Outils de fouille textuelle à partir d'annotations sémantiques

Marc Bertin, Iana Atanassova, Jean-Pierre Desclés

Outils de fouille textuelle à partir d'annotations sémantiques

Marc Bertin*, Iana Atanassova*, Jean-Pierre Desclés*

*Université Paris-Sorbonne
Maison de la recherhce
28 rue Serpente 75006 Paris
{marc.bertin | iana.atanassova | jean-pierre.descles}@paris-sorbonne.fr

Résumé. Nous proposons des outils de fouille textuelle guidée par la sémantique reposant sur des annotations produites automatiquement selon des ontologies linguistiques. Cette approche offre à l'utilisateur final de nouveaux types de produits documentaires permettant une meilleure représentation du document dans le cadre d'une recherche d'informations.

1 Introduction

L'exploitation automatique d'annotations sémantiques, effectuée selon des ontologies linguistiques contenant des points de vue tels qu'hypothèse, citation, méthode, résultat, ... permet de mettre en place de nouveaux produits documentaires, que nous appellerons *fiches de synthèse* ou *bibliographie augmentée*, offrant à l'utilisateur un aperçu structuré et synthétisé des contenus textuels correspondants à des requêtes de recherche d'informations sémantiques. Les outils de fouille textuelle que nous proposons ont pour but de fournir à l'utilisateur les briques de navigation textuelle permettant une meilleure représentation du document dans le cadre d'une recherche d'informations.

2 Système de fouille textuelle

A partir d'un corpus d'articles scientifiques, nous cherchons à extraire, par exemple, les hypothèses de travail, les méthodes utilisées, voire les conclusions, afin de à répondre à un besoin spécifique : au-delà de la simple recherche par mots clés, notre voulons proposer à l'utilisateur un accès aux contenus guidé par la sémantique textuelle. Pour cela, nous avons construit de nouveaux produits documentaires, notamment des synthèses automatiques à partir d'une analyse sémantique des textes.

Les annotations sémantiques sont obtenues par le moteur Excom (Djioua et al., 2006; Alrahabi et Desclés, 2008) qui permet d'identifier dans un texte des phrases porteuses des points de vue sémantiques. Il s'appuie sur des ressources linguistiques sous forme de marqueurs de surface et de règles d'exploration de contexte. Le moteur Excom implémente la méthode d'Exploration Contextuelle, proposée par Desclés (1997). Le système que nous avons mis en place exploite les fichiers annotés au format XML. Ces fichiers sont importés dans une base de données dédiée afin d'assurer les fonctionnalités de recherche d'informations par l'indexation du

contenu textuel et des annotations sémantiques. Le système propose un service web en utilisant la technologie Apache/PHP-Ajax/MySQL. La qualité des annotations sémantiques a été l'objet de plusieurs évaluations (Bertin, 2008, 2011).

3 Réalisations proposées

Recherche d'informations sémantique. Nous avons mis en place une interface de recherche d'informations sémantique, exploitant les points de vue annotés. L'utilisateur peut poser des requêtes selon des catégories sémantiques, associées au termes de filtrage qui l'intéressent. Par exemple, pour retrouver toutes les définitions liées à la "sémantique", il peut utiliser le point de vue définition et le terme "sémantique". Pour une approche multi-document, il est nécessaire de gérer le problèmes d'ordonnancement et de similarité entre les réponses afin de fournir à l'utilisateur une information pertinente et non-redondante tout en proposant un retour au contexte. Cette problématique a été soulevée par Atanassova et al. (2008).

Fiches de synthèse. Les fiches de synthèse représentent des extraits catégorisées en utilisant les annotations sémantiques. Il s'agit d'un produit documentaire permettant de visualiser les différents types d'informations contenues dans un document ou un corpus de documents, sous forme structurée. Les catégories d'extraction sont personnalisables par l'utilisateur. Par exemple, la figure 1 présente une fiche de synthèse utilisant les points de vue *résultat*, *hypothèse*, *méthode*, *soulignement* et *opinion*.

