DefAcro : mesure de qualité pour le choix de la définition des acronymes ambigus

Mathieu Roche, Violaine Prince

Équipe TAL, LIRMM - UMR 5506
Université Montpellier 2,
34392 Montpellier Cedex 5 - France
{mroche,prince}@lirmm.fr

RÉSUMÉ. *Cet article présente un ensemble de mesures de qualité pour déterminer le choix de la meilleure définition pour un acronyme non défini dans la page Web le contenant. L'approche contextuelle que nous proposons utilise des statistiques calculées à partir de pages Web pour déterminer la définition appropriée. Les premiers résultats sont très satisfaisants car la définition pertinente des acronymes est trouvée dans* 92 *à* 98% *des cas.*

ABSTRACT. *This paper 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 contextual approach we promote uses statistics computed on Web pages to determine the appropriate definition. The first results are very satisfactory because the relevant acronym expansion is found in* 92 *to* 98% *of the time.*

MOTS-CLÉS : *Acronyme, Désambiguïsation sémantique*

KEYWORDS: *Acronym, Word Sense Disambiguation*

1. Introduction

La majorité des méthodes de Recherche d'Information (RI) s'appuient sur des approches dites "sacs de mots". Elles consistent à regrouper les textes qui partagent souvent les mêmes mots. Cependant, certains mots identiques peuvent avoir plusieurs sens et renvoyer à des concepts différents (polysémie) posant ainsi des problèmes pour les tâches de RI ou d'extraction d'informations.

Cet article s'intéresse au problème spécifique des acronymes qui sont particulièrement propices à ce type d'ambiguïté. Un acronyme est l'abréviation d'un groupe de mots formé, en général, par les initiales de ces mots. Une distinction existe entre les *sigles* dont chaque lettre est épelée (par exemple, SNCF) contrairement aux *acronymes* qui sont prononcés comme des mots classiques (par exemple, OVNI). Cependant, cet article utilisera le même mot "acronyme" pour désigner ces deux situations qui peuvent se révéler difficiles à distinguer de manière automatique. Au même titre que les mots, les acronymes ont souvent plusieurs sens. Par exemple, l'acronyme "JO" peut être associé aux définitions "Jeux Olympiques" ou "Journal Officiel". Quelques ressources plus ou moins spécialisées existent et proposent des définitions possibles pour un même acronyme. À titre d'exemple, le site http ://www.sigles.net/ fournit une telle liste.

Le problème concerne les textes pour lesquels aucune définition d'acronymes n'est présente. La difficulté est donc de choisir de manière automatique la définition la plus adaptée. Dans ce contexte, posons a un acronyme donné (par exemple, $a =$JO). Pour chaque a dont la définition n'est pas présente dans un document d, considérons que nous avons une liste de n définitions possibles pour a : $a^1 ... a^n$ (par exemple, a^1=Jeux Olympiques, a^2=Journal Officiel). L'objectif de notre approche est de déterminer k ($k \in [1, n]$) tel que a^k soit la définition pertinente pour le document d. Pour effectuer un tel choix, nous proposons une mesure de qualité, DefAcro, qui s'appuie notamment sur les ressources du Web. La figure 1 résume le processus global appliqué.

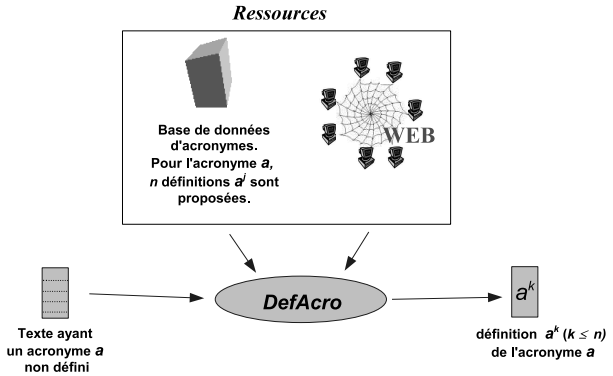

Figure 1. *Processus global.*

Dans un premier temps, la section 2 de cet article présente un résumé de l'état de l'art. La section 3 s'intéresse ensuite à la mesure de qualité DefAcro. D'une part, cette mesure est construite à partir des mesures de qualité classiquement utilisées dans la littérature. La prise en compte du contexte et des ressources du Web constituent, d'autre part, une caractéristique essentielle de cette mesure. La section 4 détaille quelques expérimentations. Enfin, des perspectives à ce travail sont proposées en section 5.

2. État de l'art

De nombreuses méthodes pour extraire les acronymes et leur définition existent dans la littérature. Nous présentons brièvement quelques approches significatives. Rappelons que nos travaux se placent dans un contexte différent car, dans notre cas, nous traitons des textes ne contenant pas la définition des acronymes. L'objectif est de sélectionner les définitions pertinentes des acronymes parmi une liste de définitions candidates.

La plupart des approches de détection d'acronyme dans les textes s'appuient sur l'utilisation de marqueurs spécifiques. La méthode de (Yeates, 1999) consiste dans un premier temps à séparer les phrases par fragments en utilisant des marqueurs spécifiques (parenthèses, points, etc.) comme frontières. Par exemple, la phrase :

- `Les JO (Jeux Olympiques) seront organisés en Chine en 2008.`

devient

- `Les JO | Jeux Olympiques | seront organisés en Chine en 2008 |`

L'étape suivante a pour but de comparer chaque mot de chacun des fragments avec les fragments précédents et suivants. Ainsi, dans notre exemple, les comparaisons suivantes sont effectuées :

`Les` avec `Jeux Olympiques` `Jeux` avec `Les JO`
`JO` avec `Jeux Olympiques` `Olympiques` avec `Les JO`, etc.

