Exit : Un système itératif pour l'extraction de la terminologie du domaine à partir de corpus spécialisés

Mathieu Roche, Thomas Heitz, Oriane Matte-Tailliez, Yves Kodratoff
roche@lri.fr, thomashe@firtech.lri.fr, oriane@lri.fr, yk@lri.fr

LRI - Université Paris-Sud - 91405 Orsay Cedex - France

Abstract

The work presented in this paper is relative to the discovery of a significant terminology in specialized texts. Our approach, partly based on statistical methods extracts the terms in an iterative way. At first, the only terms looked for are binary. The binary terms detected during this first phase are included in the corpus, and the process is iteratively repeated in order to detect very long terms, that happen often to be the most significant terms, as our experience in molecular biology has clearly shown.

Résumé

Les travaux présentés dans cet article se rapportent à l'extraction de la terminologie à partir de textes spécialisés. Notre approche, en partie fondée sur des méthodes statistiques, consiste à extraire des termes de façon itérative. À la première itération, seuls des termes binaires sont recherchés. Ces termes sont alors réintroduits dans le corpus, et ceci est répété à chaque itération. De cette façon, nous pouvons extraire des termes de longueur croissante, particulièrement pertinents pour un domaine particulier, par exemple la biologie moléculaire.

Mots-clés: Terminologie

1 Introduction

L'extraction de la terminologie du domaine à partir de textes est une tâche essentielle dans le but de construire des ontologies spécialisées. Cet article traite d'une étape importante préalable à celle de la terminologie : l'extraction des collocations (Halliday, 1976), c.-à-d. les groupes de mots. Les collocations jugées comme pertinentes sont celles qui peuvent être considérées comme des instances de concepts c'est-à-dire représenter des traces de concepts. Dans la suite, nous appellerons « termes » de telles collocations pertinentes car elles sont considérées par les experts du domaine comme des traces linguistiques de concepts de leur domaine.

Le but de nos travaux consiste à extraire les termes pertinents à partir de corpus spécialisés. Cependant, nous ne traiterons pas, dans cet article, les méthodes de regroupement des termes dans des concepts. Nous expliquerons le processus global d'extraction de la terminologie du domaine que nous proposons : le système Exit (EXtraction Itérative de la Terminologie). Une des caractéristiques essentielles de ce système provient de la méthode itérative et statistique qui le caractérise, comme nous allons le montrer dans la suite.

Nous travaillons actuellement à partir de quatre corpus de tailles, de langues et de spécialités différentes. Ainsi, nous étudions un corpus en anglais composé de 6119 résumés d'articles relatifs à la biologie moléculaire (9424 Ko). Le deuxième corpus étudié est composé de 100 textes (369 Ko) d'introductions d'articles traitant de la fouille de données. Le troisième corpus est issu du domaine des ressources humaines (Compagnie PerformanSe). Ce dernier a une taille de 3784 Ko et a été rédigé par un psychologue qui sert d'expert pour ce corpus. Enfin, nous disposons également d'un corpus de Curriculum Vitæ, d'une taille de 2470 Ko qui contient 1144 CVs (Groupe VediorBis). Dans les expérimentations que nous décrirons dans cet article, nous traiterons seulement le corpus des ressources humaines et le corpus de biologie moléculaire[1] qui sont les plus significatifs en terme de taille. Avant d'expliquer notre méthode d'extraction de la terminologie, nous résumons l'état de l'art du domaine dans la section suivante.

2 État de l'art

De multiples approches de recherche terminologique ont été développées afin d'extraire les termes pertinents à partir d'un corpus. Nous ne traiterons pas ici les approches d'aide à la structuration et au regoupement conceptuel des termes qui sont détaillés dans (Aussenac-Gilles et Bourigault, 2003).

Les méthodes d'extraction de la terminologie sont fondées sur des méthodes statistiques ou syntaxiques. TERMINO (David et Plante, 1990) est un outil précurseur qui s'appuie sur une analyse syntaxique afin d'extraire les termes nominaux. Cet outil effectue une analyse morphologique à base de règles, suivie de l'analyse des collocations nominales à l'aide d'une grammaire. Les travaux de (Smadja, 1993) (XTRACT) s'appuient sur une méthode statistique. XTRACT extrait, dans un premier temps, les collocations binaires situées dans une fenêtre de dix mots. Les collocations binaires sélectionnées sont celles qui dépassent d'une manière statistiquement significative la fréquence due au hasard. L'étape suivante consiste à extraire les collocations plus générales (collocations de plus de deux mots) contenant les collocations binaires trouvées à la précédente étape. ACABIT (Daille, 1994) effectue une analyse linguistique afin de transformer les collocations nominales en termes binaires. Ces derniers sont ensuite triés selon des mesures statistiques. Afin d'extraire la terminologie du domaine, LEXTER (Bourigault et Jacquemin, 1999 ; Aussenac-Gilles et Bourigault, 2000) s'appuie uniquement sur une analyse syntaxique, contrairement au système ACABIT, qui est fondé sur une méthode statistique. La méthode consiste à extraire les syntagmes nominaux maximaux. Ces syntagmes sont alors décomposés en termes de «têtes» et d'«expansions» à l'aide de règles grammaticales. Les termes sont alors proposés sous forme de réseau organisé en fonction de critères syntaxiques.

