ASSIST : un moteur de recherche spécialisé pour l'analyse des cadres d'expériences

Davy Weissenbacher[1] Elisa Pieri[2] Sophia Ananiadou[1] Brian Rea[1]
Farida Vis[2] Yuwei Lin[2] Rob Procter[2] Peter Halfpenny[2]
(1) National Centre for Text Mining, University of Manchester, 131 Princess Street, M1 7DN UK
(2) National Centre for e-Social Science, University of Manchester, Oxford Rd, Manchester M13 9PL,UK
{prenom.nom}@manchester.ac.uk

Résumé. L'analyse qualitative des données demande au sociologue un important travail de sélection et d'interprétation des documents. Afin de faciliter ce travail, cette communauté c'est dotée d'outils informatique mais leur fonctionnalités sont encore limitées. Le projet ASSIST[1] est une étude exploratoire pour préciser les modules de traitement automatique des langues (TAL) permettant d'assister le sociologue dans son travail d'analyse. Nous présentons le moteur de recherche réalisé et nous justifions le choix des composants de TAL intégrés au prototype.

Abstract. Qualitative data analysis requiers from the sociologist an important work of selection and interpretation of the documents. To facilitate this work, several software have been created but their functionalities are still limitated. The ASSIST project is a preliminary work to define the natural language processing modules for helping the sociologist. We present the search engine realised and justify the NLP modules integrated in the prototype.

Mots-clés : Recherche d'information, Extraction d'information, Terminologie.

Keywords: Information Retrieval, Information Extraction, Terminology.

1 L'analyse qualitative de données assistée par ordinateur

L'analyse des "cadres de l'expérience", ou *Frame Analysis*, est une méthodologie employée par la sociologie. Elle a pour objet l'étude qualitative de la rhétorique employée par les différents acteurs d'un débat. Cette analyse demande un important travail d'interprétation au sociologue. La première tâche consiste à rassembler et à trier les documents[2] pertinents pour le sujet qu'il s'est assigné. Puis, au travers un processus de lectures répétées des documents, le sociologue construit récursivement un ensemble d'annotations sémantiques qu'il associe manuellement aux fragments de textes appropriés. Ces annotations sont de natures hétérogènes et peuvent être regroupées par ensembles cohérents. Une fois interprétées par le sociologue, elles forment les

[1] Voir *http ://www.nactem.ac.uk/assist/* pour une présentation du projet.

[2] Le sociologue construit un corpus selon ses besoins. Le corpus peut être composé des documents écrits, audio et vidéo. Notre étude se limite aux documents écrits uniquement.

cadres de lectures ou *frames*. Par exemple, la frame *surveillance de l'état*, isolée au cours de l'étude du débat causé par l'introduction des cartes d'identité au Royaume-Uni[3] réalisée durant le projet ASSIST (Pieri, 2009), se voit instantiée dans le texte par les annotations *Control, Tracking, Conspiracy, ...*

Afin de faciliter ce travail d'interprétation cette communauté s'est dotée d'outils informatiques regroupés sous l'acronyme Computer Assisted Qualitative Data Analysis (CAQDAS)[4]. Bien que ce soient des outils différents et possédant des caractéristiques propres, ils proposent trois fonctionnalités communes et essentielles. La première est le stockage et l'administration de documents numériques pour une étude donnée. La seconde est la mise à disposition d'une interface conviviale pour l'annotation des documents. La dernière fonctionnalité est fournie par des interfaces complémentaires qui permettent de rechercher, lier et visualiser les annotations.

Si ces outils offrent des interfaces de gestions des documents et des annotations nécessaires au travail d'interprétation du sociologue, ils sont conçus pour un traitement humain des documents. Alors que le sociologue souhaiterait étudier un nombre toujours croissant de documents numériques, la gestion de corpus de tailles importantes (*i.e.* plus de 1000 documents) est difficilement supportée par ces logiciels. De plus, ces logiciels intègrent encore peu de techniques issues du TAL pour automatiser ou assister le sociologue dans son travail d'annotation[5]. Le projet ASSIST est une étude exploratoire afin de caractériser et d'évaluer les modules du TAL pouvant assister le sociologue dans son travail d'analyse. La section suivante décrit le système réalisé.

