AcroDef: A Quality Measure for Discriminating Expansions of Ambiguous Acronyms

Mathieu Roche, Violaine Prince

AcroDef: A Quality Measure for Discriminating Expansions of Ambiguous Acronyms

Mathieu Roche and Violaine Prince

LIRMM - UMR 5506, CNRS, Univ. Montpellier 2,
34392 Montpellier Cedex 5 - France

Abstract. This paper presents a set of quality measures to determine the choice of the best expansion for an acronym not defined in the Web page. The method uses statistics computed on Web pages to determine the appropriate expansion. Measures are context-based and rely on the assumption that the most frequent words in the page are related semantically or lexically to the acronym expansion.

1 Introduction

Named Entities Recognition (NER) has become one of the major issues in Information Retrieval (IR), knowledge extraction from texts, classification, question answering (QA), and machine aided translation (MT). The state-of-the art literature in NER mostly focuses on proper names, temporal information, specific expressions in some technical or scientific fields for domain ontologies building, and so forth. A lot of work has been done on the subject, among which on acronyms, seen as particular named entities. Acronyms are very widely used in every type of text, and therefore have to be considered as a research issue as linguistic objects and as named entities.

An **acronym** is composed from the first letters of a set of words, written in uppercase style. This set of words is generally frequently addressed, which explains the need for a shortcut. It is also a specific multiword expression, such as "named entities recognition", abbreviated into NER, sometimes completely domain dependent (as NER or NLP are) and sometimes becoming a commonly used item (such as SARS, AIDS, USA, etc.). In some cases, acronyms become proper names referring to countries or companies (like USA or IBM). However, most of the time, acronyms are domain or period dependent. They are contracted forms of multiword expressions where words might belong to the common language. As contracted forms, they might be highly ambiguous since they are created out of words first letters. For instance NER, the acronym we use for `Named Entities Recognition` might also represent `Nippon Electrical Resources` or `Natural Environment Restoration`. Those are two other possible expansions for the acronym NER. An **expansion** is the set of words that defines the acronym. The word **definition** will also be used as a synonym for expansion in this context.

In all cases, an acronym behaves like a named entity. However, the intrinsic ambiguity in most acronyms enhances the difficulty of finding which exact entity

B. Kokinov et al. (Eds.): CONTEXT 2007, LNAI 4635, pp. 411–424, 2007.

is referred by this artificial name. Literature has been addressing acronym building and expansion (see section "state-of-the art") when the acronym definition is given in the text. However, choosing the right expansion for a given acronym in a given document, if no previous definition has been provided in the text, is an issue definitely belonging to NER, and not yet exhaustively tackled. The difficulty in acronym disambiguation is to automatically choose, as an expansion, the most appropriate set of words. This article tries to deal with this issue by offering a **quality measure** for each candidate expansion. In this context, let us name a a given acronym. For every a which expansion is lacking in a document d, we consider a list of n possible expansions for a: $a^1 ... a^n$. For instance, if IR is the acronym at stake, we could have IR^1= `Information Retrieval`, and IR^2 =`Investor Relations` (in finance and communication), and IR^3 =`Infra Red` (in optics and medicine). In a multilingual context, things could become worse, IR^4 = `Impôt sur le Revenu` (the French expression for income tax). Some web resources exist for providing acronym definitions (as an example, we use the site http://www.sigles.net/, which browses more than 17,000 sites in 212 countries).

The aim of our approach is to determine k ($k \in [1, n]$) such that a^k is the relevant expansion of a in the document d. To make such a choice, we provide a quality measure, called $AcroDef$, which relies on Web resources. The figure 1 summarizes the applied global process. The presentation is structured as following: section 2 discusses the output of the related literature, section 3 focuses on the quality measure $AcroDef$, where context and web resources are essential characteristics to be taken into account. Section 4 describes some experiments and discusses their results and finally conclusion and perspectives are suggested in 5.

