Acquisition of domain conceptual dictionaries via decision tree learning.

Roberto Basili, Maria Teresa Pazienza, Fabio Massimo Zanzotto[1]

Abstract. Knowledge based systems usually rely on large size domain models needed to support reasoning and decision-making. The development of realistic models represents a critical and labour intensive phase. Automatic terminology acquisition (TA) has been proposed as the task of automatically extracting specialized dictionaries from raw texts useful for application purposes like precise information retrieval and machine translation. In this paper we argue that TA provides a significant contribution in the development of ontological components of a knowledge bases. We therefore propose an automatic knowledge acquisition architecture for the TA process based on robust methods for text processing and on algorithms for learning decision trees. An incremental semi-automatic approach is proposed to enable the first steps in the development of a domain ontology. The novel aspects of the method rely on the use of syntagmatic and lexical properties of terms combined with analogous (negative) evidences observable for non-terms. The underlying assumptions as well as the different adopted linguistic representations have been extensively investigated over a large test set. The scale of the target test data provides empirical evidence of the superiority of the method over more quantitative approaches. The proposed architecture is thus a viable approach to the development of conceptual domain dictionaries.

1 Terminology extraction and domain modelling

Knowledge based systems usually rely on large size domain models needed to support reasoning and decision-making. The development of realistic models represents a critical and labour intensive phase. Automatic terminology acquisition (TA) has been often proposed as the task of developing specialized dictionaries in target application scenarios (e.g. precise information retrieval systems). In TA complex linguistic expressions (e.g. *joint venture*) are usually discovered in texts and used to populate knowledge repositories (e.g. inheritance networks). It is worth noticing that terminology acquisition allows the detection of those grammatical structures (or surface forms) denoting complex concepts needed within a particular domain. This provides a partial support for the development of domain models in knowledge based applications. Terminological entries represent relevant concepts for the target domain.

[1] University of Rome Tor Vergata, Department of Computer Science, Systems and Production, 00133 Roma (Italy), {basili,pazienza,zanzotto}@info.uniroma2.it

Methods for TA rely on extensional descriptions of a domain usually embodied by large text collections related to it. Human experts start analysing and encoding terms by also adding normative definitions if required. These decisions *bias* their later analysis within the same texts. The corpus together with the intermediate term dictionary forms what we will hereafter call an *implicit domain model*. Clearly neither the dictionary nor the corpus represent a complete conceptual model of the domain, but they express:

- a (temporary) set of relevant concepts in a domain
- the full variety of term usages, *implicitly* expressing properties and relations of each underlying conceptual item.

The TA process represents thus an earlier phase of the development of an explicit semantic domain model useful for a number of different intelligent activities (e.g. information retrieval and decision making).

Most automatic TA methods start from the definition of *what a term is* and there is a general consensus in defining a term as a surface representation of a key domain concept [9]. This definition opens different "operational" interpretations leading to the design of different corpus-driven TA systems. An *"operational" definition* can be obtained by specifying:

- how to constrain admissible surface forms, usually via specialized NP grammars in agreement with valid natural language interpretations;
- how to establish the *relevance* of a candidate form s as denotation of a domain concept.

Generally valid surface forms are specified at the morpho-syntactic level and derive candidate forms S. Sometimes, heuristics are also employed, like stop words lists of irrelevant (e.g. temporal) expressions. Statistical models are then used over candidate in S as measures of *domain relevance*. In [7], the simple frequency $f(s)$ of surface forms in the corpus is suggested to be the most effective measure. f seems to reproduce the terminologist judgement better than other more complex measures. However, as admittedly mentioned in [7], frequency alone is still far from being a perfect discriminating "*termhood*" function. For instance, in an economic corpus made of about 13,000 newspaper articles *joint venture* and *oil price* have been found 687 and 787 times, respectively. Given such counts as the only distinctive feature, no method can distinguish the first as a true terminological expression from the second that is a general (i.e. domain independent) expression. A still frequency-driven approach that employs a contrastive measure across several domains has been recently proposed in [2], with slight performance improvements.