La section suivante présente l'approche globale que nous proposons et qui s'appuie sur des méthodes statistiques avec une approche itérative.

1. Quelques centaines de termes que nous avons extraits à partir du corpus de biologie moléculaire sont consultables à l'adresse : http://www.lri.fr/ia/Genomics/

3 EXIT : Une approche itérative pour extraire la terminologie du domaine

L'approche globale que nous proposons avec EXIT se déroule en trois phases.

Comme dans de nombreux travaux, nous nous sommes intéressés, dans la première phase, à l'extraction de la terminologie nominale. L'originalité de notre méthode vient du procédé itératif que nous proposons. Cette méthode consiste, à chaque itération, à former des termes binaires (ou ternaires pour les termes prépositionnels). Chaque terme est sélectionné selon une mesure statistique (voir section 4.2). Par exemple, sur le corpus des ressources humaines, à la première itération, nous avons extrait le terme binaire « travail administratif ». Ces termes sont réintroduits dans le corpus avec des traits d'union afin qu'ils soient reconnus comme des mots à part entière. Nous pouvons ainsi effectuer une nouvelle recherche terminologique à partir de ce corpus avec prise en compte de la terminologie du domaine acquise à l'étape précédente. Par exemple, à l'aide du terme binaire « travail administratif » trouvé à la première itération, nous pouvons extraire le terme « responsabilité de travail-administratif » à la deuxième itération. Notre méthode permet alors de détecter des termes très spécifiques (composés de plusieurs mots). Ceci est essentiel, par exemple en biologie moléculaire, où les termes les plus pertinents sont les termes composés de nombreux mots. Nous démontrerons dans la section 4.5, la terminaison de notre algorithme. De plus, comme nous le montrerons dans la section 4.6, nous avons ajouté un certain nombre de paramètres afin d'augmenter la pertinence des termes extraits.

La deuxième phase de notre travail consiste à extraire la terminologie verbale. Nous nous sommes attachés à extraire les termes du type « verbe-objet » et « sujet-verbe » comme dans les travaux de (Smadja, 1993). Par exemple, grâce à la terminologie nominale extraite dans la première phase, nous avons alors pu obtenir le terme verbal « prendre responsabilité-de-travail-administratif ». De même que pour les termes nominaux, les termes verbaux sont ordonnancés selon une mesure statistique.

Enfin, dans la dernière phase, nous avons utilisé FASTR (Jacquemin, 1996) afin de déterminer les termes variants à partir des termes nominaux extraits mais également à partir des termes verbaux.

La figure 1 illustre le schéma global de l'ensemble de notre processus. Le détail de chacune de ces phases est donné dans les sections suivantes.

4 Extraction de la terminologie nominale

4.1 Nettoyage et étiquetage

La première étape de notre travail a consisté à normaliser le corpus. Pour effectuer ce travail, nous supprimons les éléments susceptibles d'engendrer des erreurs avec les outils automatiques d'analyse des textes (étiqueteurs, analyseurs syntaxiques, etc.). Ainsi nous avons supprimé le nom des auteurs des articles que nous traitons, le nom des laboratoires, etc. Avec l'expert du domaine, nous avons également établi des règles afin d'uniformiser le vocabulaire employé. Par exemple, sur le corpus de biologie moléculaire, nous avons

