TALN 2004, Fès, 19–21 avril 2004

Extension de requêtes par lien sémantique nom-verbe acquis sur corpus

Vincent Claveau, Pascale Sébillot
IRISA - Université de Rennes 1
Campus de Beaulieu
35042 Rennes Cedex, FRANCE
{Vincent.Claveau,Pascale.Sébillot}@irisa.fr

Résumé - Abstract

En recherche d'information, savoir reformuler une idée par des termes différents est une des clefs pour l'amélioration des performances des systèmes de recherche d'information (SRI) existants. L'un des moyens pour résoudre ce problème est d'utiliser des ressources sémantiques spécialisées et adaptées à la base documentaire sur laquelle les recherches sont faites. Nous proposons dans cet article de montrer que les liens sémantiques entre noms et verbes appelés liens qualia, définis dans le modèle du Lexique génératif (Pustejovsky, 1995), peuvent effectivement améliorer les résultats des SRI. Pour cela, nous extrayons automatiquement des couples nom-verbe en relation qualia de la base documentaire à l'aide du système d'acquisition ASARES (Claveau, 2003a). Ces couples sont ensuite utilisés pour étendre les requêtes d'un système de recherche. Nous montrons, à l'aide des données de la campagne d'évaluation Amaryllis, que cette extension permet effectivement d'obtenir des réponses plus pertinentes, et plus particulièrement pour les premiers documents retournés à l'utilisateur.

In the information retrieval field, managing the equivalent reformulations of a same idea is a key point to improve the performances of existing retrieval systems. One way to reach this goal is to use specialised semantic resources that are suited to the document database on which the queries are processed. In this paper, we show that the semantic links between nouns and verbs called qualia links, defined in the Generative lexicon framework (Pustejovsky, 1995), enable us to improve the results of retrieval systems. To achieve this goal, we automatically extract from the document database noun-verb pairs that are in qualia relation with the acquisition system ASARES (Claveau, 2003a). These pairs are then used to expand the queries of a retrieval system. With the help of the Amaryllis evaluation campaign data, we show that these expansions actually lead to better results, especially for the first documents proposed to the user.

Mots-clefs – Keywords

Lexique sémantique, acquisition sur corpus, recherche d'information, Lexique génératif, extension de requête
Semantic lexicon, corpus-based acquisition, information retrieval, Generative lexicon, query expansion

1 Introduction

En recherche d'information (RI), savoir reformuler une idée par des termes différents est une des clefs pour l'amélioration des performances des systèmes de recherche d'information (SRI) existants. À ce titre, le traitement automatique des langues, et plus précisément l'acquisition de ressources sémantiques sur corpus, permet dans une certaine mesure de fournir les éléments nécessaires à ces reformulations. Il paraît en effet évident que la connaissance de relations sémantiques de noms à noms (comme la synonymie ou l'hyperonymie) permet d'associer les mots *mémoire* à *disque dur* ou *message* à *lettre*. Nous proposons pour notre part de nous attacher à des liens sémantiques moins étudiés : ceux existant entre des noms (N) et verbes (V), comme *disque dur–stocker* ou *lettre–communiquer*. Plus précisément, nous nous intéressons aux liens N-V dits qualia tels que définis par le formalisme du Lexique génératif (Pustejovsky, 1995). Ces relations qualia décrivent à l'aide de prédicats essentiellement verbaux les différentes facettes sémantiques des noms (fonction, mode de création...). Par exemple, le nom *couteau* et le verbe *couper* sont en relation qualia (*couper* représente la fonction de *couteau*) ; le couple *couteau–couper* est appelé couple qualia.

