Identification of relevant terms to support the construction of Domain Ontologies

Paola Velardi§ and Michele Missikoff‡ and Roberto Basili+

§ Università di Roma "La Sapienza"
Velardi@dsi.uniroma1.it

‡IASI-CNR, Roma
missikoff@iasi.rm.cnr.it

+ Roberto Basili
Università di Roma, Tor Vergata
basili@info.uniroma2.it

Abstract

Though the utility of domain Ontologies is now widely acknowledged in the IT (Information Technology) community, several barriers must be overcome before Ontologies become practical and useful tools. One important achievement would be to reduce the cost of identifying and manually entering several thousand-concept descriptions. This paper describes a text mining technique to aid an Ontology Engineer to identify the important concepts in a Domain Ontology.

1 Introduction

In cooperating to work together (or even in interacting in social settings), people and organizations must communicate among themselves. However, due to different contexts and backgrounds, there can be different viewpoints, assumptions and needs regarding the same domain or the same problem. They may use different jargon and terminology, sometimes even confused, overlapping, and they may use concepts and evaluation methods that are mismatched or poorly defined.

The consequence is the lack of a *shared understanding* that leads to a poor *communication* between people and organizations. In particular, when IT solutions are involved, this lack of a shared understanding impacts on:

- Effectiveness of people's cooperation
- Flaws in enterprise organization
- The identification of the *requirements* for the system specification
- The *inter-operability* among systems and
- The possibility of re-*using* and *sharing* of systems components.

The goals of an Ontology is to reduce (or eliminate) conceptual and terminological confusion. This is achieved by identifying and properly defining a set of relevant concepts that characterize a given application domain.

With respect to a *Thesaurus*:

An Ontology aims at describing concepts, whereas a *Thesaurus* aims at describing terms;

An Ontology can be seen as an enriched Thesaurus where, besides the definitions and relationships among terms of a given domain, more *conceptual knowledge*, by means of richer semantic relationships, is represented.
With respect to a *Knowledge Base* (*KB*):
An Ontology can be seen as a KB whose goal is the description of the concepts necessary for talking about domains;
A KB, in addition, includes the knowledge needed to model and elaborate a problem, derive new knowledge, prove theorems, or answer to intentional queries about a domain.
Though the utility of domain Ontologies is now widely acknowledged in the IT community, several barriers must be overcome before Ontologies become practical and useful tools for shared knowledge management.
We envisage three main areas where innovative computational solutions could significantly reduce the cost and effort of Ontology construction:

- provide effective support for collaborative development of *consensus* Ontologies, since consensus is the first condition to be met in order to obtain the desired benefits from an Ontology
- enable distributed development and access to Ontologies, since wide-spread usage of a resource outweighs the cost of development
- develop tools to identify the relevant concepts and (semi-)automatically enrich with semantic information the nodes of the Ontology, thus reducing the cost and complexity of manually defining several thousand concepts

In this paper, we describe SymOntos, an Ontology management system under development at our institution since the last several years. In designing SymOntos, we have been working to define innovative solutions concerning the three critical issues listed above. These solutions are currently being experimented in the context of the European project FETISH[1], aimed at the definition of an interoperability platform for Small Medium Enterprises in the tourism sector.
Though we will (very) briefly present SymOntos, this paper is concerned with the third issue, that is, the description of text mining methods and tools to automatically enrich the concept Ontology.
In the FETISH Project, we decided to explore the possibility to support *the extraction of initial shared/able knowledge* from on-line textual documentation accessible from the Web.

2 SymOntos: a symbolic Ontology management system

SymOntos (SymOntos 2000) is an Ontology management system under development at IASI_CNR. It supports the construction of an Ontology following the *OPAL* (*Object*, *Process*, and *Actor* modeling Language) methodology. OPAL is a methodology for the modeling and management of the *Enterprise Knowledge Base* and, in particular, it allows the representation of the *semi-formal* knowledge of an enterprise. As already mentioned, an Ontology gathers a set of concepts that are considered relevant to a given domain. Therefore, in *SymOntos* the construction of an Ontology is performed by defining a set of *concepts*. In essence, in *SymOntos* a concept is characterized by:
a *term*, that denotes the concept,
a *definition*, explaining the meaning of the concept, generally in natural language,
a set of *relationships* with other concepts.

