Apprentissage symbolique pour l'acquisition de ressources linguistiques

Vincent Claveau, Pascale Sébillot

IRISA - Université de Rennes 1
Campus de Beaulieu, 35042 Rennes Cedex
{vincent.claveau,pascale.sebillot}@irisa.fr

Résumé : Cet article présente une approche originale pour l'extraction sur corpus d'un type de ressources linguistiques issues du Lexique génératif (Pustejovsky, 1995), les couples qualia, qui présentent un grand intérêt pour la recherche d'information. Cette approche s'appuie sur la programmation logique inductive (PLI) pour inférer des patrons d'extraction à partir d'exemples de couples qualia en contexte. Cette technique d'acquisition donne des résultats très supérieurs aux techniques statistiques habituellement employées pour ce type de tâche. Elle permet aussi, grâce à l'aspect symbolique de la PLI, de préciser le concept même de rôle qualia. Nous proposons également deux modifications de cette technique afin de supprimer la phase manuelle de construction des exemples. Les deux systèmes résultants sont ainsi entièrement automatiques et présentent des résultats similaires à la version originale.

Mots-clé : Apprentissage automatique symbolique – Programmation logique inductive – Lexique génératif – Apprentissage semi-supervisé

1 Introduction

Le Lexique génératif est un modèle de lexique développé par J. Pustejovsky (Pustejovsky, 1995) dans lequel les entrées lexicales sont composées de quatre structures. L'une de ces structures, dite des qualia, donne accès à différentes propriétés sémantiques du mot à travers quatre formules prédicatives typées, essentiellement verbales : les rôles qualia (télique, agentif, constitutif et formel). La structure des qualia du nom *livre* contient par exemple les prédicats *lire* (le rôle télique, indiquant le but ou la fonction) et *écrire* (le rôle agentif, indiquant le mode de création). Chaque entrée lexicale, et plus spécialement celles des noms (N), est ainsi liée à des verbes (V) indispensables à leur interprétation syntaxique et sémantique en contexte ; ces paires N-V sont appelées couples qualia par la suite.

Outre l'intérêt des paires N-V qualia pour l'interprétation des termes composés, l'exploitation du lien nom-verbe donne également accès à certaines variantes de termes très utiles dans le domaine de la recherche d'information (Grefenstette, 1997). Il permet par exemple des reformulations de requêtes du type *magasin de disque* $\Rightarrow$ *vendre des disques* grâce au couple qualia *magasin-vendre*. Cependant, bien que leur intérêt ait été montré pour ce type de tâche (Fabre & Sébillot, 1999; Pustejovsky *et al.*, 1997), l'absence de telles ressources lexicales en empêche l'utilisation à grande échelle. De plus, les couples N-V qualia sont susceptibles de varier d'un domaine à un autre.

Certains travaux (Pustejovsky *et al.*, 1993) proposent d'acquérir ces couples, à partir de textes préalablement étiquetés syntaxiquement, à l'aide d'un jeu d'heuristiques syntaxiques. Si cette approche peut donner des résultats ayant une bonne précision, elle ne garantit pas un rappel très important et nécessite de redéfinir le jeu d'heuristiques à chaque nouveau corpus. De plus, les résultats de la méthode proposée ne sont pas clairement évalués – et ne peuvent donc être comparés à d'autres – et la question de sa portabilité (concernant notamment l'étiquetage syntaxique et la constitution du jeu d'heuristiques) n'est pas abordée. L'approche que nous présentons ici pour acquérir des couples qualia sur corpus va plus loin puisqu'elle se veut entièrement automatique et sans aucun *a priori* sur les structures verbales recherchées. Pour ce faire, notre méthode repose sur une technique d'apprentissage automatique symbolique supervisée, la programmation logique inductive (PLI), qui permet d'inférer des patrons d'extraction à partir d'exemples de couples qualia en contexte. L'expressivité de ces patrons morpho-syntaxiques et sémantiques répond également à un besoin linguistique en

fournissant un support interprétable permettant la définition du concept même de rôle qualia. Les performances du système que nous proposons sont très bonnes, tant pour la tâche d'extraction que pour la pertinence linguistique des patrons générés (Bouillon *et al.*, 2002) .

La portabilité réelle de cette méthode d'acquisition étant néanmoins compromise en pratique par l'aspect supervisé de notre méthode d'apprentissage automatique, nous proposons également deux variantes de cette technique combinant cette approche symbolique avec une approche d'extraction statistique plus classique. Les systèmes d'extraction résultants sont entièrement automatiques et ne nécessitent pas de fournir manuellement des exemples de couples qualia à l'algorithme de PLI. Ces deux variantes non-supervisées obtiennent des résultats similaires à la technique originale.