Plusieurs auteurs ont déjà souligné l'intérêt des liens sémantiques inter-catégoriels N-V pour la reformulation et la recherche d'information. Une expérience de C. Fabre & C. Jacquemin (2000) vise par exemple à prendre en compte la variation verbo-nominale des termes afin d'exploiter ce type de lien entre des termes nominaux (*e.g. méthode d'obtention*) et des formulations verbales proches (*e.g. obtenues par d'autres méthodes*). G. Grefenstette (1997) souligne quant à lui l'importance des liens N-V syntagmatiques pour aider à préciser et à désambiguïser les noms contenus dans des requêtes courtes. Il montre ainsi qu'un moyen de caractériser sémantiquement un nom est d'extraire l'ensemble des verbes utilisés avec ce nom, de manière à recenser ce qu'il permet de faire ou ce qui est fait en direction de lui (*e.g. show* et *support* pour le nom *research*). P. Bouillon *et al.* (2000) proposent, pour leur part, un moyen de systématiser cette approche utilisant les liens N-V en RI et de spécifier un critère définissant les paires pertinentes : ne retenir, parmi les paires N-V possibles pour la reformulation sémantique, que celles en relation qualia. Sur le plan pratique, la validité de l'approche consistant à exploiter les relations qualia en RI a déjà été partiellement testée. Tout d'abord, C. Fabre & P. Sébillot (1999) ont exploité ces relations sémantiques au sein d'un service télématique (annuaire du Minitel). Une autre expérience a également été menée dans un service de documentation d'une banque belge (Vandenbroucke, 2000). Il a été montré, de manière qualitative, que l'ajout de couples qualia aux requêtes des documentalistes permettait à ces derniers d'accéder à certains documents auxquels ils n'auraient ni pensé, ni pu accéder avec les requêtes originales. Ces extensions verbales sont d'autant plus intéressantes que peu manipulées par les documentalistes, ceux-ci privilégiant plus naturellement les extensions nominales. Ces résultats semblent abonder dans le sens de l'utilité d'associer des verbes qualia à des requêtes au sein d'un système de recherche documentaire. Nous proposons dans cet article d'en vérifier expérimentalement l'intérêt.

Les liens sémantiques entre noms et verbes de ce type sont malheureusement absents des bases sémantiques existantes. Il est donc nécessaire de les acquérir sur un corpus représentatif de la base documentaire, ce qui nous assure par ailleurs de mettre au jour des liens pertinents et attestés. Pour ce faire, nous utilisons ASARES (Claveau, 2003a), un outil d'acquisition symbolique permettant d'inférer des patrons d'extraction morpho-syntaxiques de couples N-V qualia.

Dans la partie suivante, nous présentons quelques travaux proches intégrant des ressources sémantiques aux systèmes de recherche d'information. Nous décrivons ensuite brièvement

ASARES, l'outil nous permettant d'acquérir sur corpus les couples N-V qualia. Enfin, dans la dernière section, nous présentons l'expérience d'extension de requêtes à l'aide des couples qualia acquis et ses résultats.

2 Extension de requêtes par ressources sémantiques

Comme nous l'avons vu, l'augmentation des performances des systèmes de recherche d'information passe notamment par le traitement du phénomène d'équivalence sémantique. Pour ce faire, plusieurs auteurs proposent l'emploi de ressources sémantiques. Leurs travaux peuvent se distinguer selon qu'ils s'appuient sur des ressources externes à la collection de textes constituant la base documentaire, ou internes, *i.e.* dérivées de cette base.

2.1 Utilisation de ressources externes

La démarche la plus souvent adoptée consiste à recourir à une base de connaissances linguistiques regroupant les mots sémantiquement proches, et structurée selon des relations hyperonymiques ou synonymiques. Les index de la requête peuvent par exemple être automatiquement propagés en suivant les liens exprimés dans la base lexicale, de manière à disposer d'une description étendue de cette requête. C'est l'option choisie par exemple à FT-R&D pour la consultation du Minitel en français (Gilloux *et al.*, 1993). On connaît le coût de construction de telles ressources, qui amène généralement ceux qui adoptent cette approche à plaider pour l'utilisation de ressources générales, mutualisables, dont WORDNET (Fellbaum, 1998) constitue le modèle (Smeaton, 1999). Le gain apporté par le recours à de telles ressources n'a toutefois pas été démontré jusqu'à présent. Dans ses expériences consistant à étendre des requêtes avec des termes appartenant aux mêmes *synsets* de WORDNET que les termes de la question, E. Voorhees (1994) constate même des dégradations de performances.

Deux aspects de cette démarche expliquent essentiellement ses limites : tout d'abord, elle fait l'hypothèse d'une ressource lexicale générale valable hors contexte. Or, les limites de l'utilisation des bases généralistes sur des domaines particuliers sont connues. On ne peut en effet pas savoir dans quelle mesure un modèle sémantique conçu *a priori* s'avère adéquat pour représenter le fonctionnement de domaines particuliers. Or, étendre la requête consiste précisément à tenter de la rapprocher des documents qu'elle cherche à explorer, en d'autres termes, à l'ancrer dans les mots réellement utilisés dans le corpus. En second lieu, il manque à l'approche par thésaurus une réflexion linguistique préalable concernant le fonctionnement sémantique des descripteurs. Elle mobilise en effet exclusivement les relations lexicales traditionnelles (hyperonymie, synonymie). Cette option témoigne d'une vision très cloisonnée du lexique. Ainsi A. Smeaton (1999) déclare n'exploiter de WORDNET que les noms, ceux-ci étant les principaux détenteurs du contenu des textes. Or, s'il est prouvé que les groupes nominaux constituent le principal mode d'expression des descripteurs, l'apport sémantique d'autres catégories de mots tels que les verbes ne doit pas être négligé pour réaliser l'enrichissement et la reformulation des index.

