

Élimination des liens inter-langues erronés dans Wikipédia

Nacéra Bennacer Seghouani, Francesca Bugiotti
Jorge Galicia, Mariana Patricio
Gianluca Quercini

LRI, CentraleSupélec, Univ. Paris-Saclay, Gif-sur-Yvette, 91190, France
{nacera.bennacer, francesca.bugiotti}@lri.fr
{jorge.galicia-ayon, mariana.patricio}@student.ecp.fr
gianluca.quercini@lri.fr

Résumé. Un lien inter-langue dans Wikipédia est un lien qui mène d'un article appartenant à une édition linguistique à un autre article décrivant le même concept dans une autre langue. Ces liens sont ajoutés manuellement par les utilisateurs de Wikipédia et ainsi ils sont susceptibles d'être erronés. Dans ce papier, nous proposons une approche pour l'élimination automatique des liens inter-langues. Le principe de base est que la présence d'un lien erroné est révélée par l'existence d'un chemin de liens inter-langues reliant deux articles appartenant à une même édition linguistique. Notre approche élimine des liens inter-langues, à partir de ceux qui ont un faible *score de correction*, jusqu'à ce qu'il n'y ait plus de chemins entre deux articles d'une même édition linguistique. Les résultats de notre évaluation sur un sous-graphe de Wikipédia consistant en 8 langues montre que l'approche est prometteuse.

1 Introduction

Les liens inter-langues (LILs) de Wikipédia permettent de naviguer facilement entre ses différentes éditions linguistiques et sont également exploités dans des applications de recherche d'information multilingue [de Melo et Weikum (2010); Sorg et Cimiano (2012)]. Cependant, les LILs sont ajoutés manuellement par les utilisateurs de Wikipédia et ainsi ils sont susceptibles d'être erronés (c.a.d. ils relient des articles qui ne décrivent pas un même concept).

Dans ce papier, nous proposons une approche pour l'élimination automatique des LILs. Les travaux existants s'attaquent principalement au problème de la détection des LILs manquants [Bennacer et al. (2015); Moreira et Moreira; Penta et al. (2012); Sorg et Cimiano (2008)]. Comme de Melo et Weikum l'ont fait remarquer, l'existence d'un chemin de LILs entre deux articles appartenant à une même édition linguistique (et, donc, décrivant deux concepts différents) révèle la présence d'un LIL erroné [De Melo et Weikum (2010)]. Autrement dit, si deux articles provenant d'une même édition linguistique appartiennent à une même composante connexe, au moins un LIL de cette composante est erroné et la composante est dite *incohérente*.

Notre approche attribue un *score de correction* aux liens d'une composante incohérente et élimine des liens de façon itérative, à partir de ceux qui ont un score faible (c.a.d., susceptibles