Récital 2001, Tours, 2-5 juillet 2001

Filtrage d'information par analyse partielle
Grammaires locales, dictionnaires électroniques et lexique-grammaire pour la recherche d'information

Antonio Balvet

Université Paris X Nanterre
200, av. de la République 92013 Nanterre
antonio.balvet@u-paris10.fr
Thales Research & Technologies
Domaine de Corbeville, 91404 Orsay Cedex
antonio.balvet@lcr.thomson-csf.com

Résumé

Nous présentons une approche de filtrage d'information par analyse partielle, reprenant les résultats de recherches issues aussi bien de la recherche documentaire que du traitement automatique des langues. Nous précisons les contraintes liées au domaine du filtrage d'information qui militent, à nos yeux, pour une approche linguistique permettant d'obtenir des performances importantes, ainsi qu'une transparence de fonctionnement. Nous présentons quelques résultats concrets pour illustrer le potentiel de l'approche décrite.

Mots-clés : filtrage d'information, TALN, analyse partielle, grammaires locales, lexique-grammaire

Abstract

We present a partial analysis approach to the problem of information filtering, based on the results of different areas of research, ranging from information retrieval to natural language processing. We lay the emphasis on the particular constraints of the information filtering activity compatible with a linguistic and user-friendly treatment (partial analysis). We also present some results measured on an actual corpus in order to illustrate the potential of the described approach.

Keywords: information filtering, NLP, partial analysis, local grammars, lexicon-grammar

1 Introduction

Le filtrage d'information automatique[1], une branche relativement jeune du domaine de la recherche d'information[2], semble se caractériser par une prépondérance des approches statistiques, basées sur l'élaboration d'un index prenant en compte les fréquences des termes présents dans un document. Dès l'apparition du terme "information filtering" (filtrage d'information), au cours des conférences internationales d'évaluation TREC[3], la plupart, voire la totalité, des approches évaluées ont eu pour base un moteur statistique d'indexation des documents (ex : PRISE). Une des questions qui ressort des différentes éditions de TREC est celle de la qualité du traitement automatique opéré ainsi que la garantie des performances. En effet, les approches statistiques du problème supposent, dans un objectif d'amélioration des performances, un paramétrage peu transparent des algorithmes employés, source de difficulté pour l'extension à de nouvelles langues ou, tout simplement, pour une utilisation immédiate (sans procédure d'apprentissage) par un utilisateur non expert en informatique. Nous présentons, pour notre part, une approche faisant essentiellement appel à des ressources linguistiques au format électronique, pour lesquelles se posent des problèmes de maintenance[4] et de mise en œuvre par des utilisateurs non experts en linguistique informatique dans un environnement de travail de type intra/internet. Dans un premier temps, nous préciserons la définition du domaine et des méthodes du filtrage de documents. Puis, nous montrerons le potentiel du recours à une approche linguistique du problème de l'IF. Enfin, nous nous pencherons sur un cas concret : le thème "les prises d'otage".

2 Précisions sur la notion de filtrage d'information

Nous examinerons, tout d'abord, les propriétés des activités de *pull* versus de *push*, puis nous rappellerons les spécificités de l'IF, tel que définies dès 1995 dans le cadre de TREC.

2.1 Activités de *pull* vs. de *push*

Dans le domaine de la recherche d'information, on distingue deux types d'activités, en fonction des critères suivants : nature du fonds documentaire traité, qualité du besoin en

---

[1] Désormais IF (Information Filtering).

[2] Nous opérons une distinction entre l'activité désignée sous le terme "Information Dissemination" (Dissémination d'Information), pratiquée par des opérateurs humains et formalisée dès 1958 (cf. Luhn H.P., 1958), de l'activité de filtrage d'information proprement dite, opérée par des systèmes automatiques. Pour plus de détails sur ces notions, voir (Oard D. & Marchionini G., 1996) ainsi que (Lewis D.D. & Hill M., 1995).

[3] Text REtrieval Conference, organisées principalement par le DARPA (Defense Advanced Research Projects Agency) et le NIST (National Institute for Standards and Technology). Le terme de filtrage d'information est apparu dans les actes des conférences en 1995 et a été inclus dans la compétition proprement dite dès TREC-5 (1996) en tant que tâche secondaire (*track*). Voir à ce sujet (Harman D., 1995), ainsi que (Voorhees E. & Harman D., 1996) et (Voorhees E. & Harman D., 1997).

