Modèle d'exploration contextuelle pour l'analyse sémantique de textes

Slim Ben Hazez, Jean-Pierre Desclés, Jean-Luc Minel

CAMS-LaLIC, UMR 8557 du CNRS, EHESS, Université de Paris-Sorbonne
96 boulevard Raspail
75 006 Paris – France
{Slim.Ben-Hazez}@paris4.sorbonne.fr

Résumé – Abstract

Nous présentons dans cet article un modèle d'exploration contextuelle et une plate-forme logicielle qui permet d'accéder au contenu sémantique des textes et d'en extraire des séquences particulièrement pertinentes. L'objectif est de développer et d'exploiter des ressources linguistiques pour identifier dans les textes, indépendamment des domaines traités, certaines des relations organisatrices des connaissances ainsi que les organisations discursives mises en places par l'auteur. L'analyse sémantique du texte est guidée par le repérage d'indices linguistiques déclencheurs dont l'emploi est représentatif des notions étudiées.

In this paper, we present a model of contextual exploration and a workstation dedicated to semantic filtering and relevant sentence extracting. The purpose is to develop and to exploit linguistics resources in order to identify in texts, independently of processed domains, some specific relations which organize knowledge and author discourse. Semantic analysis is driven by the identification of linguistic indicators which are relevant clues for the studied notions.

1 Introduction

Les techniques actuelles d'analyse sémantique de textes tendent de plus en plus à mettre en œuvre une analyse locale fondée sur le repérage d'indices textuels de certaines informations sémantiques présentes dans les textes. La stratégie généralement adoptée par les systèmes d'extraction d'information (Pazienza, 1997) repose sur une analyse locale qui utilise des techniques de pattern-matching (Grishman, 1997). En pratique, les indices textuels sont des patrons d'extraction très spécialisés qui dépendent étroitement du domaine traité. Ces systèmes ont révélé leur limites en termes de portabilité et d'évolution par rapport aux besoins des utilisateurs (Poibeau et Nazarenko, 1999). Pour chaque application et pour chaque nouveau domaine, il faut élaborer de nouveaux patrons spécifiques, les tester et construire les automates

Les évaluations réalisées sur certains systèmes de résumé automatique (Minel et al., 1997 ; Jing et al., 1998) ainsi que les travaux menés en collaboration avec les résumeurs professionnels (Enddes-Niggemeyr, 1993) ou en comparaison avec les résumés produits par

ces professionnels (Saggion et Lapalme, 1998) ont néanmoins montré la difficulté à réaliser des résumés standard, c'est-à-dire construits sans tenir compte des besoins des utilisateurs.

En dépit des performances obtenues dans les systèmes de résumé automatique et d'extraction d'information, l'étude des stratégies utilisées soulève un certains nombre de questions: Comment définir des tâches réutilisables dont les ressources et les traitements peuvent s'adapter aux besoins de l'utilisateur ? Comment modéliser les ressources linguistiques et les rendre accessibles ? Comment exploiter l'organisation du texte et explorer le contexte linguistique pour lever l'indétermination sémantique des indices linguistiques recherchés ?

L'objectif de nos travaux est de construire une plate-forme de filtrage sémantique de textes qui vise à donner quelques éléments opérationnels pour répondre à ces questions. L'originalité de notre approche (Desclés et al., 1997, Minel et al., 2001) revient à se donner les moyens d'accéder au contenu sémantique des textes, pour mieux les cibler et en extraire des séquences particulièrement pertinentes. D'une part, nous cherchons à exploiter directement l'organisation textuelle des propos de l'auteur. D'autre part, nous nous intéressons aux manifestations textuelles de certaines relations organisatrices de connaissances (relations définitoires, causales, spatiales…). Notre but est de cibler, à l'aide de marqueurs linguistiques et de certaines connaissances grammaticales, des séquences textuelles qui peuvent exprimer un certain savoir sur le monde. Ce savoir ne se réduit pas à une nomenclature (objets, propriétés, événements, etc.). Il est notamment structuré par un certain nombre de relations entre concepts, événements, etc.