Après une brève description du corpus utilisé lors de nos expérimentations, nous présentons dans la section suivante l'utilisation de la PLI comme technique d'extraction de couples qualia. La section 3 propose quant à elle les deux variantes de cette technique combinant extraction symbolique et statistique. Les performances de ces différents systèmes d'extraction sont examinées et comparées en section 4. Enfin nous concluons en revenant sur les qualités et les inconvénients des systèmes proposés et présentons quelques perspectives à ces travaux en dernière partie.

2 Acquisition symbolique de couples qualia

Cette section décrit dans un premier temps le corpus utilisé et les différents étiquetages qu'il a subis avant son utilisation par les trois systèmes d'acquisition détaillés ci-après. Nous présentons ensuite la technique d'apprentissage employée : la programmation logique inductive ; son utilisation dans notre cadre d'extraction de couples qualia est détaillée en troisième partie.

2.1 Corpus et étiquetages

Le corpus utilisé lors de nos expérimentations est un manuel de maintenance d'hélicoptères, en français, qui nous a été fourni par MATRA-CCR Aérospatiale. Il contient environ 104 000 mots, soit une taille de près de 700 Koctets. Ce corpus technique présente des caractéristiques se prêtant bien à notre tâche d'acquisition : il est très homogène (aussi bien pour ses structures syntaxiques que pour son vocabulaire) ; il contient également de nombreux termes concrets apparaissant fréquemment au sein des phrases avec des verbes indiquant leur rôle télique ou agentif.

Ce corpus a tout d'abord subi un étiquetage catégoriel automatique. À l'aide des outils développés dans le projet MULTEXT (Armstrong, 1996), il a donc été segmenté en phrases et en mots, puis lemmatisé et analysé et enfin désambiguïsé avec TATOO [1], un outil basé sur des chaînes de Markov cachées. Chaque mot a ainsi reçu une étiquette indiquant sa catégorie morpho-syntaxique, son genre, son nombre. La précision de cet étiquetage, évaluée à l'aide d'un extrait de 4 000 mots étiquetés à la main est très bonne : moins de 2% d'erreurs ont été détectées.

Suivant la méthode exposée dans (Bouillon *et al.*, 2000), un étiquetage sémantique du corpus a également été réalisé. Il est effectué sur le corpus étiqueté catégoriellement et bénéficie ainsi de la désambiguïsation des mots polyfonctionnels tels que *règle* qui peut être à la fois un verbe à l'indicatif et un nom (Wilks & Stevenson, 1996). Pour composer le jeu d'étiquettes sémantiques employé, les classes les plus génériques de WordNet (Fellbaum, 1998) ont été utilisées et adaptées à notre corpus : les classes non pertinentes (pour notre corpus) ont été supprimées, et pour les classes trop larges, une granularité plus fine a été choisie. Pour les noms, nous obtenons par exemple 33 classes, organisées hiérarchiquement comme indiqué en figure 1 (les classes non utilisées pour l'étiquetage sont en italiques, les étiquettes effectivement employées sont entre parenthèses). Une description plus détaillée du processus d'étiquetage sémantique et du jeu d'étiquettes est donnée dans (Bouillon *et al.*, 2001; Claveau *et al.*, 2001). Le taux d'erreurs détectées, mesuré à l'aide d'un extrait du corpus de 6 000 mots étiquetés à la main, est là encore très faible : 85% des ambiguïtés sont correctement résolues, soit une précision totale de 98.82%.

2.2 La programmation logique inductive

L'utilisation de techniques d'apprentissage automatique symbolique en TAL devient de plus en plus courante, offrant ainsi une alternative aux approches statistiques, plus largement employées.