[4] Révisions, corrections, extension, pérennité des bases de données …

information, contraintes sur le temps de traitement et type de décision de sélection opérée par le système. Les activités de *pull* se caractérisent par un fonds documentaire stable, une taille du fonds documentaire traité pouvant aller jusqu'au To (téraoctet), un besoin en information éphémère, une obligation de temps de traitement "acceptable" (jusqu'à quelques minutes) et une décision de sélection continue (classement des réponses selon un degré de pertinence). Les activités de *push*, de leur côté se caractérisent par un fonds documentaire dynamique (ex : actualisé toutes les minutes), une taille des documents à traiter limitée à quelques Ko (ex : dépêches journalistiques), un besoin en information de nature stable, une obligation de temps de traitement proche du temps réel (quelques millisecondes), ainsi qu'une décision de sélection continue pour le **routage**, binaire (oui/non) pour le **filtrage**. Les systèmes de routage sont le plus souvent des adaptations de moteurs d'indexation et de recherche classiques, ainsi que les conférences TREC l'ont montré. Ils constituent donc, de notre point de vue, des systèmes à la frontière du *pull* et du *push*, raison pour laquelle nous ne nous y intéresserons pas davantage dans la suite du présent article, que nous situons dans le cadre du *push* strict. De leur côté, les systèmes de **filtrage** d'information proprement dit sont quasi-inexistants au sein de TREC.

2.2 Décision de sélection binaire et évaluation des performances

Un système d'IF opère une décision de sélection binaire : tout document entrant dans la chaîne de traitement doit être attribué à un ou plusieurs profils, ou bien être rejeté. L'utilisateur final reçoit les documents filtrés dans l'ordre de leur arrivée, chaque document qui lui parvient est considéré comme pertinent. Les conséquences du caractère binaire de la décision de sélection opérée sont multiples : ainsi, étant donné la sur-représentation des approches statistiques, la plupart des systèmes destinés à l'IF doivent simuler cette décision binaire au moyen d'une fonction de seuil. De ce fait, la majorité des publications de TREC portant sur l'IF décrivent en réalité des heuristiques visant à simuler une décision binaire à partir d'un système opérant une décision continue. Par ailleurs, en ce qui concerne l'évaluation des performances des différents systèmes, un problème semble se poser. En effet, autant le principe d'une évaluation reposant sur les scores de rappel et de précision semble aller de soi pour les systèmes de *pull*, autant, en matière d'évaluation des systèmes de filtrage, un certain flou semble planer. Pourtant, dès la mise en place de la tâche de filtrage, les organisateurs de TREC avaient proposé une métrique, appelée "utilité" (*utility*) prenant en compte le gain (respectivement, le coût) que peut représenter pour un utilisateur une bonne (respectivement, une mauvaise) décision de sélection. Toutefois, cette métrique est qualifiée, par ses auteurs eux-mêmes, d'inadaptée (cf. Lewis D. 1996) et (Hull D.A.,1997). Pour toutes les raisons mentionnées ci-dessus, la présente communication ne pourra pas fournir de mesure quantitative définitive des performances du système de filtrage que nous décrivons. Nous ne disposons, au stade actuel des travaux sur le domaine, que d'éléments concernant la couverture des ressources linguistiques (grammaires locales) employées, évaluée par confrontation avec le corpus.

3 Quelle approche pour un filtrage d'information de qualité?

Ainsi que nous l'avons vu plus haut, l'IF se caractérise par des contraintes fortes portant sur le temps de traitement, ainsi que sur la qualité de la sélection automatique des documents. Or les systèmes les plus répandus, qu'ils aient le degré de maturité suffisant pour participer à TREC

ou non[5], reposent principalement sur une approche "floue". De ce fait, la plupart de ces systèmes ne prennent pas en compte les relations de dépendance existant entre des éléments linguistiques (mots, syntagmes, *chunks*, phrases, paragraphes …) et toute recherche de segments pertinents effectuée sur ces bases ne peut que contenir une part d'imprécision[6]. Pour toutes ces raisons, nous avançons que seul un moteur de filtrage opérant sur des bases linguistiques est à même de garantir la qualité attendue. Ceci nous semble devoir passer par la réduction des possibilités de bruit par une reconnaissance exacte de séquences effectuée grâce à des grammaires locales, ainsi que par une limitation des risques de silence découlant de cette reconnaissance exacte grâce à la mise en œuvre de règles transformationnelles[7].