[1] The interested reader may access the Web site reported in the bibliography

Figure 1 shows an example of filled *concept form* in the Tourism domain. The Domain Ontology is called OntoTour. Concept relationships play a key role since they allow concepts to be inter-linked according to their semantics. The set of concepts, together with their links, forms a *semantic network* (Brachman 1979).

In a semantically rich Ontology, both concepts and semantic relationships are categorized.

Semantic relationships are distinguished according to three main categories[2] namely, *Broader Terms, Similar Terms, Related Terms*, that are described below.

The *Broader Terms* relationship allows a set of concepts to be organized according to a generalization hierarchy (corresponding in the literature to the well-known *ISA* hierarchy).

With the *Similar Terms* relationship, a set of concepts that are similar to the concept being defined are given, each of which annotated with a *similarity degree*. For instance, the concept *Hotel* can have as similar concepts *Bed&Breackfast,* , with similarity degree 0.6, and *camping*, with similarity degree 0.4.

Finally, the *Related Terms* relationship allows the definition of a set of concepts that are semantically related to the concept being defined. Related concepts may be of different kinds, but they must be defined in the Ontology.

For instance, *TravelAgency*, *Customer*, or *CreditCard*, are concepts that are semantically related to the *Hotel* concept.

In SymOntos, Broader relations are also referred to as "vertical" , while Related and Similar are called "horizontal" relations.

SymOntos is equipped with functions to ensure *concept management*, *verification* and Ontology *closure*, and a web interface to help developing consensus definitions in a given user community (Missikoff and Wang, 2000). These functions are not described here since they are outside the purpose of the paper.

Hotel	
Def: *A place where a tourist can stay*	*XML tag: <htl>*
Gen: Accommodation *Spec*: Country_ Guest_ house, motel *Part-of*: receptivity _system *Has-part*: fitness_facilities, restaurant, garage	*Related-objects*: Reservation, payment, deposit *Related-actors*: Htl-manager, cashier, room_service *Related-processes*: reserving, paying, billing, airport_transfer *Similar-concepts*: B&B[0.6], camping[0.4], holiday_apartment[0.7]

Figure 1 – the Hotel concept in OntoTour

3 Text Mining tools to construct a Domain Ontology

In Section 2 we illustrated the main features of the SymOntos system, and provided an example of concept definition in the Tourism domain.

The techniques described in this Section are intended to significantly improve human productivity in the process that a group of domain experts accomplish in order to find an agreement on:

- the identification of the *key concepts* and *relationships* in the domain of interest
- providing an explicit representation of the conceptualization captured in the previous stage

To reduce time, cost (and, sometimes, harsh discussions) it is highly advisable to refer *to the documents available in the field*. In this paper we show that text-mining tools may be of great help in this task.

[2] The represented information is in fact quite more rich, but we omit a detailed description for sake of space

At the present state of the project, natural language processing tools have been used for the following tasks:

1. Identification of thesauric information, i.e. discovery of *terms* that are good candidate *names* for the concepts in the Ontology.
2. Identification of taxonomic relations among these terms.
3. Identification of related terms

For sake of space, only the first method is described in this paper. Details of the other methods may be found in (Missikoff et al. 2001).

To mine texts, we used a corpus processor named ARIOSTO (Basili et al. 1996) whose performance has been improved with the addition of a Named Entity recognizer (Cucchiarelli et al. 1998) (Paliouras et al. 2000) and a chunk parser CHAOS (Basili et al, 1998). In the following, we will refer to this enhanced release of the system as ARIOSTO+.

Figure 2 provides an example of final output (simplified for sake of readability) produced by ARIOSTO+ on a Tourism text. Interpreting the output predicates of Figure 2 is rather straightforward.

