Managing the Acronym/Expansion Identification Process for Text-Mining Applications

Article · December 2008

Source: DBLP

2 authors:

Mathieu Roche
Cirad - La recherche agronomique pour le dé…
226 PUBLICATIONS **628** CITATIONS

SEE PROFILE

Violaine Prince
Université de Montpellier
113 PUBLICATIONS **364** CITATIONS

SEE PROFILE

Int J Software Informatics, Vol.2, No.2, December 2008, pp. 163–179 E-mail: ijsi@iscas.ac.cn
International Journal of Software and Informatics, ISSN 1673-7288 http://www.ijsi.org
 Tel: +86-10-62661040

Managing the Acronym/Expansion Identification Process for Text-Mining Applications*

Mathieu Roche and Violaine Prince

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

Abstract This paper deals with an acronym/definition extraction approach from textual data (corpora) and the disambiguation of these definitions (or expansions). Both steps of our global process of acquisition and management of acronyms are precisely described. The first step consists in using markers such as brackets to identify expansion candidates. The alignment of the letters allows to select the acronym/definition couples. The second step is to define the relevant expansion of an acronym in a given context. Our method is based on statistical measurements (Mutual Information, Cubic Mutual Information, Dice Measure) and the results provided by search engines. This paper presents an evaluation of the global process from real data (general and specialized domains).

Key words: Web-mining; text-mining; natural language processing; BioNLP; named entities recognition; acronym; quality measures

Roche M, Prince V. Managing the acronym/expansion identification process for text-mining Applications. *Int J Software Informatics*, **2008, 2(2): 163–179. http://www.ijsi.org/1673-7288/2/163.pdf**

1 Introduction

The study of named entities (NE) is useful for many applications in text-mining tasks. This paper deals with a specific type of NE called *acronym*. An acronym is a set of characters corresponding to the first letters of a group of words, and is supposed to designate that group. For instance, the acronym "GIS" can be associated with the *definition* (also called *expansion*) "Geographic Information System". Acronyms are used to avoid a cumbersome repetition of compound terms that need to be frequently addressed in a text. Therefore, NE, standing for 'named entity', is an acronym that is bound to be found in this document. With the huge volumes of data available in different languages, acronyms are crucial in documents from a general field (e.g., IMF – International Monetary Fund, HIV – Human Immunodeficiency Virus, etc) as well as those from specialized domains (e.g., NLP – Natural Language Processing, IJCAI – International Joint Conference on Artificial Intelligence, etc).

Acronyms share with other words the property of being ambiguous, since they might be expanded into several distinct definitions. They are therefore addressed by the classical linguistic issue of *polysemy*, i.e., the existence of several meanings in a single linguistic form. For example GIS can mean "Geographic Information System"

* Corresponding author: Mathieu Roche, Email: mroche@lirmm.fr
Manuscript received 14 Oct., 2008; revised 4 Dec., 2008; accepted 20 Dec., 2008; published online 29 Dec., 2008.

or "Genome Institute of Singapore". These definitions are associated with different domains (geography and biology respectively). Note that specialized fields such as medicine or biology uses many acronyms (see Table 1).

Table 1 Examples of definitions of the acronym ABCD from the biomedical domain

amphotericin B colloidal dispersion
Appropriate Blood Pressure Control in Diabetes
Access to Baby and Child Dentistry
AmB colloidal dispersion
Association of British Clinical Diabetologists

Acronyms, present in a general field, are not necessarily suited to a specialized field. For example, the acronym GIS from a biomedical field has very different expansions compared to the previous definitions proposed: "Glomerular Inulin Space", "Graft Intolerance Syndrome", etc. Therefore, it is necessary to build specialized dictionaries, and to design a rather sophisticated process handling acronyms disambiguation, mostly when the acronym expansion is not present in the text that uses it. This is what happens in most scientific communities: For instance, the acronym MRI is no more explained in most medical texts. When a reader has not acquired the previous knowledge, he/she is unable to expand the acronym.

This paper addresses that particular issue, and describes a global process of acronym acquisition and expansion running in two distinct stages.

Stage 1: Acronyms and definitions are first extracted from corpora. This stage allows to build or enrich dictionaries. The method has two steps detailed in section 3: (1) Extraction of acronym/definition candidates, (2) Filtering the candidates.

Stage 2: After the dictionary acquisition process is ended, a quality measure is applied to determine the relevant definition of an acronym in a document from which its appropriate definition is absent. In this context, it is essential to have a suitable dictionary, which justifies the process first stage. Statistics from the Web are used to select the relevant definition (section 3) and the documents context is also taken into account.

Figure 1 summarizes the global process of acronyms/definitions management. Each stage will be both described and evaluated in this article. First, Section 2 offers a brief survey on the state-of-the-art literature on acronyms/definitions extraction. The following sections describe the successive stages of the acronyms acquisition process (Section 3 as well as their disambiguation (Section 4).

2 Acronym/Expansion Detection in Literature

Among the several existing methods for acronyms and acronyms expansion extraction in the literature, some significant works need to be mentioned. 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[24] involves the following steps: First, separating sentences by segments using specific markers (brackets, points) as frontiers.