In the above models the *implicit domain model* is considered just as the sample space where distributions of new candidates can be observed. However, the distributional behaviour of surface forms is not the only *observable* property. Two important kinds of information are neglected:

- *usages* of already accepted terms (or terms given as initial seeds) embodied by the corpus. The *contextual information* of terms, like grammatical relationships they establish with other words, is extensively used by terminologists to reason about term relevance
- *negative assumptions*. The refusal of frequent, but non-terminological, expressions provides information about *what a term is not*. An inductive approach (e.g. decision trees) may well exploit this as negative evidence during training and classification.

From information about typical usages of accepted (or refused) candidates we can derive an *intentional definition* of term (or non term). Several properties (i.e. *exogenous information* as in [4]) can be observable in the contexts of terms. They form an implicit definition for the acceptance new candidates. Relevant properties shared among terms should from a predictive (intentional) model able to correctly separate terms from non-terms. For example, let the following text fragments represent a scientific corpus (SC):

Example 1 *(Corpus SC)*

a) The **bread-and-butter equation** *of fluid mechanics governs the conservation of energy of everything from flows to jets and turbulence. But the* **equation** *has been hard to apply to drop formation because at the time of pinch-off, terms in the equation head off to infinity.*

b) The **generalized airfoil equation** *governs the pressure across an airfoil oscillating in a wind tunnel.*

Expressions *bread-and-butter equation* and *generalized airfoil equation* in the Example 1, are both `subject`s of the verb *govern*. This is often true of technical definitions of physical laws. Such grammatical fact (as a shared property among the two potential terms) may thus be adopted as selective criteria. If, for example, *bread-and-butter equation* has been already decided as a term, we can use *subject*-ness with the verb *govern* as a decision rule. Such grammatical similarity is typical of the underlying domain. The induction of such rules provides a truly domain-specific *intensional term definition*. Notice how this is also true for non-terms that give rise to negative classification rules in the model.

The method depicted above is an original approach to TA. Corpus-driven methods (e.g. [6, 11, 10]) usually do not develop any *unified* intentional term definition and do not use any negative evidence. The extensional term definition has been instead used for quantitative (frequency based) criteria like co-occurrences in text windows. A knowledge-intensive method based on term semantic networks is used in [1, 10] to detect new terms and organize them within the existing knowledge bases. In [4] a richer approach based on shallow syntactic analysis is proposed to support TA over "poorer" domains, i.e. domains for which lexical semantic knowledge bases are not available. The above approaches make use of contextual information observed for known (i.e. already assessed terms) but neglect the negative information about "non-terms" (general expressions available in "supervised" approaches).

In this paper the notion of domain-specific *intensional term definition* is induced via linguistic processing of a target corpus and machine learning. The result is a weakly supervised classification system that, triggered by a small amount of seeding information (terms already known in the domain), predicts the "termhood" of new surface forms as found in the corpus. A syntactically motivated model is induced by representing grammatical exogenous properties of terms (and non-terms) in contexts. The formalism adopted for exogenous information is described in Section 2. Experimental evaluation based on a decision tree learning algorithm (C4.5 [12]) is then presented in Section 3. Results suggest that the proposed corpus-driven TA method is a viable architecture that integrates natural language processing and machine learning for knowledge acquisition.

2 A syntactic-oriented notion of extensional term definitions

The implicit domain model definition is used by terminologists that read texts and decide about *termhood* or *non-termhood* of new candidates. Inductive learning of an intensional term model is inherently based on the observations over text corpora. A suitable observation model should include all those selective properties characterizing terms. One such model corresponds to a space for describing positive and negative instances.

The aim here is to select the regular behaviour of terms in corpus contexts, i.e. their exogenous information. Syntax will be used (in line with other works like [8] or [4]) as a linguistic level able to characterize similarity among contexts. If a particular grammatical relation (e.g. `subject`) frequently links positive instances (e.g. *generalized airfoil equation*) to some other textual elements (e.g. the verb *to govern*), it can be assumed as a decision rule for discovering *termhood*. As grammatical relations can be easily observed for known terms as well as for promising candidates, the resulting rule set is the target *intensional term definition*.

