extension of the current work is to try to piece these sub-ontologies together to form a skeletal ontology for the Tang dynasty. In order to carry out this full-scale work, we have already started the design and construction of automatic tools to construct domain ontology based on domain lexicons and SUMO. This will integrate the knowledge we gain from the current work as well as modules from existing systems, such as Sigma system constructed by Adam Pease. Such a working environment will facilitate the ultimate goal of the Shakespearean-garden approach. In addition, we will also try to apply the simultaneous bilingual mapping approach to construct a modern domain. Ultimately, we would like to see if it still plausible to construct ontology based on a shared upper ontology even if the background knowledge systems are drastically different.

The current work on the domain knowledge of Tang civilization willl also provide solid foundation for future work on metaphor. Based on Lakoff's contemporary theory of metaphor, Ahrens et al. (2003) shows that the crucial step in predicting and explanation of the use of linguistic metaphors lies in capturing the rules governing the mapping between source domain and target domain knowledge. For the historical poetic work such as Tang poetry, an additional challenge to the study of metaphor would be the precise characterization of the source domain knowledge. Our non-standard ontology can be viewed as the foundational work defining source domain knowledge in Tang poetry. With the source domain knowledge described, we will be able to develop in-depth study of Tang poetic metaphors in the future.

Lastly, the issue regarding the relation between a wordnet and an ontology is also touched upon. In the Shakespearean-garden approach, it is crucial that the specific domain lexicon can be obtained and annotated with correct lexical semantic information. However, how can lexical semantic relations be best used in an ontological study remains a challenging and promising issue.

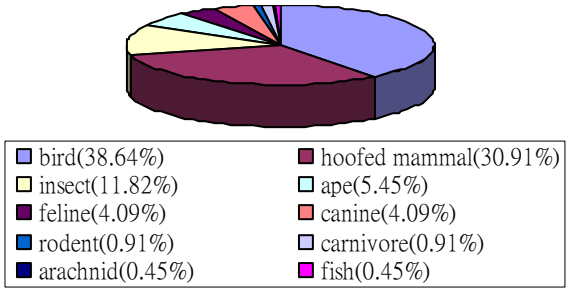


Figure 2: Distribution of animal concepts in Tang 300

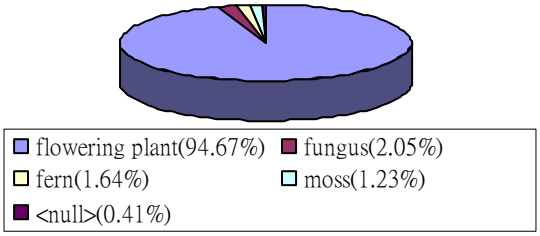


Figure 3: Distribution of plants concept s in Tang 300

Online Resources

Sinica BOW: http://BOW.sinica.edu.tw/
SUMO: http://ontology.teknowledge.com/
WordNet: http://www.cogsci.princeton.edu/~wn/
Tender Lyrics-The 300 Tang Poems (in Chinese) http://cls.admin.yzu.edu.tw/300/HOME.HTM
CKIP Segmentation and Tagging Program http://corpus.ling.sinica.edu.tw/project/LanguageArchive/lc_index.html

Reference

Ahrens, Kathleen, Chu-Ren Huang, and Siaw-Fong Chung. (2003). Conceptual Metaphors: Ontology-based representation and corpora driven Mapping Principles. Presented at the Workshop on Lexicon and Figurative Language. An ACL2003 Workshop. July 11, Sapporo, Japan.

Chang, Ru-Yng and Feng-ju Luo. (1999). Cross-platform Web-bases Learning System—the construction of Tender Lyrics-The 300 Tang Poems (in Chinese). Presented at 1999 Taiwan Symposium on Taiwan Academic network. Kaohsiung.

Chen, K.-J. and S.-H. Liu. (1992). Word Identificaiton for Chinese Sentences. Proceedings of COLING92. 501-505.

Fellbaum, Christine. Ed. (1998). WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.

Huang, Chu-Ren, Ru-Yng Chang, and Shiang-bin Li. (2004). Sinica BOW (Bilingual Ontological Wordnet): Integration of Bilingual WordNet and SUMO. To be presented at the LREC2004 conference. May26-28. Lisbon.

Huang, Chu-Ren, Li, Xiang-Bing, Hong, Jia-Fei. (2004). Domain Lexico-Taxonomy:An Approach Towards Multi-domain Language Processing. Proceedings of the Asian Symposium on Natural Language Processing to Overcome Language Barriers. March 25-26, 2004. Hainan Island.

Huang, Chu-Ren, Feng-ju Lo, Ru-Yng Chang, and Sueming Chang. (2004). Sinica BOW and 300 Tang Poems: An overview of a bilingual ontological wordnet and its application to a small ontology of Tang poetry. Presented at the Workshop on Possibilities of a Knowledgebase of Tang Civilization. Institute for Research in Humanities, Kyoto University. February 20-21.

Huang, Chu-Ren, Elanna I.J. Tseng, Dylan B.S. Tsai, & Brian Murphy. (2003). Cross-lingual Portability of Semantic Relations: Bootstrapping Chinese WordNet with English WordNet Relations. Language and Linguistics. 4(3), 509--532.

Huang, Chu-Ren. Elanna I.J. Tseng & Dylan B.S. Tsai. (2002). Translating Lexical Semantic Relations: The first step towards multilingual Wordnets. In Proceedings of the COLING2002 workshop: SemaNet: Building and Using Semantic Networks. Taipei, Taiwan.

Niles, I. & Pease, A. (2003). Linking Lexicons and Ontologies: Mapping WordNet to the Suggested Upper Merged Ontology. In Proceedings of the IEEE International Conference on Information and Knowledge Engineering. (IKE 2003), Las Vegas, Nevada.

Niles, I., & Pease, A., (2001). Toward a Standard Upper Ontology. In Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001). Ogunquit, Maine.

Pease, A., (2003). The Sigma Ontology Development Environment. In Working Notes of the IJCAI-2003 Workshop on Ontology and Distributed Systems. Volume 71 of CEUR Workshop Proceeding series.

Wilkins, J. (1668). An Essay Towards a Real Character, and a Philosophical Language. Reprinted in 2002. Thoemme Press.

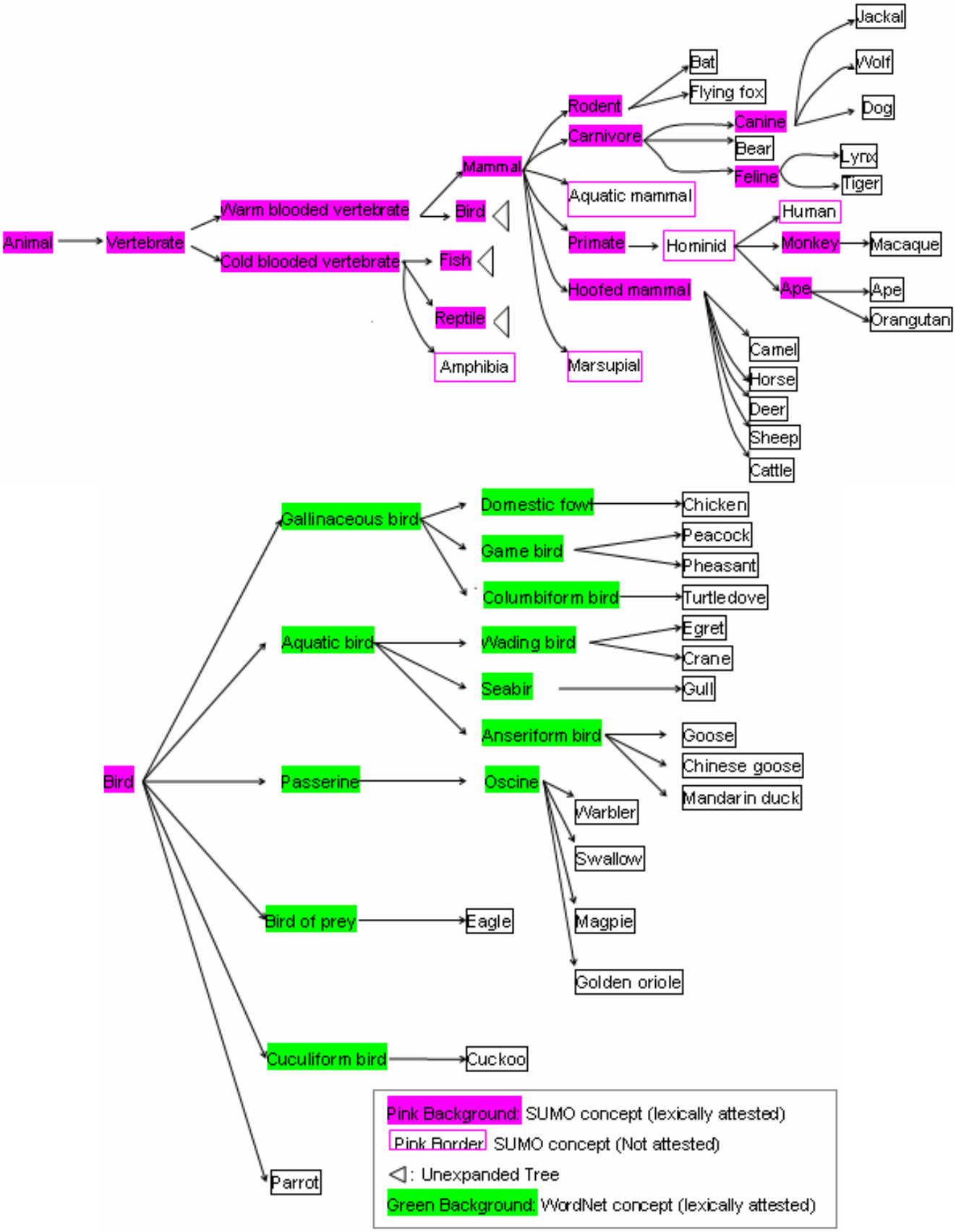


Figure 4: Tang Animals Ontology

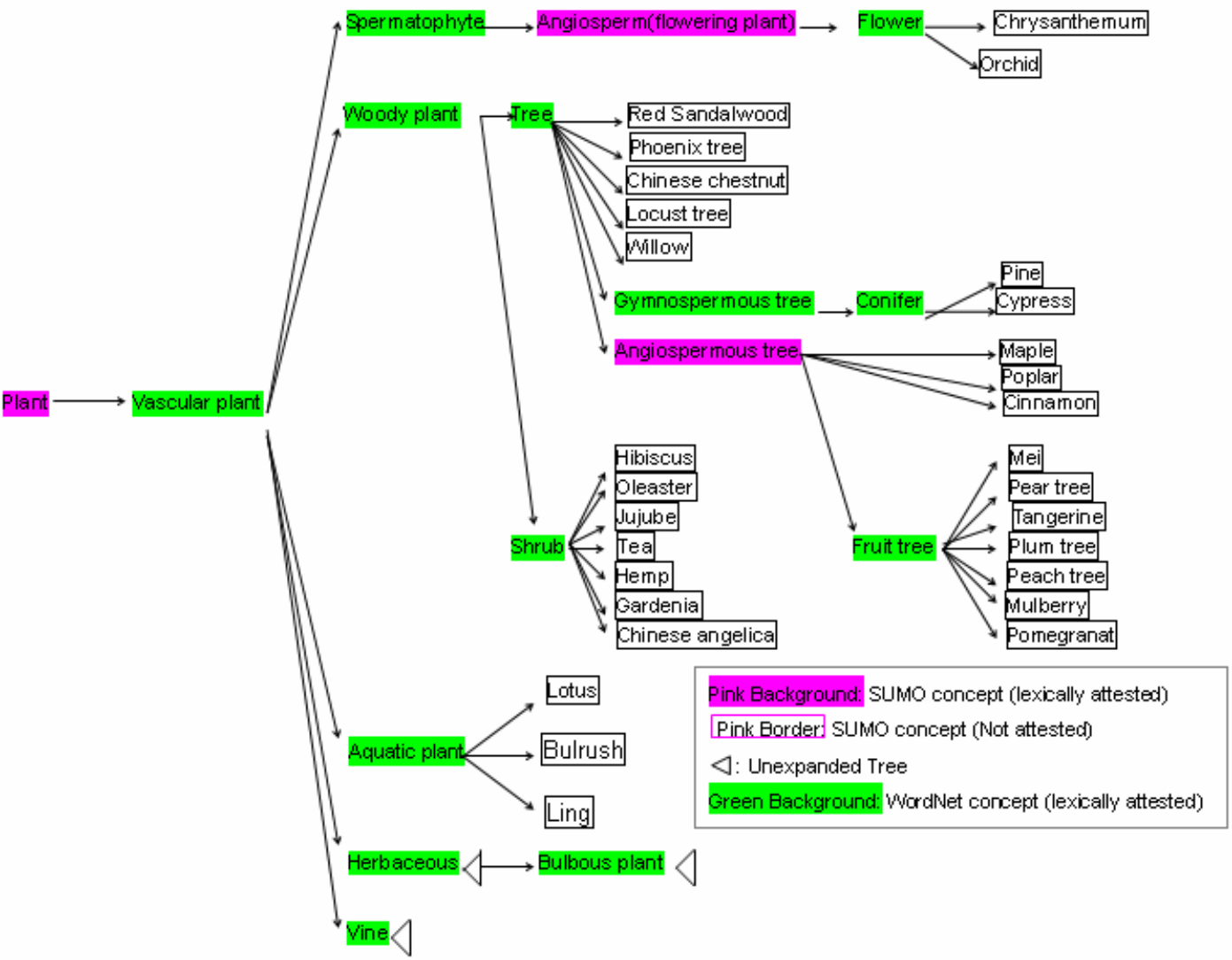


Figure 5: Tang Plants Ontology

OntoTag's Linguistic Ontologies as a Reference for Semantic Web Annotations

Guadalupe Aguado de Cea
DLACT, UPM
lupe@fi.upm.es

Inmaculada Álvarez de Mon
DLACT, UPM
ialvarez@euitt.upm.es

Antonio Pareja-Lora
SIP, UCM / LIA, UPM
apareja@sip.ucm.es

Campus de Montegancedo, s/n. 28660 – Boadilla del Monte (Spain)
Tel: (+34) 91 336 74 13 Fax: (+34) 91 336 54 72

Abstract

Following the road in-between purely linguistic annotation and solely ontology-based annotations for the Semantic Web, a hybrid (ontological and linguistic) model and platform, called OntoTag, has been created, aiming at better machine communication, interoperability and language understanding; these capabilities are derived from the incorporation into the platform of a set of linguistic ontologies, the main topic of this demonstration, which are the main referent for the generation of multi-leveled and standardized annotations of Semantic Web documents within OntoTag.

Introduction

Many are the schemas developed so far for the different kinds of annotation required in the field of **Corpus Annotation**. Besides, with the appearance of the **Semantic Web** (Berners-Lee *et al.*, 1999) many other schemas have been devised (most of them based on ***ontologies*** (Gruber 1993; Borst 1997)) for *web page annotation*. Thus far, on the one hand, Corpus Linguistics researchers are trying to cover as many levels and aspects of annotation –from a linguistic point of view– as possible to describe language phenomena (Wilson & Thomas, 1997; Schmidt, 1988); on the other hand, researchers in the Semantic Web area are focusing on achieving a sound model of semantic annotation for web pages, that is able to capture as much knowledge from these pages as possible, so that computers can process them in a much smarter way (Benjamins *et al.*, 1999, Motta *et al.*, 1999, Luke *et al.*, 2000, Staab *et al.*, 2000). However, there is an emerging road in-between, nowadays, that seeks to merge and sum up both kinds of annotations, combining them in order to bear a new, unified, multilingual, flexible, extensible and fully semantic model of annotation, useful for both communities (Aguado *et al.*, 2003a). Moreover, as shown by the ISO - TC37SC4 (2003) "there is an increasing need for new standardization as well as urgent recognition of existing *de facto* standards and their transformation into International Standards". In fact, one of the main aims of this committee is "to develop standards and related documents to maximize the applicability of language resources". The OntoTag model for Semantic Web Annotation (Aguado *et al.*, 2003b), whose Linguistic Ontologies we present here, is being developed following this in-between road aforementioned, as well as a number of guidelines hitherto published (EAGLES 1996a, 1996b; CES 1999; MILE 2003; GDA 2002), in order to achieve the goal of standardisation sought within the ISO - TC37SC4 committee.

OntoTag's Linguistic Ontologies

One of the main components of the OntoTag model is its set of linguistic ontologies, devised to represent the structure and relationships between the elements of language at different linguistic levels. The kind of elements and relationships considered in them are the ones usually included in existing annotation schemas and also those already discussed in the literature but not implemented yet (Wilson & Thomas, 1997; Schmidt, 1988) as well as some others, determined by our research team.

OntoTag's Core Linguistic Ontologies

First of all, a Linguistic Level Ontology (LLO) has been implemented both to capture the stratification of natural language analysis and generation and to simplify the study of the other elements. Then, following the EAGLES guidelines for morpho-syntactic annotation of corpora (EAGLES 1996a), but obviously broadening its scope, three different ontologies have been implemented to represent the category-attribute-value formalism at all levels of annotation (morpho-syntactic, syntactic, semantic, discourse and pragmatic): a Linguistic Unit Ontology (LUO), a Linguistic Attribute Ontology (LAO), and a Linguistic Value Ontology (LVO).

The Linguistic Unit Ontology (LUO) includes all the units (categories) identified at the different levels of annotation considered in the LLO, and incorporates an adaptation of the SIMPLE (2000) ontologies at the semantic level; the Linguistic Attribute Ontology (LAO) includes the various attributes associated to the units in the LUO; and the Linguistic Value Ontology (LVO) accounts for the possible values of the attributes in the LAO.

OntoTag's Supplementary Linguistic Ontologies

Complementing these four ontologies, a fifth one (the Linguistic Pattern Ontology, LPO) has been designed for the representation of the patterns that these units follow when combined in an utterance. Finally, the OntoTag Integration Ontology (OIO) establishes the main relationships between documents (annotated and non-annotated), units, attributes and values both in the linguistic and in the ontological areas of annotation.

OntoTag's Linguistic Ontologies: Application

The application of these six ontologies in the OntoTag annotation model is twofold: first, as discussed above, they identify the different elements (mostly linguistic, but also ontological) that are annotable in the Semantic Web field; second, once the ontology has been populated (instantiated) by the annotations obtained with OntoTag, they will also act as a repository or database of these annotations.

Further information about OntoTag's linguistic ontologies, their respective roles and interaction, as well as their properties and application (to pragmatic purposes or with automatic means, for instance) can be found in Aguado *et al.* (2004a; 2004b).

Conclusions

To conclude, we could say that, derived from the extensibility and flexibility capabilities of the Linguistic Ontologies presented here, the OntoTag model of annotation inherits these properties as well. It can also be considered as domain independent in the sense that these source ontologies can be replaced and, still, meaningful annotations would be obtained. Due to the multilingual nature of the EAGLES guidelines, followed (and broadened) in the design of the different Linguistic Ontologies, OntoTag becomes also applicable to the annotation of the languages studied in these guidelines. The consensual nature of ontologies and the sources used in their construction (EAGLES 1996a, 1996b; CES 1999; MILE 2003; GDA 2002; Dubuc & Lauriston 1997; Faber & Tercedor 2000; Mel'cuk 1996, 1988; Pustejovsky 1998) enables them (and the annotations obtained with them) so as to be considered standardised.

Acknowledgements

This research has partly been supported by the ministry of Science and Technology grant (Reference TIC2001-2745 CONTENTWEB project) and by the UPM grant (Reference 14286 PLAN-H-SEMWEB project)

References

- Aguado de Cea, G., Álvarez de Mon, I., Pareja-Lora, A. 2003a. "Primeras aproximaciones a la anotación lingüístico-ontológica de documentos de Web Semántica: OntoTag" in *Revista Iberoamericana de Inteligencia Artificial*, Vol 1, pp 37-49.
- Aguado de Cea, G., Álvarez de Mon, I., Gómez-Pérez, A., Pareja-Lora, A. 2003b. "OntoTag: XML/RDF(S)/OWL Semantic Web Page Annotation in ContentWeb" in *Proceedings of the 2nd Workshop on NLP and XML (NLPXML-2003) – Language Technology and the Semantic Web*, pp. 25-32. 10th Conference of the European Chapter of the Association for Computational Linguistics. EACL'03. Budapest, Hungary.
- Aguado de Cea, G., Álvarez de Mon, I., Gómez-Pérez, A., Pareja-Lora, A. 2004a. "OntoTag's Linguistic Ontologies: Improving Semantic Web Annotations for a Better Language Understanding in Machines" in *Proceedings of the International Conference on Information Technology (ITCC 2004)*, pp. 124–128. IEEE Computer Society:Washington, Brussels, Tokyo.
- Aguado de Cea, G., Álvarez de Mon, I., Pareja-Lora, A. 2003b. "OntoTag's Linguistic Ontologies: Enhancing Higher Level and Semantic Web Annotations" in *Proceedings of the 4th Language Resources and Evaluation Conference (LREC 2004)*. Lisbon, Portugal.
- Álvarez de Mon-Rego, I. 2003. La cohesión del texto científico-técnico: un estudio contrastivo inglés-español. Universidad Complutense de Madrid (forthcoming).
- Benjamins, V.R., Fensel, D., Decker, S., Gómez-Pérez, A. 1999. $(KA)^2$: Building Ontologies for the Internet: a Mid Term Report. *IJHCS, International Journal of Human Computer Studies*, 51, pp. 687–712.
- Berners-Lee, T., Fischetti, M. 1999. *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by its Inventor*. Harper. San Francisco.
- Borst, W. N. 1997. Construction of *Engineering Ontologies*. PhD thesis, University of Twente, Enschede.
- CES. 1999. *Corpus Encoding Standard*. http://www.cs.vassar.edu/CES/
- Dubuc, R. and Lauriston, A. 1997. "Terms and Contexts" in Wright, S.E. and G. Budin, *Handbook of Terminology mangement* Vol 1, John Benjamins:Amsterdam, Philadelphia, pp. 80-87.
- EAGLES. 1996a. *EAGLES: Recommendations for the Morphosyntactic Annotation of Corpora*. EAGLES Document EAG--TCWG—MAC/R.
- EAGLES. 1996b. *EAGLES: Recommendations for the Syntactic Annotation of Corpora*. EAGLES Document EAG--TCWG—SASG/1.8.
- Faber, P. and Tercedor, M. 2000. "Codifying conceptual information in descriptive terminology management" in *Meta*, XLVI, 1, pp. 192-204.
- GDA. 2002. Global Document Annotation Initiative: The GDA Tag Set. http://www.i-content.org/GDA/tagset.html
- Gruber, T. R. 1993. "A Translation Approach to Portable Ontologies" in *Journal on Knowledge Acquisition,* Vol. 5(2), 199-220
- ISLE. 2003. http://www.ilc.cnr.it/EAGLES96/isle/ISLE_Home_Page.htm
- ISO - TC37SC4. 2003. http://www.tc37sc4.org
- Luke S., Heflin J. 2000. *SHOE 1.01. Proposed Specification*. SHOE Project. http://www.cs.umd.edu/projects/plus/SHOE/spec1.01.htm
- Motta, E., Buckingham Shum, S. Domingue, J. 1999. "Case Studies in Ontology-Driven Document Enrichment" in *Proceedings of the 12th Banff Knowledge Acquisition Workshop*, Banff, Alberta, Canada.
- Mel'čuk, I. A. 1988. *Dependency Syntax*, New York: State University of New York Press.

- Meľčuk, I. A. 1996. “Lexical functions: a tool for the description of lexical relations in a lexicon”, in Wanner, L. *Lexical functions in lexicography and natural language processing*, John Benjamins: Amsterdam, Philadephia.
- Pustejovsky, J. 1998. *The generative lexicon,* MIT Press:Cambridge, Massachussets.
- Schmidt, K. M. 1988. Der Beitrag der begriffsorientierten Lexicographie zur systematischen Erfassung von Sprachwandel und das Begriffswörterbuch zur mhd. Epik. *Mittelhochdeutsches Wörterbuch in der Diskussion*, ed. by Bachofer, W. Max Niemeyer:Tübingen, 35–49.
- SIMPLE Project. 2000. http://www.ub.es/gilcub/SIMPLE/simple.html
- Staab, S., Angele, J., Decker, S., Erdmann, M., Hotho, A., Mädche, A., Schnurr, H.-P., Studer, R. 2000. Semantic Community Web Portals. *WWW'9*. Amsterdam.
- Wilson, A., Thomas, J. 1997. Semantic Annotation. *Corpus Annotation: Linguistic Information from Computer Text Corpora*, R. Garside, G. Leech & A. M. McEnery (ed.). Longman:London.

Sinica BOW: Integration of Bilingual WordNet and SUMO
A Demonstration

Chu-Ren Huang
Academia Sinica.
130 SEC.2 Academia Road, Nankang, Taipei, TAIWAN 11529, R.O.C
churen@sinica.edu.tw

Abstract

The Academia Sinica Bilingual Ontological Wordnet (Sinica BOW) integrates three resources: WordNet, English-Chinese Translation Equivalents Database (ECTED), and SUMO (Suggested Upper Merged Ontology). The three resources were originally linked in two pairs: WordNet 1.6 was manually mapped to SUMO (Niles and Pease 2003) and also to ECTED (the English lemmas in WordNet were mapped to their Chinese lexical equivalents). ECTED encodes both equivalent pairs and their semantic relations (Huang et al. 2003). With the integration of these three key resources, Sinica BOW functions both as an English-Chinese bilingual wordnet and a bilingual lexical access to SUMO. Sinica BOW allows versatile access and facilitates a combination of lexical, semantic, and ontological information. Versatility is built in with its bilinguality, and the lemma-based merging of multiple resources. First, either English or Chinese can be used for the query, as well as for presenting the content of the resources. Second, the user can easily access the logical structure of both the WordNet and SUMO ontology using either words or conceptual nodes. Third, multiple linguistic indexing is built in to allow additional versatility. Fourth, domain information allows another dimension of knowledge manipulation.

1. Sinica BOW: Overview

The ***Sinica BOW*** (Academia Sinica Bilingual Ontological Wordnet) is intended as a linguistic infrastructure for knowledge representation and knowledge engineering. It is built upon the relation-based structure of WordNet. On one hand, a bilingual English-Chinese wordnet is constructed with the crucial design feature of treating bilingual translation correspondences as lexical semantic relations (Huang et al. 2003). On the other hand, SUMO (Suggested Upper Merged Ontology) is adopted as the shared system of conceptual categorization (Niles and Pease 2001). SUMO is also one of the first conceptual categorization systems to be mapped to an English lexicon (Niles and Pease 2003). With the mapping between SUMO and WordNet synsets, each English sense can be assigned a position in the ontology and for applications in knowledge engineering. When this mapping is linked with the bilingual wordnet, each Chinese lemma can now be conceptually categorized by with the same upper ontology. The design of Sinica BOW with the combination of ontology and wordnet has three intended goals: 1) To assign to each linguistic form a robust and rigorously defined conceptual location, 2) To clarify the relation between conceptual classification and linguistic instantiation, and 3) To facilitate genuine cross-lingual access of knowledge.