1. Disponible à http://www.issco.unige.ch/staff/robert/tatoo/tatoo.html.

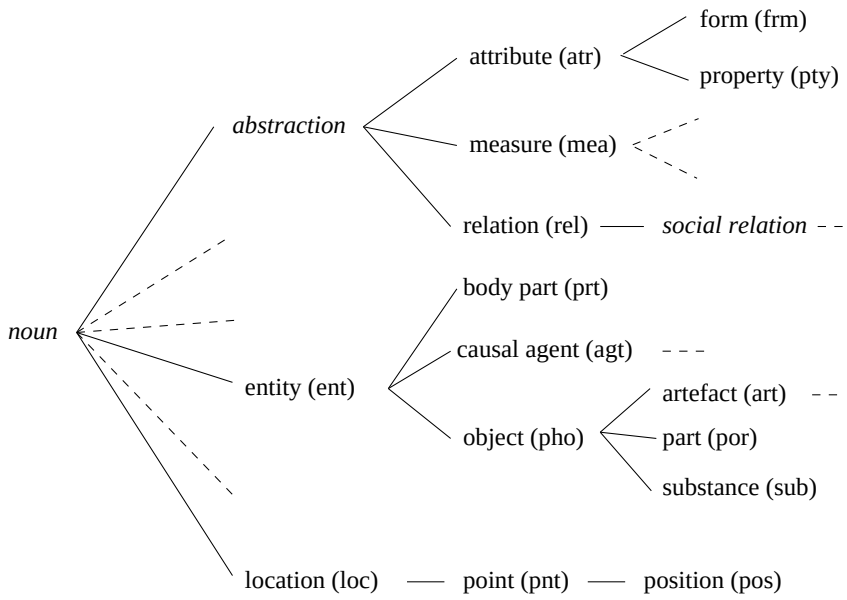

FIG. 1 – *Extrait de la hiérarchie des classes sémantiques des noms*

Parmi ces techniques, la PLI (Muggleton & De-Raedt, 1994), grâce à son expressivité et sa flexibilité, a été appliquée à de nombreux problèmes dont l'étiquetage catégoriel, la construction d'analyseurs syntaxiques, ou encore l'interrogation en langage naturel de bases de données (se reporter à (Cussens & Džeroski, 2000) pour un panorama de ce domaine).

La PLI se situe à la croisée de la programmation logique et de l'apprentissage automatique. Elle est utilisée pour produire (inférer) des règles générales (sous forme de clauses de Horn) expliquant un concept à partir d'exemples et de contre-exemples de ce concept et d'un ensemble d'informations préexistantes appelé *background knowledge*. Formellement, si l'on note E^+ l'ensemble des exemples positifs, E^- l'ensemble des exemples négatifs et $\mathcal{B}$ le *background knowledge*, la PLI cherche à induire un classifieur, c'est-à-dire un ensemble de clauses, R tel que $\mathcal{B} \wedge R \models E^+$ et $\mathcal{B} \wedge R \wedge E^- \not\models \square$ [2]. L'avantage majeur de la PLI est de permettre l'apprentissage à partir d'exemples relationnels (c'est-à-dire qu'on ne peut pas décrire par un ensemble de couples attribut-valeur) ainsi que l'apprentissage de concepts relationnels, usuellement exprimés en Prolog. C'est cette expressivité unique à la fois en entrée et en sortie du processus d'apprentissage qui rend la PLI indispensable pour traiter certains problèmes difficilement exprimables hors de ce cadre relationnel. Le logiciel de PLI utilisé lors de nos expérimentations est ALEPH, une implémentation en Prolog réalisée par Ashwin Srinivasan [3].

Dans notre cas, le concept que nous cherchons à apprendre est la nature qualia d'une paire N-V apparaissant au sein d'une phrase. On souhaite donc obtenir un classifieur permettant, d'une part, de distinguer en contexte une paire N-V qualia d'une non-qualia, mais aussi de refléter des éléments linguistiquement interprétables définissant le concept même de rôle qualia. La PLI, grâce à sa puissance explicative, se révèle un choix bien adapté pour cette tâche puisque les règles générées peuvent ensuite être directement interprétées comme des patrons linguistiques d'extraction. Par ailleurs, l'expressivité de la PLI et la possibilité d'adjoindre des informations via le *background knowledge* permettent également de décrire et d'exploiter aisément les données relationnelles telles que nos structures hiérarchiques d'étiquettes catégorielles et sémantiques. Enfin, la PLI offre la possibilité de gérer des données bruitées, ce qui se révèle indispensable pour notre tâche où des erreurs, même en taux faibles, sont inhérentes à nos processus d'étiquetage.

2. Le symbole $\models$ représente l'implication et $\square$ signifie faux.

3. Aleph est disponible l'URL http://web.comlab.ox.ac.uk/oucl/research/areas/machlearn/Aleph/ .

2.3 Apprentissage de patrons d'extraction

Comme nous l'avons expliqué précédemment, les algorithmes de PLI génèrent des règles expliquant ce qui caractérise les exemples du concept à apprendre par rapport aux contre-exemples. Dans notre cas, nous désirons discriminer les couples N-V qualia des non-qualia en fonction de leur contexte catégoriel et sémantique. Notre première tâche est donc de construire un jeu d'exemples positifs et négatifs décrivant les phrases dans lesquelles apparaissent des couples N-V qualia ou non-qualia en termes d'informations catégorielles et sémantiques.

Pour ce faire, chaque occurrence en contexte d'un sous-ensemble de couples N-V de notre corpus est manuellement examinée par un expert. Les occurrences considérées comme qualia forment ainsi l'ensemble E^+, et les non-qualia l'ensemble E^- ; leur contexte (informations catégorielles et sémantiques de tous les mots des phrases dans lesquelles elles apparaissent) est décrit dans le *background knowledge* $\mathcal{B}$. Près de 3 100 exemples positifs et 3 200 négatifs sont ainsi produits à partir du corpus MATRA-CCR.

