TALN 2003, Batz-sur-Mer, 11–14 juin 2003

Extraction de couples nom-verbe sémantiquement liés : une technique symbolique automatique

Vincent Claveau
IRISA - Université de Rennes 1
Campus de Beaulieu
35042 Rennes Cedex, France
Vincent.Claveau@irisa.fr

Mots-clefs – Keywords

Acquisition de lexique, extraction de patrons morpho-syntaxiques et sémantiques, lexique génératif, programmation logique inductive, *bootstrapping*, apprentissage semi-supervisé
Lexicon acquisition, morpho-syntactic and semantic pattern extraction, Generative Lexicon, inductive logic programming, bootstrapping, semi-supervised learning

Résumé - Abstract

Dans le modèle du Lexique génératif (Pustejovsky, 1995), certaines propriétés sémantiques des noms sont exprimées à l'aide de verbes. Les couples nom-verbe ainsi formés présentent un intérêt applicatif notamment en recherche d'information. Leur acquisition sur corpus constitue donc un enjeu, mais la découverte des patrons qui les définissent en contexte est également importante pour la compréhension même du modèle du Lexique génératif. Cet article présente une technique entièrement automatique permettant de répondre à ce double besoin d'extraction sur corpus de couples et de patrons morpho-syntaxiques et sémantiques. Elle combine pour ce faire deux approches d'acquisition — l'approche statistique et l'approche symbolique — en conservant les avantages propres à chacune d'entre elles : robustesse et automatisation des méthodes statistiques, qualité et expressivité des résultats des techniques symboliques.

In the Generative Lexicon framework (Pustejovsky, 1995), some semantic properties of common nouns are expressed with the help of verbs. These noun-verb pairs are relevant in various domains, especially in Information Retrieval. Their corpus-based acquisition is thus an interesting issue; moreover discovering the contextual patterns in which these pairs can occur is also important in order to understand the Generative Lexicon model. This paper presents a fully automated technique that allows us to acquire from a corpus both noun-verb pairs, and semantic and morpho-syntactic patterns. This technique combines two acquisition approaches—the statistical one and the symbolic one—and keeps advantages of each approach: robustness and automation of statistical methods, quality of the results and expressiveness of symbolic ones.

1 Introduction

Le Lexique génératif est un modèle de lexique proposé par J. Pustejovsky (Pustejovsky, 1995) dans lequel les entrées lexicales sont composées de quatre structures contenant des informations typées permettant différentes stratégies d'interprétation d'un mot en fonction de son contexte. L'une de ces structures, appelée structure des qualia, indique les différentes facettes sémantiques d'un mot au travers de prédicats essentiellement verbaux (les rôles qualia). La structure des qualia du nom *livre* contient par exemple les verbes *lire*, *écrire* et *contenir (de l'information)* indiquant respectivement la fonction ou but du nom, son mode de création et la relation qu'il entretient avec les classes sémantiques dont il hérite. Chaque entrée du Lexique génératif, et plus particulièrement les noms (N), est donc liée via des relations lexicales à des verbes (V). Ce sont ces couples N-V dont V appartient à la structure des qualia de N (nommés couples qualia par la suite) que nous nous proposons d'acquérir automatiquement sur corpus.

Un des principaux intérêts de la structure des qualia réside dans la possibilité qu'elle donne d'interpréter en contexte des termes composés. L'exploitation du lien nomino-verbal autorise également l'accès à des variantes de termes qui se révèlent très utiles dans le cadre de la recherche d'information (Grefenstette, 1997). Elle permet par exemple des reformulations intercatégorielles de requêtes du type *magasin de disques* ⇒ *vendre des disques* en s'appuyant sur le couple qualia *magasin-vendre*. Bien que leur intérêt ait été montré pour ce genre de tâches (Fabre & Sébillot, 1999), le manque de ressources lexicales de ce type en empêche l'utilisation à grande échelle.

Le formalisme proposé par Pustejovsky, développé uniquement dans un cadre théorique, ne suggère aucune méthode de construction des entrées lexicales. L'acquisition sur corpus d'éléments du Lexique génératif tels que les couples qualia est donc en soi un enjeu intéressant, mais, parallèlement à cette tâche d'extraction, notre objectif est également de fournir une aide à la verbalisation du concept de rôle qualia par l'obtention de règles d'extraction linguistiquement pertinentes. Par ailleurs, les différents rôles qualia d'un nom sont propres à chaque domaine ; il paraît donc essentiel de développer une méthode d'acquisition sur corpus aisément portable d'un texte à un autre, critère d'évaluation des méthodes d'extraction au même titre que la qualité des résultats qu'elles produisent.