In order to facilitate the above goals, the online system of Sinica BOW (http://BOW.sinica.edu.tw) allows lexical searches in either Chinese or English to return ontological information (again, in either language). Searches on Sinica BOW can return the following information: Sense-based English-Chinese translation equivalency, English word-sense-based ontology and inference, Chinese word-based ontology and inference, Word-sense-based domain specification.

In addition to the integration of Wordnet and ontology, it is also an important goal of Sinica BOW to integrate lexical resources. Sinica BOW's design is lemma-driven. A lexical database of word forms is first compiled by integrating multiple lexical resources. This becomes the central database for lexical management for Sinica BOW. Making use of this lexical database, a lexical search may link to either the main BOW knowledgebase or any of the corresponding entries in an online resource

2. Presentational Versatility

Sinica BOW allows versatile access and facilitates a combination of lexical semantic and ontological information. The versatility is built in with bilinguality, and lemma-based merging of multiple language sources.

The versatility and combinatory presentation is crucial to the presentation of a knowledge system.

2.1. Lexicon-driven Access

The Sinica BOW access is both lexicon-driven and knowledge-based. Since knowledge representation is the main concern of Sinica BOW, the query can be either lemma based or conceptual node based. In addition, query results are directed linked to the full knowledgebase and expandable. Each query returns a structured lexical entry, presented as a tree-structured menu, as seen in Figure 1. A keyword query returns with a menu arranged according to word senses, as shown in Figure 2. The top level information returned including POS, usage ranking, and cross-reference links. In addition to wordnet information, cross-references to up to five resources are pre-compiled for either language. For an English word, the main resource is of course the bilingual wordnet information that our team constructed. Major outside references are listed for quick hyperlink. These include corpora and both EC and CE dictionaries. For Chinese, the main resource is again our bilingual wordnet. In addition, links are established to Sinica Corpus, to Wen-Land (a learner's Lexical KnowledgeNet), and to online monolingual and bilingual dictionaries. The tree-structured query result is not only itself a menu of accessible resources, it is also a quick overview of the distributional characteristics of the query term.

The access to the ontology and the domain

taxonomy are also lexicon-driven. That is, in addition to using the pre-defined ontology or domain terms (in either English or Chinese), a query based on a lexical term is also possible. For SUMO, it will return a node where the word appears in. It can also be achieved by looking up the ontological or domain node the word belongs to.

Figure 1: Initial Return of Lemma Search

One last but critical feature of the lexicon-driven access is the possibility to re-start a query with any lexical node. When expansion reaches at the leave node and results in a new word, clicking on the word is equivalent to start a new keyword search.

2.2. Multiple Knowledge Sources

Sinica BOW preserves the logical structure of both WordNet and SUMO ontology yet links them together to allow direct accesses to the merged resources. This is shown in Figure 2. In a wordnet search, the available information is listed under each sense and include: POS, synset, sense explanation, translation, and list of lexical semantic relations. In addition, we add the domain information, translation equivalents, and link to the corresponding SUMO node. Each item is expandable to present the database content. For instance, Figure 2 shows the query return for the lemma 'fish', with the Part_Meronym and Holonym of sense 4 expanded. The field of domain and SUMO will lead directly to the corresponding node in the domain taxonomy of the ontology and allow further exploration. For instance, the menu item of the mapped SUMO node links to the SUMO representation, as well browsing of the SUMO ontology and axioms.

Figure 2: A sample lemma query result of Sinica BOW

2.3. Taking Advantages of Linguistic and Knowledge Structures

Two more aspects of versatility are achieved through the use of higher level linguistic generalizations and knowledge of domain taxonomy to organize information. Such higher level generalizations are accessed through the index pages of Sinica BOW, shown in Figure 3.

First, Sinica BOW also integrates the rich structural information of the integrated lexical resources. Glyph, phonological, and morphological structures can all be used to help access the ontological wordnet. This work has implications far beyond being convenient search tools. It is often claimed that the glyph composition (e.g. radicals) in Chinese has its semantic base. This can also be said about the morphological composition (and to a much lesser degree, phonological composition). In other words, the integration allows us to study the possible links between these lexical structure and conceptual classifications. The clustered search tools that are available now include search by prefix, suffix, POS, and frequency.

Figure 3: A sample index page of Sinica BOW

Second, one important motivation for constructing Sinica BOW is the premise that linguistic elements and conceptual atoms may not be clearly delimited. And that classification by linguistic elements or by conceptual terms may yield different, yet cognitively significant results. Hence it is important for the system to access information classified by conceptual terms as well as lexical forms. Sinica BOW now allows query by either any term defined in SUMO, or in the locally maintained domain taxonomy. For instance, choosing the domain taxonomy of 'religious music' will return all items belonging to that domain. Similarly, specifying the ontological term of 'FloweringPlants' in a query, and the system will return all entries satisfying that conceptual classification.

Lastly, all the above classification can be used in conjunction to define the exact intersection of terms a user looks for.

3. The Multilingual Properties of Semantic Relations

In addition to relying on lemmas as retrieval keys, a crucial step in establishing synergy between language and knowledge resources is to identify the conceptual atoms that apply equally effectively to knowledge and language

resources. Lexical semantic relations are exactly such a set of atoms. Sinica BOW implements this idea by encoding the lexical semantic relations between English-Chinese translation equivalent pairs. In addition to more precisely describing the relationship between two translation equivalents, this also allows better cross-lingual inferences. Explicitly allowing lexical semantic relations to be coded cross-lingually also will facilitate the transferring to a structured set of tree relations from one language to the other.

4. Domain Taxonomy and Domain Ontology

In adopting an upper ontology approach, we implicitly accept the widely received idea about lower level and domain ontologies. There are two crucial assumptions here. First, that the lower level and more detailed ontololgies contain far too many conceptual nodes to be exhaustively listed. Second, conceptual terms may be defined differently in different knowledge domains. In other words, only the shared upper nodes can be considered constant. To account for the potential variations introduced by lower ontologies as well as domain ontologies, Sinica BOW includes a domain taxonomy as well as some domain ontologies.

Our basic approach is the proposal to tag all WordNet synsets with domain information (Huang et. al 2004a). The basic motivation is that, in order to ensure cross-domain knowledge sharing, it is important to identify the lemmas that can be used in multiple domains. This is antithetic to the traditional approach of trying to identify domain specific terms. In other words, the domain classification is most useful when the domain cannot be easily determined or when the resources involved containing content belonging to more than one domain. Hence the goal is to assign domain tags to as many general lexical items listed in WordNet as possible. More than 30% of the WordNet synsets are now assigned with a domain tag. They allow versatile re-organizations of domain lexica, as well as identification of possible domain information in a general purpose archives (such as news articles or the web in general.)

One of the most immediate and perhaps most powerful application of Sinica BOW is perhaps the construction of domain specific ontologies. This will be a crucial step towards providing a feasible infrastructure to implement web-wide specific ontologies, as required by the vision of Semantic Web. It is also a critical test to see if the upper ontology approach is really applicable to a wide range and diversity of knowledge domains. And lastly, for Sinica BOW, it provides a test ground for us to show that the combination of bilingual wordnet and ontology does provide a better environment for knowledge processing.

Two first attempts have been carried out. The first is a small fish domain ontology projected from the FishBase terms. This is mapped using Sinica BOW. Part of the ontology is shown in Figure 4. We would like to explore the possibility of using this domain ontology for non-expert to extract expert knowledge from the FishBase in the future.

The second attempt, reported in Huang et al. (2004b), involves the Shakespearean-garden approach to domain ontology. In this approach, we collect domain lexicon from a target collection of texts (Tang poems in this case), and map them to the SUMO ontology. This approach allows us to examine the knowledge and/or experience of a specific domain as reflect in that collection of texts. This could be personal, historical, regional etc. This approach allows us to make generalizations based on the full knowledge structure, not just one lexical incident. For instance, we were able to confirm the Tang civilization's fascination with flying by looking at the dominance of animal references in the texts.

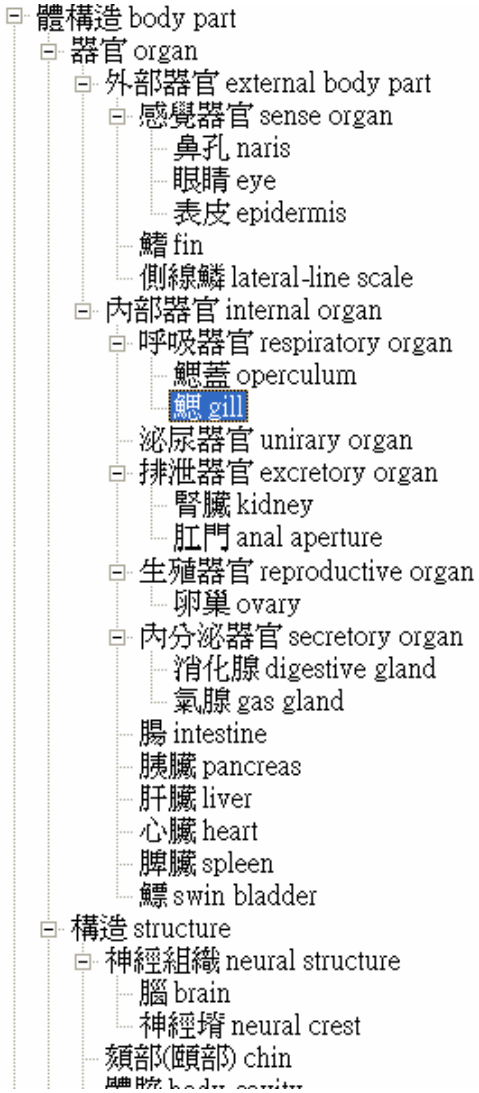


Figure 4: A sample domain ontology: Fish

5. Conclusion

Integrating and interpreting information from multiple and varying sources will be the main challenge for information processing for the current generation. Taking lexicon as the bridging knowledgebase and ontology as the overall knowledge structure seems to be a logical choice. Integrating the two resources with multilingual capacity will add to the versatility and open new possibilities.

Online Resources

Sinica BOW: http://BOW.sinica.edu.tw/
SUMO: http://ontology.teknowledge.com/
WordNet: http://www.cogsci.princeton.edu/~wn/

Referneces

Fellbaum, Christine. Ed. (1998). WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.

Huang, Chu-Ren Feng-ju Lo, Ru-Yng Chang, and Sueming Chang. (2004). Reconstructing the Ontology of the Tang Dinasty - a pilot study of the Shakespearan garden approach. To be presented at OntoLex 2004 - Ontologies and Lexical Resources in Distributed Environments. Lisbon. May 29, 2004.

Huang, Chu-Ren, Li, Xiang-Bing, Hong, Jia-Fei. (2004). Domain Lexico-Taxonomy:An Approach Towards Multi-domain Language Processing. Presented at the Asian Symposium on Natural Language Processing to Overcome Language Barriers. March 25-26, 2004. Hainan Island.

Huang, Chu-Ren, Elanna I.J. Tseng, Dylan B.S. Tsai, & Brian Murphy. (2003). Cross-lingual Portability of Semantic Relations: Bootstrapping Chinese WordNet with English WordNet Relations. Language and Linguistics. 4(3), 509--532.

Huang, Chu-Ren. Elanna I.J. Tseng & Dylan B.S. Tsai. (2002). Translating Lexical Semantic Relations: The first step towards multilingual Wordnets. In Proceedings of the COLING2002 workshop: SemaNet: Building and Using Semantic Networks. Taipei, Taiwan.

Niles, I. & Pease, A. (2003). Linking Lexicons and Ontologies: Mapping WordNet to the Suggested Upper Merged Ontology. In Proceedings of the IEEE International Conference on Information and Knowledge Engineering. (IKE 2003), Las Vegas, Nevada.

Niles, I., & Pease, A., (2001). Toward a Standard Upper Ontology. In Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001). Ogunquit, Maine.

Searching and Browsing Collections of Finnish Museums on the Semantic Web

**Eero Hyvönen, Samppa Saarela, Kim Viljanen, Eetu Mäkelä,
Arttu Valo, Mirva Salminen, Suvi Kettula, Miikka Junnila**

Helsinki Institute for Information Technology (HIIT), University of Helsinki
P.O. Box 26, 00014 UNIV. OF HELSINKI, FINLAND
{firstname.lastname}@cs.helsinki.fi
http://www.cs.helsinki.fi/group/seco/

Abstract

This paper presents the semantic portal MUSEUMFINLAND for publishing museum collections on the Semantic Web. It is shown how museums with their semantically rich and interrelated collection content can create a large, consolidated semantic collection portal together on the web. By semantic web techniques, it is possible to make collections semantically interoperable and provide the museum visitors with intelligent content-based search and browsing services to the global collection base.

1. Why MUSEUMFINLAND?

"MUSEUMFINLAND — Finnish Museums on the Semantic Web"[1] is a semantic portal that contains metadata from the collection databases of the National Museum[2], Espoo City Museum[3], and Lahti City Museum[4], and more content is being ported into the system. The application is intended for the public in the large to use.

The goals for developing the system were the following:

Global view to distributed collections It is possible to use the heterogeneous distributed collections of the museums participating in the system as if the collections were in a single uniform repository.

Content-based information retrieval The system supports intelligent information retrieval based on ontological concepts, not on simple keyword string matching as is customary with current search engines.

Semantically linked contents A most interesting aspect of the collection items to the end-user are the implicit semantic relations that relate collection data with their context and to each other. In MUSEUMFINLAND, such associations are exposed to the end-user by defining them in terms of logical predicate rules that make use of the underlying ontologies and collection metadata.

Easy local content publication The portal should provide the museums with a cost-effective publication channel.

In the following, these goals and solutions developed in our work are described. After this, main results of the work are summarized, lessons learned discussed, and directions for further research outlined.

2. Global View to Collections

Museum databases are usually situated at different locations and use different database systems and schemas. This creates a severe obstacle to information retrieval. To address the problem, the web can be used for creating a single interface and access point through which a search query can be sent to distributed local databases and the results combined into a global hit list. This "multi-search" approach is widely applied and there are many cultural collection systems on the web based on it, such as the portals Australian Museums Online[5] and Artefacts Canada[6].

A problem of multi-search is that by processing the query independently at each *local database*, the *global* dependencies, associations between objects in different collections are difficult to found. Since exposing semantic associations between collections items is one of our main goals, MUSEUMFINLAND cannot be based on the multi-search paradigm. Instead, the local collections are first consolidated into a global repository, and the queries are answered based on it (cf. figure 1). Mutually shared conceptual models, ontologies, are used for enriching the content and for making the collections interoperable. To show the associations to the end-user, the collection items are represented as web pages interlinked with each other through the

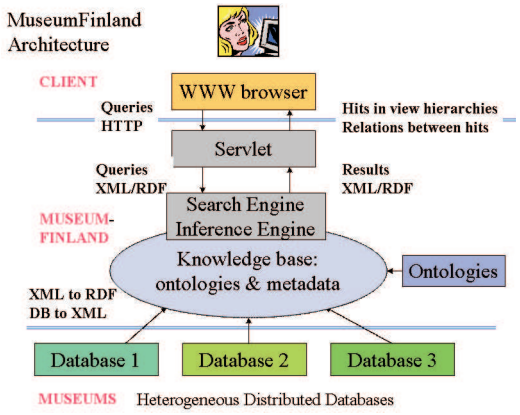


Figure 1: Information retrieval in MUSEUMFINLAND. Local database contents are first merged and the query is evaluated with respect to the global interrelated data.

[1] http://museosuomi.cs.helsinki.fi
[2] http://www.nba.fi
[3] http://www.espoo.fi /museo/
[4] http://www.lahti.fi /museot/
[5] http://amol.org.au/collection/collections_index.asp
[6] http://www.chin.gc.ca

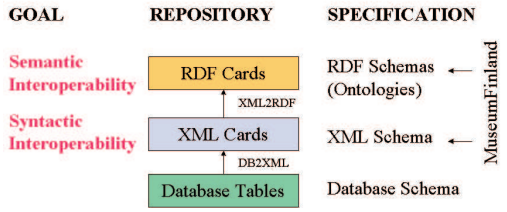


Figure 2: Data transformations in MUSEUMFINLAND.

View category	View	Ontology
Object	Artifact	Artifacts
	Material	Materials
Creation	Creator	Actors
	Location of creation	Locations
	Time of creation	Times
Usage	User	Actors
	Location of usage	Locations
	Situation of usage	Situations
Museum	Collection	Collections

Table 1: View facets in the MUSEUMFINLAND portal.

semantic associations. The MUSEUMFINLAND home page is the single entry point through which the end-user enters the global semantic WWW space.

The challenge in consolidating the collections is how to make them interoperable in syntax and, especially, in semantics. In our solution (Hyvönen et al., 2004c), the museum first transforms its collection data into XML (cf. figure 2). Each collection object is represented as an *XML card* that describes the object in terms of 22 properties whose values are strings and numbers read from the underlying database. The XML Schema used is agreed upon by the participating museums and guarantees syntactic interoperability of the collections.

Next, each XML card is transformed into an *RDF card* with similar RDF properties, but where up to 16 string values are transformed into the URIs of the corresponding classes and individuals in a set of underlying RDF(S) ontologies. This transformation is based on a set of *term cards* that map terms with ontology resources. MUSEUMFINLAND provides the museums with the ontologies and a set of term cards. The museums can adapt their terminological conventions to the portal by creating new term cards of their own. Two special tools have been developed for creating terminologies (Terminator) and RDF annotations (Annomobile) semi-automatically. Protégé-2000[7] is used for the manual editing part.

3. Multi-Facet Search Based on Ontologies

The content-based search engine of MUSEUMFINLAND is a server called Ontogator. Ontogator is based on the multi-facet view-based search paradigm developed within the information retrieval research community (Pollitt, 1998; Hearst et al., 2002; Hyvönen et al., 2004b). Multi-facet search is based on a set of *categories* that are organized into a set hierarchical, orthogonal taxonomies called subject *facets* or *views*.

A search query in multi-facet search is formulated by selecting categories of interest from the different facets. For example, by selecting the category "Furniture" from the Artifact facet, and "Eero Saarinen" from the Creator facet, the user can express the query for retrieving all kinds of furniture, such as chairs, tables, etc., created by Eero Saarinen. Intuitively, the query is a conjunctive constraint over the facets with disjunctive constraints over the sub-categories in each facet.

More formally, if the categories selected are $C_1, ..., C_n$ and the subcategories of $C_i, i = 1...n$, including C_i itself are $S_{i,1}, S_{i,2}..., S_{i,k}$, respectively, then this selection corresponds to the following boolean AND-OR-constraint:

$$(S_{1,1} \vee ... \vee S_{1,k}) \wedge (S_{2,1} \vee ... \vee S_{2,l}) \wedge ... \wedge (S_{n,1} \vee ... \vee S_{n,m}) \quad (1)$$

Facets can be used for helping the user in information retrieval in many ways. Firstly, the facet hierarchies give the user an overview of what kind of information there is in the repository. Secondly, the hierarchies can guide the user in formulating the query in terms of appropriate keywords. Thirdly, the hierarchies can be used to disambiguate homonymous query terms. Fourthly, the facets can be used as a navigational aid when browsing the database content (Hearst et al., 2002). Fourthly, the number of hits in every category that can be selected next can be computed *beforehand* and be shown to the user (Pollitt, 1998). In this way, the user can be hindered from making a selection leading to an empty result set—a recurring problem in IR systems—and is guided toward selections that are likely to constrain (or relax) the search appropriately.

Table 1 depicts the 9 views used in MUSEUMFINLAND and their underlying 7 ontologies. The Artifacts ontology is a taxonomy of the tangible collection objects such as pottery, cloths, weapons, etc. All exhibits in the system belong to some class in this ontology. The Materials ontology is a taxonomy of the artifact materials, such as steel, silk, tree, etc. The Actors ontology defines classes of agents, such as persons, companies etc., and individuals as instances of these classes. The Events ontology include intangible happenings, situations, events, and processes that take place in the society, such as farming, feasts, sports, war, etc. Locations is an ontology representing areas and places on the Earth and in Finland in particular. The Times ontology is a taxonomy of various predefined historical periods, and the Collections ontology classifies the museums and collections in the portal. The Artifacts, Materials, and Events ontologies are subsets of a larger cultural ontology called MAO (6768 classes) that we created based on the Finnish cultural thesaurus MASA (Leskinen, 1997).

Figure 3 shows the search interface of MUSEUMFINLAND. The nine facet hierarchies of table 1 are shown (in Finnish) on the left. For each facet hierarchy, the next level of sub-categories is shown as links. A query is formulated by selecting a sub-category by clicking on its name. When the user selects a category c in a facet f, the system constrains the search by leaving in the result set only

[7] http://protege.stanford.edu/

Figure 3: The search interface of MUSEUMFINLAND.

such objects that are annotated in facet f with some sub-category of c. The figure depicts the situation after selecting the sub-category Tools (”työvälineet”) from the Artifact facet (”Esinetyyppi”). The result set is shown on the right grouped by the sub-categories of Tools, such as Textile making tools (”tekstiilityövälineet”) and Tools of folk medicine (”kansanlääkinnän työvälineet”). Hits in different categories are separated by horizontal bars and can be scrolled independently in each category. In this case, all categories do not fit in the screen shot.

When answering the query, the result set for each direct sub-category in the facets seen on the screen is recomputed, and a number (n) is shown to the user after the category name. It tells that if the sub-category is selected next, then there will be n hits in the result set. For example, the number 643 in the Collection facet on the bottom (”Kokoelma”) tells that there are 643 tools in the collections of the National Museum (”Kansallismuseon kokoelmat”). A selection leading to an empty result set (n=0) is removed from its facet (or alternatively disabled and shown in gray color, depending on the user’s preference). In this way, the user can be hindered from making a selection leading to an empty result set, and is guided toward selections that are likely to constrain the search appropriately. The query can be relaxed by making a new selection on a higher level of the facets or by dismissing the facet totally from the query.

In above, the category selection was made among the direct sub-categories listed in the facets. An alternative way is to click on the link Whole facet (”koko luokittelu”) on a facet. The system then shows all possible selections in the facet with hit counts. In this way, the user can easily formulate the query using the right categories exposed to her as links, and can get easily overviews of the database contents along different classifications in different situations.

User studies (Lee et al., 2003; English et al., 2003) indicate that if the user does not precisely know what (s)he is looking for, then the multi-facet search method with its “browsing the shelves” sensation is clearly preferred over keyword search (or using a single facet search). Otherwise, a direct Google-like keyword search interface is preferred. To support word-based search, too, an additional search engine was implemented in MUSEUMFINLAND (upper left corner in figure 3). This engine is used for two purposes at the same time: for searching categories to be used in multi-facet search and for searching collection objects with matching metadata values in the conventional way. (Hyvönen et al., 2004a)

4. Semantic Linkage

One of the main goals of the MUSEUMFINLAND portal is to reveal the rich semantic linkage connecting the collection objects with each other. The links can be *explicit* or *implicit*. Explicit links correspond to the RDF statements (triples) in the underlying knowledge base and are based on the collection domain ontologies (classes and their properties) and the actual collection data (instance data). For

example, an instance of a painting may have the RDF property dc:creator linking the art work to an individual artist. Implicit links can be defined in terms of explicit ones but are not present in the RDF graph. For example, if there are explicit links linking children with their mothers and fathers, then implicit links such as "grandfather" or "cousin" can be defined.

In MUSEUMFINLAND, implicit links are defined declaratively in terms of logic by using Prolog predicates. Each predicate defines a semantic association and gives it an explanatory label, such as "cousin of". By applying such a predicate to a collection item resource, implicitly related other resources with respect to the semantic association can be found. On the HTML level in the user interface, the label of the association is used as the name for the link and the found resource as the target. For example, if the family relations of artists are known in the ontology, then such a predicate could infer links to other pages depicting paintings whose creator is of the same family.

For example, figure 4 depicts an collection object page found by multi-facet search. The object is a distaff ("rukinlapa" in Finnish) used in a spinning wheel. On the left, a photo of the object is shown. The metadata of the object is shown in the middle on top. All facet categories of the object are listed in the middle bottom as hierarchical link paths. A new search can be started by selecting any link from there. On the right, the system displays links to other recommended collection items. i.e., *semantic recommendations*.

The recommendation links provide a semantic browsing facility to the end-user. For example, in figure 4 there are links to objects used at the same location (categorized according to the name of the common location), to objects related to similar events (e.g., objects used in spinning, and decorative objects, because distaffs are usually beautifully decorated), to objects manufactured at the same time, and so on. Since a decoratively carved distaff used to be a typical wedding gift in Finland, it is also possible to recommend links to other objects used as wedding gifts, such as wedding rings. In MUSEUMFINLAND, such associations can be exposed to the end-user as link groups whose titles and link names explain to the user the reason for the recommendation. The possibilities for creating such associations are intriguing. Of course, only links that can be inferred based on the metadata and ontologies available can be created.

Recommendations are defined in terms of flexible logical predicate rules using the methods described in (Hyvönen et al., 2003). The semantic recommendation system of MUSEUMFINLAND is implemented as a logic server called Ontodella. This system is based on the HTTP server version of SWI-Prolog[8] (Wielemaker et al., 2003).

There is also a prototype implementation of MUSEUMFINLAND that can be used with WAP 2.0 compatible mobile telephones. The current prototype recreates all functionality of the web interface in a layout more suitable to the limited screen space of mobile devices. When the user makes a selection for the multi-facet search, impossible category choices leading to empty results can be pruned out. This is a very useful feature for devices that have a small screen to display choices.

[8] http://www.swi-prolog.org

5. Discussion

MUSEUMFINLAND is an application of the idea of semantic portals to solving interoperability problems of museum collection databases when publishing their content on the Semantic Web. The power of MUSEUMFINLAND comes from the use of ontologies:

Exact definitions By using ontologies, the museums can define the concepts used in cataloging in a precise, machine understandable way.

Terminological interoperability The terms used in different institutions can be made mutually interoperable by mapping them onto common shared ontologies. The ontologies are not used as a norm for telling the museums what terms to use, but rather to make it possible to tolerate terminological variance as far as the terminology mapping from the local term conventions to the global ontology is provided.

Ontology sharing Ontologies provide means for making exact references to the external world. For example, in MUSEUMFINLAND, the location ontology (villages, cities, countries, etc.) and the actor ontology (persons, companies, etc.) is shared by the museums in order to make the right and interoperable references. For example, two persons who happen to have the same name should be disambiguated by different URIs, and a person whose name can be written in many ways, should be identified by a single URI to which the alternative terms refer.

Automatic content enrichment Ontological class definitions, rules, and consolidated metadata enrich collection data semantically.

Intelligent services Ontologies can be used as a basis for intelligent services to the end-user. In MUSEUMFINLAND, the view-based search engine is based on the underlying ontological structures and the semantic link recommendation systems reveals to the end-user the underlying semantical context of the collection items and their mutual relations.