Par ailleurs, les informations relatives à la hiérarchie des classes catégorielles et sémantiques sont transcrites dans $\mathcal{B}$. Par exemple, le fait qu'un mot ayant une étiquette tc_verb_pl soit un verbe conjugué au pluriel et puisse être considéré comme un verbe conjugué, et plus généralement comme un verbe, est précisé en Prolog par :

```
conjugated_plural_verb(W) :- tags(W, tc_verb_pl, _).
conjugated_verb(W) :- conjugated_plural_verb(W).
verb(W) :- conjugated_verb(W).
```

où tags/3 est le prédicat indiquant les étiquettes catégorielles et sémantiques d'un mot.

Un *langage d'hypothèses* est également décrit dans $\mathcal{B}$; il est utilisé pour définir précisément la forme attendue des règles générées. Ce langage nous assure ainsi de n'obtenir que des règles bien formées et linguistiquement pertinentes au regard de notre tâche. Dans notre cas, ce langage d'hypothèses exploite les informations catégorielles et sémantiques des mots apparaissant dans les exemples (c'est-à-dire N, V ou leur contexte) ainsi que les informations d'ordre et de distances entre ces mots. Une description détaillée de ce langage et des techniques utilisées pour contrôler l'expressivité et l'efficacité de notre algorithme de PLI est donnée dans (Claveau *et al.*, 2003).

En fonction de ce langage, l'objectif de l'algorithme de PLI est d'inférer des règles qui couvrent (c'est-à-dire expliquent) un maximum d'exemples et aucun contre-exemple (ou très peu, un peu de bruit pouvant être autorisé pour produire des patrons plus généraux). Plus précisément, le processus d'inférence est le suivant :

1. choisir un exemple positif $e \in E^+$ à généraliser. S'il n'y en a aucun, arrêt.
2. définir un espace de recherche d'hypothèses $\mathcal{H}$ à partir de e et du langage d'hypothèses ;
3. chercher dans $\mathcal{H}$ la règle h qui maximise une fonction de score Sc ;
4. ôter tous les exemples positifs couverts par la règle choisie. Retourner à l'étape 1.

La fonction de score Sc dépend du nombre d'exemples positifs et négatifs couverts par l'hypothèse h. Ces deux ensembles sont respectivement notés E_h^+ et E_h^- et leurs cardinaux $|E_h^+|$ et $|E_h^-|$ [4].

Finalement, les règles produites, ensuite utilisées comme patrons pour extraire de nouvelles paires qualia, sont par exemple de la forme suivante :
`is_qualia(N,V) :- infinitive(V), action_verb(V), artefact(N), pred(V,A), pred(A,N).` Cette règle signifie qu'une paire composée d'un nom N et d'un verbe V est considérée qualia si N est un artefact apparaissant dans la phrase après un mot A quelconque, lui-même suivant le verbe d'action à l'infinitif V. Une telle règle – dont la succession des pred/2 traduit l'aspect relationnel difficilement capturable par une méthode d'apprentissage autre que la PLI – permet ainsi d'extraire le couple qualia *prise-alimenter* dans la phrase *"La prise peut alimenter une pompe à carburant..."* et le couple *vis-serrer* dans *"...écrous, vis : serrer au couple..."*

Les très bonnes performances de ce système, appelé système PLI supervisé par la suite, aussi bien pour la tâche d'extraction pure que pour la génération de patrons linguistiquement pertinents, sont exposées en section 4. Cependant, le coût de cette approche, reposant en grande partie dans la phase de construction par un expert des ensembles d'exemples et contre-exemples, rend cette technique

4. Dans notre cas, cette fonction Sc dépend également d'autres paramètres comme la longueur L de l'hypothèse. Plus précisément, elle est définie par le couple $\left(|E_h^+| - |E_h^-|, L\right)$; une hypothèse h_1 de score $\left(|E_{h_1}^+| - |E_{h_1}^-|, L_1\right)$ est dite meilleure qu'une hypothèse h_2 de score $\left(|E_{h_2}^+| - |E_{h_2}^-|, L_2\right)$ si et seulement si $|E_{h_1}^+| - |E_{h_1}^-| > |E_{h_2}^+| - |E_{h_2}^-|$ ou $|E_{h_1}^+| - |E_{h_1}^-| = |E_{h_2}^+| - |E_{h_2}^-| \wedge L_1 < L_2$.

difficilement portable d'un corpus à un autre. La section 3 présente deux approches permettant de supprimer ce désavantage inhérent à l'aspect supervisé de notre méthode, tout en conservant ses performances.