3.1 Vers un moteur linguistique de filtrage d'information

Nous proposons de fait une alternative aux approches statistiques : un moteur de filtrage par repérage de séquences jugées pertinentes par utilisateur donné et de leurs variantes syntaxiques, par le biais d'une analyse partielle à précision modulable, par transducteurs élaborés grâce à une boîte à outils : le système INTEX[8].

3.1.1 Architecture

Le système présenté, CORAIL, est un prototype de moteur d'IF intégrant des contraintes d'ordre linguistique, par le biais de dictionnaires, de grammaires locales, ainsi que de tables du lexique-grammaire[9]. Ces différentes ressources sont toutes traitées de façon uniforme (traduction sous forme d'automates à états finis), afin d'offrir des performances de traitement optimales proches du temps réel[10]. CORAIL se présente sous la forme d'une plate-forme distribuée (traitements par agents autonomes), développée sous Java pour une mise en œuvre en réseau (intra/internet). La plate-forme intègre des fonctions de gestion des différents profils (distinction entre profils privés/publics, mise à jour dynamique), ainsi qu'une interface-

---

[5] Il existe plusieurs systèmes de filtrage d'information en marge de TREC, relevant, par exemple, du logiciel libre. Ces systèmes sont généralement destinés à sélectionner des messages envoyés à des serveurs de groupes de discussion (*usenet*) ou à opérer un tri dans les messages électroniques de l'utilisateur, et reposent dans leur majorité sur un noyau statistique. Pour plus de renseignements sur ces systèmes, voir l'adresse internet : http ://www.enee.umd.edu/medlab/filter/software.html.

[6] Ainsi, dans la plupart des cas pour de tels systèmes, des séquences telles que "le ministre de la culture" et "la culture du ministre" pourront être jugées équivalentes.

[7] Pour plus de détails, voir (Meunier F., Balvet A. , Poibeau T., 1998).

[8] Voir (Silberztein M., 1993).

[9] L'application de techniques issues du Traitement Automatique des Langues (TAL) constitue un champ de recherche suffisamment établi pour figurer dans la liste des tâches de TREC depuis 1995, voir : (Strzalkowski T. *et al.*, 1995), (Strzalkowski T. *et al.*, 1996) et (Strzalkowski T. *et al.*, 1997).

[10] La version actuellement disponible de INTEX n'est que partiellement intégrable : l'outil était au départ destiné à un public de linguistes, en tant que concordancier. De ce fait, les temps de traitement observés ne sont "que" proches du temps réel.

utilisateur dont la réalisation, évaluée auprès d'un public non spécialiste du domaine[11], tient compte de recommandations ergonomiques.

3.1.2 Chaîne de traitements

La chaîne des traitements du système de filtrage que nous décrivons peut être schématisée comme suit.

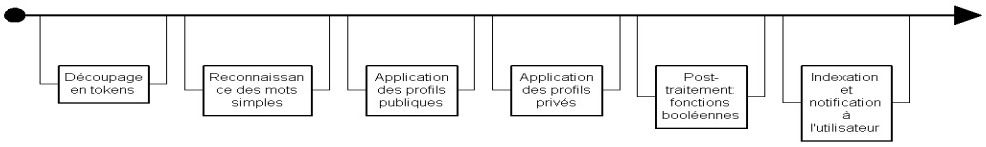

Figure 1 : chaîne de traitement d'un système de filtrage d'information

3.2 Quel format pour les ressources linguistiques ?

Une fois admis le principe d'un filtrage d'information efficace au moyen de ressources linguistiques, reste à préciser un format de ressources, assurant une maintenance optimale.

3.2.1 Vers des bases de données lexicales (ré)utilisables

Les ressources employées au sein de la plate-forme CORAIL sont de plusieurs types : dictionnaires, grammaires locales et tables du lexique-grammaire. Ces dernières constituent, à nos yeux, le format idéal. En effet, suite à leur formalisation sous l'impulsion des travaux du LADL[12] les tables du lexique-grammaire se présentent sous la forme de tableaux associant à une entrée lexicale ses contraintes syntaxico-sémantique[13] sous forme binaires.

