Filtrages syntaxiques de co-occurrences pour la représentation vectorielle de documents

Romaric Besançon, Martin Rajman
Laboratoire d'Intelligence Artificielle
Faculté Informatique et Communications
École Polytechnique Fédérale de Lausanne, (IN) Ecublens 1015 Lausanne
{Romaric.Besancon,Martin.Rajman}@epfl.ch

Mots-clefs – Keywords

Similarités textuelles, représentation vectorielle de textes, sémantique distributionnelle, contexte de co-occurrence
Textual similarities, vector space representation, distributional semantics, co-occurrence context

Résumé - Abstract

L'intégration de co-occurrences dans les modèles de représentation vectorielle de documents s'est avérée une source d'amélioration de la pertinence des mesures de similarités textuelles calculées dans le cadre de ces modèles (Rajman *et al.*, 2000; Besançon, 2001). Dans cette optique, la définition des contextes pris en compte pour les co-occurrences est cruciale, par son influence sur les performances des modèles à base de co-occurrences. Dans cet article, nous proposons d'étudier deux méthodes de filtrage des co-occurrences fondées sur l'utilisation d'informations syntaxiques supplémentaires. Nous présentons également une évaluation de ces méthodes dans le cadre de la tâche de la recherche documentaire.

The integration of co-occurrence information in the vector-space representation models for texts has proven to improve the relevance of textual similarities (Rajman *et al.*, 2000; Besançon, 2001). In this framework, the definition of what is the context considered for the co-occurrences is an important issue. In this paper, we provide the study of two methods for the filtering of the co-occurrences, both using additional syntactic information. We also present an evaluation of these methods in the framework of information retrieval.

1 Introduction

La notion de similarité entre textes est très souvent utilisée dans les applications du traitement de la langue destinées à l'exploitation de collections de documents de grande taille. Par exemple, en recherche documentaire, les documents pertinents retournés par le moteur de recherche peuvent être définis comme les plus proches de la requête selon une certaine mesure de similarité (Salton & McGill, 1983) ; de même, le regroupement incrémental de documents en classes en fonction de leurs similarités peut permettre une structuration automatique de bases de données textuelles à l'aide de techniques de classification automatique non supervisée (Salton *et al.*, 1975).

La notion de similarité entre documents est évidemment fortement liée au choix de la méthode de représentation des textes. La représentation la plus utilisée est la représentation vectorielle (mise en œuvre, en particulier, dans les systèmes de recherche documentaire tels que SMART (Salton, 1971)), dans le cadre de laquelle un document est représenté par un vecteur dans un espace vectoriel dont les dimensions sont associées à des unités linguistiques spécifiques (mot, *stems*, lemmes, etc). La similarité entre documents est alors évaluée par une mesure de similarité définie sur cet espace vectoriel.

Des améliorations peuvent également être apportées dans ce modèle de représentation par l'intégration de connaissances externes (Besançon, 2001). En particulier, dans l'optique de la sémantique distributionnelle (Rajman *et al.*, 2000), des connaissances de co-occurrences peuvent être utilisées pour intégrer plus d'informations sémantiques dans la représentation. L'objectif de cet article est d'étudier l'influence de la méthode de sélection des co-occurrences considérées sur la qualité des représentations et des mesures de similarité entre documents.

Dans la section 2, nous présentons brièvement le modèle de représentation vectoriel standard, ainsi que le modèle DSIR, qui étend le modèle standard par l'intégration de co-occurrences dans la représentation des documents. Dans la section 3, nous présentons deux méthodes de filtrage des co-occurrences utilisant des informations syntaxiques pour déterminer quelles seront les co-occurrences effectivement considérées. Enfin, dans la section 4, nous proposons une évaluation de ces méthodes de filtrage pour la tâche de la recherche documentaire.

