Évaluation de stratégies linguistiques pour le filtrage d'information

Antonio BALVET

Abstract: Automatic information filtering has become a major issue for both military and civilian activities. However, the existing systems do not always meet quality and transparency expectations, nor do they take into account users' subjectivity towards information processing. This paper describes an ongoing research focusing on the identification and the evaluation of linguistic strategies for high quality information filtering, implemented mostly by the INTEX system local grammar functionalities.

Keywords: natural language processing, information filtering, profile extension, evaluation.

Mots clés : traitement automatique des langues, filtrage d'information, extension de profils, évaluation.

1. Introduction

Le problème du contrôle de l'information textuelle se pose de façon accrue tant dans les domaines civil (veille informative) que militaire (veille stratégique), alors que l'offre en solutions viables et utilisables par le grand public reste limitée. Quelles stratégies retenir dans le cadre d'un système de filtrage d'information afin de garantir la qualité des résultats, la transparence des critères de filtrage et l'ergonomie d'utilisation ? Nous proposons des éléments de réponse à ce problème en décrivant un travail de recherche en cours visant à mettre en œuvre des outils linguistiques pour un filtrage de qualité, reposant principalement sur des grammaires locales mises en œuvre par le système INTEX.

✉ Antonio BALVET, Thomson-Csf Laboratoire Central de Recherches,
Domaine de Corbeville 91404 Orsay Cedex France.
Fax : 33 (0)1 69 33 08 65 e-mail : antonio.balvet@lcr.thomson-csf.com

Dans un premier temps, nous définirons les termes employés dans le cadre du filtrage d'information et nous proposerons une comparaison qualitative des approches linguistiques et non linguistiques du problème. Dans un deuxième temps, nous présenterons une évaluation de l'impact quantitatif d'une approche linguistique du filtrage d'information sur les performances, par rapport à un ensemble de documents pris comme référence par un échantillon de relecteurs, sur le thème des prises d'otage. Puis, nous ferons une synthèse des enseignements que nous tirons de la présente évaluation, qui sera suivie d'un aperçu des perspectives de recherche et de développement.

Ce faisant, nous visons à mettre en place des procédures et des protocoles génériques pour le traitement automatique de l'information textuelle, ainsi qu'une application de ces principes au sein d'un système opérationnel : la plate-forme CORAIL développée au LCR Thomson-Csf.

2. Le filtrage d'information

La terminologie du domaine étant encore peu stabilisée, en raison de calques par rapport à l'anglais notamment, il paraît utile de préciser le sens des différentes notions auxquelles nous aurons recours tout au long de la présente étude. Les conclusions que nous tirons du présent travail ne sont valables que dans le cadre ainsi défini.

2.1. Précisions terminologiques

2.1.1. *Filtrage*

Le filtrage d'information vise à détecter, en un temps minimal — idéalement en « temps réel » — un ensemble de thèmes, qui représentent les centres d'intérêt d'un ensemble d'utilisateurs, au sein d'un flux d'information textuelle. Les centres d'intérêt des utilisateurs sont regroupés sous le terme de « profils ». Ces profils sont stables, alors que le flux d'information est dynamique, ce qui distingue le filtrage d'information de l'activité de recherche documentaire (par moteurs de recherche), dans laquelle les centres d'intérêt sont mouvants alors que le fonds documentaire est (relativement) fixe. Cette définition semble

faire l'unanimité, au vu de la littérature existante [1].

2.1.2. *Séquence pertinente*

Une séquence pertinente se définit comme suit : une séquence dont la forme de surface correspond à un centre d'intérêt d'utilisateur, exprimé sous la forme d'un filtre (*i.e.* une grammaire locale). Dans le cadre du filtrage d'information par automates à états finis, la couverture visée d'une grammaire locale dépend de l'ensemble des séquences considérées comme pertinentes. Contrairement à des domaines comme l'extraction d'information ou la compréhension automatique de textes, le filtrage d'information n'a pas pour objectif premier d'opérer des inférences sur les documents analysés (*e.g.* : si « la victime » est composée d'un déterminant et d'un nom, alors il s'agit d'un groupe nominal). De ce fait, aucun appel à des considérations sémantiques n'est fait à propos de la pertinence telle que nous la définissons ici, ni à propos des notions connexes que sont le bruit et le silence : nous considérons que les effets sémantiques ne sont pas, à ce jour, traitables de façon satisfaisante en respectant les contraintes du filtrage d'information par des automates à états finis [2].

2.1.3. *Bruit*

Dans les termes de la recherche documentaire, le bruit se définit grossièrement comme l'ensemble des documents ne correspondant pas à l'attente d'un utilisateur exprimée par une requête. Le taux de bruit se calcule comme suit :

- Précision : Bons documents retrouvés / (Bons documents retrouvés + Fausses alarmes)
- Bruit : 1 – Précision.

---

[1] Pour un état de l'art et un traitement plus épistémologique de la notion de filtrage d'information, voir OARD et MARCHIONINI 1996.

[2] À notre connaissance, aucun système accessible au grand public ne traite de façon exhaustive les problèmes sémantiques en respectant les contraintes de traitement en temps réel. Par ailleurs, il paraît illusoire de vouloir traiter des phénomènes de cet ordre de complexité au moyen des seules ressources utilisables dans le système INTEX.