2.2 Utilisation de ressources internes

En alternative à l'utilisation de ressources généralistes, certaines études dérivent directement de la collection de documents les connaissances sémantiques ensuite exploitées dans le processus de recherche. L'utilisation de cooccurrences a notamment fait l'objet très tôt de plusieurs

travaux (Lesk, 1969; van Rijsbergen, 1977). L'idée sous-jacente sur laquelle ils s'appuient est que tout terme étroitement lié à un terme d'indexation peut lui-même être utilisé comme terme d'indexation. En pratique, ces termes « étroitement liés » sont calculés à partir des cooccurrences fréquentes des mots, par des méthodes essentiellement numériques. Un thésaurus est ainsi construit ; lors d'une interrogation, aux termes de la requête sont alors ajoutés les éléments du thésaurus qui leur sont proches, soit en considérant chaque mot de la requête indépendamment, soit en considérant l'ensemble de la requête (Qiu & Frei, 1995). L'efficacité de ces approches d'extension de requêtes par cooccurrents est variable selon les travaux mais aucune amélioration franche des résultats ne semble se dégager quelle que soit la collection de documents. H. Peat & P. Wilett (1991) expliquent ce phénomène par le fait que les méthodes utilisées pour l'extraction des cooccurrences favorisent l'acquisition de termes approximativement de même fréquence. Or si les termes de la requête sont très fréquents, les termes ajoutés sont eux aussi trop fréquents pour être discriminants.

Notons enfin que les travaux se plaçant dans le domaine du retour de pertinence (*relevance feeback*), initiés par J. Rocchio (1971), peuvent également s'interpréter comme une utilisation de ressources sémantiques internes pour l'expansion de requêtes. En effet, le principe de ces travaux (Salton & Buckley, 1990) est d'exploiter les documents retournés en réponse à une requête pour améliorer dans un second temps le résultat de la recherche. En particulier, cela peut se faire en extrayant des documents ramenés par la requête originale de nouveaux termes utilisés à leur tour pour interroger la base. Les techniques choisies pour l'extraction de ces nouveaux termes sont variées mais peuvent se rapprocher de celles utilisées pour la construction automatique des thésaurus citées précédemment.

Notre travail se situe dans ce second type de méthodes d'utilisation de ressources sémantiques. Nous nous distinguons cependant des travaux existants par l'originalité de la ressource sémantique exploitée — la relation N-V qualia — et par la technique d'acquisition utilisée pour constituer cette ressource. En effet, comme nous le précisons en section suivante, l'extraction des couples N-V qualia sur corpus se fait par une technique symbolique, ASARES, qui devrait permettre de dépasser les limites des approches numériques évoquées par H. Peat & P. Wilett (1991).

3 Acquisition sur corpus de liens nom-verbe qualia

Pour réaliser l'acquisition sur corpus des couples N-V qualia, nous utilisons le système symbolique ASARES. Nous en présentons succinctement dans cette section le principe. Le lecteur intéressé peut se reporter à (Claveau, 2003b; Claveau, 2003a) pour une description plus détaillée.

Le système ASARES repose sur une technique symbolique à la croisée de la logique et de l'apprentissage artificiel supervisé : la Programmation Logique Inductive (PLI). La PLI, grâce à son expressivité, permet d'apprendre automatiquement des patrons contextuels (clauses de Horn représentant la structure morpho-syntaxique des phrases dans lesquelles les couples N-V apparaissent). C'est une technique d'apprentissage supervisée ; elle nécessite donc en entrée des exemples de phrases contenant des couples qualia (ainsi que des contre-exemples) pour produire ces clauses. Les patrons ainsi générés sont ensuite utilisés pour extraire de nouveaux couples qualia. Cette méthode permet donc non seulement l'acquisition de couples N-V qualia — avec des taux de rappel et précision bien meilleurs que les approches statistiques classiques — mais aussi de fournir, via les clauses obtenues, un support interprétable par des linguistes des struc-

tures portant les couples qualia (Bouillon *et al.*, 2002).