Ensuite, les couples acronymes/définitions sont testés. Les candidats acronymes sont retenus si les lettres de l'acronyme sont mises en correspondance avec les premières lettres des définitions potentielles. Dans notre cas, le couple "JO/Jeux Olympiques" est un candidat acronyme. La dernière étape consiste à utiliser des heuristiques spécifiques pour retenir les candidats pertinents. Ces heuristiques s'appuient sur le fait que les acronymes ont une taille plus petite que leur définition, ils sont en majuscule, les définitions des acronymes ayant une longueur importante ont tendance à posséder davantage de mots outils (par exemple, les articles et les prépositions), etc. Dans notre cas, le couple "JO/Jeux Olympiques" qui vérifie ces heuristiques peut alors être considéré comme un acronyme. De nombreuses approches (Larkey *et al.*, 2000, Chang *et al.*, 2002) utilisent des méthodes similaires fondées sur la présence de marqueurs associés à des heuristiques spécifiques.

Quelques méthodes recherchent des définitions des acronymes en utilisant le Web. Par exemple, l'utilisation d'un moteur de recherche intervient dans les travaux de (Larkey *et al.*, 2000) afin d'enrichir un corpus initial de pages Web utiles pour la recherche d'acronymes. Pour ce faire, à partir d'une liste d'acronymes, des requêtes sont sou-

mises au moteur de recherche AltaVista[1]. Ceci permet d'acquérir des pages Web dont les URLs sont à leur tour explorées pour enrichir le corpus des pages Web.

Notre approche a des similarités avec (Larkey *et al.*, 2000) concernant l'utilisation du Web. Cependant, dans notre cas, nous ne recherchons pas les définitions des acronymes dans les textes car nous nous intéressons à déterminer les définitions des acronymes qui sont absentes des textes. Notre approche a davantage de similarités avec les travaux de (Turney, 2001) qui ne s'intéressent pas spécifiquement à la recherche des acronymes mais qui utilisent le Web pour établir une fonction de rang. L'algorithme PMI-IR (Pointwise Mutual Information and Information Retrieval) de (Turney, 2001) consiste à interroger le Web via le moteur de recherche AltaVista pour déterminer des synonymes appropriés. À partir d'un terme donné noté mot, l'objectif de PMI-IR est de choisir un synonyme parmi une liste donnée. Ces choix, notés $choix_i$, correspondent aux questions du TOEFL. Ainsi, le but est de calculer, pour chaque mot, le synonyme $choix_i$ qui donne le meilleur score. Pour ce faire, l'algorithme PMI-IR utilise différentes mesures fondées sur la proportion de documents dans lesquels les deux termes sont présents. Nous donnons ci-dessous (formule (1)) une des mesures de base utilisée dans les travaux de (Turney, 2001). Cette mesure s'appuie sur l'Information Mutuelle qui sera décrite dans la section 3.1.1.

$$score(\ choix_i\ ) = \frac{nb(\ mot\ NEAR\ choix_i\ )}{nb(\ choix_i\ )} \qquad [1]$$

– $nb(x)$ calcule le nombre de documents contenant le mot x,

– $NEAR$ (utilisé dans la rubrique "recherche avancée" d'Altavista) est un opérateur qui précise si deux mots sont présents ensemble dans une fenêtre de 10 mots.

Ainsi, la formule (1) calcule la proportion de documents contenant mot et $choix_i$ dans une fenêtre de 10 mots par rapport au nombre de documents contenant le mot $choix_i$. Plus la proportion de documents contenant ces deux mots dans une même fenêtre est importante et plus mot et $choix_i$ sont considérés comme synonymes. D'autres formules plus élaborées ont également été appliquées. Ces formules utilisent les informations sur la présence de négations dans les fenêtres de 10 mots. Par exemple, les mots « `grand` » et « `petit` » ne sont pas synonymes si, dans une même fenêtre, la présence d'une négation associée à un des deux mots est relevée.

Notre approche qui est une méthode non supervisée possède des différences majeures par rapport à la méthode de (Turney, 2001). Dans un premier temps, nous considérons que toutes les définitions associées aux acronymes sont pertinentes. Ainsi, nous avons décidé de ne pas mesurer la dépendance entre les acronymes et leur définition mais, de manière similaire aux travaux de (Daille, 1994), d'étudier la dépendance entre chacun des mots représentant les définitions afin d'ordonner ces dernières. De plus, l'Information Mutuelle utilisée par (Turney, 2001) est une mesure qui a des limites comme nous le montrerons par la suite. Ainsi, nos travaux s'appuient sur d'autres me-

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

sures de qualité. Par ailleurs, l'utilisation d'un contexte spécifique permet d'améliorer significativement les mesures de base.

Précisons que notre approche n'utilise pas de corpus d'apprentissage pour choisir la définition adaptée aux acronymes (un tel corpus pouvant représenter la base de la méthode LSA (Landauer *et al.*, 1998, Turney, 2001)). Les seules ressources utilisées sont les statistiques issues des moteurs de recherche et une liste de définitions possibles d'acronymes. Notons enfin que contrairement à de nombreux travaux liés à la désambiguïsation sémantique (Audibert, 2003), notre approche n'utilise aucune connaissance linguistique telles que les informations lexicales et/ou syntaxiques. Cependant les informations grammaticales sont souvent des critères pertinents qui peuvent se révéler intéressants à associer aux mesures statistiques décrites dans la section suivante.