A	B	C	D	E	F	G	H	I	J	K
N0=Nspec	N1=Nspec	V	N0 V	N0 V N1	Nominalisation	Synonymie	Actif	Passif	V-Support (procéder à Det vN)	vN
":NEtatSyn"	":NEtatSyn"	<faire> muter	-	+	+	":VRéformerSyn"	+	+	+	<mutation>
":NEtatSyn"	":NEtatSyn"	<moderniser>	+	+	+	":VRéformerSyn"	+	+	+	<modernisation>
":NEtatSyn"	":NEtatSyn"	<modifier>	-	+	+	":VRéformerSyn"	+	+	+	<modification>

Figure 2 : lexique-grammaire simplifié des verbes de réforme

---

[11] L'évaluation visait à estimer la lisibilité des métaphores graphiques employées au cours de l'élaboration de filtres sous la forme de grammaires locales, et non pas à quantifier les performances de filtrage proprement dites.

[12] Voir (Gross M., 1975).

[13] Par exemple : contraintes de sous-catégorisation, familles de transformation acceptées etc …

Cette table permet de spécifier, un ensemble de contraintes syntaxico-sémantiques[14] telles que : le type des actants N0 et N1 ("NSpec" : Nom Spécifique), les construction acceptées (construction absolue N0 V et construction transitive directe N0 V N1), des relations lexicales (nominalisation, synonymie), les transformations valides pour l'entrée considérée (voix active, passive et nominalisation par verbe-support), ou encore la forme verbale nominalisée (VN). Au sein de la boîte à outils INTEX, ces tables sont associées à un métagraphe qui permet d'implémenter[15], entre autres, un certain nombre de transformations syntaxiques (voir fig.4).

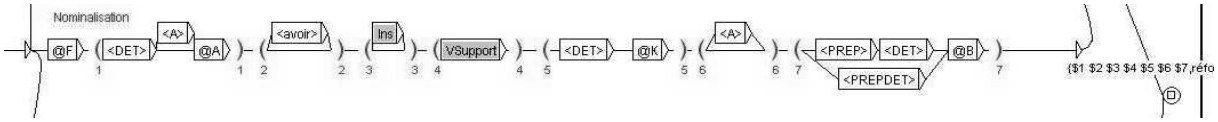

Figure 3 : extrait d'un métagraphe implémentant une transformation de nominalisation avec verbe-support

La prise en compte de variantes syntaxiques est primordiale pour notre objet. En effet, les dictionnaires et les grammaires locales disponibles en standard avec INTEX pour le français ont une couverture largement adaptée pour le niveau des mots (simples ou composés) ainsi que pour certains segments tels que l'expression des dates. Cependant les contraintes de la plate-forme CORAIL ont révélé la nécessité de disposer de bases de données telles que celle de la fig. 3 afin de traiter des segments dépassant les frontières du mot typographique, qui se situent plutôt au niveau du *chunk*. Dans cette optique, le format relativement simple (texte ASCII) des tables du lexique-grammaire, ainsi que la liberté dans l'expression et le choix des différentes contraintes jugées pertinentes pour le domaine en font un outil incontournable.

4 Une expérience de filtrage d'information : les prises d'otage

Dans cette ébauche d'évaluation du système que nous avons présenté, nous nous appuierons sur un corpus journalistique[16] collecté par interrogation manuelle du fonds documentaire. Nous montrerons à partir de ce corpus les potentialités de l'approche décrite plus haut, qui apporte, grâce à la spécification de quelques transformations syntaxiques, un début de réponse au problème de la variation naturelle des formes de surface dans les textes traités. Nous soulignons l'absence, à notre connaissance, de protocole d'évaluation établi, et reconnu par

---

[14] Les travaux décrits ici reprennent des résultats de (Abney S., 1996) et (Grefenstette G., 1996) défendant la pertinence d'une analyse partielle, (Roche E. & Schabes Y., 1997), (Roche E., 1993 a) et b)), en ce qui concerne le principe de l'intégration de tables du lexique-grammaire.