In the next Sections the formal definitions of the feature vectors representing positive and negative instances (i.e. their exogenous information) are presented.

2.1 Term occurrences and exogenous information

When collecting evidences of a given term t across a domain corpus we need to determine whether or not different contexts are indicators of its exogenous behaviour. A first possibility is to collect only contexts where a valid surface form for t[2] appears. Notice however, that in many cases terms are referred in an elliptic fashion. In the example 1.a), the second occurrence of the word *equation* is an elliptic occurrence of *bread-and-butter equation*. As a consequence the context *... the equation has been hard to apply to drop formation ...* describes the exogenous behaviour of the *bread-and-butter equation* term as well. Many simple terms (i.e. one-word terms) are elliptic references to complex terms (i.e. multi-word terms). Generally,

[2] A valid surface form is here simply intended as a morphological variant of the term, e.g. *fluid mechanics equations* vs. *fluid mechanics equation*: the t canonical form here is the latter singular.

the term grammatical *head* (e.g. *equation* in *bread-and-butter equation*) is used in elliptic references.

The syntactic exogenous behaviour of a term is driven by its semantics. The head $h(t)$ of a term t is usually its semantic carrier. This assumption is widely used in other term structuring approaches (cf. [11]). $h(t)$ is thus a good canonical candidate of t. Its occurrences in the corpus are representative of direct or elliptic occurrences of t. This is a computationally attractive approximation for counting. Moreover, as terms are expected to have unique interpretations in a coherent domain, terms t and t' such that $h(t) = h(t')$ will be considered equivalent with respect to their exogenous information. Accordingly, terms *bread-and-butter equation* and *generalized airfoil equation* are equivalent with respect to the head *equation*.

The contribution of all contexts where a given head $h(t)$ appears forms an equivalence class, $C(t)$, in the corpus. A single (collective) representation, the vector $v(t)$, for t can be thus derived from all $c \in C(t)$. Moreover, this seemingly applies to "non-terms". In the next section, definition for vectors $v(t)$, i.e. feature vectors populating the sample space, is given.

2.2 Spaces for exogenous features

The automatic induction of a model for terms (or non terms) out from the extensional domain model requires a suitable knowledge representation formalism. The overall process includes the following steps:

- Extraction of grammatical information from local contexts as *local feature vectors* using shallow parsing techniques
- Generation of a *global feature vector* for an entire term equivalence class
- Induction of the target intensional definition as a *decision tree* that separates incoming candidates into terms and non-terms via machine learning algorithms

The above process can be also modelled as an incremental approach. Newly accepted (or refused) candidates[3] allow a dynamic revision of the corresponding decision tree structure: a new learning process can be activated over the newly assessed instances. The usage of syntactic information supports the derivation of rich (decision-tree) descriptions. It supports not only decisions about the termhood of incoming surface forms but the declarative aspects (e.g. typical grammatical relations as properties motivating a given decision) may support as well more powerful inferences, like induction of semantic relations among terms.

As poor agreement exists about grammatical properties characterizing terminological structures in a domain, we will explore several alternatives. A "light model" represents only the type of the grammatical relations established by a term t with other contextual elements. For example, types like `object` or `subject` are features expressing the role of a term in an underlying context (e.g. *equation* is `subject` in Ex. 1). The resulting feature space will be hereafter called "*syntactic* (Σ)".

A more informative (deeper) model can be otherwise obtained by preserving the local lexical information. In such a "syntactic lexicalised" model (Λ), the lexical item that governs the observed grammatical relation is stored in a local vector together with the grammatical type. For example, given the context *"The equation of mechanics governs the conservation of energy."* of *equation*, we can capture *equation* as the *subject* of the verb *to-govern*.