3 Bootstrapping

De nombreux travaux actuels visent à réduire les coûts dus à l'étiquetage des exemples dans les méthodes d'apprentissage supervisé. La plupart de ces travaux s'appuient sur des variantes de *bootstrapping* (Jones *et al.*, 1999), et les techniques d'apprentissage résultantes sont alors dites semi-supervisées. À partir de principes similaires, nous proposons deux variantes de notre système d'extraction, travaillant sur le même corpus étiqueté catégoriellement et sémantiquement, permettant de supprimer la phase manuelle de construction des exemples de couples qualia grâce à une technique classique d'extraction statistique que nous présentons ci-dessous.

3.1 Extraction statistique

De nombreux travaux d'acquisition d'informations à partir de textes, et plus particulièrement d'extraction de cooccurrences, ont été menés via des approches statistiques (Manning & Schütze, 1999; Pearce, 2002). L'extraction de couples N-V qualia, vus comme une forme spéciale de cooccurrences, peut entrer ce cadre ; nous pouvons ainsi utiliser les nombreuses méthodes statistiques développées pour ce type de tâche. Dans (Bouillon *et al.*, 2002), nous présentons les résultats obtenus avec quelques-uns des indices statistiques les plus communs (Kulczinsky, Ochiai, Yule, Loglike, Simple Matching, Information Mutuelle, Information Mutuelle au cube, Φ^2). Parmi eux, le coefficient Information Mutuelle au cube (IM^3 par la suite) proposé par B. Daille (Daille, 1994) semble donner les meilleurs résultats. En utilisant les notations de la table de contingence 1 (les cooccurrences indiquées sont calculées à partir des lemmes des mots dans une fenêtre d'une phrase), le coefficient IM^3 se définit 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. 1 – *Table de contingence du couple N_i-V_j*

Les expériences décrites dans (Bouillon *et al.*, 2002) montrent que cette technique d'extraction statistique est toutefois bien moins performante que l'approche symbolique (cf. section 4). De plus, ce type de technique ne remplit pas notre condition d'interprétabilité des résultats puisqu'aucun indice n'explique pourquoi un couple est considéré comme qualia ou non. Cependant, cette approche statistique possède des avantages intéressants : elle est totalement automatique (aucune intervention humaine n'est requise), facile d'utilisation et donc tout à fait portable d'un corpus à un autre. Les deux parties suivantes exposent deux techniques pour transposer ces avantages vers notre système symbolique d'extraction.

3.2 Approche mixte séquentielle

La technique d'extraction mixte présentée dans cette partie repose sur une combinaison séquentielle des systèmes symbolique et statistique présentés précédemment. Comme il est indiqué dans l'algorithme 1, chaque système utilise itérativement en entrée les données de sortie de l'autre système. Plus précisément, la liste de paires N-V générée par un système (L_{PLI} pour le symbolique, L_{IM^3} pour le statistique) est utilisée par l'autre pour construire sa propre liste de couples. 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, tous les couples N-V apparaissant au sein d'une phrase sont considérés comme potentiellement qualia ; cela est indiqué grâce à la règle is_qualia(N,V). donnée dans la liste de patrons d'extraction L_R.

L'itération s'arrête lorsque le même ensemble de règles est obtenu lors de deux tours successifs. Lors de nos expériences, n_1 a été choisi (à chaque itération) tel que les n_1 premiers couples de L_{IM^3}

Algorithme 1 Système mixte séquentiel

Initialisation

- L_R = {is_qualia(N,V).}
- 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}
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}

soient tous ceux ayant un score d'association positif ; n_2 a quant à lui été choisi tel que $n_2 = n_1$. Le système d'extraction résultant est appelé par la suite système mixte séquentiel.

3.3 Approche mixte intégrée

Contrairement au système présenté ci-dessus dans lequel les techniques statistique et symbolique sont utilisées sans modifications majeures, le second système mixte que nous proposons combine ces deux approches plus étroitement et nécessite quelques changement dans l'algorithme de PLI.

Comme nous l'avons mentionné en section 2.3, lors de la troisième étape d'un apprentissage par PLI, une règle h est choisie parmi un espace d'hypothèses $\mathcal{H}$ si elle maximise une fonction de score Sc. Cette fonction dépend du nombre d'exemples positifs et négatifs que h couvre ; ainsi, on a :
$h = \underset{h \in \mathcal{H}}{\text{argmax}}\ Sc\left(|E_h^+|, |E_h^-|\right)$.