La première phase nécessaire à l'utilisation d'ASARES est l'étiquetage morpho-syntaxique du corpus sur lequel on doit effectuer la tâche d'acquisition. À chaque mot du corpus est donc assignée une unique étiquette indiquant sa catégorie et quelques informations morphologiques (verbe à l'infinitif, nom commun féminin au pluriel...). Dans les expériences présentées ci-dessous, cette phase est réalisée à l'aide de l'étiqueteur CORDIAL ANALYSEUR[1].

La phase de génération d'exemples consiste à trouver des phrases dans le corpus de référence contenant des occurrences des informations (unités ou relations) sémantiques recherchées. Dans notre cas, il s'agit de repérer des phrases contenant des couples nom-verbe en relation qualia. Cette phase peut être soit manuelle — un expert extrait alors de telles phrases du corpus — soit automatique. Dans ce dernier cas, des techniques statistiques d'extraction de collocations basées sur des calculs de scores d'association (information mutuelle, Loglike...) peuvent être utilisées. Ce sont donc les occurrences et leur contexte qui constituent la base d'exemples. Des contre-exemples peuvent également être trouvés de la même manière.

Une fois les ensembles d'exemples constitués, la phase la plus importante du système, celle de production de patrons, peut se dérouler. L'approche que nous avons adoptée est de considérer cette étape comme un problème d'apprentissage artificiel symbolique. Nous tentons donc d'inférer à partir des exemples et des contre-exemples, par PLI, un ensemble de règles qui serviront ensuite de patrons d'extraction. Ainsi à partir de l'exemple *allumage-déclencher* dans la phrase « *... déclenche : l'allumage du voyant 1, l'allumage du voyant alarme ...* » la clause de Horn suivante peut être inférée : is_qualia(N,V) :- precedes(V,N), singular_common_noun(N), suc(V,X), colon(X), pred(N,Y), punctuation(Y). ; celle-ci correspond au patron d'extraction V + : + (tout token)* + [:,;] + N au singulier. Les patrons obtenus à cette étape peuvent enfin permettre de détecter de nouveaux éléments conformes à ceux donnés en exemples. Le patron précédent permettrait ainsi d'extraire *capot-ouvrir* dans la phrase « *Ouvrir : le capot coulissant, le capot droit et ...* ». Tous les couples ainsi extraits forment une sorte de lexique de couples qualia que l'on va exploiter pour étendre les requêtes d'un système de recherche d'information.

4 Extension de requêtes par couples qualia

Comme nous l'avons vu, les liens sémantiques entre noms et verbes, et plus spécialement les liens N-V qualia, semblent particulièrement intéressants dans un contexte de recherche d'information. Nous nous proposons dans cette partie d'en vérifier expérimentalement l'utilité en se servant de couples N-V en relation qualia pour étendre des requêtes et ainsi mesurer l'impact sur la pertinence des documents retournés.

Nous présentons ci-après le protocole selon lequel se déroulent ces expérimentations. Nous décrivons ensuite plus spécifiquement la mise-en-œuvre de l'extension de requêtes à l'aide des relations qualia acquises automatiquement par ASARES. Enfin, nous examinons et discutons les résultats obtenus par cette technique d'extension sur notre collection de test.

4.1 Protocole expérimental

Le protocole expérimental que nous retenons pour nos expérimentations se veut le plus usuel possible. Le système de recherche d'information (SRI) que nous utilisons est SMART,

[1] CORDIAL ANALYSEUR est un produit de la société Synapse Développement ; http://www.synapse-fr.com.

développé par G. Salton (1971) pour mettre en œuvre ses idées concernant le modèle vectoriel. Dans ce modèle, à tout document est assigné un vecteur dont chaque composante représente l'importance (calculée à partir d'un schéma de pondération s'appuyant le plus souvent sur la fréquence) d'un mot dans le document.

Les données utilisées lors de nos expérimentations sont issues de la campagne Amaryllis (Landi *et al.*, 1998) d'évaluation des SRI. Nous travaillons donc sur un corpus d'articles du journal Le Monde, un jeu de requêtes et leurs réponses. Ainsi, ce sont 11 016 articles (chacun représentant un document et identifié par un numéro unique) qui composent notre base documentaire. Pour chacune des 26 requêtes du jeu de test, une liste des identificateurs des documents pertinents est fournie. Les requêtes sont structurées en un domaine général, un sujet, une question détaillée en langage naturel, une explication des documents attendus en réponse, et une liste de concepts (sous forme de mots-clés) proches ; une requête peut ainsi comporter jusqu'à 50 mots pleins. Les réponses aux requêtes servant de référence pour l'évaluation ont été établies par un groupe de juges humains.