The main principles underlying the CHAOS parsing technology are *decomposition* and *lexicalization*. Parsing is carried out in four steps: (1) POS tagging, (2) Chunking, (3) Verb argument structure matching and (4) Shallow grammatical analysis. .

Chunks are defined via *prototypes*. These are sequences of morphosyntactical labels mapped to specific grammatical functions, called *chunk types*. Examples of labels for the inner components are `Det, N, Adj, and Prep` while types are related to traditional constituents, like `NP`, `PP`, etc.

The definition of chunk prototypes in CHAOS is implemented through regular expressions.

Chunks are the first types of output shown in Figure 2. The link(..) predicates represent the result of shallow parsing. Whenever the argument structure information cannot be used to link chunks, a plausibility measure is computed, which is inversely proportional to the number of colliding syntactic attachments (see the referred papers for details). The first phase of the Ontology building process consists in the identification of the key concepts of the application domain.

The Colorado River Trail follows the Colorado River across 600 miles of beautiful Texas Country - from the pecan orchards of San Saba to the Gulf of Mexico .
[1 , Nom , [The,Colorado_River_Trail]]
[2 , VerFin , [follows]]
[3 , Nom , [the,Colorado_River]]
[4 , Prep , [across,600_miles]]
[5 , Prep , [of,beautiful,Texas_Country]]
(more follows..)

link(0,2,'Sentence').
link(2,1,'V_Sog', plaus(1.0)).
link(2,3,'V_Obj', plaus(1.0)).
link(3,4,'NP_PP',plaus(0.5)).
link(2,4,'V_PP',plaus(0.5)).
link(4,5,'PP_PP',plaus(0.3333333333333333)).
link(3,5,'NP_PP',plaus(0.3333333333333333)).
link(2,5,'V_PP',plaus(0.3333333333333333)).
(…*morefollows*…)

Figure 2. An example of parsed Tourism text

Though concept *names* do not always have a lexical correspondent in natural language, especially at the most general levels of the Ontology, one such correspondence may be naturally drawn among the more specific concept names and domain-specific words and complex nominals, like:

- Domain Named Entities (e.g., *gulf of Mexico*, *Texas Country*, *Texas Wildlife Association*)
- Domain-specific complex nominals (e.g., *travel agent, reservation list, historic site*, *preservation area*)
- Domain-specific singleton words (e.g., *hotel, reservation, trail, campground*)

We denote these singleton and complex words as Terminology.

Terminology is the set of words or word strings , which convey a single, possibly complex, meaning within a given community. In a sense, Terminology is the surface appearance, in texts, of the domain knowledge in a given domain. Because of their low ambiguity and high specificity, these words are also particularly useful to conceptualize a knowledge domain, but on the other side, these words are often not found in Dictionaries. We now describe how the different types of Terminology are captured using NLP techniques.

3.1 Detection of Named Entities

Proper names are the *instances* of domain concepts, therefore they populate the leaves of the Ontology.

Proper names are pervasive in texts. In the Tourism domain, as in most domains, Named Entities (NE) represent more than 20% of the total occurring words.

To detect NE, we used a module already available in ARIOSTO+. A detailed description of the method summarized hereafter may be found in (Cucchiarelli et al. 1998) (Paliouras et al. 2000). In ARIOSTO+, NE are detected and semantically tagged according to three main conceptual categories: *locations (objects in OPAL)*, *organizations* and *persons (actors in OPAL)* . When contextual cues are sufficiently strong (e.g. "*lake Tahoe is located*."), names of locations are further sub-categorized (*city*, *bank*, *hotel*, *geographic location*, ..), therefore the Ontology Engineer is provided with semantic cues to correctly place the instance under the appropriate concept node of the Ontology.

Named Entity recognition is based on a set of *contextual rules* (e.g. "*a complex or simple proper name followed by the trigger word authority is a organization named entity*").