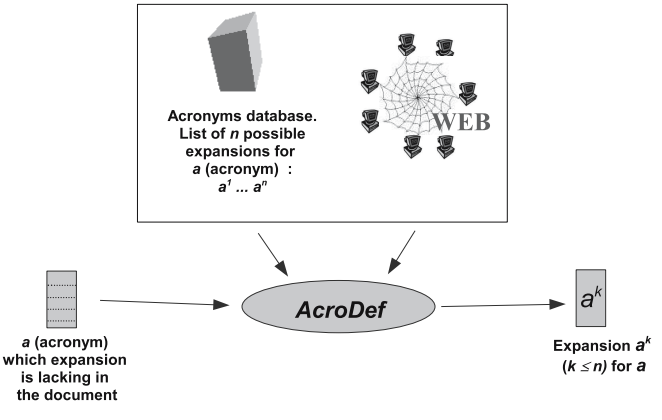

Fig. 1. Global process

2 Acronym Expansion Relevant Literature

Among the several existing methods for acronyms detection and expansion in literature, we present here some significant works. First, acronyms detection

within texts is an issue by itself. It involves recognizing a character chain as an acronym and not as an unknown or misspelled word. Most acronyms detecting methods rely on using specific linguistic markers.

Yates' method [19] involves the following steps. First, the sentences are separated by segments using specific markers (brackets, points) as frontiers. The second step compares each word of each segment with the preceding and following segments. Then the couples acronyms/expansions are tested. The candidates acronym/definitions are accepted if the acronym characters correspond to the first letters of the potential definitions words. For example, the pair "IR/Information Retrieval" is a good acronym/expansion candidate. The last step uses specific heuristics to select the relevant candidates. For example, these heuristics rely on the fact that: (1) acronyms length is smaller than their expansion length, (2) they appear in upper case, (3) long expansions of acronyms tend to use "tool-words" such as determiners, prepositions, and so forth.

Other works [4,11] use similar methods, based on the presence of markers associated to specific and linguistically oriented heuristics. Larkey *et al.*'s method [11] uses a search engine to enhance an initial corpus of Web pages useful for acronym detection. To do so, starting from a list of given acronyms, queries are built and submitted to the AltaVista search engine[1]. Queries results are Web pages which URLs are explored, and eventually added to the corpus.

Our method shares with [11] the usage of the Web. However, we do not look for existing expansions in text since we try to determine possible expansion that would be lacking in the text where the acronym is detected. From that point of view, we are closer to works like Turney's [17], which are not specifically about acronyms but which use the Web to define a ranking function. The algorithm PMI-IR (Pointwise Mutual Information and Information Retrieval) described in [17] queries the Web via the AltaVista search engine to determine appropriate synonyms to a given query. For a given word, noted *word*, PMI-IR chooses a synonym among a given list. These selected terms, noted $choice_i$, $i \in [1, n]$, correspond to the TOEFL questions. The aim is to compute the $choice_i$ synonym that gives the better score. To obtain scores, PMI-IR uses several measures based on the proportion of documents where both terms are present. Turney's formula is given below (1): it is one of the basic measures used in [17]. It is inspired from Mutual Information described in section 3.1.

$$score(\ choice_i\ ) = \frac{nb(\ word\ NEAR\ choice_i\ )}{nb(\ choice_i\ )} \tag{1}$$

- $nb(x)$ computes the number of documents containing the word x,
- $NEAR$ (used in the "advanced research" field of AltaVista) is an operator that precises if two words are present in a 10 words wide window.

With this formula (1), the proportion of documents containing both *word* and $choice_i$ (within a 10 words window) is calculated, and compared with the number of documents containing the word $choice_i$. The higher this proportion is,

[1] http://www.altavista.com/

the more *word* and $choice_i$ are seen as synonyms. More sophisticated formulas have also been applied: they take into account the existence of negation in the 10 words windows. For instance, the words "`big`" and "`small`" are not synonyms if, in a given window, a negation associated to one of these two words has been detected.

To enhance relevance to the document, our approach tries to take into account the dependencies between the words composing the possible expansions in order to rank them. In that sense, we are close to Daille's approach [7,8]. Also, as defended in next section, we use other quality measures and attempt to relate as much as possible to the context, in order to significatively enhance basic measures.

3 Defining the *AcroDef* Measure

To determine the expansion of an acronym starting from a list of co-occurrences of set of words, our aim is to provide a relevance ranking of this set using statistical measures. The most appropriate definition has to be placed at the top of the list. Therefore an overview of some existing measures is necessary to understand our choice.

3.1 Statistical Measures

Several quality measures in the literature are based on ranking function. They are brought out of various fields: Association rules detection [1,10], terminology extraction [8,13], and so forth. The following are the most widely used.

