International Journal
of
English Studies

IJES

UNIVERSITY OF MURCIA www.um.es/ijes

A statistical approach to term extraction

ROGELIO NAZAR*
Instituto Universitario de Lingüística Aplicada
Universidad Pompeu Fabra

Received: 4 June 2011 / Accepted: 3 December 2011

ABSTRACT

This paper argues in favor of a statistical approach to terminology extraction, general to all languages but with language specific parameters. In contrast to many application-oriented terminology studies, which are focused on a particular language and domain, this paper adopts some general principles of the statistical properties of terms and a method to obtain the corresponding language specific parameters. This method is used for the automatic identification of terminology and is quantitatively evaluated in an empirical study of English medical terms. The proposal is theoretically and computationally simple and disregards resources such as linguistic or ontological knowledge. The algorithm learns to identify terms during a training phase where it is shown examples of both terminological and non-terminological units. With these examples, the algorithm creates a model of the terminology that accounts for the frequency of lexical, morphological and syntactic elements of the terms in relation to the non-terminological vocabulary. The model is then used for the later identification of new terminology in previously unseen text. The comparative evaluation shows that performance is significantly higher than other well-known systems.

KEYWORDS:

English technical terminology, terminology extraction, computational terminography, quantitative linguistics.

RESUMEN

Este artículo presenta argumentos en favor de una aproximación estadística a la extracción de terminología, general a todas las lenguas pero con parámetros específicos para cada una de ellas. En contraste con la tendencia general en terminología aplicada, que suele ser específica de una lengua y un dominio de especialidad, el presente artículo adopta unos principios generales acerca de las propiedades estadísticas de la terminología especializada y un método para obtener los parámetros correspondientes a una lengua en particular. Este método se utiliza para la identificación automática de los términos en los textos, y su efectividad es evaluada en este artículo mediante un estudio empírico en el caso de la terminología médica en inglés. El modelo requiere escasa complejidad teórica y computacional, y no necesita recurrir a fuentes de conocimiento lingüístico u ontológico. Este algoritmo aprende automáticamente a identificar términos durante una fase de entrenamiento en que se utilizan conjuntos de ejemplos de unidades terminológicas y no terminológicas. Con estos ejemplos, el algoritmo elabora un modelo de los términos teniendo en cuenta la frecuencia de elementos léxicos, morfológicos y sintácticos en relación al vocabulario no terminológico. Sobre la base de este modelo, identificará luego nuevos términos en nuevos textos. El estudio comparativo demuestra que el presente algoritmo tiene un desempeño significativamente superior al de otros sistemas conocidos.

PALABRAS CLAVE:

Terminología especializada en inglés, extracción de terminología, terminografía computacional, lingüística cuantitativa.

Address for correspondence: Rogelio Nazar. Institute for Applied Linguistics. Pompeu Fabra University, Roc Boronat 138, 08018, Barcelona, Spain. Tel: 00 34 935421187; E-mail: rogelio.nazar@upf.edu

© Servicio de Publicaciones. Universidad de Murcia. All rights reserved.
ISSN: 1578-7044

1. INTRODUCTION

Technical terms, or, more broadly speaking, specialized terminology, constitute the portion of the vocabulary of a language which has a special meaning for the community targeted by the texts where terms occur. This community shares a common area of interest and a variable amount of knowledge in particular thematic domains. Abstracted from each particular text, terms refer to specific concepts in this common knowledge and these concepts are interrelated configuring a conceptual network necessary for a precise interpretation of the texts. This paper adopts terms as its object of study and proposes a methodology to separate them from the rest of the vocabulary of the texts with a comparatively high degree of precision.

As an object of study, terminology has attracted the attention of different professional groups. Evidently, the most numerous group is formed by translators working with technical literature. The greatest difficulty of this task is the proper selection of the terminology in the target language. Most probably, the translator will find terminological units in the source language which are not yet documented in terminographic resources. This never-ending need is pressing terminologists to create and update terminographic resources at a very fast pace, generating in turn the demand for computer-aided terminology management and terminology extraction from corpus, also called computational terminography (Bourigault et al. 2001).

Parallel to the practical or application-driven research on terminology, terms have been a topic of interest of national and international normalization organisms, which seek to provide order and guidelines of proper use. These organisms are informed by a prescriptive tradition represented by Wüster (1979) and followers (Felber, 1984; Arntz & Picht, 1995 and others). Terminology is also an object of study of linguistics, and there exist numerous lines of research presenting diverse theories of terminology as an object of study abstracted from its language or domain (Cabré, 1999; Gaudin, 1993; Kageura, 1995; Sager, 1990; Temerman, 2000). Linguists working in this area have to be informed on developments in the field of computational terminography because the techniques used to study corpora can be an invaluable source of data for a linguistic approach to terminology. Conversely, by elaborating a general theory of terms, linguists can contribute to the development of better terminology extraction software. This is the point of view adopted in the present paper, which is intended to explore a general theory of terms rigorous enough to have a predictive power, that is to say, a theoretical construct capable of determining how likely it is that a given segment of text is a terminological unit.

The fact that the method casts the problem of term extraction purely in statistical terms and disregards external sources of knowledge can be interpreted as the most significant difference with other approaches that can be categorized broadly as applied linguistics. In the case of this discipline, the goal is not the advancement of knowledge on a theory of terminology, but rather the production of high quality raw terminographic material for the compilation of glossaries, and for that purpose, it is natural to use all sources of information available. On the contrary, from the point of view of theoretical linguistics, exogenous sources of knowledge are not considered relevant. There are, however, common interests in this case for both disciplines, since a term extraction algorithm that does not need abundant linguistic resources has the advantage of being easily adapted for languages that are less documented, which is, unfortunately, the current situation of the vast majority of the languages of the world.

In this context, the present paper explores a mathematical approach that could help us to formalize the object of study, by formulating a model to separate the technical terms from the rest of the vocabulary as signal from noise. It is a model based on universal principles, but includes a methodology to obtain the specific parameters of the language, English in the case of the empirical study presented in this paper. This mathematical model of terms is built during a training phase that records the study of examples of both terminological units and a reference corpus of general language text. Later, in the analysis phase, this model is used to predict whether certain fragments of specialized texts, being single or multi-word units, are most likely terms, and this is done taking into consideration the frequency of certain features in contrast to the reference general language corpus. These features are lexical, morphological and syntactic patterns, which are easier to collect and process than features such as the collocational or distributional behavior of the analyzed units. While it is possible to study the distributional behavior of terms using co-occurrence graphs (see Section 2), these techniques demand more computational complexity. In comparison with these distributional proposals, the general approach presented in this paper is very efficient and its performance remains competitive. The algorithm is easy to implement and fast to execute, analyzing a million words in a few seconds, which makes the present proposal an ideal approach for web based applications, where speed of response is a critical variable. In comparison with more simple statistical algorithms for terminology extraction which are based on frequency, an important property of the present proposal is that it is not vulnerable to low frequency units. Statistical analysis, in the present case, includes hapax legomena and dis legomena. This means that the algorithm reaches high recall and, at the same time, that it can analyze minimal samples of text, as is the case of the experiment reported here.

The rest of the paper is organized as follows: Section 2 is devoted to the review of previous research on the subject of automatic terminology extraction. Section 3 will present the approach adopted in this paper. Section 4 presents the results obtained in an experiment with English medical terms and in Section 5 these results are evaluated and compared with the results obtained with different term extraction systems on the same data. Section 6, finally, draws some conclusions and lines of future work.

