EXTRACTING TERMS WITH EXTra

Lucia C. Passaro
CoLing Lab
Dipartimento di Filologia,
Letteratura e Linguistica
University of Pisa (Italy)
lucia.passaro@for.unipi.it

Alessandro Lenci
CoLing Lab
Dipartimento di Filologia,
Letteratura e Linguistica
University of Pisa (Italy)
alessandro.lenci@unipi.it

Keywords: Term recognition, Multiword expressions, Information extraction, Ontology population, Automatic Indexing

Abstract

The identification and extraction of terms play an important role in many areas of knowledge-based applications, such as automatic indexing, knowledge discovery and management, as well as in computational approaches to terminology and lexicography. In this paper, we present EXTra, a tool designed to extract and calculate the degree of termhood of multiword expressions as a function of the statistical distribution of their parts and of the presence of other sub-terms. This work describes EXTra's algorithm, and provides the results of its evaluation on a task of term extraction from an Italian corpus of documents belonging to the domain of Public Administration.

INTRODUCTION

In recent years, the development of robust approaches to terminology extraction is playing an important role in many areas of knowledge-based applications such as automatic indexing, knowledge discovery and knowledge management. The need for domain terminology extraction has emerged from different disciplines and to answer to various goals, such as dictionary and thesaurus construction, text indexing, machine translation, automatic summarization, etc.

A general definition of *term* is "a surface representation of a specific domain concept" (Jacquemin, 1997; Pazienza, 1999). In general, a term can be either a single word or a multiword unit. In this study we focus on the latter kind of terms. The bag-of-words model, based on single word terms, is in fact a simplified representation of the lexicon used in natural language processing (NLP) and information retrieval (IR). We assume that "multiword expressions" (i.e. complex terms) range from completely opaque idioms to

semantically compositional word combinations (Evert, 2008). Multiword terms are less ambiguous and less polysemous than single word terms, yielding a better representation of the document content. Moreover, the lion share of domain concepts are normally expressed through multiword terms, which represent a crucial component of natural language lexicons (Jackendoff, 1997).

Term Extraction is a key application in Information Extraction (IE) and IR, and a crucial component to tackle several NLP tasks, such as Ontology Learning and Ontology population, Key-words extraction and Document Indexing. The recognition of complex terms from texts is performed on the basis of different criteria. Major differences exist between algorithms that take into account only the distributional properties of terms, such as frequency and TF / IDF (Salton and McGill, 1983), and those using contextual information such as syntactic, terminological and semantic features as in Maynard and Ananiadou (2000), Frantzi and Ananiadou (1999), Maynard (2000), Dell'Orletta et al., 2014, and Bonin et al. (2010). The common trait of most of the strategies above is the identification of a set of ranked candidates from texts, and then the application of a filtering function to separate real terms from non-terms. In this latter phase, the candidates are usually sorted according to their association strength as an estimate of their degree of termhood.

We have organized this paper as follows: In section 1, we present the term extractor EXTra by describing its approach to the candidate selection step, its original weighting algorithm and its possible parameters. In section 2, we report the evaluation of EXTra to extract domain terminology from documents belonging to the Italian Public Administration (PA). Section 3 reports the results obtained from the validation of the terms in this case study, focusing on the precision and on the quality of the ranking produced by EXTra.

1. EXTRA

The term extractor EXTra takes into account the linguistic structure of multiword terms by implementing a candidate selection step that uses manually-defined *structured* PoS-patterns. Moreover, in order to tackle the complexity of term phrases, EXTra adopts a new association measure that promotes terms composed by one or more sub-terms. The intuition is that the degree of termhood of a candidate pattern is a function of the statistical distribution of its parts, and of the presence of highly weighted sub-terms. The last step of EXTra applies a filtering function to separate real terms from wrong candidates. EXTra also includes various parameters that allow the user to optimize the extracted terms with respect to the target corpus and domain. In particular, the user can specify the set of structured patterns that guide the extraction process, a list of stopwords, and the thresholds for the association measure and the n-gram frequency. In the configuration file, the user also selects the association measure used by the weighting algorithm. The association measures currently implemented in EXTra are the Pointwise Mutual Information (Church and Hanks, 1990), the Local Mutual Information (Evert, 2008), and the Log Likelihood Ratio (Dunning, 1993), as well as an identity function weighting the n-grams with their raw frequency. In order to assure the flexibility of the system, a further parameter affects the importance given to long terms by the weighting algorithm (cf. section 1.2). The input of EXTRA is a PoS-tagged and lemmatized text in a tab-delimited CONLL format. The output of EXTra consists of two files: the input file enriched with the extracted multiword terms, and a list of multiword terms ranked according to their termhood.

