

# A contrastive approach to term extraction

Roberto Basili, Alessandro Moschitti,  
Maria Teresa Pazienza, Fabio Massimo Zanzotto

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

---

## Résumé

Many approaches to corpus-driven terminology extraction are based on symbolic (i.e. purely syntactic), statistical, and hybrid models (Jacquemin, 1997). Different statistical measures for selecting terminological expressions among candidates observed in the source corpus have been comparatively studied in (Daille, 1994): simple frequency is suggested as the more effective for the task. However, it is still far from representing a *satisfactory* discriminating function. The wide evidence collected by previous studies suggests that term detection should make use of more information than the observable distributional behavior of candidate terms. Better models should be derived over different sample spaces rather than in the refinement of probabilistic measures in the target domain. Traditionally all the suggested measures are related to a single target domain from which distributional information is derived. In this paper a contrastive approach to statistical term extraction based upon selection/filtering criteria that capitalizes on differences among domains is proposed. The method relies on a grammatical candidate extraction component and a cross-domain statistical measure as a term selection model. Experiments over the target domain against a reference terminological database show an improvement of the proposed method over simple frequency.

## 1. Introduction

Different methodologies for addressing the automatic extraction of terms have been proposed: symbolic, statistical, and hybrid approaches (see (Jacquemin, 1997) for an interesting survey). These approaches generally rely on corpora assumed as models of the target knowledge domain. Syntax aims to extract collocations satisfying linguistic constraints, while statistical methods focus on distributional properties that are inherently bound to the corpus. Usually, the search space of the above process is limited to the model of the underlying target domain.

Several statistical measures for selecting terms among matched candidates in corpora have been investigated in (Daille, 1994). A ranking of the observed candidate list is derived upon probabilistic scores ranging from relative frequency to mutual information.

Among them, frequency is proofed as the most effective, as the terms are found within the most frequent candidates. However, it is also noticed that frequency alone is still far from a "perfect discriminating function". The scale of the above mentioned experiments suggests that better models should be defined over richer search spaces rather than by exploring novel distributional measures.

Cases related to *false positives* (i.e. candidates that are very frequently matched in the target corpora but that are not terms in the domain) as *fine rapporto* (\*end of the relation) and *via principale* (\*main way) are very common. Notice how they are always related to jargon or, even worse, largely used language-specific collocations. In the legal corpus used for our experiments they appear 111 and 86 times, respectively. Such a high number of occurrences prevent any method driven by corpus frequencies to prune them from the target terminology. These errors are the major performance pitfalls of statistical filtering techniques (Daille, 1994).

The main weakness is here given by the adoption of frequency *within* the target corpus as the only selective criteria. Notice that terms are "domain" properties, (Kageyra *et al.*, 1998), and not just "document" properties (as keywords in IR). The role of the domain in the term extraction process should be thus strengthened. Learning specific properties of an object means comparing it with other objects and generalizing the characterizing features. Similarly, assessment of (domain-specific) terms means studying them across different domains and assessing their specificity (if any). Availability of text collections for much different domains will represent a different model that seems more useful for contrastive analysis: the *farer* are the underlying topics, the more selective will be the filtering criteria. Purely language-dependent phenomena should spread similarly across different collections, while domain specific expressions should exhibit odd behaviours. Sorting terms according to a cross-domain score should realize a superior selective process. This criteria would be very important mainly for *singleton* terms, i.e. terminological entries made of a single word like *imposta* (tax), *udienza* (hearing), *causa* (lawsuit) in a legal domain. These *simple terms* are usually more polysemic than *complex terms* (i.e. multiword terms). Furthermore, (Kister, 1993), they are often elliptic occurrences of *complex terms*. Important statistical properties exist between *simple terms* (*TS*) and *complex terms* (*TC*). The activity of ranking candidate terms can be designed taking to account for these aspects as investigated in (Basili *et al.*, 1997; Pazienza, 2000). We will make the proposed contrastive model sensitive to such differences.

