Combining statistics on n-grams for automatic term recognition

Almudena Ballester[†], Ángel Martín Municio[*†], Fernando Pardos[†], Jordi Porta Zamorano[†], Rafael J. Ruiz Ureña[†], Fernando Sánchez León[†]

[†]Departamento de Lingüística Computacional
Real Academia Española
c/ Felipe IV, 4. 28071 Madrid Spain
{almu, amunicio, fernando, porta, rafa, fsanchez}@rae.es

[*]Real Academia de Ciencias
c/ Valverde, 22-24. 28004 Madrid Spain

Abstract

This paper presents the work-in-progress in the development of an automatic term recognition (ATR) system built around the *Corpus Científico-Técnico* (CCT). Terms are modeled using three non-correlated dimensions: unithood, domainhood and usage, applied to a set of n-grams automatically extracted from the corpus. These dimensions are combined with a supervised machine learning algorithm in order to classify n-grams as terms or non-terms. Results of both noise and silence are promising given the paucity of data employed for training. Moreover, error analysis on noise reveals that other information dimensions can be used for significantly reducing noise.

1. Introduction

This paper presents the work-in-progress in the development of an automatic term recognition (ATR) system built around the *Corpus Científico-Técnico* (CCT). The CCT gathers Spanish texts of scientific/technical domains organized according to a taxonomy of scientific disciplines and encoded in an XCES-compliant format.

The corpus is expected to contain 30 million words by the end of 2004 and it will be a source of information complementary to the glossaries and terminological dictionaries edited by the Spanish *Royal Academy of Sciences*, the funding institution. Up to date, the CCT contains 122 texts from chemistry, physics, biology, medicine, mathematics and telecommunications, amounting some 1.2 million words. The texts belong to the *Royal Academy of Sciences* or have been obtained through agreements with technical book and journal publishers.

The whole text acquisition and encoding process has been automatised (with some human intervention) and texts should exist in a previous electronic form to achieve an acquisition faster and less error-prone (avoiding transcription/OCR mistakes). Besides this automatisation process on the source texts, some of the Academy technical dictionaries (specifically, the *Diccionario Esencial de las Ciencias*, DEC) have been also processed in order to both encode the information in XML and extract the term list. However, due to the need to provide lexicographers with new term candidates rather than already known terms and also to the semantic dimension inherent to knowledge-based term recognition an empirical approach has been adopted, hence the extracted term list is only scarcely used in this paper.

2. Modeling terms

Most definitions of *term* lay mainly on semantics and become non-operational for computing without lexical and domain knowledge. This took us to an empirical approach where n-grams are extracted from the corpus and characterized along three different dimensions supposedly relevant for term identification. The measurements performed in each dimension have been borrowed from the fields of lexicography and information extraction (IE).

2.1. Unithood

Unithood is the degree of lexical cohesive force that is shown by the elements of an n-gram. Both complex terms and collocations from a scientific/technical corpus have similar cohesion values, so a considerable overlap may occur. Unithood is approximated by measuring the degree of association among the words contained in the n-gram[1]. Mutual information has been reported as a useful statistic for extracting collocations despite its stability problems with low frequency data (Church and Hanks, 1991).

Generalized Mutual Information (GMI), which deals with arbitrary n, has been formulated in different ways (see for instance Chien and Chun-Liang (2001)). For $n > 2$, we have adopted that of Yamamoto and Church (2001). The formulation of GMI for a given n-gram t we used in this paper is:

$$GMI(t) = \begin{cases} \text{NA}, & n = 1 \\ \dfrac{N \cdot tf(ab)}{tf(a) \cdot tf(b)}, & n = 2, t = ab \\ \dfrac{tf(a\gamma b) \cdot tf(\gamma)}{tf(a\gamma) \cdot tf(\gamma b)}, & n > 2, t = a\gamma b \end{cases}$$

where *tf* is the IE *term frequency* of an n-gram and N is the corpus size.

