Néonaute, Enrichissement sémantique pour la recherche d'information

Emmanuel Cartier[1], Loïc Galand[1], Peter Stirling[2] et Sara Aubry[2]
(1) Université Paris 13 SPC, LIPN-RCLN, UMR 7030 CNRS, Labex EFL, 99 avenue Jean-Baptiste Clément, 93430 Villetaneuse
(2) Bibliothèque nationale de France, Départements du Dépôt légal et des Systèmes d'information, Quai François-Mauriac, 75706 Paris Cedex 13
emmanuel.cartier@lipn.univ-paris13.fr,
loic.galand@lipn.univ-paris13.fr,peter.stirling@bnf.fr,sara.aubry@bnf.fr

Avec l'explosion du nombre de documents numériques accessibles, les besoins en outils pour l'enrichissement sémantique des données textuelles, ainsi que des fonctionnalités avancées de recherche et d'exploration des collections, se font sentir. Cette combinaison entre les domaines de la recherche d'information et du traitement automatique des langues est l'une des caractéristiques du projet Néonaute.

Ce projet, financé par la DGLFLF [1] en 2017 (appel Langues et numérique), regroupe la Bibliothèque nationale de France (BnF), le LIPN - RCLN (CNRS UMR 7030) et l'Université de Strasbourg (LILPA, EA 1339). Son objectif principal est de doter les observateurs de la langue française d'un moteur de recherche s'appuyant sur une collection de sites de presse d'actualité, collectés automatiquement par la BnF au titre de sa mission de dépôt légal de l'internet. Sur cette collection, le projet vise à proposer un moteur de recherche de nouvelle génération, disposant d'une indexation enrichie par l'analyse automatique des textes (analyse morphosyntaxique, entités nommées, thématiques), d'une part, et d'outils de recherche, d'exploration et de visualisation multidimensionnelle interactive des résultats, d'autre part.

Enrichissement des métadonnées par le TAL L'objectif premier du projet est d'enrichir les informations indexées dans le moteur de recherche. La BnF constitue depuis 1996 une archive du web français, qui représente fin 2017 plus de 30 milliards de fichiers et 938 To de données. Depuis décembre 2010, une centaine de sites d'actualités (presse nationale, presse régionale, portails d'information) sont collectés quotidiennement (page d'accueil et liens internes à un clic), constituant la collection dite "Actualités" qui représente 1 milliards de fichiers et 13 To de données. Cette collection est accessible dans les salles de lecture de la BnF via une interface de recherche plein texte (Archives de l'internet Labs) construite à partir du moteur Apache Solr. Le projet Néonaute effectue une analyse *linguistique* automatique des contenus textuels, comme suit :

1. filtrage des contenus collectés pour ne conserver qu'un corpus de pages à contenu textuel : pour ce faire, un certain nombre de filtres ont été développés, aboutissant à ne retenir qu'environ 10% de l'archive totale ;
2. nettoyage des pages filtrées pour ne conserver que le contenu textuel "nouveau" : pour ce faire, un état de l'art et une évaluation nous ont permis de choisir puis d'utiliser la librairie

1. délégation générale à la langue française et aux langues de France

Python *JusText* [2] (Pomikálek, 2011) ;

3. analyse morphosyntaxique du contenu textuel : pour ce faire, après un état de l'art, l'établissement de critères de choix puis une évaluation des précision et rappel sur un échantillon d'une dizaine de pages web, l'outil *Spacy* [3] (Choi *et al.*, 2015) a été retenu ;
4. détection automatique des entités nommées (personnes, lieux, organisations, autres) : la même procédure de sélection a abouti à choisir l'outil *Spacy* ;
5. détection automatique des thématiques de chaque page web : cet aspect est actuellement en cours d'évaluation, avec des techniques d'extraction de mots clés à l'aide d'une pondération des mots avec le modèle TF-IDF, ainsi que des techniques de topic modelling (Allocation de Dirichlet Latente).

Nous présenterons chacune des étapes effectuées, les résultats, leur évaluation et les problèmes non résolus à ce stade du projet.

Exploitation de l'indexation enrichie L'enrichissement *linguistique* des données permet de nouvelles exploitations dans le cadre des moteurs de recherche. En effet, les analyses linguistiques ajoutent minimalement des métainformations liées aux lexies discriminantes, aux entités nommées et aux thèmes principaux de chaque article. Ces enrichissements peuvent être exploités sous forme de *facettes* dans le cadre d'une recherche simple, soit sous forme de *visualisation multidimensionnelle interactive* (Cartier, 2017), pour explorer les données selon différentes points de vue. Par exemple, il sera possible de visualiser l'évolution temporelle des emplois de *twitterisation* sur toute la période, en tenant compte des métadonnées liées à chaque source d'informations (journal, type de presse, auteurs, etc.).

Cas d'usage En plus de l'objectif principal, qui aboutira à la mise à disposition d'un moteur de recherche de nouvelle génération, trois cas d'utilisation sont prévus, dont les premiers résultats seront présentés :

1. étude de l'implantation des néologismes sur une période temporelle de 8 ans (2010-2017) ;
2. étude de l'implantation des préconisations de termes de la DGLFLF ;
3. étude des formes de la féminisation des noms communs dans ce même corpus.

Durant la démonstration, nous présenterons les différentes phases d'analyse linguistique des textes, leur indexation et leur exploitation dans le moteur de recherche, en nous appuyant sur les cas d'utilisation.

Références

CARTIER E. (2017). Neoveille, a Web Platform for Neologism Tracking. In *Proceedings of European Chapter of the Association for Computational Linguistics 2017, Valencia, 3-7 avril 2017*.

CHOI J. D., TETREAULT J. & STENT A. (2015). It depends : Dependency parser comparison using a web-based evaluation tool. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, volume 1, p. 387–396.

2. https://github.com/miso-belica/jusText
3. https://github.com/explosion/spaCy

POMIKÁLEK J. (2011). *Removing boilerplate and duplicate content from web corpora*. PhD thesis, Masaryk university, Faculty of informatics, Brno, Czech Republic.