3. Définition de la mesure DefAcro

3.1. *Mesures statistiques*

Dans la littérature, de nombreuses mesures de qualité sont utilisées afin d'effectuer un classement par intérêt décroissant. Ces mesures sont issues de domaines variés : recherche de règles d'associations (Azé, 2003, Lallich *et al.*, 2004), extraction de la terminologie (Daille, 1994, Roche, 2004), etc. Notre approche consiste à sélectionner la définition d'un acronyme à partir d'une liste de co-occurrences de mots (définition des acronymes). Le but est donc d'effectuer un classement par pertinence en utilisant des mesures statistiques ; les définitions les plus pertinentes devant être placées en début de liste.

3.1.1. *Information Mutuelle*

Une des mesures couramment utilisée pour calculer une certaine forme de dépendance entre chacun des mots composant une co-occurrence est l'Information Mutuelle (Church *et al.*, 1990) :

$$I(x,y) \quad = \quad \log_2 \frac{P(x,y)}{P(x)P(y)} \qquad [2]$$

Une telle mesure a tendance à extraire des co-occurrences rares et spécifiques (Daille, 1994, Thanopoulos *et al.*, 2002, Roche, 2004). Notons que, dans la formule (2), l'utilisation de la fonction log_2 n'est pas nécessaire. En effet, la fonction log_2 est strictement croissante, l'ordre des co-occurrences donné par la mesure n'est donc pas affecté avec l'application ou non de la fonction log_2. Dans notre cas, $P(x,y)$ permet d'estimer la probabilité d'apparition des couples de mots (x,y) où x et y sont voisins dans cet ordre. Par exemple, avec l'acronyme JO, x peut représenter le mot "Journal" et y le mot "Officiel". Après diverses approximations, la formule (2), peut s'écrire de la manière suivante où nb représente le nombre d'occurrences des mots et des couples de mots :

$$IM(x,y) \quad = \quad \log_2 \frac{nb(x,y)}{nb(x)nb(y)} \qquad [3]$$

Cette mesure peut être adaptée aux co-occurrences ternaires de manière similaire aux travaux de (Jacquemin, 1997). Ainsi, une extension naturelle consiste à appliquer cette mesure à des définitions d'acronymes formés de n mots (formule (4)).

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

3.1.2. *Information Mutuelle au Cube*

L'Information Mutuelle au Cube (Daille, 1994) est une mesure empirique qui s'appuie sur l'Information Mutuelle mais en privilégiant davantage les co-occurrences fréquentes. Une telle mesure est définie par la formule (5).

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

Cette mesure est utilisée dans bon nombre de travaux liés à l'extraction des termes nominaux ou verbaux dans les textes (Vivaldi *et al.*, 2001, Claveau *et al.*, 2003). (Vivaldi *et al.*, 2001) ont d'ailleurs estimé que l'Information Mutuelle au Cube était la mesure qui avait le meilleur comportement. De la même manière que l'Information Mutuelle, une telle mesure peut être étendue de la manière suivante :

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

3.1.3. *Coefficient de Dice*

Dans la suite nous allons présenter une autre mesure de qualité appelée coefficient de Dice (Smadja *et al.*, 1996). Cette mesure est définie par la formule (7).

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

De manière similaire à l'Information Mutuelle au Cube, le coefficient de Dice privilégie moins les co-occurrences rares souvent non pertinentes (Roche *et al.*, 2006). La formule (7) permet de directement déduire la formule (8) qui s'appuie sur le nombre d'occurrences nb des mots et des couples de mots[2].

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

Dans les travaux de (Petrovic *et al.*, 2006), les auteurs présentent une extension de la formule d'origine de Dice à trois éléments :

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

2. en posant, $P(x) = \frac{nb(x)}{nb\_total}$, $P(y) = \frac{nb(y)}{nb\_total}$, $P(x, y) = \frac{nb(x,y)}{nb\_total}$

Nous pouvons de la même manière proposer une extension naturelle à n éléments :

$$Dice(x_1, ..., x_n) \quad = \quad \frac{n \times nb(x_1, ..., x_n)}{nb(x_1) + ... + nb(x_n)} \qquad [10]$$

Les deux sections suivantes (sections 3.2 et 3.3) présentent la mesure DefAcro que nous proposons (mesure de base et mesure contextuelle) qui s'appuie sur l'utilisation du coefficient de Dice. La section 3.4 propose la mesure DefAcro qui s'appuie sur les deux autres mesures statistiques (Informations Mutuelle et Information Mutuelle au Cube).

3.2. *Mesure DefAcro de base fondée sur le coefficient de Dice*

Notre contexte de travail est lié à l'élaboration d'une mesure de qualité qui s'appuie sur les ressources du Web. Dans ce cas, la fonction nb utilisée dans les différentes mesures qui ont été détaillées dans la section précédente représente le nombre de pages retournées par le moteur de recherche Exalead (http ://www.exalead.fr/). En effet, comme expliqué dans la section 4, la constitution du corpus de test va s'effectuer à partir du moteur de recherche google (http ://www.google.com/). Il est donc intéressant de nous appuyer sur un moteur de recherche différent pour la mesure DefAcro.

À partir de la formule (10), nous pouvons en déduire la formule (11) qui représente la mesure DefAcro de base.