In this paper, a term extraction method based on contrastive analysis across domains is proposed. It uses a syntactic approach to match candidate terms in a source corpus (Sect. 2.1) and a multi-level cross-domain statistical approach (Sect. 2.2) to select proper terminological expressions. Large-scale experiments have been run on the target domain of the Italian Civil Code: a large manually controlled term database<sup>1</sup> has been used for comparative evaluation. Results are discussed in Section 3.

## 2. A contrastive model for corpus-driven term extraction

Corpus-driven term extraction should rely on syntactic as well as distributional information. The model we propose hereafter is similarly hybrid. Firstly, natural language processing techniques are applied to the source corpora to match grammatically suitable

---

<sup>1</sup>The terminology has been developed by human experts at the European Academy of Bolzano-Italy.

expressions for simple and complex "candidate" terms. Although the method is targeted to a single domain (i.e. hereafter the *target* domain), several domains are under investigation. A separate processing for each of the related corpora is applied. Then, a statistical filtering process is run in order to exploit selective differences and accept or reject candidates for the target domain. Fig. 1 shows the general architecture of the method.

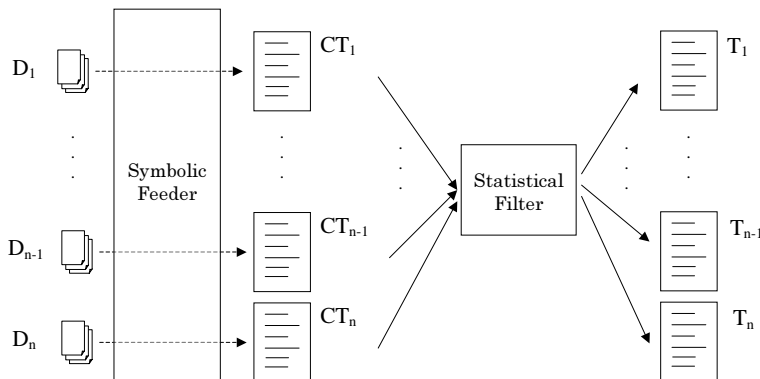


Figure 1: Comparing candidates from different domains

For each domain  $D_i$ , a collection of candidate terms  $CT_i$  is produced via shallow parsing techniques (the *symbolic feeder* in Fig. 1). These unordered candidate collections are then ranked according to the contrastive analysis by the *statistical filter*.

In the next sections details on the syntactic detection of candidate terms is first reported (Sect. 2.1). Then (Sect. 2.2), the statistical filtering technique is described, by introducing the overall processing steps (Sect. 2.2.1) and then formalizing the measures (Sect. 2.2.2 and 2.2.3).

### 2.1. The Symbolic Feeder

The process for automatically extracting terms from raw text must rely on a wide linguistics knowledge; in fact complex surface structures, representing terminological concepts, have first to be recognized. The candidate extractor used in our system is based on the CHAOS shallow parser developed within several NLP applications (Basili *et al.*, 2000). The term extractor component, called the *symbolic feeder*, is based mainly on grammatical constraints imposed on the parser output.

The CHAOS parser applies a cascade of processing modules: (a) a *tokenizer*, matching words from character streams; (b) a *yellow page look-up module* that matches named entities existing in catalogues; (c) a *morphologic analyzer* that attaches (possibly ambiguous) syntactic categories and morphological interpretations for each word; (d) a *named entities matcher* that recognizes complex named entities according to special purpose grammars; (e) a rule-based *part-of-speech tagger*; (f) a *POS disambiguation module* that resolves potential conflicts among the results of the POS tagger and the morphologic analyzer; (g) a *syntactic parser* based on modularization and lexicalization: it builds a chunk-based representation of the input text, including major grammatical dependencies among chunk heads. Details of the parser can be found in (Basili *et al.*, 2000).

