A Contrastive Approach to Multi–word Term Extraction from Domain Corpora

Francesca Bonin[*•], Felice Dell'Orletta[◇], Giulia Venturi[◇] and Simonetta Montemagni[◇]

[◇] Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC–CNR) – Pisa (Italy)
[*]Dipartimento di Informatica Pisa, Università di Pisa,
[•]Language Interaction and Computation Lab, University of Trento,
{francesca.bonin, felice.dellorletta, giulia.venturi, simonetta.montemagni}@ilc.cnr.it

Abstract

In this paper we present a novel approach to multi–word terminology extraction combining a well–known automatic term recognition approach, the C–NC value method, with a contrastive ranking technique, aimed at refining obtained results either by filtering noise due to common words or by discerning between semantically different types of terms within heterogeneous terminologies. The proposed methodology has been tested in two case studies carried out in the History of Art and Legal domains with promising results.

1. Introduction

Terminology extraction is a central field of research for a number of Knowledge Management applications, such as Ontology Learning, Text Mining, Information Retrieval, etc. Starting from the assumption that terms unambiguously refer to domain-specific concepts, a number of different methodologies has been proposed so far to automatically extract domain terminology from texts. Generally speaking, the term extraction process consists of two fundamental steps: 1) identifying term candidates (either single or multi–word terms) from text, and 2) filtering through the candidates to separate terms from non-terms. To perform these two steps, term extraction systems make use of various degrees of linguistic filtering and, then, of statistical measures ranging from raw frequency to Information Retrieval measures such as Term Frequency/Inverse Document Frequency (TF/IDF) (Salton et al., 1988), up to more sophisticated methods such as the C-NC Value method (Frantzi et al., 1999), or lexical association measures like log likelihood (Dunning, 1993) or mutual information. Others make use of extensive semantic resources (Maynard et al., 1999), but as underlined in Basili et al. (2001b), such methods face the hurdle of portability to other domains.

Another interesting line of research is based on the comparison of the distribution of terms across corpora of different domains. Under this approach, identification of relevant term candidates is carried out through inter–domain contrastive analysis ((Penas et al., 2001; Chung et al., 2004; Basili et al., 2001a)). Interestingly enough, this contrastive approach has so far been applied only to the extraction of single terms, while, multi–word terms' selection is based upon contrastive weights associated to the term syntactic head. This choice is justified by the assumption that multi-word terms typically show low frequencies making contrastive estimation difficult (Basili et al., 2001a). On the contrary, we aim at focusing our attention on the extraction of multi–word terms, which have been demonstrated to cover the vast majority of domain terminology (85% according to Nakagawa et al. (2003)); for this reason, we believe that they have to be considered independently from the head. Aware of the problem of data sparseness of multi-word terms, we propose a two–stage approach where we firstly extract a shortlist of well-formed and relevant candidate multi-word terms, and secondarily we apply a contrastive method against the selected terms only. The proposed methodology has been tested on Italian text collections belonging to two different domains, presenting different degrees of complexity: the Art History domain and the Legal domain. The latter appears to be quite challenging because of the acknowledged difficulties in discerning law terms from terminology of the regulated domain (Lame, 2005; Lenci et al., 2009).

2. General extraction method

The multi–word term extraction methodology we propose here is based on a combination of "termhood" measures, assessing the likelihood of being a valid technical term, and contrastive methods. In particular, multi–word term extraction is carried out by identifying candidate multi–word terms in an automatically POS–tagged and lemmatized text, which are then weighted with the C-NC value, currently considered as the state–of–the–art method for terminology extraction. The ranking of identified multi–word terms is then revised on the basis of a contrastive score calculated for the same terms with respect to corpora testifying general language usage. The main novelty of the proposed approach lies in the fact that, differently from previous studies, here the contrastive analysis is applied to previously identified multi–word terminology, with the aim of further filtering it. Starting from the assumption that domain relevant multi–words are unique elements, separate from single terms, we rather prefer basing multi–word extraction on their concrete frequency of occurrence in corpora. Such an approach becomes particularly useful when the domain text collection also includes particularly frequent common words which make the final result noisy or, more crucially, when the resulting terminology is highly heterogeneous as in the case of legal texts. In the following sections we describe, in 2.1., the multi-word candidates extraction process, and in 2.2. the subsequent contrastive ranking process.

2.1. Multi–word term extraction

In this section, we discuss the candidate extraction process, that makes use of: i) linguistic filters; ii) stoplist; iii) statis-

tical filters (C-NC Value).

2.1.1. Linguistic filters

The linguistic filters operate on the automatic POS–tagged and lemmatized text, making use of different kinds of linguistic feature. The POS–tagged text, obtained with the tagger described in Dell'Orletta (2009), is searched for on the basis of a set of POS patterns encoding morpho–syntactic templates of candidate complex terms covering the main nominal modification types. Specifically, for each multi–word term to be identified in texts, we used POS–restrictions constraining the start-token and final–token POSs, but also the internal–tokens POSs.