Le modèle adopté pour répondre aux besoins du filtrage sémantique de textes consiste à identifier, indépendamment d'un domaine particulier, certaines informations sémantiques adaptables en fonction des besoins des utilisateurs. Ce modèle est fondé d'une part, sur l'identification dans les textes de marqueurs linguistiques d'une catégorie (grammaticale ou discursive) ou d'une notion étudiée, et d'autre part sur une exploration du contexte[1] des marqueurs identifiés. Cette exploration permet: d'interpréter le contexte d'un marqueur linguistique; d'analyser la position d'un marqueur dans le texte (début de phrase, premier paragraphe, etc.); de manipuler les éléments structurels du texte (titres, paragraphes,...); d'identifier la structure thématique, etc. Pour mettre en œuvre ce modèle et fournir des outils qui permettent de développer et de déployer des ressources linguistiques orientées vers le filtrage sémantique de textes, nous avons développé la plate-forme *Filtext* en utilisant le modèle d'exploration contextuelle. Nous allons présenter, à travers des exemples de ressources et de tâches réalisées, la modèle d'exploration contextuelle et de données linguistiques et l'architecture de la plate-forme.

2 Modélisation des données linguistiques

2.1 Acquisition des données linguistiques

La méthode d'exploration contextuelle est issue d'une réflexion initiale sur le traitement informatique des valeurs aspecto-temporelles SECAT (Desclés et al., 1991). La méthode a été

[1] Cette analyse du contexte d'un marqueur linguistique ne se limite pas aux notions de "concaténation" ou de "contexte contiguë".

ensuite généralisée en tenant compte des indications présentes dans le contexte pour un calcul des valeurs sémantiques relevant de différentes tâches (Jouis, 1993, Berri, 1996, Desclés et al., 1997a, 1997b).

D'après le modèle d'exploration contextuelle, l'étude linguistique consiste à déterminer les valeurs sémantiques des marqueurs linguistiques d'une catégorie grammaticale ou d'une notion discursive. Selon les cas, une carte sémantique (par exemple, le réseau des relateurs de repérage) (Desclés, 1987) peut être construite pour une catégorie ou une notion étudiée. Le processus d'acquisition (figure 1) pour chaque tâche est fondé sur une étude systématique de corpus de textes pour y rechercher des indicateurs discursifs explicites dont l'emploi est représentatif de la valeur sémantique considérée ou de la notion étudiée. C'est en ce sens que ces indicateurs deviennent des marqueurs de valeurs sémantiques. Comme exemple de champ grammatical, donnons celui qui couvre l'identification des valeurs aspecto-temporelles associées aux morphèmes des temps grammaticaux du français. Pour le champ du discours, mentionnons par exemple les indicateurs discursifs des annonces thématiques, des expressions définitoires, des relations entre concepts, des relations de causalité, des relations temporelles entre événements, etc.

L'identification d'un marqueur (grammatical ou discursif) n'est cependant pas suffisante pour déterminer complètement la valeur sémantique du marqueur. En effet, un indicateur linguistique (indice déclencheur) est rarement un marqueur univoque d'une valeur sémantique unique. Ayant identifié une occurrence de marqueur sous la forme d'un indicateur répertorié, il faut, dans un deuxième temps, explorer le contexte de cette occurrence pour rechercher d'autres *indices* linguistiques, sous la forme d'occurrences d'indices complémentaires. Ceux-ci permettront dans un cas favorable de lever l'indétermination sémantique attachée *a priori* au marqueur analysé. Dans ce cas une étiquette sémantique pourra être attribuée à un segment linguistique (syntagme, phrase, paragraphe selon les cas). Les indices permettront également d'invalider les hypothèses sémantiques qui pouvaient être envisagées à propos du marqueur analysé dans son contexte. L'exploration contextuelle est gouvernée par un ensemble de règles (dite d'exploration) qui, pour un indicateur donné et une décision à prendre, recherchent d'autres indices explicites dans un espace de recherche (phrase, paragraphe,…). Par exemple, pour la proposition *j'ai pris mon cachet*, la marque du passé composé ne suffit pas elle seule pour décider de la valeur aspecto-temporelle attachée à la proposition (on parle d'indétermination sémantique). Les deux classes d'indices contextuels (*ouf, enfin, ça y est,…*) et (*ce matin,…*) contribueront à lever l'indétermination sémantique et orienteront respectivement vers la valeur "d'état résultant" et "d'événement" (Desclés et al., 1997a).