The syntactic information is gathered by CHAOS in a formalism called extended dependency graph (*XDGs*). Nodes are chunks and edges are syntactic dependencies among chunks (inter chunk dependencies, *icds*). Given a sentence  $s$  of an input text, a graph  $xgd_s = (C, L)$  is produced, where  $C$  is the set of constituents (i.e. *chunks* detected in  $s$ ) and  $L$  is the set of valid *icds*. For instance, the representation of the grammatical information extracted for the sentence:

*Le spese dell'apposizione dei sigilli, dell'inventario e di ogni altro atto dependente dall'accettazione con beneficio d'inventario sono a carico dell'eredità.*<sup>2</sup>

is:

[1/C\_Nom *Le spese*] [2/C\_Prep *dell'apposizione*] [3/C\_Prep *dei sigilli*] [4/C\_Cong ,] [5/C\_Prep *dell'inventario*] [6/C\_Cong *e*] [7/C\_Prep *di ogni altro atto*] [8/C\_VerInf *dependente*] [9/C\_Prep *dall'accettazione*] [10/C\_Prep *con beneficio*] [11/C\_Prep *d'inventario*] [12/C\_VerFin *sono*] [13/C\_Prep *a carico*] [14/C\_Prep *dell'eredità*] [15/C\_Cong .]

The derived inter-chunk dependencies are shown in Fig. 2.

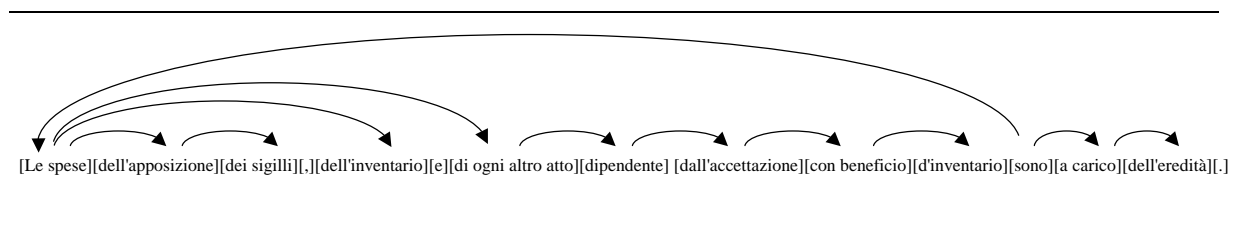


Figure 2: Sample sentence inter-chunk links

The *XDG* formalism allow a rich representation of structures interesting for candidate term matching. Connected and contiguous subgraphs  $p = (C_p, L_p)$  of a  $xgd_s = (C, L)$  (with  $C_p \subseteq C$  and  $L_p \subseteq L$ ) can be suitably constrained to suggest valid terminological expressions. We define as a *partial phrase* the subgraph  $p = (C_p, L_p)$  such that it has a *single head*, i.e.

$$\exists! \alpha \in C_p \text{ such that } \forall \beta \in C_p, \text{ with } \beta \neq \alpha \Rightarrow (\beta, \alpha) \notin L_p \quad (1)$$

Single headed partial phrases are valid candidates for terminological expressions whenever (a) they represent specific constituents (i.e. complex nominals) and (b) their inner structure is a suitable one for terminological expressions of the domain. Specific grammatical constraints can be easily imposed to the set  $L_p$  of the internal dependencies.

Information (a) can be constrained by limiting the set of chunk types to be accepted. Single headed partial phrases have just on constituent playing the role of head. Other constituents (in  $C_p$ ) are (not necessarily, head-) modifiers. We will call these latter *no-heads* constituents. For instance, in the partial phrase<sup>3</sup>

$$[1/C\_Nom \textit{Le spese}] [2/C\_Prep \textit{dell'apposizione}] [3/C\_Prep \textit{dei sigilli}] \quad (2)$$

the nominal chunk 1/C\_Nom play the role of head, while the prepositional chunks 2/C\_Prep and 3/C\_Prep are no-head constituents.

Partial phrases that are legal candidate terms can be heads or no-heads constituents. Decision rules (i.e. constraints) are imposed on the chunk type (e.g. **C\_Nom** for nominal

---

<sup>2</sup>A raw translation of the sentence: *\*All expenses to put seals, for the inventory, and for any other action consequence of accepting with reservation are at legacy expenses*

<sup>3</sup>*The expenses to put seal*

chunks, or C\_Prep for prepositional chunks) as well as on the chunk internal structure. An excerpt of the adopted constraining rules is reported in Table 1. The HEAD constraint