2. RELATED WORK

There have been reports on algorithms for terminology extraction for more than forty years, if we take into account the literature produced in quantitative linguistics and in information retrieval on the specific topic of « keyword extraction ». An example of such algorithms is the still very popular TF.IDF (Term frequency – Inverse Document Frequency, Sparck Jones, 1972). This algorithm is used as a measure of term specificity, or in other words, a method to determine how important or informative a given term is in relation to a particular document and a particular corpus. In a sufficiently large and diverse corpus –comprising documents of general and domain specific topics–, the fact that a particular term has a significantly high frequency of occurrence in a relatively short number of documents can be interpreted as a sign that this particular term has a specialized meaning. On the contrary, units with a frequency that shows uniform distribution within the entire corpus can be considered units of the general language, non-informative or non-interesting from the information retrieval and terminological perspective. The algorithm does not answer a yes/no question, but it is useful to rank the vocabulary of a given corpus according to the score each unit obtains, in order to retain those best positioned and discard the rest. The work of Juilland and Chang (1964) shows a somewhat similar intuition. In a large corpus organized by thematic domains, those units which have similar frequency in all partitions of the corpus are likely to be the core vocabulary of a language. Eliminating this vocabulary from the whole corpus, one is left with the vocabulary used in each domain.

In the late eighties, with the advancement and wide spread use of technology, publications in the field of terminology extraction experienced a steady growth, and strategies diversified into three main lines of research: 1) statistical approaches, which are mainly language-independent; 2) symbolic or rule-based language-specific approaches, and 3) hybrid approaches, combining the first two.

Within the statistically-oriented stream, strategies can again be divided into those focused on « unithood » and those on « termhood » (Kageura & Umino, 1998). The first term refers to the degree in which multi-word terminological units show strong syntagmatic association, and the second the degree in which a given term represents a specific concept in a particular domain. Some authors have applied different measures of association to try to determine the degree of unithood of multi-word terminological units, that is, applying frequency based statistics to the single-word components of multi-word terms (Dagan & Church, 1994; Daille, 1994; Enguehard & Pantera, 1994; Frantzi, 1997; Frantzi et al, 2000; Pantel & Lin, 2001). A simple extension of this idea, still in the study of the syntagmatics of multi-word terminology, is to observe how many distinct multi-word terms contain a particular single term component (Nakagawa & Mori; 2002). If a component is productive, that is, if it persistently appears as part of a limited number of word *n*-grams, this could be interpreted as indicative that these *n*-grams are terminological.

With respect to the degree of termhood, there have been attempts to study it with distributional statistics elaborating on Sparck Jones' (1972) idea. Aside from assuming that a term will have high frequency in few documents, one could also expect that the set of documents where the term occurs will have something in common, some kind of quantitatively measurable deviation with respect to the average (Hisamitsu et al., 2000). Other authors have suggested that distributional strategies can be used by comparing the frequency of term candidates in specialized corpora with their frequency in reference corpora of general language (Scott, 1997; Böhm et al, 2002; Drouin, 2003; Kilgarriff et al., 2004). Streiter et al (2002) offer an approach with a philosophy similar to the present paper. These authors report an experiment of terminology extraction based on examples instead of abstract rules. With the motivation of working in less documented languages, they describe a supervised algorithm trained with examples of terms –and not from reference corpora– which learns surface features such as affix patterns, graphic patterns and length patterns. Affix patterns are, in this case, the last letter of each component of the multi-word training set terms; graphic patterns are sequences of initial upper or lower case letters in the same training set and length patterns are the mean character length of the training set terms plus/minus three standard deviations. In terms of precision, however, they report modest results.

In recent years, a new statistical approach to both termhood and unithood that has gained much attention is the application of graphs of lexical co-occurrence. Inspired in graph theory as used in social sciences for the study of social networks (Barnes & Harary, 1983; Watts & Strogatz, 1998), these graphs characterize the profile of co-occurrence of a given unit and can therefore be used to map the meaning (or meanings) of a given term from the textual corpus and not from hand crafted lexical resources. Empirical research shows that is an effective method for word sense disambiguation and induction, term extraction and even term translation (Chatterjee et al. 2010; Madani & Yu, 2010; Matsuo et al, 2006; Nazar, 2010; Véronis, 2004; Widdows & Dorow, 2002). In this line of research, a graph for each candidate has to be elaborated from very large corpora such as the web, to afford a minimum recall. Contrary to the approach of the present article, co-occurrence networks can be used to analyze units without taking into consideration any aspect of the form of the term, but rather its distributional behavior only. It can be used, for instance, to tell that a seemingly common word, such as *difference* or *mouse*, can have a specialized use in certain domains, such as mathematics or computer hardware. Graphs can be used to cluster the contexts of occurrence of a given unit by spotting attractors or hubs, which are regions in the graph showing high density and interconnection of nodes. In one of the senses of the term *mouse,* for instance, one can see an attractor concentrating terms such as *keyword*, *screen*, *click*, *double click*, and so on. As already noticed in the introduction, such models can be interesting from scientific point of view, but, depending on the particular implementation, may also be computationally too expensive for certain applications, such as a web based terminology extraction system.

With respect to more linguistically-informed strategies, the vast majority has relied heavily on syntactic patterns, defining sequences of Part-of-Speech tags that are likely terminological (Bourigault, 1996; Jacquemin, 2001), searching for Greek and Latin roots in the candidates (Ananiadou, 1994) or including ontologies (Vivaldi & Rodríguez, 2006) or the Wikipedia (Vivaldi & Rodríguez, 2011). The latter two can be considered hybrid proposals, since they combine both statistical and rule based systems. Justeson and Katz (1995) also combined frequency information with handcrafted lists of noun phrase patterns, under the assumption that frequent noun phrases in texts are likely terminological. Patry and Langlais (2005) tried to automatically acquire such patterns directly from examples provided by a user in a supervised manner. Different combinations of statistics with syntax and semantics can also be found (Maynard & Ananiadou, 2001; Vivaldi, 2001; Vivaldi & Rodríguez, 2001).

In comparison with the above mentioned studies, and considering the balance between cost of application and quality of the results, the method reported in Nazar and Cabre (2011) on experimental terminology extraction in Spanish might be the most simple and effective solution, and it is this methodology the one that will be replicated for the present experiments in English (Section 3). This methodology has been recently adopted in the Terminus software (Cabré & Nazar, forthcoming), which is now available online. Table 1 shows the URL and references of different term extractors that are capable of analyzing English text and have online demos available. These tools where used for the evaluation in Section 5. TermFinder and Yahoo Api are proprietary algorithms. Fivefilters is a simple statistically based term extractor, meant to be an open source alternative to the Yahoo Api. The rest of the systems have already been mentioned in this section.

