Contrastive filtering of domain specific multi–word terms from different types of 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 tackle the challenging task of Multi-word term (MWT) extraction from different types of specialized corpora. Contrastive filtering of previously extracted MWTs results in a considerable increment of acquired domain specific terms.

1 Introduction

Multi-word term (MWT) extraction is a challenging and well-known automatic term recognition (ATR) subtask, aimed at retrieving complex domain terminology from specialized corpora. Although domain sublanguages are characterised by specific vocabularies, a well-defined border between specific sublanguages (SLs) and general language (GL) vocabularies is difficult to establish since lexicon shifts in a continuum from a highly specialized area to a transition area between GL and SLs (Rondeau et al., 1984). Within this continuum, Cabré (1999) identifies three types of lexical items: *a.* GL lexical items; *b.* SL terms, *c.* lexical items belonging to a borderline area between GL and SL. The proportion of these different types of lexical items varies depending on the text type. To our knowledge, automatic term recognition methods proposed so far in the literature focussed on highly specialized corpora (typically, technical and scientific literature), mainly characterized by SL terminology. However, the same ATR methods may not be equally effective when dealing with corpora characterised by a different proportion of term types; e.g. from texts such as Wikipedia articles, wich are conceived for a more extended audience, both SL terms and common words are acquired as long as they show a statistically significant distribution. In this paper, we claim that the contrastive approach to MWT extraction described in Bonin et al. (2010) can be effectively exploited to distinguish between common words and domain specific terminology in different types of corpora as well as to identify terms belonging to different SLs when occurring within the same text. The latter is the case of legal texts, characterized by a mixture of different SLs, i.e. the legal and the regulated–domain SLs (Breuker et al., 2004).

Effectiveness and flexibility of the proposed ATR approach has been tested with different experiments aimed at the extraction of domain terminology from corpora characterised by different degrees of difficulty as far as ATR is concerned, namely *i)* environmental scientific literature, *ii)* Wikipedia environmental articles, and *iii)* a corpus of legal texts belonging to the environmental domain.

2 General extraction method

The MWT extraction methodology we follow is organized in two steps, described in detail in Bonin et al. (2010). Firstly, a shortlist of well–formed and relevant candidate MWTs is extracted from a given target corpus and secondly a contrastive method is applied against the selected MWTs only. In fact, in the first stage, candidate MWTs are searched for in an automatically POS–tagged and lemmatized text and they are then weighted with the C–NC Value method (Frantzi et al., 1999). In the second stage, the list of MWTs extracted is revised and re–ranked with a contrastive score, based on the distribution of terms across corpora of different domains; in particular, the *Contrastive Selection of multi–word terms* (CSmw) function, newly introduced in Bonin et al. (2010), was used, which proved to be particularly suitable for handling variation in low frequency events. The main benefit of such an approach consists in its modularity; by first selecting valid MWTs which have significant distributional tendencies, and then by assessing their domain–relevance using a contrastive function, the MWT sparsity problem is overcome or at lest significantly reduced.

3 Experiments

The MWT extraction methodology described above has been followed in order to acquire environmental terminology from three different kinds of domain corpora. The first experiment has been carried out on a corpus of scientific articles concerning climate change research of Italian National Research Council (CNR), of 397,297 tokens, while the second experiment has been carried out on a corpus of Wikipedia articles from the Italian Portal “Ecologia e Ambiente” (Ecology and Environment) (174,391 tokens). As general contrastive corpus, we used, in both cases, the PAROLE Corpus (Marinelli et al., 2003)[1], in order to filter out GL lexical items. The third and more challenging experiment has been carried out on a collection of Italian European legal texts concerning the environmental domain for a total of 394,088 word tokens. In this case, as contrastive corpus we exploited a collection of Italian European legal texts regulating a domain other than the environmental one[2], in order to extract MWTs belonging to the environmental domain, but also to single out legal–domain terms, used in legal texts. For each acquisition corpus we followed the two–layered approach described above, selecting, firstly, a top list of 2000 environmental MWTs from the candidate term list ranked on the C–NC Value score and, secondly, re-ranking this 2000–term list on the basis of the CSmw function; then we extracted the final top list of 300 environmental MWTs. In order to assess the effectiveness of the approach against different types of corpora, we analyzed the two 300–term top lists of MWTs acquired respectively after the first and the second extraction steps. In both cases, we divided the 300–term top lists in 30–term groups which show domain-specific terms’ distribution, so that they could be easily compared. The evaluation has been carried out by comparing the lists of MWTs extracted against a gold standard resource, i.e. the thesaurus *EARTh (Environmental Applications Reference Thesaurus).*[3]. In addition, a second resource has been used in the third experiment for evaluating legal terms: the *Dizionario giuridico* (Edizioni Simone)[4]. Those terms which could not find a positive matching against the gold standard resources were manually validated by domain experts.