	Type Constraint	Structure Constraint
HEAD	C_Nom	?* Noun
	C_Nom	?* Adj Noun
	C_Prep	?* Noun
	C_Prep	?* Adj Noun
NO-HEAD	C_Prep	Prep Noun

Table 1: Examples of candidate matching rules.

definitions implies that any nominal or prepositional chunk ending with a noun (line 1), or an adjective followed by a noun (line 2) is a legal term candidate. According to rules in Table 1, the partial phrase 2 is a legal candidate: the resulting candidate term is:

*spese dell'apposizione dei sigilli*

while its head is *spese*. Note that the useless determiner *le* is eliminated given its matching with any, ?, in the constraining rule. NO\_HEAD constituents are matched similarly, according to type (first column) and structural (second column) constraints.

The adoption of a lexicalized shallow parser increase the locality of the exploited grammatical information in the control of the sentence structure<sup>4</sup> so that no limit to length and inner complexity is imposed to the matched candidates.

After the symbolic analysis is applied, a list of (structured) candidate terms and their corresponding heads are obtained. These candidate terms are naturally mapped into in complex and simple candidates, respectively.

## 2.2. A statistical filter for contrastive term selection.

Aim of the statistical filter is to prune from the candidate set those expressions that do not characterize domain specific concepts. Specificity to the target domain  $D$  should emerge from the differences in the distributional behavior throughout the independent corpora  $C_1, \dots, C_n$  used to build candidates.

In order to capture the information required for selective decisions, the model should rely on the internal structure of candidates (head-complex candidate relations) and on the distributions of simple and complex candidates in the target domains as well as in the other corpora.

The presence of modifiers (mainly C\_Prep for Italian) in the complex candidates is a meaningful mechanism for term formation in specific domains. On the other side *simple terms*, e.g. *Institute*, express a sort of *background knowledge*: this should be observable throughout the different corpora. For this reason, simple terms have higher frequency and this strengthen the reliability of inferences drawn upon their distributions. The selection of complex terms in the target domain (*foreground knowledge*, e.g. *Massachussetts Institute of Technology*, *British Film Institute*, *Institute for Contemporary Studies*) is done

<sup>4</sup>The use of lexicalized subcategorization frames in CHAOS is discussed in (Basili *et al.*, 1999b))

according to contrastive information related to simple ones. For the above reasons a cascade of statistical inferences is imposed and a complex architecture is derived.

### 2.2.1. A layered approach: the statistical filtering

The input of the statistical filter are two lists of candidates for each analyzed domain  $D_i$ : the list of simple candidate terms ( $CTS_i$ ) and the list of complex candidate terms ( $CTC_i$ ). As relevant differences in the distributional behavior are expected for the two classes of candidates, different filtering functions are used: metrics able to capture cross-domain differences are independently applied to the two lists. Accordingly, the *statistical filter* accomplishes the suitable selection of candidate terms in two steps:

- By first, simple candidate terms  $st$  are selected by a function  $w_{st}^i$  based on their distributions in the target ( $i$ ) as well in the other corpora. This *comparative statistical measure* produces a ranked list of simple candidate terms  $TS_i$  for each domain  $D_i$ .
- Then, each complex term  $ct$  is scored by a function  $cw_{ct}^i$  based upon: (1) its probability (i.e.  $f_{ct}^i$ ) in the target domain as well as (2) the comparative measure (introduced in the previous step) as observable for their heads, i.e.  $w_{h(ct)}^i$  where  $h(ct)$  is the head of the complex term  $ct$ . In other words, the global ranking of a complex term  $ct$  depends upon both on its frequency in the target corpus ( $f_{ct}^i$ ) and the comparative analysis of its head ( $cw_{h(ct)}^i$ ).

The resulting ranked lists of complex and simple terms are then pruned and form the terminological database proposed to experts for the target domain  $D$ . According to the above steps, the architecture early proposed in Fig. 1 can be revised as in Fig. 3, where the inner structure of the statistical filter is shown.