Mutual Information. One of the most commonly used measures to compute a sort of relationship between the words composing what is called a **co-occurrence** is Church's Mutual Information (MI). The formula is the following [6]:

$$I(x,y) = \log_2 \frac{P(x,y)}{P(x)P(y)} \tag{2}$$

Such a measure tends to extract rare and specific co-occurrences according to [8,13,16]. Let us notice that in this formula (2), the use of the log_2 function is not mandatory, since the latter is strictly growing. Thus, the order of the co-occurrences provided by the measure is not impacted by the application of function log_2. In the case of acronyms expansion, $P(x,y)$ measures the probability of finding couples of words (x,y) where x and y are neighbors, and in this order. For instance, with the acronym IR, x might represent the word "Information" and y the word "Retrieval". It might also be a pair such as "Investor" and "Relations". When simplified, the formula (2) could be written as follows, where nb designates the number of occurrences of words and couples of words:

$$IM(x,y) = \log_2 \frac{nb(x,y)}{nb(x)nb(y)} \tag{3}$$

This measure might be adapted to ternary co-occurrence in the way described by Jacquemin [9]. So, a natural extension of this measure would be applied to acronyms expansions that are composed of n words (formula (4)).

$$IM(x_1, ..., x_n) = \log_2 \frac{nb(x_1, ..., x_n)}{nb(x_1) \times ... \times nb(x_n)} \tag{4}$$

Cubic Mutual Information. The Cubic Mutual Information is an empirical measure based on MI, that enhances the impact of frequent co-occurrences, something which is absent in the original MI [7]. Such as measure is defined by the following formula (5). Vivaldi *et al.* have estimated that the Cubic MI was the best behaving measure [18].

$$IM3(x, y) = \log_2 \frac{nb(x, y)^3}{nb(x)nb(y)} \tag{5}$$

This measure is used in several works related to noun or verb terms extraction in texts. As for MI, the measure could be extended as follows:

$$IM3(x_1, ..., x_n) = \log_2 \frac{nb(x_1, ..., x_n)^3}{nb(x_1) \times ... \times nb(x_n)} \tag{6}$$

Dice's Coefficient. An interesting quality measure is Dice's coefficient [15]. It is defined by the following formula (7).

$$D(x, y) = \frac{2 \times P(x, y)}{P(x) + P(y)} \tag{7}$$

Similarly to the Cubic MI, Dice's coefficient weakens the impact of rare and often irrelevant co-occurrences [14]. Formula (7) leads directly to formula (8).[2]

$$Dice(x, y) = \frac{2 \times nb(x, y)}{nb(x) + nb(y)} \tag{8}$$

In Petrovic *et al.*'s article [12], the authors present an extension of the original Dice formula to three elements:

$$Dice(x, y, z) = \frac{3 \times nb(x, y, z)}{nb(x) + nb(y) + nb(z)} \tag{9}$$

In a natural way, we could extend the preceding formula to n elements as follows:

$$Dice(x_1, ..., x_n) = \frac{n \times nb(x_1, ..., x_n)}{nb(x_1) + ... + nb(x_n)} \tag{10}$$

We call it **the n extended Dice's formula**. The three measures presented before, MI, Cubic MI and Dice's Coefficient, are important for our measure *AcroDef* characterization. The two following subsections (3.2 and 3.3) describe *AcroDef* in its both variants: The basic measure and the contextual one. *AcroDef* uses Dice's coefficient. Subsection 3.4 shows another variant of *AcroDef* that involves the two other measures MI and Cubic MI.

[2] By writing $P(x) = \frac{nb(x)}{nb\_total}$, $P(y) = \frac{nb(y)}{nb\_total}$, $P(x, y) = \frac{nb(x,y)}{nb\_total}$.

3.2 Basic *AcroDef* Measure

Since our work, like many others, relies on Web resources, the *nb* function used in the preceding measures represents the number of pages provided by the search engine Exalead (http://www.exalead.fr/). The choice of Exalead has been determined by the fact that our test corpus, as explained in section 4 is built out of the Google search engine resulting pages (http://www.google.com/). It was important not to introduce a bias due to a particular engine.

Starting from the n extended Dice's formula (10), and using statistics provided by search engines we propose the basic *AcroDef* measure (formula (11)).

$$BasicAcroDef_{Dice}(a^j) = \frac{\left|\{a_i^j | a_i^j \notin M_{tools}\}_{i\in[1,n]}\right| \times nb(\bigcap_{i=1}^{n} a_i^j)}{\sum_{i=1}^{n} nb(a_i^j | a_i^j \notin M_{tools})} \text{ where } n \geq 2 \quad (11)$$

- $\bigcap_{i=1}^{n} a_i^j$ represents the set of words a_i^j ($i \in [1, n]$) seen as a string (using *brackets* with Exalead and illustrated as follows: "$a_1^j ... a_n^j$").
- M_{tools} is a set of tool-words (prepositions, determiners, etc.). The idea is to detect the pages containing these words as such, since they are not semantically discriminant.
- $|.|$ represents the number of words of the set.