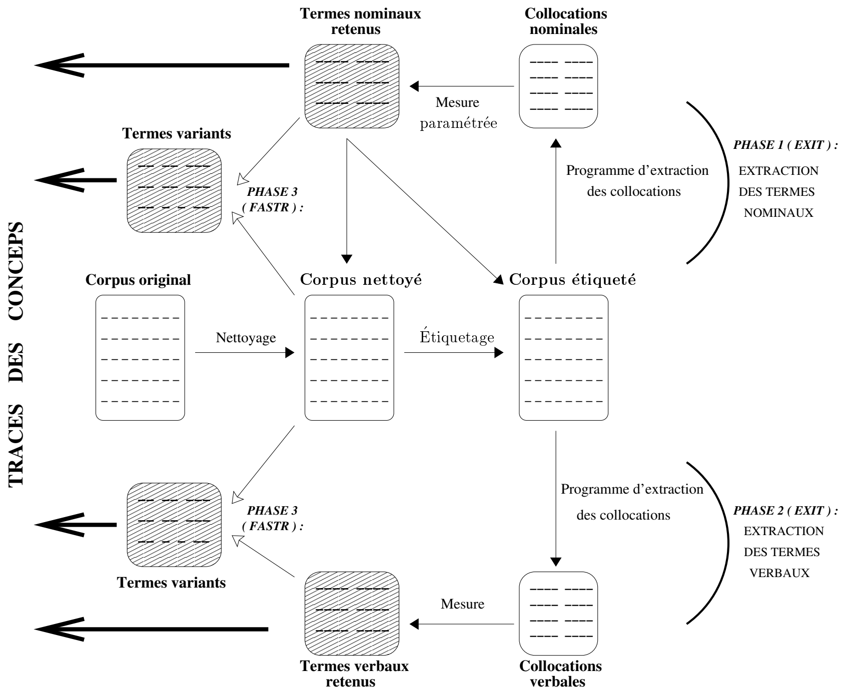


Fig. 1: *Schéma global de la recherche terminologique dans les textes avec* Exit.

remplacé les termes « carboxyl terminal », « carboxyl termini », « COOH-terminal », « C-terminal », etc. par « C-term ». L'expert du domaine a également construit une règle lexicale afin de repérer les formules dans le corpus de biologie moléculaire (par exemple, « GAL4 », « RAP1 », « FK506 », « Mcm1 », etc.). Ces formules sont souvent associées à un nom significatif pour le domaine (« gene », « protein », « promoter », « cell », etc.). Une règle de nettoyage simple (ou de pré-terminologie) a alors consisté à rassembler les formules et les mots significatifs par un trait d'union (« formule-nom_significatif »). Une telle tâche permet à l'étiqueteur de Brill que nous allons utiliser à l'étape suivante, de reconnaître de tels regroupements comme des mots à part entière. D'autres règles du même type ont été construites à l'aide de la connaissance du domaine fournie par l'expert

La deuxième étape de notre travail utilise l'étiqueteur de Brill (Brill, 1994) afin d'apposer une étiquette grammaticale à chacun des mots du corpus. Nous avons également ajouté quelques règles lexicales et contextuelles afin d'enrichir le lexique standard de Brill qui possédait bon nombre de mots inconnus. L'enrichissement du lexique est essentiel, en particulier sur un domaine très spécialisé tel que la biologie moléculaire pour lequel plus 70% des mots ne sont pas présents dans le lexique standard de l'étiqueteur de Brill. Avec l'ajout de règles lexicales et contextuelles établies par l'expert, tous les mots du corpus de biologie moléculaire sont alors reconnus par l'étiqueteur.

L'étape suivante a consisté à extraire les collocations binaires (ou ternaires pour les collocations prépositionnelles), c'est-à-dire les mots voisins ayant une étiquette spécifique.

Nous nous sommes ainsi intéressés aux collocations du type *nom-nom*, *adjectif-nom*, *nom-adjectif*[2], *nom-préposition-nom* et *formule-nom*[3] car elles sont les plus porteuses de traces de concepts.
La section suivante présente la manière dont nous allons ordonnancer ces collocations.

4.2 Mesure utilisée avec EXIT

Après avoir extrait l'ensemble des collocations binaires ou ternaires, nous ordonnançons ces dernières selon une mesure statistique. Plusieurs mesures sont décrites dans la littérature. Nous avons testé les versions traitées dans (Jacquemin, 1997 ; Daille et al., 1998 ; Lenca et al., 2003). La comparaison des résultats obtenus avec 11 mesures est détaillée dans (Roche et al., 2003). Dans cet article, nous présentons deux mesures typiques pour l'extraction de la terminologie : l'information mutuelle et le rapport de vraisemblance.

La première mesure que nous décrivons correspond à l'information mutuelle (Church et Hanks, 1990). Cette mesure est définie par la formule (1) :

$$IM(x,y) = \log_2 \frac{P(x,y)}{P(x)P(y)} \tag{1}$$

L'information mutuelle calcule une certaine forme d'indépendance des mots x et y composant la collocation. Une telle mesure a tendance à extraire des termes rares et peu fréquents.

La deuxième mesure que nous citerons dans cet article correspond au rapport de vraisemblance (Dunning, 1993). Comme le précisent (Daille et al., 1998 ; Xu et al., 2002), cette deuxième mesure est particulièrement bien adaptée pour l'extraction de la terminologie. Nous définissons le rapport de vraisemblance de la manière suivante à l'aide d'une table de contingences associée à chaque couple de mots (x_i, x_j) comme ci-dessous :

	x_j	$x_{j'}$ avec $j' \neq j$
x_i	a	b
$x_{i'}$ avec $i' \neq i$	c	d

Les valeurs a, b, c et d définissent les occurrences des couples et $a+b+c+d = N$ est le nombre total d'occurrences des couples trouvés.

