Inducing Terminology for Lexical Acquisition

Roberto Basili, Gianluca De Rossi, Maria Teresa Pazienza
Department of Computer Science, Systems and Production
University of Roma, Tor Vergata
{basili,derossi,pazienza}@info.utovrm.it

Abstract

Few attention has been paid to terminology extraction for what concerns the possibilities it offers to corpus linguistics and lexical acquisition. The problem of detecting terms in textual corpora has been approached in a complex framework. Terminology is seen as the acquisition of domain specific knowledge (i.e. semantic features, selectional restrictions) for complex terms and /or unknown words. This has useful implications on more complex text processing tasks (e.g. information extraction). An hybrid symbolic and probabilistic approach to terminology extraction has been defined.

The proposed inductive method puts a specific attention to the linguistic description of what terms are as well as to the statistical characterization of terms as complex units of information typical of domain sublanguages. Experimental evidence of the proposed method are discussed.

1 Introduction

Nowadays corpus processing techniques are widely adopted to approach the well-known lexical bottleneck problems in language engineering. Lexical acquisition methods rely on collocational analysis (pure statistics), robust parsing (syntax-driven acquisition) or semantic annotations as they are found in large thesaura or on-line dictionaries. The lexical information that trigger induction varies from simple word/tokens to syntactically annotated or semantically typed collocations (e.g. *powerful* vs. *strong tea* (Smadja,1989)), syntactic disambiguation rules (e.g. (Hindle and Rooths,1993), (Brill and Resnik,1994)) or sense disambiguation rules are usually derived. Such information is lexical as it encodes constraints (of different types) at the word level, to be thus inherited by morphologic variants of a given lemma.

This strongly lexicalized knowledge, as it is extracted from corpus data, requires lexical entries to be known in advance in some morphologic database. POS taggers or lemmatizers are generally used to suitably map tokens to lemmas. It should be noted that lemmas in a corpus depends on the underlying sublanguage and their nature and shape is not as general as it is usually encoded in a morphologic dictionary. As an example, let *studio* (i.e. *study* as a noun) be an entry in an italian morphologic dictionary. Typical information in such a database is the following:

studio `pos=noun gen=mas num=sing`

The only legal morphologic variant of *studio* is *studi* (*studies*, with `num=plur`). When searching for *studio* in a corpus of environment related texts[1], we found this kind of occurrences (e.g. short contexts):

... studi di base ..., (basic studies)
... studi di impatto ambientale ...,
*(*studies on the environmental impact)*
... studi di fattibilitá ..., (feasibility studies),
... studi di riferimento..., (reference studies)

It is very common in a corpus (not balanced, thus focused to a limited domain) to find a set of specifications of nouns that have some specific properties:

- they are not always compositional (e.g. *studio di base*);
- they describe complex concepts (e.g. *studi di fattibilitá*) in the underlying technical domain, so they are relevant for text understanding and classification/extraction

[1] Although our approach is in principle language independent, we systematically will describe rules and examples in italian as they have been derived from text corpora in Italian. The environmental corpus, called ENEA, is a collection of short scientific abstracts or newspaper articles dealing with pollution.

- they select specific and independent senses of the related term: *studi di base* refers to the abstract notion of study as an on-going reasearch, while *studi di fattibilita'* is not a reasearch but a specific engineering task ;

- the related nominal compounds show independent lexical properties. For example, all the examples are potential object of verbs like *carry out, do, ...* but only *feasibility studies* or *studies on the environmental impact* can be modelled by some techniques or policies. Furthermore, *studies on the environmental impact* have specific social and political implications that are no longer valid for the general notion of study.

In the same environmental corpus the typical short contexts of the lemma *attivitá* (*activity*) include notions like:

attivitá umana (*human activity*),
attivitá entropica (*hentropic activity*),
attivitá di costruzione (*building activity*).

These very common instances show that lexical acquisition for *attivitá* or *studio* cannot be fully accomplished without discriminating the lexical properties of such pure collocations from those related to their complex nominals. The results of lexical acquisition should thus be different for entries like *attivitá* and *attivitá entropica*.
The underlying hypothesis is that complex concepts related to a lemma do not support all the generalizations related to the source lemma. In fact, whenever a concepts is built it acquires an autonomous role within a language so it behaves in an almost independent fashion. In order to capture the essential differences we need to select the proper set of terms in a given sublanguages, formalize them into independent lexicalizations and carry out a separate lexical acquisition for each of them.