Since we were interested in nominal "chunks", which consist of nouns, adjectives and prepositions (Justeson et al., 1995), we use linguistic filters that accept only those kind of part of speech. Specifically, we identify sequences of allowed POS patterns in order to cover most of the Italian morphosyntactic multi–words structures, using the following pattern:

Noun+(Prep+(Noun|ADJ)+ |Noun|ADJ)+

The choice of linguistic filters affects the precision and the recall of the output list, e.g a restrictive filter will have a positive effect on precision and a negative effect on recall (Basili et al., 2001a). In our method, we use a filter which also constrains the maximum number of words of which a complex term can be made. In fact, we operate on the candidate terms' length (*l*) as one of the main linguistic constraints to be ruled. We believe that such a measure is to be considered as domain–dependent, being related to the linguistic peculiarities of the specialised language we are dealing with.

2.1.2. Stoplist

At this stage, linguistically filtered candidate multi–word terms are screened by using a multi–word preposition stoplist; in order to extract *domain-specific multi-word prepositions*, this list is obtained with a first run of the same term-extraction procedure operating on the corpus from which we are going to extract multi-word terms. To extract prepositional candidates, this method uses, specifically, the linguistic pattern *(Noun+Prep+Noun)*. Resulting multi–word prepositions won't be considered as a start or end element of a term: in this way, non-sense terms such as *sensi della legge* lit. 'senses of the law are avoided due to the overlapping with the multi–word preposition *ai sensi di*, 'by law'.

With these types of constraints (linguistic filtering, terms' length and stoplist filtering) the typology of multi–word term candidates is anyway quite varied, ranging from terms such as *ricerca artistica* 'artistic research', *Ministro dei Beni culturali* 'Minister of Cultural Heritage' to *piano di gestione di bacino* 'management plan of the basin'.

2.1.3. Statistical filters based on C-NC Value

As a statistical filter, we use the C-NC Value measure as described in Frantzi et al. (1999) and Vintar (2004). The C-Value method aims at bringing out those terms which tend to occur as nested terms, then, the NC-Value incorporates context information to the C-Value, aiming at improving term extraction in general.

C Value. The C-Value calculates the frequency of a term and its subterms. If a candidate term is found as nested, the C-Value is calculated from the total frequency of the term itself, its length and its frequency as a nested term; while, if it is not found as nested, the C-Value, is calculated from its length and its total frequency. Given the candidate term t , and being $|t|$ its length, the C-Value of t is given as:

$$C{-}value(t) = \begin{cases} \log_2 |t| \cdot f(t) & \text{if } t \text{ is not nested,} \\ \log_2 |t| \cdot (f(t) - \frac{1}{P(T_t)} * \sum_{b \in T_t} f(b)) & \text{otherwise.} \end{cases}$$

where $f(t)$ is the frequency of t in the corpus, T_t is the set of terms that contain t, $P(T_t)$ is the number of candidate terms in T_t, and $\sum_{b \in T_t} f(b)$ is the sum of frequencies of all terms in T_t.

NC Value. The NC-Value measure (Frantzi et al., 1999) aims at combining the C-Value score with the context[1] information. A word is considered a *context word* if it appears with the extracted candidate terms. The algorithm extracts the context words of the top list of candidates (context list)[2], and then calculates the N-Value on the entire list of candidate terms. The higher the number of candidate terms with which a word appears, the higher the likelihood that the word is a *context word* and that it will occur with other candidates. If a *context word* does not appear in the extracted context list, its weight for such term is zero. Formally, given w as a context word, its weight will be: $weight(b) = \frac{t(w)}{n}$ where $t(w)$ is the number of candidate terms w appears with, and n is the total number of considered candidate terms; hence, the N-Value of the term t will be $\sum_{w \in C_t} f_t(w) * weight(w)$, where $f_t(w)$ is the frequency of w as a context word of t, and C_t is the set of distinct context words of the term t. Finally, the general score, NC-Value, will be:

$$NCValue = \alpha * CValue(t) + \beta * NValue(t) \quad (1)$$

where, in our model, α and β are set empirically ($\alpha = 0.5$ and $\beta = 0.5$).

2.2. Multi-word terms contrastive ranking

The list of multi-word terms extracted at the processing stage described in section 2.1. is then ranked by resorting to a contrastive method.

Differently from the other contrastive methods ((Basili et al., 2001a; Penas et al., 2001; Chung et al., 2004; Kozakov et al., 2004)) that are applied only to single terms for avoiding the multi-word terms' sparsity problem, we apply the contrastive function directly to complex terms. However, being aware of such a problem, we overcome the sparsity issue by splitting the process into two different steps. First

[1] Our implementation of the C-Value score uses a context length of 3 tokens to the left and 3 tokens to the right.