On définit alors le rapport de vraisemblance par la formule (2) :

$$\begin{aligned} RV(x_i, x_j) \;=\;& a\log(a) + b\log(b) + c\log(c) + d\log(d) - (a+b)\log(a+b) \\ & -(a+c)\log(a+c) - (b+d)\log(b+d) - (c+d)\log(c+d) \\ & +N\log(N) \end{aligned} \tag{2}$$

Ces deux mesures affectent un score aux collocations (x, y). Le rapport de vraisemblance prend en compte les collocations formées avec un seul des mots x et y ainsi que les collocations formées avec aucun de ces mots x et y. A contrario, l'information mutuelle s'appuie uniquement sur l'indépendance de chacun des mots x et y composant la collocation.

2. Pour le corpus français uniquement.
3. L'étiquette « formule » est une étiquette spécifique au corpus de biologie moléculaire.

4.3 Expérimentations

Afin de comparer les différentes mesures de la littérature, nous nous appuyons sur la mesure de Précision et son complément la courbe d'élévation (« lift chart »), qui sont les seules mesures d'évaluation dont nous puissions disposer dans un cadre d'apprentissage non-supervisé. La précision, qui donne la proportion de collocations correctes parmi les collocations extraites, est définie par la formule (3) :

$$Précision = \frac{\text{nombre de collocations extraites pertinentes}}{\text{nombre de collocations extraites}} \tag{3}$$

Comme le précisent (Aussenac-Gilles et Bourigault, 2000), selon l'utilisation que l'on compte faire de la terminologie (indexation, ontologie, etc.), l'évaluation de la pertinence des termes peut différer. Pour notre part, nous jugeons qu'une collocation est pertinente si elle représente une trace de concept.

La courbe d'élévation donne la précision en fonction de la proportion de collocations proposées à l'expert. Précisons qu'il est impossible de connaître le nombre total de collocations pertinentes des corpus, c'est la raison pour laquelle nous ne calculerons ni le Rappel ni la courbe ROC (Ferri et al., 2002).

À titre d'exemple, la figure 2 présente la courbe d'élévation des deux mesures décrites dans la section précédente, appliquées au corpus de biologie moléculaire. Ainsi, la figure 2 montre que la précision est toujours plus importante avec le rapport de vraisemblance suivant le nombre de collocations proposées à l'expert. Notons que les autres corpus donnent des résultats similaires.

Plus généralement, rappelons que parmi la dizaine de mesures testées dans (Roche et al., 2003), le rapport de vraisemblance est la mesure qui a le meilleur comportement pour l'extraction de la terminologie.

Nous précisons que le critère consistant à simplement classer les collocations selon leur nombre d'occurrences peut également se révéler efficace. Cependant, il est nécessaire d'effectuer un classement des collocations ayant le même nombre d'occurences en utilisant une mesure statistique.

4.4 Sélection des collocations

Une fois les mesures statistiques effectuées, l'ensemble ordonné de ces collocations (ou un sous-ensemble d'entre elles selon le choix de l'expert) est proposé à l'expert. L'expert les classe alors en quatre catégories.

1. La première catégorie correspond aux collocations reconnues comme pertinentes par l'expert. Nous noterons $\mathcal{L}_1^i$ cet ensemble de collocations de l'itération i.

2. La deuxième catégorie correspond aux collocations reconnues comme non pertinentes par l'expert. Nous noterons $\mathcal{L}_2^i$ ces collocations de l'itération i.

3. La troisième catégorie correspond aux collocations qui ne sont pas pertinentes mais

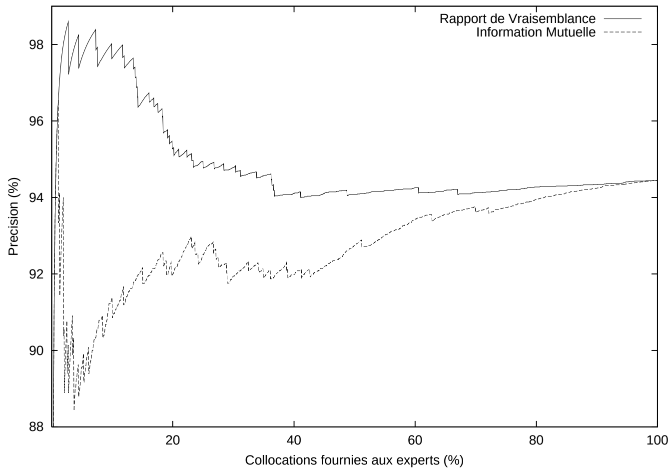


FIG. 2: *Courbes d'élévation propres au corpus de biologie moléculaire avec la relation adjectif-nom et un élagage à 4 avec l'information mutuelle et le rapport de vraisemblance.*

qui pourraient être pertinentes lors de la formation d'un terme plus long et donc plus spécifique. Nous noterons $\mathcal{L}_3^i$ ces collocations de l'itération i.