Rules are manually entered or machine learned using decision lists. If a complex nominal does not match any contextual rule in the NE rule base, the decision is delayed until syntactic parsing. A classification based on syntactically augmented context similarity is later attempted.

The NE tagger is also used to automatically enrich the Proper Names dictionary, thus leading to increasingly better coverage as long as new texts are analyzed.

As reported in the referred papers, the F-measure (combined recall and precision with a weight factor w=0,5) of this method is consistently (i.e. with different experimental settings) around 89%, a performance that compares very well with other NE recognizers described in the literature[3].

3.2 Detection of domain-specific words and complex nominals

NEs are word strings in part or totally capitalized, and they often appear in well-characterized contexts. Therefore, the task of NE recognition is relatively well assessed in literature. Other not-named terminological patterns (that we will refer hereafter again with the word "terminology" though in principle terminology includes also NEs) are rather more difficult to capture since the notion of *term* is mostly underspecified.

In the literature (see Bourigault et al. (1998) for an overview of recent research) the following steps are in general adopted:

- Detecting terminological candidates from texts
- Selecting the specific entries that can be members of a terminological glossary in the target domain of knowledge.

Candidates terminological expressions are

[3] ftp.muc.saic.com/proceedings/score_reports_index.html

usually captured with more or less shallow techniques, ranging from stochastic methods (Church, 1988) to more sophisticated syntactic approaches (e.g. Jacquemin, 1997).

Obviously, richer syntactic information positively influences the quality of the result to be input to the statistical filtering. In our research, we used the CHAOS parser to select candidate terminological patterns. Nominal expressions usually denoting terminological items are very similar to chunk instances. Specific chunk prototypes have been used to match terminological structures.

A traditional problem of purely syntactic approaches to term extraction is overgeneration. The available candidates that satisfy grammatical constraints are far more than the true terminological entries. Extensive studies suggest that statistical filters be always faced with 50-80% of non-terminological candidates.

Filtering of true terms can be done by estimating the strength of an association among words in a candidate terminological expression. Commonly used association measures are the Mutual Information (Fano, 1961) and the Dice factor (Smadja et al. 1996). In both formulas, the denominator combines the marginal probability of each word appearing in the candidate term. If one of these words is particularly frequent, both measures tend to be low. This is indeed not desirable, because certain very prominent domain words appear in many terminological patterns. For example, in the Tourism domain, the term *visa* appears both in isolation and in many multiword patterns, e.g.: *business visa*, *extended visa*, *multiple entry business visa*, *transit visa*, *student visa*, etc.…Such patterns are usually not captured by standard association measures, because of the high marginal probability of *visa*.

Another widely used measure is the inverse document frequency, *idf*.

$$\mathrm{idf_i} = \log_2 \frac{\mathrm{N}}{\mathrm{df_i}}$$

Where $\mathrm{df_i}$ is the number of documents in a domain D_i that include a term t, and N is the total number of documents in a collection of *n* domains (D_1, …, D_n). The idea underlying this measure is to capture words that are frequent in a subset of documents representing a given domain, but are relatively rare in a collection of generic documents. This measure captures also words that appear just one time in a domain, which is in principle correct, but is also a major source of noise.

Other corpus-driven studies suggested that pure frequency as a ranking score (i.e. a measure of the plausibility of any candidate to be a term) is a good metrics (Daille 1994). However, frequency alone cannot be taken as a good indicator: several very frequent expressions (e.g. *last week*) are perfect candidates from a grammatical point of view but they are totally irrelevant as terminological expressions. It is worth noticing that this is true for two *independent* reasons. First, they are not related to specific knowledge, pertinent to the target domain, but are language specific: different languages express with different syntactic structures (adverbial vs. nominal phrases) similar temporal or spatial expressions. As a result such expressions have similar distributions in different domain corpora. True terminology is tightly related to specific concepts so that their use in the target corpus is highly different wrt other corpora. Second, common sense expressions are only occasionally used, their meaning depending on factual rather than on conceptual information. They occur often once in a document and tend not to repeat throughout the discourse. Their appearance is thus evenly spread throughout documents of any corpus. Conversely, true

terms are central elements in discourses and they tend to recur in the documents where they appear. They are thus expected to show more skewed (i.e. low entropy) distributions.