The novelty of the content-based search engine with respect to other view-based systems (Pollitt, 1998; Hearst et al., 2002) is based on its capability of using RDF(S) ontologies as the basis of search. The main benefits obtained are: 1) Ontological logical inference can be employed in projecting the views from the ontologies (e.g., the location meronymy and various concept hyponymies). 2) The implicit complicated relations between view categories and the underlying data resources to be searched for can be specified flexibly in terms of logical predicates. Ontogator combines virtues of the view- and ontology-based search paradigms (Hyvönen et al., 2004b).

Figure 4: Collection item metadata with semantic recommendations.

The idea of linking collection items with semantic associations was inspired by Topic Maps (Pepper, 2000). However, in our case the links are not given by a map but are determined by logical inference using the underlying RDFS ontology and RDF metadata. Another application of this idea to generating semantically linked static HTML sites from RDF(S) repositories is presented in (Hyvönen et al., 2003). Logic and dynamic link creation on the semantic web has been discussed, e.g., in the work on Open Hypermedia (Goble et al., 2001; Dolong et al., 2003). In the HyperMuseum (Stuer et al., 2001), collection items are also semantically linked with each other. Here linking is based on shared words in the metadata and their linguistic relations, such as synonymy and antonymy. In contrast, our system is not based on words but on ontological references in the underlying RDF(S) knowledge base and the links can be defined freely in terms of logical rules. The idea of annotating cultural artifacts in terms of multiple ontologies has been explored, e.g., in (Hollink et al., 2003).

5.1. Lessons Learned

The main problem encountered in the content work was that the original museum collection data in the databases was not systematically annotated. Various conventions are in use in different museum systems and museums. Much of the metadata is not based on a keywords but is free text. The Terminator and Annomobile tools developed for the XML to RDF transformation (Hyvönen et al., 2004c) are only semi-automatic, and a human editor is often needed to make the right annotations. Due to homonymy, not even thesaurus keywords can always be to mapped unambiguously to RDF concepts by the machine. However, the homonymy problem turned out to be less severe than expected, because disambiguation could be based on the facet/ontology to which the database field was related.

The view-based search method can be implemented quite efficiently. The current system scales up to the order of 10,000 RDF cards and 10,000 ontological concepts on an ordinary PC server. From the user's perspective, the idea of multi-facet search seems useful and a natural next step a head from the single facet systems on the web today, such as Yahoo[9] and Open Directrory Project[10]. Using Prolog and RDF together for projecting the facets and for creating the semantic recommendation links was powerful and flexible. It is possible compute and store the results of some inferences before running system in order to speed up reasoning. In our case, the mappings between facet categories and RDF resources are determined in Prolog beforehand and are compiled into an RDF tree that can be used more efficiently by the view-based search engine. The semantic recommendations are currently determined dynamically.

MUSEUMFINLAND user interface was first implemented as a Java servlet using XSLT transformations. The system was then re-designed and re-implemented as a Cocoon-based server[11] that queries the Ontogator search engine server and Ontodella logic server with XML/RDF messages. It is possible to do this over HTTP. With Cocoon, the implementation could be made in a couple of months and can be modified easily. For example, the mobile telephone interface was created by modifying the PC version.

[9] www.yahoo.com
[10] www.dmox.org
[11] http://cocoon.apache.org

5.2. Further Work

We are investigating how new kind of RDF material, conforming to different ontologies, such as art collections using the Iconclass[12] system and educational videos based on the IEEE Learning Objects Metadata standard, can be merged in the portal. More work is needed in developing a set of recommendation predicates that would be of most interest to the users.

Ways of collaboration between museum content providers and portal maintenance people need to be developed in order to develop MUSEUMFINLAND from an application into a continuous publication *process*. For example, protocols for adding, modifying, and retracting RDF cards and ontology resources according to the wishes of the museums need to be developed.

The pilot version of MUSEUMFINLAND portal was opened on the public web in March 2004 at `http://museosuomi.cs.helsinki.fi`.

Acknowledgments

Our work is funded mainly by the National Technology Agency Tekes, Nokia Corp., TietoEnator Corp., the Espoo City Museum, the Foundation of the Helsinki University Museum, the National Board of Antiquities, and the Antikvaria Group consisting of some 20 Finnish museums.

6. References

Dolong, P., N. Henze, and W. Neijdl, 2003. Logic-based open hypermedia for the semantic web. In *Proceedings of the Int. Workshop on Hypermedia and the Semantic Web, Hypertext 2003 Conference, Nottinghan, UK*.

English, J., M. Hearst, R. Sinha, K. Swearingen, and K.-P. Lee, 2003. Flexible search and navigation using faceted metadata. Technical report, University of Berkeley, School of Information Management and Systems. Submitted for publication.

Goble, C., S. Bechhofer, L. Carr, D. De Roure, and W. Hall, 2001. Conceptual open hypermedia = the semantic web? In *Proceedings of the WWW2001, Semantic Web Workshop, Hongkong*.

Hearst, M., A. Elliott, J. English, R. Sinha, K. Swearingen, and K.-P. Lee, 2002. Finding the flow in web site search. *CACM*, 45(9):42–49.

Hollink, L., A. Th. Schereiber, J. Wielemaker, and B.J. Wielinga, 2003. Semantic annotations of image collections. In *Proceedings KCAP'03, Florida*.

Hyvönen, E., M. Junnila, S. Kettula, E. Mäkelä, S. Saarela, M. Salminen, A. Syreeni, A. Valo, and K. Viljanen, 2004a. Finnish Museums on the Semantic Web. User's perspective on museumfinland. In *Proceedings of Museums and the Web 2004 (MW2004), Arlington, Virginia, USA*. Http://www.archimuse.com/mw2004/papers/hyvonen/hyvonen.html.

Hyvönen, E., S. Saarela, and K. Viljanen, 2004b. Application of ontology based techniques to view-based semantic search and browsing. In *Proceedings of the First European Semantic Web Symposium, May 10-12, Heraklion, Greece, (fortcoming)*. Springer–Verlag, Berlin.

Hyvönen, E., M. Salminen, S. Kettula, and M. Junnila, 2004c. A content creation process for the Semantic Web. Proceeding of OntoLex 2004: Ontologies and Lexical Resources in Distributed Environments, May 29, Lisbon, Portugal (forthcoming).

Hyvönen, E., A. Valo, K. Viljanen, and M. Holi, 2003. Publishing semantic web content as semantically linked HTML pages. In *Proceedings of XML Finland 2003, Kuopio, Finland*. Http://www.cs.helsinki.fi/u/eahyvone/publications/xmlfinland2003/swehg_article_xmlfi2003.pdf.

Lee, K.-P., K. Swearingen, K. Li, and M. Hearst, 2003. Faceted metadata for image search and browsing. In *Proceedings of CHI 2003, April 5-10, Fort Lauderdale, USA*. Association for Computing Machinery (ACM), USA.

Leskinen, R. L. (ed.), 1997. *Museoalan asiasanasto*. Museovirasto, Helsinki, Finland.

Pepper, Steve, 2000. The TAO of Topic Maps. In *Proceedings of XML Europe 2000, Paris, France*. Http://www.ontopia.net/topicmaps/materials/rdf.html.

Pollitt, A. S., 1998. The key role of classification and indexing in view-based searching. Technical report, University of Huddersfield, UK. Http://www.ifla.org/IV/ifla63/63polst.pdf.

Stuer, Peter, Robert Meersman, and Steven De Bruyne, 2001. The HyperMuseum theme generator system: Ontology-based internet support for active use of digital museum data for teaching and presentations. In D. Bearman and J. Trant (eds.), *Museums and the Web 2001: Selected Papers*. Archieves & Museum Informatics. Http://www.archimuse.com/mw2001/papers/stuer/stuer.html.

Wielemaker, J., A. Th. Schereiber, and B. J. Wielinga, 2003. Prolog-based infrastructure for RDF: performance and scalability. In *Proceedings ISWC'03, Florida*. Springer–Verlag, Berlin.

[12] http://www.iconclass.nl

Context-related Derivation of Word Senses

Manuela Kunze and Dietmar Rösner

Otto-von-Guericke-Universität Magdeburg
Institut für Wissens- und Sprachverarbeitung
P.O.box 4120, 39016 Magdeburg
makunze|roesner@iws.cs.uni-magdeburg.de

Abstract

Real applications of natural language document processing are very often confronted with domain specific lexical gaps during the analysis of documents of a new domain. This paper describes an approach for the derivation of domain specific concepts for the extension of an existing ontology. As resources, we need an initial ontology and a partially processed corpus of a domain. We exploit the specific characteristics of the sublanguage in the corpus. Our approach is based on syntactic structures (noun phrases) and compound analysis to extract information required for the extension of GermaNet's lexical resources.

1. Introduction

One of the bottlenecks in real applications of natural language document processing is the coverage of domain-specific lexical resources. In experiments with the document suite XDOC[1], we currently are processing documents about casting technology, company profiles from web pages, and autopsy protocols. Many of the tools have an extensive need for linguistic resources. Therefore we are interested in ways to exploit existing resources with a minimum of extra work. The resources of GermaNet promise to be helpful for different tasks in the workbench.

In this paper, we will outline how the resources of GermaNet can be extended. Our methods exploit the specific characteristics of the documents in the corpus. We combine different approaches to extract new concepts from the corpus. The idea behind our approach is to generalise from structures with known GermaNet entries to structures without GermaNet entries.

This paper presents only experiments with GermaNet on German texts, but the approach can also be applied on WordNet when processing domain specific English texts.

The paper is organized as follows: The next section briefly outlines the test corpus and the integration of GermaNet in XDOC. Section 3 describes the methods for the extraction of new concepts and the results. We conclude the paper with a discussion section.

2. Document Processing with XDOC

2.1. Characteristics of the Corpus

In the following description of the approach, a corpus of forensic autopsy protocols is used, because these documents are especially amenable to processing with techniques from computational linguistics and knowledge representation.

Autopsy protocols consist of the following major document parts: *findings*, *histological findings*, *background*, *discussion*, *conclusions*, etc. Our analyses focus on the sections of *findings*, *background* and *discussion*. In the *findings* section, a high ratio of nouns and adjectives is encountered and the sentences, which can also be verbless, are mostly short. This section describes the medical findings in a common language. Here we find no domain specific (medical) terms. The *background* and *discussion* sections contain a standard distribution of all word classes and regular syntactic structures. The *background* section describes, for example, the details of a traffic accident, while the section *discussion* contains a combination of the results of the *finding* section and the facts reported in the *background* section.

2.2. Integration of GermaNet

The document suite XDOC contains methods for linguistic processing of documents in German. The focus of the work has been to offer end users a collection of highly interoperable and flexible tools for their experiments with document collections.

XDOC consists of different modules, for example, the syntactic module and the semantic module (for a more detailed description see (Rösner and Kunze, 2002b)):

For the semantic analyses of a domain using XDOC, knowledge about the domain – ideally a domain specific ontology – is needed. One possible resource for the processing of autopsy protocols could be medical thesauri like UMLS (Unified Medical Language System).[2] Many of these resources work with medical terminology, but in the corpus of forensic autopsy protocols only everyday terms are used. Thus a resource that contains everyday terms and concepts (and their relations) from the medical domain is required for the analysis. GermaNet (see (Hamp and Feldweg, 1997), (Kunze, 2001)) is intended as a model of the German base vocabulary.

However, specific terms in some particular domains, like the medical domain, are covered only partially in GermaNet.

For the semantic analysis in XDOC, the *synonymy* and the *hypernymy* relations of GermaNet are used. We found a good coverage of GermaNet's resources for terms in the corpus: section *findings* with 31 %, section *background* with 44 %, and section *discussion* with 42 % coverage (see also (Kunze and Rösner, 2003)). The reason for the poor *findings*'s result is the high frequent occurrence of medical

[1] XDOC stands for *X*ML based *doc*ument processing.

[2] http://www.nlm.nih.gov/research/umls/umlsmain.html

concepts denoted by noun compounds like *Nierengewebe (kidney tissue)* or *Halswirbelsäule (cervical spine)* that are not covered by GermaNet, whereas the individual compound words like kidney and spine have lexical entries in GermaNet.

In the next section, we will describe how new entries can be derived from entries that exist in GermaNet. We start with a corpus of autopsy protocols parsed syntactically by XDOC and with GermaNet as an initial ontology.

3. Methods for the Deduction of Word Senses

In (Rösner and Kunze, 2002a), we outlined some ideas for the exploitation of sublanguage characteristics of a corpus for lexicon creation. In this paper, we will further elaborate these ideas. This section presents how the syntactic structures of the corpus sublanguage can be useful for the extraction of new GermaNet entries.

3.1. Fundamental Idea of the Approach

In the *findings* section of the documents, high-frequency complex noun phrases can be exploited for the extension of the GermaNet resources.

The grammar fragment used in XDOC for this corpus covers the following complex noun phrases (In all cases, the first NP is a simple noun phrase.):

- NP $\text{NP}_{genitive}$,
- NP $\text{NP}_{genitive}$ *PP, and
- NP *PP.

Our experiments are based on the interpretation of complex noun phrases that are described by the syntactic structure NP $\rightarrow$ NP $\text{NP}_{genitive}$ (i.e. a simple NP modified by a genitive attribute).

In the case of a complex noun phrase, several possibilities for a semantic interpretation of this syntactic structure exist, for example, `part-of` relations in *'dermis of the hand'* or `patient-of` relation in *'the production of cars'*.

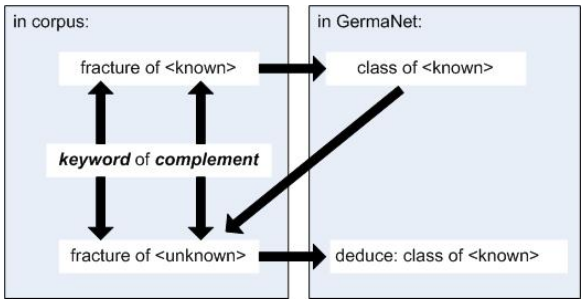


Figure 1: A Sketch of the Idea.

The idea behind the approach is based on following assumptions. A structure of the form *KEYWORD of COMPLEMENT* describes the same relation for every possible candidate of the complement, e.g., `part-of`. Further on, an assumption is that the complement candidates of a keyword have the same semantic category. The information of complement candidates available in GermaNet is used to deduce information about the semantic category of candidates that are unknown in GermaNet (see also Fig. 1).

Table 1: Some Complements of a Structure Beginning with Keyword *'Bruch'* (Fracture).

complement	occurrences	top level of GermaNet
Rippe	254	nomen.koerper
Brustbein	65	nomen.koerper
Wirbelsäule	58	nomen.koerper
Schädeldach	43	–
Oberschenkelknochen	37	–
Schädelbasis	34	–
Schlüsselbein	33	–
Schambein	30	nomen.koerper
Brustwirbelsäule	28	–
Halswirbelsäule	26	–
Schulterblatt	23	nomen.koerper

3.2. Exploiting Syntactic Structures of the Corpus

In the corpus (of 600 autopsy protocols and more than 1.5 million word forms), structures in the form of

`NP` $\rightarrow$ `NP` $\text{NP}_{genitive}$ are often encountered. For example, the phrase *'Schleimhaut des Magens'* (*mucosa of the stomach*) occurs 317 times in the corpus. The more generalised phrase *'mucosa of XXX'* occurs 836 times in the corpus. Another generalised example is the phrase *'fracture of XXX'* that occurs 749 times in 93 different forms. One example form is the class of NPs with keyword Bruch (fracture) and modified by a complement (the second noun phrase in the structure), e.g., *'Wirbelsäule' (spine)* in the phrase *'Bruch der Wirbelsäule'* (occurs 58 times) or *'Wadenbein' (fibula)* in the phrase *'Bruch des Wadenbeines'* (occurs 11 times). Other complements for the keyword *'fracture'* found in the corpus are: *'Elle' (ullna)*, *'Oberarmknochen' (humerus)*, *'Schädelgrund' (base of the skull)*, *'Schienbein' (shinbone)*, *'Unterkiefer' (lower jaw)*, *'Unterarmknochen' (radial bone)* etc.

At first, structures with high occurrence frequencies in the corpus are selected. For this task, the *findings* sections of the documents are parsed with the syntactic parser of XDOC. A domain specific grammar with ca. 40 rules is used. In the results of 18008 parsed sentences, 2808 complex noun phrases (`NP` $\rightarrow$ `NP` $\text{NP}_{genitive}$) with 1069 different keywords are encountered.

The most frequent keywords in such structures are: *'Abgang' (outlet)*, *'Bauchteil' (abdominal part)*, *'Brustteil' (chest part)*, *'Blutreichtum' (hyperemia)*, *'Fäulnis' (sepsis)*, *'Haut' (dermis)*, *'Schleimhaut' (mucosa)*, *'Gegend' (region)*, *'Schnittflächen' (cut surfaces)*, *'Unterblutung' (hematoma)*, and *'Bruch' (fracture)*.

The next step is to use regular expressions to get all occurrences of a particular combination of a keyword and a complement, because not all occurrences from the corpus can be obtained with the chart parser. The reason for this is that there are gaps in the grammar (when parsing the section *background* and *discussion*) and gaps in the morphological lexicon.

The most frequent keywords in regular expressions are used to get all phrases that begin with the keyword. The length of these phrases (text window size) is restricted to be 3 tokens (or 4 tokens, when adjectives in the complement noun phrase) are allowed.

For each structure, the GermaNet interface is used to check if information about the keyword of the complement NP is available. For the example (keyword: *fracture*), GermaNet contains 31 complement elements of the 93 complement elements found in our corpus. Most complement words of a keyword found in GermaNet have the same top level category, only a small number of words have more than one reading. For the example, following top level categories (given with its percentage related to all senses) are encountered: <nomen.Koerper>: 75 %, <nomen.Artefakt>: 16,5 %, <nomen.Menge>: 5,5 %, and <nomen.Nahrung>: 3 %. All the words with more than one sense have at least one sense with the top level category <nomen.Koerper>.

Table 1 presents a small excerpt of the complement words[3] in the corpus for the keyword *fracture*. The main top level category for the complement words is <nomen.Koerper> (WordNet category: noun.body).

The first assumption is that all complement words of a keyword in a domain will belong to the same top level category in GermaNet. That means that those words of the example which are not contained in GermaNet, like *'Oberarmknochen' (humerus)*, *'Schädelbasis' (base of the skull)*, *'Schädeldach' (calvarium)*, *'Brustwirbelsäule' (thoracic spine)*, etc., can be assigned to the same top level category: <nomen.Koerper>. In the case of the example (keyword *fracture*), this heuristic yields the correct top level category for 93,44 % of all complements.

In the next step, subclasses of the GermaNet top level category will be used , so that a word can be annotated with additional information, e.g., hypernymy relation. For this task, GermaNet's hypernymy relation is exploited. The hypernym information for all complements is selected, which do exist in GermaNet. The hypernymy relation in GermaNet can contain more than one level of hypernyms for an entry.

At first, all senses with their hypernym information are selected. Each sense and its hypernyms describe a class path and each entry in this class path names a semantic class. The occurrences of the different semantic classes for all senses (class paths) are counted. For the different forms of the phrase *'Bruch der/des XXX'* (in English: fracture of XXX), 36 senses with altogether 63 different semantic classes are encountered. Table 2 presents a partial list of all semantic classes and its number of occurrences in all the senses for the complement elements covered by GermaNet. For example, the semantic class *'Knochen' (bone)* appears in 13 senses as a hypernym, the semantic class *'Computerprogramm' (software)* only in one sense.

At this point, we don't have a clear and unique result. The highly frequent hypernym entries in all senses found in GermaNet are the entries: *'Objekt'* (object), *'Hornsubstanz'* (akeratosis), *'Knochen'* (bone), etc. These results can be enhanced when we allow only senses that describe a concept with the top level assignment of <nomen.Koerper> (see table 3). The possible senses are reduced to 27 senses with altogether 22 different semantic classes.

[3]The complement words described in table 1 occurred in the corpus in a singular or plural form.

Table 2: Hypernym Information for Complement Entries.

hypernym	number of occurrences	percentage
<nomen.Tops>=> Objekt	22	13.75
<nomen.Koerper>=> Hornsubstanz	13	8.125
<nomen.Substanz>=> Stoff1, Substanz, Materie	13	8.125
<nomen.Koerper>=> Körpersubstanz	13	8.125
<nomen.Koerper>=> Knochen, Gebein	13	8.125
<nomen.Artefakt>=> Artefakt, Werk	7	4.375
<nomen.Tops>=> Ding, Sache, Gegenstand, Gebilde	7	4.375
<nomen.Menge>=> Masseinheit, Mass, Messeinheit, Messeinheit*o	2	1.25
...	...	...
<nomen.Koerper>=> Armknochen	2	1.25
<nomen.Artefakt>=> Computerprogramm, Programm	1	0.625
<nomen.Artefakt>=> ?akustisches Gerät	1	0.625

When the basic concepts (WordNet's 'unique beginner') of GermaNet, e.g. *Objekt* is ignored, and when the most specific hypernym of all high frequent hypernyms is selected, the following partial class path results:

```
<nomen.Koerper>=> Knochen, Gebein
       <nomen.Koerper>=> Hornsubstanz
           <nomen.Koerper>=> Körpersubstanz
               <nomen.Substanz>=> Stoff, Substanz,
                    Materie
                   <nomen.Tops>=> Objekt
```

For the selection of the most specific hypernym, every level in the class path is assigned with a weighting factor (The selection process can be described by the Eq. 1). The unique beginner starts with the factor *0* (in our example *Objekt*), the next higher level get the factor *1*, and so on.

$$c_i = \arg\max_{c_i} \frac{f_i n(c_i)}{N} \quad (1)$$

For each semantic class c_i, the quotient (occurrences of the semantic class n(c_i) divided by number of all semantic classes N) is multiplied by its weighting factor f_i (see also Fig. 2). In the result above, the semantic classes got following factor assignment: $f_{Objekt} - 0$, $f_{Stoff} - 1$, $f_{Koerpersubstanz} - 2$, $f_{Hornsubstanz} - 3$, $f_{Knochen} - 4$.

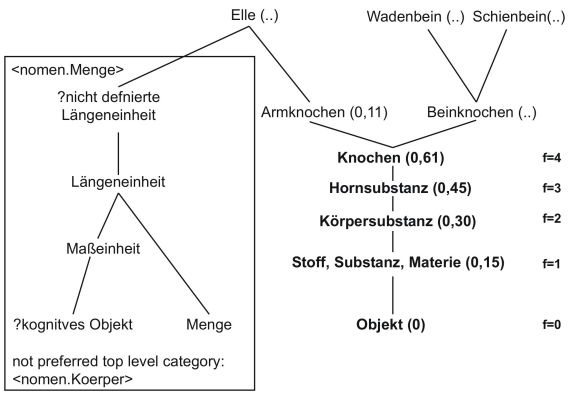


Figure 2: Weighting of Possible Semantic Classes.

The whole approach described above is sketched in the following (given a keyword *K* and a set of all complements C_S of *K*):

procedure find-entry (K, C_S):

Step 1: **for each** complement c ϵ C_S: get all (GermaNet) senses of c $\rightarrow$ H_S;

Step 2: ascertain the most frequent top level category in $H_S \rightarrow$ T;

Step 3: remove senses from H_S, which are not assigned with the preferred top level category T $\rightarrow$ $H_{Sprefer}$;

Step 4: **for each** sense s ϵ $H_{Sprefer}$: collect all semantic classes of the hypernym information of s $\rightarrow$ SC_S;

Step 5: **for each** semantic class sc ϵ SC_S: calculate

Step 5.1: occurrences of sc (n(c_i))/number of all sc (N)$\rightarrow$ sc_{ratio};

Step 5.2: sc_{ratio} times level in the hypernym tree (f_i) $\rightarrow$ sc_{weight};

Step 6: select sc with maximum of sc_{weight};

For ca. 80 % of the complement words of the keyword *fracture* this assignment is correct. Erroneous assignments result from misspelling of tokens (e.g. *Oberschenkelknorren* instead *Oberschenkelknochen*) or erroneous fragments in the results of the preprocessing steps (e.g., the treatment of German's truncations in phrases like *Bruch des Ober- und Unterarmes (fracture of upper arm and forearm)*). Another type of error occurring in the evaluation was the case when the second noun phrase can also be parsed as a complex noun phrase. For the example, only 2 forms are encountered: *Bruch der Anteile ... (fracture of parts of ...)* and *Bruch der Wandung ... (fracture of septum of ...)*. For a reliable evaluation of these results, it is necessary to consult the domain specific knowledge of a medical expert. In some cases, for a non-expert it is not clear if a derived sense is correct. For instance, the word *'Ellenbogengelenk' (elbow joint)* describes a (complex) system of bones, cartilages, connective tissues, etc.

3.3. Compound Analysis

An alternative way is to group words according to their components. In German and especially in the corpus, a lot of compounds are found, e.g., *'Armknochen' (arm bones)*, *'Oberarmknochen' (upper arm bone)*, and *'Unterarmknochen' (forearm bone)*. GermaNet contains the word *'Armknochen'*, but not the words *'Oberarmknochen'* and *'Unterarmknochen'*. For this case, a list of typical prefixes of the domain can be made of use. Prefixes in the domain are e.g., *'Unter-'*, *'Ober-'*, *'Innen-'*, *'Aussen-'*, quasi a pair list of antonyms. In this case, the hypernym information can be used directly for the new entry. For example, in GermaNet following entry of the word *Armknochen* is encountered:

```
1 sense of armknochen

Sense 1 <nomen.Koerper>Armknochen
       <nomen.Koerper>=> Knochen, Gebein
           <nomen.Koerper>=> Hornsubstanz
               <nomen.Koerper>=> Körpersubstanz
                   <nomen.Substanz>=> Stoff1, Substanz,
                        Materie
                       <nomen.Tops>=> Objekt
```

In the corpus, the complement words *'Unterarmknochen'* (3 times) and *'Oberarmknochen'* (19 times) for the same keyword are found. Both have no entry in GermaNet. The following information for the word *'Oberarmknochen'* (similar for the word *'Unterarmknochen'*) could be inserted:

Table 3: Enhanced Hypernym Information for Complement Entries.

hypernym	number of occurrences	percentage
<nomen.Tops>=> Objekt	14	16.47
<nomen.Koerper>=> Hornsubstanz	13	15.29
<nomen.Substanz>=> Stoff1, Substanz, Materie	13	15.29
<nomen.Koerper>=> Körpersubstanz	13	15.29
<nomen.Koerper>=> Knochen, Gebein	13	15.29
<nomen.Artefakt>=> Artefakt, Werk	–	–
<nomen.Tops>=> Ding, Sache, Gegenstand, Gebilde	–	–
<nomen.Menge>=> Masseinheit, Mass, Messeinheit, Messeinheit*o	–	–
...	...	...
<nomen.Koerper>=> Armknochen	2	2.35
<nomen.Artefakt>=> Computerprogramm, Programm	–	–
<nomen.Artefakt>=> ?akustisches Gerät	–	–

```
<nomen.Koerper>Oberarmknochen
   <nomen.Koerper>=> Armknochen
      <nomen.Koerper>=> Knochen, Gebein
          <nomen.Koerper>=> Hornsubstanz
              <nomen.Koerper>=> Körpersubstanz
                  <nomen.Substanz>=> Stoff1, Substanz,
                       Materie
                      <nomen.Tops>=> Objekt
```

Another kind of compound in the corpus are compounds with a prefix that describes a `body part`, e.g. *'Nierenschleimhaut'* (kidney mucosa), *'Brustwirbelsäule'* (thoracic spine). `body part` can be named a region of the body or an organ. In this case, the following restrictions should be considered by the method:

- both parts of the compound should have an entry in GermaNet and
- the parts of the compound should also appear in the corpus as a complex noun phrase: first part of the compound is the complement and the second part of the compound should be the keyword (e.g., *'Magenschleimhaut'* (stomach mucosa) vs. *'Schleimhaut des Magens'* (mucosa of stomach).

