Processus global d'acquisition et de gestion des sigles

Mathieu Roche, Violaine Prince

Processus global d'acquisition et de gestion des sigles

Mathieu Roche, Violaine Prince

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

RÉSUMÉ. *Cet article traite d'une approche d'extraction des sigles et leur(s) définition(s) à partir de données textuelles (corpus) puis de désambiguïser les définitions (ou expansions). Les deux étapes propres à notre processus global d'acquisition et de gestion des sigles sont précisément décrites. La première étape consiste à utiliser des marqueurs tels que les parenthèses pour identifier les candidats à l'expansion. L'alignement des lettres permet de sélectionner les couples sigle/définition. La seconde étape consiste à déterminer l'expansion pertinente d'un sigle dans un contexte donné. Notre méthode s'appuie sur des mesures statistiques (Information Mutuelle, Information Mutuelle au Cube, Mesure de Dice) et sur les résultats fournis par des moteurs de recherche. Cet article présente une évaluation du processus global à partir de données réelles (domaine général et spécialisé).*

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 domain).*

MOTS-CLÉS : *Sigle/Définition, Désambiguïsation, Mesures Statistiques, Terminologie*

KEYWORDS: *Acronym/Definition, Word Sense Disambiguation, Statistical Measures, Terminology*

1. Introduction

L'étude des entités nommées est une tâche utile pour de nombreuses applications en fouille de textes telles que la recherche et/ou l'extraction d'informations. Dans cet article, nous nous intéressons à une entité nommée spécifique appelée "sigle". Un sigle est un ensemble de lettres initiales servant d'abréviation, par exemple le sigle SIG peut être associé à la définition (aussi appelée expansion) "Système d'Information Géographique". Cette forme réduite des entités nommées est utile lorsque celles-ci se répètent de manière très fréquente dans les textes.

Avec la masse de données numériques aujourd'hui disponibles en différentes langues, les sigles sont très utiles et très présents aussi bien dans des textes de thème général (par exemple, SNCF, ANPE, etc) ou spécialisés (par exemple, TAL, IA, etc).

Le problème qui se pose tient au fait qu'un même sigle peut posséder plusieurs sens (problème lié à la polysémie). À titre d'exemple SIG qui peut signifier "Service d'Information du Gouvernement", "Services Industriels de Genève", "Solde Intermédiaire de Gestion", "Système d'Information Géographique". Chacune des définitions appartient à un domaine particulier (politique, industrie, banque, informatique). Notons que les domaines spécialisés tels que la médecine ou la biologie utilisent de très nombreux sigles (voir les exemples présentés dans la table 1).

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

Tableau 1. *Exemple de définitions du sigle ABCD issu du domaine biomédical.*

Précisons que les sigles issus d'un domaine général ne sont pas nécessairement adaptés pour un domaine spécialisé. À titre, d'exemple le sigle SIG issu d'un domaine biomédical a une signification extrêmement différente comparativement aux expansions précédemment proposées pour ce sigle : "strong ion gap", "small inducible gene", etc. C'est la raison pour laquelle, lors d'un traitement d'un domaine spécialisé, il est nécessaire de construire des dictionnaires spécifiques comme nous allons le décrire ci-dessous. Nos travaux issus du projet ProSigles[1] développé par le LIRMM possèdent deux phases distinctes qui sont résumées ci-dessous.

1) Les sigles et leur(s) définition(s) sont tout d'abord extraits à partir de données textuelles quelconques (domaines spécialisés ou non, différentes langues, etc). Une telle phase permet d'acquérir ou d'enrichir des dictionnaires plus ou moins spécialisés. La méthode mise en œuvre possède deux étapes qui sont détaillées dans la section 3 de cet article : extraction des sigles candidats et filtrage de ces derniers.

1. Projet financé par le Conseil Scientifique de l'Université Montpellier 2, France

2) Une fois ces dictionnaires constitués, nous avons proposé une mesure de qualité dans (Roche *et al.*, 2007)[2] qui consiste à déterminer la définition pertinente d'un sigle présent dans un document. Dans ces documents, la définition n'est cependant pas présente d'où la difficulté du traitement. Dans ce contexte, il est donc essentiel d'avoir à disposition un dictionnaire adapté, ce qui justifie la première phase du processus présenté. Notre approche décrite dans la section 4 utilise les statistiques issues du Web pour sélectionner la définition adaptée. Nos travaux consistent à mettre en place une mesure de qualité qui s'appuie sur des critères statistiques et sur la prise en compte du contexte.

