HAL
archives-ouvertes.fr

Choix du taux d'élagage pour l'extraction de la terminologie. Une approche fondée sur les courbes ROC

Mathieu Roche, Yves Kodratoff

▶ **To cite this version:**

Mathieu Roche, Yves Kodratoff. Choix du taux d'élagage pour l'extraction de la terminologie. Une approche fondée sur les courbes ROC. EGC: Extraction et Gestion des Connaissances, Jan 2006, Lille, France. pp.205-216. lirmm-00087576

HAL Id: lirmm-00087576
https://hal-lirmm.ccsd.cnrs.fr/lirmm-00087576

Submitted on 25 Jul 2006

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Choix du taux d'élagage pour l'extraction de la terminologie. Une approche fondée sur les courbes ROC

Mathieu Roche*, Yves Kodratoff**

*LIRMM - UMR 5506, Université Montpellier 2,
34392 Montpellier Cedex 5 - France
mathieu.roche@lirmm.fr
**LRI - UMR 8623, Université Paris-Sud,
91405 Orsay Cedex - France
yk@lri.fr

Résumé. Le choix du taux d'élagage est crucial dans le but d'acquérir une terminologie de qualité à partir de corpus de spécialité. Cet article présente une étude expérimentale consistant à déterminer le taux d'élagage le plus adapté. Plusieurs mesures d'évaluation peuvent être utilisées pour déterminer ce taux tels que la précision, le rappel et le F_{score}. Cette étude s'appuie sur une autre mesure d'évaluation qui semble particulièrement bien adaptée pour l'extraction de la terminologie : les courbes ROC (Receiver Operating Characteristics).

1 Introduction

Cet article présente une étude expérimentale consistant à évaluer le taux d'élagage le plus adapté pour l'extraction de la terminologie. Nous allons décrire ci-dessous notre méthode globale d'extraction de la terminologie et rigoureusement définir l'élagage.

La première phase de notre travail d'extraction de la terminologie à partir de corpus spécialisés consiste à normaliser les textes en utilisant des règles de nettoyage décrites par Roche (2004). Les corpus que nous utilisons sont décrits dans la section 3 de cet article. L'étape suivante consiste à apposer des étiquettes grammaticales à chacun des mots du corpus en utilisant l'étiqueteur ETIQ développé par Amrani et al. (2004). ETIQ est un système interactif s'appuyant sur l'étiqueteur de Brill (1994) qui améliore la qualité de l'étiquetage de corpus spécialisés. Nous pouvons alors extraire l'ensemble des collocations Nom-Nom, Adjectif-Nom, Nom-Adjectif[1], Nom-Préposition-Nom d'un corpus spécialisé. L'étape suivante consiste à sélectionner les collocations les plus pertinentes selon des mesures statistiques décrites par Roche et al. (2004c); Roche (2004). Les collocations sont des groupes de mots définis par Halliday (1976); Smadja (1993). Nous appelons *termes*, les collocations pertinentes.

Les termes binaires (ou ternaires pour les termes prépositionnels) extraits à chaque itération 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 du corpus avec prise en compte de la terminologie du domaine acquise aux étapes précédentes. Notre méthode itérative, proche des travaux de Evans et Zhai (1996), est décrite

[1] Corpus en français uniquement

par Roche et al. (2004b); Roche (2004). Cette approche permet de détecter des termes très spécifiques (composés de plusieurs mots). Ceci est essentiel, par exemple dans le domaine de la biologie, où les termes les plus pertinents sont les termes composés de nombreux mots.

Le choix du taux d'élagage consiste à établir le nombre de fois minimal où les collocations doivent être présentes dans le corpus afin que ces dernières soient extraites en étant jugées comme significatives. Ainsi, pour ne pas surcharger le travail de l'expert, ce dernier peut choisir de ne pas extraire les collocations rares.

Cet article présente, dans un premier temps, un état de l'art des méthodes d'extraction de la terminologie (section 2). Puis une description succincte des corpus utilisés est donnée en section 3 suivie de la présentation de l'application de différents taux d'élagages effectués sur ces corpus (section 4). La section 5 présente de quelle manière les collocations sont expertisées. La section 6 décrit différentes mesures d'évaluation de la terminologie centrées sur la problématique du choix du taux d'élagage. La section 7 propose enfin quelques perspectives.

2 État de l'art des méthodes d'extraction de la terminologie

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 regroupement conceptuel des termes qui sont détaillés dans les travaux de Aussenac-Gilles et Bourigault (2003).

