Keynote Speech 2

Friday, May 21th, 9:00

Chairperson: **Jan Odijk**

Challenges and Methods for Multilingual Text Mining

Ralf Steinberger

Multilingual text processing is useful because the information content found in different languages is complementary, both regarding facts and opinions. While Information Extraction and other text mining software can, in principle, be developed for many languages, most text analysis tools have only been applied to small sets of languages because the development effort per language is large. Self-training tools obviously alleviate the problem, but even the effort of providing training data and of manually tuning the results is usually considerable. In this paper, we gather insights by various multilingual system developers on how to minimise the effort of developing natural language processing applications for many languages. We also explain the main guidelines underlying our own effort to develop complex text mining software for tens of languages. While these guidelines – most of all: extreme simplicity – can be very restrictive and limiting, we believe to have shown the feasibility of the approach through the development of the Europe Media Monitor (EMM) family of applications (http://press.jrc.it/overview.html). EMM is a set of complex media monitoring tools that process and analyse up to 100,000 online news articles per day in between twenty and fifty languages. We will also touch upon the kind of language resources that would make it easier for all to develop highly multilingual text mining applications. We will argue that – to achieve this – the most needed resources would be uniform and simple multilingual dictionaries, corpora, and software tools.

Nom du document : Steinberger_abstract.doc
Répertoire : C:\Comm\Events\LREC\LREC 2010\Proceedings\BoA
Modèle : C:\Documents and Settings\Hélène\Application Data\Microsoft\Modèles\Normal.dot
Titre :
Sujet :
Auteur : Irene Russo
Mots clés :
Commentaires :
Date de création : 21/04/2010 6:28
N° de révision : 5
Dernier enregistr. le : 26/04/2010 10:48
Dernier enregistrement par : Irene Russo
Temps total d'édition : 12 Minutes
Dernière impression sur : 30/04/2010 11:03
Tel qu'à la dernière impression
Nombre de pages : 1
Nombre de mots : 253 (approx.)
Nombre de caractères : 1 464 (approx.)