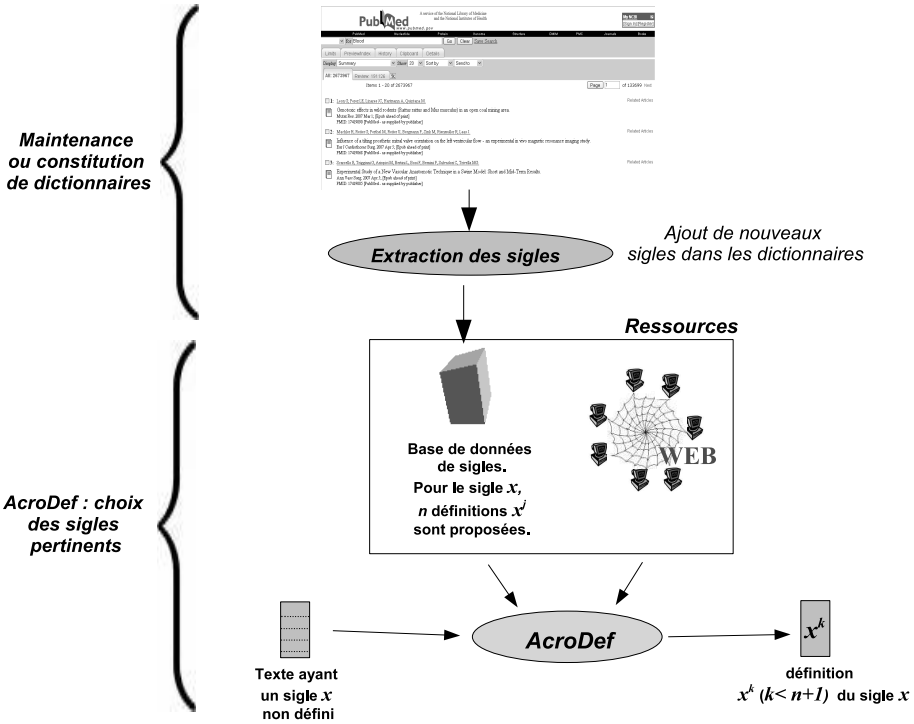

Figure 1. *Processus global d'acquisition et de gestion des sigles.*

La figure 1 résume l'ensemble de l'approche. Chacune des phases sera précisément décrite et évaluée dans cet article. Dans un premier temps, la section 2 résume l'état de l'art relatif à l'extraction des sigles/définitions. Les sections suivantes décrivent les deux phases successives de notre processus global de recherche (section 3) et de désambiguïsation (section 4) des sigles.

2. Notons qu'une évaluation plus complète de cette mesure est proposée dans cet article.

2. Acquisition de sigles/définitions à partir de données textuelles : état de l'art

De nombreuses méthodes pour extraire les sigles et leur(s) définition(s) ont été développées (Yeates, 1999, Larkey *et al.*, 2000, Chang *et al.*, 2002, Okazaki *et al.*, 2006, Xu *et al.*, 2007). La plupart des approches de détection de sigles dans les textes s'appuient sur l'utilisation de marqueurs spécifiques. La méthode développée dans (Yeates, 1999) consiste dans un premier temps à séparer les phrases par fragments en utilisant des marqueurs spécifiques (parenthèses, points, etc) comme frontières. L'étape suivante a pour but de comparer chaque mot de chacun des fragments avec les fragments précédents et suivants. Ensuite, les couples sigles/définitions sont testés. Les candidats sigles sont retenus si les lettres des sigles sont mises en correspondance avec les premières lettres des définitions potentielles. Dans notre cas, le couple "SIG / Système d'Information Géographique" est un candidat sigle. De plus, des heuristiques spécifiques pour retenir les candidats pertinents peuvent être appliquées. Ces heuristiques s'appuient sur le fait que les sigles ont une taille plus petite que leur définition, ils sont en majuscule, les définitions des sigles ayant une longueur importante ont tendance à posséder davantage de mots outils (par exemple, les articles et les prépositions), etc. De nombreuses approches (Chang *et al.*, 2002, Larkey *et al.*, 2000) utilisent des méthodes similaires fondées sur la présence de marqueurs associés à des heuristiques spécifiques.

L'utilisation des marqueurs (le plus souvent, les parenthèses) représente la base de la plupart des méthodes d'acquisition de dictionnaires sigles/définitions. Cependant, les traitements qui suivent peuvent se révéler singulièrement différents. Par exemple, certains travaux récents (Okazaki *et al.*, 2006) consistent à utiliser des mesures statistiques issues du domaine de l'extraction de la terminologie pour classer les expansions (mesures fondées sur l'approche C-value adaptée aux données biomédicales). Contrairement à notre approche qui utilise le résultat de moteurs de recherche, la mesure proposée par (Okazaki *et al.*, 2006) s'appuie sur la fréquence des termes présents dans des corpus. L'avantage de cette mesure est d'être indépendante de l'alignement des caractères car dans de nombreux cas, les couples sigles/définitions pertinents ne peuvent être mis en correspondance (par exemple lorsque les lettres du sigle sont dans un ordre différent de celles de la définition : "AW/water activity").

D'autres approches s'appuient sur des techniques d'apprentissage supervisé pour sélectionner les expansions pertinentes (Xu *et al.*, 2007). Dans ces travaux, les auteurs utilisent les approches SVM (Support Vector Machine) avec des descripteurs propres aux sigles et expansions (leur longueur, présence de caractères spéciaux, etc) et au contexte. Notons enfin que les travaux de (Torii *et al.*, 2007) présentent une étude comparative des principales approches (notamment les méthodes d'apprentissage supervisé et les approches à base de règles) en associant des connaissances du domaine à partir de données textuelles du domaine biomédical.

3. Acquisition d'un dictionnaire sigles/définitions

L'approche d'acquisition de dictionnaires de sigles que nous proposons se compose de deux étapes successives qui sont détaillées dans les sections suivantes. Une interface graphique (Matviico *et al.*, 2008) développée en Java permet de visualiser les résultats obtenus au cours de ces deux étapes.