4. La dernière catégorie est composée des collocations non évaluées par l'expert. Nous noterons $\mathcal{L}_4^i$ cet ensemble de collocations relatives à l'itération i.

Pour faciliter, le travail de l'expert, nous avons développé une interface graphique en Java (voir figure 3). Cette interface permet de sélectionner les collocations selon les catégories décrites. Nous précisons que pour faciliter le travail de l'expert, nous proposons à l'utilisateur de visualiser les contextes (phrases) dans lesquels apparaissent les collocations.

De plus, nous proposons à l'expert la possibilité de valider certaines collocations si elles respectent des règles lexicales que l'expert peut construire à l'aide de l'interface graphique. Ainsi, une partie des collocations est validée de façon semi-automatique comme collocations pertinentes ou non pertinentes.

Enfin, le système que nous avons développé utilise le principe de l'élagage progressif suivant les itérations. L'élagage correspond au fait de proposer à l'expert uniquement les collocations qui apparaissent un nombre de fois minimum dans le corpus. Afin de sélectionner les termes de plus en plus significatifs à chaque itération, l'expert peut effectuer un élagage plus important selon les itérations. Nous appelons ce principe un élagage progressif.

4.5 Terminaison de l'algorithme

Pour démontrer la terminaison de notre algorithme itératif de recherche terminologique, nous en présentons ci-dessous quelques caractéristiques.

- Soit N_i le nombre total de collocations extraites à l'itération i.
- Soit n_i le nombre total de collocations extraites après un élagage λ_i à la $i$ème itération.

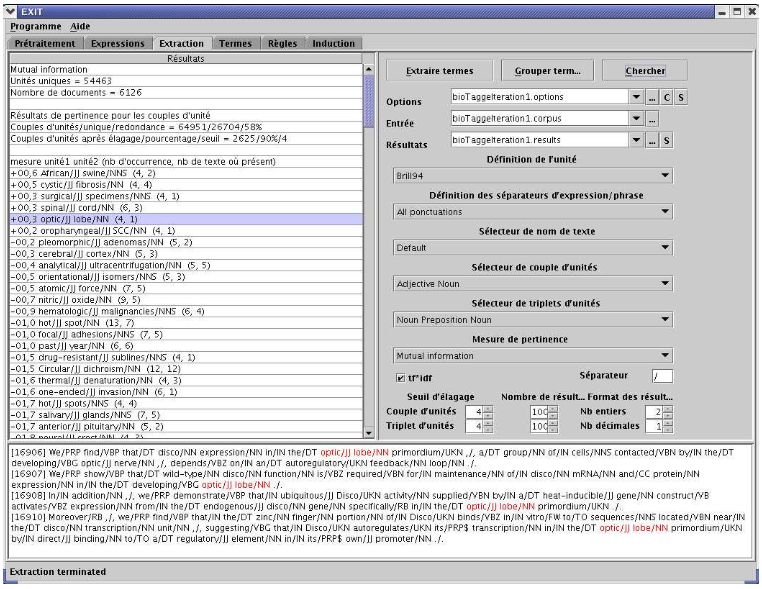


Fig. 3: *Capture d'écran du logiciel* Exit. *La partie gauche présente les différentes collocations classées selon une mesure statistique. La partie droite présente les différents paramètres (types de collocations à extraire, mesures à utiliser, élagage, etc.). La partie inférieure présente les contextes (phrases) dans lesquels nous trouvons les collocations sélectionnées par l'expert.*

– Soit α_i le nombre de collocations retenues dans les listes $\mathcal{L}_k^i$ où $1 \leq k \leq 3$ (voir section 4.4) parmi les n_i collocations extraites à l'itération i.
  Ainsi, $\alpha_i = card(\bigcup_{k=1}^{3} \mathcal{L}_k^i)$. On suppose qu'à chaque itération, au moins une collocation est retenue, c.-à-d. $\forall i, \alpha_i > 0$.

Bien entendu, $\forall i, \alpha_i \leq n_i \leq N_i$. S'il n'y a pas d'élagage ($\lambda_i = 1$) alors $n_i = N_i$.

L'algorithme s'arrête lorsque nous obtenons une liste vide de collocations après élagage. Ainsi, le nombre d'itérations de notre algorithme revient à déterminer j tel que $n_{j+1} = 0$.

