Informatique d'assistance pour l'analyse de documents numériques textuels

Vincent Perlerin

Introduction

Nos travaux en informatique ne relèvent pas de l'IA (Intelligence Artificielle) classique mais d'une IA (Informatique d'Assistance) moderne où le calcul ne s'oppose plus au sens, mais où les deux se cherchent des voies de coopération dans des sociétés désormais hybrides, faites d'hommes et de machine. Dans nos travaux, nous utilisons le modèle LUCIA (*Located User-Centred Interpretative Analyser*) qui est à la fois modèle de représentation lexicale et modèle de l'interprétation centré sur un groupe d'utilisateurs [Perlerin, 2004]. LUCIA et les logiciels qui l'implantent, sont tous dédiés à l'assistance à l'accès aux documents numériques textuels et à l'exploration (et/ou à l'analyse) de leur contenu. Le but est de fournir des outils informatiques qui ne nécessitent pas la constitution préalable de données complexes et à visée exhaustive (dictionnaires de spécialité, ontologies, *etc.*) et qui permettent d'assister l'analyse personnalisée de textes numériques.

Description de l'outil

En tant que modèle de représentation lexicale, LUCIA permet de décrire les connaissances d'un utilisateur ou d'un groupe d'utilisateurs sur le lexique d'un domaine en les organisant de façon componentielle selon leurs différences et leurs ressemblances. Ces deux critères organisationnels s'expriment à travers l'utilisation de la notion centrale d'attributs qui sont les pendants des *sèmes* de la Sémantique Interprétative [Rastier, 1987]. Par exemple, les lexies *anticyclone* et *dépression* peuvent être décrites à l'aide de l'attribut [Pression : basse *vs* haute], *anticyclone* actualisant la valeur « haute » tandis que *dépression* actualise la valeur « basse ». Plusieurs attributs peuvent être combinés pour décrire un ensemble de lexies proches. Les lexies décrites par un jeu d'attributs communs constituent une table LUCIA, dont chaque ligne correspond à une actualisation spécifique des attributs mis en jeu. La notion d'héritage d'attribut et de valeurs d'attributs se traduit alors par un lien d'une ligne vers une table. L'ensemble des tables décrivant un domaine particulier est appelé un dispositif. Les dispositifs sont exploitables pour des analyses de documents numériques textuels basées sur le calculs des différences et des récurrences.

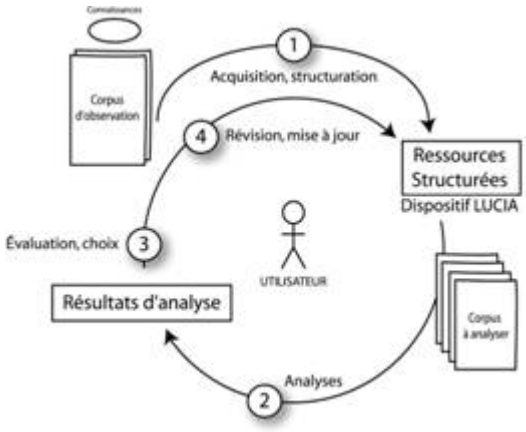

Figure 1 – Processus itératif d'utilisation du système.

L'utilisation de LUCIA s'inscrit dans un processus itératif (Figure 1). À partir d'un corpus d'observation (un ensemble de documents en rapport avec le domaine d'intérêt de l'analyse), le logiciel MEMLABOR assiste l'extraction de graphies redondantes, pertinentes pour la tâche [Perlerin, 2002] (étape 1 sur la figure 1). Le fonctionnement de MEMLABOR est principalement fondé sur des calculs statistiques. Les lexies sélectionnées peuvent faire l'objet de regroupements thématiques. Ces regroupements sont formalisés à l'aide du logiciel THEMEEDITOR qui exploite un principe de coloriage pour assister l'évaluation des associations thématiques par un retour sur le corpus initial [Ferrari et Perlerin, 2004]. Le logiciel LUCIABUILDER permet d'assister la constitution des dispositifs. L'opération, qui est une coopération à part entière, s'étale sur trois paliers : la définition des attributs (éléments de différenciation et de mises en corrélation de lexies), la mise en place des tables (structures d'attributs qui représentant une catégorie sémantique) et la confection des dispositifs (structures de tables et de leurs relations de sous-catégorisation). De telles structures permettent de définir la notion de point de vue et de jugement de leur(s) auteur(s). Elles sont les supports d'analyses automatiques qui prennent en considération à la fois les particularités de la tâche courante et celles des utilisateurs.

Une fois un ou plusieurs dispositifs constitués, les informations qu'ils contiennent sont projetées sur le corpus à analyser (étape 2 sur la figure 1). Cette projection consiste en une annotation des documents et permet d'obtenir un matériau analysable de façon automatique en terme de calculs de redondances d'attributs ou valeurs d'attributs sur des portions définies (zones textuelles, paragraphes, documents, corpus). Ces calculs permettent de mettre au jour des zones thématiques remarquables qui sont présentées à l'utilisateur à travers différents supports d'interaction qui exploitent eux aussi le coloriage (étape 3 sur la figure 1 et figures 2, 3 et 4, p.3). En fonction des résultats obtenus, il est possible de modifier les propositions initiales (étape 4 sur le figure 1). Cette possibilité rend le modèle LUCIA dynamique.