3.1. *Étape 1 : Extraction des candidats sigles/définitions*

De manière similaire aux travaux décrits dans la section précédente, notre méthode consistant à extraire les candidats sigles/définitions utilise des marqueurs (parenthèses, crochets, etc). Les marqueurs visent à identifier soit un sigle, soit une définition, ce qui nécessite la prise en compte de deux traitements différents :

Premier cas : le sigle se situe avant la définition qui se trouve entre les marqueurs (les parenthèses dans le cas le plus courant). Exemple : "`... S.I.G. (Solde Intermédiaire de Gestion)...`"

Deuxième cas : la définition se trouve avant le sigle qui se situe entre les marqueurs. Exemple : "`... les Systèmes d'Informations Géographiques (SIG)` ...". Dans ce cas, la taille de la définition est pour le moment indéterminable. C'est pourquoi, il est nécessaire de la définir arbitrairement en fonction du nombre de lettres composant le sigle. Ce choix doit tenir compte des mots qui pourraient venir s'intercaler sans pour autant nous intéresser, comme par exemple les prépositions ou les articles. Nous avons expérimentalement fixé cette taille à trois fois le nombre de lettres composant le sigle. Par exemple, la définition potentielle choisie pour l'exemple "SIG" sera composée des neuf mots qui précèdent ce sigle.

Le but de cette première étape consiste à extraire tous les candidats sigles/définitions pertinents. Bien entendu, cette phase retourne une quantité importante de bruit car cette première étape s'appuie seulement sur les marqueurs pour identifier un candidat potentiel. Ainsi, la seconde étape de notre approche consiste à filtrer les couples sigles/définitions pertinents parmi la liste retournée lors de cette première étape.

3.2. *Étape 2 : Filtrage des candidats*

Cette seconde étape de notre application utilise donc les résultats obtenus lors du premier traitement qui vient d'être décrit. Les résultats sont triés afin de supprimer les paires sigles/définitions non pertinentes et d'extraire précisément les définitions présentes dans les définitions potentielles (ces dernières pouvant être trop longues puisque coupées arbitrairement lors du second cas de la recherche des candidats).

Pour permettre un tel filtrage, nous effectuons un alignement des lettres contenues dans le sigle avec les mots de la définition. Cet alignement consiste à vérifier la correspondance entre les lettres des sigles avec les premières lettres de chacun des mots des

définitions. Dans notre méthode, si le premier caractère des mots de la définition candidate ne peut être aligné, les caractères qui suivent au sein des mots sont considérés. Par exemple, cette méthode permet de reconnaître "Extraction Itérative de la Terminologie" comme la définition du sigle EXIT dans lequel la lettre "X" a pu être alignée. Notons enfin que les mots outils (prépositions, articles, etc) sont considérés comme des mots quelconques sans traitement spécifique contrairement à certains travaux qui utilisent une telle liste dans le processus (Larkey *et al.*, 2000). Ceci a pour but d'avoir une méthode indépendante des langues et de considérer ces mots qui permettent de constituer un sigle (par exemple, la préposition "de" pour "GDF / Gaz de France").

Nous présentons dans la table 2 une évaluation de notre système d'alignement des sigles avec les définitions candidates. Pour cette évaluation, nous nous appuyons sur les données issues du site http ://www.sigles.net/ proposant 25463 sigles et leurs définitions issus de 17 langues. L'évaluation consiste à extraire aléatoirement de ce site des sigles de 2, 3 et 4 caractères et d'évaluer le taux de réussite de l'alignement (nombre de sigles alignés avec les définitions du site en utilisant la version actuelle de notre logiciel). Le tableau 2 présente les résultats de 800 cas de mise en correspondance qui se sont révélées globalement très satisfaisants (taux de réussite de 78% à 98%). Par ailleurs, ce tableau montre que les sigles longs sont plus difficiles à aligner. Ceci est par exemple dû à la présence de lettres en majuscule accentuées qui ne sont pas encore considérées par notre logiciel. De telles améliorations techniques sont assez aisées à mettre en œuvre. Cependant, notons que de nombreux cas particuliers plus difficiles à traiter peuvent exister comme l'alignement de caractères numériques / non numériques (par exemple, "3D / Trois Dimensions", "ST2I / Sciences et Techniques de l'Informatique et de l'Ingénierie").

Nb de lettres	**Nb de sigles**	**Nb de définitions**	**Nb de définitions non reconnues**	**Pourcentage de réussite**
2	100	616	11	98.2 %
3	50	157	10	93.6 %
4	20	32	7	78.1 %

Tableau 2. *Taux de réussite de l'alignement sigles/définitions.*

Le résultat des étapes de notre approche d'extraction des sigles associés à leur(s) définition(s) est illustré dans la figure 2.

La section suivante présente une première évaluation des deux étapes de notre approche d'extraction des sigles et de leur(s) définition(s).