2.2. Domainhood (distribution within the corpus)

Terms are characteristic of a domain. The distribution of most, if not all, terms in the CCT should be far from uniform due to its balanced design. A useful measure from IE

[1]When $n = 1$, and therefore it corresponds to a single word, this measure is not applicable.

that identifies good keywords from documents for retrieval is the *Inverse Document Frequency* (IDF) (Spärk Jones, 1973). A measure based on IDF and tf, the *Residual Inverse Document Frequency* (RIDF), is proposed by Church and Gale (1999) as a better measure to extract keywords. It selects those whose distribution is different from what is expected assuming a Poisson. The formulation of RIDF is:

$$\begin{aligned} RIDF(t) &= IDF(t) - \widehat{IDF}(t) \\ &= -\log_2 \frac{df(t)}{D} + \log_2(1 - e^{-\frac{tf(t)}{D}}) \end{aligned}$$

where df is the IE *document frequency* (calculated as the number of texts where an n-gram occurs) and D is the number of texts in the corpus.

2.3. General vs. specific usage

The frequency of a given term should be higher in a specialized corpus than in a non-specialized one. A subcorpus of the CREA[2] was created to represent a non-scientific/non-technical genre. In order to get maximum variability within the genre, the 250 smallest literary book texts in the CREA were selected obtaining a 17 million-word corpus. All n-grams (ranging n from 1 to 5) were generated and the tf computed with the algorithm described in §3.

This subcorpus acts as an *exclusion filter* or *blank reference* for determining usage. We use the *Relative Frequency Ratio* (r) between the CCT and this subcorpus of the CREA to compare n-gram usage[3]. This ratio is calculated by:

$$r(t) = \frac{f_{\text{CREA}}(t)}{f_{\text{CCT}}(t)}$$

In Fig. 1 it is shown a sample of 400 manually classified n-grams (see §5.1.) plotted according to the relative frequency values in the CCT and in the subcorpus of the CREA. The sample is composed of 200 terms (painted black) and 200 non-terms (painted white). A diagonal line divides the plane in two. The upper region contains n-grams whose relative frequency in the CREA is greater than in the CCT whereas the lower region contains those n-grams more frequent in the CCT. As expected, terms are located below the diagonal line.

3. N-gram statistics computation

Yamamoto and Church (2001) describe a fast algorithm to compute tf and df for all the substrings of a corpus. This algorithm makes use of suffix arrays and some properties in order to group substrings (i.e. variable length n-grams) into equivalence classes with the same tf and df. This partition of substrings leads to a drastic reduction of the number of elements and allows, with the introduction of binary search, the computation of other statistics based on tf, such as GMI.

In order to restrict the amount of n-grams extracted with the algorithm, only those with frequency over 3 and n ranging from 1 to 5 are produced. The decision to limit n to 5

[2]The CREA is a reference corpus of current Spanish containing some 130 million words and resembling the BNC in its balanced design (Martín Municio et al., 2000).

[3]Frequencies for n-grams not represented in the CREA are smoothed by adding one.

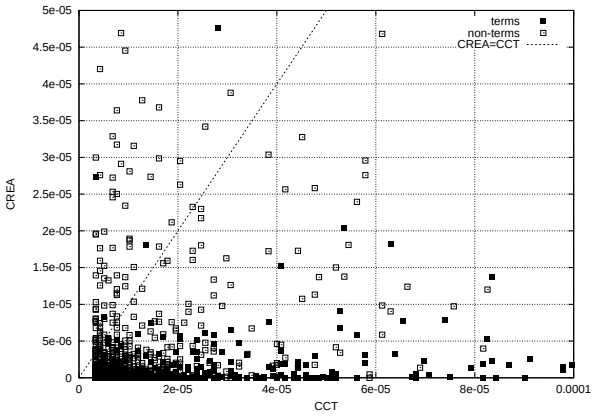

Figure 1: Relative Frequency Ratio

was motivated by the fact that the number of entries with same length listed in the DEC suffers an exponential decay as length increases (see Table 1) and 99.6% of the entries have length below 6.