Peu de travaux ont été entrepris pour la construction de structures des qualia. On peut néanmoins citer (Pustejovsky *et al.*, 1993) qui propose d'acquérir les éléments de ces structures à partir d'un texte étiqueté syntaxiquement en utilisant une extraction statistique de cooccurrences couplée à un jeu d'heuristiques sous forme de patrons syntaxiques. Malheureusement, ces travaux ne sont pas clairement évalués, et la question de la portabilité de la méthode d'un texte à un autre n'est pas abordée, notamment en ce qui concerne l'étiquetage syntaxique et les heuristiques employées. Pour notre part, des travaux antérieurs (Claveau *et al.*, 2003) nous ont permis de mettre au point et d'évaluer une technique d'extraction de couples N-V qualia basée sur une méthode d'apprentissage symbolique supervisé : la programmation logique inductive (PLI) (Muggleton & De-Raedt, 1994). La PLI sert à générer un classifieur (un ensemble de règles) de manière supervisée, à savoir à l'aide d'exemples de couples qualia en contexte et de contre-exemples (couples N-V non-qualia au sein d'une phrase). Les règles obtenues sont ensuite utilisées comme patrons d'extraction. Les résultats de cette technique sont de bonne qualité, aussi bien pour la tâche d'extraction des couples que pour la pertinence linguistique des règles générées. Cependant, ces patrons d'extraction étant *a priori* propres à chaque corpus, il est nécessaire de reconduire la phase d'apprentissage par PLI pour tout nouveau texte. Cette approche supervisée présente donc l'inconvénient intrinsèque de nécessiter la construction d'un jeu d'exemples et de contre-exemples de couples qualia propre au corpus à traiter. Cette phase de supervision, essentiellement manuelle et exigeant l'aide d'un expert, est très coûteuse.

L'approche que nous présentons ici remédie à ce problème et répond ainsi aux différentes exigences citées précédemment : bonne qualité des résultats, interprétabilité linguistique et automatisation du processus. Nous rejoignons en cela certains travaux effectués en extraction d'informations (Riloff & Jones, 1999) réduisant l'intervention humaine à l'apport de quelques données (*seed-words*) en début de processus. Notre système d'extraction se veut quant à lui entièrement automatique et sans aucun appui humain. Pour ce faire, en conjonction de notre approche symbolique supervisée, nous utilisons une technique reposant sur une approche différente de l'extraction : l'approche statistique. La collaboration de ces deux techniques permet d'obtenir un système mixte d'extraction combinant les avantages des deux approches.

Nous présentons dans la section suivante le corpus utilisé lors de nos expériences, puis nous décrivons succinctement les approches symbolique et statistique utilisées pour notre tâche d'extraction de couples qualia. La section 3 détaille le fonctionnement du système mixte combinant ces deux techniques et ses résultats. Enfin, la dernière section propose des commentaires sur cette approche mixte au regard de travaux connexes en apprentissage semi-supervisé (c'est-à-dire utilisant un nombre minimal d'exemples pour conduire l'apprentissage) et conclut en donnant des perspectives à notre travail.

2 Différentes approches de l'extraction

Après la description du corpus utilisé lors de nos expérimentations, cette section présente les deux techniques d'extraction développées sur lesquelles repose notre système mixte ; la première s'inscrit dans un cadre symbolique (le couple qualia est extrait à l'aide de patrons) et la seconde dans un cadre statistique (le couple est considéré comme une forme de cooccurrence).

2.1 Corpus et étiquetages

Le corpus utilisé lors de nos différentes expérimentations est une collection de manuels de maintenance d'hélicoptères qui nous a été fourni par MATRA-CCR Aérospatiale. Il contient environ 104 000 occurrences de mots, et sa taille avoisine 700 Koctets. L'aspect très technique de ce corpus se prête bien à notre tâche grâce à la cohérence du vocabulaire et des structures syntaxiques employés et l'utilisation fréquente de verbes indiquant la fonction de termes concrets.