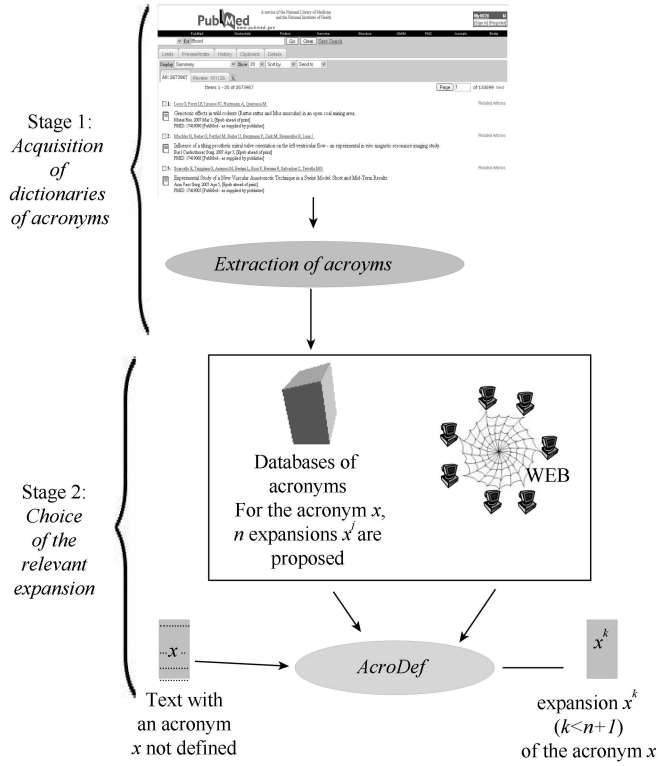

Figure 1. Global acronyms/definitions management process

Then the acronym/expansion couples are tested. The acronym/definition candidates are accepted if the acronym characters correspond to the first letters of the potential definitions words. In the GIS example, the pair "GIS/Geographic Information System" is a good acronym/expansion candidate. The last step uses specific heuristics to select the relevant candidates. These heuristics rely on the fact that acronyms length is smaller than their expansion length, that they appear in upper case, and that long expansions of acronyms tend to use stop-words such as determiners, prepositions, suffixes and so forth. Therefore, the pair "GIS/Geographic Information System" is valid according to these heuristics.

Other works[3, 14] use similar methods based on the presence of markers associated with linguistic and/or statistical heuristics. A recent article[17] employs statistical measurements from the terminology extraction field. Okazaki and Ananiadou apply the C-value measure[11, 16] initially used to extract terminology. It favors a candidate term that doesn't appear often in a longer term. For instance, in a specialized corpus (Ophthalmology), the authors discovered that the term "soft contact" was irrelevant, while the frequent and longer term "soft contact lens" is relevant. An advantage of this C-value measure is its independence from characters alignment (actually, a lot of acronyms/definitions are relevant while the letters are in a different order e.g. "AW / water activity").

Other approaches based on supervised learning methods consist in selecting relevant expansions. In Ref.[23], the authors use the SVM technique (Support Vector Machine) with features based on acronym/expansion informations (length, presence

of special characters, context, etc). Reference [20] presents a comparative study of the main approaches (supervised learning methods, rules-based approaches) by combining domain-knowledge.

Larkey *et al.*'s method[14] 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 (www.altavista.com) search engine. Queries results are Web pages which URLs are explored, and possibly added to the corpus.

The method presented in this paper shares with[14] the usage of the Web. However, it does not look for existing expansions in texts since it tries to determine a possible expansion that would be lacking in the text where the acronym is detected. From that point of view, it is closer to Turney's results in Ref.[21], which are not specifically about acronyms but use the Web to define a ranking function. The algorithm PMI-IR (Pointwise Mutual Information and Information Retrieval) described in Ref.[21] 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 (Test of English as a Foreign Language). The aim is to compute the $choice_i$ synonym that gives the best score. To obtain scores, PMI-IR relies on several measures based on the proportion of documents where both terms are present. Turney's formula is given below (formula (1)): It is one of the basic measures used in Ref.[21]. It is inspired from Mutual Information described in Ref.[5].

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

Here $nb(x)$ 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 $choice_i$. *word* and $choice_i$ are seen as synonyms when we have the higher value of $score(choice_i)$. 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, which is likely to happen, since they are antonyms (opposite meanings). Actually, the synonymy suggested by Turney is more a 'relatedness' than a precise lexical function such as synonymy.

To enhance relevance to the document, the method described in this article tries to take into account dependencies between the words composing the possible expansions in order to rank them. In that sense, it is close to Daille's approach[7] which uses statistical measures to rank terms. Also, other quality measures, defined in next section, are invoked in an attempt to relate the tested acronym as much as possible to its context. Obviously, context information significantly enhances basic measures values.

3 Acquisition of Acronym/Definition Dictionaries

Building acronyms dictionaries is a two steps procedure detailed in the following sections. An user-friendly software developed in Java[15] performs these two steps.

3.1 Step 1: Candidates acronym/definition pairs extraction

First, specific punctuation and character markers (parentheses, brackets, etc) are taken into account, like in the domain state-of-the-art described in the previous section. They enable an acronym/definition pair identification by using different processes:

First case: The acronym is before the definition identified with the markers (within parentheses for example).
Example: "`... GIS (Geographic Information System)...`"

Second case: The definition is before the acronym.
Example: "`... Geographic Information System (GIS) ...`".