Len.	# Entr.
1	10,508
2	5,357
3	2,154
4	570
5	202
6	43
7	20
8	4
9	1
10	0
11	1

Table 1: Distribution of entry lengths in the DEC

4. Distribution of terms

To test how well these statistics distinguish terms and non-terms and how they are distributed along the dimensions chosen, some scatter plots have been drawn with a random sample of 200 n-grams. In the case of plots involving GMI, the sample does not include unigrams. All plots focus on the most populated regions (sometimes excluding also outliers).

As noted by Yamamoto and Church (2001) and shown in Fig. 2, GMI and RIDF do not exhibit apparent correlation. Multiword terms are concentrated in the upper half of the plot giving GMI more discriminative power than RIDF. It can be noted that terms and non-terms are not perfectly split. This situation gets worse when more n-grams are taken into account.

Roughly the same can be said of Fig. 3, except that IDF concentrates slightly more terms on higher values.

It can be observed in Fig. 4 that n-grams with r far from zero and high GMI are terms. GMI can be low for those n-grams near zero frequency in the CREA.

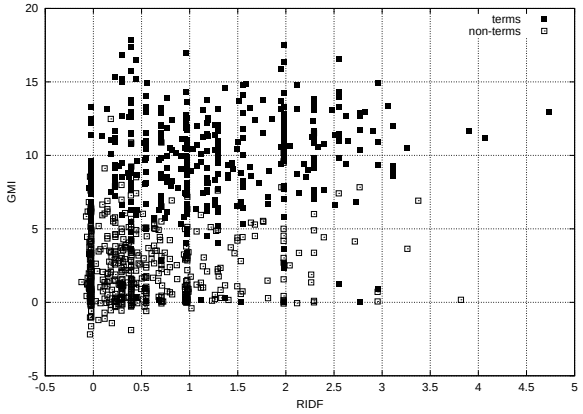

Figure 2: GMI vs. RIDF

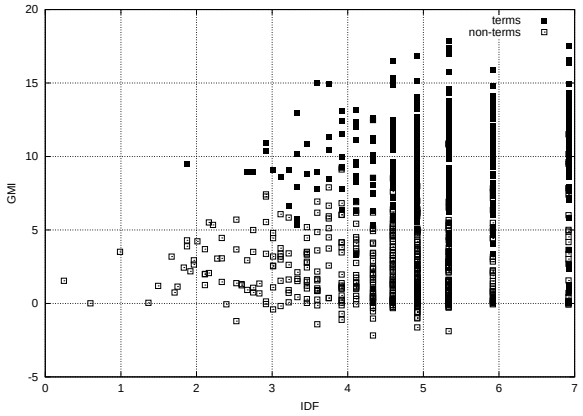

Figure 3: GMI vs. IDF

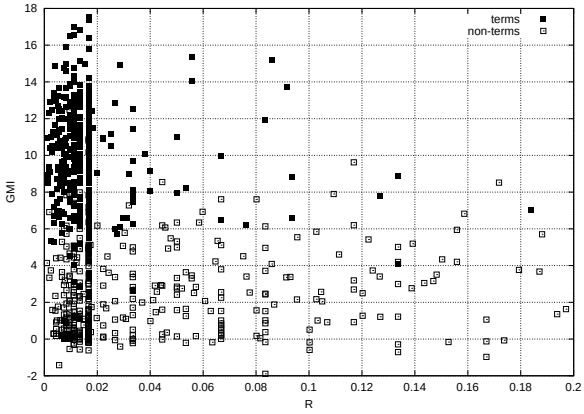

Figure 4: GMI vs. r

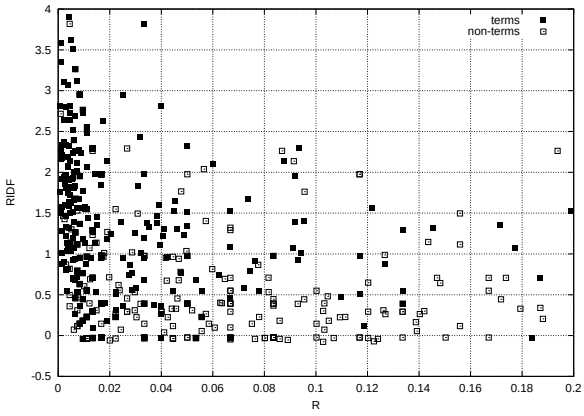

Figure 5: RIDF vs. r

Distribution of n-grams in Figs. 5 and 6 are not so clear as in previous cases. These plots represent the most confused part of the whole picture. It is doubtful whether these two statistics neatly distribute terms and non-terms but their plotting serves to show that no evident correlation seems to exist for these pairs of dimensions.