In these cases, information via GermaNet's meronym relation is deduced.

3.4. Disambiguation

The fundament of correct deduction of concepts is the selection of the correct sense of the senses available in GermaNet. In our case, the restriction to one top level category is sufficient for this analysis of forensic autopsy protocols, especially the findings section. In this section, only anatomic concepts and its findings are described. For other domains, it is necessary to use methods for a certain word sense disambiguation, e.g., methods that used selectional preference (see (Resnik, 1997) or (Abney and Light, 1999)) or conceptual density ((Agirre and Rigau, 1996)) for word sense disambiguation.

4. Related Work

The approach exploits the specific syntactic structures of a sublanguage. In the work of (Kokkonakis et al., 2000), the analyses of compounds and specific syntactic structures are used for the extension of the Swedish SIMPLE lexicon. This work exploits the advantage of the productive

compounding characteristic of Swedish to derive new lexical items (results in information about semantic type, domain, and semantic class). Furthermore, they used a raw and partially parsed corpus for the analyses of enumerative NPs (with more than three common nouns) for the derivation of co-hyponyms. The following heuristic is used for an unknown noun in an enumerative NP: if at least two nouns have the same assignment to a semantic class, then there is a strong indication that the rest of the nouns are co-hyponyms and thus semantically similar with the two already encoded nouns.

The usage of a lexical resource to learn new entries for the same resource (WordNet) is described in (Navigli and Velardi, 2002). This paper outlines an approach for the deduction of a sense of multi-word terms that is based on the senses of individual words of the multi-word terms. Another similar approach that combines corpus and WordNet information to deliver verb synonyms for high frequent verbs of a domain-specific sublanguage is described by Xiao (Xiao and Rösner, 2004). Peters (Peters, 2004) describes how new knowledge fragments can be derived and extended from synonymy, hypernymy and thematic relations of WordNet and implicit information from the (Euro)WordNet.

5. Conclusion

Linguistic resources with domain-specific coverage are crucial for the development of concrete application systems. In this paper, we proposed an approach for the extraction of semantic information, using the information available in GermaNet for the individual words that frequently occur in a specific syntactic structure of the corpus.

The results of the approach can be helpful for the corpus based semiautomatic extension of the GermaNet resources. With this approach, it is possible to extract information about a new entry (e.g., *forearm bone*) or to complete senses or hypernym information for entries existing in GermaNet (e.g., *lower leg*). The results also contain synonyms, like *'Jochbogen' (zygoma)*, *'Jochbeinknochen' (zygomatic bone)*, and *'Jochbogen' (zygomatic)*, which can be detected by an deeper context-related investigation of the elements of a complement set.

In future work, we will evaluate the approach for other syntactic structures and investigate if it is possible to deduce information about the keyword of a syntactic structure when the complements are known. Another aspect will be the exploitation of the resources of the Medical Subject Headings (MeSH).[4] The investigation points are: How many medical terms (in a more everyday language) of the forensic autopsy protocols are covered by MeSH? and What differences exist between entries of MeSH and GermaNet, because Basili et al. describe some discrepancies between entries in MeSH and WordNet (Basili et al., 2003). Further on, this paper outlines the mapping of a domain concept hierarchy (MeSH) with a lexical knowledge base (WordNet) for the building of a linguistically motivated domain hierarchy. If such an approach is necessary in the analysis of forensic autopsy protocols, it should be considered in further analyses of the corpus and the evaluation by medical experts.

[4] http://www.nlm.nih.gov/mesh/meshhome.html

6. References

Abney, S. and M. Light, 1999. Hiding a Semantic Hierarchy in a Markov Model. In *Proceedings of the Workshop on Unsupervised Learning in Natural Language Processing*. College Park, MD.

Agirre, E. and G. Rigau, 1996. Word Sense Disambiguation Using Conceptual Density. In *Proceedings of COLING'96*. Copenhagen, Danmark.

Basili, R., M. Vindigni, and F. M. Zanzotto, 2003. Integrating ontological and linguistic knowledge for conceptual information extraction. In *Proceedings of IEEE/WIC International Conference on Web Intelligence (WI'03)*. Halifax, Canada.

Hamp, B. and H. Feldweg, 1997. GermaNet – a lexical-semantic Net for German. In P. Vossen et al. (ed.), *Proc. of ACL/EACL-97 workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*. Madrid.

Kokkonakis, D., M. Toporowska Gronostaj, and K. Warmenius, 2000. Annotating, Disambiguating & Automatically Extending the Coverage of the Swedish SIMPLE Lexicon. In *Proceedings of LREC 2000*. Athens, Greece.

Kunze, C., 2001. *Lexikalisch-semantische Wortnetze*. Heidelberg; Berlin: Spektrum, Akademischer Verlag, pages 386–393.

Kunze, M. and D. Rösner, 2003. Issues in Exploiting GermaNet as a Resource in Real Applications. In A. Wagner C. Kunze, L. Lemnitzer (ed.), *GermaNet-Workshop: Anwendungen des deutschen Wortnetzes in Theorie und Praxis*. Tübingen, Germany.

Navigli, R. and P. Velardi, 2002. Automatic Adaptation of WordNet to Domains. In *Proceedings of the OntoLex 2002*. Las Palmas, Canary Islands, Spain.

Peters, W., 2004. Building and Extending Knowledge Fragments. In P. Sojka et al. (ed.), *2nd International Conference of the Global WordNet Association*. Brno, Czech Republic.

Resnik, P., 1997. Selectional Preference and Sense Disambiguation. In *ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How?*. Washington, D.C., USA.

Rösner, D. and M. Kunze, 2002a. Exploiting Sublanguage and Domain Characteristics in a Bootstrapping Approach to Lexicon and Ontology Creation. In *Proceedings of the OntoLex 2002*. Las Palmas, Canary Islands, Spain.

Rösner, D. and M. Kunze, 2002b. An XML based Document Suite. In *Proceedings of Coling 2002*. Taipei, Taiwan. ISBN 1-55860-894-X.

Xiao, C. and D. Rösner, 2004. Finding High-frequent Synonyms of a Domain-specific Verb in English Sublanguage of MEDLINE Abstracts Using WordNet. In P. Sojka et al. (ed.), *2nd International Conference of the Global WordNet Association*. Brno, Czech Republic.

Automatic Thai Ontology Construction and Maintenance System

Asanee Kawtrakul **Mukda Suktarachan** **Aurawan Imsombut**
The Specialty Research Unit of Natural Language Processing and
Intelligent Information System Technology
Department of Computer Engineering,
Kasetsart University, Bangkok, Thailand
{ak, mukda, aurawani}@vivaldi.cpe.ku.ac.th
Tel. (662)9428555 Ext. 1438

Abstract

Ontology is an essential resource to enhance the performance of Information Processing system such as information integration, document classification in taxonomies, including information retrieval and data cleaning in database system. This paper proposes three methodologies for Automatic Thai Ontology Construction and Maintenance from technical corpus, dictionary and thesaurus. For corpus based ontology construction, Shallow Parser is used for terms extraction. Syntactic-semantic constraint and Name Entities Extraction are used for ontological relation identification. For dictionary based Ontology extraction, we applied Task Oriented Parser to extract relational terms. Finally, we converted Broader/Narrower relation of the Domain Specific thesaurus to IS-A relation. The accuracy of the Automatic Thai Ontology Construction and Maintenance System based on agriculture corpus, dictionary and thesaurus is 73 %, 100% and 91% respectively. The organizing system accuracy is 87 %.

1 Introduction

Ontology is a principle of any system to represent knowledge for a given domain. It represents information from multiple heterogeneous sources in concepts and semantic relations of the concepts. Davies, et al., (2003) classified ontology in two essential aspects, one is real-world semantic ontology such as word taxonomy and another is formal semantic ontology such as task-oriented ontology.

Ontology plays the important role in increasing magnitudes with the performance of Information processing system such as information integration, document classification in taxonomies including information retrieval System. Creating ontology by the expert is an expensive task and it is endless task for ontology maintenance since its content relies on user requirement. It is also the fact that information in the real world has been increased. Especially in scientific documents, there are rapidly new terms and instance generation and difficult to follow up. By this reason, it is necessary to construct and maintain ontology automatically in order to update Ontology data.

There are three sources for Ontology Extraction: Raw text, Dictionary and Thesaurus. Raw text is a huge information source and they are frequently updated. Therefore a number of proposals have been made to facilitate ontological engineering based on unstructured text. Hearst (1992) and Landau & Morin (1999) developed a method for the automatic acquisition of hyponymy relations by identifying a set of frequently used and unambiguous lexico-syntactic patterns. Maedche and Staab (2001), Kiet et al. (2000) and Navigli et al. (2003) have used statistical techniques and machine learning to construct ontology. For dictionary based ontology extraction, there are many researches used a specialized grammar and a combination of heuristics for identifying ontological terms. Soergel et al., Clark et al. (2000) and Wielinga et al. (2001) have combined simple NLP techniques for constructing ontology from thesaurus. Many of these proposals used existing concept hierarchies from WordNet, SemCor and GermaNet to be knowledge base.

In this paper, we presented the Automatic Thai Ontology Construction and Maintenance system for Agricultural domain. There are three sources consisting of corpus, dictionary named "Thai Plant Names" Dictionary (Smitinand, 2001) and multilingual thesaurus named "AGROVOC" (http://www.fao.org/agrovoc). The system will extract only Hyponym, Meronym and Synonym relations. For corpus based ontology construction, Shallow Parser and Name Entities Recognition are used for terms extraction and syntactic-semantic constraint is used for ontological relation identification. For dictionary based Ontology extraction, we applied Task Oriented Parser to extract relational terms. Finally, we converted Broader/Narrower relation of the Domain Specific thesaurus to IS-A relation. In addition, we develop tool for expert to verify and extend the Ontology.

The paper is outlined as follows: The next section introduces the Crucial Problem for Thai Ontology Extraction. Section 3 explains in detail of the Automatic Thai Ontology Construction and Maintenance System. The Ontology Organizing System will be presented in Section 4. Section 5 shows the Ontology verification tool. Finally, we conclude the overall of the construction and maintenance system.

2 Crucial Problems for Thai Ontology Extraction

2.1 Related Terms Distance

In the text, we often found that the head word, which we interest in, stand far away from the related terms like in the sentence. For example, "***Fruits*** that provide the most nutrients help to form the foundation of a nutritious diet such as *mango, cantaloupe, apricots, kiwi fruit, strawberries, oranges and prunes*.". This kind of sentence causes the problem of reference linking between head word and its related terms.

2.2 Clue word's sense and its function disambiguation

Using clue words set for hinting relationship of terms is a technique for Ontology extraction. Nevertheless Thai language has no derivation. One word has several functions and several meanings. Thus we have to prune irrelevant clue word function and meaning by using Shallow Parser System.

3 An overview of Automatic Thai Ontology Construction and Maintenance System

Figure 1 shows the overview of the Automatic Thai Ontology construction and Maintenance System consisting of Ontology Extraction, Ontology tree organizing and Verification. In this section, we will brief only the part of Ontology Extraction.

3.1 Corpus Based Ontology Extraction

This part is a challenge task since the corpus based is unstructured data. By the observation, semantic relation expressions are both in phrase and sentence level. In the process of ontological expressions extraction, we found that there are several problems such as reference-referent identification, Semantic Relation Identification, clue word's sense and its function disambiguation.

- **The process for finding lexico-syntactic patterns**

In order to identify pattern that express semantic relations, the process was:

First Step, we select 100 pairs of IS-A concepts, 50 pairs of Part-of concepts and 50 pairs of synonym concepts from AGROVOC.

Second step, extract sentences with selected concepts pairs from the first step from the 150 Agricultural documents and subsequently deduce to lexico-syntactic pattern.

Third step, accumulate the patterns from the second step to the lexico-syntactic patterns database.

The experiment from the 150 Agricultural documents, 1,035 IS-A, 80 A-PART-OF and 80 SYNONYM occurrences were detected. There were a distinct variety of lexico-syntactic patterns. The most frequent patterns were:

Patterns	Clue-Word Meaning	Occurring (times)	%
IS-A			
NP1 เป็น NP0	is/am/are	640	67
NP0 ได้แก่ NP1, …,NPn	such as	85	9
NP0 เช่น NP1, …,NPn	for example	230	22
Other Patterns	-	80	2
A-PART-OF			
NP0 ประกอบด้วย NP1,..., NPn	consist of	80	100
SYNONYM			
NP0 ชื่อวิทยาศาสตร์ NP1	scientific name	25	31
NP0 ชื่อสามัญ NP1	formal name	20	25
NP0 ชื่อท้องถิ่น NP1	local name	15	19
NP0 ชาวบ้านเรียกว่า NP	local name	10	12.5
Other Patterns	-	10	12.5

Table1: Lexico-syntactic patterns

Based on these statistics, we decided to focus on the high eight frequency patterns as above. The problem, however, there are some of the most ambiguous hyponym relation patterns. For example,

"กาบใบเป็นสีน้ำตาล" (Leaf is brown color.)

Leaf and brown is not hyponym relation expression but it is the properties of the object. We solved this ambiguous by defining heuristic rules such as using the word list of object properties to eliminate non-concept term.

- **The process for extract corpus-based ontology**

There are three steps for corpus based ontology extraction.

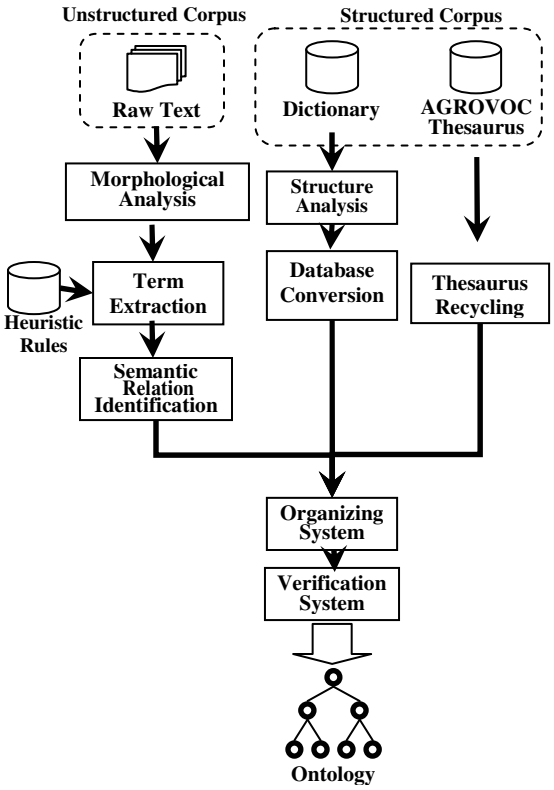


Figure 1: Thai Ontology construction and maintenance system architecture

3.1.1 Morphological Analysis Preprocessing

As other ASIAN languages, Thai text composes of a sequence of words with no delimiters. Thus, the word segmentation and POS tagging (Sudprasert & Kawtrakul, 2003) are necessary for identifying term unit with its syntactic categories.

3.1.2 Term extraction

In this process, we use the Shallow Parser with heuristic information such as clue words for signifying salient sentences and phrases.

Shallow Parser Module can be decomposed into two processes: sentence anchoring, term candidate generation and ontological terms selection.

- **Sentence Anchoring**

The sentence anchoring process suspects plausible sentences whose content bare its ontological relation. A sentence σ is said to be anchored if σ contains a clue word that is a member of C. To be more precise, an anchored sentence σ can be rewritten as a string $T_{L_{1...m}} \cdot c \cdot T_{R_{1...n}}$, where $T_{L_{1...m}}$ and $T_{R_{1...n}}$ are sequences of noun phrases or terms on the left side and on the right side respectively, $c \in C$ is a clue word, and C = {*/dai-kae/* (such as or consisting of), */chen/* (for example), */pen/* (to be), */prakop-douy/* (consists of)} is a set of clue words.

- **Term Candidate Generation**

This process shallowly parses anchored sentences and generates noun phrase as term candidates for generate ontology terms. To accomplish this process, we utilize phrase chunking as the mechanism.

Phrase chunking is to identify shallow phrase boundaries within a sentence. This process boosts up parsing speed by pruning parsing candidates. As mentioned, ontological relation can occur at any levels of constituents within a Thai sentence. Therefore, all anchored sentences are chunked to pose noun phrases, verb phrases, and subordinate clauses. In this paper, parser relies on NP and word formation rules and also lexical data. The output of this step is a list of candidate noun phrases without structural disambiguation. At this step, we can describe as 4-tuple:

{T, N, R, COMP}
where
T is the set of start symbol
(T ∈ {ncn, npn, nct, honm, pref2, pref3 ntit})
N is the set of non-terminal symbols
R is the set of rules in the grammar
(see Table 2)
COMP is the terminal symbols

Abbr.	Full word	Pattern	Example
ncn	Common Noun	ncn tv ncn	พนักงาน(ncn) ต้อนรับ(tv) receptionist
npn	Proper Noun	npn ncn	พุทธ(npn) ศาสนา(ncn) Buddhism
nct	Collective Noun	nct ncn	กลุ่ม(nct) ประเทศ(ncn) group of country
honm	Honorific maker	honm ncn	พระ(honm) พักตร์(ncn) face (king)
pref2	Prefix2	pref2 tv ncn ++	ผู้(pref2) โดยสาร(tv) passenger
pref3	Prefix3	pref3 pro2 ncn	ชาว(pref2) ต่าง(pro2) ชาติ(ncn) forienger
ntit	Title	ntit npn	นาง(ntit) กาคี(npn) Mrs. Kaki

Table 2: Noun phrase grammatical rules

After this we used Mutual Information method to extract word co-occurrence for pruning error noun phrase that was a result of the previous step. This step we emphasize to analyze noun phrases by applying both syntactic structures and also statistical technique to solve the problem of over- or under-noun phrase generation.

From the syntactic annotated corpus, we created a probabilistic of the same noun phrase by extracting the information from the document and calculating with the formula below.

$$P_{NPs}(w_i, w_j) = P_f(w_i) * P_b(w_j)$$

Where as

$P_{NPs}(w_i, w_j)$ is a noun phrase or compound noun, which w_i and w_j could be related to be a new word.
$P_f(i)$ is the frequency of the occurrence of *i* in noun phrase and follow by other words / the frequency of all *i* in the document.
$P_b(i)$ is the frequency of the occurrence of *i* in noun phrase and other words occur before it / the frequency of all *i* in the document.
For example, the word "เครื่อง" in Thai always occur in the initial position of noun phrase. Thus the value of P_b(เครื่อง) is 0, where as the value of P_f(เครื่อง) is 0.95 . While the proper name, which often occur in the final position, the value of P_b(proper name) is 0.98 where as the value of P_f(proper name) = 0

These probabilistic approaches can be applied to prune the erroneous noun phrases from the candidate noun phrases. If the noun phrases from previous part have probability more than threshold, it is likely that the noun phrase is a proper name.

- **Ontological Term Selection**

After extracting terms by Shallow Parsing, we applied several methodologies to select Ontological Term. First we use NLP Technique to deal compound candidate terms by head word consistency. Candidate term would be selected if its head word matches to head word of related terms. For example,

> **ปอกระเจา** ที่นิยมปลูกกันใน**ประเทศไทย**มี 2 ชนิด ***ได้แก่*** **ปอกระเจาฝักยาว** และ **ปอกกระเจาฝักกลม**
> (There are 2 kinds of **Jute** that are favorably plant in **Thailand *such as* Tossa Jute, White Jute**).

The system would analyze compound noun which is head word and modifier as the following rule.

NP → MOD NCN
MOD → ADJ, NCN, NPN, …

Where MOD is a modifier
NCN is Common Noun
ADJ is an adjective
NPN is a proper name

From this example, head word of Tossa Jute and White Jute are "Jute" not "Thailand" then the system will select the Jute is ontological term.

Another technique is Statistical-based Technique to analyze the co-occurrence. If there are no head word consistency terms, we will use Mutual Information to extract the co-occurrence of candidate terms and related term.

$$MI(w_1, w_2) = \log 2 \frac{P(w_1, w_2)}{P(w_1)P(w_2)}$$

Where, w_1 is a candidate term.
w_2 is a related term.
$P(w_i)$ is frequency of term w_i
$P(w_i, w_j)$ is frequency of co-occurrence of term w_i and w_j

The system will select the candidate term that has the highest mutual information value. For example,

> **สมุนไพร**หลายชนิดมีสรรพคุณเป็น**ยารักษาโรค** และมีการนำมาผลิตในระดับ**อุตสาหกรรม**แล้ว ***เช่น*** กระเทียม ใบแปะก๊วย
> (Many **herbs** can be used as **medicine** and some of them are manufactured in the **industry** level, ***such as* garlic, ginkgo biloba**)

If the word "herb" occurs with "garlic" and "ginkgo biloba" in the document more frequent than "medicine" and "industry", the system will select "herb" with "garlic" and "ginkgo biloba" to be ontological term.

3.1.3 Semantic Relation Identification

This step is to identify the correct relation (Hyponym, Meronym or Synonym). Each relation was extracted by defying in this fashion.

Hyponym Relation extraction: The NP, which occur before clue word {**ได้แก่** (*/dai-kae/* such as) and **เช่น** (*/chen/* for example) is defined as Hypernym, where the NP* occurred after clue word are Hyponym terms. The NP, which occur before clue word {เป็น (/pen/ is) is a Hyponym term and the NP occurred after clue word is a Hypernym term.

Meronym Relation extraction: The NP, which occur before clue word {ประกอบด้วย *(/prakop-douy/)*} is defined as a Whole concept, where the NP* occurred after clue word are a Part of terms.

Synonym Relation extraction: Synonym was extracted by using clue words set which is {**ชื่อวิทยาศาสตร์** */chu-wittayasart/* (scientific name), **ชื่อสามัญ** */chu-samun/* (formal name), **ชื่อท้องถิ่น** */ chu-thong-thin/* (local name) and ชาวบ้านเรียกว่า */chao-ban-reak-wa /* (local name)}.Behavior of Synonym relation with clue words **ชื่อวิทยาศาสตร์** */chu-wittayasart/* (scientific name) and **ชื่อสามัญ** */chu-samun/* (formal name) often occur across sentence by starting with a new line. Solution system, thus, mainly relies on positions of title and clue words in the document. In order to extract this kind of relation, we have to analyze position of document structure. For the clue words "**ชื่อท้องถิ่น**" */ chu-thong-thin/* (local name) and "ชาวบ้านเรียกว่า" */chao-ban-reak-wa /* (local name), we normalized pattern of sentence with those clue words to the lexico-syntactic expressions as well as Hyponym and

Meronym relation patterns. The NP before clue words was defined as the same meaning as NP after clue words.

Moreover we could use Name Entity Extraction (NE) as a process for identifying Generic-Specific terms relation. We applied (Chanlekha & Kawtrakul, 2003) to extract Name Entities from noun phrases in order to generate Generic-Specific Terms and its relative concept such as "Jasmine Rice". Jasmine will be detected as Specific term of rice.

The experiment of the Automatic Thai Ontology Construction and Maintenance System based on agriculture corpus, dictionary and thesaurus is 73 %.

3.2 Dictionary based Ontology Extraction

Since Ontology in this paper is a Domain Specific Ontology, thus Domain Specific Dictionary is a best way for extracting relational information because dictionary has certainly structure as well as clear and clean information. In this paper, the Ontology system was based on the "Thai Plant Names" Dictionary, which developed by Prof.Dr.Tem Smitinand and edited by the Forest Herbarium Royal Forest Department in 2001.

The relation that embedded in Dictionary based Ontology Construction are;

R = {HYPO, SYNO}

where

R is the Relation in the Ontology
HYPO is Hyponym Relation
SN is Synonym Relation

There are two steps for the process of Dictionary based Ontology Extraction those are Structure Analysis and Database Conversion.

2.2.1 Structure Analysis

The Dictionary Structure Analysis is an important procedure for Dictionary based Ontology Construction. This procedure will distinguish structure of word entries to sub-part. The dictionary was analyzed terms positions and pruned irrelevant part, then transferred needed parts to Hierarchical tree by using Task Oriented Parser.

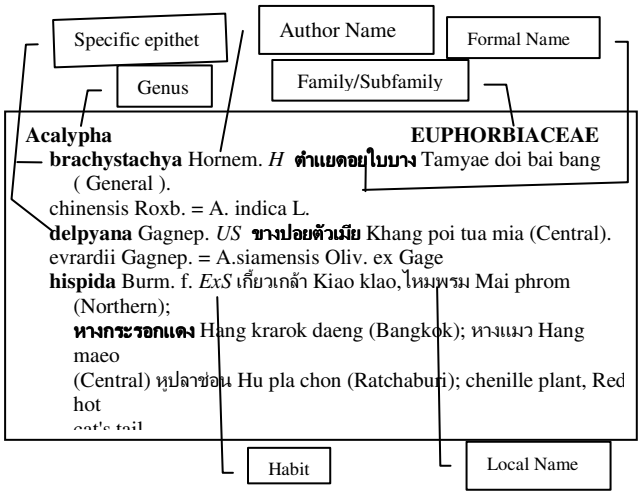


Figure 2: Dictionary Structure

The figure 2 illustrates dictionary structure analysis. We converted terms by alphabet characteristic and position of terms to relational database and prune irrelevant part and then predefine Hierarchical relation as Family, Sub-Family, Genus, Specific epithet, Formal Name and Local Name respectively.

Feature	Database field	Example
All upper case	Family/Sub-Family	EUPHORBIACEAE
Start with upper case	Genus	Acalypha
All lower case	Specific epithet	brachystachya
Thai alphabet with bold font	Formal Name	**ตำแยดอยใบบาง**
Thai alphabet	Local Name	เกี๊ยวเกล้า

Table 3: Characteristics of Dictionary Conversion.

3.2.2 Database Conversion

After parsing process, all those terms was defined identify number and then converted to relational database.The experiment on dictionary based ontology extraction is 100%.