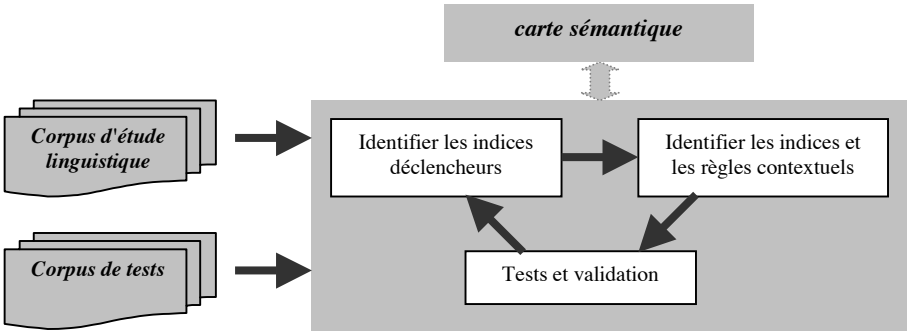

Figure 1 : Processus d'acquisition des données linguistiques

L'acquisition de ces données linguistiques nécessite le choix d'un corpus de travail qui dépend généralement de la tâche à réaliser et une fouille systématique des textes en vue d'accumuler les indicateurs, les indices contextuelles et les règles qui les combinent ; cette fouille est complétée par un travail de réflexion linguistique, afin de dégager des régularités textuelles.

2.2 Modèle de données linguistiques

Le travail de modélisation a pour objectif de capitaliser et de rendre accessibles les ressources linguistiques. Pour répondre à l'objectif d'une acquisition incrémentale et capitalisable des données linguistiques, nous avons défini un modèle et des outils qui permettent au linguiste de construire et de maintenir des bases de données linguistiques spécifiques à des tâches d'identification d'informations sémantiques (Ben Hazez et Minel, 2000). Ce modèle répond aux spécifications suivantes:

- Les données linguistiques sont organisés par tâche. Chaque tâche se voit associée une base de données de marqueurs linguistiques et de règles d'exploration contextuelle. La notion de tâche est un moyen qui permet au linguiste d'organiser ces connaissances de manière indépendante. Le modèle permet de gérer plusieurs bases linguistiques et de les organiser sous forme d'arbre. Par exemple, la base de données linguistique associée à la tâche de résumé automatique est composée de plusieurs sous-tâches de repérage: des annonces thématiques, des conclusions/récapitulations, des résultats, des hypothèses, etc.

- La stratégie d'exploration de texte est guidée par des *indices déclencheurs* (indicateurs). Contrairement à une stratégie guidée par les règles, la description et la reconnaissance de ces indicateurs ne sont pas donc codées dans les règles.

L'interface homme machine (IHM) présentée dans la figure 2 illustre un extrait des tables de création et de manipulation des données linguistiques. Cette interface permet au linguiste de constituer sa base de données linguistiques en spécifiant, les tâches, les classes d'indicateurs et les règles d'exploration contextuelle associées:

- La table "TACHE_fr" permet de déclarer ou de sélectionner une tâche. Comme le montre l'exemple de la figure 2, les données affichées représentent une vue de la base globale relative à la tâche sélectionnée « Résumé ».

- La table "CLASSES_MARQUEURS_fr" contient les déclarations des classes d'indicateurs (indices déclencheurs) et les indices contextuels. Par exemple, la classe "&Crecap1.10" sélectionnée contient des indicateurs d'énoncés de récapitulation.

- La table "MARQUEURS_fr" contient la description des marqueurs (motifs linguistiques) pour chaque classe de la tâche courante. Par exemple, les expressions linguistiques sélectionnées ("en conclusion", "pour finir", "en résumé",…) représentent des indicateurs de récapitulation de la classe "&Crecap1.10

- La table "REGLES_fr" permet de déclarer pour chaque règle, le nom de la règle, l'étiquette sémantique associée, la classe des indicateurs qui déclenche la règle et le segment textuel à étiqueté. Par exemple, la ligne sélectionnée de la table spécifie le nom de la règle déclenchée par les indicateurs de la classe "&Crecap1.10". L'action de cette règle consiste à attribuer l'étiquette sémantique "Récapitulation" à chaque phrase contenant un indicateur de la classe "&Crecap1.10" vérifiant les contraintes de la règle.

Ces tables sont organisées selon un schéma conceptuel implanté dans un système de gestion de base de données relationnelle. Pour faciliter la gestion des bases linguistiques nous avons intégré dans l'IHM des fonctions de tri, de recherche, de filtrage ainsi que des outils d'interrogation et de vérification des contraintes d'intégrités des données linguistiques.