In this case, the definition size is indeterminable. It is therefore necessary to define this size by looking at the acronym number of characters, but, at the same time, since potential definition phrases might contain articles, prepositions and other words not recorded in the acronym, it is necessary to provide a larger definition window than the one determined by the acronym characters number. An upper limit to this size is easy to fathom. An empiric study of the various corpora showed that a three times window is unlikely to be too small and is always bigger than necessary. Thus, a $3 \times$ *"number of characters of the acronym"* rule has been applied to the extraction process (e.g., in this example the potential definition of the acronym "GIS" is composed of nine words before this one). For a clarity sake, we call it *a three times rule*.

Step 1 extracts almost all the relevant acronym/definition candidates. Of course, it returns a significant amount of noise (irrelevant candidates). Thus, the second step filters acronym/definition pairs from its resulting list.

3.2 Step 2: Candidates filtering

The second step aims at removing irrelevant acronym/definition pairs and deleting irrelevant word(s) from a potential definition, which size has been determined by the 'three times rule' mentioned above, or by the appropriate marks.

For this process, we propose to align the acronym letters with the potential definition words, by mapping each acronym letter with the first character of each definition word, respecting the words order. If the first letter of the candidate definition word can not be aligned with the acronym corresponding character, the following characters (of the word) are taken into account. On the other hand, if the acronym letter is in no way matched with a first letter in candidate definition corresponding word, then the other characters of the appropriately ranked definition word are browsed and matching is tried in order to provide an alignment.

Example: This method allows to find that "Extraction Itérative de la Terminologie" is a possible definition of the French acronym EXIT. The candidate definition is a five words window, with the letters E, I ,D, L ,T being respectively the first letters

of its words. The acronym is a four letters word: E, X, I, T.

1. *Matching words letters with acronym letters*:
   E (rank 1) from definition matches with E (rank 1) from acronym. I (rank 2) matches with the I rank 3, D and L do not match, therefore, the system tries the other letters of the corresponding words: E (from the word 'de') and then A (from the word 'la') but nothing matches. Last, T (rank 5) from the definition matches with T (rank 4) in the acronym.
   Result: letters E, I, T in the acronym have been respectively aligned with word 1 ('Extraction'), word 3 ('Itérative'), and word 5 ('Terminologie'). Letter 'X' is still pending.

2. *Matching remaining acronym letters with other definition word letters*:
   The acronym letter X rank is 2, and definition word number 2 is aligned with acronym letter number 3. Since order must be mandatorily followed, then X cannot be aligned with a word which rank is superior to 1. Therefore, word number 1, 'Extraction', is analyzed, letter by letter. it happens that the acronym letter X matches with its second letter. Thus the algorithm stops.
   Result: Letter X from acronym is aligned with the second letter of word 1 ('Extraction').

3. *Global result*:
   Definition is relevant, with the following pattern:
   Capital letters are present in both definition and acronym. All other letters (and words) are neglected. '**EX**traction **I**térative de la **T**erminologie'.

In literature, the POS (part-of-speech) tag of the definition word has been studied as a filtering criterion. Reference [14] has a specific treatment for stop-words. Acronyms are believed to neglect words that do no convey 'meaning' in the sense that they address a particular object or subject in the world. However, it is not always true. For instance, several acronyms in French take prepositions into account, such as GDF, meaning 'Gaz de France', the major gas providing company. Because the filtering process has been designed as language independent (a major application aimed at is acronym translation) then it considers stop-words exactly like the others.

The results of the acronym/definition extracting process using our user-friendly software[15] is illustrated in Fig.2. Section 5.1 provides experiments of both steps of the acronym/definition extraction process.

Matching letters capture potential definitions, but not necessarily the relevant ones for a given document in a given domain. This stage has only built acronym dictionaries, and as such, several definitions are associated with a given acronym. Relevance has to be stressed out of another procedure, relying on meaning and context. This is detailed in the following section.

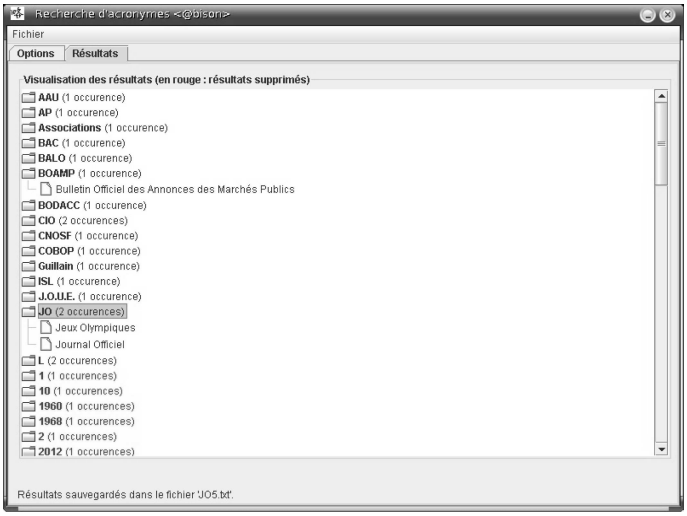

Figure 2. Visualization of the extracted acronyms with our software

4 Choosing Relevant Expansions

4.1 Motivations

The process second stage (see Fig.1 which uses acronym/definition dictionaries built during the first phase) determines relevant expansions in a document. In this case, the appropriate definition of an acronym, not expanded in a document, can be detected with the process hereby described.