The above issues suggest the application of two different evaluation (*utility*) functions. Although both are related to the widely employed notion of term probability, they capture more specific aspects and provide a more effective ranking.

3.2.1 Modeling Relevance in domains

As observed above, high frequency in a corpus is a property observable for terminological as well as non-terminological expressions (e.g. "*last week*" or "*real time*"). The specificity of a terminological candidate with respect to the target domain (Tourism in our case) is measured via comparative analysis across different domains. A specific score, called *Domain Relevance* (DR), has been defined. More precisely, given a set of *n* domains[4] (D_1, …, D_n) the domain relevance of a term *t* is computed as:

$$DR(t, D_i) = \frac{P(t \mid D_i)}{\sum_{i=1..n} P(t \mid D_i)} \quad (1)$$

where the conditional probabilities ($P(t|D_i)$) are estimated as:

$$E(P(t|D_i)) = \frac{freq(t \text{ in } Di)}{\sum_{i=1..n} freq(t \text{ in } Di)}$$

3.2.2 Modeling Consensus about a term

Terms are concepts whose meaning is agreed upon large user communities in a given domain. A more selective analysis should take into account not only the overall occurrence in the target corpus but also its appearance in single documents. Domain concepts (e.g. *travel agent*) are referred frequently throughout the documents of a domain, while there are certain specific terms with a high frequency within single documents but completely absent in others (e.g. *petrol station*, *foreign income*).

Distributed usage expresses a form of *consensus* tied to the consolidated semantics of a term (within the target domain) as well as to its centrality in communicating domain knowledge. A second indicator to be assigned to candidate terms can thus be defined. *Domain consensus* measures the distributed use of a term in a domain D_i.. The distribution of a term *t* in documents d_j can be taken as a stochastic variable estimated throughout all d_j *Di*. The entropy H of this distribution expresses the degree of consensus of *t* in D_i. More precisely, the domain consensus is expressed as follows

$$DC(t, D_i) = H(P(t, dj) = \sum_{dj \in Di} P(t, dj) \log_2 \left(\frac{1}{P(t, dj)} \right) \quad (2)$$

Where:

$$E(P(t, dj)) = \frac{freq(t \text{ in } dj)}{\sum_{dj \in Di} freq(t \text{ in } dj)}$$

Pruning of not terminological (or not-domain) candidate terms is performed using a combination of the measures (1) and (2). We experimented several combinations of these two measures, with similar results. The results, discussed in the next Section, have been obtained applying a threshold α to the set of terms ranked according to (1) and then eliminating the candidates with a rank (2) lower than β.

4 Experiments

An obvious problem of any automatic method

[4] « domains » are (pragmatically) represented by texts collections in different areas, e.g. medicine, finance, tourism, etc.

for concept extraction is to provide objective performance evaluation.

- Firstly, a "golden standard" tourism terminology would be necessary to formally measure the accuracy of the method. One such standard is not available, and determining this standard is one of the objectives of FETISH. Moreover, the notion of "term" is too vague to consider available terminological databases as "closed" sets, unless the domain is extremely specific.
- Secondly, no formal methods to evaluate a terminology are available in literature. The best way to evaluate a "basic" linguistic component (i.e. a module that performs some basic task, such as POS tagging, terminology extraction, etc.) within a larger NLP application (information extraction, document classification, etc.) is to compute the difference in performance with and without the basic component. In our case, since Ontology does not perform any measurable task, adopting a similar approach is not straightforward. As a matter of facts, an Ontology is a basic component itself, therefore it can be formally evaluated only in the context of some specific usage of the Ontology itself.