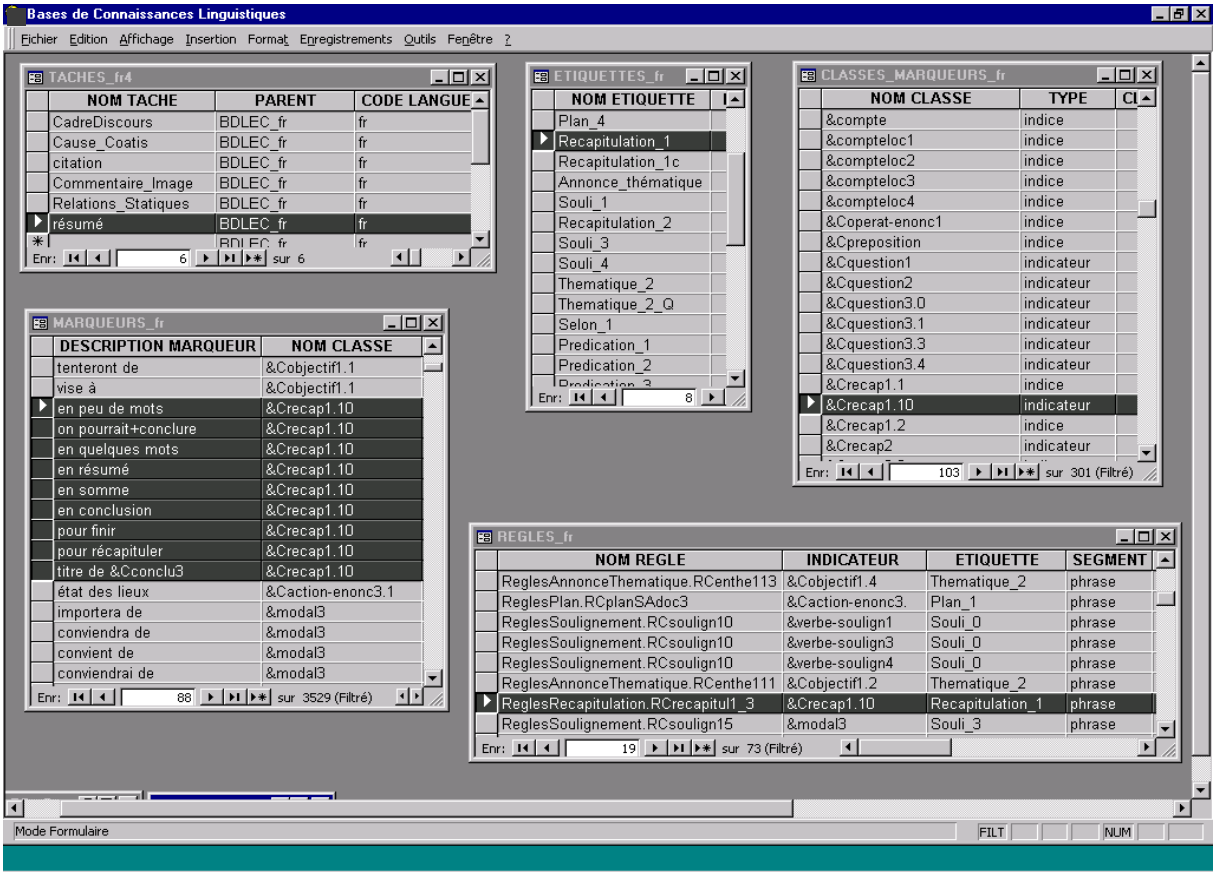

Figure 2 : Tables de définition et de manipulation des données linguistiques

2.2.1 Langage de description des marqueurs linguistiques

Les marqueurs linguistiques sont des unités lexicales simples ou composées (morphèmes, mots, expressions et locutions, …) formées à partir d'unités atomiques ("tokens") et des opérateurs de concaténation, de disjonction et de répétition. Chaque motif linguistique décrit un ensemble de réalisations possibles de séquences textuelles continues ou éventuellement discontinues. Ces marqueurs linguistiques sont regroupés dans des classes en fonction de critères syntaxiques ou sémantiques[2]. Une classe est identifiée par un nom précédé par le symbole "&". Par exemple, l'expression *il est + &importance* permet de repérer des lexies du type: *il est primordial ; il est particulièrement important ; il est,…, essentiel ;* etc. Pour une tâche donnée seules certaines flexions d'un verbe sont significatives. Ainsi, si le but est de

[2] A partir d'une première expression il est possible de construire une classe d'expressions par des opérations de synonymie, de nominalisation ou de paraphrase. L'utilisateur doit par la suite valider les expressions engendrées et éliminer les expressions incorrectes ou non équivalentes.

rechercher les annonces thématiques d'un article scientifique, le verbe *présenter* est significatif seulement lorsqu'il est employé à l'indicatif présent ou au futur, à la première personne du singulier ou du pluriel. Le langage intègre un ensemble de classes d'expressions prédéfinies (dates, énumérations, abréviations, nombres, etc.) identifiées par un "tokenizeur". L'intégration d'un étiqueteur morpho-syntaxique permet d'améliorer la qualité de la reconnaissance en effectuant des recherches par lemmes, par catégories grammaticales et par utilisation des informations flexionnelles.

