Approche d'extraction de classes interlangues à partir de documents multilingues à base de Concepts Fermés

Mohamed Chebel *, Chiraz Latiri **

* mohammedchebel@gmail.com
** chiraz.latiri@gnet.tn
Laboratoire de recherche LIPAH, Faculté des Sciences de Tunis
Campus Universitaire Tunis El Manar, 1060 Tunis, Tunisie.

1 Introduction

Dans cet article, nous proposons une nouvelle méthode de classification non supervisée de documents multilingues de corpus comparable bruité afin d'améliorer l'extraction des lexiques de traduction. Nous nous basons sur l'approche de (Rouane et al., 2007) dans la réingénierie des modèles UML et (Mimouni et al., 2012) dans la RI qui ont profité d'un couplage entre l'aspect formel et le relationnel afin de prendre en compte des relations entre les objets d'un même contexte. Nous avons choisi d'effectuer un couplage entre l'Analyse Formelle de Concepts (AFC) et les modèles vectoriels. En effet, l'AFC, appliquée dans un contexte de fouille de textes, permet d'extraire des classes de documents sous formes de CFs. D'un autre côté, les modèles vectoriels basés sur les vecteurs des extensions des CFs extraits, permettent d'aligner les CFs des différentes langues en calculant le degré de similarité des Concepts Fermés monolingues extraits, dans l'objectif de générer des CFs multilingues.

2 Extraction de Concepts Fermés à partir de corpus comparables

En classification de documents, un Concept Fermé est le couple $< T, D >$, avec T l'ensemble des termes des documents qui appartiennent à tous les documents D, et D, l'ensemble des documents qui contiennent tous les termes de T. Dans notre contexte de recherche, un Concept Fermé représente une classe de documents regroupés selon un ensemble de termes représentatifs. L'extraction des Concepts Fermés, à partir d'un corpus comparable français-anglais, est précédée par une étape de pré-traitement linguistique du corpus comparable bilingue mais aussi une réorganisation du contenu des documents du corpus en question est nécessaire. Les concepts en sortie sont de la forme : $CF =< \{t_1, t_2, \ldots, t_n\}, \{d_1, d_2, \ldots, d_m\} >$ tel que $\{t_1, t_2, \ldots, t_n\}$ (ou extension) est l'ensemble des termes qui composent un termset fermé et $\{d_1, d_2, \ldots, d_m\}$ (ou intension) l'ensemble de documents dans lesquels $\{t_1, t_2, \ldots, t_n\}$ sont apparus ensemble avec une fréquence supérieure ou égale à $minsupp$. La sortie est composée de l'ensemble des Concepts Fermés français CF_{fr} et des Concepts Fermés anglais CF_{en} séparément.