Les méthodes d'extraction de la terminologie sont fondées sur des méthodes statistiques ou syntaxiques. Le système TERMINO de 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 de Daille (1994) effectue une analyse linguistique afin de transformer les collocations nominales en termes binaires. Ces derniers sont ensuite triés selon des mesures statistiques. Contrairement à ACABIT qui est fondé sur une méthode statistique, LEXTER de Bourigault (1993) et SYNTEX de Bourigault et Fabre (2000) s'appuient essentiellement sur une analyse syntaxique afin d'extraire la terminologie du domaine. 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.

Pour discuter le choix du taux d'élagage selon le nombre d'occurrences, nous allons classer les collocations en utilisant la mesure Occ_{RV} décrite dans les travaux de Roche (2004). Cette mesure qui a le meilleur comportement comme précisé par Roche et al. (2004a,c) classe les collocations selon leur nombre d'occurrences et les collocations ayant le même nombre d'occurrences sont classées en utilisant le Rapport de Vraisemblance de Dunning (1993). Cette mesure est parfaitement bien adaptée à cette étude car le classement effectué selon le nombre d'occurrences permet de discuter le choix du taux d'élagage.

3 Description des corpus

Au cours de ces travaux, nous avons travaillé à partir de quatre corpus de spécialité, de langues et de tailles différentes.

- Le premier corpus étudié est composé de 100 introductions d'articles scientifiques (369 Ko) traitant de la Fouille de Données écrits en langue anglaise par des auteurs anglophones.
- Le deuxième corpus en langue anglaise traité est un corpus de Biologie Moléculaire (9424 Ko). Il a été obtenu par une requête au *NIH* sur Medline (PubMed)[2] avec les mots-clés *DNA-binding, proteins, yeast* obtenant de ce fait un corpus de 6119 résumés d'articles scientifiques. Outre sa taille conséquente, une des caractéristiques de ce corpus tient dans l'utilisation d'un vocabulaire très spécialisé. Une partie de ce corpus est disponible à l'adresse suivante : http ://www.lri.fr/ia/Genomics/.
- Le troisième corpus est composé de 1144 Curriculum Vitæ fournis par la société VediorBis. Ce corpus écrit en français a une taille de 2470 Ko. Une des particularités de ce corpus tient au fait qu'il est composé de phrases très courtes avec de nombreuses énumérations.
- Enfin, le dernier corpus de spécialité étudié dans cet article est composé d'un ensemble de textes écrits en français qui sont issus du domaine des Ressources Humaines (société PerformanSe). Les textes écrits correspondent à des commentaires de tests de psychologie de 378 individus. Les textes sont écrits par un seul auteur qui emploie un vocabulaire spécifique avec l'utilisation de tournures souvent littéraires. Une partie de ce corpus est disponible à l'adresse suivante : http ://www.lri.fr/∼roche/Recherche/corpusPsy.html

4 Taux d'élagage de la terminologie

Le principe de l'élagage des collocations consiste à exploiter seulement les collocations présentes un nombre de fois minimum dans le corpus. L'élagage permet d'exclure les collocations trop rares qui peuvent se révéler comme non significatives pour le domaine. Sur chacun des corpus expérimentés, le tableau 1 présente différents élagages appliqués.

La première observation à relever dans le tableau 1 tient dans le nombre d'occurrences des collocations qui diffère selon les langues. Par exemple, les collocations de type Nom-Nom sont beaucoup moins fréquentes sur les corpus en français par rapport aux corpus en anglais.

Suivant les domaines de spécialité écrits dans une même langue, les résultats peuvent également différer de manière importante. Par exemple, sur le corpus de CVs, le nombre de collocations de type Nom-Nom est beaucoup plus important que celui du corpus des Ressources Humaines également écrit en français. Le corpus des Ressources Humaines a pourtant une taille plus grande que le corpus de CVs. Ceci est dû au fait que les CVs sont écrits de manière condensée en employant un vocabulaire très spécifique : "*emploi solidarité*", "*action communication*", "*fichier client*", "*service achat*", etc.