2.2.2 Langage de description des règles

Les règles d'exploration contextuelle sont exprimées dans un langage formel de type déclaratif comme le montre l'exemple de la figure 3. Ce langage est centré sur la notion d'un espace de recherche, c'est-à-dire un segment textuel déterminé à partir de l'indicateur, espace dans lequel les indices complémentaires doivent être recherchés. Il est important de pouvoir exprimer des contraintes qui prennent en compte la dimension textuelle. La partie condition de la règle explicite les conditions que doivent vérifier les indicateurs et les indices complémentaires. Le langage permet d'exprimer différentes conditions, comme l'existence, la position dans le texte et l'agencement des indices. La partie action consiste à attribuer une étiquette à un segment textuel ou à déclencher une autre règle.

```
/* Tâche déclenchante : thématique ;
capte un schéma du type : il semble .. crucial
Indicateur : &modal3; */
E1 := Créer_espace(PhraseParent_de Indicateur);
L1:= &verbe-etat3 ; L2 := &adjectif-necessité
Condition: Il_existe_un_indice y appartenant_à E1 tel_que
classe_de y appartient_a (L1 ) ;
Condition : Il_existe_un_indice z appartenant_à E1 tel_que
classe_de z appartient_a (L2) ;
Actions : Attribuer_Etiquette ("Soulignement_Auteur ")
```

Figure 3 : Exemple d'une règle pour repérer des énonces de soulignement par l'auteur

3 Architecture de la plate-forme

La plate-forme FilText s'appuie sur le modèle conceptuel des données linguistiques et le modèle d'exploration contextuelle présentées précédemment. L'une des particularités du système est son ouverture: les utilisateurs peuvent créer de nouvelles bases linguistiques pour d'autres tâches et réutiliser des données ou des composants logiciels. Les traitements sont encapsulés dans des API[3] Java et produisent des sorties selon un format d'échange standard XML, ce qui favorise un déploiement plus facile des ressources linguistiques. Comme le montre la figure 4, la plate-forme est composé de trois sous-systèmes qui coopèrent : i) un gestionnaire des données linguistiques doté d'une IHM; ii) un moteur d'exploration contextuelle qui exploite les données linguistiques pour une ou plusieurs tâches choisies par l'utilisateur et produit une structure hiérarchique du texte avec des "décorations sémantiques"; iii) un ensemble d'agents spécialisés dotés d'IHM et de connaissances qui exploitent les décorations sémantiques générées par le moteur d'exploration contextuel.

[3] Application Programming Interface.