[2] In this work we set the top list threshold at 4.0 C-Value score.

we select well-formed, relevant multi-words, having significant distributional tendencies; afterwords we apply the contrastive function only to these pre-selected multi-word terms. With this procedure we focus, firstly, on the retrieval of valid technical terms, and secondarily on domain pertinence, in two distinct but consequent moments.
In what follows we start describing the approach used by Basili et al. (2001a) (section 2.2.1.). Then, we describe a new approach where Basili's approach is applied directly on multi-word terms showing that, multi-word terms can be treated as autonomous entities (section 2.2.2.). In this method, we differ from other TF-IDF-like approaches since we obtain a modular structure which allows us to use different functions both for multi-word extraction, and for contrastive ranking, according to the specific tasks. In this case, we focus on the double domain terminology problem, as described in 3.2., and, for this purpose, we propose a new contrastive function aiming at distinguishing double domain terminology (described in 2.2.3.).

2.2.1. Contrastive Selection via Heads

Basili et al. (2001a) proposed a Contrastive method, henceforth referred to as *Contrastive Selection via Heads* (CSvH), where the selection of multi–word terms in the target domain is done according to contrastive information related to their head. The CSvH method can be divided in two steps:

- single candidate terms are selected using a contrastive function based on their distribution in the target and contrastive corpora;
- the single weighted terms are the heads of multi–word terms and the multi–word term scores are calculated by multiplying the head contrastive value with the frequency of the multi–word term in the target domain.

The contrastive function used by Basili et al. (2001a) is a TF-IDF inspired measure. However instead of *Inverse Document Frequency*, they used *Inverse Word Frequency* (IWF):

$$IWF(st) = \log(\frac{N}{F(st)}), \tag{2}$$

where, st is a candidate single term, N is the size of the contrastive corpus and $F(st)$ is the frequency of st in all domain corpora. In the same way as the TF-IDF measure, the TF-IWF measure takes into account the frequency of the candidate term st in the target domain to avoid the penalization of high frequency terms. Therefore, the contrastive function is:

$$w_i(st) = \log(f_i(st)) * IWF(st) \tag{3}$$

where $f_i(st)$ is the frequency of st in the target domain i.
In the second step of the CSvH method, multi–word terms are weighted by multiplying the contrastive value of their head with the frequency of the term in the target domain. Hence, the contrastive weight (Cw) of the multi–word term t in the domain i is defined as:

$$Cw_i(t) = f_i(t) * w_i(h(t)) \tag{4}$$

where $f_i(t)$ is the frequency of the term t in the target domain i and $w_i(h(t))$ is the contrastive weight of the term's head.

2.2.2. Term Frequency Inverse Term Frequency

The *Term Frequency Inverse Term Frequency* (TFITF) method is a variant of Basili et al. (2001a). Differently from CSvH, the contrastive function is applied directly on a list of previously selected candidate multi–word terms. In our work we use the multi–word extraction process described in Section 2.1. for obtaining the list of candidate multi–word terms.
Given the set of multi–word terms T extracted from the target domain i, the TFITF value of term $t \in T$ is:

$$w_i(t) = \log(f_i(t)) * IWF(t) \tag{5}$$

where $f_i(t)$ is the frequency of t in the target domain i and $IWF(t)$ is defined:

$$IWF(t) = \log(\frac{N}{F(t)}). \tag{6}$$

$F(t)$ is the frequency of t in all domain corpora and N is:

$$N = \sum_{t \in T}(F(t)). \tag{7}$$

2.2.3. Contrastive Selection of Multi–word terms

Starting from the assumption that multi–word terms are less frequent than single terms, we introduce a new Contrastive method, called *Contrastive Selection of multi–word terms* (CSmw), particularly suitable for handling variation in low frequency events. As in the TFITF method, the CSmw statistical weight is assigned directly to multi–word terms.
The CSmw function is based on an arctangent function of this form:

$$w(x) = \arctan(K * x) \tag{8}$$

where K is a coefficient.
This function presents two interesting features:

- the presence of an asymptote in the point $(0, \pi/2)$,
- the higher the coefficient K the faster the knee of the function gets closer to the asymptote.

Therefore, given the set of multi–word terms T extracted from the target domain i and a set of contrastive domains C, we defined the coefficient K as:

$$K(t) = \frac{1}{\frac{F_c(t)}{N_c}}. \tag{9}$$

Where $t \in T$, $K(t)$ is the coefficient of t, $F_c(t)$ is the sum of the frequencies of t in the contrastive corpora and N_c is the sum of the frequencies of all elements of T in the contrastive corpora. More formally

$$F_c(t) = \sum_{j \in C}^{j \neq i}(f_j(t)), \tag{10}$$

and

$$N_c = \sum_{t \in T}(F_c(t)). \tag{11}$$

$K(t)$ has the property that when $F_c(t)$ increases, $K(t)$ decreases and vice–versa.
Hence the statistical function is:

$$w(t) = \arctan\Big(\frac{f_i(t)}{\frac{F_c(t)}{N_c}}\Big) \quad (12)$$

where $f_i(t)$ is the frequency of t in the domain corpus.
This function guarantees three fundamental properties for tackling our tasks, given two terms $t1$ and $t2$:

- when $(F_c(t1) == F_c(t2))$ and $(f_i(t1) > f_i(t2))$, $CSmw(t1) > CSmw(t2)$;
- when $(F_c(t1) < F_c(t2))$ and $(f_i(t1) == f_i(t2))$, $CSmw(t1) > CSmw(t2)$;
- when $(F_c(t1) < F_c(t2))$ and $(F_c(t1) - f_i(t1) == F_c(t2) - f_i(t2))$, $CSmw(t1) > CSmw(t2)$.