TokenID	Token	Lemma	PoS	Term (EXTra)
1	Registro	registro	S	_
2	Generale	Generale	SP	_
3	n.	n.	SA	_
4	961	961	N	_
5	Del	di	EA	_
6	26/09/2012	26/09/2012	N	_
1	Copia	copia	S	_
1	Determinazione	determinazione	S	_
1	n.	n.	SA	_
2	140	140	N	_
3	del	di	EA	_
4	26/09/2012	26/09/2012	N	_
1	Settore	settore	S	_
2	Servizi	Servizi	SP	_
3	Finanziari	Finanziari	SP	_
1	Oggetto	oggetto	S	_
2	:	:	FC	_
3	Affidamento	Affidamento	SP	_
4	del	di	EA	_
5	servizio	servizio	S	servizio_di_manutenzione
6	di	di	E	servizio_di_manutenzione
7	manutenzione	manutenzione	S	servizio_di_manutenzione
8	presidi	presidio	S	_
9	antincendio	antincendio	A	_
10	posti	posto	S	_
11	negli	in	EA	_
1	edifici	edificio	S	edificio_di_proprietà
2	di	di	E	edificio_di_proprietà
3	proprietà	proprietà	S	edificio_di_proprietà

Term (EXTra)	PLMI
bollettino_ufficiale_di_regione_autonomo	2059330,459
carta_libero_ad_uso_amministrativo	1363747,944
carta_libero_per_uso_amministrativo	1363747,944
originale_in_carta_libero	453971,4196
copertura_finanziaria	90748,258
attribuzione_al_dirigente_di_dotazione	78023,94
diminuzione_permanente_di_capacità_lavorativo	74245,3462
protezione_dei_dati_personali	62194,11451
trattamento_di_dati_personali	61056,1527
trattamento_di_dati_personali	61056,1527
incarico_di_dirigenza_di_settore	60877,00747
dinamica_costiero_in_unità_fisiografica	56705,82765
codice_identificativo_di_gara	52275,99217
intervento_di_assistenza_sociale	46027,35039
cambio_di_destinazione_di_uso	44029,83091
cambio_della_destinazione_di_uso	44029,83091
briglia_esistente_su_torrente_lucido	42739,01878
equiparazione_stabilito_da_legge_vigente	42473,67371
dichiarazione_sostitutivo_di_atto_notorio	42074,08231
rispetto_di_normativa_previsto	39939,49118

Figure 2. Examples of output files produced by EXTra

1.1. Candidate selection

Candidate terms are identified using manually-defined *structured PoS patterns* that represent the recursive phrase structure of terms. A *structured PoS pattern* is a bracketed list of constituents, where each constituent can be either a sequence of two *content PoS* or another bracketed constituent. This structure defines long term patterns as a composition of smaller patterns. The c*ontent PoS* are specified in the configuration file, allowing the user to exclude from the termhood computation particular classes of PoS (e.g. articles and prepositions). The following is an example of *structured PoS pattern*:

[[noun (-s), preposition (-e), noun (-s)], preposition (-ea), [noun (-s), adjective (-a)]]

It is composed by two constituents, [noun (-s), preposition (-e), noun (-s)] and [noun (-s), adjective (-a)]. This structured pattern identifies the candidate "Politica di sviluppo delle Risorse Umane" *(human resource development policy)*. Following the pattern structure and ignoring prepositions, we can isolate two embedded sub-terms: [politica-s di-e sviluppo-s] ([noun (-s), preposition (-e), noun (-s)]) and [risorse-s umane-a] ([noun (-s), Adjective(-A)]). From a computational point of view, during the candidate selection phase, EXTra first stores the statistical information of each sub-patterns (e.g., the frequencies of the embedded pairs <Politica, Sviluppo> and <Risorse, Umane>), and then stores the frequency of the aggregate pair <politica_di_sviluppo, risorse_umane>.