A further aspects that is worth to be mentioned is that *terms* are generally understood as single lexical units during syntactic recognition. They are sentence fragments already parsed. Robust methods widely employed in computational linguistics are thus sensible to a precise recognition of terms, as much of the ambiguity embedded within the term structures simply disappear after ercognition has been accomplished. Let for example be *attivitá di costruzione* or *articoli da spiaggia* (*beach articles*) two terms. Sentence fragments like

... l'inizio della attivitá di costruzione ...
the start of the building activity

or

... trasportavano articoli da spiaggia ...
they transported beach articles,

although inherently ambiguous (*l'inizio della costruzione* and *trasportavano da spiaggia* are sentence readings that also obey to selectional constraints (e.g. *to transport/bring from a place*)) can be correctly parsed when the two terms are employed before syntactic analysis is triggered. Applying syntactic driven lexical acquisition (e.g. (Grishman and Sterling,1994) or (Basili et al.,1996)) after corpus specific term recognition and extraction highly improve the precision and complexity of the parsing activity. Experimental evidence will be discussed in later sections.

In synthesis corpus driven terminology definition and recognition has positive implications on LA:

- Terms rather than words are the atomic units of information on which LA applies: more selective induction thus results in a more precise acquisition

- Terminologic variants of a given term are hints for domain specific word sense disambiguation

- Terms are sentence fragments that have been already parsed: the lower ambiguity resulting from term recognition has a beneficial effect on the later syntagmatic analysis of the corpus

2 Terminology and Lexical Acquisition.

In this framework, a *term* is more than a token or word (to be searched for) as it stands in a more subtle relation with a piece of information in a specific knowledge domain. It is a concept, as it requires a larger number of constraints on the information to be searched for in texts. Furthermore a *term* conveys a well assessed (usually complex) meaning as long as a user community agrees on its content. As long as we are interested in automatic terminology derivation, we can look at terms as surface canonical forms of (possibly structured) expressions indicating those contents.

A term is thus characterized by a general commitment about it and this has some effects on its usage. Distributional properties of complex terms (nominals) differ significantly on those of their basic elements. Deviance from usual distributional behavior of single components can be used both as marker of non compositionality and specific hints of domain relevance. The detection of complex terms

assumes a crucial role in improving robust parsing and POS tagging for lexical acquisition, thus supporting a more precise induction of lexical properties (e.g. PP disambiguation rules). This specific view extends and generalizes the classical notion of terminology as used in Information Science.

Most of the domain specific terms we are interested to are nouns or noun phrases that generally denote *concepts* in a knowledge domain. In order to approach the problem of terminological induction we thus need:

1. to extract surface forms that are possible candidates as concept markers;
2. to decide which of those candidates are actually concepts within a given knowledge domain, identified by the set of analyzed texts.

Linguistic principles characterize classes of surface forms as potential terms (step 1). Note that the notion of *terminological legal expression* here is not equivalent to that of *legal noun phrases*. Concepts are lexicalized in surface forms via a set of operations that imply semantic specifications. The way syntax operates such specification may be very complex and independent on the notion of grammatical well formedness.

The decision in step (2) is again sensible to a principled way a language expresses concept specifications but needs also to be specific to the given knowledge domain, i.e. to the underlying sublanguage. Given the body of texts, the selective extraction should be sensitive to the different observed information. In this phase statistics is crucial to control the relevance of linguistically plausible forms of all the guessed terms.

3 Integrating linguistic and statistical information for term discovery

The principled definitions of legal grammatical structures by which terms are expressed and the description of their distributional properties in a sublanguage are crucial for the automatic construction of a domain terminological dictionary. A number of methods for language driven terminological extraction and complex nominals parsing and recognition have been proposed to support NLP and lexical acquisition tasks. They mainly differ in the emphasis they give to syntactic and statistical control of the induction process. In (Church,1988) a well-know purely statistical method for POS tagging is applied to the derivation of simple noun phrases that are relevant in the underlying corpus. On the contrary more language oriented methods are those where specialized grammar are used. LEXTER (Bourigault,1992) extracts maximal length noun phrases (*mlnp*) from a corpus, and then applies a special purpose noun phrase parsing to them in order to focus on significant complex nominals. Although the reported recall of the *mlnp* extraction is very high (95%) the precision of the method is not reported. Voutilanen (1993) describes a noun phrase extraction tool (*NPtool*) based upon a lemmatizer for English (ENGTWOL) and on a Constraint Grammar parser. The set of potential well-formed noun phrases are selected according to two parsers working with different NP-hood heuristics. A very high performance of NP recognition is reported (98.5% recall, and 95% precision).