2 Le modèle de représentation DSIR

2.1 Modèle vectoriel

Dans le cadre du modèle vectoriel standard (VS), un document d est représenté par un vecteur $d^{VS} = (d_1^{VS}, \ldots, d_{|T|}^{VS})$, appelé *profil lexical*, dans lequel la j^{e} composante d_j^{VS} représente le poids (ou importance), dans le document d, du terme d'indexation t_j associé à la j^{e} dimension de l'espace vectoriel. D'une façon générale, le poids est le plus souvent une fonction de la fréquence du terme dans le document et se décompose habituellement en une pondération locale, une pondération globale et un facteur de normalisation (par rapport à la longueur du document). Pour nos expériences, nous avons utilisé le schéma de pondération *ltn* de SMART (Salton & Buckley, 1988; Singhal, 1997) :

$$d_j^{VS} = w_j = idf \times (1 + log(tf)) \tag{1}$$

où tf est la fréquence du mot dans le document et idf est le facteur de fréquence en document inverse $idf = \log \frac{1}{df}$, où df est la fréquence en documents du terme (c'est-à-dire le nombre de

documents dans lesquels le terme apparaît). Dans ce cas, le facteur de pondération locale est $1 + log(tf)$, le facteur de pondération globale est idf (ce facteur permet d'accorder un poids plus important aux termes qui apparaissent moins fréquemment dans la collection et sont donc plus utiles pour la discrimination). Aucun facteur de normalisation n'est intégré directement dans cette pondération mais une normalisation implicite est effectuée en utilisant la mesure de similarité du cosinus, indépendante de la norme.

2.2 Modèle à base de co-occurrences

Le modèle DSIR est un modèle vectoriel permettant d'intégrer des informations sémantiques supplémentaires par l'utilisation de co-occurrences (Rajman & Bonnet, 1992; Rajman *et al.*, 2000; Besançon, 2001).

Dans le cadre de ce modèle, les unités linguistiques u_i considérées sont représentées par un vecteur $c_i = (c_{i1}, \ldots, c_{i|T|})$, appelé *profil de co-occurrence*, dont chaque composante c_{ij} est la fréquence de co-occurrence de l'unité linguistique u_i avec un terme d'indexation t_j. Un document d est alors représenté comme la somme pondérée des profils de co-occurrence des unités linguistiques qu'il contient, c'est-à-dire par un vecteur $d^{DS} = (d_1^{DS}, \ldots, d_{|T|}^{DS})$ où chaque d_j^{DS} est défini par :

$$d_j^{DS} = \sum_{u_i \in d} w_i \; c_{ij}$$

où la pondération w_i est celle définie par l'équation (1).

Notons que, dans le modèle DSIR, les termes effectivement présents dans les documents ne sont pris en compte qu'indirectement, par le biais de leur profil de co-occurrence. Pour cette raison, un modèle DSIR hybride prenant en compte à la fois les occurrences et les co-occurrences des termes dans les documents a également été proposé (Rungsawang, 1997; Rajman *et al.*, 2000). Dans ce modèle un document est représenté par un vecteur dont la j^{e} composante est définie par :

$$d_j^{DS} = \alpha \; w_j + (1 - \alpha) \sum_{u_i \in d} w_i \; c_{ij} \qquad (2)$$

où α est le coefficient d'hybridation entre le modèle DSIR pur et le modèle VS.

3 Filtrage syntaxique des co-occurrences

Le calcul des fréquences de co-occurrence c_{ij} dépend bien évidemment en premier lieu des choix effectués pour ce qui est de la sélection des relations de co-occurrence considérées, et donc, en particulier, de la définition des contextes qui seront pris en compte pour le calcul de ces co-occurrences. Ces contextes peuvent être de trois types : documentaire, positionnel ou syntaxique.

L'approche la plus simple est de considérer soit un contexte positionnel, soit un contexte documentaire et donc de calculer, à partir d'un corpus de référence, toutes les co-occurrences entre toutes les unités linguistiques prises deux à deux dans une fenêtre de taille donnée (contexte positionnel) ou sur une unité documentaire donnée, comme la phrase ou le paragraphe par exemple (contexte documentaire).