Ce corpus a tout d'abord été étiqueté catégoriellement grâce aux outils développés dans le cadre du projet MULTEXT (Armstrong, 1996). La qualité de cet étiquetage morpho-syntaxique est très bonne puisque moins de 2% d'erreurs ont été détectées sur un extrait du corpus de 4 000 mots étiquetés manuellement. Un étiquetage sémantique automatique a ensuite été réalisé sur le corpus. Il s'appuie sur un jeu d'étiquettes construit à partir des classes les plus génériques de WordNet (Fellbaum, 1998) et d'étiquettes de granularité plus adaptée au corpus ; pour les noms, 33 classes sont ainsi retenues. Les étiquettes employées et le processus d'étiquetage sémantique sont décrits respectivement dans (Claveau *et al.*, 2001) et (Bouillon *et al.*, 2000). Là encore, le taux d'erreur, estimé sur un extrait de 6 000 mots étiqueté manuellement, est très faible : 85% des ambiguïtés sont résolues correctement, ce qui représente 98,82% de mots bien étiquetés.

2.2 Apprentissage supervisé de règles d'extraction de couples qualia par programmation logique inductive

Depuis quelques années, l'utilisation de méthodes d'apprentissage automatique symbolique pour des tâches relevant du traitement automatique des langues s'est développée. Parmi ces méthodes, la programmation logique inductive (PLI), grâce à son expressivité et sa souplesse d'utilisation, a été appliquée à des problèmes aussi divers que l'étiquetage morpho-syntaxique, la construction d'analyseurs syntaxiques ou encore l'interrogation de bases de données en langage naturel (voir par exemple (Cussens & Džeroski, 2000) pour un panorama de travaux dans ce domaine). Nous présentons ici succinctement l'utilisation de la PLI dans notre cadre d'extraction de couples qualia (se reporter à (Claveau *et al.*, 2003) pour une description détaillée de l'ensemble de la méthode).

La PLI vise à produire des règles générales (sous forme de clauses de Horn) expliquant un concept à partir d'exemples et de contre-exemples du concept et d'un ensemble de connaissances préalables. Les règles sont obtenues par généralisation des exemples ; les contre-exemples servent à empêcher une généralisation excessive en vérifiant que les règles produites n'en recouvrent aucun (ou très peu, un peu de *bruit* pouvant être autorisé). Dans notre cas, le concept à apprendre est la nature qualia d'un couple N-V apparaissant au sein d'une phrase. Un expert doit donc construire manuellement un ensemble d'exemples E^+ et de contre-exemples E^-, c'est-à-dire extraire du corpus des couples N-V qualia et des couples N-V non-qualia avec leur contexte (tous les mots et leurs étiquettes apparaissant avec le couple dans une phrase). Par ailleurs, un langage d'hypothèses déterminant avec précision le format des clauses à produire est défini dans l'ensemble des connaissances préalables. Cela permet d'obtenir des règles bien formées et linguistiquement pertinentes au regard de notre problématique. Dans notre cas, ce langage exploite notamment les informations catégorielles et sémantiques des mots intervenant dans les exemples (soit les composants d'un couple N-V, soit son contexte) ainsi que des informations de distances (en nombre de mots) entre N et V. Les règles obtenues, qui serviront ensuite de patrons pour extraire de nouveaux couples qualia, sont par exemple de ce type :

is_qualia(N,V) :- infinitive(V), action_verb(V), artefact(N), precedes(V,N).

Cette règle signifie qu'un couple dont le nom est N et le verbe V sera considéré qualia si N apparaît dans une phrase après V, que V est un verbe d'action à l'infinitif et N un artefact.

Comme toute technique d'apprentissage automatique, il important de contrôler le niveau de généralisation des règles obtenues. Une trop grande généralisation engendrerait une faible précision de l'extraction, et une faible généralisation (apprentissage par cœur) un faible rappel. Ce contrôle et le réglage des paramètres qui en découle — notamment le bruit autorisé — se font de manière automatique grâce à une validation croisée en 10 blocs (Kohavi, 1995) : l'ensemble des exemples et contre-exemples d'apprentissage (E^+ et E^-) est divisé aléatoirement en 10 sous-ensembles ; chaque sous-ensemble sert alternativement de jeu de test pour évaluer l'apprentissage effectué à partir des 9 autres sous-ensembles. Les résultats de chaque évaluation sont résumés dans une matrice de confusion similaire à celle de la figure 1 [1]. On peut ainsi calculer le coefficient Φ qui traduit en une seule grandeur toutes les informations de cette matrice de confusion :

$$\Phi = \frac{(TP * TN) - (FP * FN)}{\sqrt{PrP * PrN * AP * AN}}$$

1. La signification des variables est donnée par la combinaison des lettres : A signifie *actual* (réel), Pr *predicated* (prédit), T *true* (vrai), F *false* (faux), P *positive* (positif) et N *negative* (négatif).