Une collection de couples qualia nécessaires à l'extension de requêtes est construite à partir du corpus constitué de tous les documents de notre base. Pour cela, nous utilisons le système ASARES décrit en section précédente. L'ensemble des articles a donc été étiqueté, puis une première phase d'extraction statistique, utilisant le coefficient du Loglike, a ensuite été menée et a servi à définir deux ensembles d'exemples et contre-exemples. La phase d'apprentissage sur ces deux ensembles a permis d'inférer des patrons morpho-syntaxiques qui ont ensuite été appliqués aux articles pour en extraire les couples qualia. Les couples extraits sont rassemblés dans une base lexicale avec leur nombre de détections (nombre d'occurrences trouvées par les patrons inférés). Par exemple, les couples *drogue-dépénaliser* et *drogue-acheter* ont été détectés 3 fois, le couple *drogue-consommer* a été détecté 2 fois...

Pour vérifier l'intérêt éventuel des liens de type qualia en extension de requêtes, nous utilisons les mesures de performances couramment utilisées en RI. Plus précisément, nous calculons les taux de rappel R (nombre de bons documents retournés par rapport au nombre total de bons documents dans la base) et de précision P (nombre de bons documents retournés par rapport au nombre total de documents retournés). Puisqu'un score est assigné à chaque document en fonction de sa proximité avec la requête, les documents donnés en réponse sont donc ordonnés, mais il faut décider du nombre de documents que l'on propose à l'utilisateur. Ce nombre, appelé *document cut-off value* (DCV), fait bien sûr varier les taux de rappel et de précision ; ceux-ci sont donc évalués pour différents DCV (et notés R(DCV) et P(DCV)), ou rassemblés au sein de courbes rappel-précision. Des moyennes de la précision, soit interpolée (IAP), soit non-interpolée (NIAP), pour les différents DCV, représentant les performances globales du système, sont également calculées. Par ailleurs, pour s'assurer que les grandeurs mesurées ne sont pas le fruit du hasard, nous vérifions que les améliorations ou dégradations constatées sont statistiquement significatives. Pour ce faire, nous utilisons le test de Student pour données pairées et celui de Wilcoxon (Hull, 1996) ; les probabilités que les grandeurs mesurées soient dues au hasard sont respectivement notées p_S et p_W.

4.2 Description de la méthode d'extension des requêtes

Conformément aux conclusions de C. de Loupy & M. El-Bèze (2002), les requêtes que nous utilisons sont uniquement composées des sujets des requêtes Amaryllis[2] (voir ci-dessus). Le

[2]Nous utilisons donc un protocole expérimental tout à fait différent de celui utilisé pour les campagnes Amaryllis dans le seul but de mesurer clairement l'impact de l'utilisation des liens qualia pour étendre des requêtes. En

nombre de mots pleins utilisés est ainsi plus proche d'une utilisation du système de recherche dans des conditions réelles et ouvertes puisque toutes les études effectuées à partir des *logs* de moteurs de recherche du Web montrent que les requêtes formulées sont en moyenne inférieures à deux mots (Jensen *et al.*, 2000; Silverstein *et al.*, 1998).

Pour tester l'apport des verbes qualia à la recherche documentaire, nous proposons d'étendre chaque requête avec les verbes qualia correspondant aux noms communs présents dans cette requête. La stratégie utilisée en pratique pour réaliser cette extension est très simple :

– le nombre maximum de verbes ajoutés par nom, noté Nb_V, est arbitrairement fixé à 5 dans les expériences présentées ci-après ;

– tous les noms présents dans la requête sont candidats à l'extension ;

– pour un nom fixé, les Nb_V verbes qualia choisis dans la collection de couples sont ceux ayant le plus d'occurrences détectées par ASARES.

Cette extension de requête se fait donc en considérant les noms de la requête de manière disjointe, mais prend en compte, à travers le choix des verbes détectés le plus souvent, une sorte de degré de certitude fourni par notre système d'acquisition. Par exemple, si la requête originale est *la drogue en France*, les verbes *revendre, acheter, consommer, prendre, dépénaliser*, associés à *drogue* et extraits du corpus par ASARES, y seront ajoutés.

