JEP-TALN-RECITAL 2012

JEP : Journées d'Études sur la Parole
TALN : Traitement Automatique des Langues Naturelles
RECITAL : Rencontre des Étudiants Chercheurs en Informatique
pour le Traitement Automatique des Langues

Actes de la conférence conjointe JEP-TALN-RECITAL 2012

Atelier DEFT 2012: DÉfi Fouille de Textes

Éditeurs
Cyril Grouin
Dominic Forest
Gilles Sérasset

4 – 8 Juin 2012
Grenoble, France

Des versions imprimées de ces actes peuvent être achetées auprès de :

GETALP-LIG
Laurent Besacier
BP 53
38041 Grenoble Cedex 9
France
Laurent.Besacier@imag.fr

Préface

Créé en 2005 à l'image des campagnes d'évaluation internationales TREC *(Text Retrieval Conference)*, le défi fouille de texte (DEFT) propose chaque année une campagne d'évaluation francophone en fouille de texte, sur des thématiques et des corpus régulièrement renouvelés.

Plusieurs champs d'activité ont ainsi été abordés au cours des différentes éditions : les ruptures de style en 2005, la segmentation thématique en 2006, la fouille d'opinion en 2007 puis de nouveau en 2009, la classification en genres et thèmes en 2008, et la variation diachronique en 2010 et 2011.

Les campagnes ont porté sur des corpus de discours politiques (de 2005 à 2007), des corpus de critiques grand public sur des livres, des films et des jeux vidéo (en 2007), et des corpus de journaux contemporains (en 2008 et 2009) ou anciens (en 2010 et 2011).

A partir de 2011, un nouveau type de documents a été utilisé dans le défi : les articles scientifiques qui ont paru en revues dans le domaine des Sciences Humaines et Sociales. Plusieurs applications ont ainsi été envisagées sur la base de ce nouveau corpus. En 2011, une tâche d'appariement entre un article scientifique et le résumé qui lui correspond a été proposée, dans une perspective d'identification des éléments saillants d'un article à mettre en avant dans le résumé ; cette édition a permis aux participants d'obtenir d'excellents résultats. Pour cette nouvelle édition, nous proposons de travailler sur l'indexation des articles scientifiques dans une tentative d'identification des mots-clés choisis par les auteurs pour indexer leur article.

Cyril Grouin et Dominic Forest, *co-présidents du Comité de Programme*

Comités

Comité de programme

Daille, Béatrice (LINA, Nantes)

Forest, Dominic (EBSI, Université de Montréal), *co-président*

Grouin, Cyril (LIMSI–CNRS, Orsay), *co-président*

Paroubek, Patrick (LIMSI–CNRS, Orsay)

Torres-Moreno, Juan Manuel (LIA, Avignon)

Zweigenbaum, Pierre (LIMSI–CNRS, Orsay)

Comité d'organisation

Forest, Dominic (EBSI, Université de Montréal)

Grouin, Cyril (LIMSI–CNRS, Orsay)

Paroubek, Patrick (LIMSI–CNRS, Orsay)

Ponton, Claude (Lidilem, Université Stendhal Grenoble 3)

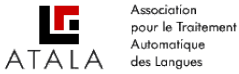
Zampa, Virginie (Lidilem, Université Stendhal Grenoble 3)

Zweigenbaum, Pierre (LIMSI–CNRS, Orsay)

Sponsors

ATALA

Association pour le Traitement Automatique des Langues.



Projet DoXa

Financement Cap Digital.

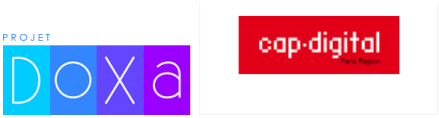


Table des matières

Présentation et résultats

Indexation libre et contrôlée d'articles scientifiques. Présentation et résultats du défi fouille de textes DEFT2012
Patrick Paroubek, Pierre Zweigenbaum, Dominic Forest et Cyril Grouin 1

Méthodes des participants

Key-concept extraction from French articles with KX
Sara Tonelli, Elena Cabrio et Emanuele Pianta 15

Acquisition terminologique pour identifier les mots-clés d'articles scientifiques
Thierry Hamon 25

Indexation à base des syntagmes nominaux
Amine Amri, Maroua Mbarek, Chedi Bechikh, Chiraz Latiri et Hatem Haddad 33

Détection de mots-clés par approches au grain caractère et au grain mot
Gaëlle Doualan, Mathieu Boucher, Romain Brixtel, Gaël Lejeune et Gaël Dias 41

Participation de l'IRISA à DeFT2012 : recherche d'information et apprentissage pour la génération de mots-clés
Vincent Claveau et Christian Raymond 49

Participation du LINA à DEFT2012
Florian Boudin, Amir Hazem, Nicolas Hernandez et Prajol Shrestha 61

Algorithme automatique non supervisé pour le Deft 2012
Murat Ahat, Coralie Petermann, Yann Vigile Hoareau, Soufian Ben Amor et Marc Bui 69

Enrichir et raisonner sur des espaces sémantiques pour l'attribution de mots-clés
Adil El Ghali, Daniel Hromada et Kaoutar El Ghali 77