Having in mind all these inherent difficulties, we performed two sets of experiments. In the first, we extracted the terminology from a collection of texts in the Tourism domain, and we manually evaluated the results, with the help of other participants in the FETISH project (see the FETISH web site). In the second, we attempted to assess the generality of our approach. We hence extracted the terminology from a financial corpus (the Wall Street journal) and then we both manually evaluated the result, and compared the extracted terminology with an available thesaurus in a (approximately) similar domain.

As a reference set of terms we used the Washington Post[5] (WP) dictionary of economic and financial terms.

To compute the Domain Relevance, we first collected corpora in several domains: tourism announcements and hotel descriptions, economic prose (Wall Street Journal), medical news (Reuters), sport news (Reuters), a balanced corpus (Brown Corpus) and four novels by Wells. Overall, about 3,2 million words were collected.

In the first experiment, we used the Tourism corpus as a “target” domain for term extraction.

The Tourism corpus was manually built using the WWW and currently has only about 200,000 words, but it is rapidly growing.

Table 1 is a summary of the experiment. It is seen that only 2% terms are extracted from the initial list of candidates. This extremely high filtering rate is due to the small corpus: many candidates are found just one time in the corpus. However, candidates are extracted with high precision (over 85%).

N. of candidate multiword terms (after parsing)	14.383
N. of extracted terms (with α=0.35 and β=0.50)	288
% correct (3 human judges)	85.20%
Number of subtrees (of which with depth>0)	177 (54)

Table 1. Summary results for the term extraction task in the Tourism domain

Table 2 shows the 15 most highly rated multiword terms, ordered by Consensus (Relevance is 1 for all the terms in the list).

Table 3 illustrates the effectiveness of Domain Consensus at pruning irrelevant terms: all the

[5]http://www.washingtonpost.com/wp-srv/business/longterm/glossary/indexag.htm

candidate terms in the list have DR>α, but DC<β.

Terms	Domain Consensus
credit card	0.846913
tourist information	0.696701
travel agent	0.686668
swimming pool	0.664041
service charge	0.640951
car rental	0.635580
credit card number	0.616671
card number	0.616671
room rate	0.596764
information centre	0.579662
beach hotel	0.571898
tourist area	0.565462
tour operator	0.543419
standard room	0.539450
video camera	0.523142

Table 2: The 15 most highly ranked multiword Tourism terms

	Domain Relevance	Domain Consensus
english cyclist	1.000000	0.000000
manual work	1.000000	0.000000
petrol station	1.000000	0.000000
school diploma	1.000000	0.000000
western movie	1.000000	0.000000
white cloud	1.000000	0.000000
false statement	0.621369	0.000000
best price	0.612948	0.224244
council decision	0.612948	0.000000
foreign income	0.441907	0.000000
gay community	0.441907	0.224244
mortgage interest	0.441907	0.000000
substantial discount	0.441907	0.224244
typical day	0.441907	0.224244

Table 3. Terms with high Domain Relevance and low Domain Consensus

In the second experiment, we used the one-million-word Wall Street journal (WSJ) and the Washington Post (WP) reference terminology.

The WP includes 1270 terms, but only 214 occur at least once in the WSJ. We used these 214 as the "golden standard" ($Test_1$), but we performed different experiments eliminating terms with a frequency lower than 2 ($Test_2$), 5 ($Test_5$) and 10 ($Test_{10}$). This latter set includes only 73 terms.

During syntactic processing, 41,609 chunk prototypes have been extracted as eligible terminology.

The Tables 4 and 5 compare our method with t with Mutual Information, Dice factor, and pure frequency. Clearly, these measures are applied on the same set of eligible candidates extracted by the CHAOS chunker. The results reported in each line are those obtained using the *best* threshold for each adopted measure[6]. For our method (DR+DC), the threshold is given by the values α and β. As remarked in the introduction, a comparison against a golden standard may be unfair, since, on one side, many terms may be present in the observed documents, and not present in the terminology. On the other side, low frequency terms in the reference terminology are difficult to capture using statistical filters. Due to these problems, the F-measure is in general quite low, though our method outperforms Mutual Information and Dice factor. As remarked by Daille (1994), the frequency emerges as a reasonable indicator, especially as for the Recall value, which is a rather obvious result.