A more statistically oriented approach is undertaken in (Daille et al,1994) where a methodology for syntactic recognition of complex nominals is described. Linguistic filters of morphological nature are also applied. Corpus driven analysis is mainly based on mutual information statistics and the resulting system has been successfully applied to technical documentation, e.g. telecommunication.

All these methods deal with the problem of NP recognition. As we are essentially interested to *NP* that are actual terms in a domain, we will need to decide *which NPs* are actual terms. We will define:

1. well formedness principia for term denotations and a description of the different grammatical phenomena related to terms of a language
2. distributional properties that distinguish terms from other (accidental) forms (e.g. non terminological complex nominals).

3.1 Grammatical descriptions of terms in Italian

It is generally assumed that a terminologic dictionary is composed of a (possibly structured) list of nouns, or complex nominals. Nominal forms are in fact lexicalization of domain concepts: proper nouns, acronyms as well as technical concepts are mostly represented as nominal phrases of different length and complexity. For this reason, we concentrated only on noun phrases analysis, as the main source of terminologic information[2]. A term is obtained by applying several mechanisms that add to a source word (generally a noun) a set of further specifications (as additional constraints of semantic nature).

[2] In lexical acquisition the role of other syntactic categories (e.g. verbs, adjectives, ...) is also very important but the set of phenomena related to them is very different as also outlined by (Basili et al.,1996b)

A detailed analysis of the role of syntactic modifiers and specifiers (De Rossi,1996) revealed that legal structures for modifiers and specifiers in Italian are mainly of two types:

1. *restrictive (or denotative) modifiers* (postnominal participial, adjectival or prepositional phrases)
2. *appositive (or connotative) modifiers* (prenominal modifiers, i.e. adjectival phrases)

Restrictive modifiers are generally used to constraint the semantic information related to the corresponding noun, via a further specification of a given *type*for that noun as in *scambi commerciali* (**exchanges commercial*): the referent noun is forced to belong to a restricted set of *exchanges* (that are in fact of commercial nature). On the contrary, appositive modifiers are used by the speaker/writer to add additional details: his own point of view or pragmatic information, as in *la bianca cornice* (*the white frame*) or *la perduta gente* (*the lost people*). Appositive modifiers do not correspond to any (shared) classification, but rather to the subjective speaker's point of view. Furthermore prenominal modifications are rather unfrequent in Italian. We thus decided to focus only on restrictive modifiers, the best candidates to bring terminological (i.e. assessed classificatory) information. The set of syntactic phenomena that have been studied as good candidates for restrictive forms are:

1. adjectival specification (via postnominal adjectives, as in *inquinamento idrologico* (**pollution hydrological*))
2. nominal specification (postnominal appositions, as in *vagone letto* (*wagon-lit*), or *Fiat Auto* (*Fiat Cars*))
3. locative phenomena (postnominal proper nouns indicating locations, as in *IBM Italia*
4. verbal specification (via postnominal past participle, as in *siti inquinati* (**sites polluted*))
5. prepositional specification (via a particular set of postnominal prepositional structures, as in *Istituto di Matematica* (*Institute of Mathematics*), or *barca a vela* (*sailing-boat*)). [3]

Given the above linguistic principles, a special purpose grammar for potential terminological structures can be sketched. With a simple language of regular expressions the grammar of adjectival, prepositional and participial restrictions can be expressed as:

Term ← *noun* *A_P**
Term ← *noun* *A_P* (*Cong* *A_P*)*
Term ← *noun**
Term ← *noun* (– *noun*)*
Term ← *noun* , *Term*
A_P ← *adjective* | *past_participle*
Cong ← '–' | *e*

Prepositional postmodifiers are modeled according to the following rules:

Term ← *noun* *P_P**
P_P ← *Prep* *noun*
Prep ← *di* | *a* | *da* | *per*

Note that the allowed structures are post nominal due to the typical role of specifications in Italian.

3.2 Distributional properties and term extraction

The recursive nature of some rules require an iterative analysis of the corpus. The following algorithm is used:

1. Select singleton nouns whose distributional properties are those for terms and insert them in the terminologic dictionary (TD)
2. Use the valid terms in TD to trigger the grammar and build complex nominals *cn*
3. Select those *cn* whose distributional properties are those for terms and insert them in TD.
4. Iterate steps 2 and 3 to build longer *cn* [4].

Note that newly found complex terms, added to TD in step 3, force a re-estimation of term probabilities obtained by a further corpus scanning, so that their heads are not counted twice.