Since we ran most of our experiments in French, we used the acronym "JO" as a basic example. With $a =$`JO`, two definitions are available on http://www.sigles.net/:

a^1: `Jeux Olympiques` (Olympic Games) and a^2: `Journal Officiel` (Official Journal)

Let us precise that the resulting pages numbers with both definitions are:

- $a_1^1 \cap a_2^1 =$ `Jeux` $\cap$ `Olympiques`: 366,508 resulting pages
- $a_1^2 \cap a_2^2 =$ `Journal` $\cap$ `Officiel`: 603,036 resulting pages

As a matter of fact, the IR acronym has given the following results on the same site:

1. IR^1: `Initiative Républicaine` (Republican Initiative). Domains: *Politics*, *society*. Language: French.
2. IR^2: `Impôt sur le Revenu` (Income Tax). Domains: *Finance*, *tax*. Language: French.
3. IR^3: `Infrarouge` (Infrared). Domains: *Research*, *sciences*. Language: French.
4. IR^4: `Insuffisance Rénale` (Renal Insufficiency). Domains: *Health*, *sciences*. Language: French.
5. IR^5: `Investor Relations`. Domains: *Communication*, *finance*. Language: English.
6. all other listed elements contain IR as a subchain either in the acronym or in its expansion.

Let us note that `Information Retrieval` does not appear on this Acronym Dictionary Portal as a well known expansion for IR. This means that domain dependent acronyms really need to be associated to an ontological choice, something that is discussed in the perspectives section of this paper.

Back to our example with "JO", the obtained values with the *BasicAcroDef* formula (11) are very close.[3]

- $BasicAcroDef_{Dice}(\texttt{JO}^1) = \frac{2 \times nb(\texttt{Jeux} \cap \texttt{Olympiques})}{nb(\texttt{Jeux}) + nb(\texttt{Olympiques})} = \frac{2 \times 366508}{116929964 + 1207545} = 0.0062$
- $BasicAcroDef_{Dice}(\texttt{JO}^2) = \frac{2 \times nb(\texttt{Journal} \cap \texttt{Officiel})}{nb(\texttt{Journal}) + nb(\texttt{Officiel})} = \frac{2 \times 603036}{178302348 + 28140994} = 0.0058$

Practically this comes back to submitting the three following queries to Exalead: `"Jeux Olympiques"` (`Jeux` $\cap$ `Olympiques`), `Jeux` and `Olympiques`. Let us note that more pages result from the query `"Journal Officiel"`, whereas the highest score is obtained with the expansion `"Jeux Olympiques"`.

In languages like French, many noun phrases contain tool words such as determiners or prepositions, and thus, several acronym expansions will be composed of such elements. So, when the definition of an acronym contains a tool word, it is neglected in the formula denominator.

This basic formula does not take the context into account. This is a severe limitation. Therefore, next subsection details a measure that relies on context to define a more relevant expansion choice for a given acronym.

3.3 Contextual *AcroDef* Based on Dice's Coefficient

In this paper, we define the **context** as a set of significant words present in the page where the acronym to expand is found. Of course, other definitions of the context notions have to be considered as extensions to this preliminary approach. However, even in this restricted point of view, several operational expressions of the context could be used:

- the n most frequent words (excepting tool words);
- the n most frequent proper name;
- the n most rare words;
- grammatical (part-of-speech tag) [3] or terminological information [2,8,13] present in the surroundings of the considered item.

A combination of these expressions could also be envisaged. The experiments presented in this article (section 4) use a context represented by the most frequent words, and give satisfying results. Other experiments using several contexts will be proposed in a future work.

Adding contextual information to *BasicAcroDef* (formula (11)) leads to formula (12). The principle underlying this formula is to apply statistical measures on a set of words of a given domain. So, the goal is not to count the dependency between the words of an acronym definition and those of the context, but to restrict the searching space. This restriction is a requirement for the word dependency computation (and not otherwise). The formula is written as follows:

$$AcroDef_{Dice}(a^j) = \frac{\left|\{a_i^j + C | a_i^j \notin M_{tools}\}_{i \in [1,n]}\right| \times nb(\bigcap_{i=1}^{n} a_i^j + C)}{\sum_{i=1}^{n} nb(a_i^j + C | a_i^j \notin M_{tools})} \quad \text{where } n \geq 2 \tag{12}$$

[3] Queries submitted in December 2006.