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

où

– $\bigcap_{i=1}^{n} a_i^j$ désigne la suite de mots a_i^j ($i \in [1, n]$) que l'on considère comme une chaîne de caractères (utilisation des *guillemets* avec Exalead que l'on peut illustrer de la manière suivante : "$a_1^j ... a_n^j$").

– M_{outils} représente une liste de mots outils (prépositions, articles, etc.). Le but est de ne pas considérer le nombre de pages possédant ces mots outils qui ne sont pas porteurs de sens.

– $|.|$ désigne le nombre de mots de l'ensemble.

Avec l'acronyme $a =$`JO`, deux définitions sont possibles (voir http ://www.sigles.net/) : a^1 : `Jeux Olympiques` et a^2 : `Journal Officiel`

Les scores obtenus avec la mesure donnée par la formule (11) sont très proches[3].

– $\text{DefAcroBase}_{Dice}(\texttt{JO}^1) = \frac{2 \times nb(\text{Jeux} \cap \text{Olympiques})}{nb(\text{Jeux}) + nb(\text{Olympiques})} = \frac{2 \times 366508}{116929964 + 1207545} = 0.0062$

– $\text{DefAcroBase}_{Dice}(\texttt{JO}^2) = \frac{2 \times nb(\text{Journal} \cap \text{Officiel})}{nb(\text{Journal}) + nb(\text{Officiel})} = \frac{2 \times 603036}{178302348 + 28140994} = 0.0058$

3. Requêtes effectuées en décembre 2006.

Dans la pratique, le premier exemple revient à effectuer avec Exalead les trois requêtes suivantes : `"Jeux Olympiques"`, `Jeux`, `Olympiques`. Notons que dans ce cas, davantage de pages sont retournées avec la requête `"Journal Officiel"`, pourtant le score le plus élevé est obtenu avec `"Jeux Olympiques"`.

Lorsque la définition d'un acronyme possède un mot outil telle qu'une préposition, il ne sera pas pris en compte dans le dénominateur de la formule. En effet, il n'est pas pertinent de calculer le nombre de pages possédant seulement les mots outils. Ainsi, l'acronyme `ADP` signifiant `Aéroports de Paris` et `Association des Diplômés de Polytechnique` est calculé de la manière suivante :

– $\text{DefAcroBase}_{Dice}(\texttt{ADP}^1) = \frac{2 \times nb(\text{Aéroports} \cap \text{de} \cap \text{Paris})}{nb(\text{Aéroports}) + nb(\text{Paris})}$

– $\text{DefAcroBase}_{Dice}(\texttt{ADP}^2) = \frac{3 \times nb(\text{Association} \cap \text{des} \cap \text{Diplômés} \cap \text{de} \cap \text{Polytechnique})}{nb(\text{Association}) + nb(\text{Diplômés}) + nb(\text{Polytechnique})}$

3.3. *DefAcro contextuelle fondée sur le coefficient de Dice*

La mesure de base proposée a une limite majeure liée au fait que le score ne prend pas en compte le contexte. Ainsi, nous proposons de considérer ce dernier pour effectuer un choix plus pertinent de la définition à associer à chaque acronyme. Dans cet article, nous définissons le contexte comme des mots caractéristiques présents dans la page dans laquelle l'acronyme à définir est présent. Plusieurs contextes C peuvent être utilisés :

– n mots les plus fréquents (sauf les mots outils).

– n noms propres les plus fréquents.

– n mots les plus rares.

– Utilisation d'informations grammaticales (noms, verbes, etc.) (Brill, 1994) et/ou de la terminologie (Daille, 1994, Bourigault *et al.*, 1999, Roche, 2004).

Une combinaison de ces contextes peut également être envisagée. Notons que les expérimentations présentées dans cet article (section 4) s'appuient sur le contexte représenté par les mots les plus fréquents qui donne des résultats très satisfaisants.

L'ajout d'informations contextuelles à la mesure DefAcro (formule (11)) permet la construction de la formule (12). Le principe de cette mesure contextuelle est d'appliquer des approches statistiques sur un ensemble qui est propre au domaine étudié. La dépendance des mots de la définition de l'acronyme est alors calculée à partir des seules pages partageant un contexte proche.

$$\text{DefAcro}_{Dice}(a^j) = \frac{\left|\{a_i^j + C; a_i^j \notin M_{outils}\}_{i \in [1,n]}\right| \times nb((\bigcap_{i=1}^{n} a_i^j) + C)}{\sum_{i=1}^{n} nb(a_i^j + C; a_i^j \notin M_{outils})} \quad [12]$$

Dans la formule (12) où $n \geq 2$, $a_i^j + C$ représente le mot a_i^j avec tous les mots du contexte C. $nb(a_i^j + C)$ retourne le nombre de pages données par le moteur de recherche avec la requête $a_i^j + C$ (utilisation de l'opérateur *AND* d'Exalead).