The two kinds of information have different feature spaces, Σ and Λ respectively. The feature system F^{Σ}, in Σ, includes the following syntactic types: $F^{\Sigma} = \{$`V-PP`, `NP-PP`, `V-Subj`, `V-Obj`, `ADJ-PP`, `ADV-PP`$\}$. In the syntactic lexicalised space Λ the different lexicalised information (`Syntactic_Type`, `governing_lemma`) will be considered as independent features. For example F_h^{Λ}=`(V-Subj,to-govern)` for t=*equation* or F_k^{Λ}=`(NP-PP,conservation)` for the t=*energy* can be derived from Ex. 1.

The above features can be obtained by shallow parsing of the corpus sentences. In the experiments we have used the CHAOS syntactic parser described in [5]. Notice that syntactic ambiguity in parsing may affect the above observations and frequency counts. Highly ambiguous (but frequent) phenomena (e.g. prepositional phrase attachments) may increase the values for irrelevant features. On the contrary, the pruning of all ambiguous relations may result in too poor evidences. In our approach we use the notion of plausibility of a grammatical relation within an eXtended Dependency Graph (XDG) representation scheme (see [3]). Ambiguous relations r in a dependency graph are given a score $pl(r)$ inversely proportional to the number of conflicting syntactic interpretations. The plausibility $pl(r)$ ranges in the $(0, 1]$ interval: $pl(r) = 1$ if r is unambiguous for the parser, and $pl(r) < 1$ otherwise. The excerpt in Ex. 1.a) generates the XDG in figure 2, where
$pl($`NP-PP,conservation-of-energy`$) = 0.5$
$pl($`VP-PP,govern-of-energy`$) = 0.5$.

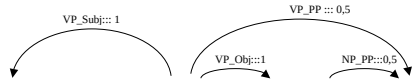

Figure 2. A sample XDG

Grammatical relations local to the source sentence s are thus quadruples (t, s, F_i, p_i) where p_i is the plausibility local to s of the relation between t and F_i. This representation is used to easily obtain the local term vectors. The values given to the features are the shown plausibility in s or 0 if they do not appear. From the graph in Fig 2 the following local vectors are obtained for t =`energy`, in Σ and Λ respectively:
$\overline{v^{\Sigma}}(t, s) = (\ldots (F_i^{\Sigma} = 0.5) \ldots (F_j^{\Sigma} = 0.5) \ldots)$
$\overline{v^{\Lambda}}(t, s) = (\ldots (F_h^{\Lambda} = 0.5) \ldots (F_k^{\Lambda} = 0.5) \ldots)$
where F_i^{Σ}=`V-PP`, F_j^{Σ}=`NP-PP` and F_h^{Λ}=`(V-PP,to-govern)`, F_k^{Λ}=`(NP-PP,conservation)`, For t =`equation` the following vectors are derived instead:
$\overline{v^{\Sigma}}(t, s) = (\ldots (F_l^{\Sigma} = 1) \ldots (F_j^{\Sigma} = 0) \ldots)$
$\overline{v^{\Lambda}}(t, s) = (\ldots (F_m^{\Lambda} = 1) \ldots (F_k^{\Lambda} = 0) \ldots)$
where F_l^{Σ}=`V-Subj` and F_m^{Λ}=`(V-Subj,to-govern)`.

Once local vectors $\overline{v}(t, s)$ are available for each sentence s in the corpus, the global feature vectors representing the global behaviour of the term in the corpus are obtained in the two spaces as follows:

$$v^{\Sigma}(t) = \sum_{s \in C(t)} \overline{v^{\Sigma}}(t, s) \qquad (1)$$

[3] Possibly manual validation at each step can be applied to the most promising choices of the system.

$$v^{\Lambda}(t) = \sum_{s \in C(t)} \overline{v^{\Lambda}}(t,s) \quad (2)$$

where $C(t)$ include the corpus contexts (i.e. the equivalency class) of t.