	Scient.Lit.		Wikipedia	
Group	C-NC	CSmw	C-NC	CSmw
0-30	22	27	27	29
30-60	28	25	28	26
60-90	24	30	25	25
90-120	19	28	23	27
120-150	25	29	23	24
Sub-TOT	118	139	126	131
150-180	25	25	22	20
180-210	23	27	20	30
210-240	24	29	23	26
240-270	23	25	24	24
270-300	21	19	15	25
TOT	234	264	230	256

Table 1: Environmental terms in the 300–term top lists from scientific articles (columns 2 and 3) and from Wikipedia (columns 4 and 5).

3.1 Discussion of results

Achieved experimental results highlight two main issues. Firstly, they show that the proposed contrastive approach to domain–specific MWTs extraction has a general good performance. As Figures 1, 2 and 3 show, the amount of envi-

[1] It is made up of about 3 million word tokens and it includes Italian texts of different types.

[2] A corpus of Italian European Directives on consumer protection domain for a total of 74,210 word tokens.

[3] http://uta.iia.cnr.it/earth.htm#EARTh%202002. Containing 12,398 environmental terms.

[4] Available online: http://www.simone.it/newdiz and including 1,800 terms.

	C–NC Value		CSmw	
Group	Env	Leg	Env	Leg
0-30	12	12	21	4
30-60	10	8	16	4
60-90	11	10	20	3
90-120	22	1	19	3
120-150	10	13	13	6
Sub-TOT	65	44	89	20
150-180	9	13	14	6
180-210	13	10	17	6
210-240	16	5	11	9
240-270	11	9	16	9
270-300	12	8	9	13
TOT	126	90	156	63

Table 2: Env(ironmental) and Leg(al) MWTs in the 300–term top list from the legal corpus.

ronmental MWTs after the contrastive stage increases with respect to the amount of MWTs acquired after the candidate MWT extraction stage carried out with the C–NC Value method. Secondly, reported results witness that such performances are differently affected by the different types of input corpora: as summarized in Table 3, the relative increment of environmental MWTs after the contrastive filtering stage ranges from 11.3% to 23.81%. Interestingly, as shown in Table 1, the results obtained in the first and second experiments show similar trends.

Type of text	% relative increment
Wikipedia	11.3%
Scientific articles	12.82%
Legal texts	23.81%

Table 3: Relative increment of environmental MWTs in the contrastive re–ranking stage

This is due to the overwhelming occurrence in the two input corpora of specialized terminology with respect to the GL items. Differently from what could have been expected, Wikipedia texts contain highly specialized terminology. However, a qualititative evaluation of MTWs extracted revealed that this latter corpus includes terms which belong to that borderline area between GL and SL (case *c.* in the Cabré (1999) classification). It follows that in the Wikipedia case the contrastive stage filtered out not only common words, such as *milione di dollari* ‘a million dollars’, but also terms such as *unità immobiliare* ‘real estate’ belonging to such borderline area of terminology; their difficult classification slightly decreases the contrastive stage performance. In the third experiment, the total amount of environmental MWTs percentually increased by 23.81% after the second stage of contrastive re–ranking. Differently from the previous experiments, in this case we faced the need for discerning terms belonging to the vocabulary of two SLs, i.e. regulated domain (i.e. environmental) terms and legal ones (e.g. *norma nazionale*, national rule): this emerges clearly from the results reported in Table 2 where it is shown that the same number of environmental and legal MWTs (i.e. 12 terms) are extracted at the first stage in the first 30–term group, and that the contrastive re–ranking allows the emergence of 21 environmental MWTs against 4 legal MWTs only. This trend can be observed in Figure 4, where the divergent lines show the different distributions of environmental and legal terms: interestingly, lines cross each other where legal terms outnumber environmental terms, i.e. in the last 30–term group. Such a relative increment with respect to the C–NC Value ranking can be easily explained in terms of the main features of the two methods, where C–NC Value method is overtly aimed at extracting domain–specific terminology (both environmental and legal terms), and the contrastive re–ranking step is specifically aimed at distinguishing the relevance of acquired MWTs with respect to the involved domains.