Software	**Reference**	**URL**
TerMine	Frantzi et al., 2000	http://www.nactem.ac.uk/software/termine/cgi-bin/termine_cvalue.cgi
TermFinder	Translated Labs, online.	http://labs.translated.net/terminology-extraction/
Yahoo API	YAHOO, online.	http://developer.yahoo.com/search/content/V1/termExtraction.html
Termostat	Drouin, 2003	http://idefix.ling.umontreal.ca/~drouinp/termostat_web/
Fivefilters	Minoukadeh, online.	http://fivefilters.org/term-extraction/
Terminus	Cabré & Nazar, forthcoming	Now hosted on a provisional URL: http://iula05v.upf.edu When finished, the official URL will be: http://terminus.upf.edu/

Table 1: Online term-extraction services

3. METHODS AND MATERIALS

As already stated in the introduction, the point of departure of the terminology extraction method presented here is a supervised statistical algorithm, with a training phase where examples of both terminological and non terminological units are provided by a user. With this training set, the algorithm will record the frequency of events that are produced on lexical, morphological and syntactic levels. The idea is that if there are certain events that are more likely to occur in the set of terms of the training data in comparison with the non-terminological set, then these events can be considered intrinsic characteristics of the terms. The following subsections explain the learning process at each of the above mentioned levels in more detail and how the learned data are used in the analysis of new data. The training material is the same on all levels: on the one side, a list of at least three thousand validated terms from the analyzed domain and, on the other side, a reference corpus of general language of at least two million words.

3.1. Training on the lexical level

The first level of the analysis is simply the frequency count of all single word types in both corpus. Word types, in this case, are defined as case-insensitive orthographic words, i.e., units between spaces or punctuation signs. Multi-word terminological units contain, thus, *n* word types defined in this way. The result is a matrix with each word type as row and the relative frequency of the type both in the specialized terms and the reference corpus as columns. Different columns are created for the frequency of the word forms and of the lemmata.

3.2. Training on the morphological level

The training at the morphological level is very similar to the previous, the only difference being that, at the morphological level, types are defined in a different manner. While in the lexical level we count the frequency of words, in this level we count the frequency of fragments of words, defined as sequences of characters both at the beginning and at the end of each single word. The length of these segments can be considered an execution parameter. Previous experiments suggest that appropriate lengths are 3, 4 and 5 letters. The segmentation of the words is applied to both word forms and lemmata. The format of the matrices is similar to 3.1, containing the relative frequency of prefixes and suffixes of different length.

It is important to note that it is type frequency and not token frequency which is counted in this level. Type frequency is a more appropriate measure of the productivity of morphemes (Baayen & Renouf, 1996), since in this way we count the number of different word types that contain them. For instance, it would be incorrect to think that the prefix *sai-* is productive in general English just because the form *said* is very frequent.

3.3. Training on the syntactic level

The idea of having the algorithm follow a syntactic training is that it can develop a model of the syntax of the terminology of the domain. There is an interesting contrast here with respect to mainstream extractors, as commented in Section 2, since there is no explicit information about terminological syntactic patterns in English. The algorithm relies on POS-tagging, however this does not mean that one can no longer claim language independence because, as shown by Schmid (1994), POS-tagging can also be set up as a language independent problem using supervised statistical algorithms, as is the case of the present proposal. Something different is the amount of time that it takes to train a POS-tagger in a new language, but what remains true is that learning is always based on examples and not on explicit rules.

Under the assumption that one can use a POS-tagger, as is of course the case in English, the training procedure consists of tagging the example terms. The purpose of the training at the syntactic level is to elaborate a matrix similar to the previous levels, now recording the frequency of the syntactic patterns (sequences of POS tags) found in the example terms. Typically, the most frequent syntactic pattern in English terminology will be the single noun; followed by adjective+noun or noun+noun, noun+preposition+noun, and so on. Eventually, if a real user applying this methodology is for some reason interested in excluding some syntactic pattern, the list of patterns could be manually edited. Naturally, no manual editing of this type was conducted for the purpose of evaluating this algorithm in the experiment reported in Section 4. With respect to the syntactic patterns of the reference corpus, they are considered not relevant and therefore are not recorded.

For technical reasons, one must present each term as a sentence to the tagger. These are not the best conditions for the application of such devices, but the effects of tagging errors are not so damaging at this phase. One can safely assume that a majority of the cases will be tagged correctly, and since the important information are the most frequent patterns, on average the model will approximate correctly to the actual syntax of the terms.

3.4. Using the training for the analysis of new data

Once the training process has finished, the analysis phase consists of evaluating the candidate terms found in an analyzed document using the information gathered during the training phase. If a given word or phrase in the text has a syntactic pattern which is frequently used in terms, if it contains words frequently used in terms or, moreover, if its morphology is typically terminological, then the algorithm would have enough clues to promote this unit as a term candidate. The idea is that the algorithm combines and balances the information learned during the training phase to rank candidate terms from the analyzed sample.

Methodologically, the first step is to segment the analyzed sample into sequences which were recorded as frequent patterns in the syntactic training. In contrast to the training phase, errors of the tagger in this phase can cause considerable damage, because at this point the tagger acts as a filter: any unit that is not encapsulated as potentially terminological from the syntactic point of view, will now be irremediably lost. If it is indeed the case that the tagger misses a correct candidate by assigning wrong POS tags, this unit will most probably not be recovered later, unless it is wrongly assigned a POS tag which is also considered terminological, in which case the candidate would apply for the next steps (even tagged with a wrong lexical category).

A resulting matrix records the relative frequency in the analyzed sample of this first batch of candidates, along with the score provided by its syntactic pattern (defined as the relative frequency of the pattern in the training set) and the score that the unit obtains in the lexical and morphological level. The score is given in Equation 1: for each element *j*, the algorithm simply takes the observed frequency $f_o(j)$ (defined in this case as the relative frequency of element *j* in the training set of terms) divided by the expected frequency, $f_e(j)$ (defined as the relative frequency of the same element in the reference corpus) plus a positive constant *s* to avoid having an expected frequency with zero value when *j* does not occur in the reference corpus.

$$w(j) = f_o(j) / (f_e(j) + s) \qquad 1$$

As said before, a candidate *j* can be weighted in this way with respect to different features. On a first run, a candidate term will obtain different scores at the lexical levels for each of the words it contains, excluding function words. In the case of the lexical coefficient, at this point it acts as a filter, discarding any candidate below an empirically determined threshold (0.05), and this takes into account both the inflected form as well as the lemma. In the case of multi-word expressions, this evaluation is conducted with its first and last component, because a legitimate multi-word term is likely to have a high frequency word as a central component but not in the extremes, as it is the case of the component *of* in the multi-word term *circle of Willis*.

On a second run, new values are added in the same way for each of the segments of three to five letters at the beginning and at the end of each word of the candidate. The final score (*T*) of a candidate *j* is computed shown in Equation 2, where the factors are the observed frequency of the candidate in the analyzed document –$f_o(j)$–, the syntactic coefficient –syn(*j*)– , the lexical coefficient –lex(*j*)– and, finally, the morphological coefficient –morph(*j*). Any candidate resulting in a final score below a threshold (0.5) is automatically discarded but, naturally, this is an adjustable parameter.

$$T(j) = f_o(j) \,.\, \mathrm{syn}(j) \,.\, \mathrm{lex}(j) \,.\, \mathrm{morph}(j) \qquad 2$$

4. RESULTS

The first step of the experiment has been to gather a reference corpus of general English and a list of terms from the medical domain. A relatively small reference corpus of English was used, taken from Quasthoff et al. (2006). This corpus comprises mainly press articles and has en extension of around two million words. Better results are to be expected when training with larger corpora. However, for the purpose of the present experiment, it is preferable to test the performance in conditions of poor training for the sake of replicability in less documented languages. With respect to the training with examples of terms, again a small amount of training examples was used, in order to test the reliability of the algorithm in non-ideal conditions. This was a random sample of 3000 medicine terms from the Mosby dictionary (Anderson et al., 1998).