The values a feature vector assigns to features F_i emphasize the strength of association between the t and F_i. Cumulative plausibility here replaces frequency counts to better model ambiguity in observations. Notice that, for the same F_i, the estimated frequency $\sum_{s \in C(t)} pl(t,s,F_i)$, produces the same ranking as mutual information $MI(t,F_i)$. Feature vectors $v^{\Sigma}(t)$ and $v^{\Lambda}(t)$ are finally normalized to obtain $\hat{v}^{\Sigma}(t)$ and $\hat{v}^{\Lambda}(t)$. These normalized vectors $\hat{v}^{\Sigma}(t)$ and $\hat{v}^{\Lambda}(t)$ are input to the decision tree learner. For sake of comparison, a frequency-based learner has been obtained (feature space Φ) by defining:

$$\hat{v}^{\Phi}(t) = (rf(t))$$

where $rf(t)$ is the relative frequency of t in the corpus. Such discrete space will simulate the behaviour of a quantitative model based on simple frequency.

The above spaces, i.e. the syntactic, the syntactic lexicalised and the frequency-based spaces can be called here "pure". As better results can be obtained if different information is integrated (as also suggested in [4]): contextual information can be used in cooperation with the term frequency. Three other spaces have been thus defined via juxtaposition of the underlying pure vectors, $v^{\Phi}(t)$, $\hat{v}^{\Sigma}(t)$ and $\hat{v}^{\Lambda}(t)$: (1) $\Phi \times \Sigma$, representing frequency and syntactic information; (2) $\Phi \times \Lambda$ that merges frequency and syntactic lexicalised information; (3) $\Phi \times \Sigma \times \Lambda$ merging the three sources of information.

3 Experimental investigation

The aim of the evaluation is to measure the impact of the different source information on the TA process as well as to verify the quality of the term intensional model embodied by the induced decision tree. The experiments have been run over a well-established implicit domain model and statistical validation of results have been obtained by n-fold cross validation. The source domain consists of a corpus of about 250,000 words on the Italian Civil laws, a corresponding thesaurus of 600 terminological expressions built by a team of expert terminologists. The corpus has been processed by the CHAOS parser [3, 5] producing about 3,000 different structures denoting potential terminological expressions. We assumed that the *only* valid term instances are those coded in the thesaurus. We have thus about 1/4 valid structure among the corpus-derived candidates. As a performance figure the error rate ϵ has been adopted as the percentage of misclassified items in the test set, i.e. wrongly accepted corpus candidates. In each 5-fold cross-validation, the system considers an 80% of the corpus candidates as training items (divided evenly between positive terms in the thesaurus and negative items, i.e. nominals that are NOT in the thesaurus). Different decision trees are built for the 5 runs deriving (for the different feature spaces) what is called a domain-specific explicit term model in section 1. The test is then run over the 20% remaining candidates and error rates are then reported as mean values. A baseline TA method assigning about 0.20 probability to each candidate and randomly selecting 1 over 5 candidates has a performance of about 32% error rate [4].

The first test has been carried out within the Φ space. The decision tree based learning looks for discriminating frequency classes among terms and non-terms. Table 1 reports the outcome of this first test. As expected pure frequency (Φ) provide a significant increase in performance ($\epsilon \simeq .17$) with respect to the baseline. However, both exogenous spaces (Σ and Λ) show superior performances. Notice that all the three learning processes make use of negative information. The better "pure" models appear to be the syntactic lexicalised space Λ. This demonstrates that a truly domain-oriented term definition outperforms a general notion of domain importance (as the one based on simple term frequency). An explicit modelling of exogenous grammatical information is very useful given a significant difference in performance (i.e.+18% wrt Φ). This confirms the initial assumption: syntactical lexicalised features capture stable relations between particular lexicals in the domain. The most important rules in one decision tree for the Λ space are for example:

```
(V-Subj,essere) <= 0.00772201 ∧ (NP-PP,contratto)
> 0.00239808 ⇒ Term

(V-Subj,essere) > 0.00772201 ∧ (V-Subj,essere) >
0.153846 ∧ (V-Obj,fare) > 0.0263158 ⇒ Non-term
```

where the `(V-Subj,essere)` is the feature representing the `subject` relation with the verb *essere* (*to be*) whilst `(V-Obj,fare)` represents the `object` relation with the verb *fare* (*to make*). These characterize specific semantic properties of terms as induced from the corpus. The space Σ, although weaker, is characterized by a less expensive training, as it includes about 6 features with respect to the 5,251 features of Λ. It is thus useful to analyse the combination of different features trying to optimise also the learning complexity. In Table 2 the performance of hybrid systems are reported. Combining different sources always outperforms "pure" systems (Table 1). Although lexicalised indicators are always superior (column 3 and 4 in Table 2), the performance over the Φ and Σ is rather good.

