E-Quotes : un outil de navigation textuelle guidée par les annotations sémantiques

Motasem ALRAHABI
Université de Paris-Sorbonne Abou Dabi, ÉAU
motasem.alrahabi@gmail.com

RÉSUMÉ

Nous présentons E-Quotes, un outil de navigation textuelle guidée par les annotations sémantiques. Le système permet de localiser les mots clés et leurs variantes dans les citations sémantiquement catégorisés dans corpus annoté, et de naviguer entre ces citations. Nous avons expérimenté ce système sur un corpus de littérature française automatiquement annoté selon des catégories sémantiques présentes dans le contexte des citations, comme par exemple la définition, l'argumentation, l'opinion, l'ironie ou la rumeur rapportées.

ABSTRACT

E-Quotes : A semantic annotations-driven tool for textual navigation

We present in this paper a semantic annotations-driven tool for textual navigation. The system allows to locate keywords and their variants in semantically categorized passages of an annotated corpus, and navigate between these passages. We tested this system on a French literary corpus automatically annotated according to semantic categories existing in the context of quotations, such as reported Definition, Argumentation, Opinion, Irony or Rumor.

MOTS-CLÉS: Annotation sémantique, citations catégorisées, navigation textuelle, fouille de textes

KEYWORDS: Semantic annotation, categorized quotations, text navigation, text mining

1 Introduction

E-Quotes[1] est un outil d'exploration de corpus et de navigation textuelle guidée par les annotations sémantiques. Le système a été réalisé et testé pour la première fois sur un corpus annoté avec les citations en arabe (Alrahabi, 2015). Dans ce papier, nous présentons de nouvelles fonctionnalités de cet outil spécifiquement mis en place pour le corpus littéraire du labex OBVIL[2]. A l'aide d'excom2, un outil d'annotation à base de règles (Alrahabi, 2010), les citations dans ce corpus[3] ont été automatiquement identifiées et catégorisées en fonction des modalités énonciatives présentes dans leur contexte : *opinion, accord, désaccord, définition, argumentation, assertion, comparaison, ironie, exemplification, observation, rumeur, critique*, etc. Plus de 600 marqueurs linguistiques (*verbes, adjectifs, adverbes*...) ont été manuellement collectés et classés dans une trentaine de catégories sémantiques et discursives. E-Quotes prend en entrée les documents segmentés et annotés par excom2. Ceux-ci sont ensuite indexés à l'aide de la plateforme Apache Solr.

1 http://www.e-quotes.net

2 L'observatoire de la vie littéraire (http://obvil.paris-sorbonne.fr)

3 Le corpus « Critique », il est composé de 300 textes du 19ième siècle (≈ *23 millions de mots*).

2 Présentation du système E-Quotes

L'interface utilisateur d'E-Quotes permet de combiner la technologie classique de recherche d'information (*requêtes de mots clés sur un sac de mots*) et une recherche autour des citations sémantiquement catégorisées. L'utilisateur peut rechercher par exemple un terme dans les citations annotés avec une catégorie bien particulière. Il a aussi la possibilité d'affiner cette recherche en choisissant de localiser le mot clé uniquement à l'intérieur des citations ou bien à l'extérieur de celles-ci. Cette fonctionnalité permet de trouver des réponses à la question :

Comment l'énonciateur présente l'attitude d'un locuteur vis-à-vis d'un propos rapporté ?

où la position de l'énonciateur (*auteur*) et l'attitude du locuteur sont représentées par l'ensemble des catégories sémantiques présentes dans le contexte des citations. E-Quotes donne aussi la possibilité d'intégrer dans les requêtes des listes de termes avec leurs synonymes ou équivalents (ex. *Flaubert, Gustave Flaubert, M. Flaubert...* ou bien *darwinisme, évolution, classification...*). L'application permet enfin d'effectuer des requêtes dans les champs "Titre", "Auteur" ou "Date" des articles du corpus et de combiner plusieurs requêtes ensemble avec les opérateurs ET, OU et NON. Toutes ces fonctionnalités permettent à l'utilisateur de réaliser des requêtes sophistiquées. Exemple : rechercher dans les articles écrits entre 1850 et 1900 les Définitions rapportées, en présence du terme « romanesque » ou ses équivalents à l'intérieur même de la citation. Voici un résultat annoté :

Acceptant cette définition de Madame Necker : « Le roman doit être le monde meilleur », Balzac ajoute : « Mais le roman ne serait rien si, dans cet auguste mensonge, il n'était pas vrai dans les détails. »

Les résultats d'une recherche sont classés par document et renvoient au contexte de la citation dans l'article d'origine. Toutes les citations catégorisées sont surlignées et les termes de la requête sont coloriés. Une fenêtre de navigation dans chaque document offre à l'utilisateur le moyen de parcourir les annotations du document sous une forme de « lecture guidée » où annotations sémantiques et mots clés sont mis en relief pour offrir une lecture optimale et interprétation pertinente. L'utilisateur peut ainsi aller d'une Définition à une Comparaison, d'une Accusation à une Indignation, d'une Opinion positive à une autre négative, etc. D'autres informations sont également fournies dans la fenêtre de navigation comme le nombre d'occurrences des termes de la requête trouvés et les mots les plus fréquents de l'article. L'utilisateur peut exporter et réutiliser les résultats d'une requête sous forme de tableur.

3 Etat de l'art et conclusion

Plusieurs travaux ont abordé la question des citations, mais, à notre connaissance, très peu de recherches ont abouti à des applications opérationnelles pour des utilisateurs finaux. Nous citons la fameuse application du Centre Commun de Recherche européen NewsExplorer (Pouliquen et al., 2007). A partir d'un nom choisi dans une liste, cette application permet, entre autres, de détecter les citations attribuées à cette personne (*locuteur*) ou qui parlent de cette personne. NewsExplorer[4] couvre différentes langues et traite quotidiennement des milliers d'articles journalistiques et dépêches. Notre système, contrairement à NewsExplorer, offre la possibilité d'effectuer des recherches dans des citations sémantiquement catégorisées et de naviguer dans les documents entre ces différentes catégories très fines au niveau du sens. E- Quotes permet en effet de rendre les informations sémantiques accessibles pour un utilisateur final (*littéraires, chercheurs en humanités numériques, journalistes...*). Cet outil montre bien la faisabilité de ce genre de techniques pour la recherche d'informations sémantiques et nous envisageons de le développer sur d'autres catégories et sur de plus grands corpus. Une évaluation est en phase finale de réalisation pour mesurer, dans un premier temps, la précision des annotations obtenues.

[4] http://emm.newsexplorer.eu/NewsExplorer/home/fr/latest.html

Références

ALRAHABI M. (2010). *EXCOM-2: plateforme d'annotation automatique de catégories sémantiques. Applications à la catégorisation des citations en français et en arabe*. Thèse de doctorat, sous la direction du Prof. Jean-Pierre Desclés, Université Paris-Sorbonne.

ALRAHABI M. (2015). E-Quotes: Enunciative Modalities Analysis Tool for Direct Reported Speech in Arabic. Actes de *The 16th International Conference on Intelligent Text Processing and Computational Linguistics (CICLING)*, Cairo, Egypt,

POULIQUEN B., STEINBERGER R., BEST C. (2007). Automatic detection of quotations in multilingual news. Actes de *Recent Advances in Natural Language Processing (RANLP)*, Borovets, Bulgaria