Bilan du Premier Défi Francophone de Fouille de Textes

Jérôme Azé*, Mathieu Roche**,
Érick Alphonse***, Ahmed Amrani*,****
Thomas Heitz*, Amar-Djalil Mezaour*****

* LRI – Université Paris-Sud
Bât. 490, 91405 Orsay Cedex
{aze,amrani,heitz}@lri.fr
**LIRMM – Université de Montpellier 2
161 rue Ada, 34392 Montpellier Cedex 5
mroche@lirmm.fr
***LIPN – Université de Paris Nord
99, avenue Jean-Baptiste Clément, 93430 Villetaneuse
Erick.Alphonse@lipn.univ-paris13.fr
****ESIEA Recherche
9 rue Vésale, 75005 Paris
amrani@esiea.fr
*****Exalead
10 place de la Madeleine, 75019 Paris
Amar-Djalil.Mezaour@exalead.com

Résumé. Le **DÉ**fi **F**ouille de **T**extes (DEFT) a consisté à supprimer les phrases non pertinentes dans un corpus de discours politiques en français. Il a eu lieu en 2005 et réuni onze équipes, totalisant une trentaine de participants. Cet article décrit les prétraitements effectués sur les corpus de F. Mitterrand et de J. Chirac dans le cadre de ce défi. Notamment, la conversion au format texte, le découpage en phrases, le classement des discours, l'introduction de phrases de F. Mitterrand dans les discours de J. Chirac et l'identification des dates et noms de personnes. Les résultats obtenus par les onze équipes participantes sont aussi présentés.

1 Introduction

Le but du défi proposé consiste à supprimer les phrases non pertinentes dans un corpus de discours politiques en français. Ce défi porte le nom de DEFT pour **DÉ**fi **F**ouille de **T**extes. Ce défi, proche de la tâche *Novelty* du challenge TREC[1] [Soboroff, Harman, 2003, Amrani *et al.*, 2004], est motivé par le besoin de mettre en place des techniques de fouille de textes permettant soit d'identifier des phrases non pertinentes dans des textes, soit d'identifier des phrases particulièrement singulières dans des textes apparemment sans réel intérêt. Cette étape est préliminaire à tout processus d'extraction d'informations.

[1] Text REtrieval Conferences : http ://trec.nist.gov