La requête étendue se compose donc des termes de la requête originale et des verbes qualia. Cette requête étendue ne remplace pas l'ancienne requête ; cette dernière est en effet également utilisée lors de la recherche grâce aux mécanismes de sous-vecteurs proposés dans l'extension du modèle vectoriel de E. Fox (1983). Dans ce modèle, les requêtes sont composées de sous-vecteurs, chacun d'eux pouvant représenter un type d'information différent appelé *ctype* pour *concept type*. La similarité d'un document et d'une requête est la somme des similarités selon chaque sous-vecteur.

4.3 Évaluation des performances de l'extension de requêtes

La figure 1 présente les courbes rappel-précision du système avec les extensions de requêtes décrites précédemment comparé au même système mais utilisant uniquement les requêtes originales (sans extension). Deux jeux de courbes sont fournis : l'un décrit les performances de ces deux systèmes en fixant le nombre de documents examinés (DCV) à 20, l'autre considère tous les documents pertinents. Il en ressort que l'extension semble avoir un effet bénéfique notable sur la précision lorsque le rappel est inférieur à 30% ; au-delà de cette limite les systèmes avec extension et sans extension se comportent de manière identique. On remarque également qu'à rappel identique, l'extension de requête améliore plus nettement les performances lorsque le DCV est fixé à 20 documents.

Cette dernière remarque est par ailleurs confirmée par la figure 2 représentant la précision du système calculée sur les 5 à 5 000 premiers documents retournés. L'amélioration amenée par l'extension de requête (notée ici par $Nb_V = 5$) par rapport au même système utilisant les requêtes non étendues (noté par $Nb_V = 0$) est flagrante pour des DCV < 20. Au-delà de ce seuil, aucune modification n'est apportée par l'ajout de verbes qualia.

Ces remarques sont également confirmées par les données recueillies dans le tableau 1. Dans ce dernier, seules les mesures jugées statistiquement significatives par l'un des deux tests employés sont indiquées. Les probabilités non indiquées sont supérieures au seuil fixé de 0.1. Toutes les différences statistiquement significatives sont positives, c'est-à-dire au bénéfice de l'extension

conséquence, nous ne comparons pas par la suite nos résultats à ceux obtenus lors de cette campagne.

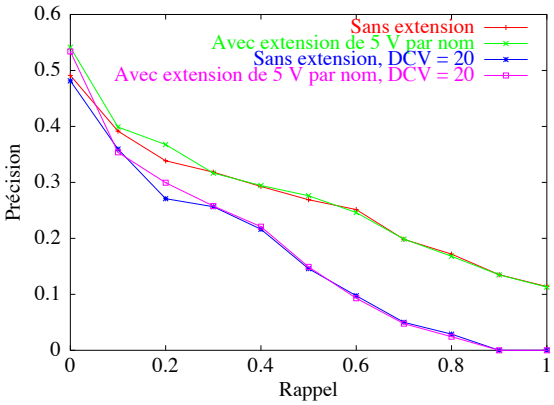

Figure 1: Courbes rappel-précision du système avec et sans extension

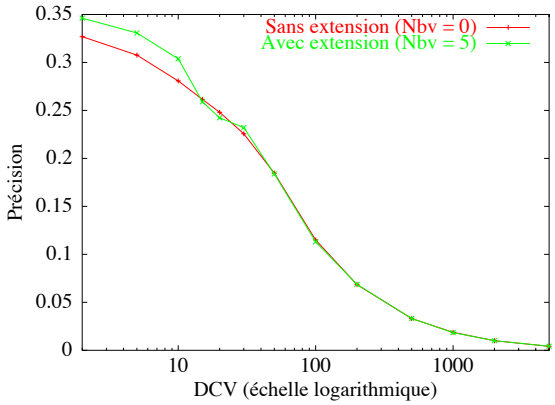

Figure 2: Précisions du système avec et sans extension selon différents DCV

	Sans extension (%)	Avec extension (%)	Amélioration (%)	Probabilité de H_0
IAP	27.01	27.77	+2.81%	$p_S < 0.05, p_W < 0.1$
NIAP	25.32	25.91	+2.33%	$p_S < 0.05$
P(5)	30.77	33.08	+7.5%	$p_S < 0.1$
P(10)	28.08	30.38	+8.22%	$p_S < 0.025, p_W < 0.06$
P(30)	22.56	23.21	+2.84%	$p_S < 0.05, p_W < 0.08$
P(5000)	0.41	0.42	+1.11%	$p_S < 0.05$
R(5)	8.14	8.61	+5.8%	$p_S < 0.1$
R(10)	15.35	16.86	+9.79%	$p_S < 0.025, p_W < 0.1$
R(30)	34.91	35.53	+1.78%	$p_S < 0.05, p_W < 0.08$
R(5000)	92.77	94.52	+1.89%	$p_S < 0.025$