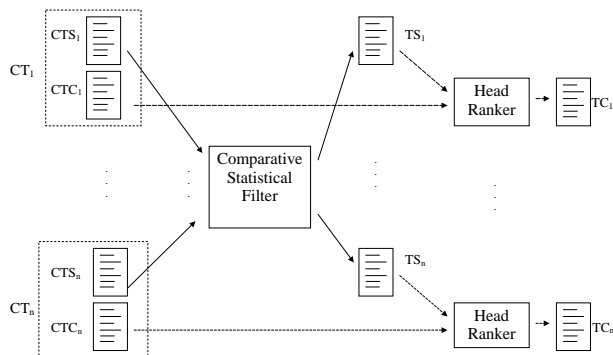


Figure 3: The contrastive statistical filter

### 2.2.2. A Contrastive weight for simple terms

In order to emphasize differences among corpora measures of relevance (as indexing scheme) in different collections of documents can be used. For instance, *Inverse Document Frequencies IDF* (Salton, 1991) emphasizes differences in the distribution of potential indexes through documents in a target bibliographic database. Similarly, *Inverse Word*

*Frequencies* ( $IWF(t)$ ), defined in (Basili *et al.*, 1999a) for a text categorization task, measures differences in the behaviour of indexes throughout a set of topics (i.e. set of training documents in given categories). For its higher robustness (as experimented in (Basili *et al.*, 1999a)), we decided to select IWF instead of IDF as our contrastive filter. More formally, given a candidate  $t$  and its cumulative frequency  $F_t = \sum_j f_t^j$  throughout all domains, the *Inverse Word Frequencies* is

$$IWF(t) = \log\left(\frac{N}{F_t}\right) \quad (3)$$

where  $N$  is the size of the corpus obtained by summing up contributions (i.e. frequencies) of all candidates in all domains. The *IWF* (as *IDF*) penalizes high-frequency (and dispersed) candidate terms although it is not related to the notion of document. As IWF is by no way related to the target domain  $D_i$ , the weighting function has also to take into account the frequency of term  $t$  in the target domain  $i$ . Similarly to the well known vector space models the weighting function can be obtained as:

$$w_t^i = \log(f_t^i) * IWF(t) \quad (4)$$

where  $f_t^i$  is the frequency of the simple term candidate  $t$  in the target domain. The function introduced by equation 4 will be hereafter called *contrastive weight*. Contrastive weights provide an effective ranking of simple term candidates as described in experimental section 3.

### 2.2.3. Ranking of complex candidate terms using the heads

The ranking of complex term candidates is based upon contrastive weights of simple terms (Eq. 4) as applied to their heads. The use of heads provide a more robust ranking according to the higher frequencies of heads: data sparseness do not allow an easy estimation of domain probabilities for complex terms that are usually very rare.

In synthesis the relevant information for the scoring of a complex term  $ct$  is thus: (1) contrastive weight of its head  $h(ct)$  with respect to the target domain  $i$ , and (2) frequency of  $ct$  in  $D_i$ . Given a complex term candidate  $ct \in CTC_i$ , a measure  $cw_{ct}$ , hereafter called *Contrastive Selection via Heads*, can be formally defined by:

$$cw_{ct}^i = w_{h(ct)}^i \cdot f_{ct}^i \quad (5)$$

where  $f_{ct}^i$  is the frequency of  $ct$  in the  $i$ -th corpus and  $w_{h(ct)}^i$  is contrastive weight of its head (Eq. 4).

## 3. Testing the performances of contrastive measures.

Our approach has been experimented on a target corpus (*JCC*), the Italian Civil Code of approximately 1,400 judgments of the Italian Corte di Cassazione (Italian Supreme Court). A second corpus (*NEWS*) has been used to contrast *JCC* and was made of a collection of 6,000 news ranging on different domains (Sports, Politics, Economics). The reference target terminology *RefT* of the *JCC* domain is given by 2,000 (manually extracted and validated) terms.

On the two corpora the symbolic feeder produced simple and complex candidate terms reported in Table 2.