Reprenons l'exemple de l'acronyme $a =$`JO`, qui possède deux définitions possibles (`Jeux Olympiques` et `Journal Officiel`). Rappelons qu'avec la mesure de base, la définition privilégiée est toujours `Jeux Olympiques` :

$$\text{DefAcroBase}_{Dice}(\texttt{JO}^1) = 0.0062 \text{ et } \text{DefAcroBase}_{Dice}(\texttt{JO}^2) = 0.0058$$

Prenons dans un premier temps $C = \{\texttt{loi}\}$. Dans ce cas, nous avons :

– $\text{DefAcro}_{Dice}(\texttt{JO}^1) = \frac{2 \times nb((\text{Jeux} \cap \text{Olympiques}) + \text{loi})}{nb(\text{Jeux} + \text{loi}) + nb(\text{Olympiques} + \text{loi})} = 0.018$

– $\text{DefAcro}_{Dice}(\texttt{JO}^2) = \frac{2 \times nb((\text{Journal} \cap \text{Officiel}) + \text{loi})}{nb(\text{Journal} + \text{loi}) + nb(\text{Officiel} + \text{loi})} = 0.159$

Dans la pratique, le premier exemple revient à effectuer les trois requêtes suivantes en utilisant Exalead : `"Jeux Olympiques" AND loi`, `Jeux AND loi`, `Olympiques AND loi`.

La mesure prenant en compte le contexte $C = \{\texttt{loi}\}$ permet de privilégier la définition `Journal Officiel` à associer à l'acronyme `JO`. Cette mesure revient à calculer le coefficient de Dice à partir des seules pages contenant le mot `loi`.

En utilisant le contexte $C = \{\texttt{sport}\}$, on obtient :

$$\text{DefAcro}_{Dice}(\texttt{JO}^1) = 0.025 \text{ et } \text{DefAcro}_{Dice}(\texttt{JO}^2) = 0.010$$

Ainsi, dans ce cas, la définition `Jeux Olympiques` est privilégiée. Bien entendu, le fait de donner un contexte encore plus riche (composés de plusieurs mots) permet d'accentuer les écarts des scores pour les deux définitions. Par exemple, avec $C = \{\texttt{sport, natation}\}$ nous avons :

$$\text{DefAcro}_{Dice}(\texttt{JO}^1) = 0.190 \text{ et } \text{DefAcro}_{Dice}(\texttt{JO}^2) = 0.008$$

Notons enfin que la mesure DefAcro_{Dice} qui est proposée dans cet article et les mesures qui sont présentées dans la section suivante sont indépendantes des langues des textes étudiés.

3.4. *DefAcro fondée sur l'Information Mutuelle et l'Information Mutuelle au Cube*

De manière similaire à la formule (12), les formules (13) et (14) présentent respectivement la mesure DefAcro fondée cette fois-ci sur l'Information Mutuelle et l'Information Mutuelle au Cube.

$$\text{DefAcro}_{IM}(a^j) = \frac{nb((\bigcap_{i=1}^{n} a_i^j) + C)}{\prod_{i=1}^{n} nb(a_i^j + C; a_i^j \notin M_{outils})} \text{ où } n \geq 2 \quad [13]$$

$$\text{DefAcro}_{IM3}(a^j) = \frac{nb((\bigcap_{i=1}^{n} a_i^j) + C)^3}{\prod_{i=1}^{n} nb(a_i^j + C; a_i^j \notin M_{outils})} \text{ où } n \geq 2 \quad [14]$$

Après avoir proposé différentes mesures pour choisir des définitions d'acronymes adaptées, la section suivante propose l'expérimentation de DefAcro sur des données réelles.

4. Expérimentations

Les sections suivantes présentent le protocole expérimental mis en œuvre pour l'évaluation du système à partir d'un corpus acquis manuellement de taille raisonnable (section 4.1) et d'un corpus de grande dimension (section 4.2). Les expérimentations ont d'abord été menées avec l'acronyme JO qui est ambigu, d'où les problèmes rencontrés pour les tâches de RI. L'utilisation du contexte afin de lever les ambiguïtés liées aux définitions possibles de cet acronyme est discutée. Enfin, la section 4.3 présente une évaluation à partir d'acronymes ayant un nombre de mots différent et un choix de définitions variable.

L'application programmée en Perl possède différents paramètres qui sont : le nombre de mots à prendre en considération dans le contexte C, la liste des mots outils, les différentes mesures statistiques.

4.1. *Expérimentations à partir d'un corpus acquis manuellement*

Nous avons constitué manuellement un corpus de 100 pages possédant l'acronyme JO. Il est formé de 50 pages Web associées au "Journal Officiel" et 50 pages propres à la définition "Jeux Olympiques". Notons que cette proportion est motivée par le fait que les premières pages retournées par le moteur de recherche google sont réparties de manière semblable[4]. Ces pages obtenues à l'aide de diverses requêtes manuelles avec le moteur de recherche google ne contiennent aucune définition de ces acronymes[5]. La première tâche a consisté à nettoyer le corpus (enlever les balises HTML, supprimer les mots outils, supprimer les ponctuations et les divers caractères spéciaux, etc.).

Pour évaluer les différentes mesures proposées, nous construisons la matrice de confusion suivante :

		Réel	
		Journal Officiel	Jeux Olympiques
Prédiction	Journal Officiel	a	c
	Jeux Olympiques	b	d

4. Expériences effectuées avec les 50 premières pages retournées en février 2007 par le moteur de recherche google avec la requête "JO". Manuellement, nous avons évalué le fait que 10 pages sont propres au "Journal Officiel" et 10 pages sont relatives aux "Jeux Olympiques".
5. Utilisation de la rubrique "pages contenant aucun des mots suivants" de la recherche avancée.

où

– a : nombre de pages correctement prédites avec la définition "Journal Officiel",

– b : nombre de pages prédites avec la définition "Jeux Olympiques" mais dont la définition réelle est "Journal Officiel",

– c : nombre de pages prédites avec la définition "Journal Officiel" mais dont la définition réelle est "Jeux Olympiques",

– d : nombre de pages correctement prédites avec la définition "Jeux Olympiques".