4.1. Results of the training phase

4.1.1. Results of training at the level of the lexicon

Tables 2 and 3 show, respectively, part of the result of the frequency count with the 15 more frequent word forms in the reference corpus and in the term examples (the rest cannot be shown for obvious space limitations). Function words were eliminated from Table 3 using the

Rank	Word	Frequency
1	the	118301
2	of	59408
3	to	56152
4	a	47661
5	and	43595
6	in	43184
7	for	20758
8	that	18849
9	is	17231
10	said	14940
11	on	14744
12	was	14172
13	with	11966
14	by	11812
15	from	10770

Table 2: Most frequent words in the general language reference corpus

Rank	Word	Frequency
1	syndrome	74
2	disease	54
3	reflex	32
4	system	30
5	blood	26
6	therapy	26
7	health	24
8	hydrochloride	22
9	cancer	22
10	sodium	20
11	fracture	20
12	anesthesia	20
13	pressure	19
14	carcinoma	18
15	artificial	18

Table 3: Most frequent words in a random sample of 3000 medical terms

© Servicio de Publicaciones. Universidad de Murcia. All rights reserved.

100 most frequent words obtained in Table 2 as a stoplist. Hapax legomena and dis legomena are excluded from both tables. The lexical training finishes here.

4.1.2. Results of training at the morphology level

As already mentioned in section 3.1.2, the training at the morphology level is very similar to lexical training, the only difference being that in this case we count the frequency of initial and final fragments of words of variable length, both from the reference corpus and from the term examples. Tables 4, 5 and 6 show the frequency of initial and final segments of words of different length in the reference corpus. Tables 7 and 8, in turn, show the most frequent initial and final segments in the examples of terms. Only a few examples are shown, but the calculation is also performed on initial and final segments of 3 and 4 letters.

Rank	**Segment**	**Frequency**
1	comp-	1138
2	comm-	828
3	mill-	764
4	cont-	689
5	stat-	679
6	mark-	671
7	inte-	666
8	mont-	628
9	coun-	624
10	shar-	601
11	cent-	558
12	cons-	548
13	bill-	496
14	pres-	478
15	offi-	465

Table 4: Most frequent 4 letter-long segments at the beginning of words in the reference corpus

Rank	**Segment**	**Frequency**
1	-tion	1243
2	-ting	1049
3	-ding	699
4	-ring	623
5	-ther	604

Table 5: Most frequent 4 letter-long segments ending a word in the reference corpus

Rank	**Segment**	**Frequency**
1	-ing	5636
2	-ers	1947
3	-ion	1574
4	-ted	1396
5	-ter	1121

Table 6: Most frequent 3 letter-long segments ending a word in the reference corpus

Observing tables 7 and 8, one finds striking how with only a limited number of terms it is possible to capture the morphology that is typical of the domain. It is easy to collect affixes using this technique in comparison with manual compilation. Adding to the difficulty of manually elaborating a complete list of affixes of the domain, one has to consider that the effort would have to be repeated with every new domain. Certainly, not all technical domains show an extensive use of morphology as medical terms, but at least some measurable degree of deviation can be expected. A definitive answer to this question will need further empirical research on different domains.

Rank	Segment	Frequency
1	trans-	19
2	inter-	16
3	hyper-	15
4	neuro-	13
5	micro-	13
6	psych-	13
7	lymph-	12
8	elect-	11
9	pseud-	11
10	osteo-	9
11	angio-	9
12	intra-	9
13	kerat-	8
14	laryn-	8
15	fibro-	8

Table 7: Most frequent 5 letter-long initial segments in the sample of 3000 English medical terms

Rank	Segment	Frequency
1	-ation	111
2	-ction	29
3	-tosis	24
4	-ology	20
5	-ative	18
6	-genic	15
7	-ional	14
8	-raphy	13
9	-ctomy	13
10	-ility	12
11	-lysis	12
12	-cular	11
13	-atory	11
14	-ement	11
15	-usion	10

Table 8: Most frequent 5 letter-long final segments in the sample of 3000 English medical terms

4.1.3. Results of training on the syntactic level

The final training on the syntactic level is performed, as explained in 3.1.3, using a POS-tagger, Schmid's (1994) TreeTagger in this case. What can be seen in Table 9 are the 10 most frequent syntactic patterns found in the sample of terms that formed the training set.

Rank	Syntactic Pattern	Frequency
1	NN	1122
2	JJ NN	658
3	NN NN	435
4	JJ	168
5	JJ NN NN	79
6	NN NN NN	64
7	JJ JJ NN	57
8	VV NN	42
9	NN PO NN	32
10	NP PO NN	26

Table 9: The ten most frequent syntactic patterns found in the sample of terms

The most frequent pattern, as was to be expected, is the single-noun (NN); the second is the adjective+noun pattern (JJ NN); the third is noun+noun (NN NN); and so on. Patterns with frequency below 5 were not registered since their variability increases exponentially and thus they are not useful for a model.

4.2. Results of the analysis phase

For this experiment, the sample of text to be analyzed is a semi-randomly selected specialized document, the first pdf document served by Google with a randomly selected English medical term, which happens to be a document of high degree of specialization. For convenience, only a small fragment of the document constitutes the sample to be analyzed: just the 285 word long abstract. From this fragment, and prior to its submission to the system, an observer manually extracted the totality of the single and multi-word terms that, according to terminological criterion, could be included as an entry in a medical glossary. This handmade list will be used later for the evaluation of this and other term extractors (Section 5).

The fragment of text can now be submitted to the extraction algorithm, which will produce, as a first step, the tokenization and POS-tagging of the text using TreeTagger. Table 10 shows a small fragment of the result of this first step.

TOKEN	**POS-TAG**	**LEMMA**
The	DT	the
neuroanatomic	JJ	neuroanatomic
substrate	NN	substrate
of	IN	of
cognitive	JJ	cognitive
deficits	NNS	deficit
in	IN	in
long	JJ	long
term	NN	term
survivors	NNS	survivor
of	IN	of
prematurity	NN	prematurity
with	IN	with
PVL	NP	PVL
is	VBZ	be
poorly	RB	poorly
...	...	...

Table 10: Fragment of the analyzed sample after POS-tagging with TreeTagger

Unfortunately, an important number of tagging errors are made by TreeTagger as a consequence of the high degree of specialization of the document. One has to expect a higher error rate in this kind of material because of the much higher number of Out-of-Vocabulary units (OOV's) which tend to be tagged as nouns (NN) by TreeTagger. For instance, the sequence *parieto occipital white matter*, which should be tagged as JJ+JJ+JJ+NN (three adjectives followed by a noun) are instead tagged as NN+NN+JJ+NN, presumably because the first two components of the term are not in the vocabulary of the tagger. The same occurs with other sequences, such as *pulvinar abnormalities*, tagged as NN+NN when it should be JJ+NN. Certainly, the POS-tagging of specialized corpora remains an open question, and it could represent an interesting line of research to find the correct tagging when one expects that a high proportion of the vocabulary will be unknown for the tagger. This may require a completely different tagging strategy, and the issue is too complex to be addressed in this paper. At the moment, errors of the tagger will have to be assumed until a better solution is found.