1.2. Weighting algorithm

The structure of the PoS patterns is also used to guide the process of statistical term weighting by following the same order of incremental composition. Following a recursive structure, the weighting algorithm assigns a termhood score to each of the embedded phrases, and then computes the global score for the complex term by combining the partial weights of its components.

EXTra's term weighting algorithm is applied recursively to the internal structure of the patterns: At the base step it measures the association strength σ of each candidate two-

word term $< w_1, w_2 >$by computing standard association measures, such as for instance Pointwise Mutual Information (PMI). The candidates whose score σ is above an empirically fixed threshold are added to the set of the terms T={t_1,...,t_n}. In the recursive step, EXTra measures the association strength σ of any n-word candidate term <c_1, c_2> by combining the association strengths of its sub-elements. The termhood of a candidate is calculated using the following formula:

$$\sigma(c_1, c_2) = S(c_1) * S(c_2)$$

If $c_i \notin T$, $S(c_i) = 1$, else $S(c_i) = (\log_2 \sigma(c_i)) / k$. As we said above, this weighting scheme formalizes the assumption that the termhood of longer terms depends on the degree of termhood of their parts. The parameter k controls the contribution of sub-terms to the weight of longer terms: The smaller the k, the higher the weight assigned to longer terms containing them.

Coming back to the previous example

[[politica (-s), di (-e), sviluppo (-s)], delle (-ea), [risorse (-s), umane (-a)]]

in the base step, EXTra measures the association strength σ of each two-word term <w_1, w_2> using standard association measures. Supposing that the selected association measure is the PMI, at the base step EXTra measures the scores for the pairs <risorse-s, umane-a> and <politica-s, sviluppo> and it stores their termhood value. In the recursive step, the system calculates the score σ between the sub-candidates <politica_di_sviluppo, risorse_umane> by applying the formula:

σ (politica_di_sviluppo, risorse_umane) = S (politica_di_sviluppo) * S (risorse_umane).

Since $\sigma(< c_1, c_2 >)$as$S(c_1)$*$S(c_2)$both the sub-terms belong to the set of accepted terms T, the termhood score σ is calculated using the formula $S(c_i) = (\log_2 \sigma(c_i)) / k$.

1.3. Filtering

Candidate multiword terms are filtered by using three main filters. First of all, an optional stoplist is used to exclude the terms containing one or more words in the blacklist during indexing. Then, patterns with a frequency below a frequency threshold are discarded before computing their strength of association. Finally, the association measure filter defines the minimum strength of association that an n-gram must have to be considered as a multiword term: The candidates whose score σ is above an empirically fixed threshold are then added to the set of terms T.

2. EVALUATING EXTRA

We have evaluated EXTra on a term extraction task in the Italian Public Administration (PA) domain. This is a particularly challenging domain because of the highly heterogeneous nature of its terminology, which typically includes domain terms belonging to the multifarious fields covered by PA, ranging from the management of schools up to urban planning and health care.

As a preliminary step, we automatically collected PAWaC! (Public Administration Web as Corpus) which contains documents extracted from the Italian online "Albo pretorio" (Council notice board) of various small and medium municipalities in Tuscany. Most of these documents are "Delibere" (Town council resolutions), "Determine" (Executive resolutions), and generic administrative acts, such as bidding processes, local regulations etc. PAWaC includes 15,321 documents, for a total of 34,725,652 tokens and 17,272,068 content words (nouns, adjectives and verbs). We PoS-tagged the corpus using the PoS-Tagger described in Dell'Orletta (2009) and we identified the list of structured PoS patterns showed in Table 5. Since we were mainly interested in extracting nominal phrases, the list of the patterns only include nouns, adjectives and prepositions (Justeson et al., 1995).