Finally, we moderated the positive effect of the low frequency of t in the contrastive corpora ($F_c(t)$) by multiplying the argument of the arctangent for the logarithm of the frequency of t in the domain corpora ($\log(f_i(t))$). So the CSmw function is:

$$CSmw(t) = \arctan\Big(\log(f_i(t)) * (\frac{f_i(t)}{\frac{F_c(t)}{N_c}})\Big) \quad (13)$$

Figure 1 illustrates the CSmw function. Given a target domain D and three terms extracted from D ($d1, d2, d3$) with three different frequencies in a contrastive domain C ($F_c(d1) = c1 = 10000, F_c(d2) = c2 = 1000, c3 = 100$), Figure 1 shows the CSmw contrastive function as the number of occurrence of $d1, d2, d3$ in the target domain changes.

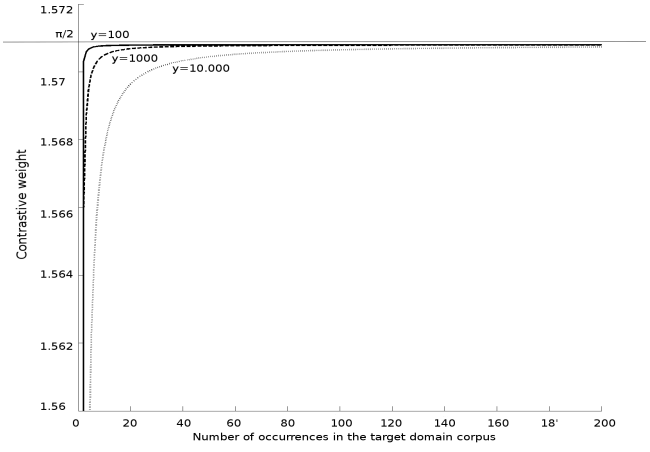

Figure 1: Contrastive Selection Multi–words function. N_c=100000.

3. Case studies

The term extraction methodology described above has been tested in two case studies carried out in the History of Art and Legal domains. The Art History corpus has been collected by a domain expert and includes texts representative of different artistic periods, for a total of 326,066 tokens. The legal corpus is constituted by a collection of European legal texts of 394,088 word tokens concerning the environmental domain; this corpus will be hereafter referred to as "Environmental Corpus". As a general contrastive corpus we used the PAROLE Corpus (Marinelli et al., 2003), made up of about 3 million words and including Italian texts of different types (newspapers, books, etc.).

3.1. Extraction of domain specific terminology from an Art History corpus

In this case study we used the Art History corpus as the target domain corpus and the PAROLE Corpus as the contrastive corpus. We selected a top list from the candidate term list ranked on C-NC Value score (2.1.3.), which was obtained by setting an empirically defined threshold: i.e. the first 600 terms of the ranked list were selected. Such a selected list turned out to include domain–specific terms (e.g. *pittura italiana*, 'Italian painting') but also open–domain ones (e.g. *ente locale*, 'local authority'). The final term list is represented by the top list of 300 terms ranked according to the contrastive score: such a list includes domain–specific terms only, without noisy common words. It should be noted that the two thresholds for top lists' cutting as well as the maximum term length can be customized for domain–specific purposes through the configuration file[3]. As it was discussed in Section 2.1.1., the length of multi–word terms is dramatically influenced by the linguistic peculiarities of the domain document collection. We empirically tested that for the Art History domain multi–word terms longer than 4 tokens introduce noise in the acquired term list.
Table 1 contains a fragment of the acquired list of 300 multi–word terms we obtained following the contrastive approach described in 2.2..

Artistic Multi-words
movimento artistico (*artistic movement*)
figura umano (*human figure*)
arte contemporaneo (*contemporary art*)
produzione artistico (*artistic production*)
pittore italiano (*Italian painter*)
mostra online (*online exhibition*)
percorso espositivo (*exhibition path*)
collezione privato (*private collection*)
arte italiano (*Italian art*)
bene culturale (*cultural heritage*)

Table 1: First 10 multi–word terms extracted from the Art History Corpus

3.2. Extraction of domain specific terminology from a legislative corpus

The second case study has been carried out on the legal domain which poses the further challenge of the highly het-

[3]In our experiments the threshold has been set empirically after several experiments at 600 terms and the maximum term length at 4 tokens.

erogeneous nature of the extracted terminology, typically including legal terms as well as terms of the domain being regulated. So far, term extraction applied to legal domain corpora results in a hybrid term glossary, including terminology of mixed nature. We believe that the proposed approach can be of some help in discriminating legal terms from regulated–domain terms, a crucial topic that - to our knowledge - has never been tackled in the terminology extraction literature. This has been achieved by iterating the contrastive process more than once against two contrastive corpora of different nature. As it is illustrated in Figure 2, the Environmental corpus we exploited in this second case study has been contrasted, first, against the open–domain PAROLE Corpus and, then, against a legal corpus belonging to a domain other than the environmental one. This latter corpus, of 74,210 word tokens, containing European law texts on consumer protection, will be hereafter generically referred to as "Legal Corpus".