La qualité du système peut alors être calculée en mesurant le taux d'erreur correspondant au nombre de pages mal classées dans la population sur le nombre de cas de la population : $TE = \frac{b+c}{a+b+c+d}$

Par exemple, sans utiliser le contexte avec DefAcro$_{Dice}$ (formule (12)), le meilleur score est toujours obtenu avec la définition "Jeux Olympiques" (voir section 3.2). Ceci a pour conséquence que toutes les pages sont classées dans la catégorie "Jeux Olympiques". Ainsi, nous avons un taux d'erreur de 50% (avec $b = d = 50$ et $a = c = 0$).

Dans la suite, nous proposons d'utiliser un contexte formé d'un à trois mots (mots les plus fréquents de chaque page qui ne sont pas des mots outils). La restriction à un contexte de trois mots maximum est motivée par le fait qu'un contexte de quatre mots ou plus ne retourne aucune page dans un nombre de cas non négligeable. Les résultats de nos expérimentations sont présentés dans les tableaux 1 et 2. Notons que ce jeu de test a nécessité 1800 requêtes à partir du moteur de recherche Exalead[6] (6 requêtes par page avec 3 jeux de test de 100 pages).

Les tableaux 1 et 2 montrent que les mesures qui donnent un résultat de bonne qualité sont l'Information Mutuelle au Cube et le coefficient de Dice. Par ailleurs, le fait d'utiliser un contexte avec davantage de mots a tendance à améliorer significativement les résultats. Ainsi, un contexte formé de trois mots donne un taux d'erreur faible avec l'Information Mutuelle au Cube et le coefficient de Dice (respectivement 8% et 9%). Notons que la grande majorité des erreurs de classement sont dues au fait que les mots les plus fréquents des pages Web sont non significatifs pour le domaine (mots assez généraux tels que "demain", "juillet", "produits", "France", etc.). Le nettoyage des pages HTML qui peut se révéler difficile dans certains cas peut également provoquer des erreurs dans la prédiction des définitions des acronymes.

Après avoir évalué les résultats obtenus à partir d'un corpus de taille raisonnable (100 textes), une expérience à plus grande échelle est présentée dans la section suivante (plus de 1300 textes).

4.2. *Expérimentations à partir d'un corpus de grande dimension*

Dans ces expérimentations, nous avons utilisé un corpus provenant du défi DEFT'06 (DÉfi Fouille de Textes). Ce deuxième défi francophone de fouille de textes

6. Expériences menées en décembre 2006

	Contexte d'1 mot	**Contexte de 2 mots**	**Contexte de 3 mots**
DefAcro$_{IM}$	47%	45%	42%
DefAcro$_{IM3}$	26%	14%	8%
DefAcro$_{Dice}$	29%	16%	9%

Tableau 1. *Taux d'erreur sur le corpus de 100 pages Web (acronyme JO).*

		Réel		**Réel**	
		Journal Officiel	Jeux Olympiques	Journal Officiel	Jeux Olympiques
INFORMATION MUTUELLE : **DefAcro**$_{IM}$		Contexte constitué d'**1 mot**		Contexte constitué de **2 mots**	
Prédiction	Journal Officiel	7	4	7	2
	Jeux Olympiques	43	46	43	48
		Contexte constitué de **3 mots**			
Prédiction	Journal Officiel	10	2		
	Jeux Olympiques	40	48		
INFORMATION MUTUELLE AU CUBE : **DefAcro**$_{IM3}$		Contexte constitué d'**1 mot**		Contexte constitué de **2 mots**	
Prédiction	Journal Officiel	31	7	45	9
	Jeux Olympiques	19	43	5	41
		Contexte constitué de **3 mots**			
Prédiction	Journal Officiel	48	6		
	Jeux Olympiques	2	44		
MESURE DE DICE : **DefAcro**$_{Dice}$		Contexte constitué d'**1 mot**		Contexte constitué de **2 mots**	
Prédiction	Journal Officiel	36	15	45	11
	Jeux Olympiques	14	35	5	39
		Contexte constitué de **3 mots**			
Prédiction	Journal Officiel	48	7		
	Jeux Olympiques	2	43		

Tableau 2. *Matrice de confusion avec différents contextes (acronyme JO).*

consistait à déterminer les segments thématiques de corpus écrits en français issus de domaines différents (politiques, juridiques, scientifiques). Dans nos expérimentations, nous nous sommes appuyés sur le corpus juridique propre à des articles de loi de l'Union Européenne[7]. Les 1303 articles (11 Mo) possédant l'acronyme JO sont pris en compte.

7. Corpus disponible à l'adresse suivante : http ://www.lri.fr/ia/fdt/DEFT06/corpus/donnees.html

Cet acronyme est généralement utilisé dans ce corpus pour faire référence à un ou des articles précis du Journal Officiel (par exemple, les références "JO 308 du 18.12.1967" ou "JO no L 249 du 8.9.1988" pour lesquelles l'acronyme JO n'est pas défini). Pour chacun des articles de loi, en utilisant DefAcro, nous mesurons si l'acronyme JO doit être associé à la définition "Journal Officiel" sans prendre en compte les spécificités précédemment décrites du corpus. Le tableau 3 présente les taux d'erreur obtenus à partir de ce corpus avec différents contextes (de un à trois mots). Notons que dans ces expérimentations, il est nécessaire d'exécuter 23454 requêtes : 1303 articles de loi et 6 requêtes par article avec 3 jeux de test (contextes de un à trois mots).