Feature Space	Φ	Σ	Λ
Error Rate (%)	16,99	16,235	14,25

Table 1. Final error rate on the Σ, Λ e Φ "pure" spaces.

Feature Space	$\Phi \times \Sigma$	$\Phi \times \Lambda$	$\Phi \times \Sigma \times \Lambda$
Error Rate (%)	15,61	13,88	13,74

Table 2. Final error rate on the $\Phi \times \Sigma$, $\Phi \times \Lambda$, and $\Phi \times \Sigma \times \Lambda$ "frequency" spaces.

A further set of measures has been carried out by simulating an incremental training process closer to the terminologist activity. Negative items are here added to the learning set in increasing portions. The result is a sequence of systems obtained by different training. As positive instances can be derived by pre-existing terminological resources in an *implicit domain model*, repositories of *non-terms* instances do not exist. For these latter the major source is the activity of the terminologists themselves that are involved in the decision

[4] It is the $prob(Refuse|Term) + prob(Accept|NonTerm) = .8 * .2 + .2 * .8 = .32$

process: refused entries enter in the repository of negative example and can be exploited during the manual activity. In order to assess the impact of the negative information, the performance has been be measured over training sets characterized by an increasing number of negative examples. The plot in Fig. 3 describes the error rate variations against percentages of the training negative samples employed. It shows that, whenever the baseline is overcome, the feature space in which a better term model can be induced is still the syntactical lexicalised space Λ. The superiority of Λ features char-

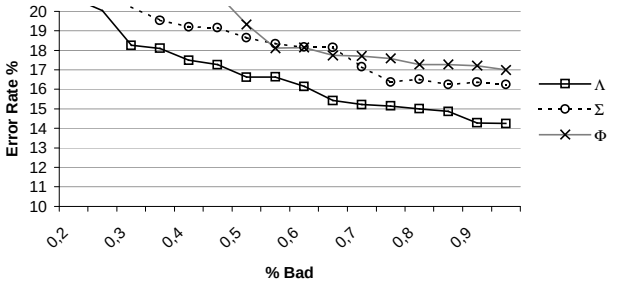

Figure 3. Error rate on the Σ, Λ e Φ "pure" spaces.

acterizes several training set sizes, and this shows that the corresponding learning converges earlier *during* the terminologist work. The best result reachable with the term frequency (i.e. space Φ) is obtained in the Λ space using roughly 60% of the negative training examples. The optimal threshold for the term frequency (i.e. the one obtained with all the training set) is not easily induced by the decision-tree learner. On the other hand, Σ is rarely more effective than simple term frequency. In Figure 4 , the same test is repeated for hybrid learners. The syntactic space becomes more interesting when

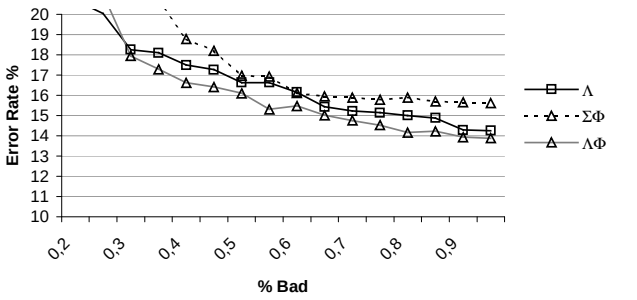

Figure 4. Error rate on the $\Sigma \times \Phi$ and $\Lambda \times \Phi$ "frequency" spaces against the Λ space.

used in combination with term frequency as already suggested by the Table 2. This is also shown by the plots of Fig. 4: the $\Sigma \times \Phi$ and $\Lambda \times \Phi$ space are comparable to Λ. Even if $\Sigma \times \Phi$ is still under the performances obtained by Λ, it is a valuable alternative to the simple frequency for its low computational training costs.