Similarly to the Art History case study, from the C-NC Value ranked terms' list, we selected a top list[4], thus obtaining a shortlist of 600 either legal (e.g. *norma europea*, 'European norm'), environmental (e.g. *emissione di gas a effetto serra*, 'emission of greenhouse gases') or open–domain terms (e.g. *direttore generale*, 'director–general').

Afterwards, we firstly contrasted a top list of 600 multi–word terms against the PAROLE Corpus, in order to reduce the noise deriving from highly frequent common words. Then, we contrasted a top list of 300 environmental–legal multi–word terms against the Legal Corpus, obtaining a final list of 300 terms ranked on the contrastive score (as described in 2.2.3.). Also in this case, it should be noted that all thresholds for top lists' cutting have been empirically defined after several experimental tests. This double contrast was aimed at discerning, in the input list, terms belonging to the two different target domains, namely environmental and legal terms: whereas the former were expected to be found at the top of the final list ranked according to the contrastive score, the latter were expected at the bottom. In this case we empirically tested that in corpora of environmental-legal texts, relevant domain–specific information is carried by multi–word terms longer than those occurring in the Art History texts; for this reason the maximum multi-words' length has been set at 6 tokens. It is the case of both *legal* terms, such as e.g. *testo della disposizione essenziale del diritto*, 'text of the law essential provision', and terms which belong to the regulated domain, such as e.g. *inquinamento atmosferico trasfrontaliero a grande distanza*, 'long distance atmosferic transfrontier pollution'.

Tables 2 and 3 report two fragments of the 300 multi–word term list we obtained by iterating the contrastive process. In particular, Table 2 contains the first 10 terms of the final list while Table 3 shows the last 10 terms.

Interestingly enough, our initial hypothesis seems to be proved: the top of the final list as reported in Table 2 contains environmental terms, while the legal terms can be found at the bottom (see Table 3). These results will be discussed more in detail in Section 4..

[4]The threshold has been empirically set at 4.0 C-NC Value score.

Environmental terms
sostanza pericoloso (*hazarous substance*)
salute umano (*human health*)
sviluppo sostenibile (*sustainable developement*)
principio attivo (*active ingredient*)
inquinamento atmosferico (*air pollution*)
valore limite di emissione (*emission limit value*)
effetto serra (*greenhouse effect*)
rifiuto pericoloso (*hazardous waste*)
corpo idrico (*water body*)
cambiamento climatico (*climate change*)

Table 2: First 10 multi–word terms extracted from the Environmental Corpus

Legal terms
funzionamento di mercato interno
(*functioning of national market*)
disposizione essenziale di diritto interno
(*essential internal provision of national law*)
diritto nazionale (*national law*)
disposizione nazionale (*national provision*)
diritto interno (*national law*)
norma nazionale (*national rule*)
disposizione legislativo (*legislative measure*)
responsabile di formulazione (*formulator*)
legislazione comunitario (*community legislation*)
disposizione comunitaria (*community provision*)

Table 3: Last 10 multi–word terms extracted from the Environmental Corpus

4. Evaluation

The evaluation of the acquired multi–word term lists was carried out by adopting similar evaluation criteria for the two case studies even though partially different extraction methodologies have been exploited.

4.1. General evaluation criteria

The multi–word term lists extracted in the case studies described in 3.1. and in 3.2. have been evaluated both against gold-standard resources and through manual validation by domain experts. These two different evaluation types were specifically aimed at dealing with two general issues of multi-word terms evaluation: *i)* the considered reference resources have a good coverage of domain specific single terms, but they do not have a proper coverage of domain-specific complex terms (e.g. *scena di genere*, 'genre works'); *ii)* many terms cannot be easily unambiguously categorized as belonging to a specific domain. As it will be discussed in Section 4.3., *ii)* is often the case of those terms that occur in legal documents but refer to objects or concepts of the real world, regulated by the law; e.g. terms such as *rifiuto pericoloso* 'dangerous waste' or *inquinamento atmosferico* 'atmosferic pollution' label environmental concepts which typically occur in environmental–specific laws. Consequently, they are included in both environmental and legal terminological re-

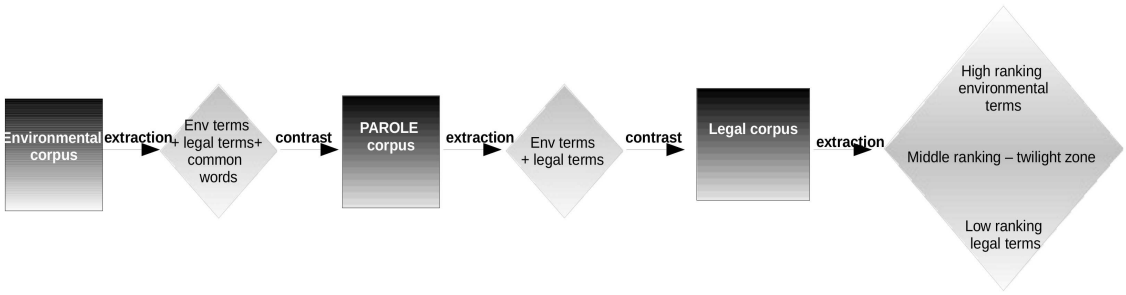

Figure 2: Double Contrast Process

sources and they can be disambiguated only through manual validation by a domain expert.

