

Enrichissement de lexique émotionnel pour l'informatique affective

Proposition de stage de fin d'études

Résumé

Proposition de stage de recherche ou de fin d'études en informatique appliqué au Traitement Automatique des Langues Naturelles, d'une durée minimale de 4 mois.

Contexte scientifique

En collaboration avec le Laboratoire d'Informatique de l'Université de Tours, le laboratoire IRISA, antenne de Lorient (56 – Morbihan) propose un sujet de stage dans le cadre du projet de recherche DAPAI-EMO financé par la société BAMSOU.

Le projet DAPAI-EMO fait suite à un projet (EmotiRob) concernant le développement d'un robot compagnon affectif pour des enfants en hospitalisation longue. Cette poursuite de travaux fait abstraction ici de sa dimension robotique pour se concentrer sur ses aspects liés à la compréhension émotionnelle de la langue. Au cours du projet EmotiRob, nous avons développé EmoLogus, un système de détection des émotions qui intervient à la suite d'un système logique de compréhension de message appelé Logus. EmoLogus utilise la structure sémantique de l'énoncé fourni par Logus pour mettre en œuvre un calcul de la valence émotionnelle portée par l'énoncé, c'est-à-dire pour savoir si celui-ci porte une émotion positive, négative ou neutre. Ce calcul logique se base principalement sur l'utilisation de normes lexicales émotionnelles qui décrivent le système de valeurs du système : à chaque mot du vocabulaire est associée une valence (positif, neutre, négatif) et une intensité (nul, faible, fort) émotive. Le lexique émotionnel sur lequel se base le système a été élaboré en collaboration avec l'équipe de psycholinguistique d'Arielle Syssau, de l'Université Montpellier 2. Basé sur des jugements évaluatifs contrôlés auprès d'une population de test échantillonnée avec soin, il nous garantit la représentativité du système de valeurs d'EmoLogus.

Le système a montré une bonne robustesse de détection dans le cadre restrictif d'une communication enfantine. Il souffre toutefois du manque de couverture de son lexique émotionnel. A l'heure actuelle, le système EmoLogus intègre en effet un lexique limité à un millier de mots, alors que la langue française générale compte entre 50 000 et 100 000 entrées lexicales. Dans le cadre de ce projet, nous proposons d'utiliser des techniques d'extension automatique de lexique émotionnel à partir d'une ressource initiale telle que celle du système EmoLogus. Parmi les méthodes proposées pour étendre automatiquement un lexique émotionnel étendu, on distingue deux types d'approches :

- celles basées sur des réseaux sémantiques comme WordNet, où sont décrits des relations de synonymies entre tous les mots d'une langue. On peut alors rechercher des synonymes des mots germes présents dans le lexique originel et leur appliquer un algorithme de propagation de valence,
- celles basées sur des techniques d'analyse de données sur des corpus textuels. Dans ce second cas, on va étudier les cooccurrences de mots dans un corpus pour calculer des similarités sémantiques (remplaçant les liens de synonymie explicites de Wordnet) et les intégrer dans le calcul de la valence des mots du lexique. Dans le cadre de ce stage, on se propose ainsi d'utiliser la technique de l'analyse sémantique latente (LSA : *Latent Semantic Analysis*) pour calculer ces proximités sémantiques et s'en servir pour estimer la valence d'un mot. Les mots germes déjà présents dans la norme lexicale émotionnelle actuelle serviront de base à l'espace vectoriel sur lequel sera opérée l'analyse de données permettant l'extension du lexique.

Ce stage visera à développer au moins une de ces deux techniques pour étendre le lexique émotionnel d'EmoLogus, et tester l'apport de cette extension sur un corpus de test. En cas d'avancée significative, ce travail pourra donner lieu à communication dans une conférence scientifique à laquelle sera invité à participer le stagiaire.

Travail à réaliser

La personne recrutée sera en charge de la conception de nouvelles techniques d'extension de lexique émotionnel, du développement d'un lexique à large couverture pour le système EmoLogus ainsi que de la réalisation de tests d'évaluation du système étendu obtenu. Le stage se déroulera en trois étapes successives :

- *Phase n°1 – Préparation des données (T0 – T0+1)* : Veille technologique sur le sujet, définition des formats d'échange entre les différentes techniques d'extension du lexique, caractérisation d'une ou plusieurs application test et définition des données de test en relation et du protocole d'évaluation final.
- *Phase n°2 – Extension de lexique par relations sémantiques (T0+1 – T0+3)* : Extension du lexique germe par analyse des relations de synonymie et d'antonymie entre éléments (synsets) de Wordnet, évaluation de l'approche sur données de test (comparaison des performances d'EmoLogus avec ou sans le lexique étendu).
- *Phase n°3 – Extension de lexique par analyse de données (T0+4 – T0+6)* : Extension du lexique germe par analyse sémantique latente. Évaluation sur tests unitaires de l'approche, évaluation de l'approche sur données de test (comparaison des performances d'EmoLogus avec ou sans le lexique étendu). Cette phase ne sera abordée qu'en cas de stage de durée supérieure à quatre mois.

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 connaissances en Traitement Automatique des Langues et en analyse de données seront appréciées, sans être un pré-requis à recrutement. Dans le cas d'un(e) étudiant(e) en Master Recherche, le sujet de stage pourra être adapté aux attentes de l'étudiant.

Rémunération

Rémunération minimale prévue par la réglementation à savoir 436,05 € par mois. Cette rémunération sera assurée dans le cadre d'un projet industriel financé par la société BAMSOO.

Durée du stage et lieu d'exercice

La personne recrutée travaillera au sein du laboratoire IRISA, dans les locaux de l'ENSIBS, à Lorient (Morbihan). Il s'intégrera dans une équipe projet composée de Jeanne Villaneau (IRISA, équipe SEASIDE) et Jean-Yves Antoine (Laboratoire d'Informatique de l'Université François Rabelais de Tours, équipe BDTLN).

La durée minimale de stage sera de 4 mois. Une prolongation de stage est envisageable à la demande du stagiaire ou de son établissement.

Contact – Dépôts de candidature

Contact : Jeanne.Villaneau@univ-ubs.fr

Dépôt des candidatures : auprès de Jeanne Villaneau. 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 développement Java sera demandé pour la sélection du candidat.

Liens utiles

Laboratoire LI, équipe BDTLN : <http://li.univ-tours.fr/equipes/equipe-bdtln-198022.kjsp>

Laboratoire IRISA, équipe SEASIDE : <http://www-seaside.irisa.fr/>