The validation of a limited set of potential surface forms as actual terms is crucial for lowering the complexity of the above algorithm. Given the grammar, we need criteria to decide which surface forms, that reflect the typical structure of a potential terms, are actual lexicalizations of relevant concepts of the corpus. The kind of observations that are available from the corpus are: (*i*) the set of lemmas met in the texts, (*ii*) the set of their well formed restrictions (i.e. complex nominals) and (*iii*) the distributional properties of entries in (*i*) and (*ii*). We firstly establish when a singleton lemma is a relevant concept by using distributional properties of nouns. Then we characterize which restrictions of those terms are valid lexicalizations of more specific concepts. We proceed as follows:

[3] The set of prepositions that have been selected to introduce typical restrictive descriptions are: *di,a,per,da*. Only postnominal prepositional phrases introduced by one of these prepositions have been allowed for term expressions.

[4] As terminological units longer than 5 words are very infrequent in any sublanguage, we decided to stop after the second iteration

1. *Select* the set of lemmas that by themselves are markers of relevant concepts in the corpus. Lemmas are detected according to their frequency in the observed language sample as well as to their selectivity, i.e. how they partition the set of documents. This phase produces an early TD dictionary of simple terminological elements.

2. *Extend* TD also with those (well-formed) restrictions, $cn(l)$, of any $l \in TD$ according to the mutual information they exchange with l.

Select and *Extend* depend on distributional properties of simple lemmas and complex nominals, respectively.

The distributional property needed for the *Select* step is the *term specificity*. Specific nouns are those frequently occurring in a corpus, but whose selectivity in sets of documents is very high, that is: they are very frequent in a (possibly small) set of documents and very rare in the rest. In order to capture such behavior we use two scores: the frequency t_{ij} of a term i in a document j and the *inverse document frequency* of a term (Salton,1989). Given a term i, its *inverse document frequency* is defined as follows:

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

where df_i is the number of documents of the corpus that include term i, while N is the total number of documents in the collection. The following criteria is defined to capture singleton terms: *if exists at least one document where a noun i is required as index (because it is relevant for that document and selective with respect to other documents) then such a noun denotes a relevant domain term (i.e. specific concept)*. In order to decide we rely on idf_i and t_{ij} as follows.

DEF: (*Singleton term*). A noun i is a *term* if at least one document j exists for which:

$$w_{ij} = t_{ij} \log_2 \frac{N}{df_i} \geq \tau \qquad (1)$$

w_{ij} captures exactly the notion of *specificity* required in the *Select* step of our algorithm. Potential heads of terminological entries are selected according to their selective power in the corpus. Even very rare words of the corpus can be captured by (1).

In the *Extend* step of the algorithm we need to evaluate the mutual information values of phrase structures like:

$head\ Mod_1$

$head\ Mod_1\ Mod_2$

...

$head\ Mod_1\ Mod_2 \cdots Mod_n$

Mutual Information between two words x and y is defined as (Fano,1961):

$$I(x,y) = \log \frac{prob(x,y)}{prob(x)prob(y)}$$

and it can be estimated by a maximum likelihood method as in (Dagan,1993):

$$\hat{I}(x,y) = \log_2 \frac{N \cdot freq(x,y)}{freq(x)freq(y)}$$

where: $freq(x,y)$ is the frequency of the joint event of (x,y), $freq(x)$, $freq(y)$ and N are the frequency of x, y and the corpus size respectively. In order to apply the standard definition of mutual information we need to extend it to capture the specific nature of the joint event head-modifier ($H \quad M_1$). Note that M_1 denotes post nominal adjectives or past participle but also prepositional phrase like *dello Stato* in *territorio dello Stato*. We decided to estimate the Mutual Information of such structures in a left to right fashion. The rightmost modifier (i.e. M_1 in (H M_1) structures, or M_n in (H $M_1 \ldots M_n$)) is considered as the right event y and every left incoming substructure (i.e. H or $H \ldots M_{n-1}$) is represented as a single event x. The generalized evaluation of Mutual information for $cn = ((H, M_1, M_2 \cdots M_{n-1}), M_n)$ is thus:

$$\hat{I}(cn) = \log_2 \frac{N \cdot freq(H, M_1, \cdots, M_{n-1}, M_n)}{freq(H, M_1, \cdots, M_{n-1})freq(M_n)} \qquad (2)$$

As an example a term like *debito pubblico* (*public debt*) receive a mutual information score according to the following figure:

$$\hat{I}(x,y) = \log_2 \frac{N \cdot freq(debito,publico)}{freq(debito)freq(publico)}$$

while *debito publico estero* (*foreign public debt*) produces to the following ratio:

$$\hat{I}(x,y) = \log_2 \frac{N \cdot freq(debito,publico,estero)}{freq(debito,publico)freq(estero)}$$