Les taux de précision et de bruit (ainsi que ceux de rappel et de silence définis plus bas) sont des indicateurs de référence en évaluation de systèmes de recherche documentaire (*e.g.* : évaluation des performances d'un moteur de recherche) ; ce sont les indicateurs utilisés au cours des campagnes d'évaluation internationales MUC[1] et TREC[2]. Toutefois, dans le cadre du filtrage d'information utilisant des automates à états finis, la question de la pertinence de cette notion de bruit se pose. En effet, bien que cet indice semble adapté dans le cas des moteurs de recherche opérant sur une base statistique, ce n'est pas le cas pour le filtrage d'information par grammaires locales. En effet, les moteurs de recherche à base statistique considèrent de façon implicite la langue (les documents) comme un « sac de mots » dans lequel l'ordre des éléments n'est pas pertinent. Or, pour le français du moins, l'ordonnancement des mots (la syntaxe) est pertinent, ce qui explique qu'à une requête exprimée sous la forme d'une conjonction de mots clés soient éventuellement associés des documents non pertinents dans lesquels les termes ne figurent pas dans le même ordre que celui de la requête, ou à des distances importantes les uns des autres. Dans le cadre du filtrage d'information par grammaires locales, en revanche, une fois un profil défini, la reconnaissance de séquences se fait de façon binaire et exacte, et non pas selon un continuum comme c'est le cas pour un moteur de recherche à base statistique.

De ce fait, le choix de la technologie des automates à états finis intégrant des contraintes linguistiques, telle qu'elle est implémentée par INTEX par exemple, implique que, dans l'acception du terme donnée ci-dessus, les profils exprimés sous la forme de transducteurs ne génèrent pas de bruit[3].

---

[1] Message Understanding Conference, dont les résultats, pour MUC 6, sont visibles à l'adresse internet http://www.cs.nyu.edu/cs/faculty/grishman/muc6.html.

[2] Text Retrieval Evaluation Conference, dont le site officiel est situé à l'adresse http://trec.nist.gov/. Lespinasse *et al.* (1999) développent, par ailleurs, une comparaison ainsi qu'une évaluation du projet TREC et de son pendant français Amaryllis. Pour un aperçu de la dernière conférence TREC 7, voir VOORHEES et HARMAN 1998.

[3] Ce point sera discuté plus en détail plus bas, dans le contexte d'une comparaison des performances du système de filtrage avec celles de relecteurs humains.

2.1.4. *Silence*

La notion de silence se définit habituellement comme : un document de l'ensemble de référence non repéré. Le taux de silence se calcule comme suit :

- Rappel : Bons documents retrouvés / (Bons documents retrouvés + Manqués)
- Silence : 1- Rappel.

Contrairement au taux de bruit, le taux de silence, dans son acception habituelle en recherche documentaire, nous apparaît comme pertinent dans le cadre du filtrage d'information tel que nous l'avons défini. Le taux de silence y est associé à une couverture insuffisante de la grammaire locale définissant un profil donné.

2.2. Comparaison entre approches du filtrage d'information

Un certain nombre d'approches du problème du filtrage de l'information textuelle existent, incarnées par des logiciels plus ou moins largement distribués. Il convient de distinguer, dans ces différentes approches d'un même problème, celles qui mettent l'accent sur la dimension linguistique des textes traités de celles qui ne considèrent les documents que comme des « sacs de mots ». Nous examinerons tout d'abord les approches non linguistiques, qui représentent la majorité (voire la quasi-exclusivité) des solutions commercialisées, puis nous nous pencherons sur les approches linguistiques, essentiellement cantonnées au domaine de la recherche. La présente comparaison des différentes approches se veut générale, elle ne tiendra pas compte des évolutions les plus récentes des logiciels particuliers relevant de telle ou telle approche. Une comparaison détaillée justifierait à elle seule l'édition de plusieurs volumes.

2.2.1. *Approches non linguistiques*

Les solutions commerciales disponibles à ce jour sont généralement des adaptations de systèmes de recherche documentaire classiques (moteurs de recherche), où les requêtes sont soumises à intervalles réguliers à un fonds documentaire actualisé en conséquence. Toutefois, dans cette approche, l'impératif de traitement en temps réel ou

minimal n'est pas garanti : le fonds documentaire, dans ce type de systèmes, n'est pas conçu pour être actualisé à des fréquences inférieures à la minute. Or, le fil de dépêches de l'AFP, qui constitue un exemple de flux documentaire relevant du filtrage d'information, fournit environ une dépêche par minute, voire plus en période de pointe. Par ailleurs, dans ce type de systèmes, aucune garantie de scalabilité des temps de traitement n'est donnée en fonction de la taille des profils. En effet, en raison de la nature même du processus de traitement, le passage d'une dizaine de profils à quelques centaines risque de provoquer une saturation de la plate-forme hébergeant le logiciel.

Ces adaptations de systèmes de recherche documentaire sont, pour les raisons évoquées ci-dessus, de piètres candidats pour un filtrage d'information de qualité. En effet, la plupart des systèmes de recherche documentaire font l'impasse sur la charge informative des verbes, des mots grammaticaux, de la ponctuation, de la constituance, etc. Pour toutes ces raisons, il paraît vraisemblable d'affirmer qu'aucun système accessible au grand public ne fournit, à l'heure actuelle, de garanties quantitatives et qualitatives suffisantes pour un système de filtrage en temps réel, prenant en compte des contraintes linguistiques[1].

2.2.2. *Approches linguistiques*

Les approches linguistiques du problème du filtrage de l'information textuelle apparaissent comme une bonne alternative aux systèmes décrits plus hauts. En effet, bien qu'il paraisse encore utopique de vouloir intégrer une profondeur d'analyse linguistique suffisante dans le cadre de l'activité de recherche documentaire sur des fonds documentaires de l'ordre du Giga-octet, le domaine du filtrage d'information opérant sur des textes courts de quelques dizaines de Kilo-octets (courrier électronique, dépêches journalistiques, messages de Usenet, etc.) paraît un candidat de choix pour une telle intégration.