Domain		Collection size	# Terms in the collection
JCC	Simple Terms	15,989	595
	Complex Terms	63,924	348
NEWS	Simple Terms	27,521	no terminology available
	Complex Terms	34,869	available

Table 2: Candidate term vs. text collections

It is worth noticing that terms of *RefT* that appear in the *JCC* corpus are less than an half (943/2000) We will call these terms as overlapping, *TG*. It is clear that *JCC* (that is a controlled and extensive legal corpus) is by no way able to fully represent the target legal domain (to which *RefT* refers). It is thus clear that statistical filters have an inherent upper bound of 0.47 coverage of the phenomena (i.e. *RefT*). Measurements are done taking the overlapping terms *TG* as the standard. Although recall and precision measures are not fully capturing the nature of the task (extracting candidates to be submitted for validation to human experts), they provide significant evidence for comparative purposes.

In order to compare the rankings obtained by contrastive weights ( $w_t^i$  and  $cw_t^i$ ) and pure frequency  $f_t^i$ , *F*-measure:

$$F = \frac{1}{0,5/p + 0,5/r} \quad (6)$$

where  $p$  and  $r$  refer to traditional recall and precision respectively, is used. The *F*-measure is computed with respect to the set of overlapping terms *TG*, according to different portions of the ranking. Partitions of the ranked lists are first created so that the  $k$ -th partition includes the first  $k$  members of the list. *F*-measure is thus computed over such  $k$  accepted candidates with respect to the *TG* golden standard. The plots in 4 and 5 report data for the simple and complex terms respectively.

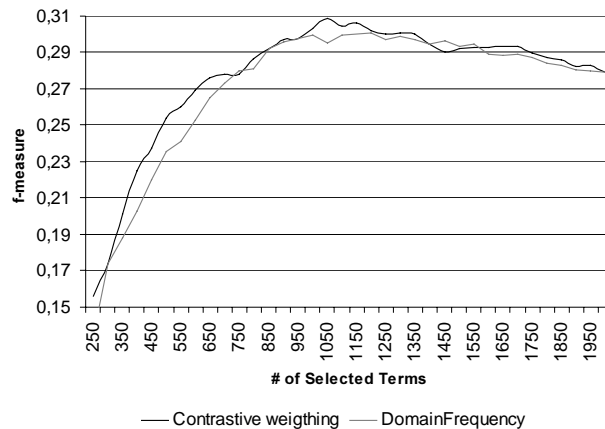


Figure 4: Simple Terms in the judgments of the Corte di Cassazione

## 4. Discussion

Several interesting implications can be drawn from the data plotted in figures 4 and 5.

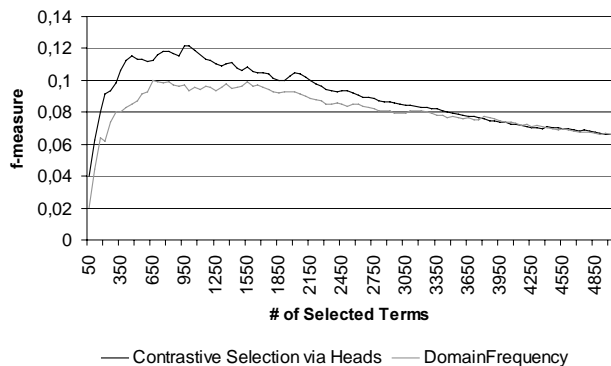


Figure 5: Complex Terms in the judgments of the Corte di Cassazione

The proposed statistical measure (Eq. 4 and Eq. 5) outperforms in both cases pure frequency: the proposed ranking seems to include better candidates earlier in the list (i.e. higher values of  $F$ -measure suggested by the first values of  $k$ ). Implications of this improvement in the work of human validators is evident: the sooner valid candidates are met in the list the faster is the overall process of term identification.

The overall performance, as quantified by the  $F$ -measure score is not striking. The pervasive phenomena of jargon, acronyms affects both measures, as over-generation of the symbolic feeder is problematic. However, the reported figures are obtained by only grammatical constraints (as those reported in Table 1). No simple heuristics is imposed like elimination of closed word classes, i.e. numbers, dates or known multiwords/stoplists. The main reason is to fully evaluate the robustness of this method on real data, something that is often neglected in "in vitro" experiments. Real applications are usually faced with a huge quantity of these phenomena and actual robustness should be always tested.