Le tableau 1 montre qu'en éliminant simplement les collocations présentes une seule fois dans le corpus, plus de la moitié des collocations sont supprimées dans tous les cas, excepté la relation Adjectif-Nom du corpus des Ressources Humaines. Ce corpus a un comportement

[2] http ://www.ncbi.nlm.nih.gov/entrez/query.fcgi

	nb total	**élagage 2**	**élagage 3**	**élagage 4**	**élagage 5**	**élagage 6**
	CORPUS DE CVS (français)					
Nom-Nom	1781	353 (80%)	162 (91%)	100 (94%)	69 (96%)	56 (97%)
Nom-Prép-Nom	3634	662 (82%)	307 (91%)	178 (95%)	113 (97%)	80 (98%)
Adjectif-Nom	1291	259 (80%)	103 (92%)	63 (95%)	44 (97%)	34 (97%)
Nom-Adjectif	3455	864 (75%)	448 (87%)	307 (91%)	222 (94%)	181 (95%)
	CORPUS DES RESSOURCES HUMAINES (français)					
Nom-Nom	98	34 (65%)	11 (89%)	5 (95%)	2 (98%)	0 (100%)
Nom-Prép-Nom	4703	2281 (51%)	1268 (73%)	787 (83%)	566 (88%)	417 (91%)
Adjectif-Nom	1290	739 (42%)	476 (63%)	326 (75%)	254 (80%)	210 (84%)
Nom-Adjectif	5768	2809 (51%)	1628 (72%)	1038 (82%)	748 (87%)	578 (90%)
	CORPUS DE FOUILLE DE DONNÉES (anglais)					
Nom-Nom	2070	475 (77%)	223 (89%)	134 (93%)	96 (95%)	70 (97%)
Nom-Prép-Nom	1170	93 (92%)	20 (98%)	8 (99%)	3 (100%)	0 (100%)
Adjectif-Nom	2411	451 (81%)	176 (93%)	96 (96%)	64 (97%)	39 (98%)
	CORPUS DE BIOLOGIE MOLÉCULAIRE (anglais)					
Nom-Nom	22241	7668 (65%)	4600 (79%)	3332 (85%)	2549 (88%)	2081 (90%)
Nom-Prép-Nom	28613	4192 (85%)	1674 (94%)	927 (97%)	617 (98%)	467 (98%)
Adjectif-Nom	23284	6816 (71%)	3781 (84%)	2547 (89%)	1951 (92%)	1545 (93%)

TAB. 1 – *Élagages des différents corpus étudiés. Les pourcentages représentent les proportions d'élagage effectuées.*

particulier. En effet, en moyenne chaque taux d'élagage à partir du corpus des Ressources Humaines supprime une proportion plus faible de collocations contrairement aux trois autres corpus. Deux raisons peuvent expliquer cette situation. Premièrement, le corpus a été écrit par un seul auteur issu de la société PerformanSe. Ainsi, le vocabulaire employé est moins varié que les trois autres corpus. Deuxièmement, les tests de psychologie décrivent des comportements et des observations qui se répètent souvent. Ainsi, de nombreuses tournures de phrases sont identiques, les collocations sont donc globalement plus fréquentes.

5 Acquisition des termes pour la classification conceptuelle

Pour construire une classification conceptuelle, les collocations évoquant des concepts du domaine sont extraites. Le tableau 2 présente des exemples de collocations associées à des concepts à partir des deux corpus en français étudiés.

Afin de valider les collocations extraites, plusieurs catégories de pertinence (ou de non pertinence) ont été identifiées :

- 1**ère catégorie :** La collocation est pertinente pour la classification conceptuelle (exemple du corpus de CVs : " *baccalauréat littéraire*")
- 2**ème catégorie :** La collocation est pertinente mais très spécifique et pas nécessairement adaptée au domaine (exemple du corpus de CVs : "*écosystème marin*")

Corpus des Ressources Humaines	
Collocations	**Concepts associés**
besoin d'information	Communication
capacité d'écoute	Communication
contexte professionnel	Environnement
lieu de travail	Environnement
responsabilité de gestion	Rôle
rôle hiérarchique	Rôle
sentiment de malaise	Stress
tension permanente	Stress
besoin de point de repère	Vous-Même
haute ambition	Vous-Même

Corpus de CVs	
Collocations	**Concepts associés**
aide comptable	Activité Gestion
gestion administrative	Activité Gestion
employé libre service	Activité Commerce
assistant marketing	Activité Commerce
chef de service	Activité Encadrement
direction générale	Activité Encadrement
BEP secrétariat	Compétence Secrétariat
BTS assistante de direction	Compétence Secrétariat
baccalauréat professionnel vente	Compétence Commerce
BTS commerce international	Compétence Commerce

Tab. 2 – *Extrait de la classification conceptuelle construite à partir d'un corpus des Ressources Humaines et de CVs.*

- $3^{\textbf{ème}}$ **catégorie :** La collocation est pertinente mais très générale et pas nécessairement adaptée au domaine (exemple du corpus de CVs : "*situation actuelle*")
- $4^{\textbf{ème}}$ **catégorie :** La collocation est non pertinente (exemple du corpus de CVs : "*jour quotidienne*")
- $5^{\textbf{ème}}$ **catégorie :** L'expert ne peut pas juger si la collocation est pertinente (exemple du corpus de CVs : "*master franchisé*").