Bibliographie augmentée. La figure 2 montre une bibliographie contenant non seulement les références mais également les catégories sémantiques de l'acte de citation ainsi que l'extrait correspondant. L'utilisateur a ainsi accès aux relations sémantiques entre les auteurs (voir les travaux de Bertin (2011)), à savoir comment ils sont cités : pour leurs résultats, pour une définition, pour leurs méthodes, ... Pour l'utilisateur, la lecture de la bibliographie se révèle éclairante sur la nature de l'article ainsi que sur l'utilisation des citations faites par l'auteur et offrant ainsi une meilleure représentation du document.

Distribution des références bibliographiques. La figure 3 montre la répartition des références bibliographiques en fonction des annotations sémantiques dans deux thèses de doctorat, l'une en philosophie [1], l'autre en informatique [2]. L'axe des abscisses indique comme valeur le numéro du segment textuel, à savoir la phrase. L'axe des ordonnées affiche les différents points de vue de fouille sémantique. La répartition des références est relativement homogène selon les différents points de vue pour la thèse en informatique alors que celle de philosophie montre une prédominance des citations de la part de l'auteur. Cela nous conduit donc à revisiter, à travers l'expérimentation, l'approche et l'utilisation des citations et nous invite à nous interroger sur le rôle de certains indicateurs présents dans les systèmes d'évaluation.

1. Thèse de P. Gauvin, "*Les notions de mouvement et de changement en relation avec les théories contemporaines de l'aspectualité et de la cognition*".

2. Thèse de I. Atanassova, "*Exploitation informatique des annotations sémantiques d'Excom pour la recherche d'informations et la navigation*".

La démonstration soulignera l'importance à proposer des outils facilitant la navigation textuelle ainsi que la capacité du système à resituer un segment dans son contexte initial, c'est-à-dire dans le document source, montrant ainsi une rupture avec les systèmes de recherche d'information traditionnels.

Références

Alrahabi, M. et J.-P. Desclés (2008). Automatic annotation of direct reported speech in Arabic and French, according to a semantic map of ennunciative modalities. In *6th Internatilnal Conference of NLP, GOTAL*, Gothenburg, Sweden.

Atanassova, I., J.-P. Desclés, A. Franchi, et F. Le Priol (2008). La plate-forme excom comme outil automatique d'annotations sémantiques des textes pour la catégorisation des informations sur le web. In *Internet : besoin de communiquer autrement, Université St. Clément d'Ohride*, Sofia, Bulgarie.

Bertin, M. (2008). Categorizations and annotations of citation in research evaluation. In *The 21st international FLAIRS Conference*, Coconut Grove, Floride. AAAI Press.

Bertin, M. (2011). *Bibliosématique : une technique linguistique et informatique par exploration contextuelle*. Ph. D. thesis, Université Paris-Sorbonne.

Desclés, J.-P. (1997). *Systèmes d'exploration contextuelle*. Presses Universitaires de Caen.

Djioua, B., J. G. Flores, A. Blais, J.-P. Desclés, G. Guibert, A. Jackiewicz, F. Le Priol, L. Nait-Baha, et B. Sauzay (2006). Excom : an automatic annotation engine for semantic information. *The 19th international FLAIRS Conference, Melbourne, Floride 1*, 285–290.