[1] Ainsi, par exemple, la totalité des logiciels de filtrage référencés à l'adresse internet suivante : http://www.ee.umd.edu/medlab/filter sont des variantes d'approches par mots clés et de vectorisation : BUCKLEY *et al.* 1994a et BUCKLEY *et al.* 1994b.

Toutefois, à notre connaissance, aucune solution commerciale de filtrage d'information sur des bases linguistiques pour grand public n'est actuellement disponible sur le marché. De ce fait, nous ne mentionnerons que les premiers résultats obtenus pour un prototype développé au laboratoire central de recherches de Thomson-Csf : le système CORAIL.

Ce prototype utilise la technologie des automates à états finis pour la reconnaissance de patrons de recherche représentant les profils d'utilisateurs[1]. Les performances, en termes de temps de traitement, sont proches du temps réel : une dépêche de l'AFP est traitée en moins d'une minute, temps qui reste inférieur à la fréquence moyenne de diffusion (une dépêche par minute). En ce qui concerne la qualité du filtrage, aucune évaluation complète n'est disponible pour l'instant, les résultats donnés ci-dessous ne sont à considérer, au niveau quantitatif, que comme des résultats préliminaires. Toutefois, au niveau qualitatif, ils semblent confirmer le choix de la technologie des automates à états finis, implémentée par le système INTEX, comme moteur de filtrage. Les performances en termes de bruit et de silence semblent, en l'état actuel des recherches, conformes aux prévisions : peu ou pas de bruit (dans l'acception donnée plus haut), une part de silence dépendante du domaine.

Les recherches en cours sur le système CORAIL visent à enrichir les filtres disponibles en respectant les contraintes linguistiques des domaines considérés, notamment par le recours à un module de reformulation et d'extension sémantique. Malgré le caractère embryonnaire du système CORAIL, les caractéristiques mentionnées plus haut devraient, au minimum, améliorer l'existant en fournissant une plateforme modulaire, réutilisable, intégrant des contraintes ergonomiques permettant de rendre les utilisateurs finaux autonomes tout en leur assurant un taux de performance satisfaisant.

[1] Ce prototype se base sur les résultats des recherches suivantes : MORIN 1999, ROCHE 1993, ROCHE et SCHABES 1997.

3. Évaluation : les prises d'otage

Une première évaluation du système CORAIL a été menée à bien, dont nous relatons ici les résultats. Cette évaluation s'est déroulée de façon relativement classique, par comparaison des résultats du système CORAIL avec ceux d'un groupe de relecteurs, annotant des dépêches journalistiques manuellement. Le corpus considéré est constitué des archives d'articles du journal *Le Monde*, de 1987 à 1996, soit 9 années, collecté grâce à l'interface de consultation livrée avec le CD-ROM, à laquelle une requête portant sur le mot clé *otage* et sur la taille des documents (courts) a été soumise. La spécification de la taille des documents (taille : court) a permis de restreindre la recherche à des textes de type « dépêche », semblables dans leur style ainsi que dans leur structuration aux dépêches « brutes » fournies par le fil de l'AFP. Cette dernière source de documents n'a pas été retenue pour la raison suivante : étant donnée la quantité de dépêches collectées par jour, il semble peu probable qu'un nombre statistiquement probant de dépêches soit consacré aux prises ou tentatives de prise d'otage au cours de la procédure d'évaluation (quelques semaines).

Les performances des différentes stratégies ont été évaluées grâce au calcul des indices définis ci-dessus (rappel, précision, bruit et silence).

3.1. Phase 1 : constitution d'une référence

La constitution d'une référence à partir d'un corpus trié par des relecteurs humains est une procédure habituelle dans le cadre de l'évaluation de systèmes de traitements de l'information (cf. protocoles d'évaluation des conférences MUC et TREC). Nous avons décidé de mettre en place une procédure légèrement différente des procédures standard, pour des raisons d'ordre pratique, en constituant un ensemble cible de documents jugés les plus pertinents, après validation par des relecteurs non experts du domaine des prises d'otage. Cette phase d'évaluation, dont les modalités précises ainsi que les résultats sont détaillés plus bas, a surtout permis de mettre en évidence l'importance de la subjectivité des relecteurs non experts face à un document ainsi que la nécessité de prendre en compte cette subjecti-

vité dans la phase de conception d'un système de traitement de l'information.

3.1.1. *Protocole d'évaluation*

Trois relecteurs ont été affectés à une tâche de classification manuelle du corpus constitué de 100 dépêches indexées dans la base des articles du *Monde* par le mot clé *OTAGE*. Les relecteurs avaient pour consigne de valider ou d'invalider une dépêche donnée, de surligner les passages jugés pertinents pour chaque dépêche et de spécifier, le cas échéant, le caractère central du document par rapport au thème « les prises d'otage ». De ce fait, une gradation dans le degré de pertinence attribué aux différentes dépêches était possible.

Ce protocole a été élaboré dans le but de comprendre les écarts éventuels de classification des dépêches en examinant les passages jugés pertinents, ainsi que dans une optique de constitution d'une référence fine permettant de procéder à une évaluation linguistiquement pertinente des stratégies de filtrage exposées plus bas. En effet, en isolant les passages pertinents, ce protocole permet d'être sûr, dans une certaine mesure, que les performances des différentes stratégies sont dues à la couverture des filtres employés et non à un effet de bord d'un algorithme ou au seul hasard.

Les trois relecteurs retenus, familiers des problématiques d'extraction/filtrage d'information, ont été laissés libres de leurs critères de classification. Autrement dit, aucune étape de concertation entre relecteurs n'a été organisée, ce afin de simuler une situation vraisemblable de filtrage d'information par des non-experts : en pratique, les opérateurs travaillent de façon majoritairement isolée. Toutefois, il convient de souligner que cette procédure diffère de celle habituellement suivie en évaluation de systèmes de recherche documentaire, qui fait généralement appel à des experts du domaine guidés dans leur tâche par des règles explicites de catégorisation [1].

