LI, Université François Rabelais de Tours
LIMSI-CNRS, Université Paris XI

Proposition de stage Master / Ecole d'Ingénieur

Techniques de fouille de données pour la recherche d'information : évaluation des ressources et traitements pour la reconnaissance d'entités nommées

Résumé

Proposition de stage de fin d'études ou de Recherche de niveau Bac+5 (Master, Ecole d'Ingénieur) en Informatique appliquée au Traitement Automatique des Langues d'une durée de 4 mois minimum.

Contexte scientifique

Le Laboratoire LI et le Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur (LIMSI) proposent un sujet de stage commun dans le cadre du projet de recherche TMH (Télécommunications, Mobilité et Handicap) financé par la société BAMSOO. Le sujet porte sur le Traitement Automatique des Langues (TAL) appliqué à des tâches d'extraction d'information textuelle à l'aide de techniques de fouille de données. Parmi les tâches relevant de l'extraction d'information se trouve la reconnaissance automatique des entités nommées (REN) qui consiste à rechercher des références à des noms de personnes, de lieux, d'organisation, d'unités monétaire ou temporelles dans de grands flux de données. Ces entités nommées peuvent présenter des formes linguistiques très variées. Par exemple, les systèmes de REN doivent reconnaître que *François Hollande*, *le président de la République*, *le président normal* ou *le locataire de l'Elysée* désignent toutes la même personne, qui est une personnalité politique. C'est sur cette tâche de reconnaissance d'entités nommées que portera spécifiquement ce stage.

Le système que nous avons développé (mXS) propose une approche de type fouille de données dont une des caractéristiques est de rechercher séparément le début et la fin de chaque entité nommée. Dans ses fondements, mXS repose sur l'énumération de motifs en s'appuyant sur des techniques bien établies en TAL (catégorisation morpho-syntaxique, lemmatisation, utilisation de lexiques à large couverture) comme en fouille de données (motifs séquentiels, hiérarchies, règles d'association). Il a obtenu de bonnes performances dans le cadre de la campagne d'évaluation ETAPE, en particulier dans des contextes bruités (transcriptions automatiques). Au delà des performances globales du système, il reste difficile de déterminer quels sont les choix de modélisation effectués qui avantagent ou pénalisent le système. Outre une étape de ré-ingénierie logicielle, ce stage a pour objectif de mener des travaux expérimentaux permettant de mieux cerner les apports de notre démarche.

Travail à réaliser

Le travail à réaliser vise à consolider le code du système existant puis à mieux étudier son comportement et éventuellement dresser un état des lieux des utilisations possibles de la fouille de données pour diverses tâches tournées vers le TAL. Il comportera deux phases successives principales :

Phase 1 (2 mois) : ré-ingénierie logicielle – Cette étape consistera à factoriser et optimiser le code existant, afin de le rendre plus évolutif, de le mettre à disposition en ligne et de favoriser sa large diffusion dans la communauté scientifique. En pratique, il s'agira ici de mieux modulariser les différents traitements TAL qui le composent :

- prétraitements (morpho-syntaxe, lexiques) pour enrichir les textes selon la langue et les outils disponibles,
- extraction de motifs séquentiels hiérarchiques,
- modèles (symboliques ou statistiques) qui exploitent les motifs pour l'annotation.

Phase 2 (2 mois miniimum) : étude du comportement du système – Une étude approfondie sera menée sur l'intérêt de rechercher des marques de début et de fin d'entités nommées, plutôt que d'adopter une approche plus classique de classification mot-à-mot (i.e. décider si chaque mot fait partie ou non d'une entité nommée). Pour cela, une analyse sera conduite sur la comparaison des performances et des sorties de différents types de systèmes : à base de règles et DAG (graphes dirigés sans cycle) comme le système CasEN développé également au LI, mais aussi de CRF. A terme, ce travail permettra de définir les perspectives d'évolution les plus prometteuses pour les systèmes traitant cette tâche.

Phase complémentaire (si extension de stage) – En cas d'avancée satisfaisante du travail, on cherchera à étudier comment mieux manipuler les motifs à des fins d'extraction de connaissances. Ce travail commencera par se doter des outils nécessaires à la caractérisation d'un corpus à partir des motifs qui en ont été extraits automatiquement. L'objectif étant d'être à même de sélectionner les motifs d'intérêt par utilisation de méthodes formelles, les motifs étant organisés au sein de treillis.

Profil recherché

La personne recrutée sera en cycle terminal d'études en informatique, de niveau Bac+5 (Master informatique professionnel, recherche ou indifférencié, école d'ingénieur). Des compétences en Traitement Automatique des Langues et/ou en Fouille de Données seront appréciées. Dans le cas d'un(e) étudiant(e) en Master Recherche, le sujet de stage pourra être adapté aux attentes de l'étudiant. Potentiellement, ce travail pourra donner lieu à communication dans des conférences scientifiques.

Rémunération

Rémunération maximale prévue par la règlementation à savoir 436,05 € par mois, pour une durée de 4 mois de stage minimum (prolongation de la durée du stage jusqu'à 6 mois à la demande de l'étudiant ou de son établissement). Cette rémunération sera assurée dans le cadre d'un projet industriel financé par la société BAMSOO.

Lieu d'exercice

Le stage se déroulera dans les locaux du Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur (LIMSI-CNRS), Université Paris-Sud, Rue John von Neumann, 91403 Orsay, au sein de l'équipe ILES (Information, Langue Ecrite et Signée). Le stage sera encadré part Damien Nouvel, ingénieur de recherche au LIMSI et Jean-Yves Antoine, professeur de l'Université François Rabelais de Tours (équipe BDLTN.

Contact – Dépôts de candidature

Contact : damien.nouvel@limsi.fr

Dépôt des candidatures : auprès de Damien Nouvel. Merci de déposer un CV détaillé de vos activités passées, accompagné d'une lettre de motivation et de vos relevés de notes des deux dernières années d'études. Un petit travail de développement sera demandé pour la sélection du candidat.

Liens utiles

- Système mXS : http://damien.nouvels.net/fr/mXS
- Laboratoire LIMSI (groupe ILES) : http://www.limsi.fr/Scientifique/iles/
- Laboratoire LI (équipe BDTLN) : http://li.univ-tours.fr/equipes/equipe-bdtln-198022.kjsp
- Campagne Etape : http://www.afcp-parole.org/etape.html