Rank	Candidate	Frequency	Syntax	Lexicon	Morphology	Total
1	thalamus	6	1.00	1.00	74.71	448.33
2	abnormality	3	1.00	1.88	71.04	400.74
3	atrophy	1	1.00	2.82	135.79	383.09
4	cerebral	1	0.15	11.73	192.10	337.41
5	parieto	2	1.00	1.00	151.27	302.55
6	efferent myelinated	1	0.59	2.82	169.86	281.09
7	parieto occipital	2	0.39	1.00	133.67	103.66
8	method	1	1.00	3.23	30.49	98.41
9	cortex	1	1.00	1.91	47.79	91.31
10	microstructural abnormality	2	0.59	1.44	50.48	85.29
11	efferent	1	0.15	4.64	98.85	68.73
12	abnormal	1	0.15	4.28	80.59	51.64
13	diffusivity	1	1.00	1.00	48.81	48.82
14	neuronal loss	1	0.59	1.05	74.62	45.99
15	lambda	1	1.00	0.99	43.61	43.27
16	neuroanatomic substrate	1	0.59	1.00	64.95	38.09
17	cognition	1	1.00	1.00	34.98	34.99
18	neuronal	1	0.15	1.91	122.11	34.94
19	microstructure	1	1.00	1.00	30.72	30.73
20	pulvinar abnormality	1	0.39	1.80	40.30	28.13

Table 11: The 20 best rated candidates extracted from a 285 word long medical abstract (all syntactic patterns together)

Once the POS-tagging of the analyzed sample is finished, the algorithm extracts sequences of units with syntactic patterns which were registered as frequent during the syntactic training phase. With this first batch of term candidates, the algorithm can produce the values of each candidate using the information gathered during the training phase. The best 20 candidates are shown in Table 11, irrespective of their syntactic pattern. Optionally, the candidates can also be presented in a breakdown of patterns, as in the examples shown in Tables 12, 13 and 14.

A close examination of Table 11 offers a first impression of the quality of the results and the different weight contributed by each factor. Almost all the 20 best rated candidates are indeed medical terms, with few exceptions. Possibly, candidates such as *parieto* (position 5), *efferent myelinated* (6) and *parieto occipital* (7) might seem to be segmentation errors, as they are parts of a term acting as modifiers instead of being well-formed phrases with a head. Adjectives and modifiers represent frequent patterns within the set of training examples of terms. The reason for including units with these syntactic patterns in a terminological dictionary is that they are relevant from a terminological perspective. These elements are recorded to account for their occurrence in different terms modifying different types of heads, which is the purpose of the breakdown in syntactic patterns that will follow, where there are lists of, for instance, adjectives, which are found as part of different terms. As mentioned in Section 3.3, if in a real application a terminologist applying this method is not interested in such patterns, the list of patterns that feeds the algorithm could eventually be edited after the training. The relation between modifiers and heads, as the number and proportion of heads per modifier or vice-versa, is still an open question and a very interesting line of research, similar to Dagan and Church (1994) or Nakagawa and Mori (2002).

A candidate like *method*, in position 8, is instead terminologically less interesting, and therefore in this experiment it is considered a false positive, even when the Mosby dictionary reserves an entry for this unit. Other units that received high ranking by the algorithm are statistical terms such as *lambda*, *p* or *positive correlation*. It may be disputable if these units can be considered medical terms, since they are more properly terms from statistics. However, they are frequently used in medical literature. Should they be included in a terminological dictionary? Despite the statistical origin of these terms, most terminologists would agree. The terms *lambda*, *p-value*, and *correlation* are also listed as entries in the Mosby dictionary.

With respect to the breakdown of the candidates in syntactic patterns, only a few examples are shown. The classification is useful for practical purposes, but for the theoretical purpose of this article, this information is of secondary importance. More important is the precision of the selection. The cases of errors that are found are at times caused by previous tagging errors, as is the case with the previously mentioned *parieto* tagged as a noun. Candidates such as *child*, *volume*, *reduction*, *effect* or *patient* are less relevant for a terminologist and in this experiment they are considered false positives as well, irrespective of the fact that, again, the Mosby dictionary also includes these five examples as entries. Other clear false positives are, for instance in the Adjective+Noun pattern, segmentation errors (*efferent myelinated*) and genuine scoring errors (*long term*, *significant reduction*, *overall volume*, *major finding*).

Rank	Candidate	Frequency	Lexicon	Morphology	Total
1	thalamus	6	1.00	74.71	448.33
2	abnormality	3	1.88	71.04	400.74
3	atrophy	1	2.82	135.79	383.09
4	pulvinar	1	1.50	45.36	10.19
5	parieto	2	1.00	151.27	302.55
6	myelinated	1	1.50	140.51	4.88
7	method	1	3.23	30.49	98.41
8	cortex	1	1.91	47.79	91.31
9	diffusivity	1	1.00	48.81	48.82
10	lambda	1	0.99	43.61	43.27
11	cognition	1	1.00	34.98	34.99
12	microstructure	1	1.00	30.72	30.73
13	child	3	1.30	7.14	27.95
14	injury	1	3.06	8.01	24.53
15	term	2	0.42	25.85	21.84

Table 12: The 15 best rated nouns

Rank	Candidate	Frequency	Lexicon	Morphology	Total
1	efferent myelinated	1	2.82	169.86	281.09
2	microstructural abnormality	2	1.44	50.48	85.29
3	pulvinar abnormality	1	1.80	40.30	28.13
4	neuronal loss	1	1.05	74.62	45.99
5	neuroanatomic substrate	1	1.00	64.95	38.09
6	secondary effect	1	2.88	8.32	14.09
7	cognitive deficit	1	0.79	21.85	10.06
8	microstructural damage	1	0.70	17.37	7.13
9	visual processing	1	0.79	14.46	6.72
10	positive correlation	1	1.19	8.85	6.17
11	preterm child	1	1.15	7.80	5.28
12	long term	1	0.28	25.12	4.12
13	extensive interconnection	1	0.85	7.95	3.95
14	significant reduction	1	0.79	6.75	3.13
15	overall volume	1	0.97	5.27	3.00

Table 13: The 15 best candidates with the pattern Adjective + Noun

Rank	Candidate	Frequency	Lexicon	Morphology	Total
1	cerebral	1	11.73	192.10	337.41
2	efferent	1	4.64	98.85	68.73
3	abnormal	1	4.28	80.59	51.64
4	neuronal	1	1.91	122.11	34.94
5	axonal	1	1.00	134.50	20.14
6	neuroanatomic	1	1.00	122.95	18.41
7	atrophic	1	1.50	64.06	2.06
8	cognitive	2	0.99	34.82	10.35
9	pulvinar	1	1.50	45.36	10.19
10	periventricular	1	1.50	37.62	1.21
11	afferent	1	1.00	52.09	7.80
12	secondary	1	4.46	9.00	6.02
13	quantitative	1	2.73	13.97	5.71
14	microstructural	1	1.00	29.93	4.48
15	visual	1	0.91	26.81	3.65

Table 14: The 15 best rated Adjectives.

5. EVALUATION AND COMPARISON WITH OTHER SYSTEMS

The evaluation of a new methodology needs more than simply to run an experiment and record precision and recall. The evaluation has to be carried out with reference to other parameters, usually a baseline algorithm representing a classic or basic solution to the same problem and, when possible, a direct comparison with the results of other algorithms on the same data set. The manual extraction of the terms from the sample conducted at the first phases of the experiment (Section 4.2) is now used as a gold-standard for the comparison with the selection of terms proposed by each algorithm. The systems listed in Table 1 of Section 2 plus a baseline algorithm that will be described below, are used to extract term candidates from the same sample of text.

