



Sujet :

Analyse d'opinions ou de sentiments exprimés dans des textes

Niveau : Bac + 5 - 3^{ème} année d'école d'ingénieur – Master

Durée du stage : 4 à 6 mois

Contexte

Les blogs sont de nos jours un moyen de plus en plus utilisé par les internautes pour diffuser de l'information. Ils représentent l'état d'esprit d'une population donnée à un instant T sur un sujet particulier (politique, scientifique, médical, culinaire, etc) ou sur un produit donné (nouveau modèle de voiture par exemple, l'image de marque d'une entreprise, etc). Cette image est importante pour la Société. Toutes les personnes et toutes les entreprises pour lesquelles l'opinion publique est importante ont besoin de se tenir au courant de leur image et des sujets qui intéressent la population qu'elles ont pour cible pour s'adapter à ses besoins, anticiper ses besoins futurs ou simplement évaluer l'impact d'une campagne médiatique.

Pour ces raisons, l'analyse des sentiments est devenu un sujet actif, dans la communauté scientifique du TAL et de la sémantique comme en témoigne l'introduction de la tâche 14 (Affective Text) de la campagne d'évaluation Semeval sur la sémantique. C'est aussi un domaine émergent que se disputent les acteurs du data mining et où l'utilisation d'une analyse plus profonde des textes se justifie.

Le stage se situe dans un contexte double, celui de l'extraction d'information et celui de l'analyse des sentiments ou des opinions. L'extraction d'information consiste à consolider une connaissance structurée et semi-formelle à partir de documents non structurés (typiquement du texte en langage naturel). Elle est utilisée par exemple pour remplir des formulaires (*templates*) prédéfinis à partir de textes. L'analyse des opinions ou des sentiments consiste à étiqueter un texte pour savoir s'il traduit de façon plus ou moins marquée de la part de son auteur des sentiments de colère, d'indignation, de joie, etc.

A partir d'un extrait de texte tel que :

Bien sûr c'est une question de goûts, mais entre ce nouveau Koleos et les récents Citröen C-Crosser et Peugeot 400Z, mon choix irait vers la marque au Losange. Et vous?

issu d'un blog d'essais automobile, le système devrait pouvoir extraire les modèles de voiture, les marques et les expressions d'opinion et de sentiments qui leur sont attachés, et donc indiquer que l'auteur a une préférence pour le modèle 4x4 de Renault.

Objectifs du stage.

L'objectif du stage sera de définir un modèle sémantique sur un sujet d'actualité restreint (description d'un produit, d'un service ou d'une situation avec une ontologie et le lexique associé), de se servir de ce modèle pour paramétrer le système d'extraction d'information de notre plate-forme, puis d'exploiter les données extraites pour analyser les tendances de l'opinion ou les sentiments qui se dégagent d'un ensemble de textes traités. En dernier lieu les tendances doivent être visualisées selon les différents axes sémantiques de l'ontologie.

Le laboratoire LIC2M du CEA LIST dispose déjà pour réaliser cette tâche d'un environnement de développement permettant de réaliser une analyse morpho syntaxique de texte et de l'annotation sémantique dans des domaines de spécialité. Le laboratoire dispose par ailleurs aussi de ressources lexicales enrichies pour de l'analyse de sentiments.

Bibliographie :

Guillaume Pitel, Gregory Grefenstette (2008)
Semi-automatic Building Method for a Multidimensional Affect Dictionary for a New Language
6th Conference on Language Resources and Evaluation (LREC 2008), Marrakech, Morocco



Carlo Strapparava and Rada Mihalcea. 2007 Semeval-2007 task 14: Affective text. In Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007), pages 70–74, Prague, Czech Republic, June. Association for Computational Linguistics.

Domaines de spécialité requis :

Informatique, Traitement automatique des langues,
apprentissage statistique, analyse opinion

Moyens informatiques mis en œuvre :

Langages : C++, langage de script de type Perl ou Python sous Linux

Autres moyens mis en œuvre (expériences, méthodes d'analyses, autres...)

Centre : Fontenay-aux-Roses

Pôle ou Direction : DRT - LIST

Dépt/Service/Labo : DTSI/SRCI/LIC2M :

Laboratoire d'Ingénierie de la Connaissance Multimédia Multilingue

Web : <http://www-list.cea.fr/>

Encadrement

Nom de l'ingénieur responsable :	Romaric Besançon Olivier Ferret	Tél. :	+33 1 46 54 80 16 +33 1 46 54 96 45
E-mail :	romaric.besancon@cea.fr olivier.ferret@cea.fr	Fax :	+33 1 46 54 75 80
Secrétariat :	DTSI/SRCI	Tél. :	+33 1 46 54 91 17
Nom du chef de laboratoire :	Olivier Mesnard	Tél. :	+33 1 46 54 73 38