4.2. Evaluation results of the Art History case study

The first phase in the evaluation of the 300 multi–word term lists, extracted from the Art History Corpus, was carried out by automatically comparing the acquired list against a Art glossary[5]. Afterwards, the results of this first evaluation phase have been manually validated by a domain expert. Eventually, we obtained four lists of 300 validated terms, further divided in 30-term groups which show domain-specific terms' distribution. Contrastive-based methods have in general better performances in extracting domain-specific multi-words. Table 4, in fact, reports the amount of domain specific terms for each group. Even though, the four extraction methods have similar results in the first group, the CSmw method has the best Sub-TOT, with 124 artistic terms out of 150 candidate terms. TFITF approach extracts 119 artistic terms out of 150, having a slightly better performance than the CSvH. This result witnesses that a better extraction of multi-word terms can be carried out by applying TFITF measure, directly on complex terms (see Section 2.2.), instead of on single terms. The CSvH method gets the higher total number of artistic multi-word terms, but these terms are uniformly spread on the entire range. On the contrary, the CSmw function shows considerable better results in the top list, being able to discriminate artistic terms on top. As well, the TFITF function is able to group artistic specific domain terms at the top of the list, maintaining anyway good scores on the entire list. Figure 3 shows the trends of the four functions in retrieving artistic terms. Finally, it is interesting to notice that, the CSmw method, acting directly on multi-words, turns out to extract those terms which are not only domain-specific terms, but also domain-specific terms for the analyzed text; on the other hand, although CSvH extracts good domain specific terms, these terms are not necessarily relevant in the considered text. It is the case of *arte concettuale* ('conceptual art') which is an artistic term with high rank in CSvH, but with very low frequency in the analyzed text.

[5]The glossary has been provided by the Art History Department of the University of Pisa, to which has been added an online resources for a total amount os 1048 terms (http://www.babelearte.it/glossario.asp?ini=a, http://www.marcolla.it/glossario/a/a.htm).

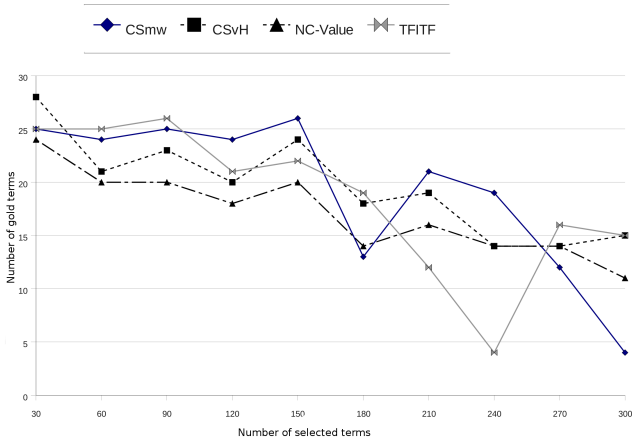

Figure 3: Double contrastive process

Group	NC-Value	CSmw	CSvH	TFITF
0-30	24	25	28	25
30-60	20	24	21	25
60-90	20	25	23	26
90-120	18	24	20	21
120-150	20	26	24	22
Sub-TOT	102	124	116	119
150-180	14	13	18	19
180-210	16	21	19	12
210-240	14	19	14	4
240-270	14	12	14	16
270-300	11	4	15	15
TOT	171	193	196	185

Table 4: Evaluation of Art Domain

4.3. Evaluation results of the legislative case study

Due to the heterogeneous nature of the acquisition corpus, the evaluation of the 300 multi–word term list extracted from the Environmental Corpus was carried out against two different gold standard resources. Namely, the thesaurus *EARTh (Environmental Applications Reference Thesaurus)*[6], containing 12,398 terms, was used as a reference resource for what concerns the environmental domain, and the *Dizionario giuridico* (Edizioni Simone) available on-

[6]http://uta.iia.cnr.it/earth.htm#EARTh%202002

line[7], including 1,800 terms, was used for the legal domain. According to the general evaluation criteria, we compared the four multi–word term lists, extracted following the NC–Value, the CSvH, the TFITF and the CSmw approaches, against the two aforementioned gold standard resources. Afterwards, the term lists have been manually validated by a legal and an environmental expert.