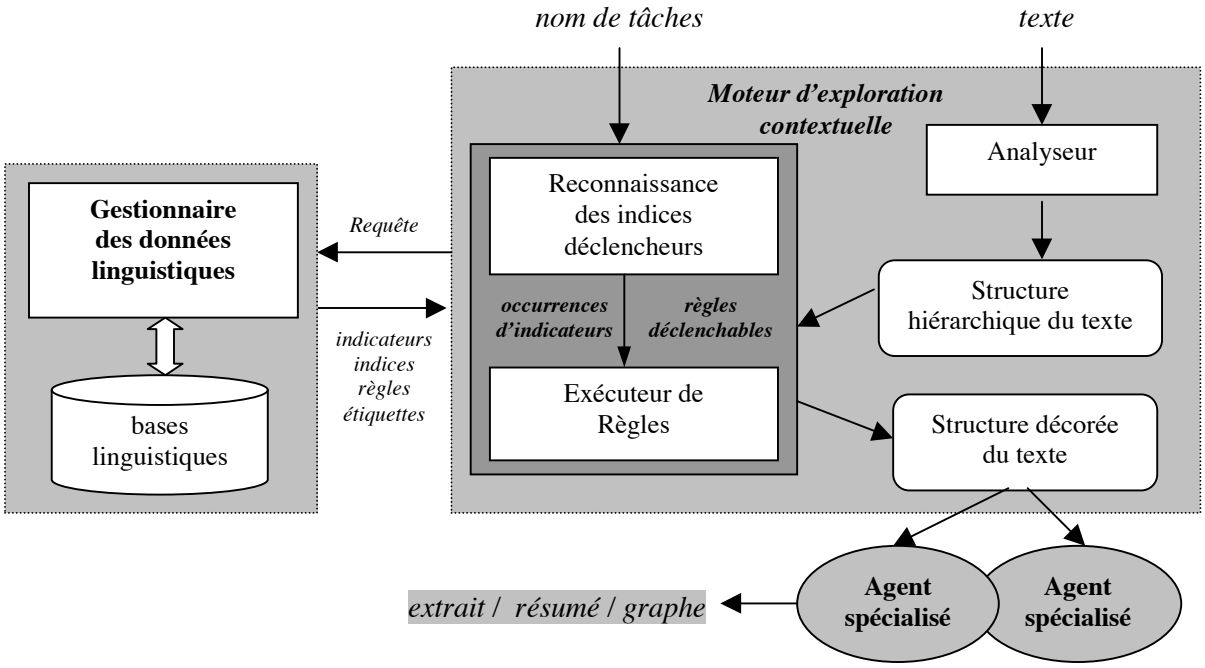

Figure 4 : Architecture générale de la plate-forme

3.1 Gestionnaire des données linguistiques

Les données linguistiques sont capitalisées dans un système de gestion de bases de données relationnelles (SGBDR). Le gestionnaire des données linguistiques est implémenté sous forme d'une API Java qui communiquent avec le SGBDR via la passerelle JDBC (Java Database Connectivity). Cette API offre une vue objet de la base de données permettant le développement d'une IHM et de répondre à des requêtes lancées par les autres sous-systèmes de la plate-forme. Ce module implémente un simple langage de requête qui permet différentes opérations sur la base de données linguistiques: importation de données, sélection, modification, etc. Il intègre aussi un module de *reconnaissance de motifs linguistiques (pattern-matching)* utilisé par les autres composants de la plate-forme. Ce dernier est un automate qui permet de chercher dans un segment textuel (séquence de tokens, phrase, etc.) toutes les occurrences d'un motif linguistique donné. Le résultat de recherche d'un marqueur linguistique dans un segment S est représenté sous forme d'un ensemble de paires de valeurs notées S(u, v); u et v représentent respectivement les positions du premier et dernier éléments de chaque occurrence du motif.

3.2 Moteur d'exploration contextuelle

Le moteur d'exploration contextuelle permet l'étiquetage sémantique des segments textuels ; il est composé de trois modules qui coopèrent :

- *L'analyseur de textes* qui construit une première représentation qui reflète l'organisation structurelle du texte. La construction de cette structure hiérarchique s'appuie sur le texte

segmenté en unités structurelles: sections, paragraphes, phrases. L'analyseur fait appel à trois modules indépendants: un segmenteur en unités structurelles; un tokeniseur qui effectue le découpage en tokens et permet d'identifier la catégorie de chaque token (date, énumération,…); un constructeur du modèle objet du texte qui représente la structure hiérarchique du texte.

- *Le module de reconnaissance des indicateurs* qui a pour tâche de compiler les indicateurs (indices déclencheurs) pour l'ensemble des tâches déclenchées et de les appliquer sur le texte. Cette étape permet de rechercher toutes les occurrences des indicateurs et de déterminer quelles sont les règles à déclencher.

- *Exécuteur de règles* qui permet d'exécuter les règles associées à chaque occurrence d'un indicateur trouvée. Les règles sont considérées comme indépendantes. Ce mode de fonctionnement correspond à l'hypothèse que, pour une tâche donnée, certains marqueurs sémantiques ne sont pas exclusifs entre eux. Par exemple, la présence d'une négation dans une phrase conclusive n'implique pas que cette phrase ne soit pas par ailleurs une « conclusion ». D'autre part, une phrase étiquetée comme « définitoire » peut aussi être étiquetée comme « conclusion ». Toutes les déductions effectuées par les règles sont attribuées aux éléments qui composent la hiérarchie du texte et produisent ainsi une structure hiérarchique « décorée » par des informations sémantiques.