Par exemple, les collocations qui sont des instances de concepts peuvent être utilisées pour découvrir des règles d'association entre concepts présents dans les textes. Les concepts peuvent également être utilisés pour construire des patrons d'extraction utiles pour la recherche d'informations.

Par exemple, pour découvrir des règles d'association, les concepts utilisés doivent être précis afin de déterminer des associations éventuelles. Ce travail a des similarités avec les approches de Srikant et Agrawal (1997) qui consistent à utiliser une taxonomie pour généraliser des règles d'association extraites. Dans nos travaux précédents (Azé et Roche (2003); Kodratoff et al. (2003)) et dans la thèse de Azé (2003), les règles d'associations découvertes sont de la forme $concept_1 ... concept_{n-1} \rightarrow concept_n$ où n est le nombre de concepts impliqués dans les règles d'association extraites. Le détail de l'algorithme est présenté dans les travaux de Azé (2003).

À titre d'exemple, nous donnons deux règles d'association extraites à partir du corpus des Ressources Humaines :

"Stress" → "Environnement"

Cette règle signifie que le stress s'exerce par l'intermédiaire de l'environnement. Cette règle d'association, bien que correcte, n'a pas été jugée comme particulièrement intéressante (n'apportant pas d'informations nouvelles) lors de la phase de validation.

Un autre exemple de règle qui a été extraite est donné ci-dessous.

"Implication dans l'entreprise" → "Environnement".

Une telle règle exprimant des informations liées à l'implication dans l'entreprise a été jugée comme intéressante mais qui demande une expertise plus approfondie.

L'extraction des règles d'association s'effectue avec des concepts très précis qui intéressent l'expert. Ainsi, en reprenant les différentes catégories évoquées au début de cette section, les collocations pertinentes afin de découvrir des règles d'association entre concepts sont les collocations de la catégorie 1. Les collocations issues des catégories 2, 3 et 4 sont jugées comme non pertinentes. Enfin, les collocations de la catégorie 5 ne sont pas prises en considération car l'expert n'a pas été apte à les valider. Cette dernière catégorie correspond en fait à des collocations qui sont ambiguës ou qui n'ont pas pu être évaluées par méconnaissance partielle du domaine.

6 Évaluation de la terminologie et taux d'élagage

Nous allons évaluer les résultats sur le seul corpus de CVs. En effet, pour ce seul corpus, toutes les collocations Nom-Adjectif quel que soit le taux d'élagage établi ont été évaluées par un expert.

6.1 Expertise de la terminologie

Le tableau 3 donne le nombre de collocations de type Nom-Adjectif associées à chaque catégorie d'expertise. Nous rappelons que chacune des catégories a été décrite dans la section 5 de cet article.

Le tableau 3 présente les résultats des expertises effectuées selon différents taux d'élagage. Plus les collocations fréquentes sont privilégiées en appliquant un taux d'élagage important et plus la proportion de collocations de la catégorie 1 correspondant aux collocations pertinentes est élevée. À titre d'exemple, si toutes les collocations sont conservées (élagage à un), la proportion de collocations pertinentes est de 56.3% contre plus de 80% en faisant un élagage de quatre, cinq ou six.

Élagage	**catégorie 1**	**catégories 2 et 3**	**catégorie 4**	**catégorie 5**	Total
1	1946 (56.3%)	919 (26.6%)	395 (11.4%)	195 (5.6%)	3455
2	631 (73.0%)	151 (17.5%)	58 (6.7%)	24 (2.8%)	864
3	348 (77.7%)	73 (16.3%)	17 (3.8%)	10 (2.2%)	448
4	256 (83.4%)	36 (11.7%)	8 (2.6%)	7 (2.3%)	307
5	185 (83.3%)	29 (13.1%)	3 (1.3%)	5 (2.2%)	222
6	152 (84.0%)	23 (12.7%)	2 (1.1%)	4 (2.2%)	181

TAB. 3 – *Nombre de collocations dans chacune des catégories.*

6.2 Les mesures de précision, de rappel et de F_{score}

La précision est un critère d'évaluation parfaitement adapté à un cadre de travail non supervisé. La précision permet de calculer la proportion de collocations pertinentes extraites parmi les collocations extraites. En utilisant les notations du tableau 4, la précision est donnée par la formule $\frac{VP}{VP+FP}$. Une précision de 100% signifie que toutes les collocations extraites par le système sont pertinentes.