Le système a déjà été exploité dans deux champs d'applications distincts. Le premier relève de la linguistique informatique et correspond à l'étude d'un fait de langue particulier (une métaphore

conventionnelle) [Beust et al., 2003]. Le second s'inscrit dans le champ de la veille documentaire et fait actuellement l'objet d'un projet industriel. Dans ce dernier cadre, les dispositifs permettent de filtrer et de réordonnancer des ensembles de documents proposés comme réponses à une requête soumise à plusieurs moteurs de recherche.

Les outils que nous proposons sont exploitables dans de nombreux cadres de recherche où la prise en considération des particularités praxéologiques dans lesquelles s'inscrivent les textes à analyser est primordiale. L'analyse assistée par l'informatique de forums de discussions constitue l'un des champs d'application possible de nos travaux. Les diversités constatées (thématiques, communautés concernées, *etc.*) entre les forums nécessitent dans un cadre d'analyse informatique, soit l'utilisation de ressources pré-existantes qui présentent nécessairement des lacunes et des inadéquations par rapport aux pratiques à prendre en considération, soit la constitution de grandes quantités des ressources ; tâche généralement coûteuse en temps et en investissement humain. La voie que nous proposons est celle de l'utilisation de ressources limitées en quantité pour assister l'analyse des spécialistes. Cette limitation permet une personnalisation des données pour prendre en considération à la fois les objectifs particuliers de la tâche en cours et les spécificités des conditions de production du matériau sujet d'étude.

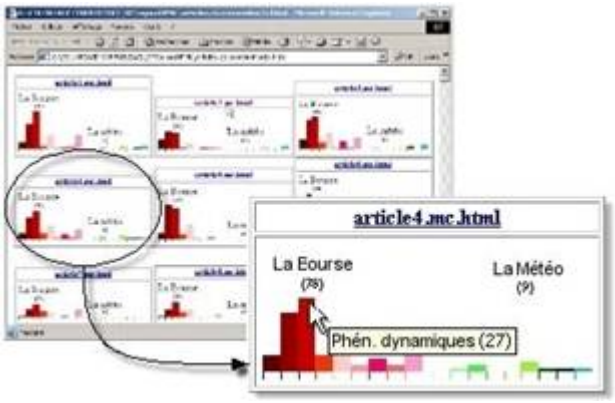

Figure 2 – Interface de parcours des documents analysés pour le projet d'étude de fait de langue.

Les graphiques en histogrammes permettent d'apprécier les proportions de lexies associés aux catégories sémantiques d'un dispositif donné. Dans l'exemple, deux dispositifs ont été utilisés pour l'analyse, un en rapport avec le bourse et l'autre avec la météorologie.

Figure 3 – Interface de lecture d'un document analysé.

Chaque lexie d'un dispositif est coloriée en fonction de la couleur associé à la table dans laquelle est apparaît.

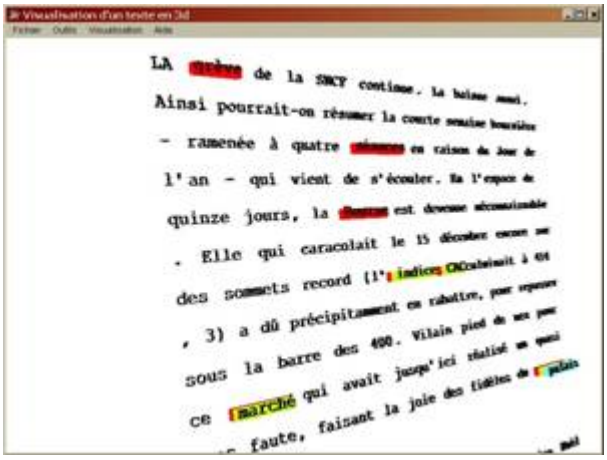

Figure 4 – 3D-LUCIAVizualizer : vue en trois dimensions des redondances de valeurs d'attributs.

La 3D permet d'apprécier simultanément toutes les redondances de valeurs d'attributs au sein d'un texte.

REFERENCES

[Beust et al.,2003] Beust, P., Ferrari, S., et Perlerin, V. (2003). *NLP model and tools for detecting and interpreting metaphors in domain-specific corpora*. Proceedings of the Corpus Linguistics 2003 conference. Archer, D., Rayson, P., Wilson, A. and McEnery, T. (eds.). Lancaster, UK, pp. 114123. UCREL technical paper n°16. UCREL, Lancaster University.

[Ferrari et Perlerin, 2004] Ferrari, S. et Perlerin, V. (2004). *Modèle sémantique et interactions pour l'analyse de documents*. Actes du 7e colloque international sur le document électronique, CIDE 7. La Rochelle. pp. 231-251. Europia, Paris.

[Perlerin, 2004] Perlerin, V. (2004). *Sémantique légère pour le document – Assistance personnalisée pour l'accès au document et l'exploration de son contenu.* Doctorat en informatique de l'université de Caen / Basse-Normandie. 271 p.

[Perlerin, 2002] Perlerin, V. (2002). *Memlabor, un environnement de création, de gestion et de manipulation de corpus de textes*. Actes de la 9e conférence internationale sur le Traitement Automatique des Langues Naturelles, TALN/RECITAL. Tome 1. pp. 507-516. Nancy.

[Rastier, 1987] Rastier, F. (1987). *Sémantique interprétative*. PUF, Paris