A first useful application of this process is that it can be adapted for query expansions from general or specialized domains. For example, a biologist could query a search engine with the acronym "TU" in order to return all the documents using this acronym. Several definitions are possible for this acronym.* Thus, the process would significantly improve the information retrieval task by expanding the original query with the relevant definition of "TU". For example, this expansion could be a disjunction ("OR") of the acronym and its definition. This one returns a large amount of relevant documents. The conjunction of the acronym and the expansion ("AND") enables to return a lower number of documents. But the returned documents are more relevant (i.e. the precision is improved).

Another useful application is that this process helps understanding automatically retrieved texts or text portions, where an acronym is used and its definition is not present in the same portion. We consider an acronym x (e.g., x = GIS), without its definition, in a document d. From the acronym dictionaries described before, a list of possible definitions for the acronym x (e.g., "Geographic Information System", "Genome Institute of Singapore") is available. The process goal is determining the relevant definition x^k (e.g., x^1 = Geographic Information System, x^2 = Genome

* Expansions returned by the Acromine software (www.nactem.ac.uk/software/acromine/): testosterone undecanoate, thiourea, thiouracil, tuberculin units, toxic unit, Tetranychus urticae, T undecanoate, transcription unit, traumatic ulcers, transrectal ultrasonography, temperature, transvaginal ultrasonography

Institute of Singapore) to the document d. To select it, we propose a quality measure based on Web resources.

4.2 Designing an appropriate definition relevance measure: AcroDef

Literature offering statistical measures is abundant, but mainly three measures appear to be the most likely evaluation tools.

- Mutual Information, used by Turney in Ref.[21] with PMI-IR.
- Cubic Mutual Information proposed by Ref.[6] for terminological knowledge acquisition.
- Dice's coefficient extended to n elements.

Selecting the relevant definition needs an adaptation of these measures to the acronym expansion issue, and mostly, the introduction of context, which is a very specific feature of this approach. Therefore, three versions of a quality measure, called $AcroDef$, have been designed and a first set of experiments has been performed, detailed in Ref.[19]. The rationale was the following:

- The ranking functions based on Mutual Information are simple and effective because they require little information. Indeed, these measures are based on the number of examples (in our case, the number of pages returned with the word definitions) without the need to identify negative examples (used with many quality measures: Loglikelihood[9], Conviction[2], J-measure[12], Contradiction[1], etc). In our unsupervised context using only statistical information of the Web, the negative examples are often more complex to determine.
- Cubic Mutual Information, defined as the cube of the preceding measure, favors frequent co-occurrences (words that appear together) compared to the original Mutual Information (MI) proposed by Ref.[5]. Cubic Mutual Information is used in many works related to term extraction[22] or NE[8] tasks. Reference [22] considers that the Cubic Mutual Information gives the best behavior.
- Last, Dice's coefficient favors frequent co-occurrences too. This measure favors the words that appear together giving less importance to the presence of these words alone.

4.2.1 Mutual Information and *AcroDef*

Mutual Information (MI) has the following formula:

$$MI = \frac{nb(x_1, \ldots, x_n)}{nb(x_1) \times \ldots \times nb(x_n)} \tag{2}$$

where $x_1 \ldots x_n$ are words in a given window, and nb their occurrence number.

$AcroDef$, based on MI, can be written as:

$$AcroDef_{MI}(x^j) = \frac{nb((\bigcap_{i=1}^{n} x_i^j) + C)}{\prod_{i=1}^{n} nb(x_i^j + C; x_i^j \notin M_{stop})} \quad \text{where } n \geqslant 2 \tag{3}$$

The nb function (formula (3)) is the number of pages returned with the n words x_i^j ($i \in [1, n]$) of the definition x^j. Then, to calculate $AcroDef_{MI}$, we use quotation marks with the Exalead search engine: "$x_1^j \ldots x_n^j$" ($\bigcap_{i=1}^{n} x_i^j$, $i \in [1, n]$). Exalead (www.exalead.fr) search engine is here operated and the test corpus is extracted by applying queries with the Google search engine (www.google.fr). For example, nb(Geographic ∩ Information ∩ System) is the number of pages returned with the query `‘‘Geographic Information System’’`.

The context C, represented by the most frequent (non stop) words of the page containing the acronym to define, is combined with the definition words. $x_i^j + C$ is the word x_i^j with all the words of the context C. Then $nb(x_i^j + C)$ returns the number of pages applying query $x_i^j + C$ using the AND operator of Exalead with the word x_i^j and the context C. For instance, nb(Geographic ∩ Information ∩ System + cartography + map) is the number of pages returned with the query `‘‘Geographic Information System’’ AND cartography AND map`. Here the context C consists of two words (cartography and map). These words are the most frequent non stop-words in the page where the acronym appears without its definition. At this stage of the process, the context definition is based on the frequent keywords of the documents. Future work is intended to integrate richer contexts using language knowledge (vocabulary, syntax, etc.) like in Word Sense Disambiguation (WSD) research[13].

With all other measures, $AcroDef$ will follow the same pattern.

Very quickly, experiments[19] have shown that this simple measure was limited and exhibited comparatively less good results than the others. Therefore, it is not really detailed in this article.