Une autre mesure typique du domaine de l'apprentissage est le rappel qui calcule la proportion de collocations pertinentes extraites parmi les collocations pertinentes. Le rappel est donné par la formule $\frac{VP}{VP+FN}$. Un rappel de 100% signifie que toutes les collocations pertinentes sont extraites. Ce critère d'évaluation est adapté aux méthodes d'apprentissage supervisées pour lesquelles l'ensemble des exemples positifs (collocations pertinentes) est connu.

	Collocations pertinentes	Collocations non pertinentes
Collocations évaluées comme pertinentes par le système	VP (vrais positifs)	FP (faux positifs)
Collocations évaluées comme non pertinentes par le système	FN (faux négatifs)	VN (vrais négatifs)

TAB. 4 – *Table de contingence à la base des mesures d'évaluation.*

En règle générale, il est important de déterminer un compromis entre le rappel et la précision. Pour cela, nous pouvons utiliser une mesure prenant en compte ces deux critères d'évaluation en calculant le F_{score} (Van-Risbergen (1979)) :

$$F_{score}(\beta) = \frac{(\beta^2 + 1) \times Précision \times Rappel}{(\beta^2 \times Précision) + Rappel} \quad (1)$$

Le paramètre β de la formule (1) permet de régler les influences respectives de la précision et du rappel. Il est très souvent fixé à 1 pour accorder le même poids à ces deux mesures d'évaluation.

Le tableau 5 montre que la précision la plus élevée est établie après un élagage important (un élagage supérieur à quatre fournit une précision supérieure à 85%). Cependant, avec une telle précision assez élevée le rappel est souvent très faible, c'est-à-dire que relativement peu de collocations pertinentes sont extraites. Avec $\beta = 1$, nous pouvons relever dans le tableau 5 que le F_{score} est le plus élevé sans appliquer d'élagage. Ceci est dû au fait que le rappel est peu important dès qu'un élagage même faible est effectué. En effet, comme précisé dans le tableau 1, un seul élagage de deux permet d'empêcher l'extraction de 75% des collocations Nom-Adjectif du corpus de CVs.

Le tableau 6 montre que le fait de faire varier β afin de donner un poids plus important à la précision donne un F_{score} logiquement plus élevé dans le cas d'élagages importants. Ceci montre les limites d'un tel critère d'évaluation car les résultats du F_{score} peuvent singulièrement différer selon la valeur de β. Ainsi, la section suivante présente un autre critère d'évaluation global fondé sur les courbes ROC.

Élagage	**Précision**	**Rappel**	F_{score}
1	59.7%	100%	74.8%
2	75.1%	32.4%	45.3%
3	79.4%	17.9%	29.2%
4	85.3%	13.1%	22.8%
5	85.2%	9.5%	17.1%
6	85.9%	7.8%	14.3%

TAB. 5 – *Précision, Rappel et F_{score} avec $\beta = 1$.*

β	1	1/2	1/3	1/4	1/5	1/6	1/7	1/8	1/9	1/10
1	74.8%	64.9%	62.2%	61.1%	60.6%	60.4%	60.2%	60.1%	60.0%	59.9%
2	45.3%	59.5%	66.4%	69.7%	71.5%	72.5%	73.2%	73.6%	73.9%	74.1%
3	29.2%	47.0%	59.1%	66.1%	70.2%	72.7%	74.3%	75.4%	76.2%	76.8%
4	22.8%	40.7%	55.1%	64.5%	70.5%	74.3%	76.9%	78.7%	80.0%	80.9%
5	17.1%	32.9%	47.4%	58.0%	65.2%	70.1%	73.5%	75.9%	77.7%	79.0%
6	14.3%	28.6%	42.9%	54.1%	62.0%	67.6%	71.6%	74.4%	76.5%	78.1%

TAB. 6 – *F_{score} selon différentes valeurs de β et divers taux d'élagage.*

6.3 Les courbes ROC

Dans cette section et dans les travaux de Ferri et al. (2002), les courbes ROC (Receiver Operating Characteristics) sont présentées. La notion de courbe ROC est initialement issue du traitement du signal. Les courbes ROC sont couramment utilisées dans le domaine de la médecine pour évaluer la validité des tests diagnostiques. Les courbes ROC présentent en abscisse le taux de faux positifs (dans notre cas, taux de collocations non pertinentes) et en ordonnée le taux de vrais positifs (taux de collocations pertinentes). Les courbes ROC sont adaptées aux approches supervisées. Par ailleurs, l'aire sous la courbe ROC (*AUC - Area Under the Curve*), peut être vue comme la mesure globale de l'efficacité d'une mesure d'intérêt. Précisons que le critère relatif à l'aire sous la courbe est équivalent au test statistique de Wilcoxon-Mann-Whitney (voir les travaux de Yan et al. (2003)).