3.3 Thesaurus Recycling to Ontology system

Thesaurus is a high-quality source for Ontology Construction and Maintenance. In this paper, we emphasize to recycling AGROVOC Thesaurus (http://jodi.ecs.soton.ac.uk/incoming/Soergel/JoDI_FAO_Soergl_revC.html), which is a multilingual agricultural thesaurus in English, French, Spanish, Portuguese, etc. that has been developed by FAO and the Commission of the European Communities in the early 1980s. It contains 16,607 descriptors and more than 10,000 non-descriptors. Each descriptor has an equivalent in the other languages.

For maintenance and adding more value of the existing ontology, we chose the AGROVOC thesaurus to recycle the original relation into ontology relationship.

The relation BT/NT in AGROVOC could be re-analyzed to semantic relation like "IS-A" relation in ontology. For example,

PLAINS

NT Coastal plains

NT Floodplains

However, not all BT/NT relation in AGROVOC could defined to "IS-A" relation. Their semantic could be defined as Ingredient of and Property of (Fisseha, 2003). For example,

Ingredient of (*MILK/ Milk Fat*)

MILK

NT Milk Fat

NT Colostrum

NT Cow Milk

Property of (*MAIZE/ sweet corn*)

MAIZE

NT popcorn

NT soft maize

NT sweet corn

Then, we will resolve this problem by heuristic method. Compound Noun in Narrower Term will be process to find head noun and if head noun is consistent to Broader term, their relation will defined as "IS-A" relation. If they are not compatible, they will be unselected to a related term.

The experiment on AGROVOC based ontology extraction is 91%.

4. Organizing System

In this step, we united the related word/phrase pairs that we collected from three parts: Corpus Based Ontology Extraction, Dictionary Based Ontology Extraction and Thesaurus Recycling to Ontology system. This system, we use AGROVOC Ontology to be a core tree because AGROVOC thesaurus has a number of concept hierarchies more than the other sources and it cover sub-domain of agriculture such as animal, plant, plant/animal pathology, chemicals etc. While the dictionary based is only plant names. Moreover the source from corpus is inappropriate for being core tree because the concept hierarchy was incomplete in itself and the concept hierarchy was not united. Then the Ontology trees from Corpus and Dictionary will be added to the AGROVOC core tree by matching techniques. There are two steps for organizing system.

4.1 Consistent terms organizing

This step, we used directly term matching for replacing exactly similarity parent concept by comparing term parents from corpus and dictionary to the core ontology. By this process, we found that there are four characteristics for tree merging.

The first case is that when the parent node is consistency to core ontology but child node disappears in core tree, child node and its relation will be added automatically to core tree. For example, the term "fruit", which consistence both in AGROVOC source and Corpus source will be merged and replaced to the core ontology tree and child node of each source will be combined as the figure 3a.

The second case, if there are several consistent parent nodes in the core tree, the system will provide child node from every sources and its relation to every parent nodes as in figure 3b.

The third case, if the parent node from different sources is a consistent node but child nodes have different hierarchical concept level, the system will allow the child nodes to be sister node as in figure 3c. If the knowledge has been increased, the level of those child nodes will be adjusted later.

The fourth case, if the grandparent and terminal node were found consistently, but one source has no parent node, the longest relation node will be kept as in figure 3d.

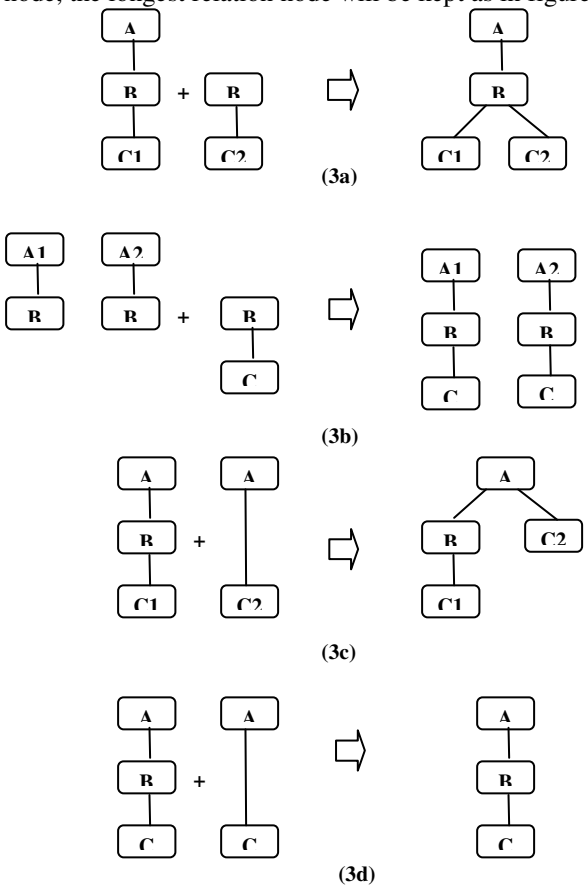


Figure 3: Terms Matching Process

4.2 Inconsistent terms organizing

The second methodology used for inconsistent terms. This step we will find concept relation of terms from head word consistency of compound noun. If a head of one

noun is consistent to another such as A and AB, AB would be defined as subclass of A.

For example, if we found that compound words that extracted from corpus often combines concept terms together such as "พืชไร่น้ำมัน" (field-oil crops), which composes of field crops and oil crops. In the other hand, the core ontology has extremely separated the terms field crops from oil crops. By the process methodology, the result then was allowed the terms "พืชไร่น้ำมัน" (field-oil crops) is a hyponym of field crops, which is a head of that compound noun. This may causes an error result.

The remaining terms will be kept for the expert to add and maintain later.

4.3 Result

The experiment in this step, we used 3,720 terms with 3,312 relations from corpus, 37,110 terms with 21,620 relations from dictionary and 27,540 terms with 15,628 relations from AGROVOC thesaurus. We found that 43,073 terms with 31,387 relations were united. By random checking with 1,000 united terms, the accuracy of the system is 87 %. The error result caused by the result of corpus extraction terms.

5. Verification Tool

Verification is required to ensure the high quality system and to guide the expert to maintain the existing Ontology. In this part, we developed user interface for the expert to verify output and add additional related word pair to the original Ontology that is the AGROVOC Ontology. Moreover the expert can move and delete Ontology node in Relational Tree as well. This is an easy and fast technique for the expert to construct and maintain the existing Ontology and apply it to some applications.

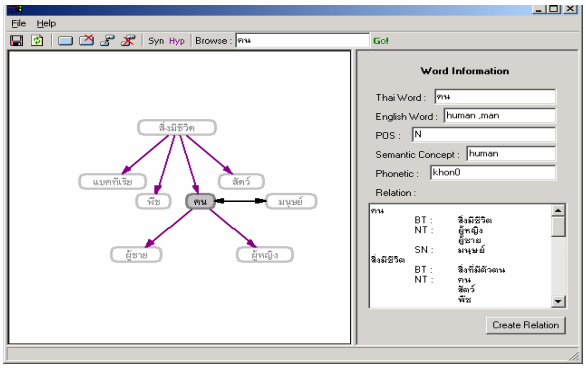


Figure 4: Ontology Verification Tool

6. Conclusion

This paper presents the system of Automatic Thai Ontology Construction and Maintenance by automatically extracting Thai Ontology on Agriculture corpus and recycling existing resources such as dictionary and thesaurus. The accuracy of the Automatic Thai Ontology Construction and Maintenance System on Corpus Based Ontology Extraction, Dictionary Based Ontology Extraction and Thesaurus Recycling to Ontology system is 73 %, 100% and 91% respectively. The organizing system accuracy is 87 %.

References

- Davies, J., et al. (2003). Towards the Semantic Web: Ontology Driven Knowledge Management. John Wiley & Sons Inc.
- http://www.fao.org/agrovoc/
- http://jodi.ecs.soton.ac.uk/incoming/Soergel/JoDI_FAO_Soergl_revC.html
- Hearst, M. (1992). Automatic acquisition of hyponyms from large text corpora. In Proceedings of the 14th International Conference on Computational Linguistics.
- Landau, M. F. & Morin, E. (1999). Extracting Semantic Relationships between Terms: Supervised vs. Unsupervised Methods. Proc. International Workshop on Ontological Engineering on the Global Information Infrastructure. Dagstuhl Castle, Germany
- Maedche, A. & Staab, S. (2001). Ontology Learning for the Semantic Web. IEEE Intelligent Systems, vol. 16, no. 2.
- Kietz, J. U., et al. (2000). A Method for Semi-Automatic Ontology Acquisition from a Corporate Intranet, Proceedings of EKAW-2000 Workshop "Ontologies and Text", Juan-Les-Pins, France.
- Navigli, R., et al. (2003). Ontology Learning and its application to automated terminology translation. IEEE Intelligent Systems, vol. 18, n.1, January February 2003.
- Yamaguchi, T. (1999). Constructing domain ontologies based on concept drift analysis. Proc. IJCAI-99 Workshop on Ontologies and Problem-Solving Methods, August, Stockholm, Sweden.
- Jannink, J. (1999). Thesaurus Entry Extraction from an On-line Dictionary. In Proceedings of Fusion '99, Sunnyvale CA.
- Ketsuwan, C., et al. (2000). Automatic Thesaurus Extraction for Thai Text Retrieval Enhancement", WAINS 7: E-Business for the new Millennium, Bangkok, Thailand.
- Clark P., et al. (2000). Exploiting a Thesaurusbased Semantic Net for Knowledge-based Search", Proc. Of IAAI-2000.
- Wielinga, B. J., et al. (2001). From Thesaurus to Ontology. Internation Conference on Knowledge Capture, Victoria, Canada.
- Smitinand, T. (2001). Thai Plant Names Dictionary. The Forest Herbarium Royal Forest Department, Thailand.
- Sudprasert, S. & Kawtrakul, A. (2003). Thai Word Segmentation Based on Global and Local Unsupervised Learning, Proc. Of NCSEC2003, Chonburi, Thailand.
- Abney, S. P. (1991). Parsing by Chunks. In Robert C. Berwick, Steven P. Abney, and Carol Tenny, editors, Principle-Based Parsing: Computation and Psycholinguistics (pp. 257—278). Kluwer Academic Publishers, Boston.
- Chanlekha H. & Kawtrakul, A. (2003). Thai Name Entity Extraction by Using Maximum Entropy Model Knowledge base, NCSEC'2003, Chonburi, Thailand.
- Fisseha, F., et al. (2003). Reengineering AGROVOC to Ontologies Step towards better Semantic Structure, NKOS Workshop.

AN ONTOLOGY FOR MULTILINGUAL TREATMENT OF PROPER NAMES

Mickaël Tran[1], **Thierry Grass**[2], **Denis Maurel**[1]
[1]LI (Laboratoire d'Informatique de l'Université de Tours)
[2]Groupe de recherche Langues et représentation (Université de Tours)
mickael.tran@etu.univ-tours.fr, {denis.maurel, thierry.grass }@univ-tours.fr

Abstract

This paper deals with the ontology of a multilingual database of proper nouns, for machine translation, computer aided translation, information retrieval and spelling dictionaries. This database is based on a five-layered ontology and a set of language independent and language dependent relations.

Keywords: Ontology, Proper Names, Electronic Dictionaries, Machine Translation, Computer Aided Translation

1 Motivations

This work takes place within the *Prolex* proper names processing project and more specifically within the *Technolangue* project supported by the French Ministry of Industry together with two companies, *Systran* and *Exalead*. It aims to create a multilingual database of proper names, the *Prolexbase*, with some linguistic information for natural language processing: machine translation, computer aided translation, information retrieval and spelling dictionaries.

Even if the database can be seen as an electronic dictionary of proper names, it is not only a list of words, because it is based on an ontology including different relations. This paper describes this ontology (layers, section 3, and relations, section 4), after a short description of the interface of the database (section 2). The logical model will be presented in section 5.

An ontology, according to (Temmerman Rita, 2003) "represents an agreed upon conceptualisation of the real world". Our ontology aims at modeling the linguistic class of proper names. However, (Gruber, 1995) writes that "a conceptualization is an abstract, simplified view of the world that we wish to represent for some purpose…An ontology is an explicit specification of a conceptualization". So we have introduced a conceptual proper name (the pivot, see section 2.4) that is the referent from different points of view. Our ontology consist in five different layers: the languages layer (a source language), the instances layer (the proper names as they appear in texts in the specific language selected), the linguistic layer (the canonical form of the instances), the conceptual layer (pivots) and the meta-conceptual layer (types and supertypes). See Figure 1 for one example.

Prolex proper names processing project, based at the University of Tours (France), started in 1994 (Maurel et al., 1996). It associates both linguists and computer scientists.

A previous paper, presented during the *Portal Conference* (Grass et al., 2002) held in Faro (Portugal), showed the inadequacies of existing classifications of proper names for NLP (i.e. (Bauer, 1998), (Koß, 1999) and (Zabeeh, 1968)), and then went on to adapt and improve these classifications to construct an ontology. This paper shows the implementation of this ontology and the new approach which results.

2 Ontology: a five-layered hierarchy

As we mentioned in our introduction, our ontology is structured in five layers that will be described below.

2.1 The languages layer

The selection of a source language is the first step to identify proper names in a text. Each language treated in the database appears with its ISO 639 Language Codes in a table, i.e. *fr* for *French* or *en* for *English*. Because of the multilingual dimension, the Unicode Standard (UTF 16) is used (http://www.unicode.org).

That means that two homographs in two different languages are duplicated. For example, there are two items *France*, one for French and one for English. The following relations justify this choice. Nowadays, the languages being implemented in the data base are English, French, German, Polish, Russian and Serbian.

2.2 The instances layer

A proper name, as found in a text, can have different instances for scriptural or morphological reasons in a particular language. All possible instances are not listed in our database, but only paradigms in two tables: one for aliases and one for inflexions.

Alias is a (0,1) relation that associates a proper name to a pattern indicating the rules of alias formation. Aliases refer to the *internal structure*, as defined by (MacDonald, 1996). Aliases are:

- Variants between uppercase and lowercase (*Siemens* or *SIEMENS*).
- Acronyms (*UNO* or *United Nations Organization*).

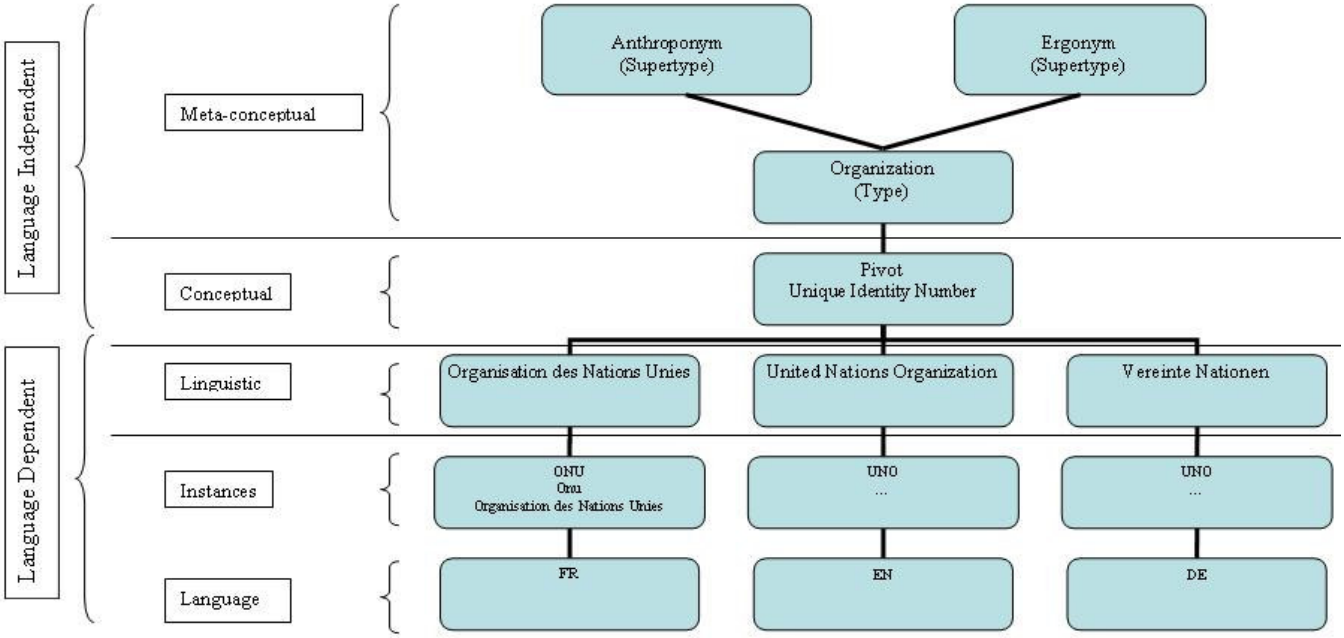


Figure 1: The five layers of the *Prolexbase*

- Abbreviation (*George Walter Bush* has for aliases *George W. Bush*, *George Bush*, *Bush*… *Microsoft Corporation* has for aliases, *Microsoft Corp.*, *Microsoft*…).
- Transcription (*Чехов –Chekhov–*can be at least written in two different manners in French *Tchekhov* or *Tchékhov*).

Inflection is a (0,1) morphological relation between a canonical form (lemma) of a proper name and a paradigm. It includes, depending on the language: gender, number and case. For example, the name of the French president, *Chirac*, invariable in French, can take the genitive mark in German (*Chiracs*) and in Polish (*Chiraca*), the dative mark in Polish (*Chiracowi*)…

2.3 The linguistic layer

The aliases and the inflected words for the same proper name in one language are linked to one unique canonical form (lemma).

2.4 The conceptual layer

Pivots are unique identity (ID) numbers which are associated to the canonical form of each proper name in every language as defined in the linguistic layer.

Concept is the link between a pivot and a canonical form of a proper name in one language. It is a (1,n) relation. A conceptual proper name is not the linguistic referent, but an historical concept, like "this proper name in a period" (see the relation Is_Renamed, section 3), a register concept, like "this proper name in slag" (see the relation Synonymy, section 3), etc.

Translation is a relation inferred by the Concept relation. As explained in the previous section, a proper name is a translation of another one if both are sharing the same pivot.

2.5 The meta-conceptual layer

Types and supertypes are two different levels of semantic information integrated in one table.

The four supertypes are:

- Anthroponyms (personal and collective names)
- Toponyms (place names)
- Ergonyms (object and work names)
- Pragmonyms (event names)

This first level gives minimal information about a proper name. It is generally possible to recognize the supertype from the linguistic context, without any human supervision. We have already developed some automatic procedures based on finite-state transducer cascades with 93,2% recall and 94,4% precision for the recognition of personal, organization and place names in French (Friburger, 2002).

The second level is more precise and consists of twenty-nine types. These types are lexical classes, determined by homogeneous semantical characteristics. They can be useful for computer aided translations. Due to the fact that a pivot can be directly related to a supertype in an automatic procedure, types and supertypes are listed in the same table (see (Grass et al., 2002) for a detailed list of types).

These types are linked to the supertypes by a relation (0, n) of **Hyponymy**: this means that supertypes are hyperonyms of several types and types are hyperonyms of conceptual proper names (pivots). This relation between types and pivots is independent of language.

Each proper name is associated to only one type, otherwise we consider them as homonyms and we duplicate their pivots, i.e. Washington as a town/toponym and Washington as a celebrity/anthroponym. That means that two homonyms get two different pivots.

3 Ontology: relations

There are two different kinds of relations within the ontology. On the one hand, there are language independent relations between the pivots and on the other hand, there are language dependent relations between the canonical forms

3.1 The language independent relations

Synonymy is a (0,n) relation between two pivots. One of them is considered as a standard form, defined for translation purposes. For example, there are two pivots for an inhabitant of Poland, one in standard language (*Polish*) and one in slang (*Polack*). This couple of synonyms also exists in French and in German, but probably not in the Polish language. If the synonymy relation is language independent, the existence of synonyms is language dependent. Another example is *Parisian* (an inhabitant of Paris) that has no synonym in English or German, but only in French: *Parisien* in standard French and *Parigot* in French slang. There are not only sociolinguistic differences entering in the concept of synonymy, but also the situation of communication. For example, *France* and *French Republic* are only synonyms in a political context. You wouldn't say that you spend your holidays on the beaches of the French Republic…

Meronymy is a (0,n) relation (Miller et al., 1990) between pivots. For example, *Lisbon* is included in *Portugal*. This relation also concerns different companies belonging to a same group. We have added to the meronymy relation the particular relation between a toponym and a anthroponym (the inhabitant names). Generally, it is also a linguistic morphological derivation, as the couple *Portugal*/*Portuguese*, but not always, as the couple *Saint-Etienne*/*Stéphanois* (*Saint-Etienne* is a French city, well-known for its football club). In an international football match, *French* can be used instead of *Stéphanois* because *Saint-Etienne* is a meronym of *France*.

Cap is a (0,n) relation, first introduced by (Mel'čuk, 1984, 1988, 1992); it means *the chief of* or *the head of* and applies to different pivots, such as a company or a state and their chiefs/heads. It also includes the relation between a country or a region and its capital.

Is_Renamed is a diachronic (0,n) relation between two pivots whose name changed for historical reasons. The city of *Karl-Marx-Stadt* in the former GDR has been renamed with its first name *Chemnitz*, after the German Reunification. *Zaire* has been renamed *Democratic Republic of Congo* after a putsch.

3.2 The language dependent relations are

Determination is a (0,1) relation that indicates whether a proper name is generally constructed with an article or not. For example, *Portugal* always takes an article in French (*le Portugal*), but not in English.

Blark (Basic LAnguage Resources Kit) (Cucchiarini et al., 2000) is an indicator of how well known a proper name is, even it is only well known for a certain period. *Buddha*, *Socrates*, *Mozart*, *Tokyo* belong to a fund of proper names which is well-known all over the world.

Extended context is a (0,n) relation between a canonical form of a proper name and typical words appearing with it like *President* in *President Bush*. It can also be a (0,n) relation between two pivots linked by the Cap relation, like *President* in *Bush, President of the USA*. Extended context is linked with information about the position and the syntactic structure in order to build local grammars. It refers to *external structure*, as defined by (MacDonald, 1996).

4 Logical model

This ontology is represented in the conceptual data model of Figure 2:

The LANGUAGE table: the languages layer.

The INTERNAL_STRUCTURE and the INFLECTION tables: the instances layer.

The PROPER_NAME table: the linguistic layer.

The PIVOT table: the conceptual layer.

The TYPE table: the meta-conceptual layer.

5 Conclusion and future prospects

This paper dealt with the ontology of a multilingual database of proper names, for machine translation, computer aided translation, information retrieval and spelling dictionaries. The core of the database is constituted by a table of pivots that represents the concepts of the proper names, without commitment to any language. This ontology differentiates between language dependent and language independent relations.

The data are saved with the open source database *MySQL*. The interface itself was implemented using the programming language *Java* and the general-purpose scripting language *PHP* that is especially suited for Web development and can be embedded into *HTML*. The interface allows users to add new data, lists of data or relations and modify data or relations, or view data and bilingual translation. A free query of the *Prolexbase* is available at http://tln.li.univ-tours.fr/.

Another goal of the project is to use the *Prolexbase* to create linguistic tools for Natural Language Processing, such as transducers for tagging purposes or local grammars…

Nowadays, the French table counts more than 323 000 entries and 55 000 links of relation; the German table counts about 13 000 entries with translation links into French. Other languages (English, Polish, Russian and Serbian) are in progress.

References

Bauer G. (1998), *Namenkunde des Deutschen*, Berlin, Germ. Lehrbuchsammlung Band 21.

Cucchiarini C., Daelemans W., Strik H. (2000), Strengthening the Dutch Human Language Technology Infrastructure, http://www.elda.fr/article48.html.

Friburger N. (2002), Reconnaissance automatique des noms propres ; application à la classification automatique de textes journalistiques, thèse de doctorat en informatique, Université de Tours.

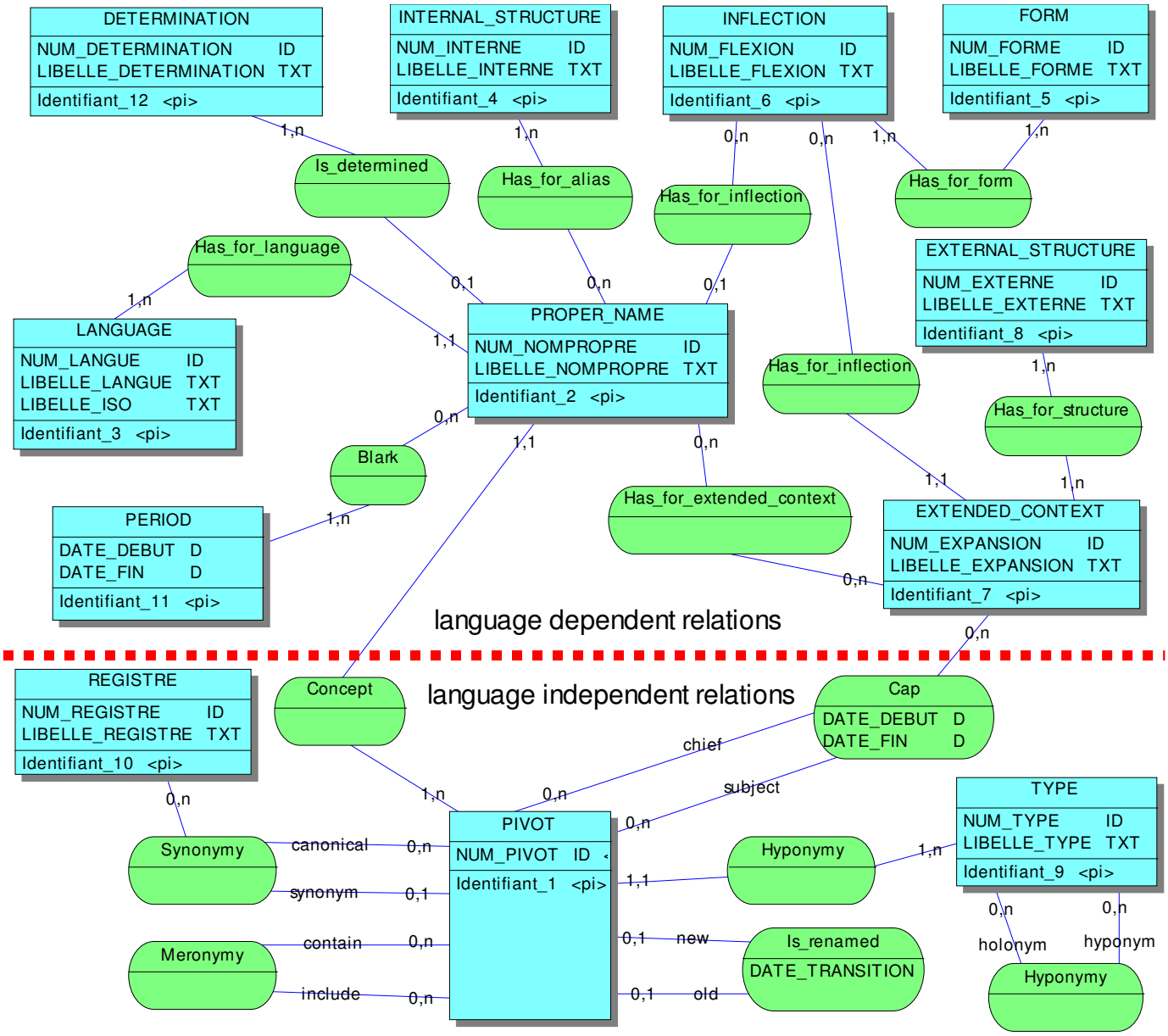


Figure 2: Conceptual data model

Grass T., Maurel D., Piton O. (2002), Description of a multilingual database of proper names, *PorTal 2002*, Faro, Portugal, 23-26 juillet, in *Lecture Notes in Computer Science*, 2389:137-140.