In this formula, $a_i^j + C$ represents the pages containing the word a_i^j with all the words of the context C. For this we use the Exalead AND operator. If we take our example $a =$`JO` with its two possible expansions (`Jeux Olympiques` and `Journal Officiel`), the favored definition with $BasicAcroDef$ is still `Jeux Olympiques` since it scores 0.0062 against the 0.0058 value for `Journal Officiel`. If we take as a first context the following $C = \{$`loi`$\}$ (meaning *law*) then in this case we have:

- $AcroDef_{Dice}(\texttt{JO}^1) = \frac{2 \times nb((\texttt{Jeux} \cap \texttt{Olympiques}) + \texttt{loi})}{nb(\texttt{Jeux} + \texttt{loi}) + nb(\texttt{Olympiques} + \texttt{loi})} = 0.018$
- $AcroDef_{Dice}(\texttt{JO}^2) = \frac{2 \times nb((\texttt{Journal} \cap \texttt{Officiel}) + \texttt{loi})}{nb(\texttt{Journal} + \texttt{loi}) + nb(\texttt{Officiel} + \texttt{loi})} = 0.159$

Now, the choice of Dice's coefficient for $AcroDef$ either basic or contextual could be questioned as such. Dice's coefficient is known to favor frequent associations, but so does the Cubic MI. And what about MI in the case of acronym expansions? What are its advantages or liabilities? These questions have lead us to attempt a comparison between fundamental measures as variables in the $AcroDef$ quality metrics and is the subject of the following subsection.

3.4 An MI and Cubic MI Based *AcroDef*

In order to provide comparisons between basic measures, the formulas (13) and (14) define the $AcroDef$ measures, respectively based on MI and Cubic MI.

$$AcroDef_{IM}(a^j) = \frac{nb(\bigcap_{i=1}^{n} a_i^j + C)}{\prod_{i=1}^{n} nb(a_i^j + C | a_i^j \notin M_{tools})} \text{ where } n \geq 2 \tag{13}$$

$$AcroDef_{IM3}(a^j) = \frac{nb(\bigcap_{i=1}^{n} a_i^j + C)^3}{\prod_{i=1}^{n} nb(a_i^j + C | a_i^j \notin M_{tools})} \text{ where } n \geq 2 \tag{14}$$

These different measures that are language independent are tested in the following section dedicated to the experimentation of $AcroDef$ on real data.

4 Experiments

The application, programmed in Perl, contains different parameters, that are: The number of words in the context C, the tool words list, the different quality measures. The following subsections describe the experimental protocol implemented for the system evaluation, with a corpus of a sensible length (see section 4.1) manually built, and a large corpus (see section 4.2). The first is a pre-evaluation corpus, evaluating the feasibility of the method and the measures soundness, and the second is a real "live" corpus, which results correspond to what is expectable from our system.

4.1 Experimenting on a Manually Built Corpus for a Pre-evaluation

To test both feasibility and soundness, we have focused on the study of the "JO" French-based acronym explained before. We have browsed a set of a 100

Web pages containing this acronym, split into 50 pages with "JO" abbreviating `Journal Officiel`, and the 50 remaining for `Jeux Olympiques`. These pages have been obtained as a result of several manual queries with Google's search engine. They contain no expansion of the "JO" acronym, as required for our working hypothesis.[4]

The first task was to clean the corpus by removing the HTML tags and the tool words, deleting punctuation marks and various special characters. Then, to evaluate the various measures defined before, we built the *contingency evaluation matrix* provided in table 1.

Table 1. Contingency evaluation matrix

		Real	
		Journal Officiel	Jeux Olympiques
Prediction	Journal Officiel	a	c
	Jeux Olympiques	b	d

where

- a is the number of pages correctly predicted with the expansion `Journal Officiel`,
- b is the number of pages predicted with the expansion `Jeux Olympiques` but which real expansion is `Journal Officiel`,
- c is the number of pages predicted with the expansion `Journal Officiel` but which real expansion is `Jeux Olympiques`,
- d is the number of pages correctly predicted with the expansion `Jeux Olympiques`.