	Nombre d'acronymes correctement associés	TE
	Contexte d'1 mot	
DefAcro$_{IM}$	190	85.4%
DefAcro$_{IM3}$	1040	20.2%
DefAcro$_{Dice}$	842	35.4%
	Contexte de 2 mots	
DefAcro$_{IM}$	434	66.7%
DefAcro$_{IM3}$	1234	5.3%
DefAcro$_{Dice}$	1200	7.9%
	Contexte de 3 mots	
DefAcro$_{IM}$	650	50.1%
DefAcro$_{IM3}$	1281	1.7%
DefAcro$_{Dice}$	1274	2.2%

Tableau 3. *Taux d'erreurs sur le corpus juridique de DEFT'06.*

Le tableau 3 montre que notre méthode donne des résultats de très bonne qualité avec le corpus de DEFT'06, plus particulièrement dans le cas d'un contexte plus riche c'est-à-dire constitué de deux ou trois mots. Nous pouvons confirmer que le coefficient de Dice et l'Information Mutuelle au Cube donnent un taux d'erreur faible respectivement de 2.2% et 1.7% avec un contexte de trois mots.

Le fait que l'Information Mutuelle au Cube et le coefficient de Dice donnent des résultats de bonne qualité à partir des deux corpus étudiés s'explique par le fait ces mesures privilégient les co-occurrences fréquentes. Dans notre cas, le nombre de pages Web partageant la définition d'un acronyme associée à un contexte pertinent est important. Ceci a pour conséquence d'accorder un score élevé à ces mesures qui sont relatives à un grand nombre de pages.

Notons que dans le cas où le nombre de pages retourné pour différentes définitions est du même ordre, la fréquence n'est pas toujours un critère pertinent. Dans ce cas, les mesures statistiques peuvent se révéler plus adaptées. Par exemple, avec un contexte de trois mots relatif à un des documents sur les Jeux Olympiques issu du corpus étudié en section 4.1, 70 et 55 pages sont respectivement retournées pour le thème "Journal Officiel" et "Jeux Olympiques". L'Information Mutuelle au Cube accorde quant à elle un score plus élevé au deuxième thème qui doit être associé à la page en question (score de 0.054 pour "Journal Officiel" et 0.088 pour "Jeux Olympiques").

Précisons enfin que les résultats présentés dans cette section sont de meilleure qualité comparativement aux expériences précédentes (corpus de 100 textes). Ceci pourrait s'expliquer par le fait que le corpus de DEFT'06 est beaucoup plus spécifique et ainsi les mots les plus fréquents constituant les contextes sont plus adaptés au domaine juridique. Par exemple, certaines pages Web issues du corpus expérimenté dans la section précédente (section 4.1) peuvent se révéler plus ambiguës (par exemple, les textes traitant des conséquences économiques de l'attribution des Jeux Olympiques).

4.3. *Expérimentations avec différents couples acronymes/définitions*

Dans les expérimentations qui suivent, nous nous sommes intéressés à l'étude des acronymes relatifs aux principaux partis politiques français. L'objectif est d'étudier les différentes mesures de qualité avec un nombre variable de définitions proposées. Par ailleurs, les acronymes peuvent être composés de plusieurs mots. Dans ces expérimentations, nous avons, dans un premier temps, relevé les différentes définitions proposées pour les acronymes LCR, PCF, PS, UDF, UMP, FN dans le site http ://www.sigles.net/. Ces définitions sont données dans le tableau 4[8].

Acronymes politiques	Définitions	
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	

Tableau 4. *Définitions des acronymes. Les définitions en gras sont relatives aux partis politiques.*

Par la suite, nous avons effectué des requêtes via le moteur de recherche google avec chacun de ces acronymes (pages ne comportant pas les mots issus de la définition des acronymes). Pour chacun des acronymes, nous avons extrait manuellement les premiers sites propres aux partis politiques (une dizaine par acronyme). Le corpus constitué de dix pages par acronyme a une taille de 500 Ko après nettoyage (suppression des balises HTML). Nous avons alors calculé le taux d'erreur de la méthode sur ce jeu de test afin de mesurer le nombre de pages qui ne sont pas associées à la définition du domaine politique. Les résultats que nous avons obtenus confirment les taux d'erreur faibles établis dans les sections précédentes même avec un nombre de mots du contexte réduit. À titre d'exemple, avec un contexte d'un seul mot le taux d'erreur est de moins de 4% avec DefAcro$_{IM3}$.

8. Les acronymes UDF et UMP ne possédant qu'une seule définition, nous avons ajouté des définitions réelles non présentes dans le site http ://www.sigles.net/.

5. Conclusion et perspectives

La mesure DefAcro proposée dans cet article a pour objectif de déterminer automatiquement la définition pertinente d'un acronyme présent dans une page Web. La méthode utilise les statistiques issues du Web pour sélectionner la définition adaptée. Les premiers résultats obtenus sont tout à fait satisfaisants car la définition pertinente de l'acronyme est retrouvée dans 92 à 98% des cas (avec un contexte de trois mots).