Gruber T. R. (1995), Toward Principles for the Design of Ontologies Used for Knowledge Sharing, Int. Journal of Human-Computer Studies, Vol. 43, 907-928.

Koß G. (1999), Was ist 'Ökonymie'?, *Beiträge zur Namensforschung*, 34-4, Heidelberg, Universitätsverlag C. Winter, 373-444.

MacDonald D. (1996), Internal and external evidence in the identification and semantic categorisation of Proper Names, *Corpus Processing for Lexical Acquisition*, 21-39, Massachussetts Institute of Technology.

Maurel D., Belleil C., Eggert E., Piton O. (1996), Le projet PROLEX, séminaire Représentations et Outils pour les Bases Lexicales, Morphologie Robuste de l'action Lexique du GDR-PRC CHM, (Actes p. 164-175), Grenoble, 13-14 novembre.

Mel'čuk I. (1984-I, 1988-II, 1992-III), *Dictionnaire explicatif et combinatoire du français contemporain*, Les presses de l'Université de Montréal.

Miller G., Beckwith R., Fellbaum C., Gross D., Miller K. (1990), Introduction to WordNet : an on-line lexical database, *International Journal of Lexicography*, n°3, p. 235-244.

Temmerman R. (2003): The ontology shift in terminography, Seminar on Multilingual Terminography: Towards Intelligent Dictionaries ?

Zabeeh, F. (1968), What's in a Name? - An Inquiry into the Semantics and Pragmatics of Proper Names, Den Haag, Martinus Nijhoff.

Integrating Semantic Lexicons and Domain Ontologies

Roberto Basili, Michele Vindigni, Fabio Massimo Zanzotto

Department of Computer Science
University of *Roma, Tor Vergata*, Roma (Italy)
{basili,vindigni,zanzotto}@info.uniroma2.it

Abstract

The cross-fertilization between advanced data modeling technologies (as aimed in the Semantic Web) and NLP is a very interesting research line. In this paper, we investigate a solution of a particular problem in this area: the intregration between a concept hierarchies (DCH) with a general-purpose linguistic knowledge base (LKB). The method we propose relies only on the taxonomical knowledge of the DCH and on the topology of the lexical knowledge base. Performances of the proposed method have been analysed on a case study (the integration of the MeSH, Medical Subject Headings, and WordNet).

1. Introduction

The Semantic Web (Lee, 2001) represents a new direction in the area of knowledge representation as the attempt to make information, texts and knowledge shareable throughout the different actors (e.g. producer/provider vs. consumer/user) involved in the scenario of a Web application. For language and text processing, the Semantic Web is a new challange and an opportunity.

On the one hand, linguistic competences are needed at least in two major perspectives. First, a strong linguistic ground is needed for expressing domain ontologies. An *ontology* is seen as a theory about a domain that obeys to specific logical contraints. However, it is usually expressed in linguistic terms as entities and relations are firstly *named* by the knowledge engineer. Most of the work done in knowledge representation and modeling for language understanding (e.g. (Gruber, 1993)) is analogous and it is targeted to a very similar task: disambiguation of texts. Any research about treatment of semantics over the Web should thus take into account all the cumulated experiences (principles, formalisms, and results, i.e. existing knowledge bases) and language oriented resources (e.g. dictionaries, lexicons and thesauri). A second aspect makes the Semantic Web dependent on research in NLP. Before entering in the scenario of interoperable services, any Web document is (at a large extent) a *textual object* and, as such, it obeys to laws that are linguistic in nature. Mapping any such instance into a (set of) semantically interoperable *data object*(s) clearly involves linguistic capabilities. As already noticed some years ago, the task of information extraction (IE) (Basili and Pazienza, 1997) can be formulated as the activity of *matching/discovery stuctured information* (i.e. the target templates) *where such a structure is only implicitly present* (i.e. in texts). Clearly, this applies to most of the information currently available within the Semantic Web. Although the IE mapping can be under the responsibility of the provider, every textual object needs to be rewritten for semantic interoperability. This calls for *robust* and *large scale* NLP capabilities.

On the other hand, advantages in the opposite direction also exist. First of all, the Semantic Web has pushed in the recent years the advent of a number of formalisms (e.g. RDF, OIL, DAML+OIL) expressedly defined for knowledge representation and exchange (Fensel et al., 2003). The benefits of this initiative are certainly in the area of logic, data modeling and knowledge representation, although some progress was undoubtedly due to previous research in ontological semantics in AI. Nowaday researchers, developers and practitioners in the Semantic Web area can rely on very powerful infrastructures (e.g. the support of XML for storing and exchanching complex data) and languages for sharing knowledge of a realistic size. As a result, there is an increasing availability of semantic resources focused on real applications and thus specific domains. Most of them (e.g. DAML ontology library) are already interoperable as they share basic principles and formalisms. Semantic interoperability increases the rate by which large knowledge bases can be produced and exchanged. This trend will (and in fact *is* currently) offer(ing) a number of resources that are large scale world models with underlying systematic principles and semantics. The awareness about the utility and effectiveness of such models in contemporary IT (e.g. *Web services* and *knowledge management*) is increasing so that a growing number of these resources can be realistically expected to be delivered in the near future. This state-of-affairs opens research directions for NLP. Ontologies will play the role of semantic models of application domains and the more systematic will be their design the more usable and powerful they will result. Here, the target issue is not their possibility of becoming a standard for their target application domains. It is more interesting their role in near-future NLP services. Most of the linguistic ambiguity, typical of NL texts, is strikingly reduced whitin a domain-specific semantic model. The ability of integrating these by-products, i.e. extensive domain ontologies, with their own lexical framework will make NLP applications (e.g, IE systems) very effective. First, they will be able to rely on ontological contraints for ambiguity resolution and this will increase basic performances. Second, and more importantly, the semantic interpretation process of any piece of text will be able to exploit both systems: a domain ontology (with ts own inheritance and other inference mechanims) and a semantic lexicon. Although still non deterministic (as a many-to-many relationship can be exected in general to hold between ontological and lexical concepts), the inter-

pretation process will be more robust with respect to lacks in domain or lexical knowldge: failures (or missing information) in one semantic system will be compensated by the other. Near-future NLP applications will be thus more and more demanding relatively to the integration of ontological and lexical knowledge. As this mappings are crucially domain and application dependent, methods for learning them from texts and/or existing resources is a very important research line.

The cross-fertilization between advanced data modeling technologies (as aimed in the Semantic Web) and NLP is thus a very interesting research line. In this paper, we then discuss a method for a solution of a particular problem in this area: the intregration between a concept hierarchy (DCH) typical of an ontology with a general-purpose linguistic knowledge base (LKB). The method we propose relies only on the taxonomical knowledge of the DCH and on the topology of the lexical knowledge base (Sec. 2.). A case study, i.e. the integration of the MeSH, Medical Subject Headings (MESH), and WordNet (Miller, 1995), will be then presented as a proof of the effectiveness and accuracy of the overall approach (Sec. 3.).

2. Integrating an Ontological Hierarchy with a Lexical Network

A suitable mapping between a domain specific resource, i.e. a *domain concept hierarchy* (**DCH**) of a Semantic Web ontology, and a domain-independent lexical knowledge base (**LKB**) is beneficial both for NLP and for Semantic Web (SW). To explain textual phenomena language oriented "is_a" hierarchies spanning over different domains are required. This mapping deals with a general many-to-many correspondence between the DCH ontological concepts and the LKB word senses.

In the rest of the section we will often discuss notions like concepts (in DCH), word senses (in LKB) and their structural properties in the underlying resources. Any concept C in the domain hierarchy (DCH) is characterized by its linguistic label hereafter noted as t_C. This label t corresponds either to a singleton word or to a multiword expression. This information can be used as a reference within the lexical knowledge base LKB. We will denote LKB entries by means of Greek letters, e.g. α. Those LKB senses that correspond to possible linguistic meanings of label t will be denoted as α_t. Sometimes α_t may not exist for technical concepts as they are not present in the domain independent LKB [1]. In general, a label t will correspond to more than one sense.

As DCH and LKB have both an internal structure some other useful properties can be introduced. First of all we will call *linguistic extension* of a DCH concept C, denoting it as $ext(C)$ the set of the labels for C or for one of its descendants C' as follows:

$$ext(C) = \{t_{C'} | C \text{ subsumes } C' \text{ according to } DCH\} \quad (1)$$

For example the linguistic extension of *Tissues* in MeSH (MESH) includes words and terms like "*Articular Cartilage*", "*Corneal Endothelium*".

Given its extension, a DCH concept C can be interpreted in LKB via its *linguistic generalization set*, that is the set of generalizations, α_t, in LKB for the labels $t \in ext(C)$. It will be denoted by $lgen(C)$ that is defined as

$$lgen(C) = \{\alpha \in LKB | \exists t \in ext(C) \text{ and } \alpha_t \text{ is subsumed by } \alpha \text{ in } LKB\} \quad (2)$$

Due to language ambiguity the generalization set $lgen(C)$ includes more senses in LKB than those strictly needed to represent C. The lexical ambiguity of terms $t \in ext(C)$ implies that its possibly irrelevant senses $\alpha_t \in LKB$ can be included in $lgen(C)$. In the next two sections the model to constraint generalizations in LKB by means of DCH information will be defined aiming to reduce the overall ambiguity and detect the correct LKB sense assignments(s) to DCH elements.

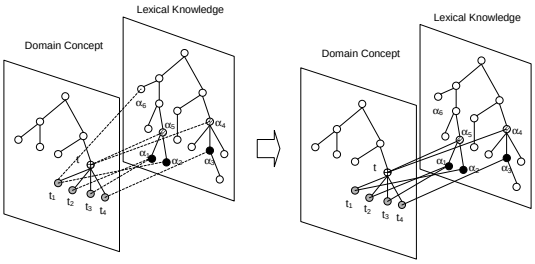


Figure 1: Integration of domain and general-purpose knowledge

2.1. Inspiring principles

One of the aims of the proposed integration is to constraint the search for word sense assignment (i.e. navigation in the LKB) through information provided by the domain resource. Vice versa the LKB structure will be used to bias the search of DCH meanings, i.e. explain linguistically the nature of the DCH primitives: for example *Cardiovascular System* has just one sense in WordNet, under the "*body_part*" sub-hierarchy; however, as a MeSH topics, it is also related to functionalities and physiological processes not coded as "*body_part*"s.

Cross-corresponding concepts between a DCH and a lexical model LKB can be detected by exploiting in combination both constraints. We will rely on the following two principles:

(**P1**) (*Extensional Nature of DCH*). Given a domain concept hierarchy DCH, whatever the nature of its basic

[1] Notice that in this case we could relax the search of the multiword expression, e.g. *Common Hepatic Duct*, and try to match senses for sub-expressions obtained by neglecting some modifier, e.g. *Hepatic Duct*. The longest expressions corresponding to one LKB entry would be retained as a possible linguistic interpretation.

unit is, *subsumption throughout the hierarchy has always an extensional interpretation*, i.e. for each couple of concepts C' and C'' subsumed by a common ancestor C in DCH, there is always a linguistically consistent concept $\alpha \in LKB$ such that the linguistic expressions $t' = t_{C'}$ and $t'' = t_{C''}$ have senses $\alpha_{t'}$ and $\alpha_{t''}$ both subsumed by α in LKB[2].

(**P2**) (*Intentional strength in LKB*). A set of linguistic denotations $W = \{w_i\}$ [3] whose senses are all subsumed by a given $\alpha \in LKB$ has an *intentional strength* for W that is a function of the senses of w_i and of their distribution in the LKB sub-hierarchy dominated by α. α represents the trade-off between the generalization required to represent all the denotations w_i and their specialization, i.e. the capability of separating the individual different senses of the w_i's. Any monotonic non-decreasing function of such a trade-off is a valid measure of the intentional strength of α with respect to words w_i.

DCH nodes C are in a many-to-many mapping to LKB senses. As a consequence sets $W = ext(C)$ may not receive a unique $\alpha \in LKB$ but are usually covered by more than one generalization (i.e. there is no α that is a common ancestor for all $t \in W$, implying that the intentional strength is 0). In this case an alternative can be found by partitioning W in more coherent (and possibly overlapping) subsets W_i. These will give independently rise to common generalizations, α_i: each one is a trade-off as the higher in the hierarchy is α_i, the larger is the size of the corresponding W_i.

2.2. Mapping domain concepts to lexical senses

The aim of the mapping between DCH and LKB is to find the correct attachment site of each DCH concept in LKB. This can be a complete equivalence whenever the DCH concept is not represented in the LKB or a "cross-generalisation" among the two hierarchies for partially represented concepts (see Fig. 1). Properties **P1** and **P2** drive the mapping in the following way. A domain concept C receives the minimal set of word senses in $lgen(C)$ with the *maximal intentional strength* as subsumers of non-trivial subsets of the linguistic extension of C, i.e. $ext(C)$. Usually specific entries in DCH (e.g. *Tissues*) are mapped into one ore more LKB senses (e.g. *'body_part'* and *'epithelium'* in WordNet). Vice versa one sense may be tagged by several DCH primitives (e.g. *'body_part'* as *'Digestive System'*, *'Cardiovascular System'*, *'Tissues'*, ...).

In our model the notion of *conceptual density* (cd), as introduced by (Agirre and Rigau, 1996), is used as a measure of intentional strength (principle **P2**). The conceptual density aims to state why and how much a set of words should be considered similar according to a reference lexical hierarchy, LKB[4]. Given a set W of words (eventually with multiple senses) and a specific node α in the lexical hierarchy dominating at least one sense for each $w \in W$, the conceptual density $cd^W(\alpha)$ is a real value associated to the common ancestor α: it is proportional to the number of covered senses of $w \in W$ and inversely proportional to the size of sub-tree rooted at α. Therefore, the smaller the sub-tree (i.e. the more specific is α as a generalization of the senses of w's), the higher is the cd value. Although its application in the ontology engineering framework proposed in this paper is new, this measure has been widely applied to word sense disambiguation problems ((Basili et al., 2004)).

The model we propose here requires that a triggering set T of DCH concepts C (with category labels t_C) has been previously selected. This set will drive the application of the principle **P1** and **P2** over the DCH. Then, for each concept $C \in T$, the corresponding set of linguistic expressions ($ext(C)$ in Eq. (1)) is determined by DCH. $ext(C)$ and the conceptual density are then used to derive an *optimal* set of LKB senses within the linguistic generalizations of C (i.e. $lgen(C)$ in Eq. (2)): this set is optimal as it is made of the intentionally strongest senses α_i that generalize all the expressions of $t \in ext(C)$. By means of a greedy technique, the generalizations α_i of non-trivial subsets $W_i \subset ext(C)$ are selected according to decreasing values of conceptual density until the entire set is not completely covered. In this way, each $C \in DCH$ is mapped to an $\alpha_C \in lgen(C)$ characterized by the highest intentional strength (i.e. $cd()$). More details can be found in (Basili et al., 2004)

The algorithm that maps the DCH into the LKB is triggered by the subset of concepts T and is sketched in the following.

procedure merge(DCH,LKB,T)
for each $C \in T$
  (Step 1) Determine the linguistic extensions $lgen(C)$
    in DCH made of all descendants of C
  (Step 2) Compute the optimal mapping $G(C) \subset lgen(C)$,
    by a greedy selection that maximizes conceptual density
  (Step 3) Attach t_C to senses in $G(C)$
  (Step 4) **for each** $t \in ext(C)$
    Attach t to $\alpha \in LKB$ iff:
      α is a sense for t in LKB and
      $\exists \beta \in G(C) | \beta$ subsumes α in LKB

The subset T of the domain concepts in DCH is therefore an input parameter. For example, the top levels of DCH can play the role of T[5]. To better explain the algorithm we

[2] The extensional interpretation α may not be unique. In fact, given a bipartite set of C descendants $\{C'_1, ..., C'_n, C''_1, ...C''_m\}$, then two (or more) concepts may exist, $\alpha' \neq \alpha''$, such that $\forall i = 1, ..., n \quad t_{C'_i}$ generalizes in α' and $\forall j = 1, .., m \quad t_{C''_j}$ generalizes in α''

[3] The use of w_i here emphasizes the difference with respect to the previously adopted notion of t_i. w_i are linguistic symbols that independently from any domain are referential in the world. t_i are terminological labels of DCH concepts and their semantics is NOT exhaustively determined on a linguistic ground. Principle **P1** focuses on the interpretation of domain symbols t_i by means of the DCH hierarchy. As **P2** focuses on purely linguistic information determined by LKB, a different notation is required.

[4] WordNet has been used as the underlying reference taxonomy for the definitions and experiments related to the conceptual density

[5] A limited semantic dictionary for which wide extensional evidence is available can improve the mapping accuracy

will make reference to the example in Fig. 1. Let us concentrate on the iteration that considers the concept $C \in T$ having $t_C = t$ (i.e. the node t of the DCH in figure). The following process is carried out: first linguistic expressions $t_i \in ext(C)$ of C descendants in the DCH hierarchy are determined in (Step 1), e.g. $ext(C) = \{t_1, ..., t_4\}$ in fig. 1. Linguistic descriptions t_i are analysed against the lexical semantic hierarchy LKB. Different subsets are derived $W_1 = \{t_1\}, W_2 = \{t_2, t_3\}$ and $W_3 = \{t_4\}$ as they receive different interpretations, i.e. activate senses $\alpha_1, ..., \alpha_6$. All elements t_i are "somehow" more specific of t_C. (Step 2) allows to select the optimal generalizations $G(C)$ as word senses. These are the valid generalizations of subsets of $ext(C)$ having the higher cd and covering the entire set $ext(C)$. In the example, $G(C) = \{\alpha_5, \alpha_4\}$ as they are enough general to represent all t_i and enough specific to refuse some useless senses. The DCH concept C is thus used to annotate senses α_5 and α_4 (Step 3). Finally, the linguistic labels of DCH concepts are attached to the related senses in the LKB (Step 4). It is easy to see that this information reduces the ambiguity of each t_i. For instance, the interpretation α_6 of t_1 is discarded: its conceptual density is too low and other senses are sufficient to cover the entire set $\{t_1, ..., t_4\}$.

The result of the above process is a *Concept Hierarchy* that links denotations of domain concepts to their linguistic counterparts: the former will support disambiguation in language processing, while the latter will favour linguistically consistent generalizations of general (i.e. non domain-specific) surface forms.

3. Mapping MeSH to WordNet: a case study

We investigated the power of the proposed algorithm over a *complex* framework: the mapping of a *not-so-structured* domain concept hierarchy, i.e. the Medical Subject Heading, to a principled linguistic knowledge base, i.e. WordNet. The fact that MeSH is not a proper *is-a* hierarchy can be shown by an example. Consider the term *dendrite* appearing three times in the MESH hierarchy, i.e. (1) "*Dendrites → Neurons → Nervous System*", (2) "*Dendrites → Cell Surface Extensions → Cellular Structures → Cells*", and (3) "*Dendrites → Neurons → Cells*". It should be noticed how the nature of the above arrows alternates between *part_of* and *is_a*: "*Dendrites* in fact *is_a Cell Surface Extensions*" while "*Dendrites* are *part_of Neurons*". Other technical relationships are also present in MeSH: for example, the term *Movement* is a very specific sort of *Physiological Mechanism* such that an improper is_a seems represented by the branch *Movement → Physiological Mechanism*. Given the extremely varying nature of the MeSH structure the mapping to WordNet appear as an extremely challanging task for the learning algorithm of Section 2.2.. If a reasonable performance can be obtained over this case, we can expect a significant accuracy in scenarios characterized by more principled domain concept hierarchies.

In order to demonstrate that the proposed algorithm can select or suggest the correct attachments (if any) of domain concepts to LKB nodes, a critical step is the definition of the golden standard. In the selection of the target MeSH concepts to be used in testing, we imposed on two properties:

Category	D26.664.255.165.810
Term	Suspension
Hierarchy Branch	Colloids; Dosage Forms; Pharmaceutical Preparations; Specialty Chemicals and Products;
Siblings	aerosol;emulsion;gel

Table 1: MeSH information for the term *Suspension*

(1) the term t denoting the concept should be represented in the LKB, i.e. at least one synset in the LKB should be activated by t; (2) t should be ambiguous in the LKB. The first constraint has been satisfied imposing that concepts whose surface forms are fully represented in WordNet have a corresponding synset. Even if complete representation of the surface form t does not completely guarantee semantic completeness (as the MeSH intended meaning for t can be missing in WordNet), this approximation helps to select a portion of concepts significant for testing. Constraint (2) is used to test the algorithm over non trivial test cases and also to measure its accuracy contrastively to the task complexity (i.e. number of senses). The application of the two constraints to MeSH, selects a set of about 1,500 MeSH nodes. This set will be hereafter referred as the *Represented and Ambiguous* ($ReprAndAmb$) set of terms.

The selected portion of the hierarchy has been given to 3 annotators asked to manually assign to each MeSH term t (as observed in its hierarchy branch *B1:B2:...:Bn*), the correct WN synset s_1,...,s_n. The additional information shown to annotators for deciding is the set of siblings of the target term t in MeSH (see Tab. 1 for an example). Although the full test is still on going, here we present partial results achieved at the time of this submission. The portion analysed by all the three annotators consists currently of a set of 500 concepts, hereafter called $Annotated$ set. The inter-annotator agreement on it is 90.2%.

In the validation we exploited also the probabilistic interpretation of the mapping algorithm of Section 2.2.. Given a term t in $ext(C)$ and given one of its senses s (i.e. $senses(t) \in WN$), let us define $cd(\sigma) = \sum_{\alpha \quad generalises \quad \sigma} cd^{ext(C)}(\alpha)$ where α are generalisations of senses σ in Wordnet as found by the algorithm of section 2.2. An estimate of the probability $P(s|t)$, that the method assigns to a sense s, can be thus defined, as a posterior probability, by:

$$post\_prob(\sigma) = \frac{cd(\sigma)}{\sum_{\alpha \in senses(t)} cd(\alpha)} \qquad (3)$$

The posterior probability should be compared with the prior probability $prior\_prob(\sigma)$, computed as $prior\_prob(\sigma) = \frac{1}{|senses(t)|}$.

Tab. 2 reports figures about the initial complexity of the task in the polisemy and entropy scores of the *a-priori* column. The reduction in the overall ambiguity, implied by the method, is reported in the *a-posteriori* column. The polysemy and entropy are both averaged over the two different test sets, $ReprAndAmb$ (1,500 cases) and $Annotated$ (500 cases). The average entropy is computed considering the probabilistic interpretation of the terms-to-senses map-

		a-priori	*a-posteriori*
$ReprAndAmb$	Avg. Polysemy	3.22	2.827
	Avg. Entropy	1.039	0.448
$Annotated$	Avg. Polysemy	3.20	2.846
	Avg. Entropy	1.026	0.444

Table 2: Preliminary analysis

	All Senses	Only Best Senses
Recall	0.920	0.700
Precision	0.347	0.630
F-measure($\alpha = 0.5$)	0.504	0.663
Avg. Information Score	0.680	

Table 3: Results on the annotated portion

ping and it suggests that (even in a technical domain as in MeSH) terminological expressions are a priori quite ambiguous (over 3 senses on average). After the application of the method, the average polysemy does not fall dramatically as not so many senses are pruned out. However, lower probabilities are given to most of them as the fall in the overall entropy, caused by more skewed distributions, suggest. The *a posteriori* distribution seems to assign to few senses most of the probability, i.e. high degree of confidence. Unfortunately, entropy does not give much information about the method's accuracy, so other figures (e.g. recall) are needed.

Tab. 3 describes the performances of the algorithm over the $Annotated$ set. Recall, precision, and f-measure are reported for two decision strategies: (1 - Second Column) *All Senses* where all the possible senses receiving non zero probability are considered as potential solutions and (2 - Third Column) *Only Best Senses* where only senses s for which $post\_prob(s) > prior\_prob(s)$ hold are retained.

The high values of recall show that the algorithm almost always prefers the correct sense although the low precision suggests that, without imposing any threshold to probability scores, too many senses still survive. When another strategy is used to decide the mapping is still effective with an acceptable precision: an increase of the f-measure of around 16% is observed, mainly justified by the doubling of the precision score. In order to better study the adherence of the posterior probability (i.e. the system preferences) to the oracle, we computed the average information score, as early introduced in (Kononenko and Bratko, 1991). This index measures the increase of information between prior (i.e. uncertain) and posterior probability. It is better suited for multiclass classification tasks based on a probabilistic model assigning rewards when increase of posterior probability is observed for the correct decisions[6] and penalties when posterior probs are higher for wrong classifications. The value of information score indicates the number of bits gained by the posterior probability in coding the correct decision model. A positive values expresses how many useless senses (i.e. wrong) have been removed with respect

[6]or dually when lower posterior probabilities characterize wrong decisions.

WNSynset	*Gloss*	P_{prior}	P_{post}
12923333	a time interval during which there is a temporary cessation of something	0.143	0.002
12316056	a mixture in which fine particles are suspended in a fluid where they are supported by buoyancy	0.143	0.915
773686	the act of suspending something (hanging it from above so it moves freely); "there was a small ceremony for the hanging of the portrait"	0.143	0.007
3805196	a mechanical system of springs or shock absorbers connecting the wheels and axles to the chassis of a wheeled vehicle	0.143	0.071
163152	a temporary debarment (from a privilege or position etc)	0.143	0
6132706	an interruption in the intensity or amount of something	0.143	0.005
11815922	temporary cessation or suspension	0.143	0

Table 4: A *good* example: synsets, prior and posterior probability for the term *Suspension*

to the prior distribution. Notice how the value obtained against the $Accepted$ oracle (0.68) is more or less equal to the decrease of entropy in the same set (as reported in Table 2): this suggests that, in general, those senses removed by the algorithm (or, better, considered unlikely) are in fact the wrong ones. As an example of the resulting system behavior the posterior probability for the term *Suspension* is the presented in Tab. 4 where it can be seen that the algorithm selects the correct sense giving the highest probability (0.915). Senses receiving 0 as posterior probability are the ones filtered out by the greedy algorithm calculating the conceptual density and the minimal coverage set.

4. Conclusions

In this paper we discussed the role of lexical information in ontology engineering and presented a method for mapping an underlying domain concept hierarchy into the entries of a lexical semantic resource. A large scale empirical investigation is on going within a complex scenario, i.e. the mapping between a medical terminological hierarchy and WordNet. The results reported on a non trivial subset show that the proposed method is very effective. It can serve as a viable approach to the building of a domain semantic dictionary, i.e. as a first step in a language-driven ontology learning process.