	qualia réel	non-qualia réel	Total
prédit qualia	TP	FP	PrP
prédit non-qualia	FN	TN	PrN
Total	AP	AN	S

TAB. 1 – *Matrice de confusion*

Une moyenne de Φ est calculée à partir des 10 matrices obtenues sur les 10 sous-ensembles de test. Pour chaque valeur possible des paramètres de l'algorithme de PLI, une validation croisée est donc effectuée et un Φ moyen calculé. Les valeurs des paramètres retenues sont celles qui maximisent ce Φ moyen. L'apprentissage est alors relancé avec ces réglages et la totalité des exemples.

Comme il est souligné dans (Bouillon *et al.*, 2002), cette méthode (*PLI avec expert* par la suite) donne d'excellents résultats pour l'extraction de couples qualia mais permet surtout, à travers les règles produites, de fournir un support linguistiquement pertinent à la verbalisation du concept de rôle qualia. Cependant, le coût de cette méthode, résidant essentiellement dans la phase de constitution par un expert des exemples et contre-exemples nécessaires à l'apprentissage, rend cette méthode peu portable d'un corpus à l'autre.

2.3 Extraction statistique de couples qualia

Beaucoup de travaux ont été effectués dans le domaine de l'extraction de cooccurrences par des approches statistiques (Manning & Schütze, 1999). Notre problème d'extraction peut s'inscrire dans ce cadre, les couples N-V qualia étant vus comme un type spécial de cooccurrences. Les expériences rapportées dans (Bouillon *et al.*, 2002) présentent les résultats obtenus pour quelques indices statistiques parmi les plus communément employés pour ce type de tâche (Kulczinsky, Ochiai, Yule, Loglike, *Simple Matching*, Information Mutuelle, Information Mutuelle au cube, Φ^2). Parmi ceux-ci, le coefficient d'Information Mutuelle au cube (IM^3 par la suite) proposé par B. Daille (Daille, 1994), dont nous rappelons la définition ci-après, semble donner les meilleurs résultats. Étant donnée la table de contingence 2 (les cooccurrences sont calculées dans une fenêtre d'une phrase à partir des lemmes des mots), le coefficient IM^3 du couple N_i-V_j est donné par : $log_2 \frac{a^3}{(a+b)(a+c)}$.

	V_j	V_k , $k \neq j$
N_i	a	b
N_l , $l \neq i$	c	d

TAB. 2 – *Table de contingence de la paire nom-verbe* N_i-V_j

Les résultats d'extraction de couples qualia avec cette technique (Bouillon *et al.*, 2002) se révèlent moins bons que l'approche PLI avec expert (jusqu'à 58% de précision en moins pour un rappel fixé). Par ailleurs, ce type de méthodes d'extraction ne fournit aucun élément de compréhension sur les résultats produits et ne peut donc répondre directement à notre problématique. En revanche, leur entière autonomie (aucune intervention humaine n'est requise) et leur facilité d'utilisation (seul le corpus est nécessaire) sont des qualités que l'on tente de conserver dans l'approche que nous proposons ci-après.

3 Approche mixte

Pour remplir notre double tâche de construction d'éléments du Lexique génératif et de constitution de patrons d'extraction linguistiquement motivés, nous proposons de combiner les avantages des deux méthodes précédentes en un système d'extraction mixte. Ce système conserve notamment l'aspect non-supervisé et donc entièrement automatique du cadre statistique tout en gardant l'aspect explicatif du cadre symbolique grâce à la production de règles pertinentes. La première partie de cette section en décrit le fonctionnement sous la forme d'un algorithme. Une évaluation des performances obtenues par ce système et une comparaison aux autres approches de l'extraction sont réalisées en seconde partie.