DEF. (*Complex Terms*) A Complex nominal $cn = ((H, M_1, M_2 \cdots M_{n-1}), M_n)$ is selected as *term* (and thus included in TD) if the following condition holds:

$$\hat{I}(cn) \geq \delta(H) \qquad (3)$$

The threshold $\delta(H)$ depends on noun H as it is evaluated according to the statistical distribution of every complex nominals headed by H [5]. The set of singleton terms is exactly the same set that a classical indexing model (Salton,1989) obtains from the document collection (i.e. the corpus). The *Extend*

[5] In the experimental tests best values for δ have been obtained as a function of mean and variance of the $\hat{I}$ distribution over the set of *cn* headed by H

phase allows to capture all the relevant specifications of the singleton terms, compile a more appropriate dictionary (for the corpus) and structure it in hierarchically organized entries.

4 Implementation Issues

The model described in the previous section has been used to implement a system for terminology derivation from a corpus. The system relies upon the POS tagging activity as it is carried out within a LA framework (e.g. the ARIOSTO system (Basili et al.,1996)) and extracts a full terminologic dictionary TD of:

1. simple terms (i.e. nouns) as seeds of a terminological structured dictionary (selected according to (1)
2. complex nominal forms of some of those seeds, generated by the grammar and filtered according to (3).

Terminology extraction is triggered after POS tagging. Morphologic analysis is rerun according to the compiled TD. This feedback allows the system to exploit complex term extraction before activating syntactic recognition, in order to prune out significant components of grammatical ambiguity. This improves the overall ability of the linguistic processor and supports term oriented rather than lemma oriented lexical acquisition.

A dedicated subsystem has been developed to support manual validation of single terms. In Figure 1 a screen dump of the graphical interface that supports the interactive validation (or removal) of terms in TD is shown. TD is hierarchically organized in separate *sections* where singleton terms dominate all their specified subconcepts. A *section* is the set of terms that share the same term head. A term like *smaltimento dei rifiuti* (*garbage collection*), has the noun "*smaltimento*" (*garbage*) as its term head. A specific section includes terms like *smaltimento dei rifiuti*, *smaltimento di materiale tossico*, *smaltimento di gas di scarico*, ...). In Figure 1 the head noun *debito* (*debt*)) is reported: the section related to *debito* includes all its validated specifications (e.g.*debito pubblico* (*public debt*), *debito pubblico estero* (*foreign public debt*) ...).

5 Experimental Set-Up

In this section we describe the experimental set-up used to evaluate and assess the described model of terminological derivation.

The method has been tested over two corpora of italian documents. The first corpus (ENEA) is a

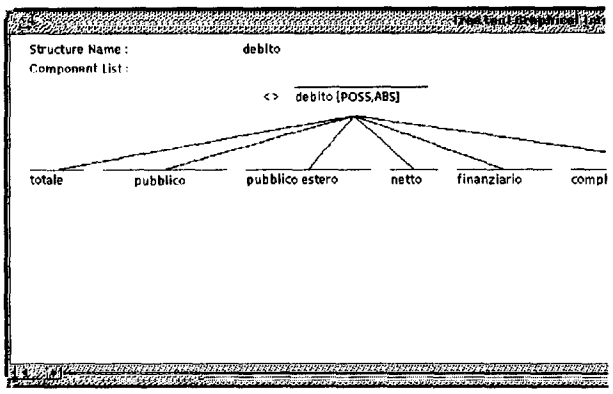

Figure 1: User Interface for Terminology Validation

Table 1: Distribution of indexes headed by *attività*

Index	1to20	21to40	41to60	61to80	81to106
Method RI: *attività*	3	2	5	1	4
Method TI: *attività*	1		5	1	3
entropica	1				1
di costruzione		1			
produttiva		1			
umana	1				

collection of scientific abstracts on the environment, made of about 350.000 words. The second corpus (Sole24Oore) is an excerpt of financial news from the *Sole 24 Ore* economic newspaper, of about 1.300.000 words. The terminology extraction have been run over both the corpora. From the ENEA corpus we derived a dictionary of about 2828 words. From the Sole24Ore corpus 5639 terms have been extracted. In order to carry out the experiments we used a subset of the ENEA corpus in order to measure performance over manually validated documents. The specific nature of our tests required the definition of particular performance evaluation measures. In fact, together with the classical notion of *recall* and *precisions*, we used also *data compression*, as the percentage of incorrect syntactic data that are no longer produced when specific terminology is used. A further index is the *average ambiguity* defined according to the notion of collision set (Basili et al., 1994). In order to accomplish the task further reference information has been used: two standard domain specific thesaura have been used for comparing the result of the terminology extraction in the environmental domain (ENEA corpus).