The system quality is measured by estimating the **error ratio** (ER) corresponding to the number of ill-classified pages divided by the total number of predictions, $ER = \frac{b+c}{a+b+c+d}$. For instance, when using only the $BasicAcroDef$ formula (based on Dice's coefficient, and without context) (formula (11)), the best score is always obtained with the `Jeux Olympiques` expansion (see section 3.2). This implies that all pages are classified into the category "Olympic Games", and thus leads to an error ratio of 50% (with $b = d = 50$ and $a = c = 0$). This is why we suggest to use a context composed of one to three words (the most frequent words, different from tool words, in every page). Restricting the evaluation to a maximum of three words context is motivated by the fact that with four words, many queries get no pages as a result.

The results of this preliminary experiments are detailed in table 2. This test set has required 1800 queries to the Exalead search engine[5] with 6 queries per page and 3 test sets of 100 pages each. The workload for building such a test set is heavy and explains why, as a first exploratory task, we restricted our preliminary evaluation to one acronym.

[4] We have used for this the subfield "pages containing none of the following words" of the "advanced research" Google functionality.

[5] Experiment lead in December 2006.

Table 2. Error ratio on a pre-evaluation test corpus of 100 Web pages (acronym "JO")

	1 word Context	**2 words Context**	**3 words Context**
$AcroDef_{IM}$	47%	45%	42%
$AcroDef_{IM3}$	26%	14%	8%
$AcroDef_{Dice}$	29%	16%	9%

Table 2 shows that measure with a low error ratio are Cubic MI and Dice's Coefficient (as expected). However, the use of both measures is here context-dependent, and the larger the context is, the better the measure behaves. A three words context has a low error ratio with Cubic MI and Dice's measure (respectively 8% and 9%). Most classification errors are caused by the most frequent words that are not related to the domain (words like "tomorrow", "july", "France", etc.). Further, cleaning the HTML pages might be difficult in some cases and might also provoke errors in the expansion prediction.

However, this first evaluation is interesting because it highlights two phenomena:

- It definitely dismisses a simple MI measure, regardless of any context: The error ratio with such a measure is 3 to 5 times the error ratios of its fellow measures.
- The context width has a significant impact on results and the best measures (Cubic MI and Dice) are more sensitive to it than MI.

Since an error ratio of 8% corresponds to a success ratio of 92% then it seems that *a Cubic MI with a three words context might be the best quality measure for an acronym expansion candidate*, when this expansion is absent from the considered document.

This "conclusion" seen as working hypothesis needs to be reinforced. So we tested the three measures on a much larger scale to see whether it still holds. Next section presents an experiment on 1303 texts.

4.2 Experimenting on a Larger Corpus

For this experiment we have used a corpus provided by the Evaluation Conference DEFT'06 (*DÉfi Fouille de Textes*, meaning *Text Mining Challenge*), which is a francophone equivalent of the TREC Conferences. This second edition of the Text Mining Challenge consisted in providing a thematic text segmentation for French written corpora belonging to various domains (politics, law, science). We particularly focused on the law corpus, composed of law articles of the European Union.[6] The 1303 articles (11 Mb) containing the JO acronym are selected. This acronym is generally used in this corpus to refer to precise articles of the Official Journal (for example, references "JO 308 du 18.12.1967" or "JO no L 249 du

[6] Corpus available at the following address:
http://www.lri.fr/ia/fdt/DEFT06/corpus/donnees.html

8.9.1988" where JO acronym is not defined). For every law article, we measure if the JO acronym has to be associated with the `Journal Officiel` expansion by using the *AcroDef* measures. Table 3 details the error ratios obtained with this corpus, with a context width varying from one to three words. In this experiment we had to submit 23,454 queries computed as such: 1303 articles, 6 queries per article and 3 test sets for the three context width values (one to three words).

Table 3. Error ratios of the three *AcroDef* measures and the three context widths, using DEFT06 law corpus

	Number of correctly associated acronyms	*ER*
	1 word Context	
DefAcro$_{IM}$	190	85.4%
DefAcro$_{IM3}$	1040	20.2%
DefAcro$_{Dice}$	842	35.4%
	2 words Context	
DefAcro$_{IM}$	434	66.7%
DefAcro$_{IM3}$	1234	5.3%
DefAcro$_{Dice}$	1200	7.9%
	3 words Context	
DefAcro$_{IM}$	650	50.1%
DefAcro$_{IM3}$	1281	1.7%
DefAcro$_{Dice}$	1274	2.2%

Table 3 shows that our method improves its results on a large corpus: With the DEFT06 corpus, the obtained results are very satisfying with our best error ratios around 2%. The context width impact is confirmed: Errors are significantly reduced with a 2 or 3 words context. Moreover, the capabilties of Cubic MI and Dice's coefficient are also confirmed over simple MI: With a 3 words context, their error ratios are respectively 1.7% and 2.2%. These two measures favor frequent co-occurrences. In our case, the number of Web pages sharing an expansion associated to a relevant context is important. As a consequence, a high score is given to measures that return a high number of pages.