3.1 Description du système d'extraction mixte

Le système mixte que nous proposons repose sur une mise en séquence (détaillée dans l'algorithme 1) des approches statistique et symbolique : chaque méthode utilise en entrée les informations données en sortie de l'autre méthode. Plus précisément, chacune des méthodes exploite la liste des couples extraits par l'autre pour construire sa propre liste de couples (L_{PLI} pour la PLI, L_{IM^3} pour l'extraction statistique). La seule contrainte est de débuter cette itération avec la méthode statistique puisqu'elle ne nécessite aucune donnée autre que le corpus. À l'initialisation de l'algorithme, tous les couples N-V présents dans une phrase du corpus sont considérés comme potentiellement qualia ; cela est indiqué par la règle *is_qualia(A,B).* fournie dans la liste des patrons d'extraction applicables au corpus L_R.

Algorithme 1 Système mixte

Initialisation

- L_R = {is_qualia(A,B).}
- application des règles de L_R au corpus ; les couples N-V extraits et leur nombre d'occurrences détectées sont insérés dans L_{PLI}

Itération

1. pour tout couple $N_i - V_j$ de L_{PLI}
   - construction de la table de contingence de $N_i - V_j$ avec les nombres d'occurrences indiqués dans L_{PLI}
   - calcul du score de $N_i - V_j$ selon IM^3
   - insertion, suivant son score, du couple dans la liste triée décroissante L_{IM^3}

   fin pour tout
2. constitution de l'ensemble E^+ (respectivement E^-) à partir de toutes les occurrences dans le corpus des n_1 (resp. n_2) premiers (resp. derniers) couples de L_{IM^3}
3. apprentissage par PLI avec E^+ et E^- ; les règles obtenues sont regroupées dans L_R
4. application des règles de L_R au corpus ; les couples N-V extraits et leur nombre d'occurrences détectées sont réunis dans L_{PLI}

L'arrêt de l'itération est conditionné par l'obtention des mêmes règles lors de deux tours successifs. Lors de nos expérimentations, n_1 a été choisi (à chaque itération) tel que les n_1 premiers couples de L_{IM^3} soient les couples dont le score d'association est supérieur à 0, et n_2 tel que $n_2 = n_1$.

3.2 Évaluation et comparaison des performances

L'évaluation du système d'extraction obtenu est réalisée dans des conditions réelles d'utilisation grâce à un jeu de test établi par quatre experts du Lexique génératif. Le corpus de test est un extrait de 32 000 mots du corpus MATRA-CCR. En dépit de sa taille relativement petite, examiner toutes les paires N-V de ce sous-corpus est impossible ; nous nous sommes donc concentré sur sept noms particulièrement représentatifs du domaine : *vis, écrou, porte, voyant, prise, capot, bouchon*. Pour ne pas fausser les mesures, aucun de ces sept noms n'a servi lors de la phase de construction du système. Un programme Perl recherche dans le sous-corpus tous les couples N-V apparaissant au sein d'une phrase dont N est l'un des sept noms cités. Les experts annotent alors manuellement chacun de ces couples comme qualia ou non-qualia, conformément aux définitions de (Busa *et al.*, 2001). Les divergences entre les experts sont discutées jusqu'à ce qu'un agrément unanime se dégage. Finalement, parmi les 286 paires ainsi examinées, 66 sont considérées qualia. Ce jeu de test nous permet d'évaluer les systèmes d'extraction en comparant les couples N-V extraits à ceux des experts.

Afin de confronter ce jeu de test aux résultats des systèmes d'extraction, et puisque les systèmes basés sur une mesure statistique (telle IM^3) donnent un score à chaque couple N-V, il faut choisir un score-seuil à partir duquel un couple extrait sera considéré comme qualia. Il en est de même pour une méthode symbolique comme la PLI (et donc pour notre système mixte) puisque l'on peut décider de ne considérer une paire comme qualia que lorsque le nombre de ses occurrences détectées par les règles apprises dépasse un certain seuil (noté s par la suite). Ainsi, les taux de rappel et de précision de ces systèmes, calculés grâce à notre jeu de test, s'expriment en fonction de s (mêmes notations que précédemment) par :

$$R(s) = \frac{TP(s)}{TP(s) + FN(s)},\ P(s) = \frac{TP(s)}{TP(s) + FP(s)}.$$

Pour l'approche statistique, symbolique ou mixte, un seuil bas favorise le rappel au détriment de la précision et un seuil élevé produit l'effet l'inverse. Pour représenter les performances de tels systèmes en fonction des différentes valeurs de s possibles (on note $\mathcal{S}$ cet ensemble), on utilise usuellement les courbes rappel-précision dans lesquelles chaque point représente la précision du système étant donné son rappel pour un seuil s donné. Ces courbes, données en figure 1 pour les trois systèmes, semblent montrer une grande similitude de comportement entre les systèmes PLI avec expert et mixte. On constate en effet que les précisions de ces deux systèmes sont presque identiques quel que soit le taux de rappel fixé. Ces deux systèmes sont par ailleurs toujours meilleurs que le système basé uniquement sur IM^3, notamment lorsque le rappel est élevé.