[15] Cet extrait de métagraphe décrit la transformation "nominalisation par verbe-support" : les appels à la table fig. 3 sont marqués par l'opérateur "@" suivi de l'indice de la colonne (A .. ZZ …), la sortie du transducteur (avant-dernier état : "{$1 $2 $3 $4 $5 $6 $7,réforme.N+Réforme}") reprend le format INTEX pour les ressources lexicales. Ce métagraphe constitue un patron syntaxique, que les informations particulières contenues dans les tables associées permettent d'instancier. La transformation de nominalisation ainsi décrite peut se gloser comme suit : un SN sujet constitué d'un Dét. et d'un NSpec, avec insertion possible d'un Adj. entre le Dét. et le N., un SV centré autour d'un verbe-support (décrit dans le sous-graphe VSupport) et d'un SN (Dét. + VN), puis un SN complément (N1) centré autour d'un NSpec.

[16] Edition électronique des archives du journal *Le Monde*.

l'ensemble de la communauté, pour un système d'IF tel que décrit ici. Nous proposons une évaluation classique en termes de taux de rappel.

4.1 Lexique-grammaire du thème "les prises d'otage"

N0 =: NHum	N0 =: N-hum	N0 =: Nprédicatif	Nprédicatif	V	N0 V N1	N1 =: Nhum	N0 V Prep otage N1	N0 V PREP DET SN "moyen coercitif"	Actif	Passif sans agent	Passif avec agent	Nominalisation par Vsupp	V Nominalisé	VNominalisé Prep otage	Complément Durée
+	-	-	-	-	-	+	-	-	+	+	+	+	<rapt>	-	-
+	-	-	-	<bloquer>	-	+	-	-	+	-	+	+	<blocage>	-	+
+	-	-	-	<capturer>	-	+	+	+	+	+	+	+	<capture>	+	-

Figure 4 : table du lexique-grammaire pour le thème "prises d'otage"

Conformément aux principes du lexique-grammaire exposé plus haut[17], nous avons spécifié un certain nombre de contraintes syntaxico-sémantiques sous la forme de traits binaires.

4.1.1 Transformations permises

Nous avons spécifié les transformations suivantes : forme active, forme passive, avec ou sans agent, nominalisation avec ou sans verbe-support (ex : "organiser la capture"). D'autres contraintes ont été définies telles que : le type des N0 et N1 possibles (Humain, Prédicatif …), ou encore des constructions particulières[18] (ex : "N0 garder en otage N1", ou "N0 maintenir sous la menace d'une arme").

4.1.2 Instanciation par métagraphe

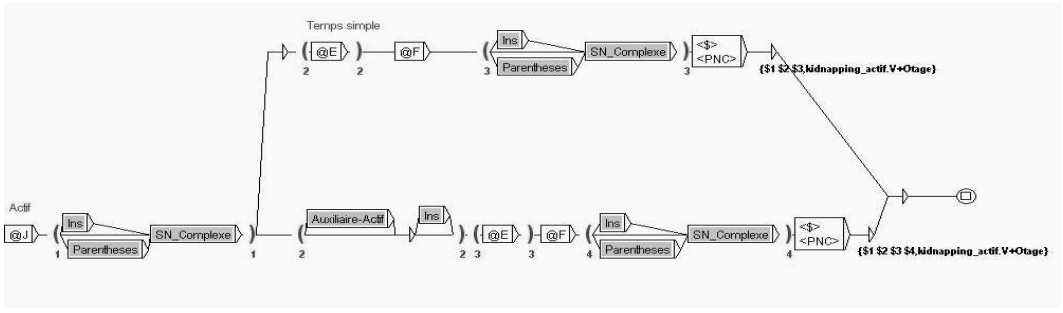

Figure 5 : métagraphe[19] décrivant la voix active, associé à la table fig.5

---

[17] Voir fig. 3.

[18] Le choix de ces contraintes est le fruit de notre expérience du travail sur corpus journalistique.

[19] Dans ce métagraphe associé à la table de la fig. 5, nous spécifions un patron syntaxique pour la voix active, pour laquelle nous distinguons entre forme simple (branche supérieure) et forme composée (branche

La fonctionnalité décrite, implémentée récemment par M. Silberztein, permet d'aborder le problème de l'analyse partielle par cascades de transducteurs de façon plus déclarative que précédemment, dans le sens où les données linguistiques (tables) sont séparées de leur mise en application (grammaire décrite par un métagraphe). Nous pensons également que cette fonctionnalité permet de dépasser les limites tenant à la maintenance d'une base de ressources linguistiques en vraie grandeur pour des applications industrielles.