One of the possible explanations for such good results on this corpus could be related to the specificity of the DEFT06 corpus: It belongs to a given domain, and the most frequent words constituting contexts are representative of the domain of law. Whereas the pre-evaluation corpus pages, derived from the Web directly through queries, could show up some ambiguities (for instance, texts dealing with the economical consequences of Olympic Games). However, experiments tend to show that, whatever the nature of the corpus is, the *AcroDef* measures with Cubic MI and Dice's coefficient are rather efficient and meaningful.

4.3 Extending Experimentation to Different Couples of Acronyms/Definitions

Ambiguous acronyms are naturally very frequent, and this first study with the French ambiguous "JO" acronym has lead us to attempt a further investigation about the acronyms of the principal French political parties (as one knows, they are rather numerous). The goal of it was to examine the various quality measures with a variable number of suggested definitions. Moreover, the acronyms could be built out of several words (and not only two as in our first set of experiments).

To start the process, we have imported different definitions for the acronyms LCR, PCF, PS, UDF, UMP, FN on the site http://www.sigles.net/ to build an "acronym thesaurus." Without any specific context, these acronyms are naturally recognized by people as political parties names. These definitions are detailed in table 4.[7] The political parties names are in bold.

Table 4. French Political Parties Acronyms Expansions. The parties full names are in bold.

Political acronyms	Expansions	
LCR	**Ligue Communiste Révolutionnaire**	Lettre de Change Relevé
PCF	**Parti Communiste Français**	Paysage Cinématographique Français
	Paysage Culturel Français	Press Club de France
PS	**Parti Socialiste**	Post Scriptum
	Police Secours	Poste de Secours
	Prise de Sang	Premier Secours
	Préfecture de la Sarthe	Préfecture de la Savoie
	Préfecture de la Somme	Passage Supérieur
UDF	**Union pour la Démocratie Française**	Union des Dentistes Français
UMP	**Union pour un Mouvement Populaire**	Urgences Médicales de Paris
FN	**Front National**	Fabrique Nationale
	Fondation Napoléon	

Then we have sent queries to the Google search engine with each of the acronyms and select pages that do not contain their expansions. Then, for each acronym, we have manually extracted the first sites belonging to political parties (about ten per acronym). We have then computed the error rate of this test corpus in order to estimate the number of pages not associated to the definition in the political domain. The obtained results validate these low error rates obtained in the precedent experiments, even with a reduced number of context words. As an example, with a one word context only, the error ratio is less than 4% with the Cubic MI $AcroDef$ measure.

5 Conclusion and Perspectives

Acronyms are widely used words that act as proper names for organizations or associations, or as shortcuts in denominating very frequent concepts or notions.

[7] The acronyms UDF and UMP having only one expansion on this site, we have explored the Web to find other sites with other definitions.

As such, they are representative of the named entities issue under study in the text mining scientific community. Acronyms recognition is one part of the issue, but ambiguous acronyms expansion, especially when the acronym definition is not present in the considered document, is another. This article offers a set of quality measures to determine the choice of the best expansion for an acronym not defined in the Web page that uses it, the *AcroDef* measure. The method uses statistics computed on Web pages to determine the appropriate definition. Measures are deeply **context-based** and rely on the assumption that the most frequent words in the page are related semantically or lexically to the acronym expansion. The first results are very satisfactory since the relevant acronym expansion is found in 92 to 98% of the time, with a context of three words.

Even few, the errors are explained by the fact that they originate from too general words within contexts. If the most frequent words in the page are highly polysemous, too widely used, or vague, this has an impact on the best expansion choice, since the semantic constraint is looser. If the corpus in which acronyms have to be expansed belongs to a given domain, an interesting perspective would be to use as heuristics domain-based descriptors (proper names, terms), or even better, a domain ontology. As an example, the very specific proper name "Beijing", if added to the measure context, could be very relevant to find pages on Olympic Games (to characterize the Olympic Games in China in 2008). The proper name "China" would be also appropriate but "Beijing" strikes better.