4 Discussion

In this paper, terminology acquisition (TA) has been modelled as an inductive process that generates an explicit term model from an implicit domain model, i.e. a corpus plus a (possibly partial) terminology database. The adoption of linguistically principled descriptions of corpus examples has enabled the derivation of features relevant to the TA activity. The resulting learner (based on decision trees) has been extensively measured. Results suggest that lexicalised and grammatical information about term contexts are effective information in the target model. Performance obtained outperforms previously published results on the same tasks. Moreover, given the richness of the proposed (and implemented) feature representation scheme, the method can be easily adaptable to more complex task. In particular, the availability of term lists enriched with syntagmatic descriptions enable the study of semantic properties over them. As proposed elsewhere, learning of a semantic (rather than simply syntagmatic) model of the underlying terminological concepts is a step towards terminology structuring and ontological induction. The overall high performances obtained by the proposed learning architecture are thus the basis for the on-going research activity in the area ontological induction from texts.

REFERENCES

[1] R. Basili, G. De Rossi, and M.T. Pazienza, 'Inducing terminology for lexical acquisition', in *Preoceedings of the Second Conference on Empirical Methods in Natural Lanaguge Processing, Providence, USA*, (1997).

[2] Roberto Basili, Alessandro Moschitti, Maria Teresa Pazienza, and Fabio Massimo Zanzotto, 'A contrastive approach to term extraction', in *Proc. of the Conference on Terminology and Artificial Intelligence, TIA2001*, Nancy, France, (2001).

[3] Roberto Basili, Maria Teresa Pazienza, and Fabio Massimo Zanzotto, 'Customizable modular lexicalized parsing', in *Proc. of the 6th International Workshop on Parsing Technology, IWPT2000*, Trento, Italy, (2000).

[4] Roberto Basili, Maria Teresa Pazienza, and Fabio Massimo Zanzotto, 'Modelling syntactic context in automatic term extraction', in *Proc. of the 3th Conference on Recent Advances in Natural Language Processing, RANLP2001*, Tzigov Churck, Bulgaria, (2001).

[5] Roberto Basili and Fabio Massimo Zanzotto, 'Parsing engineering and empirical robustness', *Natural Language Engineering*, **to appear**, (2002).

[6] Anne Condamines and Josette Rebeyrolle, 'Ctkb: A corpub-based approach to terminological knowledge base', in *Proceedings of the First Workshop on Computational Terminology COMPUTERM'98, held jointly with COLING-ACL'98*, Montreal, Quebec, Canada, (1998).

[7] Beatrice Daille, *Approche mixte pour l'extraction de terminologie: statistque lexicale et filtres linguistiques*, Ph.D. dissertation, C2V, TALANA, Universitè Paris VII, 1994.

[8] Gregory Grefenstette, 'Evaluation techniques for automatic semantic extraction: Comparing syntactic and window based approaches', in *Proceedings of the Workshop on Acquisition of Lexical Knowledge from Text*, Columus, OH, USA, (1993).

[9] Christian Jacquemin, *Variation terminologique : Reconnaissance et acquisition automatiques de termes et de leurs variantes en corpus. Mémoire d'Habilitation Diriger des Recherches en informatique fondamentale.*, Université de Nantes, Nantes, France, 1997.

[10] Diana Maynard and Sophia Ananiadou, 'Term extraction using a similarity-based approach', in *Recent Advances in Computational Terminology*, eds., Didier Bourigault, Christian Jacquemin, and Marie-Claude L'Homme, (2000).

[11] Emmanuel Morin, *Extraction de liens sémantiques entre termes à partir de corpus de textes techniques*, Ph.D. dissertation, Univesité de Nantes, Faculté des Sciences et de Techniques, 1999.

[12] J.R. Quinlan, *C4.5: Programs for machine learning*, Morgan Kaufmann, San Francisco, CA, 1993.