The manual extraction from the sample resulted in 49 different terminological units. With this list, one can calculate precision as the proportion of candidates in the ranked list proposed by each algorithm which are also in the list of manually extracted terms; as well as recall, as the proportion of manually extracted terms found also in the ranked list proposed by each algorithm. This evaluation method can possibly be criticized for relying on the judgment of a single observer, when a specialist –or, better, a group of them– could do a more precise selection of the terminology. However, it is to be expected that a single observer will be able to identify at least the majority of the terms correctly, which should be enough to evaluate a term extractor.

The first step of the evaluation is the comparison with the baseline algorithm, which is the frequency rank of case-insensitive *n*-grams from the sample, undoubtedly the most simple method for extracting terminology from text. In this case, the baseline includes unigrams, bigrams and trigrams, and applies a stoplist consisting of the 100 most frequent words in the reference corpus. Any *n*-gram including a member of the stoplist at the initial or final position is eliminated from the ranking. Hapax legomena and *n*-grams with punctuation signs are also filtered out. In general, these kinds of strategies tend to have low recall since, following a Zipfean distribution, a large proportion of the terminological units of a corpus have low frequency.

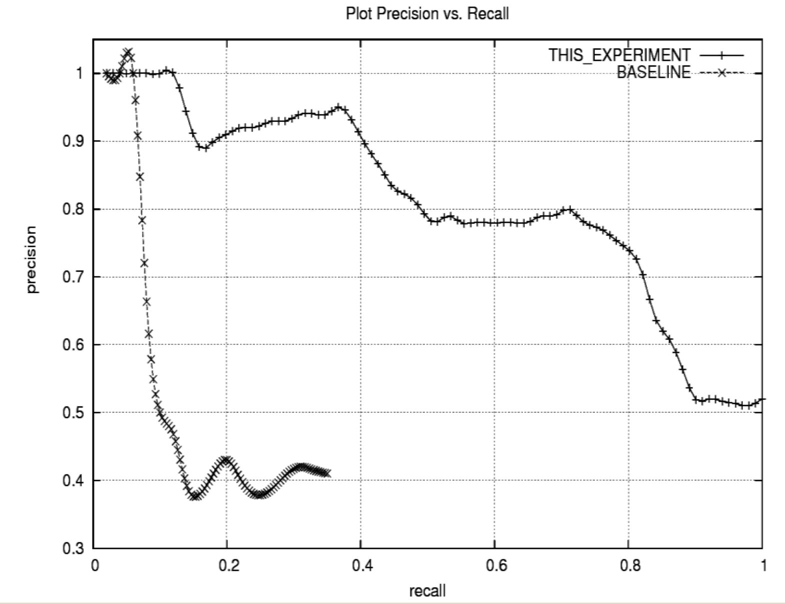

Figure 1: Precision and recall plot comparing the results of this experiment with the baseline algorithm

Using the ranking produced by the algorithm described in this paper and the one produced by the baseline algorithm, it is possible to plot Figure 1, with precision at the vertical axis and recall at the horizontal axis. Ideally, an algorithm should produce both high precision and recall, thus it can be interpreted that the surface behind the curve is a measure of the quality of the output. Normally, however, an algorithm will not produce 100% recall. Some systems are prone to offer few candidates which are most surely terminological, rather than offering many less reliable units. It is, in any case, a matter of balance between noise and silence. As already mentioned, silence or low recall in the case of the baseline algorithm is a consequence of Zipf law. Depending on the size, one would expect that between a third and close to a half of the terms in a specialized document will be hapax legomena.

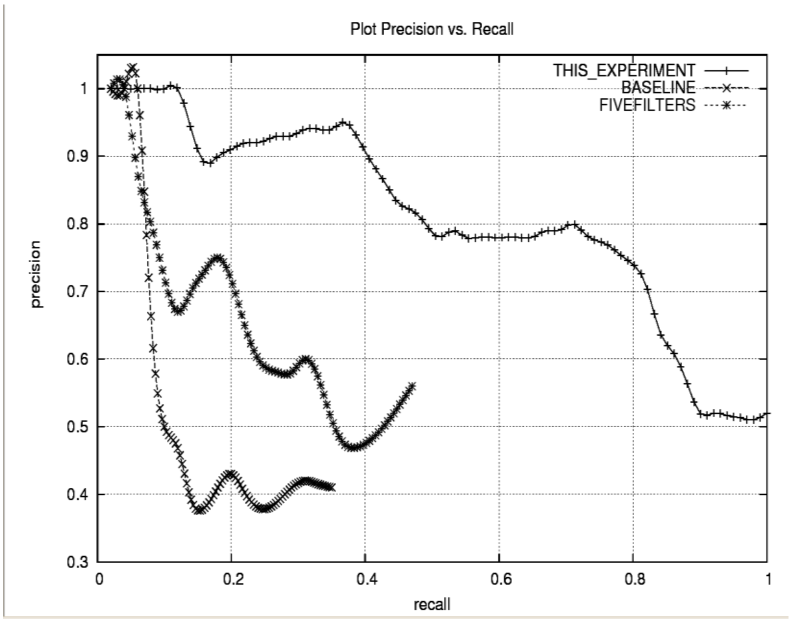

Figure 2: Precision and recall plot in comparison to FIVEFILTERS

In the case of the experiment reported in the present paper, one can see that the algorithm is able to maintain 75% precision at 80% recall, which is a good balance considering the human effort necessary to post-process the ranked lists. In comparison, FIVEFILTERS (Figure 2) performs significantly better than the baseline, but it is still closer to the baseline than to the performance achieved in this experiment. YAHOO, in turn, shows better precision than FIVEFILTERS but, comparatively, exhibits lower performance than the results reported here (Figure 3). TERMFINDER (Figure 4) is between FIVEFILTERS and YAHOO. TERMINE (Figure 5) is, from all the evaluated algorithms, the one that has the best performance in precision if we take into consideration only the segment of best rated candidates. Precision in TERMINE decreases rapidly, however, in relation to recall. Recall, in any case, is below 40%. TERMOSTAT (Figure 6) follows the baseline, which is possibly a consequence of being a statistical approach not designed for working with samples of reduced size.

Considering the balance between precision and recall, the results reported in this paper are clearly superior to those produced by the baseline and the rest of the algorithms, and this is said only taking into account the results produced by each system. Other parameters for evaluation could have also been included, such as the time and effort required to implement each solution for each language and domain, especially when they are based on linguistic and ontological knowledge. In the case of this experiment, the whole process (training + analysis) took only a few minutes. This is assuming that a POS-tagger is already trained for that language and that the rest of the training material (the list of terms and the reference corpus) is available. Fortunately, in the present days it is easy to find this material on the web.