Every method has its limitations and needs to be enhanced. This approach has difficulties in building a context for *AcroDef* when the Web page in which the acronym has been found only contains a short text (a few lines for instance). Context extraction relies on words frequency as a cornerstone for thematic detection, and if words are not numerous, frequency becomes meaningless. An interesting perspective would be to represent documents as semantic vectors defined in [5] to get a thematic information on the text. These vectors project the document on a Roget-based ontology and thus do not need quantities of words to sketch a thematic environment for the acronym. That complementary information, associated with *AcroDef*, would help predicting acronym definitions in the case of short texts.

References

1. Azé, J., Kodratoff, Y.: A study of the effect of noisy data in rule extraction systems. In: Proceedings of EMCSR'02, vol. 2, pp. 781–786 (2002)
2. Bourigault, D., Jacquemin, C.: Term extraction + term clustering: An integrated platform for computer-aided terminology. In: Proceedings of the European Chapter of the Association for Computational Linguistics, pp. 15–22 (1999)
3. Brill, E.: Some advances in transformation-based part of speech tagging. In: AAAI, vol. 1, pp. 722–727 (1994)
4. Chang, J., Schtze, H., Altman, R.: Creating an online dictionary of abbreviations from medline. Journal of the American Medical Informatics Association 9, 612–620 (2002)
5. Chauché, J.: Détermination sémantique en analyse structurelle: une expérience basée sur une définition de distance. In: TA Information, pp. 17–24 (1990)

6. Church, K.W., Hanks, P.: Word association norms, mutual information, and lexicography. Computational Linguistics 16, 22–29 (1990)
7. Daille, B.: Approche mixte pour l'extraction automatique de terminologie: statistiques lexicales et filtres linguistiques. PhD thesis, Université Paris 7 (1994)
8. Daille, B.: Study and Implementation of Combined Techniques for Automatic Extraction of Terminology. In: The Balancing Act: Combining Symbolic and Statistical Approaches to Language, pp. 49–66. MIT Press, Cambridge (1996)
9. Jacquemin, C.: Variation terminologique: Reconnaissance et acquisition automatiques de termes et de leurs variantes en corpus. In: Mémoire d'Habilitation à Diriger des Recherches en informatique fondamentale, Université de Nantes (1997)
10. Lallich, S., Teytaud, O.: Evaluation et validation des règles d'association. Numéro spécial Mesures de qualité pour la fouille des données, Revue des Nouvelles Technologies de l'Information (RNTI), RNTI-E-1 pp. 193–218 (2004)
11. Larkey, L.S., Ogilvie, P., Price, M.A., Tamilio, B.: Acrophile: An automated acronym extractor and server. In: Proceedings of the Fifth ACM International Conference on Digital Libraries, pp. 205–214. ACM Press, New York (2000)
12. Petrovic, S., Snajder, J., Dalbelo-Basic, B., Kolar, M.: Comparison of collocation extraction measures for document indexing. In: Proc of Information Technology Interfaces (ITI), pp. 451–456 (2006)
13. Roche, M., Azé, J., Kodratoff, Y., Sebag, M.: Learning interestingness measures in terminology extraction. a roc-based approach. In: Proceedings of ROC Analysis in AI Workshop (ECAI 2004), pp. 81–88 (2004)
14. Roche, M., Kodratoff, Y.: Pruning Terminology Extracted from a Specialized Corpus for CV Ontology Acquisition. In: Meersman, R., Tari, Z. (eds.) On the Move to Meaningful Internet Systems 2006: CoopIS, DOA, GADA, and ODBASE. LNCS, vol. 4276, pp. 1107–1116. Springer, Heidelberg (2006)
15. Smadja, F., McKeown, K.R., Hatzivassiloglou, V.: Translating collocations for bilingual lexicons: A statistical approach. Computational Linguistics 22(1), 1–38 (1996)
16. Thanopoulos, A., Fakotakis, N., Kokkianakis, G.: Comparative Evaluation of Collocation Extraction Metrics. In: Proceedings of LREC'02, pp. 620–625 (2002)
17. Turney, P.D.: Mining the Web for synonyms: PMI–IR versus LSA on TOEFL. In: Flach, P.A., De Raedt, L. (eds.) ECML 2001. LNCS (LNAI), vol. 2167, pp. 491–502. Springer, Heidelberg (2001)
18. Vivaldi, J., Màrquez, L., Rodríguez, H.: Improving term extraction by system combination using boosting. In: Proceedings of the 12th European Conference on Machine Learning (ECML), pp. 515–526 (2001)
19. Yeates, S.: Automatic extraction of acronyms from text. In: New Zealand Computer Science Research Students' Conference, pp. 117–124 (1999)