Dans le cas correspondant au classement des collocations en utilisant des mesures statistiques, une courbe ROC idéale correspond au fait d'obtenir toutes les collocations pertinentes en début de liste et toutes les collocations non pertinentes en fin de liste. Cette situation correspond à une AUC de 1. La diagonale correspond aux performances d'un système aléatoire, progrès du taux de vrais positifs s'accompagnant d'une dégradation équivalente du taux de faux positifs. Une telle situation correspond à AUC = 0.5. Enfin, si les collocations triées par intérêt décroissant sont telles que toutes les collocations pertinentes sont précédées par les non pertinentes, alors AUC = 0. Une mesure d'intérêt efficace pour ordonner les collocations consiste donc à obtenir une aire sous la courbe ROC la plus importante possible ce qui est strictement équivalent à minimiser la somme des rangs des exemples positifs.

L'avantage des courbes ROC provient de l'absence de sensibilité dans certains cas où un déséquilibre entre le nombre d'exemples positifs et d'exemples négatifs est rencontré. Illustrons ce fait avec l'exemple suivant. Supposons que nous ayons 100 exemples (collocations). Dans le premier cas, nous avons un déséquilibre entre les exemples positifs et négatifs avec 1 seul exemple positif et 99 exemples négatifs. Dans le second cas, nous avons 50 exemples positifs et 50 exemples négatifs. Supposons que pour ces deux cas, les exemples positifs soient tous placés en tête de listes établies par des mesures statistiques, c'est-à-dire que les collocations pertinentes sont toutes situées en début de liste.

Dans les deux cas, les courbes ROC sont strictement identiques avec une AUC égale à 1 (voir figure 1, (a) et (b)). Ainsi, le fait d'avoir les collocations pertinentes en début de listes est mis en valeur par les courbes ROC et les AUC. L'intérêt principal des courbes ROC est le fait de ne pas tenir compte d'un éventuel déséquilibre entre le nombre de collocations pertinentes et non pertinentes. Dans le cas du calcul du F_{score} (avec $\beta = 1$) qui prend en compte les mesures

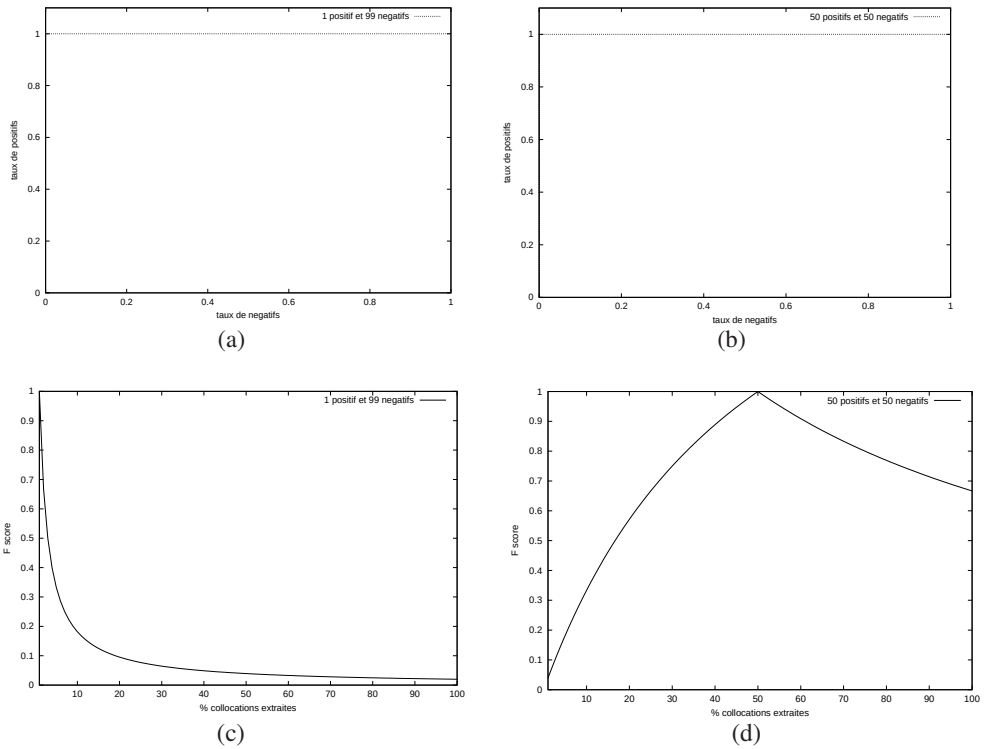

FIG. 1 – *Courbe ROC (a) et F_{score} (c) avec 1 exemple positif placé en tête de liste et 99 exemples négatifs placés en fin de liste. Courbe ROC (b) et F_{score} (d) avec 50 exemples positifs en tête de liste et 50 exemples négatifs en fin de liste. Pour le calcul du F_{score}, le paramètre β est fixé à 1.*

