Un outil multilingue d'extraction de collocations en ligne

Luka Nerima[1], Violeta Seretan[2], Eric Wehrli[1]
(1) Laboratoire d'analyse et de technologie du langage (LATL), CUI
(2) Département de traitement informatique multilingue (TIM), FTI
Université de Genève
{luka.nerima, violeta.seretan, eric.wehrli}@unige.ch

RÉSUMÉ

Cette démonstration présente la version web d'un outil multilingue d'extraction de collocations. Elle est destinée aux lexicographes, aux traducteurs, aux enseignants et apprenants L2 et, plus généralement, aux linguistes désireux d'analyser et d'exploiter leurs propres corpus.

ABSTRACT

This demo shows the web version of a multilingual collocation extraction tool. It is intended for lexicographers, translators, L2 teachers and learners and, more generally, for linguists who want to analyze and to exploit their own corpora.

MOTS-CLÉS : Extraction de collocations, analyse linguistique, multilingue, application web.

KEYWORDS : Collocation extraction, linguistic parsing, multilingual, web application.

1 Introduction

Les ressources lexicales jouent un rôle clé dans les applications de TAL. En particulier, les dictionnaires d'expressions à mots multiples sont très précieux pour l'analyse syntaxique et la traduction automatique. Par contre, leur constitution est un travail difficile et de longue haleine. C'est ce qui nous a poussés a développé FipsCo, un outil d'acquisition de collocations à partir de corpus (Seretan, 2011), que avons largement utilisé comme assistance pour peupler nos lexiques, en validant toutefois manuellement chaque insertion. Les linguistes non-informaticiens ont également montré un vif intérêt pour cet outil. C'est à leur usage que nous en avons développé une version web, facilement accessible et ne nécessitant aucune installation sur le poste de l'usager.

2 La méthode

FipsCo est un extracteur de collocations basé sur un analyseur syntaxique profond, Fips (Wehrli, Nerima, 2015). Brièvement, la méthode d'extraction est la suivante : après l'analyse d'une phrase du corpus, l'arbre syntaxique produit est parcouru et toutes les paires de mots qui sont dans une des configurations syntaxiques données[1] sont considérées comme des collocations potentielles et collectées dans une liste de candidats. A la fin de l'analyse du corpus, on applique une mesure d'associativité sur les termes des candidats. Les collocations candidates sont alors affichées dans

[1] Nom-Adj, Adj-Nom, Nom-Nom, Nom-Prep-Nom, Sujet-Verbe, Verbe-Objet, Verbe-Adverbe, etc.

l'ordre décroissant de leur score. Plusieurs mesures d'associativité ont été implémentées dans FipsCo, la mesure par défaut étant celle du log de la fonction de vraisemblance ou *log-likelihood* (voir p.ex. Dunning, 1993).

3 La version Web

La version Web de FipsCo reprend la technologie de l'extracteur de base. L'interface personne machine a été simplifiée au maximum: l'usager choisit le corpus de l'extraction (un fichier local) ainsi que les paramètres de la langue[2], de la mesure d'associativité, du score et du nombre d'occurrences minimum pour les collocations résultat. Comme l'extraction peut prendre plusieurs dizaines de minutes, l'usager peut entrer son adresse électronique, lancer l'extraction et fermer son navigateur. Une notification lui sera envoyée par courriel dès que l'extraction est terminée. La notification contient une adresse web lui permettant d'accéder directement aux résultats de son extraction. Une fois les résultats affichés, l'usager peut interactivement restreindre l'affichage des collocations en choisissant le type de collocation souhaité ou encore cliquer sur une collocation pour afficher les contextes d'occurrence (phrases).

La Figure 1 montre l'écran de lancement de l'extraction et le résultat pour les collocations de type Verbe – Objet ayant obtenu les meilleurs scores. L'extraction a été effectuée sur un corpus du journal Le Monde d'environ 250K mots.