Ces deux approches simples peuvent néanmoins s'avérer insuffisantes. En effet, dans l'une comme dans l'autre, des co-occurrences non linguistiquement pertinentes peuvent être prises en compte. Prenons par exemple le contexte représenté par la phrase suivante :

« *L'acteur porte un masque grimaçant de théâtre antique.* »

Une première phase de pré-traitement permet d'identifier les unités linguistiques qui composent la phrase. Par exemple, si les unités considérées sont les lemmes associés à leur étiquette morpho-syntaxique, on obtient :

$le|_{\mathrm{Ds}}$ $acteur|_{\mathrm{Ncms}}$ $porter|_{\mathrm{Vs}}$ $un|_{\mathrm{Dms}}$ $masque|_{\mathrm{Ncms}}$ $grimaçant|_{\mathrm{Ams}}$ $de|_{\mathrm{S}}$ $théâtre|_{\mathrm{Ncms}}$ $antique|_{\mathrm{As}}$

Si l'on calcule alors les co-occurrences sur l'ensemble de la phrase, des co-occurrences pertinentes comme (*masque–grimaçant*) ou (*théâtre–antique*) seront effectivement sélectionnées, mais des co-occurrences croisées comme (*acteur–antique*) ou (*théâtre–grimaçant*), qui semblent en revanche beaucoup moins pertinentes (et, en tout cas, ne sont pas suggérées par la structure de la phrase) seront également prises en compte.

L'utilisation, dans la définition des contextes, d'une information supplémentaire sur les dépendances syntaxiques entre les unités linguistiques de la phrase permet une définition plus fine des co-occurrences à considérer (Rajman, 1995; Rungsawang, 1997).

Nous présentons dans les deux sections suivantes deux approches possibles : la première repose sur l'idée d'un *filtrage* des co-occurrences, l'objectif étant d'éliminer certaines des co-occurrences non souhaitées, sans faire d'hypothèses sur les co-occurrences restantes ; la seconde repose sur l'idée d'une *sélection* des co-occurrences, l'objectif étant cette fois-ci de ne garder que les co-occurrences syntaxiquement fondées, et de rejeter toutes les autres.

Dans les deux cas, les relations de co-occurrence prises en compte seront synthétisées dans un *graphe de co-occurrences*, dans lequel les nœuds sont associés aux unités linguistiques considérées et les arcs représentent les relations de co-occurrence. Un exemple de tels graphes est donné dans le tableau récapitulatif de la figure 1 à la fin de la section 3.

3.1 Filtrage par les groupes syntaxiques

Dans l'exemple donné ci-dessus, une catégorie de co-occurrences qui paraissent clairement non pertinentes sont les co-occurrences entre un nom et un adjectif qui qualifie un autre nom de la phrase (des co-occurrences du type (*acteur-antique*) ou (*théâtre-grimaçant*)). L'objectif de la méthode de filtrage proposée dans cette section est donc d'éliminer ce type de co-occurrences.

Pour ce faire, nous utilisons un analyseur syntaxique de surface (*shallow parser*) pour produire les groupes syntaxiques élémentaires correspondant à la structure de la phrase, associés chacun à une unité linguistique particulière représentant la tête du groupe (en pratique, la tête d'un groupe nominal est le nom de ce groupe, et la tête d'un groupe verbal est le verbe principal – *i.e.* pas les auxiliaires). Les seules co-occurrences considérées sont alors les co-occurrences entre unités linguistiques d'un même groupe syntaxique ou entre têtes de différents groupes syntaxiques (Besançon *et al.*, 1999). Cela permet effectivement d'éviter de prendre en compte des co-occurrences entre des unités linguistiques qui seraient toutes deux des dépendances dans des groupes syntaxiques différents, ou entre des unités linguistiques qui seraient l'une la tête d'un groupe syntaxique et l'autre une dépendance dans un autre groupe syntaxique.

Après lemmatisation, un découpage en groupes syntaxiques de la phrase d'exemple pourrait

être le suivant[1] :

((le|Ds **acteur*|Ncms *) (*porter*|Vs *) (un*|Dms **masque*|Ncms *grimaçant*|Ams *) (de*|S **théâtre*|Ncms *antique*|As *))*

où les groupes syntaxiques sont délimités par les parenthèses et les têtes des groupes sont identifiées par le symbole « * » antéposée.

3.2 Sélection par les relations syntaxiques

La seconde approche proposée repose sur l'idée de sélectionner les « bonnes » co-occurrences, *i.e.* les co-occurrences à conserver en raison de leur pertinence syntaxique.