Table 1: Performances de l'extension de requête

de requêtes. Le résultat principal que l'on observe est une nette amélioration des performances, à la fois en termes de rappel et de précision, pour des faibles DCV (5, 10 et 30 documents), grâce à l'ajout des verbes qualia à la requête. La précision globale du système, mesurée par la précision moyenne interpolée et non interpolée, en bénéficie également avec une augmentation légère mais significative. Les taux de rappel et de précision mesurés sur les 5 000 premiers documents sont eux aussi en très légère hausse avec l'utilisation des requêtes étendues. Des expériences en cours semblent par ailleurs montrer que la taille de l'extension influe peu sur ces différents résultats.

5 Discussion des résultats et perspectives

Les résultats précédents sont très intéressants et confirment en partie l'intérêt du lien N-V qualia en recherche d'information. Plus précisément, il semble que l'utilisation des verbes qualia permette de retrouver plus rapidement des documents qui auraient finalement été proposés à l'utilisateur, mais à des rangs prohibitifs. Ainsi, l'extension concentre en tête de liste les doc-

uments pertinents, plus qu'elle n'agit sur la précision au détriment du rappel comme cela est souvent affirmé en recherche documentaire.

D'un point de vue pratique, cette extension de requêtes par ressources sémantiques — à savoir des verbes qualia liés aux noms contenus dans la requête — permet donc d'améliorer légèrement les performances globales d'un système de recherche documentaire standard. Mais cette extension est particulièrement performante, et donc intéressante à mettre en œuvre, pour les systèmes nécessitant une bonne précision et un bon rappel dès les premiers documents retournés à l'utilisateur. Un tel processus d'extension serait par exemple particulièrement profitable à des systèmes grand public pour lesquels on sait qu'en moyenne seuls les 20 premiers documents retournés sont consultés par les utilisateurs. Ces résultats sont notamment à mettre en perspective avec ceux, plutôt négatifs, obtenus par d'autres expériences d'extension à l'aide de ressources lexicales externes (Voorhees, 1994, par exemple) ou dérivées de la base documentaire (voir (Peat & Willett, 1991) pour une discussion sur ces résultats).

L'extension à l'aide de verbes est aussi intéressante à un autre titre. Les utilisateurs de SRI, même avertis, ont tendance à spécifier naturellement une requête en y ajoutant de nouveaux noms. L'extension par verbes qualia permet donc de faire émerger des documents qui n'auraient pas nécessairement été trouvés par une extension manuelle.

Beaucoup de perspectives restent ouvertes à l'issue de ces expériences. Tout d'abord, notre mise-en-œuvre de l'extension de requête est relativement rustique. Il serait en effet intéressant de choisir les noms à étendre à l'aide de critères linguistiques (noms têtes de syntagmes par exemple) plutôt que de tous les considérer. De la même façon, le nombre de verbes ajoutés à chaque nom, ici arbitrairement fixé à 5, pourrait varier suivant le nom à étendre. Il serait d'ailleurs intéressant d'étudier les variations de performances du système de recherche suivant ce nombre. Enfin, on peut également vouloir inclure ces ressources sémantiques non plus seulement à l'interrogation du système de recherche, mais dès la phase d'indexation de la base documentaire. Cela nécessite alors de définir une représentation du document plus complexe que la simple représentation vectorielle usuellement utilisée et reste un problème ouvert.

Références

BOUILLON P., CLAVEAU V., FABRE C. & SÉBILLOT P. (2002). Acquisition of Qualia Elements from Corpora - Evaluation of a Symbolic Learning Method. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation, LREC'02*, Las Palmas de Gran Canaria, Espagne.

BOUILLON P., FABRE C., SÉBILLOT P. & JACQMIN L. (2000). Apprentissage de ressources lexicales pour l'extension de requêtes. *TAL (Traitement automatique des langues), numéro spécial Traitement automatique des langues pour la recherche d'information*, 41(2), 367–393.

CLAVEAU V. (2003a). *Acquisition automatique de lexiques sémantiques pour la recherche d'information*. Thèse de doctorat, Université de Rennes 1, France.

CLAVEAU V. (2003b). Extraction de couples nom-verbe : une technique symbolique automatique. In *Actes de la conférence Traitement Automatique des Langues Naturelles, TALN'03*, Batz-sur-Mer, France.