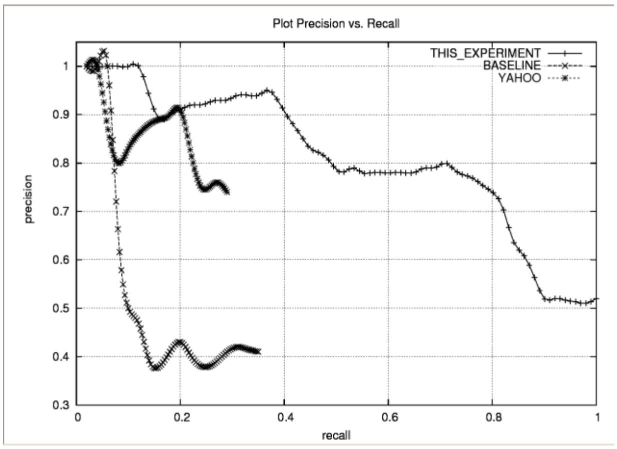

Figure 3: Precision and recall plot in comparison with YAHOO

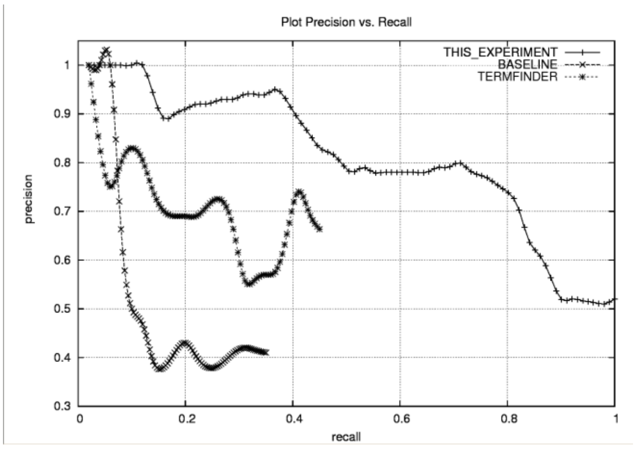

Figure 4: Precision and recall plot in comparison with TERMFINDER

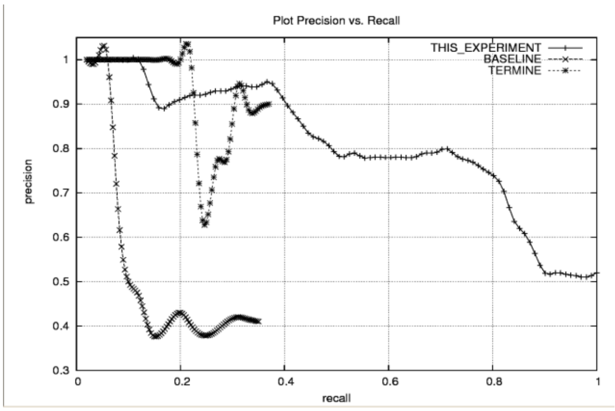

Figure 5: Precision and recall plot in comparison with TERMINE

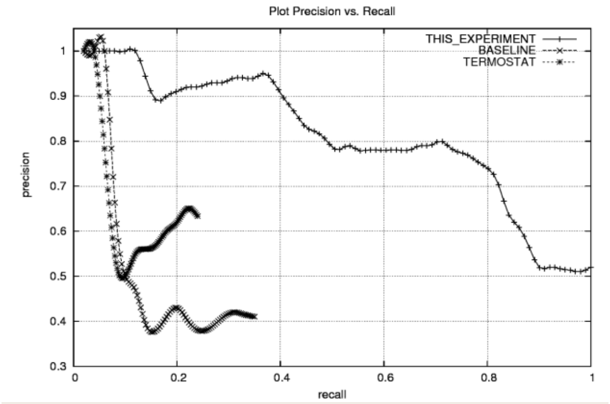

Figure 6: Precision and recall plot in comparison with TERMOSTAT

6. CONCLUSIONS AND FUTURE WORK

This paper has reported the application of a language and domain independent statistical methodology for terminology extraction in the particular case of English medical terms. Leaving aside the savings in effort of manually coding linguistic or specialized knowledge, the fact that the method does not rely on external sources has two advantages: 1) it can be used for other, less documented languages and domains; 2) the automatic recollection of the needed (implicit) knowledge avoids the risk of subjective bias that can be found in knowledge coded in the form of symbolic rules. The statistical inference performed by this algorithm seems, in contrast to rule-based algorithms, more objective and systematic. Other positive attributes of the proposal is that it is computationally efficient and inexpensive and that it is also easy to understand and implement.

A question that remains open for future versions of this algorithm is whether to develop a fully language independent POS-tagger. The current problem with using TreeTagger is that a large sample of annotated data is needed to train the program in a new language and, in practice, this is difficult to produce in a short period of time. Unsupervised methods, however, could be explored. There have been reports on solutions to the problem of inferring the lexical categories without previous knowledge (Biemann, 2010). In this scenario, there is no need to tag the words with a proper lexical category name. Instead, lexical categories are identified by arbitrary codes -i.e., numbers- as a result of a process of clustering of words based on distributional properties. If an attempt in this line proves successful, it would considerably reduce the effort of the user supervising the process.

As for future work, the most immediate task would be to improve on some technical issues of the experiments, apart from the strictly mathematical aspect. A useful implementation needs to meet diverse problems such as to develop a careful text-handling procedure, preserving the integrity of the texts after they have been converted from different file formats to plain text, guessing character encodings and identifying titles, paragraphs, proper nouns and so on.

Better evaluation figures could be obtained finding new term extractors to repeat the experiments, preferably with larger samples of text. A group of terminologists and specialists in the domain should also be part of the experiment, instead of a single observer doing the manual extraction. In this new experimental design, one can also check agreement statistics among the observers, including a group of random extractors as reference. Clustering the observers according to their selection in a dendrogram, one should see, ideally, that specialists and terminologists are clustered together with those that can be considered good term extraction systems, while those less reliable should lie apart from them, closer to the random extractors. This could be, possibly, the most objective measure for the evaluation of term extraction algorithms.

ACKNOWLEDGEMENTS

This research was funded by a contract as technician by the Spanish Ministry of Science and Innovation in the Institute for Applied Linguistics of Pompeu Fabra University (Barcelona), in the scope of Project RICOTERM-3. « Enriquecimiento de ontologías para la extracción de información de discursos de especialidad », lead by Mercè Lorente (HUM2007-65966-C02-00). The author would also like to thank the anonymous reviewers for their comments.

REFERENCES

Ananiadou, S. (1994). A Methodology for Automatic Term Recognition. *Proceedings of Coling 1994, 15th International Conference on Computational Linguistics* (pp. 1034-1038). Kyoto, Japan.

Anderson, K. N., Anderson, L. E. & Glanze, W. D. (1998). *Mosby's medical, nursing, and allied health dictionary* (5th ed.). St. Louis, MO: Mosby.

Arntz, R., Picht, H. (1995). Introducción a la terminología, Fundación Germán Sánchez Ruipérez. Madrid: Pirámide.

Baayen, H., Renouf, A. (1996) Chronicling the Times: Productive Lexical Innovations in an English Newspaper. *Language*, *72*(1), 69–96.

Barnes, J.A. & Harary, F. (1983). Graph Theory in Network Analysis. *Social Networks, 5*, 235–244.

Biemann, C. (2010). Unsupervised Part-of-Speech Tagging in the Large. Research on Language and Computation, *7*(2), 101–135.

Böhm, K., Heyer, G., Quasthoff, U. & Wolff, C. (2002). Topic Map Generation Using Text Mining. *Journal of Universal Computer Science, 8*(6), 623–633.