However pure frequency implies the problems outlined in the previous section. Upon manual inspection, we found that, as obvious, undesired terms increase rapidly in the frequency ranked term list, as the frequency decreases. Manually inspecting the first 100 highly ranked terms produced a score of 87,5 precision for our method, and 77,5 for the frequency measure. For the subsequent 100 terms, the discrepancy gets much higher (18%).

Note that the precision score is in line with that obtained for the Tourism corpus. Notice also

[6] as a matter of fact, for our method we are not quite using the best value for β, as remarked later.

that the values of α and β are the same in the two experiments. In practice, we found that the threshold α=0,35 for the Domain Relevance is a generally "good" value, while a little tuning may be necessary for the Domain Consensus. In the Tourism domain, where statistical evidence is lower, a lower value for β produces higher precision (+1, 2%).

Method	Threshold	Prec	Recall	F
DR+DC	0.35 0.49	17.18	17.61	17.39
MI	0.00009	6.68	32.08	11.05
Dice	0.034	7.48	23.90	11.39
Freq	22	14.19	25.79	18.30

Table 4 WSJ/WP experiment on $Test_1$

Method	Threshold	Prec	Recall	F
DR+DC	0.35 0.57	23.80	19.42	21.39
MI	0.00009	6.42	47.57	11.30
Dice	0.057	8.22	23.30	12.15
Freq	22	14.19	39.81	20.92

Table 5 WSJ/WP experiment on $Test_5$

5 Conclusion and Future Work

The text mining techniques proposed in this paper are meant to increase the productivity of an Ontology Engineer during the time consuming task of populating a Domain Ontology. The work presented in this paper is in part well assessed, in part still under development. We are designing new algorithms and techniques to widen the spectrum of information that can be extracted from texts and from other on-line resources, such as dictionaries and lexical taxonomies (like EuroWordnet, a multilingual version of Wordnet). An on-going extension of this research is to detect similarity relations among concepts on the basis of contextual similarity. Similarity is one of the fields (see Figure 1) in a concept definition form that are currently filled by humans.

One admittedly weak part of the research presented in this paper is evaluation: we could produce a numerical evaluation of certain specific subtasks (extraction of Named Entities and extraction of thesauric information), but we did not evaluate the overall effect that our text mining tools produce on the Ontology. However, we are not aware of any assessed Ontology evaluation methodology in the literature, besides (Farquhar et al. 1996) where an analysis of Ontology Server user distribution and requests is presented. A better performance indicator would have been the number of users that access Ontology Server on a regular basis, but the authors mention that regular users are only a small percentage[7]. As remarked in Subsection 3.1.2, an objective evaluation of an Ontology as a stand-alone artifact is not feasible: the only possible success indicator is the (subjective) acceptance/rejection rate of the Ontology Engineer when inspecting the automatically extracted information. An Ontology can only be evaluated in a context in which many users of a community (e.g. Tourism operators in our application) access the Ontology on a regular basis and use this shared knowledge to increase their ability to communicate, access prominent information and documents, improve collaboration. Though a field evaluation of OntoTour is foreseen during the last months of the project, we believe that wide accessibility and a long-lasting monitoring of user behaviors would provide the basis for a sound evaluation of the OntoTour system.

[7] The system described by Farquhar and his colleagues, however, is not a specific Ontology, but a tool, Ontology Server, to help publishing, editing and browsing an Ontology.

6 References

E. Agirre, O. Ausa, E. Havy and D. Martinez "Enriching very large ontologies using the WWW" ECAI2000 workshop on Ontology Learning, http://ol2000.aifb.uni-karlsruhe.de/, Berlin, August 2000

R.J. Brachman "On the epistemological status of semantic networks"; in "Associative Networks - Representation and use of Knowledge by Computers", N.V.Findler (Ed.); Academic Press, New York, 1979.