4.2.2 Cubic Mutual Information and Dice's coefficient $AcroDef$ formulas

As explained before, the Cubic Mutual Information[6] calculates the dependency between words $x_1 \ldots x_n$ in a given window. It is formulated as follows:

$$MI3 \quad = \quad \frac{nb(x_1, \ldots, x_n)^3}{nb(x_1) \times \ldots \times nb(x_n)} \tag{4}$$

The $AcroDef$ measure (formula (5)) based on Cubic Mutual Information ($MI3$) computes a score for each definition x^j:

$$AcroDef_{MI3}(x^j) = \frac{nb((\bigcap_{i=1}^{n} x_i^j) + C)^3}{\prod_{i=1}^{n} nb(x_i^j + C; x_i^j \notin M_{stop})} \quad \text{where } n \geqslant 2 \tag{5}$$

Formula (5) is the same as (3) but with its numerator raised to the power 3. $AcroDef_{MI3}$ calculates words dependency like terminology extraction research[7,18,22,8]. This score is obtained by using information given by search engines. In addition, the definition words dependency is computed in the same context.

The formula $AcroDef$ based on the Dice's coefficient extended to n elements (formula (6)) is detailed in Ref.[19]:

$$Dice(x_1, \ldots, x_n) \quad = \quad \frac{n \times nb(x_1, \ldots, x_n)}{nb(x_1) + \ldots + nb(x_n)} \tag{6}$$

Thus, we can build the $AcroDef_{Dice}$ measure based on Dice's coefficient and the

context C:

$$AcroDef_{Dice}(a^j) = \frac{\left|\{a_i^j + C; a_i^j \notin M_{stop}\}_{i\in[1,n]}\right| \times nb((\bigcap_{i=1}^{n} a_i^j) + C)}{\sum_{i=1}^{n} nb(a_i^j + C; a_i^j \notin M_{stop})} \quad (7)$$

where $n \geqslant 2$

and $\left|.\right|$ represents the number of words of the expansion

5 Experiments

This section deals with the evaluation of both stages of the global process: Acronym/ Expansion extraction from corpora (section 5.1) and evaluation of the $AcroDef$ measures (see subsection 4.2). Figure 2 shows the result of the process first stage. Two definitions have been identified for the French acronym "JO" extracted in the documents. This case will be used to experiment the $AcroDef$ measures in this section.

5.1 Evaluating the acronym/expansion acquisition

5.1.1 Acronym/Expansion alignment

Table 2 shows an evaluation of the acronym/definition candidates alignment. For these experiments, a real-world data set has been extracted from the site "sigles.net" (www.sigles.net). The latter provides 30,012 acronyms and their definitions from 30 languages.

Table 2 Acronyms alignment (depending of the number of characters) / expansions

Nb of characters	**Nb of acronyms**	**Nb of expansions**	**Nb of expansions not found**	**% of success**
2	100	616	11	98.2 %
3	50	157	10	93.6 %
4	20	32	7	78.1 %

First, applying a random system, we extract acronyms of two, three, and four characters from this Web site. Then we evaluate the success rate of the alignments (number of acronyms aligned with the definitions of the site using the current version of our software). Table 2 presents the results of 800 matchings. The results are very satisfactory (success rate: 78% to 98%). In addition, this table shows that long acronyms are more difficult to align. The uppercase letters with accents not yet taken into account by our software can explain the difficulty to match long acronyms. However, many cases more difficult to process may exist as the alignment of numeric/non-numeric characters (for instance the French pairs: "3D / Trois Dimensions", "ST2I / Sciences et Techniques de l'Informatique et de l'Ingénierie").

5.1.2 Global evaluation of the acronym/ expansion dictionaries acquisition from a corpus

The global results of the acronym/expansion extraction software from a corpus having a reasonable size (7,465 words) have been analyzed. After applying the process first step, nearly 100 couples have been returned.

To evaluate the results, standard evaluation measures from data mining domain are used: Precision, recall, and F-measure (i.e. harmonic average between precision and recall):

$$\text{Precision} = \frac{\text{number of relevant pairs returned}}{\text{nb of pairs returned}} \tag{8}$$

$$\text{Recall} = \frac{\text{number of relevant pairs returned}}{\text{nb of relevant pairs}} \tag{9}$$

$$\text{F-measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision+Recall}} \tag{10}$$

Table 3 shows that the first step returns not interesting results (low F-measure). Actually this first step extracts all the relevant acronym/definition couples. But the list of the extracted pairs is very noisy. This is translated by a recall of 100% with a low precision. Fortunately, the second step, filtering, significantly reduces noise (precision climbs up to 66.7%), although introducing some silence (recall drops to 80% but is still high), but the resulting F-measure is quite satisfactory (72.7%, that is almost three times its previous value).

Table 3 Evaluating the steps of the acronym/expansion extraction system

	Precision	**Recall**	**F-measure**
Extraction of couples (step 1)	15.2%	100%	26.4%
Extraction of couples + filtering (step 1 + step 2)	66.7%	80%	72.7%

Discussion: This technique cannot be compared with the supervised approaches (e.g. SVM) of the literature because it is unsupervised one, with the pros (more portable to other corpora) and the cons (less tuned to its learning corpus) associated to such a method. Moreover, because of the use of specific heuristics, the experimental comparison with other unsupervised methods in the literature is difficult to perform.