La méthode retenue pour cette approche repose sur l'utilisation des résultats d'une analyse syntaxique produisant différentes relations syntaxiques entre les unités linguistiques de la phrase, comme par exemple les relations de type sujet-verbe (SUJ), verbe-objet (OBJ), complément de nom (CNOM), ou qualification d'un nom par un adjectif (ADJ). Les seules co-occurrences qui sont alors considérées sont celles entre les unités qui sont effectivement reliées par une relation syntaxique identifiée.

Par exemple, les relations syntaxiques produites par l'analyseur syntaxique XeLDA de Xerox (Xerox, 1990) pour la phrase d'exemple sont :

SUJ(*acteur, porter*)
OBJ(*porter, masque*)
ADJ(*masque, grimaçant*)
ADJ(*théâtre, antique*)
CNOM(*masque, théâtre*)

Les graphes de co-occurrences produits pour la phrase d'exemple pour chacune des méthodes présentées sont indiqués dans le tableau récapitulatif de la figure 1. Dans tous les cas, un prétraitement ne gardant que les lemmes des noms, des verbes et des adjectifs a été réalisé.

Deux constatations peuvent être faites sur la base de l'exemple traité : d'une part, les méthodes de filtrage/sélection réduisent comme prévu le nombre de co-occurrences prises en compte, en éliminant les co-occurrences entre termes non liés syntaxiquement. D'autre part, on remarque qu'elles présentent néanmoins certains inconvénients : en particulier, la sélection sur la base des relations syntaxiques supprime les co-occurrences entre sujet et objet de l'action (ici « *acteur* » et « *masque* »), de même que des co-occurrences comme « *acteur–théâtre* », qui pourraient toutes deux sembler sémantiquement pertinentes.

4 Évaluation

L'évaluation des méthodes de filtrage syntaxique des co-occurrences a été effectuée pour la tâche de la recherche documentaire. Les résultats présentés ici ont été obtenus sur des données provenant de la seconde campagne d'évaluation AMARYLLIS (Coret *et al.*, 1997; Landi *et al.*, 1998), présentées dans le tableau 1.

[1] Ce parenthésage a été produit par l'analyseur syntaxique Sylex.

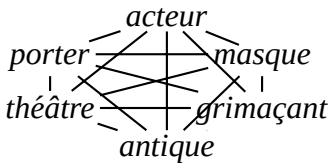
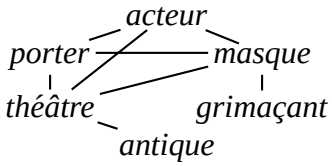
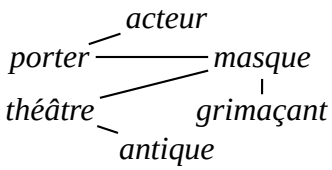
	« *L'acteur porte un masque grimaçant de théâtre antique* »	
	contexte	co-occurrences
(a) contexte positionnel	*acteur porter masque grimaçant théâtre antique*	
(b) filtrage par les groupes syntaxiques	*(*acteur)* *(*porter)* *(*masque grimaçant)* *(*théâtre antique)*	
(c) sélection par les relations syntaxiques	SUJ(acteur,porter) OBJ(porter,masque) ADJ(masque,grimaçant) ADJ(théâtre,antique) CNOM(masque,théâtre)	

FIG. 1 – Exemples de graphes de co-occurrences , (a) toutes les co-occurrences, (b) avec filtrage des co-occurrences par les groupes syntaxiques, (c) avec sélection des co-occurrences sur les relations syntaxiques.

Corpus	Sujet	type	nom	Nb docs	Nb mots
LRSA	extraits de livres sur la Mélanésie	documents requêtes	md1 mt1	355 15	428803 1301
OFIL	articles extraits du journal « *Le Monde* »	documents requêtes	od1 ot1	11016 26	4915890 1412
INIST	notes bibliographiques	documents requêtes	od1 ot1	163308 30	13678485 2022