3.3. *Évaluation de l'extraction des sigles*

Nous avons analysé les résultats obtenus à partir d'un corpus quelconque de taille raisonnable permettant d'effectuer une évaluation manuelle (7465 mots). Après application de la première étape de notre approche, près de 100 couples sigles/définitions

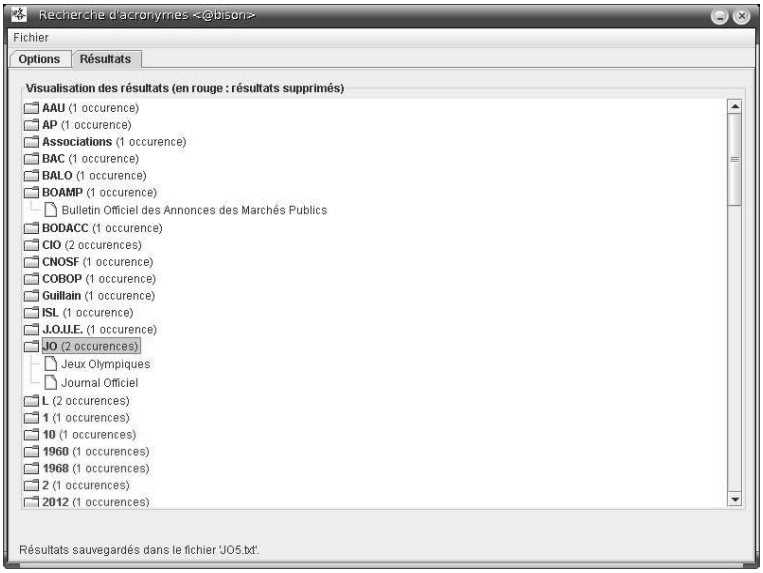

Figure 2. *Visualisation des sigles extraits (sigles/définitions retenus ou non après les deux étapes du processus)*

ont été retournés. Bien entendu, de nombreux couples sont erronés (bruit important). Cependant, tous les couples corrects ont pu être extraits (absence de silence).

Pour évaluer nos résultats, nous allons nous appuyer sur les mesures d'évaluation classiques issues de la fouille de données : précision, rappel et F-mesure représentée par la moyenne harmonique entre la précision et le rappel (formules ci-dessous).

$$\text{Précision} = \frac{\text{nombre de couples corrects retournés}}{\text{nombre de couples retournés}} \quad [1]$$

$$\text{Rappel} = \frac{\text{nombre de couples corrects retournés}}{\text{nombre de couples corrects}} \quad [2]$$

$$\text{F-mesure} = \frac{2 \times \text{Précision} \times \text{Rappel}}{\text{Précision+Rappel}} \quad [3]$$

La table 3 qui présente les résultats obtenus avec les deux étapes de notre approche montre que la F-mesure de la première étape retourne un résultat de qualité globale faible (faible F-mesure). Ceci est justifié par le fait que cette première étape a pour but de restituer l'ensemble des couples sigles/définitions pertinents même si cette liste est fortement bruitée. Ainsi, le rappel de la première étape est de 100% alors que la précision possède une valeur extrêmement faible. Le but de la seconde étape de notre approche consiste à filtrer les candidats et donc à réduire le bruit de la liste fournie après l'étape 1 du processus. Ainsi, après l'application des deux étapes (étape 1 +

étape 2, voir table 3), nous obtenons un rappel qui reste élevé (80%). La précision est relativement importante également, ainsi la valeur de la F-mesure est assez élevée (72.7%).

	Précision	**Rappel**	**F-mesure**
Extraction des couples (étape 1)	15.2%	100%	26.4%
Extraction des couples + Filtrage (étape 1 + étape 2)	66.7%	80%	72.7%

Tableau 3. *Évaluation des étapes du système d'extraction des sigles/définitions.*

4. Choix des sigles pertinents

4.1. *Motivations*

La seconde phase de notre approche (voir figure 1) qui utilise les dictionnaires sigles/définitions acquis lors de la première phase consiste à désambiguïser les sigles. Le but est alors de sélectionner les définitions les plus pertinentes dans un contexte donné (domaine). La définition adaptée d'un sigle qui n'est pas défini dans un document peut alors être retrouvée par notre approche. Ceci peut être utile pour les tâches de classification de textes. Une deuxième tâche appropriée à notre approche serait d'enrichir des requêtes de domaines généraux ou spécialisés dans le cadre de la recherche documentaire. Par exemple en biologie, un utilisateur pourrait effectuer une requête avec le sigle "TU" dans une base de données bibliographique spécialisée telle que Medline. Plusieurs définitions sont possibles pour ce sigle[3]. Ainsi, en déterminant la définition adaptée, notre méthode permettrait d'améliorer significativement la recherche documentaire par l'expansion de la requête originale. Cette expansion pourrait par exemple être une disjonction (opérateur "OR") du sigle et de sa définition afin de retourner un nombre de documents plus important (amélioration du rappel). La conjonction du sigle et de la définition (opérateur "AND") permettrait quant à elle d'obtenir des documents plus pertinents (amélioration de la précision).

Plaçons nous dans le contexte dans lequel aucune définition de sigle n'est présente. Le but est alors de choisir, de manière automatique, la définition adaptée. Dans ce contexte, posons un sigle x donné (par exemple, JO) dont la définition n'est pas présente dans un document d. Considérons que nous avons une liste de définitions possibles pour le sigle x (par exemple, "Jeux Olympiques", "Journal Officiel"). L'objectif de notre approche est de déterminer la définition x^k (par exemple, x^1=Jeux Olympiques ou x^2=Journal Officiel) pertinente pour le document d. Pour effectuer un

3. Définitions données par le logiciel Acromine (http ://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

tel choix, nous proposons une mesure de qualité, $AcroDef$, qui s'appuie notamment sur les ressources du Web.