de rappel et de précision, avec ces deux mêmes situations, nous obtenons deux courbes extrêmement différentes (voir figure 1, (c) et (d)). Ainsi, les déséquilibres entre exemples positifs et négatifs influencent fortement le F_{score} contrairement aux courbes ROC.

D'un élagage à l'autre, le taux de collocations pertinentes et non pertinentes peut se révéler fort différent, ce qui signifie que nous sommes en présence d'un déséquilibre entre les classes. Par exemple, en appliquant un élagage de six, 84% des collocations sont pertinentes contre 56% si aucun élagage n'est effectué (voir tableau 3). Le tableau 7 calcule les différentes AUC en choisissant différents taux d'élagage. Dans ce cas, l'utilisation du critère d'évaluation fondé sur les courbes ROC et les AUC qui ne sont pas sensibles au déséquilibre entre les classes et qui prennent explicitement en compte le rang des collocations extraites est particulièrement bien adapté.

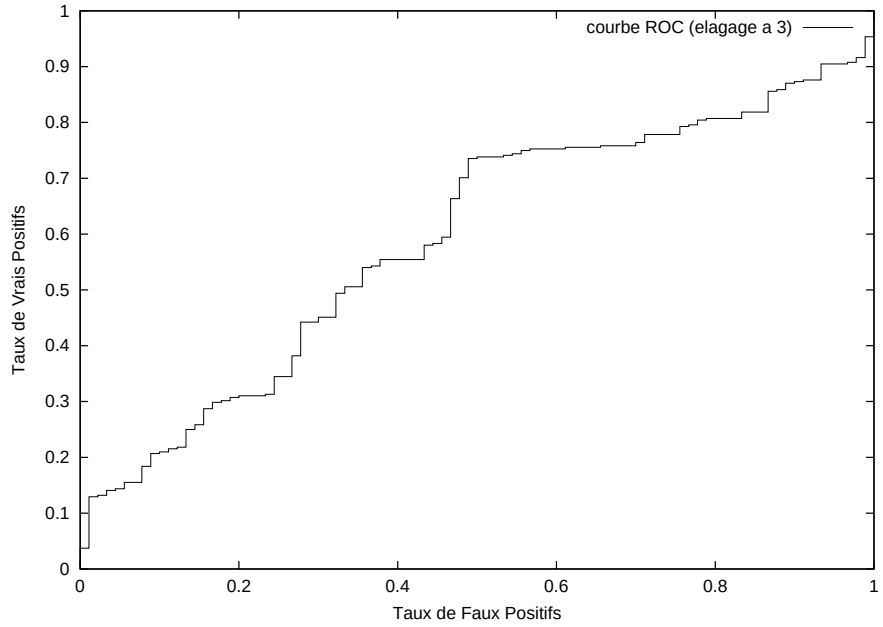

FIG. 2 – *Courbe ROC avec un élagage à 3.*

Le tableau 7 montre que l'élagage le plus adapté en terme d'AUC correspond à un élagage à trois pour les collocations Nom-Adjectif du corpus de CVs. La figure 2 montre la courbe ROC relative à un élagage à trois. Ce critère objectif fondé sur l'AUC correspond au choix souvent empirique d'élagage à trois appliqué dans les travaux de Jacquemin (1997); Thanopoulos et al. (2002).

Élagage	**AUC**
1	0.4538
2	0.5324
3	**0.5905**
4	0.5012
5	0.5432
6	0.5447

TAB. 7 – *AUC selon le nombre d'occurrences.*

7 Conclusion et perspectives

L'étude expérimentale menée dans cet article permet de discuter le choix du taux d'élagage pour la terminologie. Différents critères d'évaluation existent tels que la précision, le rappel et bien entendu le F_{score} qui permet de prendre en compte ces deux critères. Le défaut de ce dernier critère réside dans le choix pas toujours évident du paramètre le plus adapté pour privilégier la précision ou le rappel dans le calcul du F_{score}. Ainsi, dans cet article, nous proposons d'utiliser les courbes ROC et l'aire sous celles-ci afin d'évaluer au mieux le choix de l'élagage à effectuer. Ce critère permettant d'estimer rigoureusement l'élagage le plus adapté n'est pas sensible au déséquilibre entre les classes. Ceci est important car selon l'élagage effectué la proportion de collocations pertinentes et non pertinentes peut être très différente.