TAB. 1 – Données utilisées pour les tests, provenant de la campagne AMARYLLIS

Filtrage des co-occurrences par les groupes syntaxiques Dans une première phase, les corpus ont été analysés à l'aide d'un analyseur syntaxique (SYLEX, de INGENIA-LN (Constant, 1995)) pour déterminer les catégories morpho-syntaxiques des mots ainsi que leurs lemmes. Des lexiques ont alors été extraits, comprenant les lemmes des noms, verbes et adjectifs apparaissant dans les corpus. Ces lexiques forment l'ensemble U des unités linguistiques qui seront considérées. L'ensemble T des termes d'indexation a été créé à partir de U en sélectionnant les unités linguistiques de fréquence en documents comprise entre $\frac{|D|}{100}$ et $\frac{|D|}{10}$, avec D l'ensemble des documents du corpus. Les matrices de co-occurrence (de dimension $|U|\times|T|$) ont été construites pour les corpus LRSA, OFIL et INIST, en utilisant des contextes positionnels (matrices C_m, C_o, C_i respectivement), puis en utilisant le filtrage par les groupes syntaxiques (matrices C_m^{gs}, C_o^{gs}, C_i^{gs}). Le tableau 2 présente la taille des lexiques et des matrices de co-occurrences, en indiquant également pour ces dernières le taux de remplissage t_r (c'est-à-dire le pourcentage de co-occurrences de fréquence non nulle parmi les $|U| \times |T|$ co-occurrences possibles), et le taux t_d^{gs} de diminution du nombre de co-occurrences prises en compte par rapport à l'approche sans filtrage syntaxique. Il est à noter que l'utilisation du filtrage syntaxique permet une réduction d'environ 25 à 30% des matrices manipulées.

Les performances obtenues avec le modèle de représentation DSIR hybride (avec un coefficient d'hybridation $alpha = 0.5$) sont présentés pour les trois corpus dans les tableaux 2. Les mesures d'évaluation choisies sont les suivantes : précision moyenne (notée *avg_p*), R-précision[2] (notée *R_p*), les précisions à plusieurs points de coupure (la précision à N documents est notée pN), le nombre total de documents pertinents retournés par le système (notée *relret*), et le rappel final à 1000 documents (noté r1000). Les résultats sont présentés en indiquant les pourcentages d'amélioration par rapport à un résultat de base (présent dans la première colonne des tableaux), avec le coefficient de risque p_w du test de Wilcoxon qui lui est associé (Van Rijsbergen, 1979). Cette valeur indique la confiance que l'on accorde au fait que la différence mesurée n'est pas due au hasard (plus la valeur de p_w est petite, plus l'hypothèse que la différence médiane est nulle peut être rejetée, et on peut donc conclure que les résultats sont significativement différents).

taille des lexiques et des matrices de co-occurrences

	LRSA	OFIL	INIST
unités			
$\|U\|$	9762	28691	23390
$\|T\|$	4635	2796	3398
matrice	C_m	C_o	C_i
taille	3 Mo	26 Mo	10 Mo
t_r	1.66%	8.11%	3.06%
matrice	C_m^{gs}	C_o^{gs}	C_i^{gs}
taille	2 Mo	20 Mo	7 Mo
t_r	1.14%	6.16%	3.04%
t_d^{gs}	31.7%	24.1%	23.8%

LRSA

	$C_m\ \alpha = 0.5$	$C_m^{gs}\ \alpha = 0.5$
avg_p	0.3942	**0.4048** (+2.69%)$_{p_w=0.012}$
R_p	0.3993	**0.4087** (+2.35%)$_{p_w=0.44}$
p5	0.6533	0.6667 (+2.05%)$_{p_w=0.75}$
p10	**0.6333**	0.62 (-2.15%)$_{p_w=0.38}$
p15	0.5422	0.5422 (+0%)$_{p_w=1}$
p20	0.47	**0.4867** (+3.55%)$_{p_w=0.16}$
p30	0.3911	**0.4022** (+2.84%)$_{p_w=0.31}$
p100	0.19	0.1927 (+1.42%)$_{p_w=0.58}$
p200	0.1197	**0.12** (+0.251%)$_{p_w=0.31}$
p500	0.0509	**0.0515** (+1.18%)$_{p_w=0.12}$
p1000	0.0255	**0.0257** (+0.784%)$_{p_w=0.12}$
relret	382	**386** (+1.05%)$_{p_w=0.12}$
r1000	0.9031	**0.9125** (+1.05%)$_{p_w=0.12}$