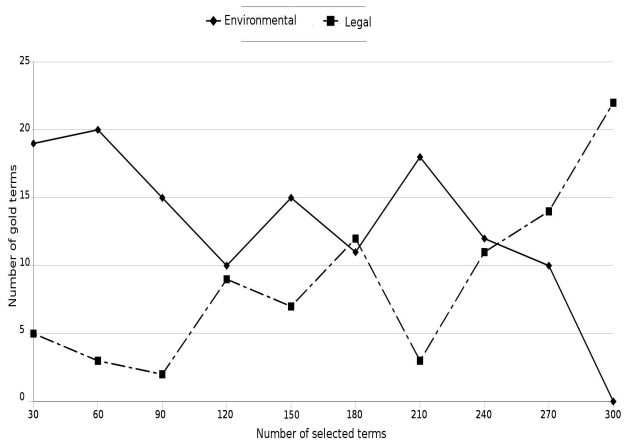

Figure 4: CSmw: double contrast trend

Group	NC-Value	CSmw	TFITF
	Env/Leg	Env/Leg	Env/Leg
0-30	11/10	19/5	17/5
30-60	11/10	20/3	13/7
60-90	14/5	15/2	13/8
90-120	12/5	10/9	9/7
120-150	6/12	15/7	13/7
150-180	12/5	11/12	7/9
180-210	16/10	18/3	14/9
210-240	9/11	12/11	12/9
240-270	14/4	10/14	9/12
270-300	11/8	0/22	9/15
TOT	116/80	130/88	116/88

Table 5: Evaluation of legislative case study

Table 5 reports the amount of environmental (referred to as Env) and legal (referred to as Leg) terms for each 30–term groups we computed.
As we can see, the CSmw method is able to distinguish clearly environmental terms from legal terms. In the first group we see 19 environmental terms against 5 legal terms; in the last: 22 legal terms, and no environmental terms. This trend is pointed out in Fig. 4, where the divergent lines show the different distributions of environmental and legal terms. The central zone of the chart, with lines crossing each other, shows a twilight zone of terms which contains both environmental and legal terms and terms that can refer to both domains (such as *politica ambientale*, 'environmental policy'). Fig. 5 sketches the absolute value of the difference between environmental and legal terms for

[7]http://www.simone.it/newdiz

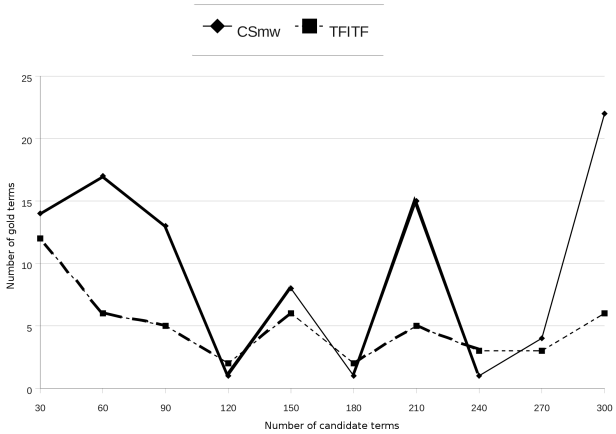

Figure 5: Absolute value of the difference between environmental and legal terms with CSmw

every group. The continuous line shows the CSmw trend, while the dashed one shows the TFITF trend, and in both lines the bold part refers to predominance of environmental terms. As we can see, the two peaks at the extremities, due to high differences in values, point out the function's success in distinguishing double domain terminology. The CSvH method turned out not to be suitable for this task, since this method cannot deal with double domain terminology by discerning different term types. In the first group of terms, as we can see from Table 6, the function seems to respect the general trend extracting more environmental than legal terms. But setting the usual threshold at 300, the proportion of environmental terms is still higher than legal term. For this reason, in order to find a turning point of this trend, where the legal terms would have been more than the environmental ones, we keep analyzing sample groups around 600 terms. At this point we see that there is still a stable ratio between terms belonging to the two different domains. We stop our evaluation where the list becomes too noisy for being analyzed. A possible explanation is that, since the CSvH method extracts multi–word terms from the single head term previously acquired, it extracts all complex terms which share the same single head term, including complex terms which are not relevant for that particular text. Namely, it could be the case that both *principio attivo* 'active ingredient' and *principio di sussidiarietà* 'principle of subsidiarity' were extracted since they share the single head term, i.e. *principio* 'principle'. However, we cannot discriminate that the first one belongs to the environmental domain while the second one to the legal domain.