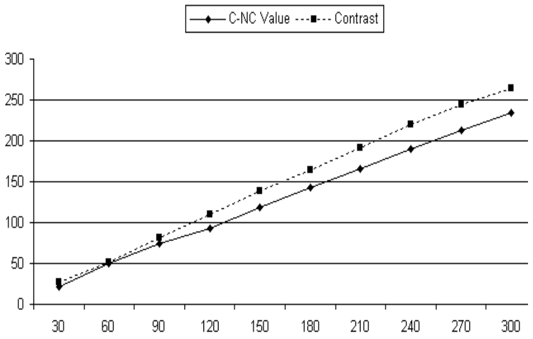

Figure 1: Scientific articles: comparative progressive trend of environmental extracted terms

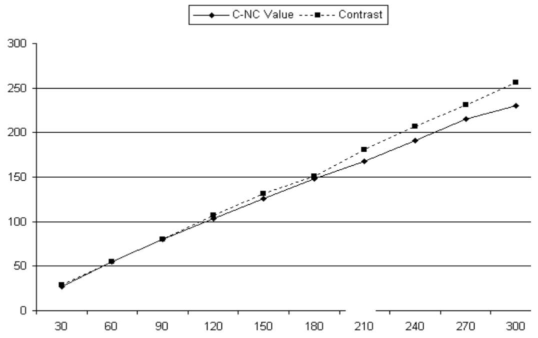

Figure 2: Wikipedia articles: comparative progressive trend of environmental extracted terms

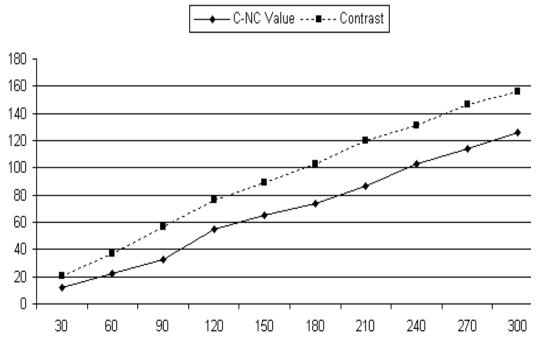

Figure 3: Legal texts: comparative progressive trend of environmental extracted terms

4 Conclusion

In this paper we tackled the challenging task of MWT extraction from different kinds of domain corpora, characterized by different types of terminologies. We demonstrated that the multi-layered approach proposed in Bonin et al. (2010) can be successfully exploited in distinguishing between GL and SL items and in assessing the domain–relevance of extracted terms. The latter is the case of type of multi–domain corpora, characterized by the occurrence of terms belonging to different SLs (e.g. legal texts). Moreover, the results obtained from different text types proved that the performance of the contrastive filtering stage is dramatically influenced by the nature of the acquisition corpus.

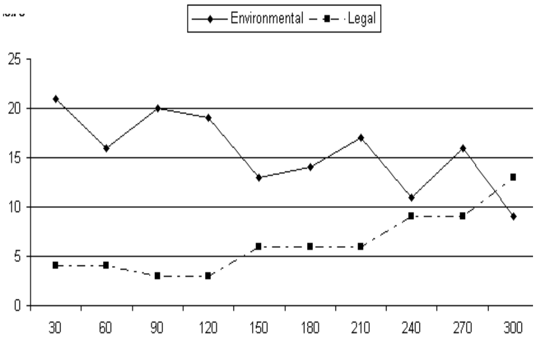

Figure 4: Legal texts: trend of contrastive function

5 Acknowledgments

The research has been supported in part by a grant from the Italian FIRB project RBNE07C4R9. Thanks are also due to A. D'Angelo (Scuola Superiore Sant'Anna, Pisa) and P. Plini (EKOLab, CNR, Rome), who contributed as domain experts to the evaluation.

References

F. Bonin, F. Dell'Orletta, G. Venturi, and S. Montemagni, 2010. *A Contrastive Approach to Multi–word Term Extraction from Domain Corpora,* in Proceedings of LREC 2010, Malta, 19–21 May, pp. 3222–3229.

J. Breuker, and R. Hoekstra, 2004. *Epistemology and Ontology in Core Ontologies: FOLaw and LRI-Core, two core ontologies for law,* in EKAW04 Proceedings, Northamptonshire, UK, pp. 15-27.

M.T. Cabré. 1999. *The terminology. Theory, methods and applications.* John Benjamins Publishing Company.

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.

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

G. Rondeau, J. Sager. 1984. *Introduction à la terminologie (2nd ed.), Chicoutimi, Gatan Morin.*