PoS Structured pattern	**Example**
[noun, adjective]	delibera comunale (*municipal resolution*)
[noun, preposition, noun]	presidente del consiglio (*Prime Minister*)
[[noun, adjective], preposition, noun]	delibera comunale di giunta (*municipal council resolution*)
[[noun, adjective], preposition, [noun, adjective]]	gestione provvisoria delle risorse finanziarie (provisional management of financial resources)
[noun, preposition, [noun, adjective]]	ordine di regolarità contabile (*accounting consistency order*)
[noun, preposition, [noun, preposition, noun]]	approvazione del verbale di gara (approval of the bidding process)
[[noun, preposition, noun], preposition, [noun, adjective]]	politica di sviluppo delle risorse umane (*human resource development policy*)
[noun, preposition, [noun, preposition, [adjective, noun]]]	Attestazione del responsabile del servizio finanziario (*declaration of the financial service manager*)

Table 5. Structured PoS patterns

We have made several experiments (section 2.1.1) with EXTra, in which we used the same set of PoS patterns but with different association measures and different k values in order to assign a different boost to long terms. Although the patterns include prepositions, they are not considered in the computation of termhood, which is calculated only considering the strength of association between nouns and adjectives.

2.1.1. Experiments

We tested EXTra with three different association measures (one of which is the raw term frequency, which we used as a baseline) and three values for the parameter k, in order to control the importance assigned to long terms in the ranking. For each configuration, from left to right Table 6 shows the association measure, the value of the parameter k , the minimum n-gram frequency and the number of extracted terms.

Configuration	Association measure	*k*	Min. Freq	#Terms
Frequency.Knull.3	Frequency	-	3	65,120
PLMI.K1.3	Positive Local MI	1		58,380
PLMI.K5.3		5		58,380
PLMI.K10.3		10		58,380
PPMI.K1.15	Positive Pointwise MI	1	15	13,032
PPMI.K5.15		5		13,032
PPMI.K10.15		10		13,030

Table 6. Configuration

In the case of the baseline configuration, we do not provide any boost to long terms, hence the parameter k is not specified. In our experiments, we used the Positive LMI (PLMI) and the Positive PMI (PPMI), in which negative scores are changed to zero, and only positive ones are considered. Following Evert (2008), PMI has been calculated as $log_2 (O/E)$ and the LMI has been calculated as $O * log_2 (O/E)$, where O is the observed co-occurrence frequency and E is the expected frequency under the null hypothesis of independence (i.e. complete absence of association). For PPMI and PLMI models, we specified the following values of k: $k = 1$ (maximum boost for long terms), $k = 5$ (medium boost) and $k = 10$ (low boost). The frequency threshold has been set to 3 in the models based on frequency and PLMI, but it was increased to 15 in the PPMI ones, because of the well-known bias of this association measure towards low-frequency n-grams.

3. RESULTS

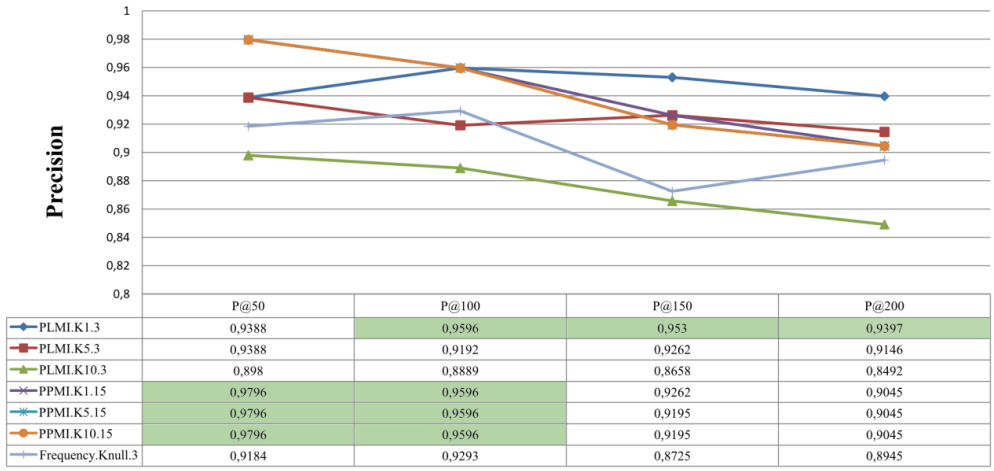
In order to evaluate the precision of EXTra, a domain expert judged the top 200 terms produced by EXTra for each configuration in Table 6. The annotator was asked to decide whether a candidate term was both a valid multiword expression and a domain-specific term in the field of PA. For example, the candidates "Ponte levatoio" (drawbridge) was discarded because it is not a domain term, while the candidate "Documento di identità in corso" was discarded because it was a truncation of the term "Documento di identità in corso di validità" (Valid identity document). Both the previous candidates were therefore labeled as False Positives (FP). On the contrary, the candidate "Esercizio finanziario" (fiscal year) complies with both requirements, and therefore was considered a True Positive (TP).