INIST

	$C_i\ \alpha = 0.5$	$C_i^{gs}\ \alpha = 0.5$
avg_p	0.1095	**0.111** (+1.37%)$_{p_w=0.39}$
R_p	0.1581	0.1595 (+0.886%)$_{p_w=0.53}$
p5	0.3	0.3133 (+4.43%)$_{p_w=0.55}$
p10	0.24	**0.2633** (+9.71%)$_{p_w=0.15}$
p15	0.2156	0.2133 (-1.08%)$_{p_w=0.64}$
p20	**0.2083**	0.2067 (-0.774%)$_{p_w=0.89}$
p30	0.1833	**0.18** (-1.83%)$_{p_w=0.45}$
p100	0.103	**0.107** (+3.88%)$_{p_w=0.18}$
p200	0.071	**0.0748** (+5.35%)$_{p_w=0.018}$
p500	0.0317	**0.0324** (+2.21%)$_{p_w=0.12}$
p1000	0.0159	**0.0162** (+1.89%)$_{p_w=0.074}$
relret	476	**486** (+2.1%)$_{p_w=0.1}$
r1000	0.3383	**0.3454** (+2.1%)$_{p_w=0.039}$

OFIL

$C_o\ \alpha = 0.5$	$C_o^{gs}\ \alpha = 0.5$
0.196	0.1867 (-4.98%)$_{p_w=0.84}$
0.2333	0.2308 (-1.08%)$_{p_w=0.81}$
0.3538	0.3308 (-6.95%)$_{p_w=0.5}$
0.2769	0.2846 (+2.78%)$_{p_w=0.64}$
0.2462	0.2333 (-5.53%)$_{p_w=0.23}$
0.2269	0.2154 (-5.34%)$_{p_w=0.18}$
0.1949	0.1949 (+0%)$_{p_w=0.76}$
0.1008	0.0969 (-4.02%)$_{p_w=0.11}$
0.0604	0.0581 (-3.96%)$_{p_w=0.22}$
0.0248	0.0246 (-0.813%)$_{p_w=0.73}$
0.0124	0.0123 (-0.813%)$_{p_w=0.82}$
322	320 (-0.625%)$_{p_w=0.57}$
0.5486	0.5451 (-0.625%)$_{p_w=0.57}$

TAB. 2 – Résultats sur les corpus LRSA, OFIL et INIST pour le filtrage des co-occurrences par les groupes syntaxiques (les résultats en gras indiquent les résultats significativement meilleurs ($p_w < 0.5$))

[2]La R-précision est la précision obtenue pour un nombre de documents retournés correspondant au nombre de documents pertinents présents dans la base. Donc en particulier, dans ce cas, la précision est égale au rappel.

Une première analyse globale de ces résultats indique que le filtrage par les groupes syntaxiques ne change pas les performances de façon très significative. Ce résultat est en lui-même intéressant car il montre que, malgré une réduction de l'information de co-occurrence de l'ordre de un quart à un tiers, le système ne subit aucune dégradation significative des performances. Il semble donc que la méthode de filtrage choisie est efficace, et permet de n'éliminer majoritairement que des co-occurrences qui n'apportent pas d'autres d'informations utiles pour la représentation que celles déjà prises en compte par les co-occurrences conservées.

Une analyse plus attentive montre même qu'on observe en fait de légères améliorations ($< 5\%$) sur les corpus LRSA et INIST, pour lesquelles les coefficients de Wilcoxon indiquent qu'elles sont significatives (*i.e.* elles ne sont pas dues au hasard). Cela montre que les co-occurrences supprimées peuvent également correspondre à du « bruit », *i.e.* de l'information non utile pour la représentation des documents et dont la suppression permet donc une amélioration des performances. Notons que, pour le corpus OFIL, les performances sont plutôt légèrement dégradées, mais les coefficients de Wilcoxon associés indiquent que cette dégradation n'est souvent pas significative.

Sélection des co-occurrences par les relations syntaxiques Pour cette seconde méthode, la production des relations syntaxiques a été réalisée à l'aide de l'analyseur syntaxique XeLDA de Xerox (Xerox, 1990). L'ensemble des unités linguistiques et l'ensemble des termes d'indexation sont donc différents de ceux utilisés dans les évaluations précédentes. Les principales caractéristiques des données utilisées sont présentées dans le tableau 3 où l'on trouve les tailles des ensembles d'unités linguistiques et celles des matrices de co-occurrences construites sur le corpus OFIL. C_{ox} représente la matrice sans filtrage syntaxique et C_{ox}^{rs} la matrice avec sélection sur les relations syntaxiques. Le pré-traitement syntaxique étant relativement long, les tests sur les autres corpus n'ont pas été effectués.