5.1 Linguistic analysis of corpus data

In Table 1 the section headed by *attività*, as it has been derived from the ENEA corpus, is shown. The

specific nature of the corpus is well reproduced by the data. Here two specific senses of the lemma *attività* are captured: natural and biological activity as in *attività entropica* and human activities (like *attività produttiva* (*productive activity*) or *attività di costruzione* (*building activity*). These latter have specific implications (for what concerns artificial pollution) in the environment.

Table 1 reports also the distribution of the term in a set of 106 documents. In method RI terms have been selected by classical inverse document frequency (Salton,1989) applied to singleton lemmas (i.e.*attività*). In Method TI we run inverse document frequency after a terminology driven lemmatization of documents (i.e. using complex terms as source lemmas). The two sections of the table show that no index has been lost by the TI method (all of the 15 indexes have been found). This result is more general: TI method produces more indexes. Over the 106 documents MI extracts 476 simple indexes while TI extracts 732 (terminological) indexes.

Again in Table 1, 5 of the fifteen indexes found by the TI method are complex nominals. In the set of documents from 1 to 20 (*1to20* column) these allow to discriminate between *attività* and *attività antropica*.

Such an higher discriminating power is required not only for document classification/retrieval but, first of all for lexical acquisition: in this technical domain in fact it seems necessary to rely on the information that *attività* is typically carried out by **humans** while *attività antropica* is not. We are convinced that these are the typical selectional constraints to be captured by corpus driven lexical acquisition methods. Finer lexicalizations (like *attività antropica*) are the only way to provide a better input to the target acquisition tasks.

5.2 Experiment 1: Effectiveness of the terminology extraction

The aim of this experiment was to test the ability of the method to capture relevant concepts in the sublanguage. We run this test on the environmental domain (ENEA corpus). The reference term dictionary was manually compiled by a team of three domain experts, culturally heterogeneous. We got a complete list of terms (simple nouns as well as complex nominals) to be used as a test-set (RT). The reference document set was a collection of 106 documents. The experts compiled a set of 482 terms organized in 155 sections (i.e. relevant head nouns). Each section thus includes 3.12 terms. For sake of completeness we selected two large hand-coded thesaura for the environment: the CNR dictionary (CNR,1995)(that includes 9613 terms) and the AIB dictionary (AIB,1995). Both these dictionaries as well as the automatically generated dictionary *TD* have been compared with the reference *RT*. The comparison has been carried out throughout the different aligned sections. The alignment of the section related to the head *smaltimento* is reported in Table 2 ("X" means the presence of the term in the corresponding dictionary, while "." denotes its absence):

Table 2: *Smaltimento* in different dictionaries

RT	CNR	AIB	TD
smaltimento dei fanghi	.	.	X
smaltimento dei rifiuti	X	X	X
smaltimento delle scorie	.	.	X

Table 3: Global Performance of different dictionaries

Dictionary	CNRD	AIB	TD
# of Relevant Terms	41	45	331
# of Terms	880	180	472
Recall	8,87%	9,74%	71,56%
Precision	4,66%	23,94%	70,13%

Any dictionary D can thus be evaluated by measuring *precision*, i.e.

$$precision = \frac{RTterms \cap Dterms}{Dterms}$$

and *recall*, i.e.

$$recall = \frac{RTterms \cap Dterms}{RTterms}$$

For example within the section related to the head *smaltimento*, we have 3 RT terms, of which 1 is in CNR and AIB respectively and 3 are in TD. When applying the recall and precision definition to every sections of the RT dictionary we obtained the average performance scores reported in Table 3 over the three dictionaries.