A dictionary acquisition method has been designed to be embedded in a global process of acronym/expansion management. But the main focus of this work is the use of these dictionaries as disambiguation tools to choose the relevant definition of a given acronym (by the application of *AcroDef* measures) in IR or queries expansion tasks. The next sections deal with the evaluation of *AcroDef* measures without the presence of the definition in the documents. Then we can not use the corpus of this section in the following experiments. We propose to use two corpora: A corpus from a general field (section 5.2) and corpus from a specialized domain (section 5.3).

5.2 Evaluating AcroDef on a general field

The first set of experiments was run on the French language and undertook the study of the acronym "JO". In French, this acronym means "Jeux Olympiques" (Olympic Games) or "Journal Officiel" (Official Journal). We have browsed a set

of 100 Web pages having this acronym, split into 50 pages with acronym "JO" for "Journal Officiel", and 50 pages for "Jeux Olympiques". These pages are the result of several manual queries with the Google search engine. They contain no expansion of the "JO" acronym (we have used the "Advanced Search" of Google).

The corpus was cleaned by removing HTML tags, stop-words, punctuation marks, and various special characters, to enable the extraction of frequent words in the document, and to define their context C. A context based on one to three words was selected, containing the one to three most frequent words of each page. A three word maximum was set because an experiment with four showed that no results were returned for an important number of queries. The results quality, presented in Table 4, can be estimated by precision, recall, and F-measure criteria.

Table 4 Evaluation of the $AcroDef$ measure

AcroDef	Context	Precision	Recall	F-measure
$MI3$	1 word	75.4%	74.0%	74.7%
	2 words	86.2%	86.0%	86.1%
	3 words	92.3%	92.0%	**92.1%**
MI	1 word	57.7%	53.0%	55.2%
	2 words	65.3%	55.0%	59.7%
	3 words	68.9%	58.0%	63.0%
$Dice$	1 word	71.0%	71.0%	71.0%
	2 words	84.5%	84.0%	84.2%
	3 words	91.4%	91.0%	91.2%

The $AcroDef$ measures require an important amount of queries in our experiments: $1,800$ queries: Each document requires to test two definitions ("Jeux Olympiques" and "Journal Officiel") with 3 queries by definition. This calculation represents $3 \times 2 = 6$ queries per document. The experiments were conducted with 3 experiments (three kinds of contexts) for 100 documents. We have therefore made $6 \times 3 \times 100 = 1,800$ queries using the Exalead search engine. Table 4 shows that the overall result is quite satisfactory. Using a rich context represented by three words associated with $AcroDef_{MI3}$ returns the F-measure at 92.1%. This first set of experiments led to a primary conclusion expressed as follows:

1. The quality measure $AcroDef$ based on Cubic Mutual Information gives better results because it encourages frequent occurrences (number of pages with the definition to predict). Note that the $AcroDef_{Dice}$ measure also gives good results.

2. For all the quality measures, the results are better with a larger context (bigger number of words used to define the context).

So, after experimenting measures and expansion disambiguation on a general field in French, the next section offers more difficult experiments conducted in a specialized domain in English.

5.3 Evaluating AcroDef on a specialized corpus

Experiments were shifted from French to English, where we focused on a classification of biological data definitions, provided by the Acromine (www.nactem.ac.uk/

software/acromine/) application. For any given acronym in this area, Acromine provides a list of its possible expansions. 102 pairs acronym/definitions have been randomly extracted from Acromine, which provided, for each tested item, from 4 to 6 possible definitions. The current acronyms can be either two, three or four character strings. For instance, JA, PKD, and ABCD are possible acronyms, and for the latter, their definitions are described in Table 1. As one can see, it might range from medicine to biochemistry, dentistry, etc.

For each of these pairs, articles abstracts have been extracted from the specialized bibliographical data base Medline (www.ncbi.nlm.nih.gov/PubMed), containing acronyms and their expansions. This base provides 204 documents (two documents per couple acronym/expansion, manually extracted). The goal of this experiment is to determine whether, for each document, the definition could be correctly predicted by classifying the candidate definitions with the *AcroDef* quality measures. The distribution of the 204 documents according to the number of plausible candidate expansions for acronyms is given in Table 5. This experiment needed to run of 3,500 queries.

Table 5 Number of possible acronym definitions for the 204 documents

Nb of documents	**Nb of possible expansions per document**
12	6
120	5
72	4

5.3.1 Measures and results

Table 6 presents the results of these experiments. For each of the three *AcroDef* measures:

- The first line value is the number of times where the correct definition has been given, as a first item,
- the second line value corresponds to the number of times it has been predicted among the two first definitions (rank 1 and 2 according to the measure classification),
- and the third value corresponds to the number of times it appears among the first three.

Experiments have been led with a one-word context only, i.e., the most frequent word in each document. Working on a specialized domain, queries with more than one word have null pages results with a general search engine such as Exalead. We have chosen this engine in order to be as close as possible to the conditions of preceding experiment, which relies on a general search (and not on intelligent and dedicated engines).