2 Un moteur de recherche spécialisé pour l'analyse des frames

Pour étudier la polémique soulevée par l'introduction des cartes d'identité au Royaume-Uni, nous avons extrait 4889 articles de journaux nationaux de la base documentaire LexisNexis[6]. Notre corpus couvre la période 2003-2008 et a été construit à partir de la requête *"ID card, Identity card, National Identity Register, National Identity scheme"*. Le système réalisé lors du projet ASSIST est un moteur de recherche intégrant différents modules de TAL (Rea & Ananiadou, 2007). *Lucene*[7] est au cœur de notre système. L'API de ce moteur de recherche offre les opérateurs usuels de requêtage, *i.e.* les opérateurs booléens et les caractères joker. Nous avons défini de nouveaux opérateurs capables de requêter les métadonnées extraites des articles (*i.e.* titres, auteurs, dates de publication et sources) et d'interroger les informations sémantiques ajoutées par les modules de TAL. *Lingo*, un algorithme de clustering non-supervisé[8] , a été ajouté pour classer les documents retournés en réponse à une requête donnée. Ces deux composants sont importants puisqu'ils simplifient le travail initial de filtrage des documents pour toute nouvelle étude. Un module de reconnaissance des Entités Nommées (EN) a été intégré au système en vue de faciliter la lecture du sociologue en révélant les acteurs et les lieux principaux des do-

[3]Un résumé de l'étude est disponible à l'adresse : *http ://www.ncess.ac.uk/research/hub_research/idcards/*

[4]Un inventaire des outils peut être consulté à l'adresse *http ://caqdas.soc.surrey.ac.uk/*

[5]A notre connaissance, le plus complet est l'*add-on Wordstat* adjoint au logiciel *QDA Miner*. Grâce à un anti-dictionnaire de mots outils, une lemmatisation et un clustering des documents, l'*add-on* améliore le calcul des fréquences des mots et facilite la recherche des annotations.

[6]Sous réserve du respect des droits de propriété intellectuelle, cette base est accessible à l'adresse *http ://w3.nexis.com/sources/*

[7]Une documentation de l'API est disponible à l'adresse *http ://lucene.apache.org/java/docs/*

[8]Nous avons utilisé l'implémentation de l'algorithme disponible dans l'API *Carrot*[2], largement documentée à l'adresse http ://project.carrot2.org/.

cuments. Les classes de l'*enamex*[9] sont trop générales pour définir des opérateurs sémantiques précis et utiles. Nous avons donc sélectionné un sous-ensemble de catégories adaptées à notre sujet[10] parmi les catégories d'EN définies par (Sekine & Nobata, 2004). Nous avons ensuite entraîné un module de reconnaissance d'EN probabiliste reposant sur les Champs Conditionnels Aléatoires (Okanohara *et al.*, 2006) pour annoter et indexer les EN de notre corpus. Un module d'extraction de termes, *TerMine*(Frantzi *et al.*, 2000), vient compléter le précédant module en calculant les concepts les plus saillants dans notre corpus (*ex.* "illegal immigration" ou "DNA database"). Les termes traduisent les thèmes des documents et sont, de fait, appropriés pour la classification et la recherche de documents similaires. Le module d'analyse des sentiments, nommé HYSEAS (Piao *et al.*, 2009), est employé pour estimer automatiquement l'opinion de l'auteur regardant un fait ou un événement narré dans son article. Le module applique un lexique subjectif pour reconnaître la tonalité des mots présents dans une phrase (*ex. intrude* : fortement négatif, *achieve* : faiblement positif) et utilise un scoreur pour attribuer un score général à la phrase. Ce module souligne les passages réthoriques d'un document et, dans le contexte de cette étude, devrait faciliter l'étude des arguments avancés par les opposants et les partisans de la carte d'identité.

3 Conclusion

Dans cet article nous signalons le besoin de logiciels adaptés à l'analyse qualitative de documents textuels. Le projet ASSIST est une première réponse. Travaillant sur une application concrète, l'analyse du débat autour de l'introduction des cartes d'identité au Royaume-Uni, nous avons sélectionné puis intégré à un moteur de recherche un ensemble de modules de TAL pour faciliter l'accès au contenu des documents. Le système est actuellement en cours d'évaluation.

Références

FRANTZI K., ANANIADOU S. & MIMA H. (2000). Automatic recognition of multi-word terms. *International Journal of Digital Libraries*, **3(2)**, pp.117–132.

OKANOHARA D., MIYAO Y., TSURUOKA Y. & TSUJII J. (2006). Improving the scalability of semi-markov conditional random fields for named entity recognition. In *Proceedings of COLING/ACL*.

PIAO S., TSURUOKA Y. & ANANIADOU S. (2009). *HYSEAS : A HYbrid SEntiment Analysis System*. Rapport interne, National Center for Text Mining (NaCTeM).

PIERI E. (2009). The introduction of id cards in the uk : A snapshot of the debate in the press. In *Panel on Media, Communication and Cultural Studies Association Conference (MeCCSA)*.

REA B. & ANANIADOU S. (2007). Text mining services to support e-research. In *UK e-Science All Hands Meeting*.

SEKINE S. & NOBATA C. (2004). Definition, dictionaries and tagger for extended named entity hierarchy. In *Conference on Language Resources and Evaluation*.

[9] *e.i.* les noms de personnes, d'organisations, de lieux et les références numériques.

[10] Par exemple la classes des célébrités comme *Tony Blair* ou celle des personnages de fiction comme *Big Brother* qui occupent un rôle important dans notre corpus.