5.3 Experiment 2: Shallow parsing with terminological knowledge

Consulting a terminologic dictionary before activating a shallow syntactic analyzer is helpful to solve several morphological and syntactic ambiguities. For example, given the sentence [6]

L'ufficiale della Guardia di Finanza visitó l'aereoporto di Fiumicino
(The officer of Finance Guard visited the Fiumicino airport)

a typical shallow syntactic analyzer (SSA) (Basili et al., 1992) produces the following elementary syntactic links (*esl*), due the syntactic ambiguity of prepositional phrases (PP), e.g. ((*di finanza*), (*di Fiumicino*)):

N_P_N ufficiale della guardia

N_P_N ufficiale di finanza

N_P_N guardia di finanza

[6] This sentence has been extracted from the Sole24Ore corpus

N_V ufficiale visitó

V_N visitó aeroporto

N_P_N aereoporto di fiumicino

V_P_N visitó di fiumicino

As each sentence reading cannot assign more than a single referent to each PP, we can partition the set of *esl* into several *collision sets* (i.e. sets of *esl* that cannot belong to the same sentence reading according to (Basili et al, 1994)). The sample sentence gives rise to the following collision sets:

{ (ufficiale di finanza) (guardia di finanza) }

{ (ufficiale visitó) }

{(aereoporto di fiumicino) (visitó di fiumicino) }

{ (ufficiale della guardia) }

{ (visitó aeroporto) }

When terminology is available many complex nominals are retained as single tokens and several ambiguity disappear. In the Sole24Ore corpus our method produced both the terms *guardia di finanza* and *aeroporto di Fiumicino* so that the final list of *esl* reduces to

N_P_N ufficiale della guardia_di_finanza

N_V ufficiale visitó

N_V guardia_di_finanza visitó

V_N visitó aeroporto_di_fiumicino

and no ambiguous (i.e. not singleton) collision set remains. We have two positive effects on the parsing activity. The first is *data compression*. In fact the overgeneration typically due to the shallow grammatical approach is significantly limited. In our example the early 7 elementary syntactic groups obtained in absence of terminology reduced to 4 with an overall data compression of ((7-4)/7) 42.8%. An extended experimentation has been carried out on a subset of 500 sentences of the corpus. The use of terminology reduces the number of elementary syntactic links from 500 to 403 with a corresponding 20% of overall data compression. Furthermore, the detection of a term carried out over single tokens that are morphologically ambiguous improves also the morphological recognition. In fact the detection of a chain of tokens that are part of the same term implies a specific choice on the grammatical category of each token, thus augmenting the selectivity of POS tagging. Over the same subset of the corpus we measured a decrement of 4% in the number of morphological derivations produced with terminology against the recognition carried out in absence of any terminological knowledge.

A second positive aspect of having an available domain specific terminology is the *reduction of the underlying syntactic ambiguity* and increase of the parser precision. As shown in the example many PP ambiguity disappears as soon as a set complex nominals is detected. This has a strong implication on shallow (or robust as widely accepted in literature) parsing. We conducted a systematic analysis of correct parsing results by contrasting a parser with and without access to domain terminology. The analysis of the results has been performed by comparing collision sets obtained by the two runs over a set of 100 sentences. Four performance scores have been evaluated: the *degree of ambiguity* (i.e.the ratio between the number of ambiguous esl's over the total number of derived *esl's*); the *average ambiguity* (expressed by the average cardinality of the collision sets (i.e. the number of reciprocally ambiguous *esl's*); finally, *precision* and *recall* have been measured according to a hand validation of the derived syntactic material [7]. The analysis has been carried out specifically for prepositional *esl's* (i.e. noun-preposition-noun, verb-preposition-noun, adjective-preposition-noun links). Results are reported in Table 4 where separate columns express the scores for the different runs: a simple parser (SP), and a terminology driven parser (TP). As a result the simple parser obtains several complex nominals but only as syntactic structures so that it fails in detecting higher order syntactic links (i.e. syntactic relations between complex nominals and other sentence segments). In these cases we penalized also the recall of the SP method, so that the difference between the two methods relies not only in amount of persisting ambiguity (i.e. precision), but also in coverage (better captured by recall).

Table 4: Performance evaluation of terminology driven parsing

Parser	Ambiguity	#Collisions	Recall	Precision
SP	0.60	3.2	0.65	0.67
TP	0.55	2.9	0.68	0.71

6 Conclusions

In this paper a method for the automatic extraction of terminological (possibly complex) units of information from corpora is presented. The proposed method combines principle of grammatical correctness with statistical constraints on the distributional

[7] *Precision* is the number of detected *correct esl's* over the total number of detected *esl's*, while *recall* is the number of detected *correct esl's* over the number of *correct esl's*

properties of the detected domain terms. In an incremental fashion NPs are first selected as possible candidates for term denotation and then inserted in an incremental terminological dictionary according to their mutual information value. The experimental test has been difficult as a precise notion of what is a relevant term in a domain is very vague and subjective. Tests against a domain specific user oriented dictionary have been carried out, in comparison with large scale thesaura in the domain. The significant improvement against this standard sources is very successful. The method has been widely applied to different corpora and it demonstrated to be easily portable without any heavy customization. As it relies upon simple POS tagging, it is widely portable to other languages, as soon as NP grammars are available. Feedback of the terminological extraction process to the morphologic analysis has been also designed. A measure of the improvement that terminological NP recognition implies over the activity of a shallow parser for LA has been carried out. The result is an overall improvement: data compression is around 5% while syntactic ambiguity elimination is about 10%. Recall and Precision of the syntactic analysis is consequently higher.