Ces résultats ont été obtenus en mettant en place une mesure de qualité qui s'appuie sur des critères statistiques et sur la prise en compte du contexte. Cependant, la majorité des erreurs proviennent de l'utilisation de mots généraux pour représenter les contextes. Ainsi, une perspective intéressante serait de déterminer des descripteurs (noms propres, termes, etc.) davantage représentatifs pour le domaine. À titre d'exemple, le fait d'utiliser le nom propre très spécifique "Beijing" dans le contexte se révèle très pertinent pour retrouver les pages des Jeux Olympiques (pour caractériser les Jeux Olympiques en Chine en 2008). De plus, l'utilisation des mots les plus fréquents pour représenter les contextes, comme cela est illustré dans cet article, pourrait être améliorée de deux manières. Premièrement, les contextes pourraient s'appuyer sur les mots les plus fréquents dans une fenêtre donnée centrée autour de l'acronyme. Deuxièmement, nous pouvons nous appuyer sur des critères numériques plus élaborés, en particulier la mesure TF-IDF pour caractériser les contextes.

Notre approche a des limites pour les documents de taille réduite rendant la construction d'un contexte pour DefAcro difficile. Une autre perspective intéressante consiste à représenter les documents sous forme de vecteurs sémantiques (Chauché, 1990) afin d'avoir des informations sur la thématique des textes. Cette information supplémentaire associée à la mesure DefAcro permettrait alors d'aider la prédiction des définitions pertinentes propres aux textes courts.

Notre article s'appuie sur des mesures statistiques de base (Information Mutuelle, Information Mutuelle au Cube, Dice). Nos futurs travaux peuvent s'orienter vers l'étude de critères statistiques plus riches pour la mise en œuvre de la mesure de qualité DefAcro (Daille, 1994, Azé, 2003, Lallich *et al.*, 2004, Roche, 2004).

6. Bibliographie

Audibert L., « Étude des critères de désambiguïsation sémantique automatique : résultats sur les cooccurrences », *Actes de la conférence TALN*, p. 33-44, 2003.

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

Bourigault D., Jacquemin C., « Term Extraction + Term Clustering : An Integrated Platform for Computer-Aided Terminology », *Proceedings of the European Chapter of the Association for Computational Linguistics*, p. 15-22, 1999.

Brill E., « Some Advances in Transformation-Based Part of Speech Tagging », *AAAI, Vol. 1*, p. 722-727, 1994.

Chang J., Schütze H., Altman R., « Creating an Online Dictionary of Abbreviations from MEDLINE », *Journal of the American Medical Informatics Association*, vol. 9, p. 612-620, 2002.

Chauché J., « Détermination sémantique en analyse structurelle : une expérience basée sur une définition de distance », *TA Information*, vol. 1/1, p. 17-24, 1990.

Church K. W., Hanks P., « Word Association Norms, Mutual Information, and Lexicography », *Computational Linguistics*, vol. 16, p. 22-29, 1990.

Claveau V., Sébillot P., « Apprentissage symbolique pour l'acquisition de ressources linguistiques », *Actes de l'atelier « Acquisition, apprentissage et exploitation de connaissances sémantiques pour l'accès au contenu textuel » de la plateforme AFIA*, 2003.

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

Jacquemin C., « Variation terminologique : Reconnaissance et acquisition automatiques de termes et de leurs variantes en corpus », *Mémoire d'Habilitation à Diriger des Recherches en informatique fondamentale, Université de Nantes*, 1997.

Lallich S., Teytaud O., « Évaluation et validation des règles d'association », *Numéro spécial "Mesures de qualité pour la fouille des données", Revue des Nouvelles Technologies de l'Information (RNTI)*, vol. RNTI-E-1, p. 193-218, 2004.

Landauer T. K., Foltz P. W., Laham D., « Introduction to Latent Semantic Analysis », *Discourse Processes*, vol. 25, p. 259-284, 1998.

Larkey L. S., Ogilvie P., Price M. A., Tamilio B., « Acrophile : An automated Acronym Extractor and Server », *Proceedings of the Fifth ACM International Conference on Digital Libraries*, p. 205-214, 2000.

Petrovic S., Snajder J., Dalbelo-Basic B., Kolar M., « Comparison of collocation extraction measures for document indexing », *Proc of Information Technology Interfaces (ITI)*, p. 451-456, 2006.

Roche M., Intégration de la construction de la terminologie de domaines spécialisés dans un processus global de fouille de textes, Thèse de Doctorat, Univ. de Paris 11, Déc., 2004.

Roche M., Kodratoff Y., « Pruning Terminology Extracted from a Specialized Corpus for CV Ontology Acquisition », *Proceedings of onToContent Workshop - OTM'06, Springer Verlag, LNCS*, p. 1107-1116, 2006.

Smadja F., McKeown K. R., Hatzivassiloglou V., « Translating collocations for bilingual lexicons : A statistical approach », *Computational Linguistics*, vol. 22, n° 1, p. 1-38, 1996.

Thanopoulos A., Fakotakis N., Kokkianakis G., « Comparative Evaluation of Collocation Extraction Metrics », *Proceedings of LREC'02*, vol. 2, p. 620-625, 2002.

Turney P., « Mining the Web for Synonyms : PMI–IR versus LSA on TOEFL », *Lecture Notes in Computer Science*, vol. 2167, p. 491-502, 2001.

Vivaldi J., Màrquez L., Rodríguez H., « Improving Term Extraction by System Combination Using Boosting », *Proceedings of ECML*, p. 515-526, 2001.

Yeates S., « Automatic Extraction of Acronyms from Text », *New Zealand Computer Science Research Students' Conference*, p. 117-124, 1999.