L'algorithme ci-dessous illustre le fonctionnement du moteur d'exploration contextuelle. La première étape consiste à compiler les données linguistiques pour une ou plusieurs tâches à déclencher, puis à déterminer la liste des indicateurs à appliquer sur le texte. Dans cette exemple le moteur applique une analyse du texte par phrase. L'étape suivante du moteur consiste d'abord à chercher dans chaque phrase du texte toutes les occurrences de la liste des indicateurs, puis à exécuter les règles associées à chaque occurrence trouvée.

Fonctions fournies par l'interface (API) du système de gestion des données linguistiques:

T ← compile(nom tâches) : compiler l'ensemble de données linguistiques des tâches déclenchées dans une vue T.
M ← selectIndicateurs(T) : renvoie l'ensemble M des classes d'indicateurs de T.
R ← selectRegles(T, M_j) : renvoie l'ensemble des règles d'exploration R associées a un indicateur M_j de M.
OCC ← identifier_Indicateurs(S_i, M_j) : identifie dans un segment S_i toutes les occurrences de l'indicateur M_j.

Algorithme du moteur:

```
DEBUT
 T ← compile(nom de tâches);
 M ← selectIndicateurs(T) ;
 POUR TOUT  phrase Si FAIRE
        POUR TOUT  indicateur Mj de M FAIRE
                OCC← identifier_Indicateurs(Si, Mj);
                SI  (OCC <> vide)  ALORS  R ← selectRegles(T, Mj);
                        POUR TOUT  occurrence Si(u, v)  de OCC FAIRE
                                Executer_Regle(R, Si(u, v));
                                                //appliquer l'ensemble des règles R  sur
                                                //chaque occurrence de l'indicateur Mj  et
                        FINPOUR                 //étiqueter le segment Si
                FINSI
        FINPOUR
 FINPOUR
FIN
```

3.3 Agents spécialisés

Les agents spécialisés ont pour tâche d'exploiter les « décorations sémantiques» du texte en fonction des objectifs définis par l'utilisateur. Chaque agent possède ces propres connaissances et une IHM de présentation des résultats.

Ainsi, l'agent résumeur-filtreur[4] exploite des connaissances linguistiques qui permettent de repérer des énoncés structurants, des définitions, des relations causales, etc. Par exemple, l'étiquette "annonce thématique" est attribuée aux phrases exprimant le sujet, le thème d'un segment textuel quelconque. L'étiquette "récapitulation/conclusion thématique" est attribuée aux phrases explicitant les conclusions et enseignements généraux du texte. Certaines phrases étant étiquetées, il devient possible de construire des extraits qui répondent aux besoins spécifiques d'un utilisateur en appliquant différentes stratégies de sélection. Cependant, cette extraction brise la cohérence du texte source et peut même introduire des contresens. Des heuristiques simples ont été définies pour détecter potentiellement certains liens anaphoriques, manipuler les marqueurs d'intégration linéaire (comme *en premier lieu, en second lieu, etc),* exploiter la structure et les énumérations du texte, etc. En outre le développement d'IHM qui permettent d'afficher l'arborescence du texte et d'offrir à l'utilisateur des moyens pour naviguer entre l'extrait et le texte source constitue une des réponses offerte par Filtext. Plutôt que de chercher à produire un résumé autonome, indépendant du texte source, en appliquant une analyse profonde pour résoudre les problèmes d'anaphore, et repérer les liens de cohésion et de cohérence, l'objectif se déplace vers la production d'un texte réduit aux informations jugées saillantes pour le lecteur, et la construction de liens qui permettent au lecteur, au vu des informations partielles qui lui sont présentées, de fouiller, à la demande, le texte source.

D'autres agents sont intégrés dans la plate-forme : identification des citations dans les textes (Mourad, 2000) et d'extraire des relations sémantiques (localisation, composition, partie-tout, attribution,…) entre concepts (Le Priol, 2000). Les résultats de ce dernier système sont représentés sous forme de graphes ou de tables relationnelles.

4 Conclusion

Nous avons présenté un modèle d'exploration de texte et une plate-forme logicielle permettant de développer des ressources linguistiques pour extraire certaines informations sémantiques. Les ressources linguistiques peuvent prendre la forme de base de données de marqueurs et de règles, ou de textes avec des "décorations" sémantiques. L'une des particularités du système est son ouverture : il permet aux utilisateurs de développer et manipuler des bases de données linguistiques pour de nouvelles tâches d'identification, l'enrichir en fonction des informations qu'ils recherchent, réutiliser des données et des fonctionnalités du système, etc.

Le modèle envisagé doit être adapté aux besoins des utilisateurs et applicable à tout type de texte. Nous travaillons actuellement à l'intégration d'une analyse thématique fondée d'une part sur des critères statistiques et d'autre part sur le repérage des introducteurs thématiques (Ferret et al., 2001).