Comparative evaluation with other proposed measures has been indirectly carried out. As (Daille, 1994) has proofed frequency as the simpler and most effective scoring function, we can capitalize on this and assume the contrastive scheme of Eq. 4 and Eq. 5 also outperforms several other domain-confined techniques. For example, the exploitation of distributional information from other domains is relevant for decreasing scores of language dependent patterns, e.g. *fine rapporto* (\*end of the relation) moving from position 47 to 85 in the ranking. On the contrary, *atto di pignoramento* (\*act of distaining) and *consulente tecnico* (*technical consultant*), i.e. two terms in the reference terminology  $RefT$ , gain 100 and 50 positions, respectively.

The last point to be discussed is the overall evaluation framework that is very difficult to assess in term extraction. We assumed the  $RefT$  as a golden standard. The desirable property of this resource is that it is highly controlled, as teams of legal experts have validated the outcome of the terminologist work. The problem with corpus-driven methods is over-generation: although most of the system decisions (after statistical filtering) appear reasonable, good suggested terms are not in the reference standard. For example, *procedimento penale* (criminal proceeding) is brought by cross-domain ranking from position 156 to 139 in the resulting list: it appears to non experts as a perfect domain specific concept (and possibly it is something we would search for a definition when reading legal prose). It is indeed missing from the golden standard. As a result measures based on  $RefT$  and

on recall/precision scores are inadequate to fully capture the nature of the task.

In conclusion the proposed hybrid approach to terminology extraction from corpus processing has been here experimented in the best possible conditions given the available resources. Although more insight is needed on the definition of suitable syntactic constraints for term recognition in corpora, the main outcome of this work is the assessment of the role of contrastive analysis in term selection. The benefits of cross-domain analysis is confirmed by all the reported measures. However, the bias of the adopted resources has to be stressed: documents dealing with legal judgments cover any real world aspect. The term extraction process is thus much complex in such a domain. More experimental evidence over other domains is needed and will be part of future research activity.

## Références

- BASILI R., DE ROSSI G., PAZIENZA & M.T. (1997). Inducing terminology for lexical acquisition. In *Proceedings of the Second Conference on Empirical Methods in Natural Lanague Processing, Providence, USA*.
- BASILI R., MOSCHITTI A. & PAZIENZA M. (1999a). A text classifier based on linguistic processing. In *Proceedings of IJCAI 99, Machine Learning for Information Filtering*, <http://www-ai.cs.uni-dortmund.de/EVENTS/IJCAI99-MLIF/papers.html>.
- BASILI R., PAZIENZA M. T. & ZANZOTTO F. M. (1999b). Lexicalizing a shallow parser. In *Proc. of the Traitement Automatique de la Langue Naturelle, TALN99*, Cargese, FR.
- BASILI R., PAZIENZA M. T. & ZANZOTTO F. M. (2000). Customizable modular lexicalized parsing. In *Proc. of the 6th International Workshop on Parsing Technology, IWPT2000*, Trento, Italy.
- DAILLE B. (1994). *Approche mixte pour l'extraction de terminologie: statistque lexicale et filtres linguistiques*. PhD thesis, C2V, TALANA, Université Paris VII.
- JACQUEMIN C. (1997). *Variation terminologique : Reconnaissance et acquisition automatiques de termes et de leurs variantes en corpus. Mémoire d'Habilitation Diriger des Recherches en informatique fondamentale*. Nantes, France: Université de Nantes.
- KAGEYRA K., YOSHIOKA M., KOYAMA T. & NOZUE T. (1998). Towards a common testbed for corpus-based computational terminology. In D. BOURIGAULT, C. JACQUEMIN & M.-C. L'HOMME, Eds., *Proc. of the 1st Workshop on Computational Terminology COMPUTERM'98, held jointly with COLING-ACL'98*, Montreal, Quebec, Canada.
- KISTER L. (1993). *Groupes nominaux complexes et anaphores: possibilité de reprise pronominale dans "N1 de (dét.) N2 "*. PhD thesis, Sciences du Language, Université de Nancy.
- PAZIENZA M. T. (2000). A domain-specific terminology-extraction system. *Terminology*, **5:2**.
- SALTON G. (1991). Development in automatic text retrieval. *Science*, **253**, 974–980.