Bourigault, D. (1996). LEXTER, a Natural Language tool for terminology extraction. *Proceedings of Seventh EURALEX International Congress* (pp. 771-779). Göteborg, Sweden.

Bourigault, D., Jacquemin, C. & L'Homme, M-C. (Eds.) (2001). *Recent Advances in Computational Terminology*. Amsterdam: John Benjamins.

Cabré, M. T. (1999). *La Terminología: Representación y Comunicación*. Barcelona: Institut Universitari de Lingüística Aplicada.

Cabré, M. T. & Nazar, R. (forthcoming). Terminus: a Workstation for terminology and corpus management. *Proceedings of TOTH 2011 Conference*. Annecy, France.

Chatterjee, D., Sarkar, S., Mishra, A. (2010). Co-occurrence Graph Based Iterative Bilingual Lexicon Extraction From Comparable Corpora. *Proceedings of the 4th Workshop on Cross Lingual Information Access, COLING 2010 workshop* (pp. 35-42). Beijing, China.

Dagan, I., Church, K. (1994) Termight: identifying and translating technical terminology. *ANLC '94 Proceedings of the fourth conference on Applied natural language processing* (pp. 34-40). Association for Computational Linguistics.

Daille, B. (1994). Approche mixte pour l'extraction automatique de terminologie : statistiques lexicales et filtres linguistiques. Doctoral Dissertation. Université Paris 7.

Drouin, P. (2003). Term extraction using non-technical corpora as a point of leverage. *Terminology*, *9*(1), 99–117.

Enguehard, C. & Pantera, L. (1994). Automatic Natural Acquisition of a Terminology. *Journal of Quantitative Linguistics*, *2*(1), 27–32.

Felber, H. (1984). *Terminology Manual*. Paris: Unesco/Infoterm.

Frantzi, K.T. (1997). Incorporating context information for extraction of terms. *Proceedings of the Association for Computational Linguistics (ACL/EACL)* (pp. 501-503). Madrid, Spain.

Frantzi, K., Ananiadou, S. & Mima, H. (2000). Automatic recognition of multi-word terms. *International Journal of Digital Libraries*, *3*(2), 117– 132.

Gaudin, F. (1993). Pour une socioterminologie: Des problèmes pratiques aux pratiques institutionnelles. Université de Rouen.

Hisamitsu T., Niwa, Y, Nishioka, S., Sakurai, H., Imaichi, O., Iwayama, M. & Takano, A. (2000). Term Extraction Using A New Measure of Term Representativeness. *Proceedings of the Second International Conference on Language Recources and Evaluation (LREC 2000). Workshop Proceedings on: Terminology Resources and Computation* (pp. 13–20). Athens, Greece.

Jacquemin, C. (2001) *Spotting and Discovering Terms through Natural Language Processing*. MIT Press.

© Servicio de Publicaciones. Universidad de Murcia. All rights reserved.

Juilland, A., Chang-Rodríguez, E. (1964). *Frequency Dictionary of Spanish Words*. The Hague, Mouton.

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

Kageura, K. (1995) Toward the theoretical study of terms. *Terminology, 2*(2), 239–257.

Kageura, K. & Umino, B. (1998). Methods of Automatic Term Recognition. *Terminology, 3*(2), 259–289.

Kilgarriff, A., Rychly, P., Smrz, P., Tugwell, D. (2004). The Sketch Engine. *Proceedings of Euralex* (pp. 105-116). Lorient, France.

Madani, O. & Yu, J. (2010). Discovery of numerous specific topics via term co-occurrence analysis. *Proceedings of the 19th ACM international conference on Information and knowledge management (CIKM '10)* (pp. 1841-1844). ACM, New York, NY, USA.

Matsuo, Y., Sakaki, T., Uchiyama, K. & Ishizuka, M. (2006). Graph-based word clustering using a web search engine. *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP '06). Association for Computational Linguistics* (pp. 542-550). Stroudsburg, PA, USA.

Maynard, D. & Ananiadou, S. (2001). TRUCKS: a model for automatic term recognition. *Journal of Natural Language Processing*, *8*(1), 101–125.

Minoukadeh, K. (online). FIVEFILTERS.org. http://fivefilters.org/term-extraction/

Nakagawa, H., Mori, T. (2002). A Simple but Powerful Automatic Term Extraction Method. *Proceedings of the second International Workshop on Computational Terminology (COMPUTERM 02)* (pp. 29–35). Morristown, USA.

Nazar, R. (2010). *A Quantitative Approach to Concept Analysis*. Doctoral Dissertation, Universitat Pompeu Fabra.

Nazar, R. & Cabré, M.T. (2011) Un experimento de extracción de terminología utilizando algoritmos estadísticos supervisados. *Revista Debate Terminológico*, *7,* 36–55.

Pantel, P. & Lin, D. (2001). A Statistical Corpus-Based Term Extractor. *Proceedings of the 14th Biennial. Conference of the Canadian Society on Computational Studies of Intelligence* (pp. 36-46). London, UK: Springer-Verlag.

Patry, A. & Langlais, P. (2005). Corpus-Based Terminology Extraction. *Proceedings of the 7th International Conference on Terminology and Knowledge Engineering* (pp. 313-321), Copenhagen, Danemark.

Quasthoff, U., Richter, M. & Biemann, C. (2006). Corpus portal for search in monolingual corpora. *Proceedings of the LREC 2006* (pp. 1799-1802). Genoa, Italy.

Sager, J. (1990). *A Practical Course in Terminology Processing*. Amsterdam/Philadelphia: John Benjamins.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of the International Conference on New Methods in Language Processing* (pp. 44-49). Manchester, UK.

Scott, M., (1997). PC Analysis of Key Words and Key Key Words. *System*, *25*(1), 1–13.

Sparck Jones, K. (1972). A Statistical Interpretation of Term Specificity and its Application in Retrieval. *Journal of Documentation*, *28*(1), 11–21.

Streiter, O., Zielinski, D., Ties, I. & Voltmer, L. (2002) Example-based Term Extraction for Minority Languages: A case-study on Ladin. *Paper presented at Sociolinguistics and Language Planning*, Ortisei, December 12-14.

Temmerman, R. (2000). *Towards New Ways of Terminology Description: the Socio-Cognitive Approach.* Amsterdam: John Benjamins.

Véronis, J. (2004). HyperLex: Lexical Cartography for Information Retrieval. *Computer Speech & Language*, *18*(3), 223–252.

Vivaldi, J. (2001/2004). *Extracción de Candidatos a Término Mediante Combinación de Estrategias Heterogéneas*, Barcelona: IULA, Sèrie Tesis 9.

Vivaldi, J.; Rodríguez, H. (2001). Improving term extraction by combining different techniques. *Terminology*, *7*(1), 31–48.

Vivaldi, J. & Rodríguez, H. (2011). Extracting terminology from Wikipedia. *Procesamiento del lenguaje natural*, *47,* 65–73.

Watts, D. J., Strogatz, S. (1998). Collective dynamics of 'small-world' networks. *Nature, 393* (6684), 440–442.

Widdows, D. & Dorow, B. (2002). A Graph Model for Unsupervised Lexical Acquisition. *Proceedings of the 19th International Conference on Computational Linguistics* (pp. 1093-1099), Taipei.

Wüster, E. (1979). *Introduction to the General Theory of Terminology and Terminological Lexicography*. Vienna: Springer.

© Servicio de Publicaciones. Universidad de Murcia. All rights reserved.