4.1.3 Quelques résultats préliminaires

Les résultats présentés ici sont à considérer comme une ébauche de validation de l'approche présentée. En effet, ainsi que nous l'avons exposé plus haut, des problèmes spécifiques se posent en ce qui concerne l'évaluation de systèmes d'IF, nous poussant à proposer une méthode visant essentiellement à évaluer la couverture des grammaires mises en œuvre. Par ailleurs, le corpus sur lequel nous nous appuyons est d'une taille limitée, pour des raisons qui tiennent autant au thème retenu qu'au genre textuel particulier que sont les dépêches journalistiques, dont l'effectif est réduit par rapport à l'ensemble des différents genres contenus dans les archives du Monde (ex : chroniques, édito, articles de fond …). Ce corpus a été obtenu par consultation d'un fonds de dépêches extrait des archives numérisées du journal *Le Monde* (de 1987 à 1996) et par validation manuelle des réponses (près de 200) à une requête portant sur le thème "les prises d'otage". Les 14 dépêches retenues constituent celles qui, à nos yeux, traitent effectivement du thème retenu[20]. Ce sous-ensemble comporte les différentes formes suivantes : 19 occurrences de formes du type N0 V N1 (désormais K), où N0 et N1 sont des syntagmes nominaux (humains en l'occurrence), et V l'ensemble des synonymes du verbe "kidnapper" (voir table fig. 5), 6 occurrences de formes figées N0 V N1 (désormais O) en otage (idem), enfin 10 occurrences de formes nominalisées (désormais N) avec ou sans verbe support (ex : "prise d'otages"). D'après nos mesures, les performances de rappel que permettent d'atteindre les règles décrites au moyen des tables et métagraphes présentés plus haut sont les suivantes : 100% pour les formes de type O, 80% pour N et 60% pour K. Nous avons estimé la proportion de formes K, O ou N correctement reconnues parmi celles repérées manuellement ; les ressources ayant servi au repérage des séquences recherchées ont été constituées manuellement.

Ces premiers résultats nous amènent aux conclusions suivantes. (1)Le repérage des formes se rapprochant d'une expression figée ("N0 prendre N1 en otage") est la tâche la plus aisée, ce qui peut s'expliquer, dans notre expérience, par le caractère peu ambigu de telles séquences, malgré la diversité des formes possibles notamment par insertion de compléments entre les

---

inférieure) du verbe principal. La structure ainsi spécifiée est relativement canonique : un SN sujet, suivi d'un SV et d'un SN objet ; les appels à la table fig. 5 ("@E" : forme verbale, "@F" : construction transitive directe) permettent d'étendre le patron ainsi spécifié à l'ensemble des entrées de la table. Par ailleurs, nous permettons, par défaut, l'insertion plus ou moins contrôlée (sous-graphes "Ins" et "Parentheses") de séquences se rapprochant des syntagmes nominaux aux frontières de constituants (ex : entre un verbe et un SN).

[20] Nous soulignons la particularité du filtrage d'information par rapport à d'autres activités de recherche d'information : les documents retenus reflètent l'expérience, en d'autres termes la subjectivité, de l'opérateur humain, alors que des activités reposant sur un principe de sélection "floue", telles que le routage par exemple, doivent viser une objectivité maximale.

éléments du patron "N0 V N1 en otage". (2)Le défaut de couverture pour les formes nominalisées peut être attribué principalement à un défaut de couverture du thesaurus employé. Ce défaut devrait se résorber de lui-même par l'intégration de ressources plus abouties (ex : Semiograph de Memodata, EuroWordnet …). (3)Les formes les plus difficiles à repérer sont celles se rapprochant du niveau de la phrase (phrases construites autour de synonymes de "kidnapper"). Ce constat est peu surprenant ; le niveau de la phrase reste, en raison de la multitude de constructions possibles, le niveau d'analyse le plus complexe pour une approche par grammaires locales.