Basili R., M.T. Pazienza, F.M. Zanzotto (1998), *A Robust Parser for Information Extraction*, Proceedings of the European Conference on Artificial Intelligence (ECAI '98), Brighton (UK), August 1998.

Bourigault, D. , C. Jacquemin and M.C. L'Homme (1998), Eds. *Proceedings of the first Workshop on Computational Terminology*, jointly held with COLING-98, Montreal, 1998

A. Cucchiarelli, D. Luzi and P. Velardi "Semantic tagging of Unknown Proper Noun"s in Natural Language Engineering, December 1998.

V. Paliouras Cucchiarelli A., Karkaletsis G. Spyropolous C. Velardi P. "Automatic adaptation of Proper Noun Dictionaries through cooperation of machine learning and probabilistic methods" 23rd annual SIGIR, Athens, June 2000

Basili, R., M.T. Pazienza, P. Velardi, An Empyrical Symbolic Approach to Natural Language Processing, Artificial Intelligence, 85, 59-99, August 1996

R. Basili, G. De Rossi, M.T. Pazienza "Inducing Terminology for Lexical Acquisition" Proc. of the Second Conference on Empirical Methods in Natural Lanaguge Processing, Providence, USA, August 1997

R. Basili, M.T. Pazienza F. Zanzotto, "Customizable Modular Lexicalized Parsing Extraction" proc. of Int. Workshop on Parsing Technology, Povo (Trento) February 2000

B. Daille "Study and Implementation of Combined Techniques for Automatic Extraction of Terminology" Proc. of ACL94 Workshop "The Balancing Act: combining Symbolic and Statistical Approaches to Language" , New Mexico State University, July 1994.

R. Fano "Trasmission of Information, MIT press, 1961

Farquhar, R. Fikes, W. Pratt, J. Rice "Collaborative Ontology Construction for Information Integration" http://www-ksl-svc.stanford.edu:5915/doc/project-papers.html

FETISH Groupware (2001) http://liss.uni.net/QuickPlace/trial/Main.nsf?OpenDatabase

Jacquemin, C. (1997). *Variation terminologique*. Memoire d'Habilitation Diriger des Recherces and Informatique Fondamentale. Université de Nantes, Nantes, France.

Justenson J.S. and S.M. Katz (1995) Technical terminology: some linguistic properties and an algorithm for identification in text. Natural language engineering, Vol. 1, Part 1, March 1995

Klavans, J (2001). *Text Mining Techniques for Fully Automatic Glossary Construction*, Proceedings of the HTL2001 Conference, San Diego (CA), March, 2001.

Miller A. "WordNet: An on-line lexical resource" Special issue of the Journal of Lexicography, 3(4) 1990

M.Missikoff, XF. Wang, "*Consys* – A Web System for Collaborative Ontology Building", submitted, Dic. 2000.

Missikoff M., Velardi P. and Fabriani P. (2001) "Text Mining Techniques to Automatically Enrich a Domain Ontology" to appear on Applied Intelligence, Special Issue on Text and Web Mining.

A. Maedche and S. Staab "Learning Ontologies for the Semantic Web" http://www.aifb.uni-karlsruhe.de/WBS/ama/publications.html

Pustejovsky "The generative lexicon : a theory of computational lexical semantics" MIT press 1993

Smadja, F, K. McKeown and V. Hatzivassiloglou (1996) *Translating Collocations for Bilingual Lexicons: a statistical approach*, Computational Linguistics, 22:1

SymOntos (2001), a Symbolic Ontology Management System. http://www.symontos.org

A. Wagner "Enriching a Lexical Semantic Net with Selectional Prefernces by means of Statistical Corpus Analysis" ECAI2000 workshop on Ontology Learning, ibidem

Y. Wilks, B. Slator and L. Guthrie "Electric Words: Dictionaries, Computers, and Meaning", MIT Press, Cambridge, MA, 1999