5. Learning to identify term candidates

A supervised machine learning algorithm is used for the task of classifying n-grams into terms and non-terms combining their statistical measures. A decision tree is induced using *C4.5* (Quinlan, 1993) from a classified list of n-grams. A recent similar approach is explained by Vivaldi et al. (2001), who propose a combination of different classifiers using *boosting*.

An initial training-set was obtained by random sampling the entire list of n-grams. Preliminary experiments showed that induced trees better classified non-terms because of the unequal distribution of classes (most non-terms and a few terms). Precision on the training-set for terms was only of 50% vs. 90% of non-terms that were correctly classified. Thus, more terms were manually extracted from texts and added (up to 1,740) to the training-set in order to overcome this distribution problem.

Evaluation was carried out on a test-set of 14,777 examples (1,801 terms) not previously seen by the algorithm during the training phase. Experiments were ran 100 times to obtain average figures. Table 2 displays the confusion matrix delivered by *C4.5* where predicted terms and non-terms are represented in columns $\widehat{+}$ and $\widehat{-}$, respectively. As usual in evaluating ATR systems, instead of precision ($\frac{a}{a+b}$)/recall ($\frac{a}{a+c}$) (common in IR), we used the complementary measures *silence* and *noise*. Silence is a measure of true terms not detected by the system ($\frac{c}{a+c}$), whereas noise is the measure of false terms proposed as term candidates ($\frac{b}{a+b}$).

Parallel experimentation for two systems has been carried out —one using GMI, RIDF, r and n, and another using IDF instead of RIDF. Noise and silence, despite of their variability, have a downward trend as the training-set grows (see Figs. 7 and 8). Slightly better results for noise are achieved with IDF while silence gives worse results with RIDF.

The system was finally trained considering GMI, RIDF, r, tf and n with all terms (3,541) and 29,000 non-terms. With this training-set distribution, both noise and silence levels, measured on the same training-set plus the rest of non-terms, were similar: noise = 31.54% and silence = 32.54%. The analysis of these errors can be found in §6.

Even though *C4.5* outputs a CF (certainty coefficient) associated to the class decision, thresholds on CF had not been used because the training-set has not been considered