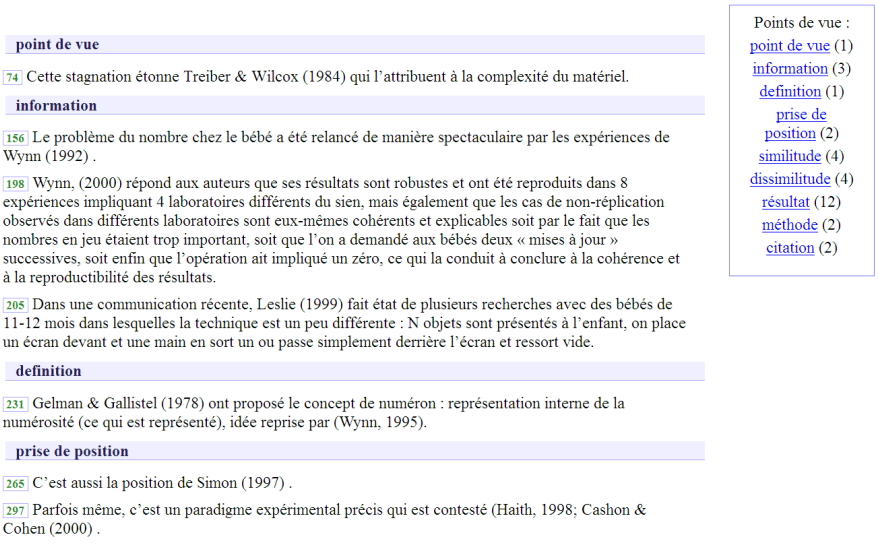

FIG. 1 – *Fiche de synthèse produite par le système*

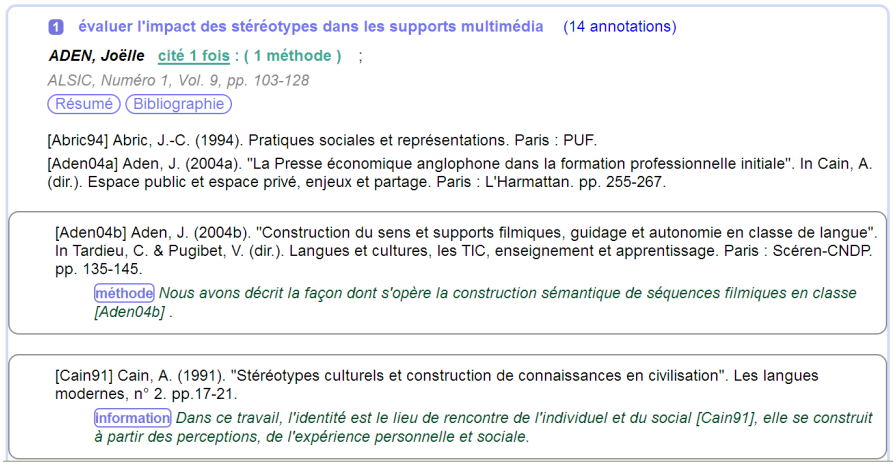

FIG. 2 – *Bibliographie augmentée*

Summary

We propose text exploration tools based on semantic analyses using annotations which are carried out automatically according to linguistic ontologies. By this approach, the final user can access new types of documentary products, that offer a better document representation in an information retrieval context.

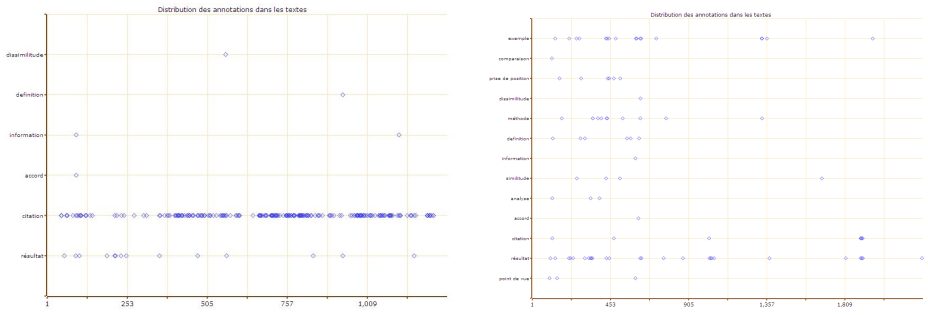

FIG. 3 – *Répartition des références bibliographiques pour une thèse de philosophie (à gauche) et pour une thèse d'informatique (à droite)*