Table 6 shows encouraging figures, particularly for the two last variants of *AcroDef* measure, based on Cubic Mutual Information and Dice's measure. *AcroDef* states that the true definition of an acronym has from 76.5% to 81.9% chance to be found in the first three definitions. However, these results are less striking than those obtained in the pre-evaluation experiments. This might be explained by the complexity

Table 6 Definition prediction of acronyms from medline abstracts

Measure	$AcroDef_{MI}$
Nb of correct definitions	62
on rank 1	(30.4%)
Nb of correct definitions	116
on ranks 1 or 2	(56.9%)
Nb of correct definitions	156
on ranks 1, 2 or 3	(76.5%)
Measure	$AcroDef_{MI3}$
Nb of correct definitions	**74**
on rank 1	**(36.3%)**
Nb of correct definitions	118
on ranks 1 or 2	(57.8%)
Nb of correct definitions	**167**
on ranks 1, 2 or 3	**(81.9%)**
Measure	$AcroDef_{Dice}$
Nb of correct definitions	72
on rank 1	(35.3%)
Nb of correct definitions	**122**
on ranks 1 or 2	**(59.8%)**
Nb of correct definitions	164
on ranks 1, 2 or 3	(80.4%)

of biological data processing, which will be detailed further in next paragraph. Some important facts:

- **Which are the best quality measures?**
  The $AcroDef_{MI3}$ measure has the two best values, one for correct first definitions, and one for correct definitions among the first three. $AcroDef_{Dice}$ has the best one for correct definitions among the first two. $AcroDef_{MI}$ has none, therefore, one can focus on the $MI3$ and $Dice$ based measures. In order to determine more precisely the quality of these measures, we have computed the sum of relevant definitions ranks. The best measure is the one that has the smallest sum. This method, while evaluating rank functions, is equivalent to approaches based on ROC (Receiver Operating Characteristics) and to the calculus of surfaces under them (AUC, standing for Area Under the Curve)[10]. Therefore, Table 7 confirms, similarly to texts in a general context, that $AcroDef_{MI3}$ and $AcroDef_{Dice}$ behave as the best two measures (respectively) in specialized documents belonging to biomedicine.

- **Significance of results:**
  $AcroDef_{MI3}$ hits the good definition on rank 1 in 36.3% of the cases. This is significantly better than a random prediction, which scores 22%. Random prediction was computed as such 1 chance over 4 to put the relevant definition as the first one in 72 cases, 1 over 5 in 120 cases, and 1 over 6 in 12 cases, which are the number of documents with respectively 4, 5, and 6 possible definitions (in Table 5).

- **Restricting the definition space:**
  The high predictive values for the first three definitions ranked by *AcroDef* restricts the search space. It is useless to go down further in the list, and in the 204 documents where more than 4 definitions occur, it would be efficient to restrict to the first three chosen by our measures, and give the user the opportunity of choosing the best one. Further, they might be close definitions as we will show it in a deeper study of the data content.

Table 7 Sums of relevant definitions ranks

$AcroDef_{MI}$	$AcroDef_{MI3}$	$AcroDef_{Dice}$
500	**472**	473

5.3.2 Data properties

The retrieved definitions has led us to formulate some comments. Among the difficulties encountered in NLP research in the biomedical domain, the fact that several terms could address the same or very similar concepts is a very classical issue. For instance, when the acronym ZO was retrieved, we had the following definitions: `zonula occludens`, `zona occludens`, `zonulae occludentes`. As one can see, these are either flexions of the same term (plural vs singular) or very close terms (*zonula* meaning 'small zone' vs *zona*). Variations are explained by linguistic functions or properties. Therefore, quite a fair amount of prediction errors could be caused by linguistic variations on the same basic lexical item.

On the other hand, some equivalent definitions cannot be fathomed without the help of a domain expert. If *terminal* and *termini* could be seen as Latin flexions in the following example: `carboxy terminal`, `carboxy termini`, or in the pair `COOH-terminal`, `COOH-termini`, or in `CO2H-terminal`, `CO2H termini`, the idea that *COOH*, *CO2H* and *carboxy* are equivalent forms (which makes all these pairs totally equivalent to each other) is not automatically deductible and needs expertise. The first and the fourth definitions of Table 1 is an other example of equivalent forms of expansions.

5.3.3 First enhancements

Merging definitions that appeared as linguistic flexions of the same basic terms was the first task to be performed in order to enhance definition disambiguation. Experiments were run again on the same corpus, and merging brought the level of correct definitions found as first definitions by $AcroDef_{MI3}$ up to 43%, which is an absolute enhancement of 7%, and a relative enhancement of 19% when compared to results shown in 6. This means that a superficial enhancement as obvious as this has a noticeable impact on definition prediction. In future works, a thorough study, undertaken with a domain expert will lead to a better definition analysis, and an appropriate merging of similar definitions.

6 Conclusion and Future Work

This paper has presented a complete process handling acronym expansion as a disambiguation task. The process is divided into two major tasks:

- A first stage in which an acronym/definition dictionary is built using a corpus as an input, and satisfactorily evaluated (good F-measure value for results).

- A second stage consisting in determining a relevant definition of an acronym lacking in a document. The selection of the relevant expansion is based on a ranking function: The *AcroDef* measure. This one gives good results from general domains and satisfactory results from specialized corpora.

The described approach relies on achievements present in the state-of-the-art literature: It scrolls the Web to build acronym dictionaries being thus akin to Ref.[14], but is even closer to Peter Turney's[21] which uses the Web to establish a ranking function (PMI-IR). Actually, the PMI-IR (Pointwise Mutual Information and Information Retrieval) algorithm queries the Web with the AltaVista search engine to determine the relevant synonym for a given word (TOEFL – Test of English as a Foreign Language). Like terminology extraction techniques[7], it measures the dependency between each word definitions to order them.