Dans le meilleur des cas, cas trivial pour démontrer la terminaison de notre algorithme, nous sélectionnons toutes les collocations de la première itération parmi les listes $\mathcal{L}_k^1$ où $1 \leq k \leq 3$. Dans ce cas $j = 1$ et on a $n_1 = \alpha_1$.

Dans le pire des cas, on suppose qu'il n'y a pas d'élagage progressif (c.-à-d. $\forall (i,j), \lambda_i = \lambda_j$) et que l'on ne sélectionne qu'une seule collocation par itération ($\forall i, \alpha_i = 1$). Dans ce cas, l'algorithme se termine après $j = n_1$ itérations. Ce nombre n_1 étant nécessairement fini, la terminaison de notre algorithme est donc assurée dans le pire des cas.
Notons qu'ici, nous supposons qu'à l'itération i, lorsqu'un seul terme est validé parmi les n_i collocations extraites, à l'itération suivante, nous avons $n_i - 1$ collocations qui sont extraites. Ainsi, nous n'avons pas pris en compte le cas rare où la sélection d'un terme permet de construire une nouvelle collocation binaire qui ne pouvait être établie lors des itérations précédentes (par exemple, avec la structure « nom adjectif1 adjectif2 » où seule

la collocation «nom adjectif1» est proposée à l'itération i puis «nom-adjectif1 adjectif2» à l'itération $i+1$). La garantie de terminaison de l'algorithme reste cependant vérifiée car le nombre de termes formés dans ce cas est nécessairement borné (la borne correspond alors au nombre fini de mots moins 1 présents dans les phrases rencontrant un tel cas).

Dans le cas général, et s'il n'y a pas d'élagage progressif suivant les itérations, l'algorithme se termine après j itérations, où j vérifie $\sum_{i=1}^{j} \alpha_i = n_1$. S'il y a un élagage progressif, l'algorithme se termine avec j_0 tel que $j_0 \leq j$.

4.6 Paramètres supplémentaires

Dans cet article, nous détaillons deux paramètres qui sont utilisés pour l'extraction de la terminologie. Nous présentons ici les paramètres que nous avons modifiés comparativement aux versions présentées dans (Roche, 2003).

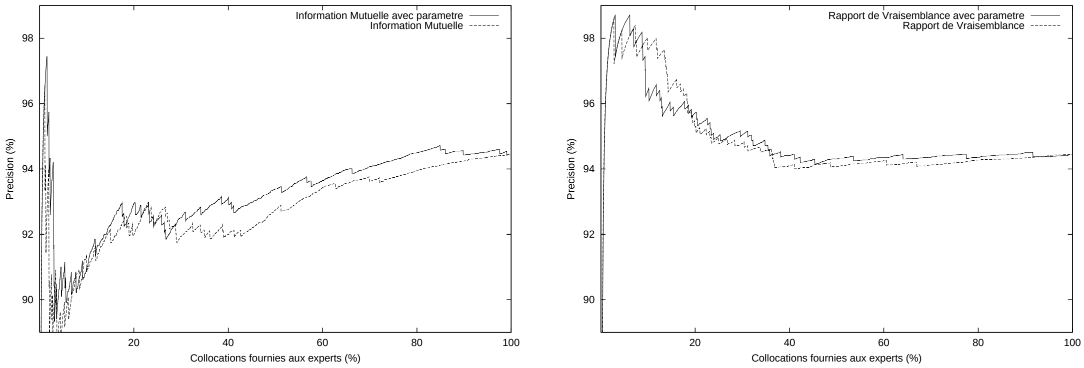
Un des paramètres que nous avons développé consiste à sélectionner automatiquement les termes déjà validés par les auteurs du corpus. Ce principe consiste à considérer, lors de la première itération, les termes du corpus qui possèdent un trait d'union (par exemple, «data-mining») comme des termes pertinents. Les mêmes collocations que l'on trouve sans trait d'union (exemple, «data mining») sont également reconnues comme des termes valides et sont introduits automatiquement dans la liste $\mathcal{L}_1^1$. Ce principe revient à utiliser l'expertise implicite de certains auteurs. Bien entendu, l'utilisateur peut éliminer de $\mathcal{L}_1^1$ tout terme qu'il juge non pertinent. Dans la phase de nettoyage (voir section 4.1), nous avons, par exemple, identifié des termes de type «formule nom_spécifique» que nous avons remplacés par le regroupement «formule-nom_spécifique». Ces termes sont *a priori* pertinents car ils ont été détectés semi-automatiquement grâce à des règles lexicales construites par l'expert du domaine durant la phase de nettoyage. Le paramètre que nous décrivons ici permet alors de placer automatiquement de tels termes dans la liste $\mathcal{L}_1^1$ des termes valides.