L'analyse mise en œuvre et l'acquisition des données linguistiques dépendent étroitement de la nature du corpus et de l'information recherchée. La couverture du corpus, le choix des bons indices textuels et la réalisation de tâches réutilisables constitue une problématique fondamentale de l'analyse textuelle qui nécessite une étude linguistique plus fouillée.

---

[4] Le lecteur trouvera dans (Minel et al., 1997) une évaluation détaillée et dans (Minel et al., 2001) un exemple de résumé produit par cet agent.

Références

Ben Hazez, S., Minel J-L. (2000). Designing Tasks of Identification of Complex Patterns Used for Text Filtering. *RIAO'2000*, Paris, 1558-1567.

Berri, J. (1996). Contribution à la méthode d'exploration contextuelle. Applications au résumé automatique et aux représentations temporelles. Réalisation informatique du système SERAPHIN. Thèse de doctorat, Université Paris-Sorbonne, Paris.

Desclés, J-P. (1987). Réseaux sémantiques : la nature logique et linguistique des relateurs, *langages, Sémantiques et Intelligence Artificielle,* n° 87, p. 55-78.

Desclés, J-P., Jouis, C., Oh, H-G., Reppert, D. M. (1991). Exploration Contextuelle et sémantique : un système expert qui trouve les valeurs sémantiques des temps de l'indicatif dans un texte. In *Knowledge modeling and expertise transfer*, pp.371-400, D. Herin-Aime, R. Dieng, J-P. Regourd, J.P. Angoujard (éds), Amsterdam.

Desclés, J-P. (1997a). *Systèmes d'exploration contextuelle. Co-texte et calcul du sens*. (ed Claude Guimier), Presses Universitaires de Caen, 215-232.

Desclés, J-P., Cartier, E., Jackiewicz, A., Minel, J-L. (1997b). Textual Processing and Contextual Exploration Method. In *CONTEXT'97*, Rio de Janeiro, Brésil.

Endres-Niggemeyer, B. (1993). An empirical process model of abstracting. In *Workshop on Summarizing Text for Intelligent Communication*, Dagsthul, Germany.

Ferret, O., Grau, B., Minel, J.-L., Porhiel, S. (2001). Repérage de structures thématiques dans les textes. *TALN 2001*, Tours.

Grishman, R. (1997) . Information extraction : techniques and challenges. In : Pazienza, M.T. éd. *Information extraction*. Berlin : Springer verlag, 10-27.

Jing, Hongyan, Regina Barzilay et Kathleen McKeown. (1998). Summarization evaluation methods : Experiments and analysis. In *Symposium on Intelligent Text Summarization,* Stanford, CA.

Jouis, C. (1993). Contribution à la conceptualisation et à la modélisation des connaissances à partir d'une analyse linguistique de textes. Thèse de doctorat, EHESS, Paris.

Le Priol, F. (2000). Extraction et capitalisation de connaissances à partir de documents textuels. SEEK-JAVA : Identification et interprétation de relation entre concepts. Thèse de doctorat, Université Paris-Sorbonne, Paris.

Minel , J-L., Nugier, S., Piat, G. (1997). How to appreciate the Quality of Automatic Text Summarization. *Workshop Intelligent Scalable Text Summarization,* EACL 97, Madrid, 25-30.

Minel J-L., Cartier, E , Crispino, G., Desclés J-P, Ben Hazez S, Jackiewicz, A. (2001) Résumé automatique par filtrage sémantique d'informations dans des textes. *Technique et Science Informatiques,* Paris, n° 3.

Mourad, G. (2000). Présentation de connaissances linguistiques pour le repérage et l'extraction de citations. TALN (*RECITAL'2000),* Lausanne, p 495-501.

Pazienza, M.T. (1997) (éd.). Information extraction (a multidisciplinary approach to an emerging information technology*), International Summer School, SCIE'97*, Springer Verlag (Lectures Notes in Computer Science).

Poibeau, T., Nazarenko, A.(1999). L'extraction d'information, une nouvelle conception de la compréhension de texte ?, T.A.L. vol 40., n°2, p 87-115.

Saggion, H., Lapalme, G. (1998). Where does information come from ? Corpus Analysis for Automatic Abstracting. *RIFRA'98, Rencontre Internationale sur l'Extraction le Filtrage et le Résumé automatiques,* Sfax, Tunisie.