Pour faciliter les comparaisons entre systèmes, on cherche donc parfois à leur assigner une mesure unique. Nous utilisons dans ce but la mesure Φ définie précédemment, qui synthétise en une seule grandeur les caractéristiques d'un système pour un seuil donné. Nous calculons également la F-mesure, fréquemment utilisée dans le domaine de la recherche d'information. Elle est définie comme étant la moyenne harmonique pondérée du taux de rappel ($R(s)$) et du taux de précision ($P(s)$). Dans le cas où un poids égal est donné au rappel et à la précision, la F-mesure s'écrit : $F(s) = \frac{2P(s)R(s)}{P(s)+R(s)}$. Le tableau 3 compare les taux de rappel, de précision, la F-mesure et le coefficient Φ de notre système mixte avec ceux des systèmes PLI avec expert

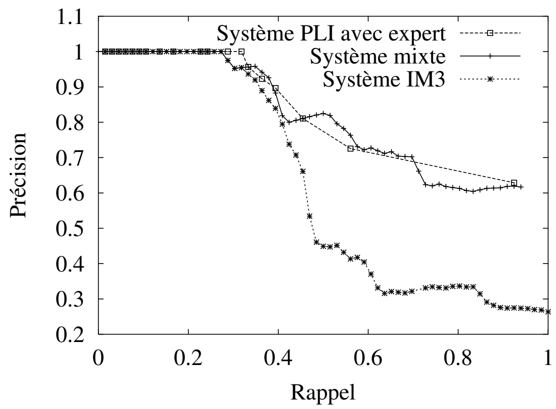

FIG. 1 – *Courbes rappel-précision des systèmes IM^3, PLI avec expert et mixte*

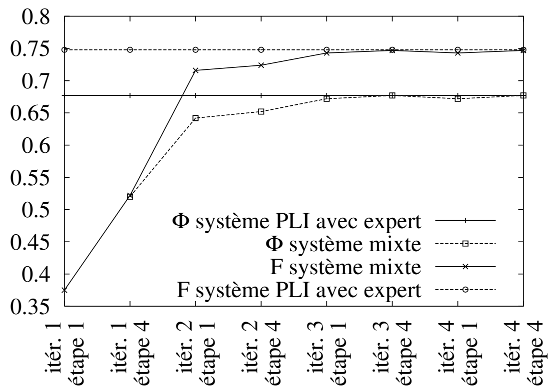

FIG. 2 – *Évolution des performances du système mixte au cours de l'algorithme*

et IM^3. Le seuil s_{opt} retenu pour chacun des systèmes est choisi tel que $\Phi(s_{opt})$ soit maximal. ($s_{opt} = \underset{s \in \mathcal{S}}{\text{argmax}}(\Phi(s))$).

	rappel (%)	précision (%)	F-mesure	coeff. Φ
PLI avec expert	92.4	62.9	0.748	0.677
IM^3	36.4	92.3	0.522	0.520
mixte	93.9	62.0	0.747	0.677

TAB. 3 – *Performances des méthodes PLI avec expert, IM^3 et mixte*

La figure 2 représente l'évolution des performances du système mixte au cours de sa construction, c'est-à-dire lors du déroulement de l'algorithme 1 sur le jeu de test décrit précédemment. Plus précisément, après chaque étape 1 et 4 de l'algorithme, on recherche comme précédemment le seuil s_{opt} maximisant Φ à partir des listes de couples qualia L_{IM^3} et L_{PLI}, et on calcule également la F-mesure pour ce même seuil. On constate qu'il suffit de trois itérations pour approcher les performances du système d'extraction PLI avec expert ; la convergence empirique de l'algorithme est donc très rapide.