4.2. *La mesure* $AcroDef$

$AcroDef$ est une mesure de qualité fondée sur des approches statistiques comme l'Information Mutuelle au Cube (Daille, 1994). Cette dernière est une mesure qui calcule une certaine forme de dépendance des mots $x_1, ..., x_n$ constituant les définitions des sigles. Une telle mesure est définie par la formule (4).

$$IM3 \quad = \quad \frac{nb(x_1, ..., x_n)^3}{nb(x_1) \times ... \times nb(x_n)} \qquad [4]$$

Le numérateur au cube de la formule $IM3$ permet de privilégier les co-occurrences fréquentes (mots présents ensemble afin de constituer la définition) comparativement à l'Information Mutuelle d'origine (IM) proposée par (Church *et al.*, 1990). L'Information Mutuelle au Cube est utilisée dans bon nombre de travaux liés à l'extraction des termes (Vivaldi *et al.*, 2001) ou des entités nommées (Downey *et al.*, 2007) dans les textes. (Vivaldi *et al.*, 2001) ont d'ailleurs estimé que l'Information Mutuelle au Cube était la mesure qui avait le meilleur comportement. Nos expérimentations avec d'autres mesures statistiques seront présentées plus loin.

La mesure $AcroDef$ (formule (5)) s'appuie sur l'Information Mutuelle au Cube ($IM3$) mais également sur les statistiques issues du Web afin de calculer un score pour chaque définition x^j.

$$AcroDef\text{-IM3}(x^j) = \frac{nb((\bigcap_{i=1}^{n} x_i^j) + C)^3}{\prod_{i=1}^{n} nb(x_i^j + C; x_i^j \notin M_{outils})} \text{ où } n \geq 2 \qquad [5]$$

M_{outils} représente une liste de mots outils (articles, prépositions, etc)

La mesure $AcroDef$-IM3 calcule la dépendance des mots constituant la définition de manière similaire aux travaux relatifs à la terminologie (Daille, 1994, Petrovic *et al.*, 2006, Roche, 2004, Vivaldi *et al.*, 2001, Downey *et al.*, 2007). Cette dépendance est calculée par rapport aux données retournées par le Web. Ainsi, nb est le nombre de pages retourné avec les n mots x_i^j ($i \in [1, n]$) propres à la définition x^j en utilisant un moteur de recherche. Nous utiliserons le moteur de recherche Exalead[4] pour le calcul de la mesure car le corpus d'évaluation est constitué à l'aide de requêtes effectuées avec le moteur de recherche Google[5]. $\bigcap_{i=1}^{n} x_i^j$ désigne la suite des mots x_i^j ($i \in [1, n]$) que l'on considère comme une chaîne de caractères. Pour cela, nous utilisons des guillemets avec le moteur de recherche Exalead : "$x_1^j ... x_n^j$". À titre d'exemple, nb(Jeux $\cap$ Olympiques) correspond au nombre de pages retourné avec la requête `"Jeux Olympiques"`.

4. http ://www.exalead.fr/
5. http ://www.google.fr/

Par ailleurs, la dépendance des mots issus de la définition est calculée relativement aux mots partageant le même contexte. Pour cela, nous utilisons le contexte C représenté par les mots les plus fréquents de la page contenant le sigle à définir. Dans ce cas, $x_i^j + C$ représente le mot x_i^j avec tous les mots du contexte C. De manière concrète, $nb(x_i^j + C)$ retourne le nombre de pages donné par le moteur de recherche avec la requête $x_i^j + C$ (utilisation de l'opérateur AND d'Exalead entre le mot x_i^j et les mots du contexte C). Par exemple, nb((Jeux $\cap$ Olympiques) + sport + natation) désigne le nombre de pages retourné avec la requête `"Jeux Olympiques" AND sport AND natation`. Le contexte C de cet exemple est constitué de deux mots (sport et natation). Ces mots sont les plus fréquents qui ne sont pas des mots outils (articles, prépositions, etc) dans la page où le sigle à définir est présent.

Dans cette première approche, la définition du contexte est fondée sur les mots les plus fréquents. Notons que nos futurs travaux pourront s'appuyer sur des contextes plus riches et donc plus pertinents utilisant notamment des informations linguistiques (lexicales, syntaxiques, etc) comme dans les travaux relatifs à la désambiguïsation sémantique (Audibert, 2003).

Notre approche a des similarités avec les travaux de (Larkey *et al.*, 2000) concernant l'utilisation du Web. Cependant, dans cette phase de notre approche, nous ne recherchons pas les définitions des sigles dans les textes car nous nous intéressons à déterminer les définitions des sigles qui sont absents des textes. Notre approche a davantage de similarités avec les travaux de Peter Turney (Turney, 2001) qui ne s'intéressent pas spécifiquement à la recherche des sigles mais qui utilisent le Web pour établir une fonction de rang. En effet, 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.

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 sigles peuvent se révéler pertinentes. Ainsi, nous avons décidé de ne pas mesurer la dépendance entre les sigles 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. Ainsi, nos travaux s'appuient sur d'autres mesures statistiques telles que l'Informations Mutuelle au Cube et la mesure de Dice (décrite dans la section suivante). Par ailleurs, l'utilisation d'un contexte spécifique permet d'améliorer significativement les mesures de base.

