Identification d'articles parallèles dans Wikipedia

Florian Boudin

Recherche d'information cross-lingue - 2015

1 Introduction

Wikipedia est un projet d'encyclopédie universelle et multilingue (291 langues mi-2015). Lorsqu'un article n'est pas disponible dans la langue souhaitée ou qu'il est incomplet, l'utilisateur peut décider de rechercher l'information dans une autre langue à l'aide des liens inter-langue de Wikipedia. Ces liens sont créés manuellement par les utilisateurs. Par conséquent, de nombreux liens sont manquants et leur mise à jour est une tâche extrêmement chronophage.

2 Travail demandé

Votre tâche consiste à développer un système permettant l'identification automatique des liens inter-langue dans Wikipedia. Pour cela, vous disposez de trois ensembles de documents extraits de Wikipedia en trois langues (français, anglais et allemand, `http://filex.univ-nantes.fr/get?k=Sqp6QNB4eJhD72SF4qz`). Ces données ont été utilisées dans le cadre de la *shared task* de BUCC 2015 [1].

Dans un premier temps, vous devez implémenter la méthode basée sur les hapax présentée dans [2] (lien direct : `http://www.aclweb.org/anthology/N07-2008.pdf`). Bien que simple, cette méthode donne de bons résultats (30%+ de MAP). Pour évaluer la performance de votre système, vous utiliserez le logiciel `trec_eval` (`http://trec.nist.gov/trec_eval/`) et les fichiers de référence disponibles pour le cours (e.g. `fr-en-train.qrels`).

Dans un second temps, réfléchissez aux moyens d'améliorer cette méthode. Proposez et implémentez vos idées d'amélioration dont vous vérifierez ensuite l'impact sur les données fournies.

Références

[1] Serge Sharoff, Pierre Zweigenbaum and Reinhard Rapp, *BUCC Shared Task : Cross-Language Document Similarity*. Proceedings of the Eighth Workshop on Building and Using Comparable Corpora (BUCC), 2015.

[2] Jessica Enright and Grzegorz Kondrak, *A Fast Method for Parallel Document Identification*. Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics ; Companion Volume, Short Papers, pages 29–32, 2007.