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.

3 Déploiement des Concepts Fermés pour la CDM

Nous proposons d'étudier l'apport des Concepts Fermés dans le domaine de la classification de documents multilingues, apprécié en terme de comparabilité. Pour cela, nous proposons de : **(*i*) Traduire les termes des extensions des CFs** ; consiste à enrichir chacun des termes de l'extension d'un CF dans sa langue source L_s avec des termes dans la langue cible L_c. En effet, les extensions sont traduites par un SYSTÈME DE MT EN LIGNE [1] du français vers l'anglais et inversement, puis enrichies par toutes leurs traductions potentielles possibles. Le but de cette étape est donc de permettre de faire la correspondance des termes t_i de l'extension d'un CF de la langue source L_s dans l'intension d'un CF de la langue cible L_c afin de permettre de construire les vecteurs d'espace vectoriel des CFs de chaque langue. **(*ii*) Aligner les CFs monolingues français et anglais extraits** ; consiste à rassembler les CFs les plus similaires extraits de chaque langue afin de générer des Concepts Fermés bilingues CF_{fr-en} ou CF_{en-fr} (*i.e.*, classes bilingues) et cela en utilisant un modèle d'espace vectoriel basé sur la mesure de similarité vectorielle Cosinus. Le but de cet alignement est de construire des classes de documents bilingues dont la comparabilité est meilleure comparée à celle du corpus initial. Le vecteur de chaque concept est composé des termes représentatifs des extensions des CFs.

Remerciements.

Ce travail est partiellement financé par le projet franco-tunisien DGRST-CNRS n° 14/R 1401, intitulé "Fouille de textes pour la construction de Lexiques interlangues et la Recherche d'Information Multilingue".

Références

Mimouni, N., A. Nazarenko, et S. Salotti (2012). Classification conceptuelle d'une collection documentaire. intertextualité et recherche d'information. *CORIA 2012, 9th French Information Retrieval Conference, Bordeaux : France*.

Rouane, M. H., M. Huchard, A. Napoli, et P. Valtchev (2007). A proposal for combining formal concept analysis and description logics for mining relational data. In *Proceedings of the 5th international conference on Formal concept analysis, ICFCA 2007, LNAI*, pp. 51–65.

Summary

In this article, we highlight the interest and usefulness of Formal Concept Analysis (FCA) in multilingual document clustering. We propose a statistical approach for clustering multilingual documents based on Closed Concepts and vector model partition the documents of one or more collections.An experimental evaluation was conducted on the collection of bilingual documents French-English of CLEF' [2] 2003 and showed the merits of this method and the interesting degree of comparability of the obtained bilingual classes.

1. https ://translate.google.com/
2. Cross-Language Evaluation Forum | http://www.clef-initiative.eu.