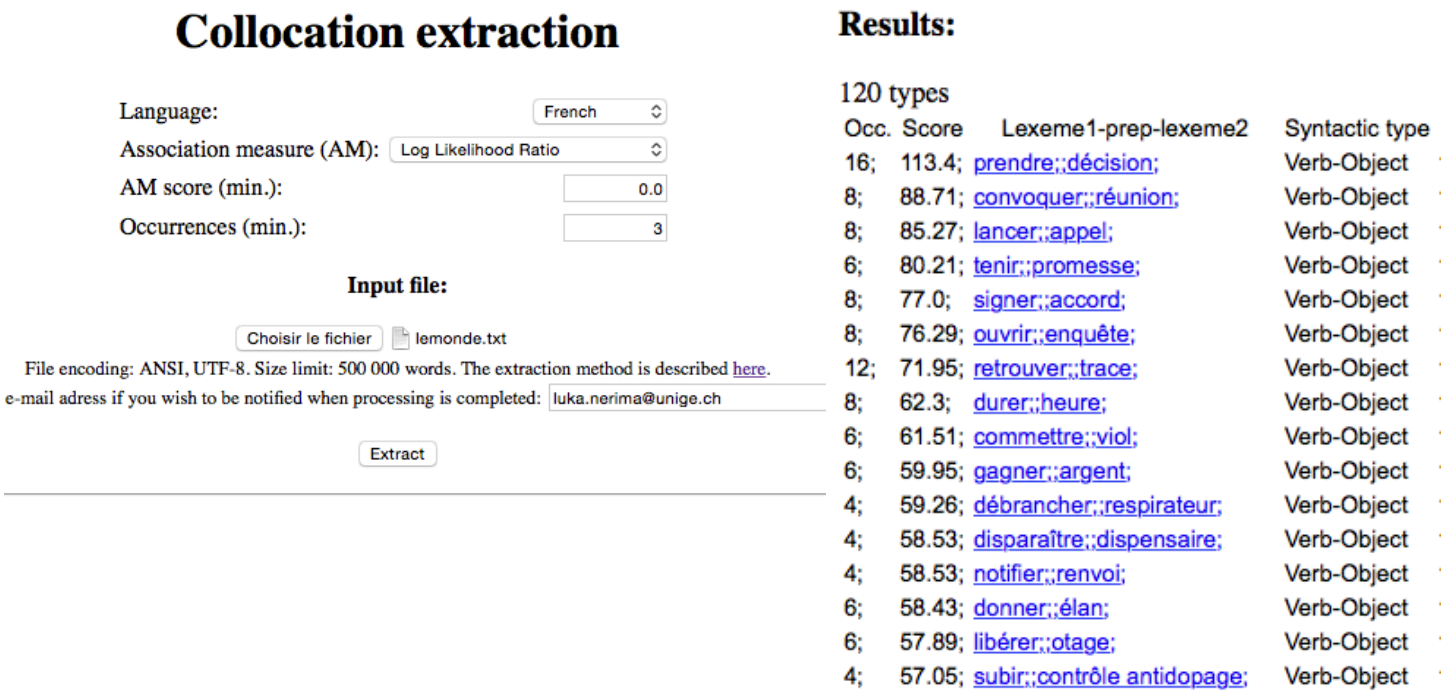

FIGURE 1 : L'écran de lancement de l'extraction et les meilleurs scores pour le type Verbe - Objet

Les outils d'extraction et d'interrogation à partir de corpus sont relativement nombreux, p.ex. ScienQuest (Falaise et al., 2011), AnaText, MWEtoolkit (Ramisch, 2015), fivefilters.org, SketchEngine et CQPweb, bien que cette dernière n'offre pas d'interface conviviale pour linguistes non-informaticiens (Evert & Hardie, 2011) ; mais à notre connaissance aucun d'entre eux n'utilise une analyse syntaxique profonde, ce qui entraîne une diminution de la précision et du rappel.

[2] Français, anglais, allemand, italien, espagnol, portugais ou grec.

Références

DUNNING T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.

EVERT S., HARDIE A. (2011). Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium. In Proceedings of the Corpus Linguistics 2011 Conference, Birmingham, UK.

FALAISE A., TUTTIN A., KRAIF, O. (2011). Une interface pour l'exploitation de corpus arborés par des non informaticiens : la plate-forme ScienQuest du Projet Scientext". TAL, 52(3):241–246.

SERETAN V. (2011). Syntax-Based Collocation Extraction, Springer.

RAMISCH C. (2015). Multiword Expressions Acquisition: A Generic and Open Framework, *Theory and Applications of Natural Language Processing* series XIV, Springer.

WEHRLI E., NERIMA L. (2015). The Fips Multilingual Parser, in Text, Speech, Language Tech., Vol. 48, Gala, Núria, Rapp, Reinhard and Bel-Enguix, Gemma (Eds.): Language Production, Cognition, and the Lexicon. Springer, pp. 473 – 490

Adresses URL des outils cités

ANATEXT : http://olivier.kraif.u-grenoble3.fr/anaText/ consulté le 18.05.2016

CQPWEB : https://cqpweb.lancs.ac.uk/ consulté le 24.05.2016

FIPSCO : http://latlapps.unige.ch/, consulté le 27.05.2016

FIVEFILTERS : http://fivefilters.org/term-extraction/, consulté le 24.05.2016

MWETOOLKIT : http://mwetoolkit.sf.net/, consulté le 24.05.2016

SKETCHENGINE : https://www.sketchengine.co.uk, consulté le 24.05.2016