Le principe de notre seconde méthode mixte est de pondérer les exemples selon leur score statistique. Les hypothèses sont donc désormais évaluées à partir des poids des exemples (que nous définissons ci-dessous) qu'elles couvrent. Les ensembles d'exemples et contre-exemples sont donc issus des résultats d'extraction de la méthode d'extraction IM^3 : toutes les occurrences dans le corpus des couples ayant les plus hauts scores sont codées dans E^+, et inversement, celles ayant les scores les plus faibles sont placées dans E^- ; un poids w, calculé à partir des scores IM^3, est assigné à chacun de ces exemples. Ainsi, plus un exemple est considéré comme pertinent (c'est-à-dire ayant un score important) par la méthode statistique, plus il influencera le choix des hypothèses. Finalement, les règles choisies sont celles maximisant $Sc(h)$ redéfinie par :

$$h = \underset{h \in \mathcal{H}}{\text{argmax}}\, Sc\left(\sum_{e^+ \in E_h^+} w(e^+), \sum_{e^- \in E_h^-} w(e^-)\right)$$

Avec ces paramétrages et les ensembles E^+ et E^- générés automatiquement, l'algorithme de PLI modifié se déroule comme indiqué en 2.3 et produit ainsi des règles utilisées ensuite comme patrons d'extraction. Cette technique est appelée par la suite système mixte intégré.

4 Évaluation des performances

Cette section présente les performances des trois systèmes proposés en conditions réelles d'acquisition de ressources linguistiques grâce à un jeu de test que nous décrivons ci-dessous. La validité

linguistique et l'expressivité des clauses produites sont discutées en dernière partie.

4.1 Construction du jeu de test

Le jeu de test sur lequel nous évaluons les performances des systèmes d'acquisition de couples qualia est un extrait de 32 000 mots du corpus MATRA-CCR étiqueté catégoriellement et sémantiquement. Malgré sa taille relativement petite, examiner manuellement toutes les occurrences de couples N-V de ce sous-corpus pour les annoter comme qualia ou non-qualia est impossible. Nous nous sommes donc concentrés sur 7 noms particulièrement représentatifs du vocabulaire du corpus : *vis, écrou, porte, voyant, prise, capot, bouchon.* Pour ne pas fausser les mesures, aucun de ces noms n'a évidemment été utilisé lors des phases d'apprentissage.

Un programme Perl présente toutes les occurrences de couples N-V, où N est l'un des 7 noms recherchés, apparaissant au sein d'une phrase du sous-corpus, à quatre experts qui annotent alors ces paires comme qualia ou non-qualia. Les divergences sont discutées jusqu'à ce qu'un accord se dégage. Finalement, parmi les 286 couples différents trouvés, 66 d'entre-eux sont notés comme étant qualia. Ce jeu de test est ensuite utilisé pour comparer les résultats d'extraction des systèmes à ceux des experts.

4.2 Évaluation des résultats empiriques

La comparaison entre le jeu de test et les couples obtenus par les systèmes d'extraction sur le sous-corpus se fait à l'aide de matrices de confusion telles que celle donnée en table 2 [5].

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

TAB. 2 – *Matrice de confusion*

Comme tout système d'extraction basé sur une mesure statistique, le système IM^3 assigne une valeur à chaque couple N-V rencontré. Il est donc nécessaire de définir une valeur-seuil (s par la suite) à partir de laquelle un couple est considéré comme qualia ; les paires dont le score statistique est inférieur à s sont alors considérées non-qualia. De la même manière, les systèmes d'extraction symbolique (et donc les systèmes PLI supervisé et mixtes) nécessitent la définition d'un seuil puisqu'une paire N-V peut n'être considérée qualia que lorsqu'au moins s de ses occurrences sont retrouvées par les patrons d'extraction. Les valeurs indiquées dans les matrices de confusion, établies à partir de notre jeu de test, sont donc fonction de s, de même que les taux de rappel (R) et précision (P) qui sont alors définis par : $R(s) = \frac{TP(s)}{TP(s)+FN(s)}$, $P(s) = \frac{TP(s)}{TP(s)+FP(s)}$.

Les courbes rappel-précision, dans lesquelles chaque point représente la précision du système en fonction de son rappel pour un seuil s fixé, indiquent les performances de ces systèmes pour toutes les valeurs de s. La figure 2 présente ces courbes pour nos trois systèmes d'extraction symbolique ; le système IM^3 sert de référence. On remarque que les performances des systèmes symboliques sont nettement supérieures à celles du système statistique, notamment lorsque le rappel est élevé : la précision du système IM^3 est ainsi jusqu'à 58% plus faible que celle de nos trois systèmes pour un rappel fixé. Par ailleurs, les versions mixtes semblent se comporter de manière quasi identique à la version supervisée de notre approche symbolique. Nos trois systèmes permettent donc une acquisition de ressources linguistiques à la fois complètes (c'est-à-dire avec un bon rappel) et fiables (avec une bonne précision).