Les règles obtenues par le système mixte sont analogues à celles issues du système PLI avec expert (se reporter à (Bouillon *et al.*, 2002)), ce qui explique les similitudes de performances des deux systèmes. On obtient donc des patrons exploitant essentiellement les informations morpho-syntaxiques du N ou du V (verbe à l'infinitif par exemple) et des informations de proximité (le nom doit être séparé du verbe par au plus un mot par exemple). En revanche, peu d'informations sémantiques sont utilisées, sauf en ce qui concerne les verbes (les verbes d'action sont privilégiés). Ces règles font par ailleurs émerger des indices de surface généralement délaissés par les linguistes tels que les ponctuations, particulièrement présentes et structurantes dans notre corpus.

4 Discussion et perspectives

Beaucoup de travaux cherchent à améliorer les coûts des algorithmes d'apprentissage supervisé, non pas en travaillant directement sur ces algorithmes mais en les utilisant, ainsi que les classifieurs qu'ils produisent, de manière particulière. Certains d'entre eux, rejoignant en cela notre problématique, visent à utiliser le moins possible d'exemples annotés ; les techniques d'apprentissage sont alors dites semi-supervisées. La plupart de ces techniques reposent sur des variantes de *bootstrapping* (Jones *et al.*, 1999) : un petit nombre d'exemples annotés est utilisé pour produire une première version du classifieur ; cette dernière sert alors à annoter des exemples supplémentaires qui aident à générer une deuxième version du classifieur, *etc.*

Des versions évoluées de *bootstrapping* telles que le *co-training* (Blum & Mitchell, 1998) ou celle proposée dans (Yarowsky, 1995) assurent des propriétés théoriques intéressantes, mais au prix de conditions fortes sur les données. Le *co-training* impose par exemple que les données d'apprentissage puissent être représentées selon deux *vues* conditionnellement indépendantes, deux algorithmes d'apprentissage travaillant ensuite chacun sur une vue des données [2]. Cette très forte condition d'indépendance est en fait rarement atteinte dans les données réelles (Abney, 2002) et empêche l'exploitation des résultats théoriques de ces algorithmes semi-supervisés bien que leur efficacité soit empiriquement avérée. C'est néanmoins une condition analogue qui sous-tend intuitivement notre système mixte. Il faut en effet éviter que la phase d'apprentissage par PLI ne soit biaisée et ne produise des patrons identiques à ceux ayant directement servis à extraire les exemples. Pour permettre l'introduction de nouveaux schémas contextuels dans le processus, les couples extraits par les patrons appris par PLI sont donc filtrés selon un critère indépendant de ces patrons (la mesure IM^3) et toutes les occurrences de ces couples dans le corpus servent ensuite d'exemples. Cette indépendance entre les informations utilisées par les approches symbolique (contexte sémantique et morpho-syntaxique) et statistique (occurrences) est mise à mal en pratique car notre corpus comporte de nombreuses instructions répétées à l'identique. Cependant, la ressemblance des patrons produits finalement par notre système mixte et le système PLI avec expert semble montrer une bonne tolérance de l'algorithme à ce propos.

La méthode d'acquisition de couples qualia et de patrons d'extraction présentée dans cet article répond à nos exigences de qualité des couples extraits, d'interprétabilité linguistique et de portabilité. Elle combine en cela les avantages des deux approches sur lesquelles elle repose : l'approche statistique permettant l'automatisation du processus et l'approche symbolique garantissant une bonne qualité des résultats et des règles d'extraction produites. Cette méthode d'extraction mixte n'échappe cependant pas aux coûts inhérents aux différents étiquetages du corpus, notamment l'étiquetage sémantique qui nécessite un important travail en amont sur les données. Comme il est montré dans (Claveau *et al.*, 2001) pour le système PLI avec expert, ce coût est néanmoins contrôlable (en n'effectuant qu'un étiquetage sémantique partiel par exemple) en fonction de la qualité attendue des résultats.

De nombreuses perspectives sont ouvertes sur différents aspects de ce travail. Tout d'abord, d'un point de vue applicatif, notre méthode mixte peut être aisément adaptée à l'extraction d'autres types de données que les couples N-V qualia (collocations ou éléments plus complexes tels que les fonctions lexicales (Kahane & Polguère, 2001)), notamment grâce à la souplesse d'utilisation de la PLI. Sur le plan technique, la collaboration des techniques statistiques et symboliques peut être réalisée différemment. Un couplage plus fin des deux approches peut notamment être envisagé en intégrant au sein même de la phase d'apprentissage par PLI les résultats statistiques (en pondérant les exemples nécessaires à la PLI en fonction de leur score statistique par exemple).