5. References

Agirre, Eneko and German Rigau, 1996. Word sense disambiguation using conceptual density. In *Proceedings of the 16th International Conference on Computational Linguistics*.

Basili, R. and M.T. Pazienza, 1997. Lexical acquisition for information extraction. In M.T. Pazienza (ed.), *Information Extraction a multidisciplinary Approach to an emerging Information Technology*. Springer-Verlag - Lecture Notes.

Basili, Roberto, Marco Cammisa, and Fabio Massimo Zanzotto, 2004. A semantic similarity measure for unsupervised semantic disambiguation. In *Proceedings of the Language, Resources and Evaluation LREC 2004 Conference*. Lisbon, Portugal.

Fensel, Dieter, Frank Van Harmelen, and Ian Horrocks, 2003. Oil and daml+oil: Ontology languages for the semantic web. In John Davies, Dieter Fensel, and Frank Van Harmelen (eds.), *Towards the Semantic Web*. West Sussex, UK: John Wiley and Sons.

Gruber, T.R., 1993. Towards principles for the design of ontologies used for knowledge sharing. In N. Guarino and R. Poli (eds.), *International Workshop on Formal Ontology*. Padova, Italy.

Kononenko, I. and I. Bratko, 1991. Information-based evaluation criteria for classifier's performance. *Machine Learning*, 6(1):67–80.

Lee, T.B., 2001. The semantic web. In *Scientific American*, volume May.

MESH (ed.). *Medical Subject Headings*. www.nlm.nih.gov/mesh/meshhome.html.

Miller, George A., 1995. WordNet: A lexical database for English. *Communications of the ACM*, 38(11):39–41.

Ontological Knowledge and Language in Modelling Classical Architectonic Structures

Amedeo Cappelli, Emiliano Giovannetti, Patrizia Michelassi
Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo",
Knowledge Discovery and Delivery Laboratory - CNR
Via G. Moruzzi 1, 56124, Pisa, Italy
{Amedeo.Cappelli, Emiliano.Giovannetti, Patrizia.Michelassi}@isti.cnr.it

Abstract

This article will concern the specification of the conceptual and linguistic constraints for the construction of a knowledge base in classical architecture, an operation that involves the structuring of the concepts of the domain to which appropriate linguistic terms must be associated. Our approach will take into account models of conceptual dictionaries proposed in computational linguistics as well as knowledge representation and ontological formalisms proposed in artificial intelligence and ontological engineering.

1. Introduction

This article will concern the specification of the conceptual and linguistic constraints for the construction of knowledge bases to be used for applications in human language technology. We shall focus on a particular domain, the representation of classical architectonic structures, for which a knowledge base has been developed.

The construction of a knowledge base is an operation that involves the structuring of the concepts of a domain to which appropriate linguistic terms must be associated – this is often called an ontology.

As a principle, we assume that the knowledge base must exhibit a high degree of clearness, coherence, and correctness mandatory to develop applications involving an advanced treatment of its content, as required by many knowledge based applications. These characteristics can be obtained by controlling the process of creation, either human or automatic, by imposing a set of integrity constraints regarding the ways of structuring concepts and associating terms to them.

Our approach will take into account certain models of conceptual dictionaries proposed in computational linguistics as well as knowledge representation and ontological formalisms proposed in artificial intelligence and ontological engineering. We are confident that the representation of the lexicon will benefit from the integration of different methodologies capable of providing more insight about the complex relationships between lexicon and knowledge.

The relation between language and knowledge is one of the major problems studied for years in linguistics, psychology, philosophy and, recently, in computational linguistics and artificial intelligence, in particular, in knowledge representation and ontological engineering.

In cognitive science, the distinction of the reality into classes of objects is used in order to study the human process of acquisition of knowledge (Keil, 1989). Part-whole relations have been investigated in order to account for the conceptual processes underlying linguistic terms used for expressing the concept of "part". The result of this analysis has been the specification of a taxonomy of part-whole relations and of the logic underlying them (Winston et al., 1987; Cruise, 1979).

Computational linguistics is interested in finding a global organization of the lexicon into classes related to each other, in order to represent word meanings and to improve natural language understanding systems. Different approaches have been adopted which combine linguistic, cognitive and lexicographic aspects. However, the results are sometimes far from being coherent with clear logical and ontological assumptions (Hirst, 2004). The methods underlying certain conceptual dictionaries, like WordNet (Miller et al., 1990; Vossen, 1998) and Dicologique (Dutoit, 1992), are the results of a number of investigations trying to integrate multidisciplinary issues in the representation of the lexicon

The design of knowledge representation formalisms frequently integrates conceptual and linguistic considerations. As an example, the semantic network has been designed to represent word meaning and knowledge representation languages based on this formalism are relevant tools for the representation of the semantic aspect of the lexicon (Brachman and Schmolze, 1985; Caligaris et al., 1992; Cappelli & Mazzeranghi 1994; Patel-Schneider et al., 1996; Woods and Schmolze 1992).

Recently, a new generation of knowledge representation languages has been introduced, in which general abstract means for structuring knowledge can interact with a set of ontological constraints regarding the inner content of concepts. As an example, OWL is a language for defining structured, Web-based ontologies which can enable richer integration and interoperability of data across application boundaries. With OWL it is possible for information contained in documents to actually be processed by applications, rather than just presented to humans. OWL facilitates greater machine interpretability of Web content than that supported by XML, RDF, and RDF Schema (RDF-S) by providing additional vocabulary along with a formal semantics. OWL has three increasingly-expressive sublanguages: OWL Lite, OWL DL, and OWL Full (Bechofer et al., 2004).

Ontology is a fundamental field of research, critical of many advanced applications in computer science and information science, but also in medicine, education, and industries.

In the "Stanford Glossary of Ontology Terms", ontology is "an explicit specification of some topic, … a formal and declarative representation which includes the vocabulary (or names) for referring to the terms in that subject area and the logical statements that describe what terms are, how they are related to each other, and how they can or cannot be related to each other. Ontology therefore provides a vocabulary for representing and communicating knowledge about some topic and a set of relationships that hold among the terms in that vocabulary".

In this definition, philosophical intentions are as important as the conceptual and linguistic modelling of a specific domain for practical applications. Conceptual and linguistic modelling are important for representing and communicating knowledge, in other words, for implementing knowledge-based systems more immediately applicable to industrial problems.

Following Gruber (1995), "An ontology is a formal explicit specification of a shared conceptualization". For Gruninger and Lee (2002), this notion of conceptualization refers to an abstract model of how people think about things in the world, usually restricted to a particular subject area. In this way, an ontology can be intended as a formal representation of a domain, instead of as a formal characterization of what exists in the world.

A philosophical and clear notion of ontology is invoked by Guarino and Welty (2000; 2002a; 2002b), who define ontology: a discipline of philosophy that deals with what *is*. Recognizing that "the accepted industrial meaning of 'ontology' makes it synonymous with 'conceptual model' and is nearly independent of its philosophical antecedents", they draw the distinction between "conceptual model", as an "actual implementation of an ontology that has to satisfy the engineering trade-offs of a running application", and "ontology design", whose only goal "is to specify the conceptualization of the world underlying such application" and is "independent from run-time considerations". In this perspective, they try to specify a methodology (OntoClean) based on formal philosophical notions general enough to be used independently of a particular domain. In other words, formal notions are used to define a set of metaproperties, which are used to characterize relevant aspects of the intended meanings of the properties, classes, and relations. Formal notions, such as identity, essence, unity, and dependence are defined in order to specify a logical framework to make an intended meaning of a taxonomy more explicit.

Because of the interest for ontology in many application sectors, a new discipline has emerged, ontological engineering, whose goal is to investigate the entire ontology life cycle, which is composed of the following steps: design, evaluation, validation, and revision (Holsapp & Joshi, 2002).

A formal methodology for the entire life cycle of ontology building is Methontology (Blazquez et al., 2002). The aim of this methodology is to bridge the gap between how people think about a domain and the language into which the ontology is formalized. Intermediate representations, whose conceptual model is implicit, are constructed and translated into a coherent ontology.

The possibility to share and reuse ontology for different applications, directly or with minor modifications, is another goal of ontological engineering. Besides, sharability and reusability are considered two important characteristics of ontology built for industrial applications.

Our model has been specified for a specific domain. However, we are now investigating its applicability to other domains, i.e. legal, economical and social security for which we have developed applications in conceptual information retrieval and in text generation. The comparison of the result of the testing of the model on these fields, which have very different characteristics (architecture is characterised by functional artefacts, the others on nominals), will help in finding the variant and the invariant functional and material characteristics of each domain.

This problem is not only fundamental from a theoretic point of view, but it is also very relevant for applicative purposes. The solution of this problem should suggest a realistic view of reusability, especially in the development of efficient and robust methodologies for the total or partial modification of the body of knowledge for new knowledge based applications.

2. Description of the Domain

The modelling of the classical architectonic structures consists in the representation of very articulated and fine-grain descriptions of complex artefacts in accordance with a very subtle degree of granularity and by using descriptive parameters concerning, among others, the morphology of the composition of the subparts, their form, the material they are made of, and their function (Allsopp, 1965).

This domain is composed of complex artefacts with a precise identity.

A rich terminology, progressively specialized and structured following a rich tradition of classification and interpretation studies, enables us to precisely refer to the subtle descriptive distinction among objects, their functions, and their uses. The lexicon is then strictly related to the conceptual aspect of the domain, since it has been modelled in accordance with the "observable" structure of the artefacts. In other terms, by structuring the lexicon, it is also possible to account for the majority of the knowledge of the domain, since lexical items precisely refer to specific classes of objects or to specific descriptive parts of the objects themselves.

Let us introduce an example in order to highlight some characteristics of the domain. In classical architecture, the structure of the temple is composed of three parts: *stylobate*, *colonnade*, and *entablature*.

The *colonnade* is a range of columns.

The *column* is composed of a *base*, a *shaft*, and a *capital*.

The *base* is the lowest member of a column and therefore usually appears only in the Ionic and Corinthian order, rarely also in the Doric.

So, the *Doric column*, has, in general, no base and is composed only of a shaft and a capital.

The *shaft* is the main body of a column or a pier, in general, which is between the base and the capital. In the most ancient buildings, monolithic shafts can be found, but in general, in the classic period, the shaft was composed of several *drums*. A *drum* is one of the cylindrical sections or courses of a column shaft. A shaft also has some *flutes*, vertical channels, segmental, elliptical, or semicircular, in a horizontal section. The flutes, twenty in general, in the classical period, were separated one from the other by an *arris* in the Greek Doric and early Ionic orders, and by a *fillet* in the developed Ionic and Corinthian orders. In Doric columns, the flute was usually segmental, or in order to emphasize the *arris*, it was formed of three arcs constituting what is known as false ellipse. A deeper curve was given to the

flutes in Greek Ionic and Corinthian columns and, in later work, the flute was semicircular. In rare examples, the flutes were carried spirally round the columns.

In the flutings of the Doric column, the *arris* was present: a sharp edge formed by two surfaces meeting at an external angle.

From this sequence of descriptions taken from the literature, certain regularities can be extracted, regarding both the nature of the objects and their mutual relationships.

The objects of the domain have an intrinsic structure made of parts described following precise descriptive parameters. The distinction between the objects and their assignment to well-defined classes are performed by the evaluation of specific descriptive parts and modalities, which vary in accordance to structural, historical, and cultural parameters. The domain is thus structured in classes and subclasses, which generalize descriptions of specific objects.

Due to this very cohesive conceptual organization, the lexicon is strongly structured in a rich technical terminology; its terms precisely refer to the objects of the conceptual organization. Objects and parts are univocally identified on the basis of subtle distinctions and have specific names. This helps in individuating singular descriptive characteristics in the vast variability of the artefacts, which can correspond to a specific sign of a period, school, or stylistic movement.

3. Formal Model of Knowledge

Starting from this representation, a formal model of the organization of knowledge has been specified, which explicitly accounts for all inherent characteristics of the knowledge. This model integrates knowledge representation techniques, lexical representation tools, and ontological engineering techniques which, together, contribute to the formal representation of the taxonomy of objects, of the association between objects and lexical terms, and of the typology of properties which describe objects as the grammar in Figure 1 shows (Cappelli et al., 2003).

3.1. Epistemological Parameters

We consider epistemology as the specification of certain basic means to structure knowledge independently of any content. In other terms, they constitute a general abstract grammar to organize knowledge (Brachman, 1979). The basic data structure of our representation is the concept, which aggregates information concerning its description, which is realized by the specification of its local descriptive subparts and its collocation inside the terminology.

A concept is an intensional representation of a class and has a structure, as shown in the following. A concept has a unique identifier, which unambiguously identifies it in the map and can aggregate a list of terms in different languages, which are the synonyms with respect to the concept. Concepts can be related the one to the other in order to specify their topological position inside the conceptual map, in terms of:

1. Superconcept. Between two concepts belonging to the same inheritance chain, one of which is more "general" than the other (*column* / *Doric column*); once inserted in the generalization chain, a concept follows the logic of subsumption
2. Thematic. Between two concepts associated by a sort of "point of view" relation, which cannot be defined in terms of a precise logic; thematic relations can be used for establishing relationships between different "semantic fields" for instance, the fact that the Doric order is characterized by the Doric column is represented putting in relation "Doric order" and "Doric column" by the relation "characterized by" which is not a clear meaning to be specified in terms of a precise type of semantics.

<Concept> → <Concept identifier> <Synonyms> <Superconcept>
<Thematic relation> <Descriptive parts> <Glossa>
<Concept identifier> → <Integer>
<Synonyms> → <Terms>*
<Terms> → <Lexicalized terms> <Non-lexicalized terms>
<Lexicalized terms> → <Word>
<Non-lexicalized terms> → <Extracted terms> <New categorizations>
<Extracted terms> → <Idioms>
<Idioms> → <Multiword>
<Multiword> → <Word> <Word>+
<New categorization> → <Expression>
<Word> → <String of characters>
<Expression> → <Text>
<Superconcept> → <Concept identifier>
<Thematic relation> → <Label> <Concept identifier>
<Label> → is characterized by | is studied by | . . .
<Descriptive parts> → <Non meronimy parts> <Meronimy parts>
<Non meronimy parts> → <Part name> <Concept identifier>
<Part name> → form | aim | stuff | . . .
<Meronimy parts> → <Components> <Place>
<Components> → <Concept identifier> <Parameters>* <Nexus>
<Parameters> → <Descriptive parameters> <Structural parameters>
<Descriptive parameters> → <Cardinality> <Dimensions>
<Cardinality> → <Cardinality label> <Integer>
<Cardinality label> → atleast | atmost | exactly
<Dimension> → <Dimension label> <Measure>
<Dimension label> → height | depth | length | diameter | . . .
<Measure> → <Real>
<Structural parameters> → <Function> <Position>
<Function> → <Predicate> <Arguments>*
<Predicate> → carry | decorate | link | channel | throw | . . .
<Position> → <Preposition> <Arguments>*
<Prepositions> → upon | under | between | in | behind | in front of | . . .
<Arguments> → <Meronimy parts>
<Place> → <Concept identifier> <Nexus>
<Nexus> → <Concept identifier> <Chain>
<Chain> → <Concept identifier> <Meronimy parts>
<Glossa> → <Text>

Figure 1: Formal grammar

3.2. Linguistic Parameters

Terms can be lexicalized or not lexicalized; for lexicalized we intend words (*temple*) present in a dictionary and multiwords (*Doric column*), which correspond to significant co-occurrences of words found in the literature. In this way, a concept represents one-word meaning (*column*) or that of a multiword expression (*Doric column*). Non-lexicalized terms are those sequences of words used as names of new concepts, which correspond to conceptualizations used for introducing technical and sharable distinctions, i.e., "the Doric Temple in B.C. 400".

3.3. Ontological Parameters

Concepts are described by the declaration of their parts, which are related to the following concepts in accordance to the following types of links:

- Meronimic, which follow the logic of meronimy;
- non-meronimic, for all the others.

This distinction is very shallow, but it enables one to clearly separate those parts that can be manipulated in accordance with a well-defined logic and those which cannot.

Non-meronimic parts are those that are not constrained by a specific type of logic. They declare an association between a concept and another concept as one of its proper descriptions.

They can be used in order to explain certain standard properties, such as, for instance: the form (*a cover tile is normally semicircular or triangular*), the stuff (*the cover tile is made of terracotta or marble*), or the aim (*a palaestra is a training school for physical exercises*). They can also be used to express any other properties with no precise semantics, such as, for instance, the date of a building, its position in a catalogue, etc.

Meronimic parts follow the meronimic logic. We give examples about two types of meronimic relationships: component/object, which covers all the rich typology of structural descriptions of architectonic structures, and place/area, which allows us to distinguish, as an example, between buildings of the same type built in different areas (the Doric temple in Greece and in Sicily).

The place/area type has not yet been structured: only a nexus with a part of the same type, belonging to another concept is specified, since it follows the logic of transitivity.

Component/object type can have some descriptive and structural parameters. The descriptive parameters are:

- Cardinality: used to explain a fact, such as, for instance, *that a shaft bears sixteen flutes*.
- Dimensions: used to explain the dimension of a part in terms of a metric measure (*a column is four meters high*).

Structural parameters are used to express structural relationships between parts of concepts and realize a simplified notion of structural description of classical KL-One. They express:

- Position: to define the relative position of a part with respect to other parts, for instance: *the architrave is carried from the top of one column or pier to another*; the position is expressed by using preposition and declaring some parts as their proper arguments.
- Function: to define the role of a part in its relationship with another, as *in a pier has the function of carrying an entablature or arch*; the function is expressed by using predicates with their arguments.

Given that meronimic parts follow the logic of meronimy, they can be linked to each other in order to create long-distance association chains between descriptive parts of concepts to be exploited by a meronimic reasoning process. This enables us to compute the well known meronimic syllogism based on transitivity, for instance, if a frieze is part of an entablature and an entablature is part of a temple, then a frieze is part of a temple.

4. Conclusions and Future Works

This model has been applied in the creation of a knowledge base in archaeology starting from a list of terms extracted from glossaries and specialized texts. A question arises whether or not this model could be successfully applied to other domains. So, we are currently investigating the application of the model to politics, social security, law, economics and space science, in which we have already produced knowledge bases in accordance to other models not so deeply formalized as the one presented in this paper.

By comparing the work for creating a conceptual dictionary about politics and social security (Bagnasco et al., 2000), and the organization of knowledge for the development of a system for the semi-automatic generation of legal contracts, in which we have developed a grammar which integrates textual (structure of the document), linguistic (lexicon and syntax), and conceptual parameters, certain preliminary conclusions can already be drawn.

The distinctions we have introduced prove to be functional for the individuation of the generic and invariant aspects of a domain and for the specification of the means used for the creation of the concepts and of their relative lexical realizations.

As we have already noted, in classical architecture, a rich terminology has been created which enables one to precisely refer to the subtle descriptive distinctions among objects, their functions, and their uses. In other words, terms have been created by multiplying words or by applying morphological processes. In social security and, partly, in law, concepts are created by using synonymy or quasi-synonymy and by the creation of multiwords, through the variation of syntactic connectors (complex noun phrases).

Epistemological parameters appear to be very relevant for the dynamic and flexible construction of the topological structure of knowledge. Besides, being the most abstract part of the grammar for representing knowledge, they proved to be useful for the representation of those conceptual processes regarding, for instance, the organization of the world from the legislator's point of view, in particular, in definitions, attributions of rights, duties, etc.

Concerning the ontological aspects, some distinctions we have introduced, especially between meronimic and non-meronimic parts, proved to be particularly important, due to the relevant presence of nominal objects in politics and law. In these domains, the tendency is to make use of shallow associations between parts and whole not subjected to a precise type of logic.

This aspect should be more deeply explored, in particular for investigating the possibility of individuating abstract classes of concepts characterized by specific types of properties. This has been the aim of many approaches to ontology, in particular in the specification of the structure of a general and standard "top level" which, in our opinion, sometimes appear to be too rigid. We are now investigating the problem by using, as a paradigm, the flexible creation of concepts as permitted by new knowledge representation languages.

5. References

Allsopp, B. (1965). A History of Classical Architecture. London: Sir Isaac Pitman & Sons Ltd.

Bagnasco, C.; Cappelli, A. & Magnini, B. (2000). A Dialogue Environment for Accessing Public Administration Data: The TAMIC-P System. In Proceedings of the ECAI2000 (pp. 686--690). Amsterdam: W. Horn (Ed.): IOS Press.

Bechofer, S.; van Harmelen, F.; Hendler, J.; Horrocks, I.; McGuinness, D.; Patel-Schneider, P. F. & Stein, L. A. (2004). OWL Web Ontology Language Reference. W3C Reccom., http://www.w3.org/TR/owl-ref/.

Blazquez, M.; Fernandez, M.; Garcia-Pinar, T. M. & Gomez-Perez, A. (2002). Building Ontologies at the Knowledge Level Using the Ontology Design Environment, Laboratorio de Inteligencia Artificial, Universidad Politecnica de Madrid.

Brachman, R. J. & Schmolze, J. G. (1985). An Overview of the KL-ONE Knowledge Representation System. Cognitive Science, 9(2), 171--216.

Brachman, R. J. (1979). Epistemological Status of Semantic networks. In N. V. Findler (Ed.), Associative Networks: Representation and Use of Knowledge by Computers (pp. 3--50). New York: Academic Press.

Caligaris, C.; Cappelli, A.; Catarsi, M. N.; & Moretti, L. (1992). An Integrated Environment for Lexical Analyses. In Proceedings of COLING-92, vol. III (pp. 935--939). Nantes.

Cappelli, A. & Mazzeranghi, D. (1994). An Intensional Semantic for a Hybrid Language. Data & Knowledge Engineering, 12, 31--62.

Cappelli, A.; Catarsi, M. N.; Michelassi, P. & Moretti, L. (2003). Conceptual and Linguistic Constraints for the Construction of a Knowledge Base in Archaeology. Applied Artificial Intelligence, 17(8-9), 835--858.

Cruise, D. A. (1979). On the Transitivity of the Part-Whole Relation. Journal of Linguistics, 15, 29--38.

Dutoit, D. (1992). A Set-theoretic Approach to Lexical semantics. In Proceedings of COLING-92 (pp. 982--987). Nantes.

Gruber, T. (1995). Towards Principles for the Design of Ontologies Used for Knowledge Sharing. International Journal of Human-Computer Studies, 43(5-6), 907--928.

Gruninger, M. & Lee, J. (2002). Ontology Applications and Design. Communication of ACM, 45(2), 39--41.

Guarino, N. & Welty, C. (2000). Identity, Unity and Individuality: Towards a Formal Toolkit for Ontological Analysis. In W. Horn (Ed.), Proceedings of the ECAI2000 (pp. 219--223). Amsterdam: IOS Press.

Guarino, N. & Welty, C. (2002a). Evaluating ontological decisions with ONTOCLEAN. Communication of ACM, 45(2), 61--65.

Guarino, N. & Welty, C. (2002b). Identity and subsumption. In R. Green, C. A. Bean & S. H. Myaeng (Eds.), The Semantics of Relationships: An Interdisciplinary Perspective (pp. 111--125). Dordrecht: Kluwer.

Hirst, G. (2004). Ontology and the Lexicon. In S. Staab & R. Studer (Eds.), Handbook on Ontologies (pp. 209--230). Berlin: Springer.

Holsapp, C. W. & Joshi, K. D. (2002). A Collaborative Approach to Ontology Design. Communication of ACM, 45(2), 42--47.

Keil, F. C. (1989). Concepts, kinds and Cognitive Development. Cambridge, MA: The MIT Press.

Miller, G. A.; Beckwith, R.; Fellbaum, C. & Gross, D. (1990). Introduction to WordNet: An On-line Lexical Database. International Journal of Lexicography, 3(4), 235--244.

Patel-Schneider, P. F.; Abrahams, M.; Resnick, L. A.; Mcguinness, D. L. & Borgida, A. (1996). NeoClassic Reference Manual: Version 1.0., Artificial Intelligence Principles Research Department, AT&T Bell Labs.

Vossen, P. (1998). Introduction to EuroWordNet. In N. Ide, D. Greenstein, & P. Vossen (Eds.), Computers and the Humanities, 32(2–3), 73--89.

Wiston, M. E.; Chaffin, R. & Herrmann, D. (1987). A Taxonomy of Part-Whole Relations. Cognitive Science, 11, 417--444.

Woods, W. A. & Schmolze, J. G. (1992). The KL-One Family. Computers & Mathematics with Applications, 23(2–5), 133--177.

Comparison of Principles Applying to Domain-Specific versus General Ontologies

Bodil Nistrup Madsen, Hanne Erdman Thomsen, Carl Vikner

Department of Computational Linguistics, Copenhagen Business School

Bernhard Bangs Allé 17B, DK-2000 Frederiksberg, Denmark
{bnm, het, cv}.id@cbs.dk

Abstract

This paper presents three principles constraining the ontological structure permitted by our support system for terminological concept modelling, CAOS, namely: the principle of uniqueness of dimensions, the principle of uniqueness of primary feature specifications and the principle of grouping by subdividing dimensions. We demonstrate that these principles are not applied in the general ontology set up in WordNet and that the violations of the principles often reveal weak points in the analysis. The conclusion is therefore that these principles are probably relevant not only to domain-specific but also to general ontologies.

1 Introduction

We are building a tool for terminological ontology structuring[1], i.e. structuring of domain-specific ontologies. The backbone of this system is constituted by characteristics modelled by formal feature specifications, i.e. attribute-value pairs[2]. The functionality of the tool rests on certain principles governing the description of the relationships between concepts based on their characteristics. In our paper for LREC 2004[3] we give a presentation of ten principles. Some of these principles apply not only to domain-specific ontologies, but also to general ontologies such as WordNet or the SIMPLE is-a hierarchy, whereas other principles probably do not apply to general ontologies. In this paper we will discuss whether these principles apply also to general ontologies. We will focus on the following three principles, since these are the most interesting in a comparison between domain-specific ontologies and general ontologies:

- uniqueness of dimensions
- uniqueness of primary feature specifications
- grouping by subdividing dimensions

Some of the other principles, such as for example polyhierarchical structure and inheritance of feature specifications, are well known, and should apply to both terminological and general ontologies. Some other principles are specific to the technical functionality of the CAOS system and will not be discussed here.

In order to illustrate the principles of terminological ontologies we present some examples, which have been developed for CAOS. Concerning the principles of general ontologies we have chosen some examples from WordNet, which is one of the most outstanding ontologies for general language, cf., for example, Gómez-Pérez et al. (2004: 79).

[1] CAOS – Computer-Aided Ontology Structuring (cf. www.id.cbs.dk/~het/idterm/CTO/caos/index.html).

[2] This approach to the modelling of characteristics was proposed in Thomsen (1998, 1999) and Madsen (1998).

[3] Principles of a system for terminological concept modelling (to be presented at LREC, Lisbon, May 2004).