L'analyse de ces données indique de façon très claire que la sélection par les relations syntaxiques est beaucoup plus restrictive que le filtrage par les groupes syntaxiques. Le nombre de co-occurrences conservées est en effet réduit de plus de 80%.

Pour ce qui est des performances, les résultats de la recherche obtenus pour le corpus OFIL sont présentés dans le tableau 3. L'analyse de ces résultats permet de conclure à une forte diminution (statistiquement significative) des performances. Cette diminution est très probablement liée à une trop forte réduction de l'information de co-occurrence utilisée et peut être également due à la qualité moyenne des relations syntaxiques extraites, qui ne sont pas toutes identiquement fiables.

Il apparaît donc que, si un filtrage syntaxique peut être bénéfique, la mise en œuvre de contraintes de sélection trop fortes entraîne une diminution des performances. Une interprétation possible de cet état de fait peut être que certaines co-occurrences, qui ne reposent pas sur des relations syntaxiques, peuvent malgré tout correspondre à une information sémantique sous-jacente, et qu'un filtrage trop brutal entraîne la perte de cette information utile à la bonne représentation des documents.

Le mécanisme de filtrage des co-occurrences doit donc trouver un juste équilibre entre la nécessaire élimination de co-occurrences inutiles ou génératrices de bruit dans la représentation d'une part et la conservation de la majorité de l'information utile à la représentation. Si le filtrage devient trop sélectif, la matrice de co-occurrence se creuse et devient donc plus discriminante, mais au-delà d'un certain seuil de réduction, le pouvoir de discrimination accru ne semble plus

taille des lexiques et des matrices de co-occurrences

	OFIL
unités	
$\|U\|$	20158
$\|T\|$	1992
matrice	C_{ox}
taille	17 Mo
t_r	10.6%
matrice	C_{ox}^{rs}
taille	3 Mo
t_r	2.0%
t_d	81.1%

	C_{ox} $\alpha = 0.5$	C_{ox}^{rs} $\alpha = 0.5$	
avg_p	**0.1077**	0.0717	(-50.2%)$_{p_w=0.0018}$
R_p	**0.1509**	0.1051	(-43.6%)$_{p_w=0.0099}$
p5	**0.192**	0.152	(-26.3%)$_{p_w=0.11}$
p10	**0.188**	0.136	(-38.2%)$_{p_w=0.037}$
p15	**0.1653**	0.1227	(-34.7%)$_{p_w=0.02}$
p20	**0.154**	0.108	(-42.6%)$_{p_w=0.0052}$
p30	**0.1293**	0.0987	(-31%)$_{p_w=0.015}$
p100	**0.0684**	0.052	(-31.5%)$_{p_w=0.001}$
p200	**0.0426**	0.0324	(-31.5%)$_{p_w=0.0002}$
p500	**0.0179**	0.0141	(-27%)$_{p_w=0.00017}$
p1000	**0.009**	0.007	(-28.6%)$_{p_w=0.0002}$
relret	**224**	176	(-27.3%)$_{p_w=0.00017}$
r1000	**0.3875**	0.3045	(-27.3%)$_{p_w=0.00034}$

TAB. 3 – Résultats sur le corpus OFIL pour la sélection des co-occurrences par les relations syntaxiques

contrebalancer la perte d'information. Dans nos expériences, le filtrage des co-occurrences par les groupes syntaxiques semble constituer un compromis efficace alors que la sélection des co-occurrences par les relations syntaxiques privilégie trop la réduction d'information et mène donc à une dégradation des performances.