2. Dans le domaine du repérage d'entités nommées, ces deux vues des données peuvent être, par exemple, l'ensemble des mots composant l'entité nommée et l'ensemble des mots composant son contexte.

Références

ABNEY S. (2002). Bootstrapping. In *40th Annual Meeting of the Association for Computational Linguistics, ACL*, Philadelphia (PA), USA.

ARMSTRONG S. (1996). Multext: Multilingual Text Tools and Corpora. In H. FELDWEG & W. HINRICHS, Eds., *Lexikon und Text*.

BLUM A. & MITCHELL T. (1998). Combining Labeled and Unlabeled Data with Co-training. In *COLT: Proceedings of the Workshop on Computational Learning Theory*, Madison (WI), USA.

BOUILLON P., BAUD R. H., ROBERT G. & RUCH P. (2000). Indexing by Statistical Tagging. In *Journées d'Analyse statistique des Données Textuelles, JADT2000*, Lausanne, Switzerland.

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

BUSA F., CALZOLARI N. & LENCI A. (2001). Generative Lexicon and the SIMPLE Model: Developing Semantic Resources for NLP. In F. BUSA & P. BOUILLON, Eds., *Generativity in the Lexicon*.

CLAVEAU V., SÉBILLOT P., BOUILLON P. & FABRE C. (2001). Acquérir des éléments du lexique génératif : quels résultats et à quels coûts ? *TAL (traitement automatique des langues), numéro spécial Lexiques sémantiques*, **42**(3).

CLAVEAU V., SÉBILLOT P., FABRE C. & BOUILLON P. (2003). Learning Semantic Lexicons from a Part-of-Speech and Semantically Tagged Corpus using Inductive Logic Programming. *Journal of Machine Learning Research, special issue on ILP*. À paraître.

J. CUSSENS & S. DŽEROSKI, Eds. (2000). *Learning Language in Logic*. LNAI. Springer Verlag.

DAILLE B. (1994). *Approche mixte pour l'extraction de terminologie : statistique lexicale et filtres linguistiques*. PhD thesis, Université Paris 7.

FABRE C. & SÉBILLOT P. (1999). Semantic Interpretation of Binominal Sequences and Information Retrieval. In *International ICSC Congress on Computational Intelligence: Methods and Applications, CIMA, Symposium on Advances in Intelligent Data Analysis AIDA*, Rochester (NY), USA.

C. FELLBAUM, Ed. (1998). *WordNet: An Electronic Lexical Database*. The MIT Press.

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

JONES R., MCCALLUM A., NIGAM K. & RILOFF E. (1999). Bootstrapping for Text Learning Tasks. In *IJCAI-99 Workshop on Text Mining: Foundations, Techniques and Applications*, Stockholm, Sweden.

KAHANE S. & POLGUÈRE A. (2001). Formal Foundation of Lexical Functions. In *Workshop on Collocation: Computational Extraction, Analysis and Exploitation, 39th Annual Meeting and 10th Conference of the European Chapter of the Association for Computational Linguistics*, Toulouse, France.

KOHAVI R. (1995). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In *14th International Joint Conference on Artificial Intelligence, IJCAI 95*, Montréal, Canada.

MANNING C. D. & SCHÜTZE H. (1999). *Foundations of Statistical Natural Language Processing*. The MIT Press.

MUGGLETON S. & DE-RAEDT L. (1994). Inductive Logic Programming: Theory and Methods. *Journal of Logic Programming*, **19-20**.

PUSTEJOVSKY J. (1995). *The Generative Lexicon*. The MIT Press.

PUSTEJOVSKY J., ANICK P. & BERGLER S. (1993). Lexical Semantic Techniques for Corpus Analysis. *Computational Linguistics, special issue on Using Large Corpora*, **19**(2).

RILOFF E. & JONES R. (1999). Learning Dictionaries for Information Extraction by Multi-level Bootstrapping. In *Sixteenth National Conference on Artificial Intelligence, AAAI*, Orlando (FL), USA.

YAROWSKY D. (1995). Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. In *33rd Annual Meeting of the Association for Computational Linguistics, ACL*, Cambridge (MA), USA.