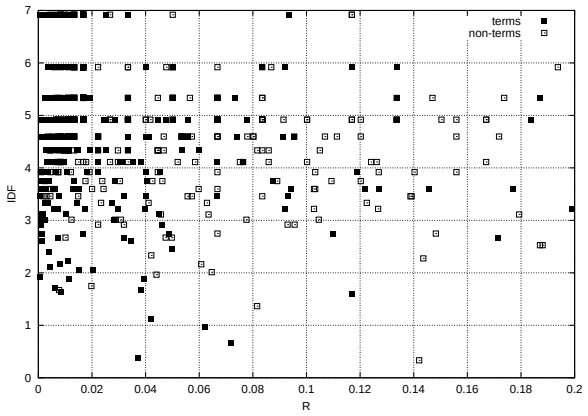

Figure 6: IDF vs. r

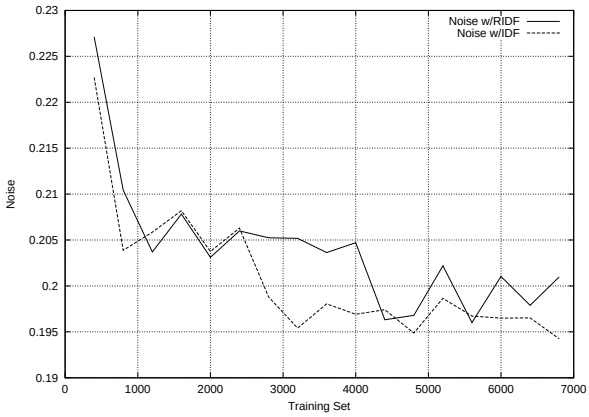

Figure 7: Noise

	$\widehat{+}$	$\widehat{-}$
$+$	a	b
$-$	c	d

Table 2: Confusion matrix

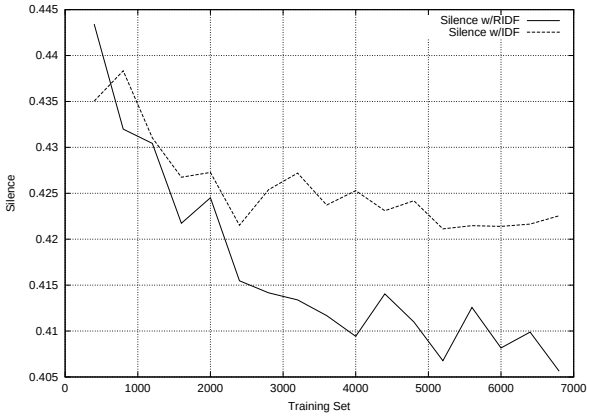

Figure 8: Silence

to have the necessary size for reliable CF calculation. In the framework of ATR, these CFs could be interpreted as a *termhood* index that can be used to rank term candidates to be presented to the lexicographers.

5.1. Manually selected candidates vs. dictionary extracted terms

Terms automatically extracted from the DEC (an upper-intermediate level technical dictionary) provide the algorithm with only positive examples. The training set was obtained by random sampling of the list of n-grams and then manual classification was carried out by linguists with no special background in science or terminology and no instruction on what a term is (only very polysemous n-grams, usually corresponding to single words, were said to be classified as non-terms).

Surprisingly, scatter plots of manually selected candidates vs. dictionary extracted ones (Fig. 9) exhibit what we interpret as a strong correlation. Termhood judgments by linguists tend to choose n-grams with measures similar to those extracted from the DEC and found in the corpus.

6. Analysis of errors

We have concentrated in the analysis of noise, since it is easier to build a set of trivial filters to reduce it, rather in the reasons for silence. The number of false terms is 1,169, being most of them unigrams (790). Given that the most discriminative dimension (GMI) is missing for unigrams, filters have been tested on the remaining errors (32.4% of noise errors).

The first set of filters exclude term candidates including punctuation (79 are excluded, 20.8% of the remaining noise errors), numbers (72, 19% of the errors), a given stop word at the beginning or end of the candidate (42, 11.1%)[4] and a given English stop word at the beginning or end of the candidate (14, 3.7%). The reduction is significant since it falls to 201 of the remaining 379 noise errors.

The second set of filters is based on other linguistic requirements expressed as negative constraints. We could have forced the candidates to conform to a given category sequence, as many authors propose (Cabré et al., 2001), but then many potentially terminological chunks would have been lost. Thus, the constraints are conservative and are based on observations over term candidates as well as over manually selected terms. All of them try to sharpen the fuzzy border between collocation and term. The first has to do with coordination. Frequently coordinated (IE) terms show a strong lexical relationship, a cohesive force that may better match the notion of collocation than that of termhood, as seen in the variety of relationships, that includes words in ordered series (`XVI y XVII`, `B y C`), unordered sets (`forma y función`, `masculino y femenino`) or co-hyponyms (`lagartos y serpientes`, `ovejas y cabras`). None of them are terms but, in some cases, coordinated terms. This is also true for most the manually selected terms including a coordinated conjunction. This filter amounts for a 9% reduction

[4]This filter is built around a couple of dozens of grammatical words and it is based on the observation that any term candidate must be a full (non-determinised) constituent, thus prepositions and determiners, for instance, can neither be the first nor the last element of a candidate.