4.3 Évaluation linguistique

Comme nous l'avons dit précédemment, le double but de ces travaux était de définir une technique performante d'extraction de couples N-V qualia, mais aussi de fournir un support linguistique

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

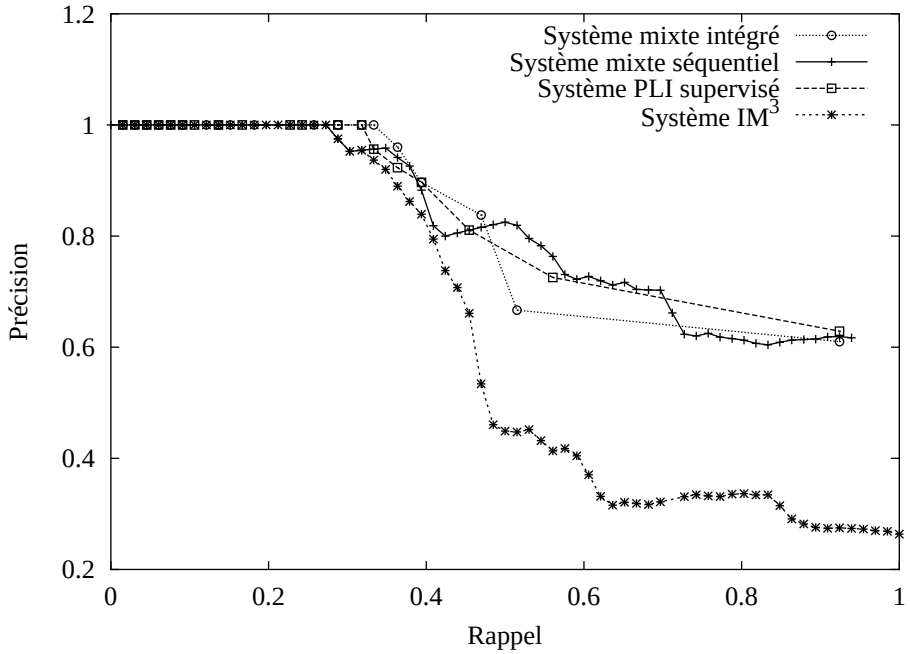

FIG. 2 – *Courbes rappel-précision des 4 systèmes*

au concept de rôle qualia. La validation des systèmes symboliques présentés ci-avant passe donc également par une évaluation de l'intérêt linguistique des règles générées.

À ce titre, on note tout d'abord de très grandes similarités entre les règles produites par nos trois systèmes, cela expliquant la similitude de leurs performances pour la tâche d'extraction. En particulier, ces règles font ressortir des schémas très généraux de proximité entre les constituants des couples (au plus un mot doit séparer le N du V) ou de position (V doit apparaître avant N dans la phrase). Par ailleurs, peu d'informations sémantiques sont utilisées à ce niveau de généralisation, à l'exception notable des verbes (les verbes d'action sont privilégiés). D'autres informations surfaciques généralement délaissées des linguistes, comme les ponctuations, particulièrement présentes et structurantes dans notre texte très technique, sont également exploitées dans des schémas plus spécifiques à notre corpus. Ainsi, la clause suivante met en évidence les structures de listes très nombreuses dans le corpus : `is_qualia(N,V):- pred(C,V), colon(C), pred(N,D), punctuation(D), singular_common_noun(N).`, et couvre par exemple les phrases de la forme *"... Verbe : ... , Nom au singulier ..."*. L'emploi de nombreux verbes à l'infinitif, typique des instructions composant le corpus, est également souligné dans la plupart des règles générées. Les clauses obtenues présentent donc, d'un point de vue linguistique, un certain nombre de caractéristiques très génériques mais aussi, comme nous nous y attendions, certaines spécificités propres à notre corpus. Une description détaillée des patrons inférés par l'approche supervisée sur le corpus MATRA CCR et des phénomènes linguistiques qu'elle permet effectivement de prendre en compte est donnée dans (Bouillon *et al.*, 2002).

5 Discussion et conclusion

Nous avons présenté une approche symbolique pour l'acquisition sur corpus de ressources linguistiques qui présente l'avantage de fournir un classifieur interprétable, contrairement aux méthodes statistiques communément employées pour ce type de tâche. L'utilisation de la PLI comme technique d'apprentissage présente le double avantage de bénéficier d'une grande expressivité à la fois pour définir notre problème et pour la génération de patrons d'extraction linguistiquement fondés. Comme nous l'avons montré avec nos deux systèmes mixtes, l'obstacle majeur à l'utilisation de cette technique d'apprentissage supervisé, à savoir la phase manuelle de construction des exemples, peut être contourné par l'emploi d'une technique d'extraction statistique servant à amorcer le système

symbolique.