4.3. *D'autres mesures statistiques*

Soulignons que les mesures de qualité fondées sur l'Information Mutuelle sont simples et efficaces car elles nécessitent peu d'informations. En effet, ces mesures s'appuient sur un nombre d'exemples (dans notre cas, le nombre de pages retournées

avec les mots des définitions) sans nécessité de déterminer les contre-exemples[6]. En effet, ces derniers sont souvent plus complexes à déterminer dans le cadre d'approches non supervisées sur la base de données statistiques issues du Web.

Deux autres mesures associées à $AcroDef$ fondées sur le nombre d'exemples sont l'Informations Mutuelle qui est donnée par la formule (6) et le coefficient de Dice qui est décrit plus loin. Ces mesures seront évaluées dans la section 5 de cet article.

$$AcroDef\text{-IM}(x^j) = \frac{nb((\bigcap_{i=1}^{n} x_i^j) + C)}{\prod_{i=1}^{n} nb(x_i^j + C; x_i^j \notin M_{outils})} \text{ où } n \geq 2 \quad [6]$$

La formule $AcroDef$ fondée sur le coefficient de Dice que nous allons brièvement présenter ci-dessous est décrite de manière précise dans (Roche *et al.*, 2007). $AcroDef$ s'appuie sur la mesure de Dice étendue à n éléments (formule (7)) :

$$Dice(x_1, ..., x_n) \quad = \quad \frac{n \times nb(x_1, ..., x_n)}{nb(x_1) + ... + nb(x_n)} \quad [7]$$

Ainsi, nous pouvons nous appuyer sur une telle mesure de qualité pour représenter $AcroDef$-Dice par la formule (8) :

$$AcroDef\text{-Dice}(a^j) = \frac{\left|\{a_i^j + C; a_i^j \notin M_{outils}\}_{i\in[1,n]}\right| \times nb((\bigcap_{i=1}^{n} a_i^j) + C)}{\sum_{i=1}^{n} nb(a_i^j + C; a_i^j \notin M_{outils})} \quad [8]$$

où $n \geq 2$

et $\left|.\right|$ représente le nombre de mots propres à la définition

Dans la figure 2 issue de la première phase du processus global d'acquisition et de gestion des sigles, nous pouvons remarquer que deux définitions ont été identifiées pour le sigle JO. Nous allons dans un premier temps nous appuyer sur ce cas pour évaluer la mesure de qualité $AcroDef$ dans la section suivante.

5. Expérimentations

5.1. *Évaluation de la mesure* $AcroDef$ *sur un domaine général*

Dans le cadre de notre protocole expérimental, nous avons constitué manuellement un corpus en français de 100 pages possédant le sigle 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

6. utiles pour de nombreuses mesures de qualité : Rapport de Vraisemblance (Dunning, 1993), Conviction (Brin *et al.*, 1997), J-mesure (Goodman *et al.*, 1988), Moindre Contradiction (Azé, 2003), etc.

par le moteur de recherche Google sont réparties de manière semblable[7]. Ces pages obtenues à l'aide de diverses requêtes manuelles avec le moteur de recherche Google ne contiennent aucune définition de ces sigles. La première tâche a consisté à nettoyer le corpus (suppression des balises HTML, des mots outils, des ponctuations et divers caractères spéciaux, etc).

La qualité du système peut alors être évaluée en nous appuyant sur les mesures de précision, rappel et F-mesure. La précision et le rappel sont calculés pour chaque classe (classe correspondant aux définitions). La précision calcule le nombre de pages correctement prédites (c.-à-d. correspondant à la définition correcte) sur le nombre de pages prédites. Le rappel de chaque classe calcule le nombre de pages correctement prédites sur le nombre de pages réelles appartenant à la classe. La F-mesure s'appuie sur la moyenne harmonique entre la précision et le rappel. Pour cette évaluation, 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.

	Contexte	**Précision**	**Rappel**	**F-mesure**	**Taux d'erreur**
$AcroDef$-IM3	1 mot	75.4%	74.0%	74.7%	26%
	2 mots	86.2%	86.0%	86.1%	14%
	3 mots	92.3%	92.0%	**92.1%**	**8%**
$AcroDef$-IM	1 mot	57.7%	53.0%	55.2%	47%
	2 mots	65.3%	55.0%	59.7%	45%
	3 mots	68.9%	58.0%	63.0%	42%
$AcroDef$-Dice	1 mot	71.0%	71.0%	71.0%	29%
	2 mots	84.5%	84.0%	84.2%	16%
	3 mots	91.4%	91.0%	91.2%	9%

Tableau 4. *Évaluation de la mesure* $AcroDef$.