Nos expérimentations ont alors montré sur un corpus de CVs, qu'un élagage de trois semble bien adapté. Dans nos prochains travaux, nous proposons de comparer ce résultat à partir des autres corpus étudiés. Pour cela, il sera nécessaire d'effectuer une expertise complète des collocations d'autres domaines, ce qui demandera un travail conséquent aux experts.

Références

Amrani, A., Y. Kodratoff, et O. Matte-Tailliez (2004). A semi-automatic system for tagging specialized corpora. In *Proceedings of PAKDD'04*, pp. 670–681.

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

Azé, J. et M. Roche (2003). Une application de la fouille de textes : l'extraction des règles d'association à partir d'un corpus spécialisé. *Revue RIA-ECA numéro spécial EGC03 17*, 283–294.

Azé, J. (2003). *Extraction de Connaissances dans des Données Numériques et Textuelles*. Ph. D. thesis, Université de Paris 11.

Bourigault, D. (1993). Analyse syntaxique locale pour le repérage de termes complexes dans un texte. *T.A.L. 34*(2), 105–118.

Bourigault, D. et C. Fabre (2000). Approche linguistique pour l'analyse syntaxique de corpus. *Cahiers de Grammaires 25*, 131–151.

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

Daille, B. (1994). *Approche mixte pour l'extraction automatique de terminologie : statistiques lexicales et filtres linguistiques*. Ph. D. thesis, Université Paris 7.

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

Dunning, T. E. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics 19*(1), 61–74.

Evans, D. et C. Zhai (1996). Noun-phrase analysis in unrestricted text for information retrieval. In *Proceedings of the ACL*, Santa Cruz, US, pp. 17–24.

Ferri, C., P. Flach, et J. Hernandez-Orallo (2002). Learning decision trees using the area under the ROC curve. In *Proceedings of ICML'02*, pp. 139–146.

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

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., J. Azé, M. Roche, et O. Matte-Tailliez (2003). Des textes aux associations entre les concepts qu'ils contiennent. *Dans les actes des XXXVièmes Journées de Statistique (résumé) Volume 2, p599-602 et dans le numéro spécial de la revue RNTI "Entreposage et Fouille de données" 1*, 171–182.

Roche, M. (2004). *Intégration de la construction de la terminologie de domaines spécialisés dans un processus global de fouille de textes*. Ph. D. thesis, Université de Paris 11.

Roche, M., J. Azé, Y. Kodratoff, et M. Sebag (2004a). Learning interestingness measures in terminology extraction. A ROC-based approach. In *Proceedings of "ROC Analysis in AI" Workshop (ECAI 2004), Valencia, Spain*, pp. 81–88.

Roche, M., T. Heitz, O. Matte-Tailliez, et Y. Kodratoff (2004b). EXIT : Un système itératif pour l'extraction de la terminologie du domaine à partir de corpus spécialisés. In *Proceedings of JADT'04*, Volume 2, pp. 946–956.

Roche, M., O. Matte-Tailliez, et Y. Kodratoff (2004c). Étude de Mesures de Qualité pour Classer les Termes Extraits de Corpus Spécialisés. In *Actes de INFORSID'04*, pp. 371–386.

Smadja, F. (1993). Retrieving collocations from text : Xtract. *Computational Linguistics 19*(1), 143–177.

Srikant, R. et R. Agrawal (1997). Mining generalized association rules. *Future Generation Computer Systems 13*(2–3), 161–180.

Thanopoulos, A., N. Fakotakis, et G. Kokkianakis (2002). Comparative Evaluation of Collocation Extraction Metrics. In *Proceedings of LREC'02*, Volume 2, pp. 620–625.

Van-Risbergen, C. (1979). *Information Retrieval*. 2nd edition, London, Butterworths.

Yan, L., R. Dodier, M. Mozer, et R. Wolniewicz (2003). Optimizing classifier performance via an approximation to the Wilcoxon-Mann-Whitney statistic. In *Proceedings of ICML'03*, pp. 848–855.

Summary

The choice of the pruning rate is crucial for acquisition of a good terminology of a specialized corpus. This paper presents an experimental study to determinate the pruning rate well adapted. Several evaluation measures can be used such as the precision, the recall and F_{score}. This study is based on another evaluation measure which seems well adapted for the terminology extraction : the ROC curves (Receiver Operating Characteristics).