[1] Ainsi, par exemple, un guide destiné aux relecteurs dans la tâche de repérage d'entités nommées de la conférence MUC 6, comportant une vingtaine de pages.

3.1.2. *Résultats*

Les résultats obtenus pour les dépêches retenues comme pertinentes (valides et centrales) à la majorité par les trois relecteurs sont exposés ci-dessous. La subjectivité et le flottement dans les critères de catégorisation des relecteurs sont visibles dans la proportion de dépêches retenues comme pertinentes à la majorité sur l'ensemble des documents : 33 %. Si l'on ne considère que les dépêches retenues comme pertinentes à l'unanimité, le score est encore plus faible : 23 %. De même si l'on ne considère que les dépêches considérées comme centrales : seules 20 % des dépêches sont retenues à l'unanimité. Or, la proportion de dépêches validées par au moins un des trois relecteurs s'élève à 87 %. Ces chiffres sont à rapporter aux chiffres annoncés par les conférences internationales d'évaluation des systèmes de traitement automatique de l'information telles que MUC et TREC : les scores sont généralement proches des 80 % d'accord entre juges. Pour indication, le taux d'accord entre juges obtenu lors de cette première évaluation est à peine supérieur à des scores qui seraient le fait du hasard.

3.1.3. *Subjectivité et traitement de l'information*

Les premiers résultats obtenus au cours de l'établissement de cette référence sur le thème « les prises d'otage » ont permis de dégager les enseignements suivants, qui couvrent aussi bien des aspects quantitatifs que qualitatifs de l'activité de filtrage d'information par des non-experts.

Le premier de ces enseignements est la prégnance de la subjectivité face à la catégorisation de documents écrits, pour une population de non-experts. Cette subjectivité n'est pas propre à l'activité de filtrage, elle est inhérente à toute activité de catégorisation et a fait l'objet de recherches. Les experts, de leur côté, tendent à stabiliser leurs critères de catégorisation au moyen de l'explicitation de ces critères, y compris sous forme de règles qui assurent un taux de performances quasi constant. Par conséquent, ces premiers résultats militent de fait pour l'explicitation des critères de catégorisation sous la forme d'un guide pour annotateurs, ce qui constitue la procédure standard des évaluations MUC ou TREC.

Le second enseignement, qui découle directement du premier, est la variabilité des critères de catégorisation au cours du temps et au fur et à mesure de l'évaluation, soulignée par l'ensemble des évaluateurs. Cette variabilité semble imputable à plusieurs facteurs, d'après les retours d'expérience fournis par les relecteurs. Un des facteurs prépondérants semble être l'apprentissage des habitudes stylistiques des rédacteurs de dépêches : ainsi, une dépêche uniquement informative est généralement complétée par plusieurs dépêches développant chacune d'autres aspects de l'événement.

Un autre facteur semble être la constitution progressive, par lecture/évaluation de chaque dépêche, d'une expertise du domaine couvert par les dépêches (*i.e.* les prises d'otage). Cette expertise influe considérablement sur le processus d'évaluation dans le sens d'une baisse du nombre de dépêches acceptées, traduisant la stabilisation des critères de catégorisation.

La mise en évidence, par le biais d'une démarche expérimentale, d'un processus d'interprétation face à un thème donné permet de préciser les spécifications d'un système de traitement de l'information utilisable aussi bien par des experts que par des utilisateurs en cours d'acquisition d'une expertise sur un domaine donné (tous les opérateurs du filtrage ne sont pas des experts accomplis). Cette prépondérance de la subjectivité des non-experts face à l'information milite pour la conception de systèmes intégrant cette donnée et laissant le maximum de latitude à l'utilisateur dans l'élaboration de ses ressources pour le filtrage. En d'autres termes, les systèmes de traitement de l'information doivent offrir une souplesse suffisante pour s'adapter à chaque stratégie et à chaque type d'utilisateur, tout en préservant les performances et la facilité d'utilisation. De ce fait, un système de filtrage, tel que nous le concevons, se doit de disposer d'une « intelligence linguistique » certaine (*i.e.* intégrer des contraintes linguistiques) bien qu'invisible à l'utilisateur, uniquement sensible dans les choix proposés à l'utilisateur par le biais de l'interface d'utilisation.

Par ailleurs, d'un point de vue quantitatif, les résultats obtenus au cours de cette première évaluation montrent que, en l'état, les résultats d'une validation menée sur une population de non-experts ne fournissent pas de données exploitables tant les critères de catégorisation

sont peu stabilisés. Toutefois, quelle que soit l'exploitabilité des résultats obtenus au cours de cette première évaluation, il semble que la procédure même d'évaluation doive être soumise à caution. En effet, la stabilisation des critères de catégorisation au moyen d'un guide, dans le cadre d'une évaluation sur une population d'experts, peut être vue comme l'introduction d'un biais dans la procédure : ne seront considérés comme pertinents que les documents dont les caractéristiques se laissent analyser de façon explicite ; les documents se prêtant mal à l'analyse seront de fait écartés. A contrario, une procédure se basant sur les résultats d'un seul expert reconnu apparaît également peu probante en termes statistiques. Par ailleurs, il nous apparaît nécessaire de souligner les spécificités de l'activité de filtrage telle que nous l'avons définie par rapport à celle de recherche documentaire telle qu'elle est habituellement pratiquée, qui interdisent d'étendre sans précautions les procédures d'évaluation développées pour un domaine (la recherche documentaire) à un autre (le filtrage d'information par grammaires locales).