The global Precision was calculated for the top 200 evaluated terms as TP / (TP + FP). Table 7 shows the results. The baseline model (Frequency.Knull.3) obtained a precision score of 0.89. The best model is PLMI with the maximum weight for long terms ($k = 1$). The worst model, outperformed by the baseline, is PLMI with the lowest salience assigned to long terms ($k = 10$). The precision for PPMI models is stable with respect to the parameter k, scoring 0.905. On average, the models reach a precision score of ~0,9.

Configuration	k	Min. Freq	#Terms	Precision
Frequency.Knull.3	-	3	65,120	0.89
PLMI.K1.3	1	3	58,380	**0.935**
PLMI.K5.3	5	3	58,380	0.915
PLMI.K10.3	10	3	58,381	0.85
PPMI.K1.15	1	15	13,032	0.905
PPMI.K5.15	5	15	13,032	0.905
PPMI.K10.15	10	15	13,030	0.905

Table 7. Global precision

The quality of the termhood ranking produced by EXTra has been evaluated by considering the Precision@n with $1 \leq n \leq 200$. Figure 3 reports the Precision@n for $n \in \{50, 100, 150, 200\}$. We can observe that for the top 50 terms, the best performing models use PPMI, reaching a Precision of 0.98. Considering top 100 terms, the precision decreases for PPMI models, and increases for the best PLMI one. Going down in the ranking, the precision of all models decreases, as expected. If we consider the quality of the ranking (Figure 3), we can notice that for the top 50 terms, the discriminating factor lies in the type of association measure. In fact, all PPMI models reached a P@50 equal to 0.98. PLMI models, on the contrary, reached a score ranging from 0.90 and 0.94. In addition, we can observe that PLMI, but not PPMI models are influenced by the *k* parameter. The results concerning the PLMI models show a trend in which precision seems to be inversely proportional to the parameter *k*. In other words, the precision of the model decreases when we reduce the importance of sub-terms. This trend is not evident in the PPMI models, in which the precision seems to be independent of the parameter *k*.



	P@50	P@100	P@150	P@200
PLMI.K1.3	0,9388	0,9596	0,953	0,9397
PLMI.K5.3	0,9388	0,9192	0,9262	0,9146
PLMI.K10.3	0,898	0,8889	0,8658	0,8492
PPMI.K1.15	0,9796	0,9596	0,9262	0,9045
PPMI.K5.15	0,9796	0,9596	0,9195	0,9045
PPMI.K10.15	0,9796	0,9596	0,9195	0,9045
Frequency.Knull.3	0,9184	0,9293	0,8725	0,8945

Figure 3. Ranking evaluation

Presumably, the contrast between PPMI and PLMI determines the different behavior of the models. In fact, PPMI favors more idiosyncratic, low-frequency expressions, while PLMI has a greater bias towards frequent expressions (Evert 2008). This might be the reason why the weighting algorithm works better with PLMI, in which the boost given to long terms is more evident.

Error analysis shows that a great portion of the FPs depends on the fact that the candidates were correct multiword terms that did not belong to the domain of the PA. This fact prompts us to enrich EXTra with additional features to identify genuine domain terms, for instance by computing a confidence score based on the distribution of the terms in domain vs. general corpora (Penas et al., 2001; Chung et al., 2004; Basili et al., 2001).

4. CONCLUSIONS

In this paper, we have introduced EXTra, a term extractor designed in order to identify multiword terms taking into account both their linguistic structure and their internal complexity. In EXTra, the degree of termhood of a candidate pattern is a function of the statistical distribution of its parts, and of the presence of highly weighted sub-terms. EXTra only requires a PoS-tagged corpus and a set of PoS-patterns defining the phrase structure of candidate terms. Therefore, it can easily be adapted to different languages and domains in an economic and very scalable way.