2 Uniqueness of dimensions

2.1 Dimensions and feature specifications

We distinguish two kinds of feature specifications: primary and inherited. A primary feature specification is assigned directly to a given concept, whereas an inherited feature specification is inherited from the concept's superordinate concepts. In CAOS attributes and values of all the primary feature specifications of subordinate concepts must be registered on the superordinate concept. This is done by creating dimension specifications on the mother concept in question consisting of a dimension and a list of values:

(DIMENSION : [value1, value2, ...]).

The principle of uniqueness of dimensions says that a given dimension may only occur on one concept in an ontology. This means that primary feature specifications with the same attribute must always occur on sister concepts.

Uniqueness of dimensions contributes to create coherence and simplicity in the ontological structure because concepts that are characterized by means of primary feature specifications with the same dimension must appear as coordinate concepts on the same level having a common superordinate concept.

Figure 1 below shows an extract of a terminological ontology for printers, where the dimensions are shown by means of boxes covering the relevant branches.

The dimension CHARACTER TRANSFER occurs only on the concept *printer*, i.e. only the coordinate concepts *impact printer* and *nonimpact printer* are distinguished by means of primary feature specifications comprising this dimension.

However the dimension OUTPUT is not unique in the hierarchy shown in figure 1, since it occurs both on the concept *printer* and on the concept *high-speed printer*. When the user tries to insert [OUTPUT : page-per-page] on the concept *high-speed page printer*, CAOS will report that OUTPUT is a dimension already found on the concept *printer*, and therefore the attribute OUTPUT can only be part of primary feature specifications on daughters of *printer*. The user can then choose to introduce a new concept *page printer* as a daughter of *printer*, so that *high-speed page printer* - in a polyhierarchical structure - can inherit the feature specification [OUTPUT : page-per-page]

from this new concept. This will result in a structure like the one in figure 2.

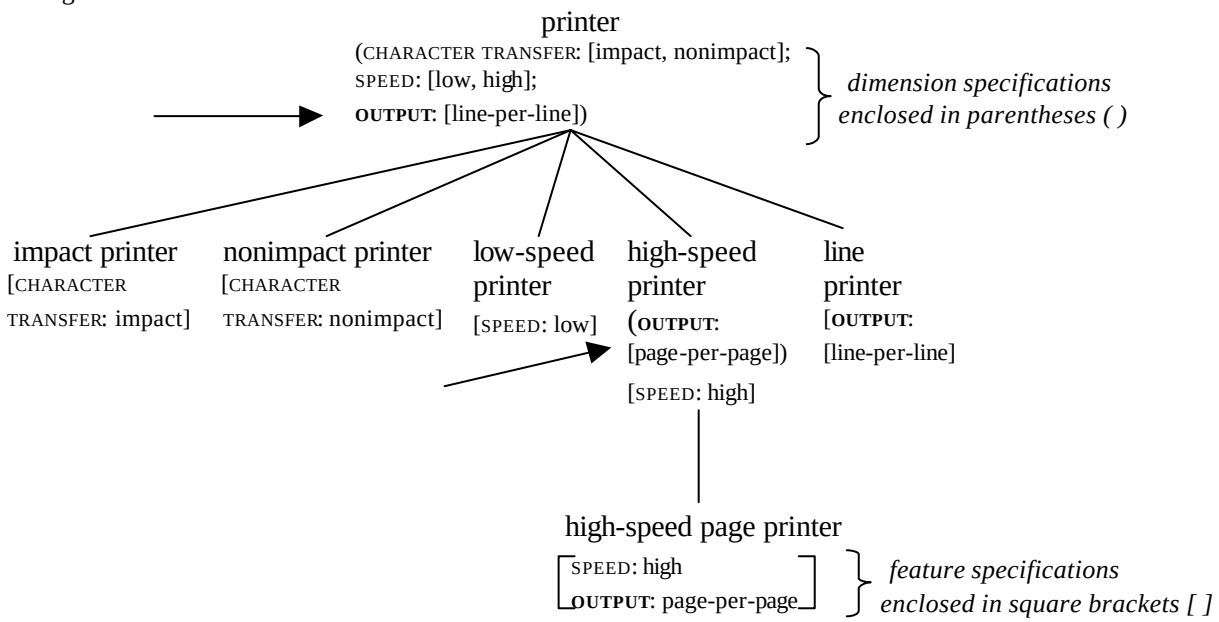


Figure 1: Extract of a terminological ontology for printers (one dimension not unique)

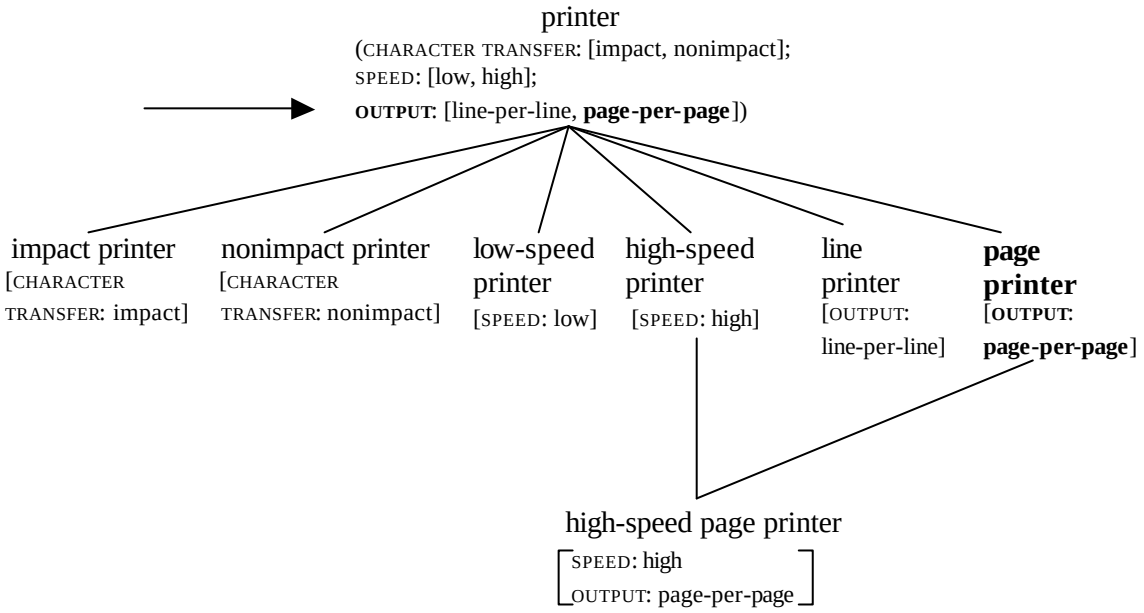


Figure 2: Extract of a terminological ontology for printers (all dimensions unique)

Sometimes it is difficult to comply with the principle of uniqueness of dimensions, because it will feel natural to use the same attribute in different places of the ontology, also where it is not possible to resolve the problem by establishing polyhierarchical inheritance. For instance, in the description of an ontology for printers, a need might arise to introduce the attribute SIZE both in dealing with different printer types and in dealing with different paper types. A problem like this can be solved by using more specific attribute names, e.g. PRINTER SIZE and PAPER SIZE, each with their own set of possible values. Most likely, however, the concepts concerning printer types and paper types will belong to two different terminological ontologies.

As a matter of fact individual domain-specific terminological ontologies in CAOS are not linked together in one common ontology. This is so because the intended users (terminologists or translators) in companies or institutions will only develop ontologies, which are relevant for the domain of their company or institution. An ontology will typically comprise between 50 and 100 concepts (though there is no limit to the number of concepts).

2.2 WordNet's *speech* hierarchy

We will use examples from WordNet in order to compare the principles used in general ontologies with the principles used in terminological ontologies in CAOS. Concepts in CAOS correspond to synsets in WordNet. The information about each synset is not given in the form of feature specifications, but in the form of verbal definitions or explanations. On the basis of these definitions one can, however, deduce dimensions and feature specifications. In WordNet all synsets are included in one ontology, which probably makes it difficult to comply with the principle of uniqueness of dimensions, unless one chooses very specific dimensions. However it is possible to isolate subhierarchies and to consider them as separate ontologies. Let us give an example of such a separate onotology. The noun *talk* has five senses, of which sense 1 is presented as follows:

talk, talking –
(an exchange of ideas via conversation; "let's have more work and less talk around here")

This sense of *talk* is found in the following hierarchy, where => means 'is a hyponym of' (= 'is a subconcept of'):

talk => conversation => speech => auditory communication => communication

On the basis of the information about the relations between synsets given in WordNet one can set up the following tree diagram:

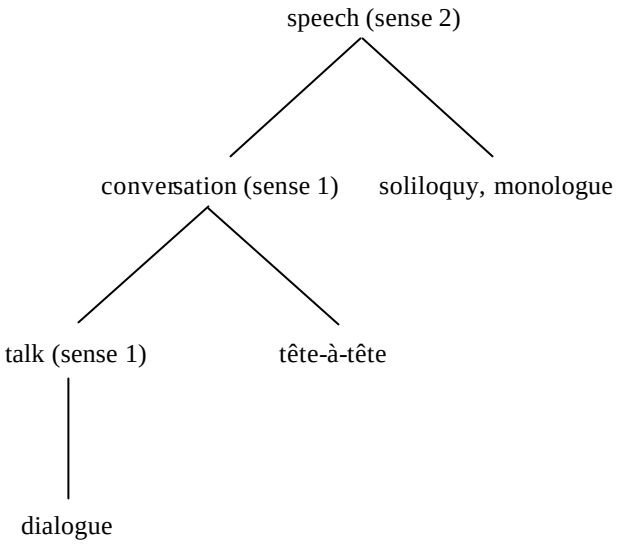


Figure 3: The *speech* hierarchy in WordNet

In several places in this hierarchy we find hyponyms of *speech* distinguished by means of the number of persons participating in the speech. Thus under sense 1 of *talk*, *dialogue* is described as follows:

dialogue, dialog, duologue --
(a conversation between two persons)

Under sense 1 of *conversation*, *tête-à-tête*, is described like this:

conversation --
(the use of speech for informal exchange of views or ideas or information etc.)
=> tete-a-tete –
(a private conversation between two people)

Here => means 'has as hyponym' (= 'has as subconcept'). As hyponym to *speech* in sense 2 we find for example *soliloquy*, *monologue*:

=> soliloquy, monologue -- (speech you make to yourself)

These descriptions make implicit reference to the same characteristic, which could be represented by the dimension NUMBER OF INTERLOCUTORS. However, this would amount to introducing the same dimension on three different concepts (on *speech*, *conversation* and *talk*). Consequently, we note that WordNet's descriptions do not observe the principle of uniqueness of dimensions.
In this particular case, however, the non-application of the principle seems to point to a deficience in the structuring of the ontology. Probably, the *speech* hierarchy could be restructured as shown in figure 4, where *tête-à-tête* is analyzed as a hyponym of *dialogue*.

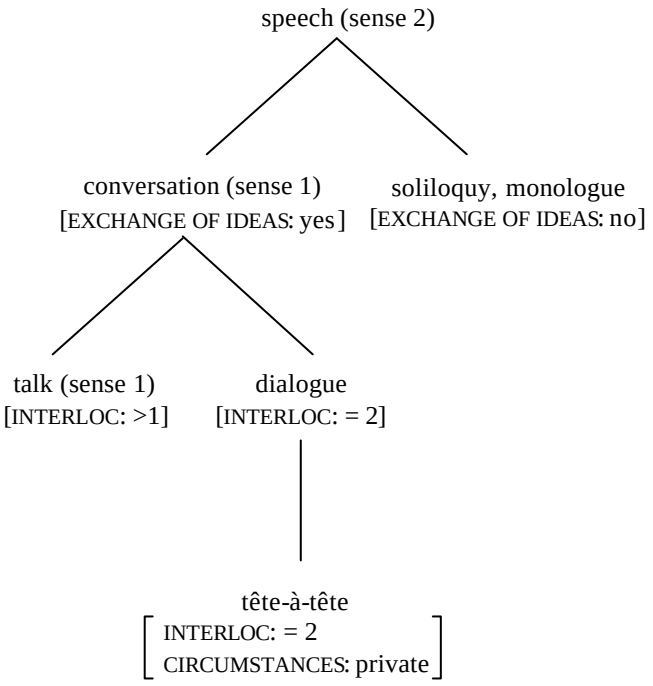


Figure 4: Restructuring of the *speech* hierarchy

2.3 WordNet's *writing* hierarchy

Let us have a look at two other synsets, namely *report* and *letter*, which belong to the following hierarchies:

report => document => writing, written material => written communication => communication

letter => text => matter => writing, written material => written communication => communication

Again the information found in WordNet may be represented as a tree diagram:

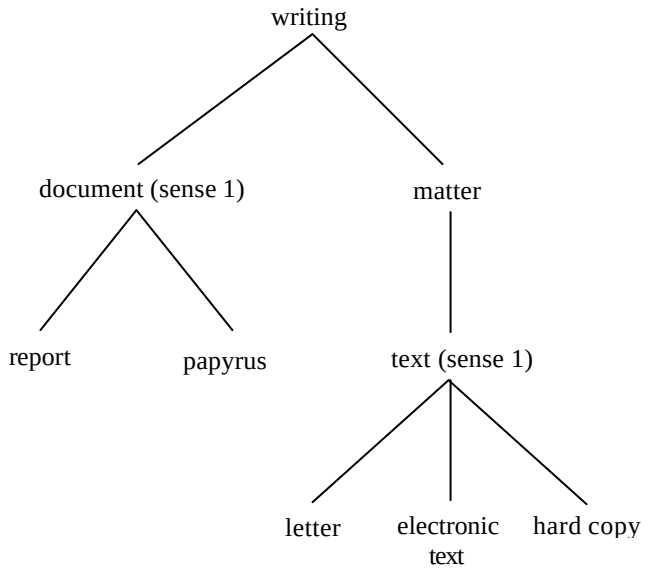


Figure 5: The *writing* hierarchy in WordNet

Under sense 1 of *text* we find the hyponyms *electronic text* and *hard copy*:

text, textual matter --
(the words of something written; "there were more than a thousand words of text"; "they handed out the printed text of the mayor's speech"; "he wants to reconstruct the original text")
=> electronic text --
(text that is in a form that a computer can store and display on a computer screen)
=> hard copy --
(text that is typed or printed on paper; "he ran off a hard copy of the report")

which are distinguished by means of the medium in which the text is represented. Under sense 1 of *document* we find *papyrus*, which is also distinguished from other kinds of documents by means of the medium:

document, written document, papers --
(writing that provides information (especially information of an official nature))
=> papyrus -- (a document written on papyrus)

This characteristic could be represented by the dimension MEDIUM, but this would again entail the introduction of the same dimension on more than one concept in the ontology, i.e. on both *text* and *document*. This problem could be solved by replacing MEDIUM by more specific variants:

MEDIUM OF A TEXT

MEDIUM OF A DOCUMENT

But in this case too, the appearance of multiplication of dimensions reveals a possible weak point in the ontological analysis. In fact, it seems rather arbitrary to make *papyrus* a hyponym of *document*, while *electronic text* and *hard copy* are hyponyms of *text*, and, correspondingly *letter* is analysed as a coordinate concept to *electronic text* and *hard copy*, but it seems to have more in common with *report*. That is, one could imagine a restructuring of the *writing* hierarchy as outlined in figure 6, where the boxes show subdividing dimensions as explained in section 4 below.

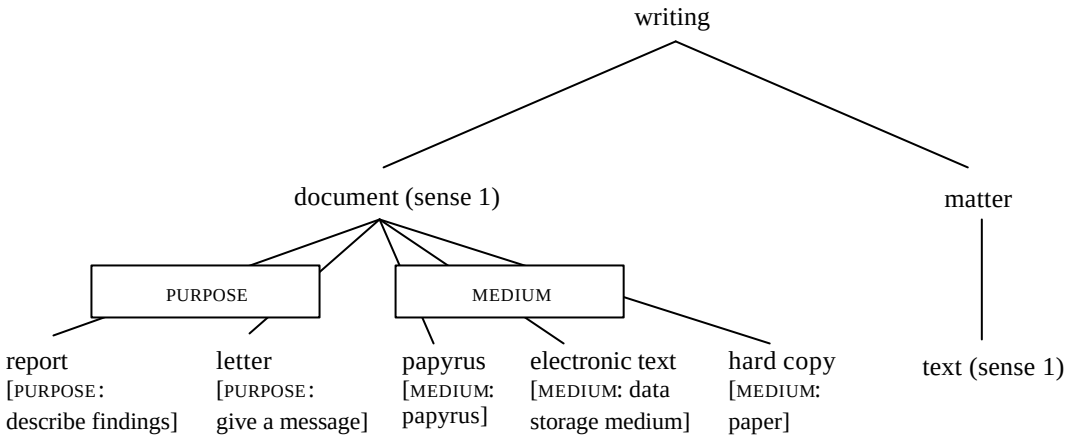


Figure 6: Restructuring af the *writing* hierarchy

We may conclude that within sub-hierarchies such as those displayed in figure 3 and 5 the application of the principle of uniqueness of dimensions could be useful by revealing weaknesses or inconcistencies in the ontological structure and thereby prompting an attempt to improve the analyses proposed.

However, on the basis of these examples alone, we cannot conclude anything about the applicability of the principle on a larger scale. It is in fact difficult to obtain information about the sense descriptions in WordNet on a larger scale. In order to ascertain multiple occurrences of the same dimension it would be necessary to find word senses with partly identical (or at least similar) definitions irrespective of their positions in the ontology. Since it is not possible to search for words or phrases in definitions in WordNet, it is. however, a very difficult task to find similar definitions in distinct parts of the network.

3 Uniqueness of primary feature specifications

The principle of uniqueness of feature specifications stipulates that a feature specification may only occur once in a terminological ontology as primary. A primary feature specification is entered on a concept directly by the terminologist, as opposed to inherited feature specifications, which are inherited from superordinate concepts.

Uniqueness of dimensions (the previous principle) means that a given primary feature specification can only appear on concepts that are daughters of the concept containing the relevant dimension. Uniqueness of primary feature specifications means that a given primary feature specification can only appear on one of these daughters.

The WordNet descriptions do not distinguish between primary and inherited feature specifications, but, as shown in section 2.2, *dialogue* and *tête-à-tête* are hyponyms of different synsets and both are characterized as being "conversation between two people". This description could be reformulated as the ascription of the primary feature specification [NUMBER OF INTERLOCUTORS: 2] to each of the two concepts. If this reformulation is correct, then the original WordNet descriptions do not follow the principle of uniqueness of primary feature specifications.

In this case, the 'violation' of the principle points to one of the same problems as those mentioned in section 2.2, above, namely that perhaps *tête-à-tête* is not in its optimal place in the ontology. This problem could probably be avoided by making *tête-à-tête* a hyponym of *dialogue*, possibly distinguished from *dialogue* by the private character of a tête-à-tête, which might be represented by a feature specification like e.g. [CIRCUM-STANCES: private] as already shown in figure 4 above.

4 Grouping by subdividing dimensions

One or more of the dimensions of a concept must be chosen as subdividing dimensions.

The terminologist is free to choose as subdividing dimensions those dimensions that seem to her to be the most crucial for defining the concepts involved. However subdividing dimensions must be chosen in such a way that each daughter concept has one and only one feature specification containing as an attribute a subdividing dimension of the mother concept (i.e. one and only one delimiting feature specification). That is, there can be no overlapping subdividing dimensions.

In figure 7 it is illustrated that - after registering the character-transfer feature specification - the terminologist has found information telling that impact printers are noisy and may produce multiple copies, whereas nonimpact printers are quiet and can only produce one single copy. This information is represented by means of the two dimensions NOISE and COPY.

In such a case the terminologist might be tempted to choose all three dimensions as subdividing dimensions, but this choice will not be permitted by CAOS, because it would result in the presence of three delimiting feature specifications on each of the two subconcepts. Thus in this case, the terminologist will have to choose between the three dimensions and in this way assign one as the subdividing dimension. Here the terminologist chooses CHARACTER TRANSFER because she will realize that this is the essential one, in that the characteristics associated with this dimension determine

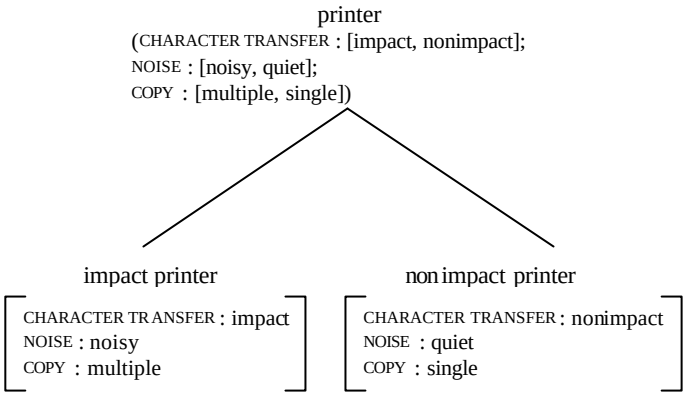


Figure 7: Three dimensions on one concept

the other characteristics: the characteristics of being noisy or quiet and the capacity for single or multiple copies are consequences of the kind of character transfer employed.

Figure 8 below is an extension of the hierarchy shown in figure 2. Only delimiting feature specifications are given and the subdividing dimensions are shown by means of boxes covering the relevant branches. As can be seen from the illustration subdividing dimensions group sister concepts according to the attributes contained in their delimiting feature specifications. They will often prove helpful to the user because they help significantly to give a clearer overview of the field.

As already pointed out WordNet does not mention sub-dividing dimensions, but a description like the following:

> => monologue (a long utterance by one person
>
> (especially one that prevents others from participating
>
> in the conversation))

seems to suggest a characterization by means of two delimiting features, i.e. *long* and *by one person*. And the following description also seems to use two delimiting features, *private* and *between two people*:

> => tete-a-tete –
>
> (a private conversation between two people)

This might indicate that the restriction on subdividing dimensions is not observed. However, in the latter case a reanalysis of *tête-à-tête* as proposed in section 3 above would eliminate the 'violation' of the principle and would suggest replacing the definition of *tête-à-tête* with 'a private dialogue'.

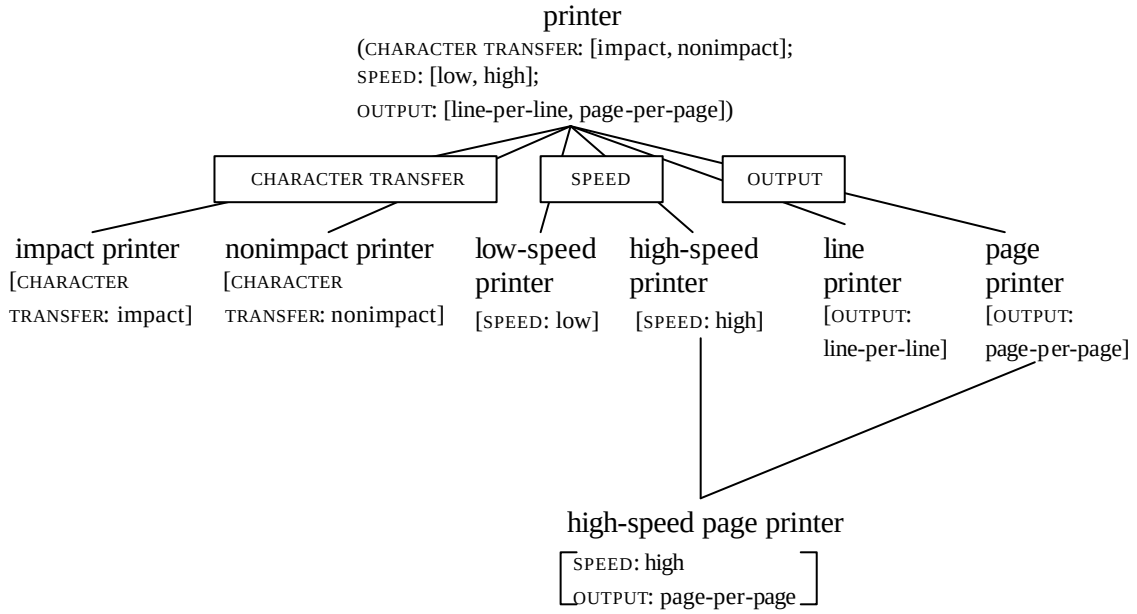


Figure 8: Subdividing dimensions grouping sister concepts

5 Conclusion

We have argued in our paper "Principles of a system for terminological concept modelling" for LREC 2004 that the principles dealt with in that paper are useful in formal descriptions of domain-specific ontologies in that they contribute to achieve simplicity, precision and clarity. But we cannot take for granted that all of these principles will be equally useful when applied to general ontologies.

In the present paper we have argued that the general ontological descriptions of WordNet do not seem to follow the three principles that we have been discussing here. This might be taken as evidence that the principles are irrelevant for general ontologies, but there is also the possibility that the descriptions given by WordNet could be reformulated and restructured in a way that would observe these principles.In the preceding sections we have suggested some restructrurings of this kind, which in our opinion result in descriptions that are simpler and clearer than the current ones in WordNet. We therefore find it probable that the principles are relevant also to general ontologies.

References

Carpenter, Bob. 1992. The Logic of Typed Feature Structures. Cambridge, Massachusetts: Cambridge University Press.

Copestake, Ann. 1992. The Representation of Lexical Semantic Information. University of Sussex at Brighton, Cognitive Science Research Papers, CSRP 280.

Gómez-Pérez, Asunción, Mariano Fernández-López & Oscar Corcho. 2004. Ontological Engineering – with examples from the areas of Knowledge Management, e-Commerce and the Semantic Web. London: Springer Verlag.

Madsen, Bodil Nistrup. 1998. Typed Feature Structures for Terminology Work - Part I. In: *LSP - Identity and Interface - Research, Knowledge and Society. Proceedings of the 11th European Symposium on Language for Special Purposes.* Copenhagen, August 1997, Copenhagen Business School: 339-348.

Madsen, Bodil Nistrup. 1999. *Terminologi 1. Principper og metoder*. Copenhagen, Gads Forlag.

Madsen, Bodil Nistrup, Hanne Erdman Thomsen & Carl Vikner. 2002. Computer Assisted Ontology Structuring. In: Melby, Alan (ed.): *Proceedings of TKE '02 - Terminology and Knowledge Engineering*, INRIA, Nancy: 77-82.

Thomsen, Hanne Erdman. 1998. Typed Feature Structures for Terminology Work - Part II. In: *LSP - Identity and Interface - Research, Knowledge and Society. Proceedings of the 11th European Symposium on Language for Special Purposes.* Copenhagen, August 1997, Copenhagen Business School: 349-359.

Thomsen, Hanne Erdman. 1999. Typed Feature Specifications for establishing Terminological Equivalence Relations. In: *World Knowledge and Natural Language Analysis. Copenhagen Studies of Language*, vol.23, Copenhagen: Samfundslitteratur: 39-55.

WordNet: Web WordNet 2.0, cf. http://www.cogsci.princeton.edu/cgi-bin/webwn.