Ces systèmes semi-supervisés n'échappent cependant pas aux coûts des étiquetages, notamment sémantiques, du corpus. Néanmoins, comme le suggère les expériences rapportées dans (Claveau *et al.*, 2001) et la faible utilisation des informations sémantiques dans les règles produites, ce coût peut être contrôlé (en n'effectuant qu'un étiquetage sémantique partiel, sur quelques catégories de mots, par exemple) sans modifier profondément les résultats d'extraction. Une autre limite d'utilisation de ce type de système d'acquisition réside dans la taille requise des corpus à traiter. Si, par essence, la méthode symbolique fonctionne quel que soit le volume de texte (pourvu qu'il soit homogène dans ses structures sémantiques et morpho-syntaxiques), l'utilisation d'un *bootstrap* statistique impose une taille minimale au corpus pour obtenir des résultats de cooccurrences fiables.

D'un point de vue plus théorique, les deux systèmes semi-supervisés se rapprochent de certaines versions évoluées de *bootstrapping* telles que le *co-training* (Blum & Mitchell, 1998) ou celle proposée par Yarowsky (Yarowsky, 1995), sans en partager cependant les propriétés formelles. Ces deux dernières techniques assurent en effet des résultats théoriques intéressants d'apprenabilité, mais au prix de conditions contraignantes sur les données. Par exemple, le *co-training* impose que les données d'apprentissage puissent être représentées selon deux *vues* conditionnellement indépendantes. Malheureusement, cette forte condition d'indépendance est rarement avérée dans les données réelles (Abney, 2002). Nos deux techniques semi-supervisées reposent néanmoins implicitement sur une condition analogue : pour éviter que la phase d'apprentissage par PLI ne soit biaisée, nous supposons que les différentes occurrences des couples apparaissent dans des structures sémantiques et morpho-syntaxiques variées qui donneront naissance à nos patrons d'extraction. Or, notre corpus comporte de nombreuses instructions répétées à l'identique ; cette hypothèse d'indépendance entre les couples extraits statistiquement et les patrons générés est donc en partie invalidée. Cependant, la ressemblance entre les règles générées par le système supervisé et les deux versions semi-supervisées semble montrer une bonne tolérance de nos algorithmes à ce propos.

De nombreuses perspectives sont ouvertes sur ces travaux. La variabilité des patrons produits et de la qualité des résultats d'extraction selon les domaines et les genres des textes traités doit être analysée. Pour ce faire, nous avons débuté des expériences similaires sur un corpus plus généraliste composés d'articles de journaux comptant plus de 6 millions de mots. D'un point de vue applicatif, ces systèmes peuvent être aisément adaptés à l'acquisition d'autres types d'éléments textuels (termes composés, collocations et autres informations lexicales). Enfin, sur un aspect plus théorique, l'utilisation de statistiques (vues comme une sorte de distribution de probabilités sur les ensembles d'exemples E^+ et E^-) en PLI, comme cela est fait dans le système mixte intégré, soulève d'intéressantes problématiques (Muggleton, 1994).

Références

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

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

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

BOUILLON P., CLAVEAU V., FABRE C. & SÉBILLOT P. (2001). Using Part-of-Speech and Semantic Tagging for the Corpus-Based Learning of Qualia Structure Elements. In *First International Workshop on Generative Approaches to the Lexicon, GL'2001*, Genève, Suisse.

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, Espagne.

CLAVEAU V., SÉBILLOT P., BOUILLON P. & FABRE C. (2001). Acquérir des éléments du lexique génératif : quels résultats et à quels coûts ? *Traitement automatique des langues (TAL), 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, 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.

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

MUGGLETON S. (1994). Bayesian inductive logic programming. In *7th Annual ACM Conference on Computational Learning Theory*, USA.

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

PEARCE D. (2002). A Comparative Evaluation of Collocation Extraction Techniques. In *3rd International Conference on Language Resources and Evaluation, LREC 02*, Las Palmas de Gran Canaria, Espagne.

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).

PUSTEJOVSKY J., BOGURAEV B., VERHAGEN M., BUITELAAR P. & JOHNSTON M. (1997). Semantic indexing and typed hyperlinking. In *American Association for Artifical Intelligence Conference, Spring Symposium, NLP for WWW.*, Stanford University, USA.

WILKS Y. & STEVENSON M. (1996). *The Grammar of Sense: is Word-Sense Tagging much more than Part-of-Speech Tagging?* Rapport interne, University of Sheffield, UK.

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