Dans la suite de cet exposé, nous développerons des stratégies de filtrage inspirées de stratégies couramment adoptées en recherche documentaire et implémentées grâce au logiciel INTEX. Nous aborderons également les résultats de stratégies proprement linguistiques, rarement présentes dans des applications commerciales. Nous avons choisi de mener nos expériences sur un fonds documentaire stable, ce qui est contraire à la définition du filtrage d'information telle que nous l'avons donnée plus haut. Cependant, pour des raisons tenant à l'évaluation des différentes stratégies, il n'est pas possible de procéder à un filtrage dans des conditions réalistes sur un fonds documentaire dynamique. De ce fait, les considérations concernant le temps de traitement ne feront pas partie de l'évaluation, bien que ce paramètre soit crucial pour un système de filtrage en temps réel. L'évaluation des performances des différentes stratégies sera faite par comparaison avec le corpus de référence décrit plus haut, pour lequel nous soulignons que toute interprétation se doit d'être prudente.

Par ailleurs, dans la suite de cet exposé, nous prendrons pour acquise l'idée que la recherche d'information à base de mots clés isolés ne prenant en compte aucune contrainte d'ordre linguistique est

une stratégie dépassée. De ce fait, l'évaluation portera sur différentes stratégies faisant appel à des outils proprement linguistiques (*i.e.* des grammaires locales), qu'une interface de consultation à base de mots clés seuls ne permettrait pas de mettre en œuvre.

3.2. Phase 2 : extension morphologique

Une des premières stratégies que nous aborderons est l'extension morphologique. Cette stratégie fait partie des procédures courantes en recherche documentaire. Toutefois, elle ne possède habituellement pas le degré de granularité que permettent d'atteindre les dictionnaires à large couverture fournis avec INTEX.

3.2.1. *But de l'expérience*

Le but de l'extension morphologique est de réduire le silence induit par la variabilité des séquences de caractères à rechercher, variabilité en partie imputable, au niveau des mots, aux règles morphologiques propres à chaque langue[1]. Dans la présente expérience, nous ne nous cantonnerons pas au simple mot clé *otage* : nous mettrons à profit l'ensemble des ressources disponibles en standard avec INTEX et nous utiliserons des filtres dépassant les limites du mot isolé. Le filtre évalué intègre une part importante des contraintes pesant sur le terme *otage* au sein de syntagmes : essentiellement les séquences *prise d'otage* et *prendre N1en otage*.

En ce sens, les filtres élaborés dans la suite de cet exposé sont plus que des simples collections de mots clés; ils intègrent une part des contraintes linguistiques inhérentes au domaine (choix d'expressions, de tournure, etc.).

3.2.2. *Expression de la requête*

Le filtre, tel que nous l'avons élaboré, est le suivant :

[1] D'autres sources de variabilité sont, par exemple, les erreurs de transcription ou les variantes locales.

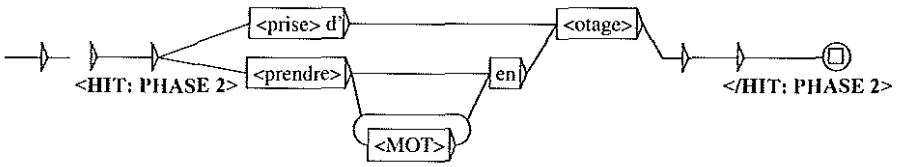

Figure 1. Phase 2 : extension morphologique

L'élaboration de ce filtre a pris environ 10 minutes; toutefois, cette durée est purement indicative : elle peut varier en fonction de l'expertise de l'utilisateur — expertise du domaine, expertise linguistique autant qu'expertise en recherche documentaire.

3.2.3. *Résultats*

Après évaluation par comparaison avec la référence, les performances peuvent être évaluées comme suit :

– taux de bruit : 0.3
– taux de silence : 0.5.

Les taux calculés ci-dessus ne sont qu'une mesure de la distance entre la référence et les performances du système automatique : comme nous l'avons montré plus haut, la notion de bruit est problématique. Par contre, le taux de silence nous paraît significatif : la stratégie d'extension morphologique que nous décrivons ici apparaît comme rapide à mettre en œuvre (placer des chevrons : < > autour d'un terme), notamment en raison de la disponibilité de ressources importantes pour le français. Cependant, une dépêche pertinente sur deux n'est pas repérée.

3.3. Phase 3 : extension sémantique

L'extension sémantique constitue une des stratégies classiques en recherche documentaire afin de limiter le silence. Cette extension consiste généralement à adjoindre de façon automatique des synonymes aux différents termes discriminants de la requête (respectivement, du filtre). Les noms communs contenus dans la requête sont généralement les termes jugés discriminants. La procédure d'extension sémantique par synonymie peut être pondérée par des contraintes terminologiques définies pour un domaine donné. À l'extension

synonymique, ou extension horizontale, peut s'ajouter une extension verticale : intégrer à la requête initiale des termes plus génériques (hyperonymes) ou plus spécifiques (hyponymes), ainsi que des termes reliés à ceux de la requête par d'autres relations sémantiques (métonymie, causalité, pragmatique, etc.).

Les principaux enseignements des expériences d'extension automatique de requêtes, tels que la littérature en fait état[1], sont les suivants :

- une extension sémantique non ou mal contrôlée dégrade les performances, notamment en ce qui concerne le bruit ;
- il est nécessaire de prendre en compte les contraintes inhérentes au domaine. Si cette condition n'est pas remplie, l'opération d'extension sémantique est faite en pure perte : elle allonge le temps de traitement et dégrade les performances du système, tout en rendant la tâche de l'utilisateur plus difficile (les résultats sont plus difficiles à comprendre).

En l'état actuel des recherches, nous proposons une phase d'extension sémantique manuelle : la sélection des termes pertinents sera faite par l'utilisateur. Toutefois, nous préparons un module d'extension sémantique automatique dans le cadre de notre travail de recherches.

