Atelier des conférences TALN'05/RECITAL'05

DEFT'05 (DÉfi Fouille de Textes)

Page Web : http://www.lri.fr/ia/fdt/DEFT05/

10 juin 2005, 14h-17h30, Dourdan (91)

1 Motivations

Le défi proposé 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.

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 (**DÉ**fi **F**ouille de **T**extes).

Dans les corpus spécialisés (biologie, médecine, etc.) un travail conséquent est dédié à l'identification des phrases pertinentes pour ensuite y rechercher des informations spécifiques. Ce type de tâche consistant à effectuer un premier filtrage des textes est une étape préliminaire essentielle à effectuer pour la constitution de corpus pertinents et homogènes. Le défi DEFT'05 est relatif à une telle tâche.

L'étape suivante peut consister à rechercher des informations précises dans ces textes filtrés. Le défi proposé ne s'intéresse pas à ce travail qui fait l'objet d'autres défis telle que la tâche *questions/réponses* du défi international TREC[1].

Le défi que nous proposons est plus proche de la tâche *Novelty* du challenge TREC. La première partie de la tâche *Novelty* de TREC consiste à identifier les phrases pertinentes

[1] http ://trec.nist.gov

puis, parmi celles-ci, les phrases nouvelles d'un corpus d'articles journalistiques. DEFT'05 qui consiste à supprimer les phrases non pertinentes d'un corpus de discours politiques est assez proche du travail d'identification des phrases pertinentes de la tâche *Novelty* du challenge TREC.

Nous proposons ici une liste non exhaustive de tâches similaires à celle proposée dans DEFT'05 et pour lesquelles il devrait être possible de réutiliser avec peu de modifications les approches mises en œuvre pour répondre à DEFT'05.

- détection des passages les plus singuliers dans des textes quelconques (rupture de style, changement de contexte) ;
- détection de plagiats possibles dans des textes ;
- détection des informations générales dans des corpus techniques.

2 Tâches à réaliser pour DEFT'05

Un corpus de textes, issus de la Présidence de Jacques Chirac (1995-2005), est fourni aux participants. Ce corpus est composé d'allocutions officielles du Président. Dans ce corpus, des passages issus d'un corpus d'allocutions du Président de la République François Mitterrand (1981-1995) sont insérés. Les passages d'allocutions de F. Mitterrand insérés sont composés d'au moins deux phrases successives.

Les passages de F. Mitterrand introduits traitent d'une thématique différente. Par exemple, dans les allocutions de J. Chirac évoquant la politique internationale, les phrases de F. Mitterrand introduites sont issues de discours traitant de politique nationale. Ainsi, la rupture thématique peut être une des manières de détecter les phrases issues du corpus de F. Mitterrand.

Un corpus avec des passages extraits d'allocutions de F. Mitterrand introduits dans les textes de J. Chirac est alors constitué. Certaines informations sont supprimées de ce corpus (années et noms de personnes) afin de constituer les données ci-dessous :

- **Corpus 1 :** Corpus sans la présence d'années ni de noms de personnes : les années et les noms de personnes sont remplacés par les balises `<date>` et `<nom>`.

- **Corpus 2 :** Corpus sans années : les années sont remplacées par la balise `<date>`.

- **Corpus 3 :** Corpus avec la présence des années et des noms de personnes.

Le but du défi consiste à déterminer les phrases issues du corpus de F. Mitterrand introduites dans le corpus composé d'allocutions de J. Chirac.

3 Comités

3.1 Comité d'Organisation

Responsables : Jérôme Azé (LRI - IA) et Mathieu Roche (LRI - IA)

Membres :

- Thomas Heitz (LRI - IA)
- Amar-Djalil Mezaour (LRI - IASI)
- Érick Alphonse (INRA - MIG)
- Ahmed Amrani (ESIEA & LRI - IA)

3.2 Comité de Programme

Présidents : Violaine Prince (LIRMM - TAL) et Yves Kodratoff (LRI - IA)

Membres :

- Nathalie Aussenac-Gilles (IRIT)
- Valérie Beaudouin (France Telecom)
- Catherine Berrut (CLIPS - MRIM)
- Béatrice Daille (LINA - LeC)
- Patrick Gallinari (LIP6 - Connexioniste)
- Éric Gaussier (Xerox Research)
- Thierry Hamon (LIPN - RCNL)
- Fidélia Ibekwe (ERSICOM)
- Michèle Jardino (LIMSI - LIR)
- Éric Laporte (IGM-LabInfo - Informatique linguistique)
- Josiane Mothe (IRIT, SIG)
- Xavier Polanco (INIST - URI)
- Pascal Poncelet (LGI2P - EMA)
- Christian Retoré (LABRI - SIGNES)
- Christophe Roche (LISTIC - Condillac)
- Pascale Sébillot (IRISA - TEXMEX)
- Yannick Toussaint (LORIA - Orpailleur)
- François Yvon (ENST)

Remerciements

Nous remercions les organisateurs de TALN'05 de nous avoir permis de mettre en œuvre cet atelier dans les meilleures conditions possibles. Nous remercions l'association **AFIA**[2] (Association Française d'Intelligence Artificielle) pour le prix de 300 euros attribué au gagnant de DEFT'05 ainsi que Jérémie Mary pour la conception du logo de ce défi. Enfin, n'oublions pas de remercier et surtout de féliciter les onze équipes issues de neuf laboratoires différents qui ont participé à DEFT'05.

[2] http ://afia.lri.fr/