The main result of this method is to support finer lexicalization, in form of complex nominals, for lexical acquisition. Lexical acquisition based on collocations between terms (and not simple lemmas) provides more granular information on lexical senses as well as (syntactic or semantic) selectional constraints. The success of this method allow to design automatic methods for taxonomic (thesaurus-like) knowledge generation. Distributional, as well syntactic, knowledge is a crucial source of information for large scale similarity estimation among detected terms.

References

AIB, 1995, Ensoli A., Marconi G., Sistema di Classificazione dei Documenti di Interesse Ambientale, Rapporti AIB-7, ISSN 1121-1482

Basili 1992, R.Basili, M.T.Pazienza, P.Velardi A Shallow Syntactic Analyzer to extract word association from corpora, Literary and Linguistic Computing, 1992, vol.7, n.2, 114-124

Basili, R., A. Marziali, M.T. Pazienza, Modelling syntactic uncertainty in lexical acquisition from texts, Journal of Quantitative Linguistics, vol.1, n.1, 1994.

Basili et al 1996a. Basili R., M.T. Pazienza, P. Velardi. An Empirical Symbolic Approach to Natural Language Processing. Artificial Intelligence, Vol.85, August 1996.

Basili et al.,1996b. Basili, R., M.T. Pazienza, P.Velardi, Integrating General Purpose and Corpus-based Verb Classifications, Computational Linguistics, 1996.

Bourigault, D., 1992, Surface Grammatical Analysis for the Extraction of Terminological Noun Phrases, Proc. of COLING 1992, Nantes, France, pp.977-981.

Brill E., Resnik P.,1994, *A rule-based approach to prepositional phrase attachment disambiguation*, in Proc. of COLING 94, 1198-1204

Church, K., 1988, A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text, Proc. of 2nd Conf. on Applied Natural Language Processing, Austin, pp. 136-143

CNR, 1995. Thesaurus Italiano Generale per l'Ambiente, Consiglio Nazionale delle Ricerche (CNR), Rapporto Scientifico 10/95, Ed. Bruno Felluga, Edizione 31/07/1995, Roma.

Dagan 1993, I.Dagan, K.Church Identifying and Traslating Technical Terminology, IJCAI 1993

Daille et al, 1994, Daille B., Gaussier E., Lange', J.M., Towards Automatic Extraction of Monolingual and Bilingual Terminology, COLING-94, August, Kyoto, Japan, 1994.

De Rossi, 1996, Elaborazioni Satistiche di Corpora Testuali mirate all'Acquisizione di Conoscenza per la Costruzione di Thesaura, Faculty of Engineering, University of Roma, Tor Vergata, 1996.

R. Fano, Transmission of Information, Cambridge, Mass., MIT Press, 1961

Hindle D. and Rooth M., 1993, *Structural Ambiguity and Lexical Relations*, Computational Linguistics, 19(1): 103-120.

Salton G., Automatic Text Processing: the Transformation, Analysis and Retrieval of Information by Computer, Addison-Wesley Publ., 1989.

7 Appendix 1: Excerpt of Terminological Dictionaries from two domains

Noun	ENEA terms	Sole 24 Terms
Fonte (Source)	[fonte,di,inquinamento] [fonte,principale] [fonte,dei,rifiuti] [fonte,di,energia] [fonte,di,inquinamento] [fonte,di,materia,prima] [fonte,energetica] [fonte,eolica] [fonte,idrica] [fonte,informativa] [fonte,nucleare] [fonte,primaria,di,energia] [fonte,principale,dei,rifiuti]	[fonte,energetica,primaria] [fonte,normativa] [fonte,normativa,citata] [fonte,principale]
Rischio (Risk)	[rischio,ambientale] [rischio,cancerogeno] [rischio,chimico] [rischio,climatico] [rischio,di,area] [rischio,di,crisi] [rischio,di,inquinamento] [rischio,erosivo] [rischio,industriale] [rischio,reale,connesso] [rischio,relativo] [rischio,sanitario] [rischio,sismico] [rischio,tecnologico] [rischio,tossicologico]	[rischio,aziendale] [rischio,standard]