3.3.1. *But de l'expérience*

Dans le but de réduire le silence, nous intégrerons plusieurs termes sémantiquement reliés à ceux du filtre élaboré pour la phase 2. Cette extension sémantique se basera sur un dictionnaire de synonymes au format électronique élaboré à l'Université de Caen. Nous avons choisi d'étendre le profil de la phase 2 en intégrant des synonymes, des hyperonymes, ainsi que des hyponymes. La tâche d'extension sémantique est rendue plus délicate par le statut de certaines expressions relevant du thème considéré. Ainsi l'expression « prendre N1 en otage » paraît quasi figée, du moins dans le corpus étudié : aucune

[1] Sur l'utilisation de ressources lexicales (thesaurus) pour la recherche documentaire voir notamment MANDALA *et al.* 1998.

insertion de modifieurs entre les constituants de cette séquence n'a été constatée sur les documents cibles (*e.g.*, « prendre subrepticement les usagers en otage »).

3.3.2. *Expression de la requête*

Le filtre, tel qu'il a été élaboré, est le suivant :

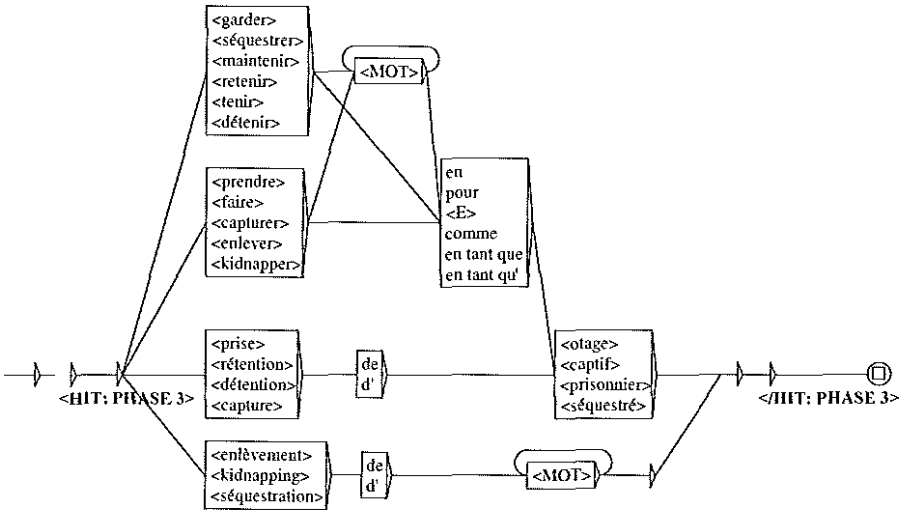

Figure 2. Phase3 : extension sémantique

Le filtre ci-dessus constitue une version étendue du filtre de la phase 2. En effet, la structure est la même; toutefois, elle a été complétée de synonymes choisis par le concepteur en fonction de son expérience du domaine. Le temps d'élaboration (manuelle) de ce filtre est d'environ 30 minutes. Toutefois, ce temps est susceptible de varier en fonction du concepteur et de sa connaissance du domaine. La qualité attendue influe également notablement sur le temps de conception, en fonction du coût que représente le taux de silence pour l'utilisateur.

3.3.3. *Résultats*

Les résultats, après comparaison avec la référence, sont les suivants :

– taux de bruit : 0.4
– taux de silence : 0.3

De ces résultats, on peut tirer les conclusions suivantes : l'extension sémantique apporte un gain de performances en ce qui concerne le taux de silence, par rapport à l'ensemble des documents cibles. À ce stade, on peut affirmer que les différences de performance sont à apprécier en fonction de la stratégie de l'utilisateur et du coût que représente pour lui une augmentation du bruit : on a pu voir que le taux de silence avait tendance à baisser entre les phases 2 et 3.

3.4. Phase 4 : reformulation

Cette phase de reformulation s'inscrit dans la logique des stratégies présentées plus haut : minimiser le silence dû à la variabilité des formes de surface. Les stratégies précédentes s'attachaient à gérer cette variabilité au niveau du mot isolé : les extensions sémantiques proposées sont des termes isolés; dans le meilleur des cas elles sont des expressions quasi figées équivalentes (*e.g.*, « prise d'otage » et « enlèvement »). Dans la suite de cet exposé, nous tenterons de franchir le seuil des mots simples pour atteindre celui des syntagmes, voire de la phrase. En effet, nous proposons d'étendre les filtres existants au moyen d'un ensemble de transformations syntaxiques pertinentes pour le domaine (*e.g.*, la passivation : « un touriste a été pris en otage par un commando d'hommes masqués » et « un commando d'hommes masqués a pris un touriste en otage »).

3.4.1. *But de l'expérience*

Les transformations que nous tenterons de mettre en œuvre sont :

- la passivation
- l'insertion de modifieurs (ADV, incises, compléments circonstanciels, etc.).

L'hypothèse que nous suivons est la suivante : les transformations retenues sont susceptibles d'améliorer les performances, notamment en termes de taux de silence, par la prise en compte de variantes syntaxiques, et non plus seulement lexicales, des formes de surface susceptibles de développer le thème « prise d'otage ». Bien qu'un certain nombre d'opérations de reformulations soient susceptibles d'être étudiées, il nous est apparu, à l'examen du corpus, que les transforma-

tions de passivation et d'insertion de modifieurs étaient les plus pertinentes pour le domaine des prises d'otage. Ceci ne signifie toutefois nullement qu'il s'agit là des seules transformations pertinentes, tous domaines confondus.