Les résultats de nos expérimentations sont présentés dans la table 4. Outre la F-mesure, ce tableau présente également le taux d'erreurs (taux de définitions qui n'ont pas été correctement prédites). Notons que, sur ce jeu de données, les différents calculs propres à la mesure $AcroDef$ ont nécessité un nombre de requêtes assez conséquent : 1800 requêtes[8] à partir du moteur de recherche Exalead. Cette table 4 montre que le résultat global est tout à fait satisfaisant avec, en particulier, une valeur de F-mesure très élevée avec un contexte représenté par trois mots associé à la mesure $AcroDef$-IM3 (92.1%). Ce tableau nous permet d'établir deux conclusions : (1) La mesure de

7. 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".
8. Chaque document nécessite un test de deux définitions ("Jeux Olympiques" et "Journal Officiel") avec 3 requêtes par définition pour le calcul des mesures de qualité. Ceci représente $3 \times 2 = 6$ requêtes par document. Des expérimentations ont été menées avec 3 contextes à partir de 100 documents. Nous avons donc effectué $6 \times 3 \times 100 = 1800$ requêtes.

qualité $AcroDef$ fondée sur l'Information Mutuelle au Cube donne de meilleurs résultats car elle favorise les occurrences fréquentes (nombre important de pages ayant la définition à prédire). Notons que la mesure $AcroDef$-Dice donne également des résultats de bonne qualité. (2) Plus le contexte est important (en terme de nombre de mots) et plus les résultats sont satisfaisants pour l'ensemble des mesures de qualité.

Après l'étude d'un domaine général, la section suivante propose des expérimentations menées à partir d'un domaine de spécialité en anglais.

5.2. *Évaluation de la mesure $AcroDef$ sur un domaine de spécialité*

Dans nos expérimentations, nous nous sommes appuyés sur le classement de définitions en anglais propres aux données biologiques. L'application Acromine[9] permet d'obtenir une liste de définitions possibles pour un sigle donné (Okazaki *et al.*, 2006). Nous avons alors extrait aléatoirement 102 couples sigles/définitions issus de l'application Acromine (pour chacun des sigles expérimentés, le site Acromine propose de 4 à 6 définitions possibles). Notons enfin que les sigles que nous étudions sont composés de deux, trois ou quatre caractères (par exemple, JA, PKD et ABCD dont les définitions sont données dans la table 1 en introduction de cet article). Pour chacun de ces couples, nous avons extrait des résumés d'articles issus de la base de données bibliographique spécialisée Medline[10] qui contiennent les sigles et les définitions. Cette base est constituée de 204 documents (deux documents par couple sigle/définition extraits manuellement). Le but de nos expérimentations est de déterminer si, pour chaque document, la définition est correctement prédite (en classant les définitions possibles avec nos mesures de qualité).

La répartition des 204 documents selon le nombre de possibilités d'expansions pour les sigles à définir dans les documents est donnée dans le tableau 5. Ces expérimentations ont nécessité l'exécution de 3500 requêtes[11]. Le tableau 6 présente le résultat de nos expérimentations pour lesquelles nous donnons le nombre de fois où la définition correcte a été donnée au premier rang, parmi les deux premières définitions et enfin parmi les trois premières définitions en nous appuyant sur les mesures de qualité $AcroDef$. Ces expérimentations ont été menées avec un seul mot pour le contexte (mot le plus fréquent de chaque document). En effet, travaillant sur un domaine très spécialisé, les requêtes avec plus d'un mot retournent très souvent aucune page avec le moteur de recherche généralise Exalead. Notons que nous avons choisi d'utiliser le moteur de recherche Exalead dans le but d'avoir une approche généraliste et afin d'avoir un protocole expérimental semblable à celui présenté dans la section précédente.

Le tableau 6 montre des résultats globalement satisfaisants, en particulier pour la mesure $AcroDef$ fondée sur l'Information Mutuelle au Cube et la mesure de Dice.

---

9. http ://www.nactem.ac.uk/software/acromine/
10. http ://www.ncbi.nlm.nih.gov/PubMed/
11. Expériences menées en février 2008.

Néanmoins, ces résultats qui restent moins intéressants comparativement à ceux obtenus dans les expériences précédentes s'expliquent par la complexité du traitement des données biologiques comme nous allons l'illustrer plus loin. Notons cependant dans le tableau 6 que la définition pertinente est située au premier rang avec $AcroDef$-IM3 dans 36.3% des cas. Ce résultat est significativement meilleur qu'une prédiction aléatoire ayant un score de 22%[12]. Enfin, le tableau 6 montre que les définitions pertinentes sont très rarement données en fin de listes après l'application de nos mesures de qualité. Notre méthode donne donc la possibilité de proposer à l'utilisateur plusieurs choix de définitions pertinentes (qui peuvent d'ailleurs être synonymes comme nous allons le préciser ci-dessous) ; ce qui permet d'éliminer automatiquement bon nombre de définitions non pertinentes.

Nb de documents	12	120	72
Nb de définitions possibles par document	6	5	4

Tableau 5. *Nombre de définitions possibles pour les sigles des 204 documents .*

Mesure	$AcroDef$-IM3	$AcroDef$-IM	$AcroDef$-Dice
Nb de définitions correctes situées au rang 1	**74 (36.3%)**	62 (30.4%)	72 (35.3%)
Nb de définitions correctes situées aux rangs 1 ou 2	118 (57.8%)	116 (56.9%)	**122 (59.8%)**
Nb de définitions correctes situées aux rangs 1, 2 ou 3	**167 (81.9%)**	156 (76.5%)	164 (80.4%)

Tableau 6. *Prédiction des définitions issues des résumés.*

Les difficultés majeures du traitement automatique du langage naturel d'un domaine de spécialité tel que la Biomédecine tient au fait que les concepts véhiculés peuvent se révéler extrêmement proches (par exemple, à partir du site Acromine, le sigle ZO issu de nos expérimentations a pour définitions : `zonula occludens`, `zona occludens`, `zonulae occludentes`, etc). Ainsi, une quantité importante d'erreurs de prédictions peuvent s'expliquer par la présence de définitions très proches voire synonymes. À titre d'exemple, l'analyse avec un expert du domaine permet de dire que les définitions suivantes font référence au même concept, ce qui rend la désambiguïsation automatique difficile : `carboxy terminal`, `carboxy termini`, `carboxyl terminal`, `carboxyl termini`, `COOH-terminal`, `COOH-termini`, `CO2H-terminal`, etc).