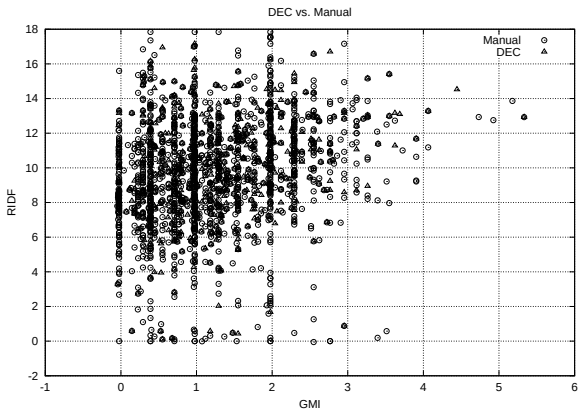

Figure 9: DEC extracted terms vs. Manually selected candidates

(18 excluded candidates) of the new 201 error set.

The error list also contains complex verbs, periphrases and control structures that can be easily excluded with a chunker (or even using an exclusion filter with a not so literary bias, since the structures are mainly informal). With this filter, 13 (6.5%) candidates are excluded.

Careful inspection of the rest of noise errors unveils some errors in manual selection. This affects to 29 cases (14.4%), that where considered non-terms by a linguist. This fact highlights the domain dependency of termhood. As Nunan (1993, 30) puts it: "Collocational patterns will only be perceived by someone who knows something about the subject at hand."

Overall noise for n-grams (where $n >= 2$) has been reduced to 37.2% with these simple filters. Moreover, the rest of false terms include frequent technical collocations (`uso masivo`, `prestigiosa revista`, `trabajos pioneros`, `niveles elevados`, `alta radiactividad`) and common head-internal argument sequences (`tiro un dado`, `valió el premio`, `expresan telomerasa`, `medir distancias`, `sufren metamorfosis`).

7. Conclusions and future work

This paper presents a language- and domain-independent methodology for ATR based on the combination of statistical measures on n-grams.

Results of both noise and silence are promising given the paucity of data employed for training. The explored dimensions can be tuned substituting current approximations for unithood by other proven useful statistics like log-likelihood ratio or MI^3 (after generalising their formulae to arbitrary length n-grams) and for domainhood by other statistics of dispersion.

Moreover, error analysis on noise reveals that a set of trivial filters significantly reduces noise, thus opening the possibility for new dimensions to be taken into account.

8. References

Mª. Teresa Cabré, Rosa Estopá, and Jordi Vivaldi. 2001. Automatic term detection: A review of current systems. In *Recent Advances in Computational Terminology*, volume 2 of *Natural Language Processing*, pages 53–87. John Benjamins.

Lee-Feng Chien and Chen Chun-Liang. 2001. Incremental extraction of domain-specific terms from online text resources. In *Recent Advances in Computational Terminology*, volume 2 of *Natural Language Processing*, page 89. John Benjamins.

Kenneth W. Church and William Gale. 1999. Inverse Document Frecuency (IDF): A Measure of Deviation from Poisson. In *Natural Language Processing Using Very Large Corpora*, volume 11 of *Text, Speech and Language Technology*, pages 283–295. Kluwer Academic Publishers.

Kenneth W. Church and Patrick Hanks. 1991. Mutual Information and Lexicography. *Computational Linguistics*, 16(1):22–29.

Ángel Martín Municio, Guillermo Rojo, Fernando Sánchez León, and Octavio Pinillos. 2000. Language Resources Development at the Spanish Royal Academy. In *Proceedings of the 2^{nd} International Conference on Language Resources and Evaluation (LREC 2000)*, volume II, pages 1265–1270, Athens, Greece.

David Nunan. 1993. *Introducing discourse analysis.* Penguin.

J. R. Quinlan. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA.

Karen Spärk Jones. 1973. A Statistical Interpretation of Term Specificity and its Application in Retrieval. *Journal of Documentation*, 28(1):11–21.

Jordi Vivaldi, Lluis Màrquez, and Horacio Rodríguez. 2001. Improving Term Extraction by System Combination using Boosting. In *Proceedings of the joint ECML-PKDD'01 Conference*, Freiburg, Germany.

Mikio Yamamoto and Kenneth W. Church. 2001. Using Suffix Arrays to Compute Term Frequency and Document Frequency for All Substrings in a Corpus. *Computational Linguistics*, 27(1):1–30, March.