3.4.2. *Expression de la requête*

Le filtre, tel que nous l'avons élaboré, est le suivant :

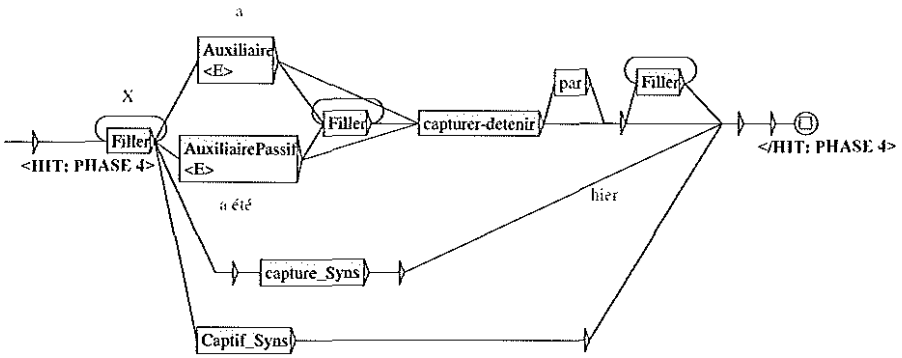

Figure 3. Phase 4 : reformulation

3.4.3. *Résultats*

Après comparaison avec le corpus de référence, les performances peuvent être évaluées comme suit :

– taux de bruit : 0.6
– taux de silence : ≈ 0

Dans le cas précis des transformations implémentées par le filtre ci-dessus, il est pertinent de parler de bruit. En effet, ce filtre tolère, dans cette version, l'insertion de chaînes de caractères peu contraintes (correspondant à l'expression régulière : (<MOT>*+<NB>*+<PNC>*)). Pour plus de pertinence, il conviendrait de spécifier plus avant les insertions possibles pour la terminologie du domaine des prises d'otage. Par contre, on peut remarquer que le taux de silence est proche de zéro. Autrement dit, il semble que la stratégie décrite ci-dessus se situe à l'opposé de la première stratégie décrite dans la phase 2 de la présente évaluation : elle dépasse les frontières du mot

isolé en intégrant des contraintes du niveau du syntagme, elle est relativement longue à mettre en œuvre (principalement en raison de l'absence de ressources disponibles) mais assure une couverture quasi parfaite du thème des prises d'otage malgré un fort taux de bruit (les deux tiers des dépêches repérées ne sont pas considérées pertinentes par les relecteurs humains).

4. Conclusions

Nous avons tenté de montrer quelles pistes semblaient les plus prometteuses pour l'amélioration des performances d'un système de filtrage d'information utilisant des grammaires locales. Nous avons, pour ce faire, mené une première phase d'évaluation, dont les résultats préliminaires sont principalement de deux ordres [1].

Premièrement, en ce qui concerne le processus de filtrage lui-même, il apparaît nécessaire, ainsi que la littérature consacrée au domaine de la recherche documentaire le souligne, de prendre en compte les spécificités du corpus traité, ainsi que celles du domaine (le thème) de filtrage. En effet, nous avons vu qu'une stratégie permettant l'insertion non contrôlée de modifieurs (verbaux comme nominaux), telle que nous la présentons dans la phase 4 de la présente évaluation, était susceptible de dégrader les performances en ce qui concerne le taux de bruit. Ceci est dû au fait que la grammaire locale, telle que nous la présentons ci-dessus, n'est pas suffisamment contrainte. Ceci pose le problème, au-delà des performances pour le thème des prises d'otage, de la réutilisabilité et de la généricité de telles ressources. La solution de ce problème nous paraît passer par une étude terminologique semi-automatisée poussée, donc par la mise au point ou l'intégration d'un extracteur terminologique dont les modalités restent à préciser (généralisation de règles de façon algorithmique, par

[1] Des informations complémentaires sur les initiatives d'évaluation de systèmes linguistiques de traitement de l'information sont développées dans les *Cahiers d'études et de recherches francophones Langues*, notamment LESPINASSE *et al.* 1999, MARIANI 1999, POIBEAU 1999 et POPESCU-BELIS 1999. Les articles cités développent différentes facettes de la problématique de l'évaluation appliquée aux systèmes de traitement automatique du langage naturel.

apprentissage, etc.) et qui soit compatible avec les outils de filtrage existants.

Deuxièmement, malgré la difficulté exposée ici d'aboutir à un ensemble de documents de référence pour le thème considéré, il est apparu que les indices habituellement utilisés en évaluation de systèmes de recherche documentaire à base de mots clés, tels que les taux de bruit et de silence, n'étaient pas directement utilisables, sans un minimum de précautions, dans le cadre du filtrage d'information à base de grammaires locales. Précisons toutefois que ceci ne remet pas en cause la pertinence des indices ni les procédures habituelles d'évaluation en recherche documentaire. Cependant, il nous apparaît que les résultats de la présente évaluation, tant positifs que négatifs, montrent surtout les spécificités de l'activité de filtrage, notamment celles du filtrage d'information prenant en compte des contraintes linguistiques (constituance notamment). De ce fait, il nous apparaît qu'une évaluation pertinente des performances d'un tel système doit plus passer par la prise en compte de la subjectivité des utilisateurs face à l'information, ainsi que par une étude de satisfaction individuelle, que par des calculs tendant à objectiver l'activité de filtrage d'information, subjective par essence du point de vue humain. Autrement dit, il nous apparaît nécessaire, au vu des premiers résultats ici exposés, d'inclure un volet qualitatif aux futures évaluations et de souligner les spécificités de l'activité de filtrage telle que nous l'avons définie.