5 Conclusion

Nous avons présenté dans cet article l'étude de deux méthodes de filtrage syntaxique pour le calcul des co-occurrences prises en compte dans une représentation vectorielle distributionnelle des documents : d'une part, une méthode prenant en compte un filtrage reposant sur les groupes syntaxiques, et d'autre part une méthode prenant en compte un mécanisme de sélection fondé sur les relations syntaxiques. Ces deux méthodes ont été testées dans le cadre de la recherche documentaire et ont montré que le filtrage du premier type permet non seulement d'éliminer un nombre important de co-occurrences dont la disparition n'altère pas les performances, mais également de filtrer des co-occurrences qui introduisent du bruit dans la représentation et dont l'élimination est donc bénéfique. Le filtrage du deuxième type semble quant à lui trop restrictif et détériore de façon significative les performances du système.

Du fait du faible gain en performance observé dans nos expériences, il apparaît que les informations sur les groupes syntaxiques permettent essentiellement de réduire le volume des données de co-occurrences à manipuler sans dégradation des résultats. Pour ce qui est des relations syntaxiques, cette information nous paraît intéressante malgré les résultats négatifs observés lors de nos expériences et d'autres méthodes d'intégration devraient être envisagées pour ces données. Une piste de recherche intéressante pourrait par exemple être de typer les co-occurrences, de façon par exemple à préserver dans la représentation la distinction entre les co-occurrences de type sujet-verbe ou verbe-objet, ou entre les co-occurrences de type tête-tête ou tête-dépendance. Une autre piste pourrait être d'étudier l'apport de représentations sémantiques du type prédicat-argument pour la sélection des co-occurrences (cela permettrait en particulier

de retrouver les co-occurrences sujet-objet à travers un prédicat verbal).

D'autre part, les résultats présentés restent des résultats quantitatifs globaux et une étude qualitative plus fine serait nécessaire pour permettre de confirmer les intuitions dégagées de cette première évaluation.

Références

BESANÇON R. (2001). *Intégration de connaissances syntaxiques et sémantiques dans les représentations vectorielles de textes*. PhD thesis, École Polytechnique Fédérale de Lausanne.

BESANÇON R., RAJMAN M. & CHAPPELIER J.-C. (1999). Textual similarities based on a distributional approach. In *Proceedings of the Tenth International Workshop on Database and Expert Systems Applications (DEXA'99)*, p. 180–184, Firenze (Italy).

CONSTANT P. (1995). *Manuel de développement SYLEX-BASE*. INGÉNIA-LN, Paris, France.

CORET A., KREMER P., LANDI B., SCHIBLER D. & SCHMITT L. (1997). Towards a methodology for evaluating information retrieval systemsadapted to textual documents in the french language : the amaryllis exploratory cycle. In *SALT Workshop on Evaluation in Speech and Language Technology*, Sheffield, UK.

LANDI B., KREMER P. & SCHMITT L. (1998). Amaryllis : an evaluation experiment on search engine in a french-speaking context. In *Proceedings of the First International Conference on Language Resources and Evaluation (LREC)*, Granada, Spain.

RAJMAN M. (1995). *Apports d'une approche à base de corpus aux techniques de traitement automatique de langage naturel*. PhD thesis, ENST, Paris.

RAJMAN M., BESANÇON R. & CHAPPELIER J.-C. (2000). Le modèle DSIR : Une approche à base de sémantique distributionnelle pour la recherche documentaire. *Traitement Automatique des Langues*, **41**(2), 549–578.

RAJMAN M. & BONNET A. (1992). Corpora-base linguistics : new tools for natural language processing. In *1st Annual Conference of the Association for Global Strategic Information*, Bad Kreuznach, Germany.

RUNGSAWANG A. (1997). *Recherche Documentaire à base de sémantique distributionnelle*. PhD thesis, ENST, Paris.

G. SALTON, Ed. (1971). *The SMART Retrieval System – Experiments in Automatic Document Processing*. Prentice Hall.

SALTON G. & BUCKLEY C. (1988). Term weighting approaches in automatic text retrieval. *Information Processing and Management*, **24**, 513–523.

SALTON G. & MCGILL M. (1983). *Introduction to Modern Information Retrieval*. McGraw Hill.

SALTON G., WONG A. & YANG C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, **18**(11), 613–620.

SINGHAL A. (1997). *Term Weighting Revisited*. PhD thesis, Department of Computer Science, Cornell University.

VAN RIJSBERGEN C. (1979). *Information Retrieval*. London : Buttherwords.

XEROX (1990). Xelda : Xerox linguistic development architecture.
http ://www.xrce.xerox.com/ats/xelda/.