Pour comparer nos trois mesures de qualité de manière globale sur ce domaine de spécialité, nous allons calculer la somme des rangs des définitions pertinentes. La mesure qui donne les meilleurs résultats possède la somme la plus faible. Cette méthode qui permet d'évaluer les fonctions de rang est équivalente aux approches fondées sur

12. Résultat d'une prédiction alétoire calculée sur la base suivante : 1 chance sur 4 de classer la définition pertinente au rang 1 dans 72 cas, 1 chance sur 5 dans 120 cas et 1 chance sur 6 dans 12 cas.

les courbes ROC (Receiver Operating Characteristics) et le calcul de l'aire sous ces dernières (Roche *et al.*, 2006). Ainsi, le tableau 7 confirme que, de manière similaire aux textes du domaine général, les mesures $AcroDef$-IM3 et $AcroDef$-Dice se comportent bien sur les données spécialisées de biologie.

Mesure	$AcroDef$-IM3	$AcroDef$-IM	$AcroDef$-Dice
Somme des rangs	**472**	500	473

Tableau 7. *Somme des rangs des définitions pertinentes.*

6. Conclusion et perspectives

L'approche globale présentée dans cet article consiste dans un premier temps à constituer ou enrichir de manière automatique un dictionnaire sigles/définitions. Cette application utilise en entrée un corpus qui peut être ou non spécialisé. L'évaluation des deux étapes constituant notre application montre des résultats satisfaisants en terme de F-mesure. Bien entendu, dans ces dictionnaires plusieurs définitions peuvent être spécifiées pour un même sigle (par exemple, SIG peut signifier "Solde Intermédiaire de Gestion" ou "Système d'Information Géographique"). Le but de la seconde phase de notre approche est, pour un sigle donné dans un document ne possédant pas sa définition, de donner l'expansion adaptée. Pour effectuer ce choix la mesure de qualité $AcroDef$ a été proposée. Cette mesure donne des résultats parfaitement satisfaisants sur des données de domaine général et spécialisé. Dans nos futurs travaux, nous proposons d'associer de manière automatique le domaine de chaque sigle/définition pour avoir un contexte plus riche. Ce contexte fondé sur la désambiguïsation sémantique (Audibert, 2003) et/ou les vecteurs sémantiques (Chauché, 1990) pourront aider à déterminer la thématique des textes utiles pour la désambiguïsation des définitions. Enfin, un contexte plus riche sur la base de traits linguistiques précis (contexte formé d'entités nommées, de mots ayant une fonction grammaticale spécifique, etc) devra être pris en compte dans nos futurs travaux.

Remerciements

Les auteurs remercient V. Matviico et N. Muret pour le développement du logiciel d'extraction des sigles dans les textes effectué dans le cadre d'un stage en Master "Intégration de Compétences". Une démonstration logicielle de cette application a été présentée à la conférence EGC'08 (Matviico *et al.*, 2008).

7. 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.

Brin S., Motwani R., Silverstein C., « Beyond market baskets : generalizing association rules to correlations », *Proceedings of ACM SIGMOD'97*, p. 265-276, 1997.

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.

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

Downey D., Broadhead M., Etzioni O., « Locating Complex Named Entities in Web Text », *Proceedings of IJCAI'07*, p. 2733-2739, 2007.

Dunning T. E., « Accurate Methods for the Statistics of Surprise and Coincidence », *Computational Linguistics*, vol. 19, n° 1, p. 61-74, 1993.

Goodman M., Smyth P., « Information-theoretic rule induction », *Proceedings of ECAI'88 (European Conference on Artificial Intelligence)*, p. 357-362, 1988.

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.

Matviico V., Muret N., Roche M., « Processus d'acquisition d'un dictionnaire de sigles », *Actes de la conférence EGC'08 (session démonstrations)*, p. 231-232, 2008.

Okazaki N., Ananiadou S., « Building an abbreviation dictionary using a term recognition approach », *22*, vol. Bioinformatics, n° 24, p. 3089-3095, 2006.

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.

Roche M., Prince V., « *AcroDef :* A Quality Measure for Discriminating Expansions of Ambiguous Acronyms », *Proceedings of CONTEXT, Springer-Verlag, LNCS*, p. 411-424, 2007.

Torii M., Hu Z., Song M., Wu C., Liu H., « A comparison study on algorithms of detecting long forms for short forms in biomedical text », *BMC Bioinformatics*, 2007.

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.

Xu J., Huang Y., « Using SVM to Extract Acronyms from Text », *Soft Comput.*, vol. 11, n° 4, p. 369-373, 2007.

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