Proposition de recherche doctorale

Stylistique automatique et identification d'auteurs

Mots clés :

- **Directeur de thèse** : Jean-Gabriel Ganascia
- **Co-encadrant(s)** :
- **Unité de recherche** : Laboratoire d'informatique de Paris 6
- **Ecole doctorale** : École Doctorale Informatique, Télécommunications, Électronique de Paris
- **Domaine scientifique principal**: Divers

Résumé du projet de recherche (Langue 1)

Cette thèse porte sur la {{stylistique automatique}} à l'aide de techniques d'apprentissage automatique. Il s'agit de caractériser l'auteur, le genre ou l'époque. Dans le passé, beaucoup de travaux ont porté sur l'attribution de paternité de textes. Il s'agit de reprendre ces études avec deux ambitions: -* {{renouveler les méthodes}}: à la différence des méthodes classique fondée sur la lexicométrie, l'approche proposée fera appel à une extraction de motifs syntaxiques. -* {{identifier les caractéristiques du style}}: les motifs syntaxiques devraient aider à expliciter les caractéristiques du style de tel auteur, de tel type de texte ou de telle époque. Pour mener à bien ce travail, on aura recours à -* des techniques de {traitement automatique des langues} (étiquetage syntaxique, analyse syntaxique, etc.) -* de la {fouille de textes} (extraction de motifs récurrents) -* de l'{apprentissage supervisé} Les recherches se poursuivrons dans le cadre du {{Labex OBVIL}} qui fait collaborer l'équipe ACASA du LIP6 avec les équipes de littérature de l'université Paris-Sorbonne. Cela permettra de valider les approches proposées en contact avec des équipes de littérature. Cela fournira aussi des corpus pertinents. A titre d'illustration, une première validation doit porter sur les romans érotiques du XIXe siècle, dont les auteurs sont en partie anonymes.

Résumé du projet de recherche (Langue 2)

Les défis scientifiques de cette thèse sont doubles: -# Il faut améliorer les techniques d'{{attribution de paternité}} de textes. Pour cela on fera des comparaisons sur des corpus connus avec les techniques classiques fondées sur la lexicométrie, c'est-à-dire sur des vecteurs de mots. -# Le second défi porte sur le {{renouveau de la stylistique}}: il faut induire des caractéristiques lisibles, c'est-à-dire des motifs syntaxiques suffisamment longs. Cela exige le recours à des analyseurs syntaxiques automatiques, puis à des approches d'extraction de motifs.

Informations complémentaires (Langue 1)

Comme dit plus haut, ce projet se réalisera dans le cadre du Labex OBVIL. Ce Labex entretien des liens privilégié avec le {{projet ARTFL de l'université de Chicago}} avec qui nous aurons plusieurs échanges. Une thèse de littérature en co-tutelle entre le Labex OBVIL et une {{équipe de linguistique informatique de l'université de Saint-Petersburg}} (Russie) se déroule actuellement. Elle porte sur l'identification des auteurs de plusieurs romans érotiques du XIXe siècle. Cette thèse se fera en collaboration avec cette équipe. Enfin, dans le cadre du Labex OBVIL, il existe des partenariats européens et avec les Etats-Unis.