Un autre paramètre essentiel que nous avons également utilisé consiste à priviligier les collocations qui apparaissent dans des textes différents, jugées comme davantage représentatives du domaine, comparativement aux collocations qui sont présentes dans peu de textes.

Pour cela, nous allons nous appuyer sur la pondération classique dans le domaine de la recherche d'informations qui est le tf × idf (term frequency - inverse document frequency). Cette pondération est donnée par la formule (4).

$$w_{ij} \quad = \quad \mathrm{tf}_{ij} \times \log_2 \frac{N}{n} \tag{4}$$

- w_{ij} est le poids du terme T_j dans le document D_i ;
- tf_{ij} est la fréquence du terme T_j dans le document D_i ;
- N est le nombre de documents dans la collection ;
- n est le nombre de documents où le terme T_j apparaît au moins une fois.



(a) (b)

FIG. 4: *Information mutuelle (a) et Rapport de Vraisemblance (b) avec paramètre.*

Cette formule combine l'importance des termes pour un document (tf) et le pouvoir de discrimination de ce terme (idf). Ainsi, un terme qui a une valeur de tf $\times$ idf élevée doit être à la fois important dans ce document et il doit apparaître peu dans les autres documents. C'est le cas où un terme correspond à une caractéristique importante et unique d'un document. Une telle mesure est souvent utilisée en recherche documentaire. Dans notre cas, nous souhaitons obtenir une réponse inverse. Ainsi, nous souhaitons privilégier les collocations fréquentes qui apparaissent dans de nombreux documents. Pour cela, nous avons modifié la formule classique du tf $\times$ idf de la manière suivante (5) :

$$w'_{ij} \quad = \quad \frac{\mathrm{tf}_{ij}}{\log_2 \frac{N}{n}} \tag{5}$$

Nous avons estimé l'amélioration de la précision obtenue en remplaçant le nombre d'occurrences des collocations et des mots par la pondération décrite précédemment. Globalement, la précision est améliorée en utilisant notre nouveau paramètre. À titre d'exemple, avec l'information mutuelle appliquée à la relation adjectif-nom du corpus de biologie moléculaire, la précision est sensiblement améliorée (voir figure 4.a). Pour le rapport de vraisemblance (voir figure 4.b), la précision est globalement meilleure en utilisant notre paramètre. Cependant, si nous fournissons aux experts un faible pourcentage de collocations (moins de 20%), l'utilisation de notre paramètre n'améliore pas toujours les résultats en terme de précision.

5 Extraction de la terminologie verbale

Dans cette étape, nous nous sommes intéressés à l'extraction des collocations verbales de type « verbe-objet » et « sujet-verbe » comme dans les travaux de (Smadja, 1993). Cette extraction des collocations s'effectue grâce à des règles contextuelles consistant à trouver les collocations binaires formées avec les étiquettes relatives aux noms et aux verbes. La différence entre collocations verbales et collocations nominales tient au fait que certaines collocations verbales peuvent se révéler erronées. Ainsi, nous avons estimé que le taux des collocations verbales syntaxiquement correctes est à plus de 80%. Ces collocations binaires sont ensuite ordonnancées selon le score donné par le rapport de vraisemblance.

6 Extraction des termes variants

Nous avons évalué les termes variants donnés par Fastr (Jacquemin, 1996). Ainsi, (Jacquemin, 1996) définit, par exemple, trois types de termes variants :

- $X_2X_3 \mapsto X_2X_4C_5X_3$. Par exemple, avec le corpus de biologie moléculaire, nous avons *Ty1 transposition* $\mapsto$ *Ty1 transcription and transposition*
- $X_2X_3 \mapsto X_2X_4X_3$. Par exemple, *close proximity* $\mapsto$ *close chromosomal proximity*
- $X_2X_3 \mapsto X_3P_4X_5X_2$. Par exemple, *protein sequence* $\mapsto$ *sequence for dimeric protein*

Nous avons effectué une évaluation des termes variants du corpus de biologie moléculaire. Sur 1200 termes variants donnés par Fastr à partir des termes nominaux extraits à la première itération, 85% des termes ont été jugés pertinents. Cette expertise donne des résultats similaires à ceux cités dans (Jacquemin, 1996) correspondant à un corpus médical. Avec ce corpus, les termes variants obtenus atteignaient également une précision de près de 85%.

7 Conclusions et perspectives

L'approche que nous proposons dans cet article est une approche globale pour extraire la terminologie à partir de corpus spécialisés. Notre méthode s'appuie essentiellement sur une méthode itérative et des mesures statistiques. Nous avons validé notre approche sur des corpus de langues, de tailles et de domaines différents.