Group	No. of multi–word terms
	Env/Leg
0-30	16/5
270-300	13/6
450-480	8/8
600-630	6/6

Table 6: Evaluation of CSvH method

5. Conclusion

In this paper we presented a novel approach to multi–word terminology extraction combining a well–known automatic term recognition approach, the C–NC value method, with a contrastive ranking technique, aimed at refining obtained results either by filtering noise due to common words or by discerning between semantically different types of terms within heterogeneous terminologies (as in the legal case). In the framework of this study, two new contrastive functions have been proposed, called TFITF and Contrastive Selection Multi–words function, which turned out to be particularly suitable for handling variation in low frequency events, typically represented by multi–word terms. The proposed methodology has been tested in two case studies carried out in the History of Art and Legal domains respectively. The evaluation of achieved results showed that the proposed two–stage approach improves significantly multi–word term extraction results. For what concerns the legal domain, the proposed approach provides an answer to a well known problem in the semi–automatic construction of legal ontologies, namely that of singling out law terms from terms of the specific domain being regulated; as a matter of facts, ontology learning efforts in the legal domain mainly focus on the latter (Francesconi et al., 2010). Current directions of reseach include: *i)* the definition of new functions for an in depth analysis of the 'twilight zone' described in Section 4.3. as part of the "double terminology" extraction task, and *ii)* the use of this approach to identify neologisms from a comparative analysis of diacronic corpora of newspapers texts.

6. Acknowledgments

The research reported in the paper has been supported in part by a grant from the Italian FIRB project "Piattaforma di servizi integrati per l'Accesso semantico e plurilingue ai contenuti culturali italiani nel web". The authors would like to thank Angela D'Angelo of the Scuola Superiore Sant'Anna of Pisa, Paolo Plini of the Institute of Atmospheric Pollution, Environmental Terminology Unit (CNR, Rome), and Elena Lazzarini of the Department of Art History of the University of Pisa, who contributed as domain experts to the evaluation process.

7. References

R. Basili, A. Moschitti, M. T. Pazienza, F. M. Zanzotto. 2001. A contrastive approach to term extraction. In *Proceedings of the 4th Conference on Terminology and Artificial Intelligence (TIA–2001)*, Nancy.

R. Basili, M. Pazienza, and F. Zanzotto. 2001. Modelling syntactic context in au- tomatic term extraction. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, Bulgaria.

T. M. Chung, P. Nation. 2004. Identifying technical vocabulary. *System, 32*, 251–263.

F. Dell'Orletta. 2009. Ensemble system for Part-of-Speech tagging. *Proceedings of Evalita'09*, Reggio Emilia, December 2009.

T. Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, vol. 19, no. 1:61–74.

E. Francesconi, S. Montemagni, W. Peters, D. Tiscornia. 2010. Integrating a Bottom–Up and Top–Down Methodology for Building Semantic Resources for the Multilingual Legal Domain. In E. Francesconi et al. (eds.), *Semantic Processing of Legal Texts*, Springer, LNAI 6036, pages 95–121.

K. Frantzi, S. Ananiadou. 1999. The C–value / NC Value domain independent method for multi–word term extraction. In *Journal of Natural Language Processing*, 6(3):145–179.

J. S. Justeson, S. M. Katz. 1995. Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, 1:9–27

L. Kozakov, Y. Park, T. Fin, Y Drissi, Y. Doganata, T. Cofino. 2004. Glossary extraction and utilization in the information search and delivery system for IBM technical support. *IBM Syst. J.* 43, 3:546-563.

G. Lame. 2005. Using NLP techniques to identify legal ontology components: concepts and relations. In Benjamins et al. (eds.), *Law and the Semantic Web. Legal Ontologies, Methodologies, Legal Information Retrieval, and Applications*, Lecture Notes in Computer Science, Volume 3369, pages 169–184.

A. Lenci, S. Montemagni, V. Pirrelli, G. Venturi. 2009. Ontology learning from Italian legal texts. In J. Breuker et al. (eds.), *Law, Ontologies and the Semantic Web – Channelling the Legal Information Flood, Frontiers in Artificial Intelligence and Applications*, Springer, Volume 188, pages 75–94.

R. Marinelli, et al. 2003. The Italian PAROLE corpus: an overview. In A. Zampolli et al. (eds.), *Computational Linguistics in Pisa*, Special Issue, XVI–XVII, Pisa–Roma, IEPI. Tomo I, pp. 401–421.

D. Maynard, S. Ananiadou. 1999. Term extraction using a similarity-based approach. In *Recent Advances in Computational Terminology*. John Benjamins, 1999.

H. Mima, S. Ananiadou. 2000. An application and evaluation of the C-NC Value approach for the automatic term recognition of multi–word units in Japanese. In *Journal of Natural Language Processing*, 6(2):175-194.

H. Nakagawa, T. Mori. 2003. Automatic Term Recognition based on Statistics of Compound Nouns and their Components. *Terminology*, Vol.9, No.2:201–219.

A. Penas, F. Verdejo, J. Gonzalo. 2001. Corpus-Based Terminology Extraction Applied to Information Access. In *Proceedings of Corpus Linguistics 2001*, 458–465.

G. Salton, C. Buckley. 1988. Term–Weighting Approaches in Automatic Text Retrieval. *Information Processing and Management*, vol. 24, n. 5, 513-523.

Š. Vintar. 2004. Comparative Evaluation of C–value in the Treatment of Nested Terms. In *Proceedings of Memura 2004 – Methodologies and Evaluation of Multi-word Units in Real-World Applications*, (LREC 2004 Workshop), 54–57.

T. Vu, A. Aw, M. Zhang. 2008. Term extraction through unithood and termhood unification. *IJCNLP*, Jan 2008.