En définitive, les premiers résultats obtenus au cours de la présente évaluation ne permettent pas, en l'état actuel de nos recherches, de déterminer précisément l'apport d'une approche linguistique du problème du filtrage d'information[1]. En effet, l'introduction d'outils linguistiques automatisés semble poser davantage le problème des attentes des utilisateurs d'un tel système que celui de la pertinence de l'approche linguistique elle-même. Autrement dit, malgré des résultats encourageants pour certains domaines, il est possible que l'apport de la linguistique au problème du filtrage d'information de qualité soit inégal en fonction des domaines et des attentes des utilisateurs. Ainsi,

[1] Cette difficulté n'est pas propre au domaine du filtrage, elle est également valable pour le domaine de la recherche documentaire : ZHAI *et al.* 1996.

par exemple, il n'est pas exclu de penser qu'une activité de veille sur une personnalité (*e.g.* : Bill Gates) ne nécessitera pas le recours à l'ensemble des outils linguistiques exposés ci-dessus[1] (extension morphologique, sémantique et reformulation), ce qui ne sera pas forcément le cas du domaine des changements de position (nominations, démissions, licenciements, etc.).

Par ailleurs, les perspectives de recherche ouvertes par cette première phase d'évaluation sont essentiellement axées autour de l'automatisation des procédures d'extension linguistique des filtres, ainsi qu'une étude plus poussée des processus de paraphrase, en faisant éventuellement appel à des recherches proprement linguistiques du phénomène[2]. Autrement dit, une fois la faisabilité démontrée d'un système de filtrage sur des bases linguistiques, nous voulons donner plus de poids au versant proprement linguistique des recherches exposées ici, notamment en ce qui concerne les opérations de reformulation les plus utiles en fonction d'un domaine donné. Par ailleurs, les problèmes de généricité et de réutilisabilité des ressources nous paraissent mériter une attention particulière dans une optique d'application en grandeur réelle.

Bibliographie

BUCKLEY (Chris), ALLAN (James), SALTON (Gerard) : 1994, « Automatic Routing and Ad-Hoc Retrieval using SMART : TREC 2 », TREC 2.

BUCKLEY (Chris), SALTON (Gerard), ALLAN (James), SINGHAL (Amit) : 1994, « Automatic Query Expansion using SMART : TREC 3 », TREC 3.

FUCHS (Catherine) : 1980, *Paraphrase et théories du langage ; contribution à une histoire des théories linguistiques contemporaines*

[1] Dans le contexte du présent article, toute dépêche mentionnant Bill Gates traitera forcément du démantèlement de Microsoft.

[2] Notamment MEL'ČUK 1989, KAHANE 1999 pour le modèle Sens-Texte, ou encore FUCHS 1980 et FUCHS 1982, ainsi que TIMIMI et BEN ALI 1998 pour une application au domaine de l'extraction et du filtrage d'information.

et à la construction d'une théorie énonciative de la paraphrase (Université Paris VII).

FUCHS (Catherine) : 1982, *La Paraphrase* (Paris : Presses Universitaires de France).

KAHANE (Sylvain) : 1999, « The Meaning Text Theory », *Dependency and Valency. An International Handbook of Contemporary Research* (Berlin : De Gruyter).

LESPINASSE (Karine), KREMER (Patrick), SCHIBLER (Danielle), SCHMITT (Laurent) : 1999, « Évaluation des outils d'accès à l'information textuelle : les expériences américaines (TREC) et française (Amaryllis) », *Cahiers d'études et de recherches francophones Langues* (Aupelf-Uref, John Libbey Eurotext : France).

MANDALA (Rila), TOKUNAGA (Takenobu), TANAKA (Hazumi), OKUMURA (Akitoshi), SATOH (Kenji) : 1998, « Ad Hoc Retrieval Experiments using WordNet and Automatically Constructed Thesauri », TREC 7.

MARIANI (Joseph) : 1999, « Traitement automatique de la langue française utilisant le paradigme d'évaluation », *Cahiers d'études et de recherches francophones Langues* (Aupelf-Uref, John Libbey Eurotext : France).

MEL'ČUK (Igor A.) : 1989, « Semantic Primitives from the Viewpoint of the Meaning-Text Theory », *Quaderni di Semantica*, Italie.

MORIN (Emmanuel) : 1999, « Acquisition de patrons lexico-syntaxiques caractéristiques d'une relation sémantique », *T.A.L* 40, n°1, p 143-166.

OARD (Douglas W.), MARCHIONINI (Gary) : 1996, « A conceptual Framework for Text filtering", Technical Report CS-TR-3613 (University of Maryland).

POIBEAU (Thierry) : 1999, « Évaluation des systèmes d'extraction d'information : une expérience sur le français », *Cahiers d'études et de recherches francophones Langues* (Aupelf-Uref, John Libbey Eurotext : France).

POPESCU-BELIS (Andrei) : 1999, « L'évaluation en génie linguistique : un modèle pour vérifier la cohérence des mesures », *Cahiers d'études et de recherches francophones Langues* (Aupelf-Uref, John Libbey Eurotext : France).

ROCHE (Emmanuel) : 1993, « Une représentation par automate fini des textes et des propriétés transformationnelles des verbes », *Lingvisticae Investigationes* (Amsterdam/Philadelphia : John Benjamins).

ROCHE (Emmanuel), SCHABES (Yves) : 1997, *Finite State Language Processing* (The Massachussets Institute of Technology).

TIMIMI (Ismaïl), BEN ALI (Salaheddine) : 1998, « La paraphrase entre l'analyse et la génération (Système 3AD) », RIFRA' 98, Rencontre Internationale sur l'extraction, le filtrage et le résumé automatique (Tunisie).

VOORHEES (Ellen M.), HARMAN (Donna) : 1998, « Overview of the Seventh Text REtrieval Conference (TREC-7) », TREC 7.

ZHAI (Chengxiang), TONG (Xiang), MILIĆ-FRAYLING (Nataša), EVANS (David A.) : 1996, « Evaluation of Syntactic Phrase Indexing - CLARIT NLP Track Report », TREC 5.