La méthodologie décrite ici a deux perspectives immédiates. Premièrement, l'expert examine de nombreuses phrases contenant les termes détectés afin de les valider. Lorsqu'il juge que tous les termes intéressants pour lui ont été trouvés, il peut alors conserver ces phrases, créant ainsi un sous-corpus « parfaitement » étiqueté sur lequel des mesures ROC et de Rappel peuvent être effectuées. Cette option est implantée dans notre système mais n'a pas encore été systématiquement utilisée par les experts.

Deuxièmement, la base étiquetée constitue un ensemble d'apprentissage (bruité en général mais non bruité pour un sous-corpus) à partir duquel un outil d'induction automatique peut créer de nouvelles règles dont la pertinence pourra être jugée par l'expert. Le problème majeur posé par cette approche est celui de la définition des descripteurs de ce qui est un « bon » terme. Nos résultats préliminaires montrent que nous aurons besoin d'outils semblables à ceux utilisés pour la détection des traces linguistiques de concepts, à savoir « l'induction extensionnelle » définie dans (Kodratoff, 2004).

Références

Aussenac-Gilles N. and Bourigault D. (2000). The Th(IC)2 Initiative: Corpus-Based Thesaurus Construction for Indexing WWW Documents. In *Proceedings of the EKAW'2000 Workshop on Ontologies and Texts, Vol-51.*

Aussenac-Gilles N. and Bourigault D. (2003). Construction d'ontologies à partir de textes. In *Actes de TALN03*, volume 2, pages 27–47.

Bourigault D. and Jacquemin C. (1999). Term Extraction + Term Clustering: An Integrated Platform for Computer-Aided Terminology. In *Proceedings of the European Chapter of the Association for Computational Linguistics (EACL '99), Bergen.*, pages 15–22.

Brill E. (1994). Some advances in transformation-based part of speech tagging. In *AAAI, Vol. 1*, pages 722–727.

Church K. W. and Hanks P. (1990). Word Association Norms, Mutual Information, and Lexicography. In *Computational Linguistics*, volume 16, pages 22–29.

Daille B. (1994). Approche mixte pour l'extraction automatique de terminologie : statistiques lexicales et filtres linguistiques. In *Thèse de Doctorat en Informatique Fondamentale. Université Paris 7.*

Daille B., Gaussier E., and Langé J. (1998). An Evaluation of Statistical Scores for Word Association. In *J.Ginzburg, Z. Khasidashvili, C. Vogel, J.-J. Levy, and E. Vallduvi (eds) The Tbilisi Symposium on Logic, Language and Computation: Selected Papers, CSLI Publications*, pages 177–188.

David S. and Plante P. (1990). De la nécéssité d'une approche morpho syntaxique dans l'analyse de textes. In *Intelligence Artificielle et Sciences Cognitives au Quebec*, volume 3, pages 140–154.

Dunning T. E. (1993). Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, 19(1):61–74.

Ferri C., Flach P., and Hernández-Orallo J. (2002). Learning decision trees using the area under the ROC curve. In *9th International Conference on Machine Learning, ICML'02*, pages 139–146.

Halliday M. A. K. (1976). *System and Function in Language.* Oxford University Press, London.

Jacquemin C. (1996). A symbolic and surgical acquisition of terms through variation. In Springer, editor, *S. Wermter, E. Riloff, and G. Scheler, editors, Connectionist, Statistical and Symbolic Approaches to Learning for Natural Language Processing*, pages 425–438.

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

Kodratoff Y. (2004). Induction Extentionnelle : définition et application à l'acquisition de concepts à partir de textes. *RNTI, numéro spécial EGC04, à paraître.*

Lenca P., Meyer P., Picouet P., Vaillant B., and Lallich S. (2003). Critères d'évaluation des mesures de qualité des règles d'association, n° spécial "entreposage et fouille des données". *RNTI (Revue des Nouvelles Technologies de l'Information), CEPADUES*, pages 123–134.

Roche M. (2003). Extraction paramétrée de la terminologie du domaine. *RSTI série RIA-ECA, numéro spécial EGC03*, 17:295–306.

Roche M., Matte-Tailliez O., Azé J., and Kodratoff Y. (2003). Extraction de la Terminologie du Domaine : Étude de Mesures sur un Corpus Spécialisé Issu du Web. In *Actes des Journées Francophones de la Toile 2003*, pages 279–288.

Smadja F. (1993). Retrieving collocations from text : Xtract. *Computational Linguistics*, 19:143–177.

Xu F., Kurz D., Piskorski J., and Schmeier S. (2002). A Domain Adaptive Approach to Automatic Acquisition of Domain Relevant Terms and their Relations with Bootstrapping. In *LREC 2002, the third international conference on language resources and evaluation.*