DE LOUPY C. & EL-BÈZE M. (2002). Managing Synonymy and Polysemy in a Document Retrieval System Using WordNet. In *Proceedings of the LREC'02 Workshop on Using Semantics for Information Retrieval and Filtering*, Las Palmas de Gran Canaria, Espagne.

FABRE C. & JACQUEMIN C. (2000). Boosting Variant Recognition with Light Semantics. In *Proceedings of the 18th International Conference on Computational Linguistics, COLING'00*, Saarbrücken, Allemagne.

FABRE C. & SÉBILLOT P. (1999). Semantic Interpretation of Binominal Sequences and Information Retrieval. In *Proceedings of the International ICSC Congress on Computational Intelligence: Methods and Applications, Symposium on Advances in Intelligent Data Analysis, AIDA'99*, Rochester, États-Unis.

C. FELLBAUM, Ed. (1998). *WordNet: An Electronic Lexical Database*. Cambridge, Massachusetts, États-Unis : The MIT Press.

FOX E. A. (1983). *Extending the Boolean and Vector Space Models of Information Retrieval with P-norm Queries and Multiple Concept Types*. PhD thesis, Cornell University, New-York, États-Unis.

GILLOUX M., LASSALLE E. & OMBROUCK J.-M. (1993). Interrogation en langage naturel du Minitel guide des services. *Écho des recherches*, 146, 1–20.

GREFENSTETTE G. (1997). SQLET: Short Query Linguistic Expansion Techniques, Palliating One-Word Queries by Providing Intermediate Structure to Text. In *Actes de la Conférence Recherche d'Informations Assistée par Ordinateur, RIAO'97*, Montréal, Québec, Canada.

HULL D. A. (1996). Stemming Algorithms: A Case Study for Detailed Evaluation. *Journal of the American Society for Information Science*, 47(1), 70–84.

JENSEN B. J., SPINK A. & SARACEVIC T. (2000). Real Life, Real Users, and Real Needs: A Study and Analysis of User Queries on the Web. *Information Processing and Management*, 36(2), 207–227.

LANDI B., KREMER P. & SCHMITT L. (1998). Amaryllis: an Evaluation Experiment on Search Engine in a French-Speaking Context. In *Proceedings of the 1st International Conference on Language and Resources Evaluation, LREC'98*, Grenade, Espagne.

LESK M. E. (1969). Word-Word Association in Document Retrieval Systems. *American documentation*, 20, 27–38.

PEAT H. J. & WILLETT P. (1991). The Limitations of Term Co-Occurrence Data for Query Expansion in Document Retrieval Systems. *Journal of the American Society for Information Science*, 42(5), 378–383.

PUSTEJOVSKY J. (1995). *The Generative Lexicon*. Cambridge, Massachusetts, États-Unis : The MIT Press.

QIU Y. & FREI H.-P. (1995). *Improving the Retrieval Effectiveness by a Similarity Thesaurus*. Rapport interne 225, ETH Zurich, Department of Computer Science, Zurich, Suisse.

ROCCHIO J. J. (1971). Relevance Feedback in Information Retrieval. In G. SALTON, Ed., *The* SMART *Retrieval System: Experiments in Automatic Document Processing*, p. 313–323. Prentice-Hall, Englewood Cliffs.

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

SALTON G. & BUCKLEY C. (1990). Improving Retrieval by Relevance Feedback. *Journal of the American Society fo Information Science*, 41(4), 288–297.

SILVERSTEIN C., HENZINGER M., MARAIS H. & MORICZ M. (1998). *Analysis of a Very Large AltaVista Query Log*. Rapport interne 1998-014, Systems Research Center, Digital Equipment Corp.

SMEATON A. F. (1999). Using NLP or NLP Resources for Information Retrieval Tasks. In T. STRZALKOWSKI, Ed., *Natural Language Information Retrieval*, p. 99–111. Kluwer Academic Publishers.

VAN RIJSBERGEN C. J. (1977). A Theoretical Basis for the Use of Co-occurrence Data in Information Retrieval. *Journal of Documentation*, 33(2), 106–119.

VANDENBROUCKE L. (2000). Indexation automatique par couples nom-verbe pertinents. Mémoire de DES en information et documentation, Université Libre de Bruxelles, Belgique.

VOORHEES E. M. (1994). Query Expansion Using Lexical-Semantic Relations. In *Proceedings of ACM SIGIR'94*, Dublin, Irlande.