The proposed methodology has been tested on the domain of Italian PA achieving very good results. However, we are aware that a better evaluation of EXTra requires us to compare the extracted terms against domain-specific terminological resources such as ontologies or thesauri, which we plan to do in the near future. Moreover, we aim at implementing additional association measures, a more efficient way of specifying the structured PoS patterns and statistical filters to single out genuine domain multiword expressions from general ones.

Acknowledgments

Lucia C. Passaro received support from the Project SEMantic instruments for PubLIc administrators and CitizEns (SEMPLICE), funded by Regione Toscana (POR CReO 2007-2013). We wish to thank Anna Gabbolini for helping us in validating EXTra on the domain of Italian Public Administration.

References

BASILI R., MOSCHITTI A., PAZIENZA M. T. AND ZANZOTTO F. M. (2001). *A contrastive approach to term extraction*. In Proceedings of the 4th Conference on Terminology and Artificial Intelligence (TIA–2001), Nancy.

BONIN, F., DELL'ORLETTA, F., MONTEMAGNI, S., VENTURI, G. (2010). *A Contrastive Approach to Multi-word Extraction from Domain-specific Corpora*. In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10). Valletta (Malta), May 19-21 2010: European Language ResourcesAssociation (ELRA).

CHUNG T. M. AND NATION P. (2004). *Identifying technical vocabulary*. System, 32, 251–263.

CHURCH, K. W. AND HANKS, P.(1990). *Word association norms, mutual information, and lexicography*. Computational Linguistics, 16(1), 22–29.

DELL'ORLETTA F.,VENTURI G., CIMINO, A., MONTEMAGNI, S., (2014). *T2K2: a System for Automatically Extracting and Organizing Knowledge from Texts*. In Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC'14). Reykjavik (Iceland), May 26-31, 2014: European Language Resources Association (ELRA).

DELL'ORLETTA, F. (2009). *Ensemble system for Part-of-Speech tagging*. In Proceedings of EVALITA 2009, Reggio Emilia, Italy.

DUNNING, T. E. (1993). *Accurate methods for the statistics of surprise and coincidence*. Computational Linguistics, 19(1), 61–74.

EVERT, S. (2008). *Corpora and collocations*. In A. Lüdeling and M. Kytö (eds.), Corpus Linguistics. An International Handbook, article 58. Mouton de Gruyter, Berlin.

FRANTZI K.T. AND ANANIADOU S., 1999. *The C-Value/NC-Value domain independent method for multi-word term extraction*. Journal of Natural Language Processing, 6(3):145–179.

JACKENDOFF R. (1997). *The Architecture of the Language Faculty*. Cambridge, MA: The MIT Press.

JACQUEMIN, C., (1997). *Variation terminologique: Reconnaissance et acquisition automatiques de termes et de leurs variantes en corpus*. Mémoire d'Habilitation à Diriger des Recherches en informatique fondamentale, Universitè de Nantes, France (1997).

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

KAGEURA, K. AND UMINO, B. (1996). *Methods of automatic term recognition: a review*. Terminology, 3(2), pp. 259–289.

MAYNARD D.G AND ANANIADOU S., 2000. *Identifying terms by their family and friends*. In Proc. of 18th International Conference on Computational Linguistics (COLING). Saarbrücken, Germany, July 31 - August 4, 2000: Association for Computational Linguistics.

MAYNARD D.G., 2000. *Term Recognition Using Combined Knowledge Sources*. PhD thesis, Manchester Metropolitan University, UK, 2000.

PAZIENZA, M.T., (1999). *A domain specific terminology extraction system*. In: International Journal of Terminology. Benjamin Ed., Vol.5.2 (1999) 183-201.

PENAS A., VERDEJO F. AND GONZALO J. (2001). *Corpus-Based Terminology Extraction Applied to Information Access*. In Proceedings of Corpus Linguistics 2001, 458–465.

SALTON,G. AND MCGILL M. J., 1983. *Introduction to Modern Information Retrieval*. New York: McGraw-Hill, 1983.