5 Conclusion et perspectives

Nous avons tenté de montrer la viabilité du principe d'une analyse partielle pour le filtrage d'information. Pour ce faire, nous avons présenté quelques résultats d'un travail de recherche en cours. Ce travail s'était, jusqu'à présent, concrétisé par l'implémentation d'un moteur d'IF reposant sur des grammaires locales "simples", dans une optique de traitement en temps quasi-réel, de transparence de fonctionnement et de contrôle des principes de traitement. Dans le présent exposé, nous avons présenté une avancée par rapport au prototype opérationnel évoqué : un moteur d'IF intégrant des contraintes syntaxico-sémantiques par le biais de tables du lexique-grammaire adaptées au domaine, permettant d'instancier un certain nombre de règles transformationnelles permettant de traiter en partie le problème de la variation des formes de surface pour un thème donné. En prenant l'exemple du thème "les prises d'otage", nous avons donné un aperçu du potentiel de notre approche grâce à quelques résultats préliminaires mesurés sur un corpus journalistique homogène. Nous considérons nécessaire, bien que coûteux, le passage à un corpus réaliste afin d'évaluer les aspects liés au temps de traitement. Par ailleurs, l'intégration des ressources existantes nous paraît une piste intéressante, notamment en vue d'une paramétrisation semi-automatisée des ressources à la tâche. De façon plus large, nous avons tenté de montrer la viabilité d'une approche non statistique de l'IF : le corpus servant de base à nos expérimentations est, par essence, de taille limitée. Ceci nous permet de souligner un des avantages d'une approche par grammaires locales sur toute approche nécessitant une phase d'apprentissage : le prototype présenté permet la détection de "signaux faibles", ainsi qu'une mise en œuvre immédiate des ressources élaborées. Par ailleurs, le prototype CORAIL nous apparaît plus transparent, dans son fonctionnement explicite, qu'un système reposant sur un ensemble d'algorithmes dont le paramétrage serait délicat. De plus, par sa conception même (environnement distribué), CORAIL permet à un utilisateur donné de mettre en œuvre des ressources dont il n'est pas l'auteur, ce qui le place, de fait, dans le cadre du filtrage collaboratif.

Références

Abney S. (1996), Partial parsing via finite-state cascades, Proceedings of the *ESSLLI'96 Robust Parsing Workshop*.

Grefenstette G. (1996), Light Parsing as Finite-State Filtering, Workshop on *Extended Finite State Models of Language, ECAI'96*.

Gross M. (1975), *Méthodes en syntaxe*, Paris, Hermann.

Harman D. (1995), Overview of the fourth Text REtrieval Conference (TREC-4).

Hearst M., Pedersen J., Pirolli P., Schütze H. (1995), Xerox site report : four TREC-4 tracks.

Hull D.A. (1997), The TREC-6 filtering track : description and analysis, Actes des conférences TREC-6.

Lewis D. (1996), The TREC-5 filtering track.

Lewis D.D., Hill M. (1995), The TREC-4 filtering track.

Luhn H.P. (1958), A business intelligence system, *IBM Journal of Research and Development*, Vol 2(4), pp. 314-319.

Meunier F., Balvet A., Poibeau T. (1998), Projet *CORAIL*, *Linguisticae Investigationes*, tome XXII, pp. 369-381.

Oard D.W., Marchionini G. (1996), A Conceptual Framework for Text Filtering, *Technical Report CS-TR-3613*, University of Maryland.

Roche E. (1993), *Analyse syntaxique transformationnelle du français par transducteurs et lexique-grammaire*, Université Paris VII.

Roche E. (1993), Une représentation par automate fini des textes et propriétés transformationnelles des verbes, *Linguisticae Investigationes* tome XVII :1, pp. 189-222.

Roche E., Schabes Y. (1997), *Finite State Language Processing*, Cambridge, MIT Press.

Silberztein M. (1993), *Le système INTEX, Dictionnaires électroniques et analyse automatique des textes*, Paris, Masson.

a) Silberztein M. (1999), *Documentation du système INTEX*, Paris, LADL.

b) Silberztein M. (1999), Traitement des expressions figées avec *INTEX*, *Linguisticae Investigationes*, tome XXII, pp. 425-449.

Strzalkowski T., Perez Carballo J. (1995), Natural language information retrieval : TREC-4 report.

Strzalkowski T., Guthrie L., Karlgreen J., Leistensnider J., Lin F., Perez Carballo J., Straszheim T., Wang J., Wilding J. (1996), Natural language information retrieval : TREC-5 report.

Strzalkowski T., Lin F., Perez Carballo J. (1996), Natural language information retrieval : TREC-6 report.

Voorhees E., Harman D, (1996), Overview of the fifth Text REtrieval Conference (TREC-5).

Voorhees E., Harman D. (1997), Overview of the fifth Text REtrieval Conference (TREC-6).