At the same time, it has its own original features and presents some differences with the mentioned related works. It is unsupervised. It builds Web-based dictionaries, going beyond corpora extraction[14], it expands acronyms, thus dealing with compound terms as a rule and not as an infrequent equivalent candidate[21], it goes beyond terminology[7]'s domain since it deals with IR and query expansion. Moreover, it multiplies measures, defines a quality measure appropriate for acronyms, and experiments in more than one language. It has shown that Mutual Information used by Ref.[21] has less good results than more sophisticated measures, when handling this specific task (i.e. acronym disambiguation). One of the *AcroDef* measures peculiarities is that it studies the impact of context and context length to improve basic measures.

In our future work, we propose to automatically associate a semantic context for each acronym. A semantic context means not only words, but also concepts indexing those words, and relationships between words of a given sentence. Then the Semantic Vector approach described by Ref.[4] may help to determine the topic of texts, or texts segments, to restrict furthermore the number of relevant definitions. Finally, a richer context based on linguistic features (context based on NE, grammatical knowledge, and so forth) would hopefully improve results.

References

[1] Azé J. Extraction de Connaissances dans des Données Numériques et Textuelles. Thèse de Doctorat, Univ. de Paris 11, Déc. 2003.

[2] Brin S, Motwani R, Silverstein C. Beyond market baskets: Generalizing association rules to correlations. In: Proc. of the ACM SIGMOD'97. 1997. 265–276.

[3] Chang J, Schtze H, Altman R. Creating an online dictionary of abbreviations from medline. Journal of the American Medical Informatics Association, 2002, 9: 612–620.

[4] Chauché J. Détermination sémantique en analyse structurelle: une expérience basée sur une définition de distance. TA Information, 1990, 1/1: 17–24.

[5] Church KW, Hanks P. Word association norms, mutual information, and lexicography. Computational Linguistics, 1990, 16: 22–29.

[6] Daille B. Approche mixte pour l'extraction automatique de terminologie: statistiques lexicales et filtres linguistiques. Thèse de Doctorat, Univ. de Paris 7, 1994.

[7] Daille B. Study and Implementation of Combined Techniques for Automatic Extraction of Terminology. The Balancing Act: Combining Symbolic and Statistical Approaches to Language. MIT Press, 1996. 49–66.

[8] Downey D, Broadhead M, Etzioni O. Locating complex named entities in web text. In: Proc.

of IJCAI'07. 2007. 2733–2739.

[9] Dunning TE. Accurate methods for the statistics of surprise and coincidence. Computational Linguistics, 1993, 19(1): 61–74.

[10] Ferri C, Flach P, Hernandez-Orallo J. Learning decision trees using the area under the ROC curve. In: Proc. of 9th International Conference on Machine Learning, ICML'02. 2002. 139–146.

[11] Frantzi K, Ananiadou S, Mima H. Automatic recognition of multi-word terms: the C-value/NC-value method. International Journal on Digital Libraries, 2000, 3(2): 115–130.

[12] Goodman MFR, Smyth P. Information-Theoretic rule induction. In: Proc. of ECAI'88 (European Conference on Artificial Intelligence). 1988. 357–362.

[13] Ide N, Véronis J. Word sense disambiguation: The state of the art. Computational Linguistics, 1998, 24: 1–40.

[14] Larkey LS, Ogilvie P, Price MA, Tamilio B. Acrophile: An automated acronym extractor and server. In: Proc. of the Fifth ACM International Conference on Digital Libraries. 2000. 205–214.

[15] Matviico V, Muret N, Roche M. Processus d'acquisition d'un dictionnaire de sigles. In: Proc. of EGC'08 (demo session). 2008. 231–232.

[16] Nenadic G, Spasic I, Ananiadou S. Terminology-Driven mining of biomedical literature. Bioinformatics, 2003, 19(8): 938–943.

[17] Okazaki N, Ananiadou S. Building an abbreviation dictionary using a term recognition approach. Bioinformatics, 2006, 22(24): 3089–3095.

[18] Petrovic S, Snajder J, Dalbelo-Basic B, Kolar M. Comparison of collocation extraction measures for document indexing. In: Proc. of Information Technology Interfaces (ITI). 2006. 451–456.

[19] Roche M, Prince V. AcroDef: A quality measure for discriminating expansions of ambiguous acronyms. In: Proc. of CONTEXT. Springer-Verlag, LNCS, 2007. 411–424.

[20] Torii M, Hu ZZ, Song M, Wu CH, Liu H. A comparison study on algorithms of detecting long forms for short forms in biomedical text. BMC Bioinformatics, 2007.

[21] Turney PD. Mining the Web for synonyms: PMI–IR versus LSA on TOEFL. In: Proc. of ECML, LNCS, 2001. 491–502.

[22] Vivaldi J, Màrquez L, Rodríguez H. Improving term extraction by system combination using boosting. In: Proc. of ECML. 2001. 515–526.

[23] Xu J, Huang Y. Using SVM to extract acronyms from text. Soft Comput., 2007, 11(4): 369–373.

[24] Yeates S. Automatic extraction of acronyms from text. In: New Zealand Computer Science Research Students' Conference. 1999. 117–124.