Rapport d'examinateur sur une demande
de subvention d'un programme de recherche
CRSNG – mars 2008

Examinateur : Malek Boualem (France Télécom Orange Labs, France)
No NIP du CRSNG : 256425

No du dossier : 363705
No du comité : 96

Candidats : XXX, XXX (Université de Montréal), en collaboration avec la société X Inc

Titre de la proposition :
Amélioration de recherche de correspondances dans XXX.

Résumé de la proposition :

XXX ressemble à un moteur de recherche "à la Google" appliqué à des millions des paires de phrases qui sont des traductions l'une de l'autre. Une requête, exprimée sous forme d'un ou plusieurs mots, retourne non seulement toutes les phrases contenant ces mots mais aussi les phrases correspondantes dans l'autre langue. Cet outil permet donc à un traducteur ou un rédacteur de retrouver des solutions toutes faites à une grande variété de problèmes de traduction. XXX, utilisé quotidiennement par des milliers de traducteurs, a été développé et commercialisé au cours des dernières années grâce à une collaboration entre le laboratoire de Recherche appliquée en linguistique informatique (XXX) à l'Université de Montréal et XXX, une entreprise d'Ottawa qui commercialise des outils d'aide à la traduction. XXX doit maintenant être mis à jour afin de profiter des nouveaux développements informatiques et algorithmiques dans le traitement de la langue naturelle. La recherche portera surtout sur le "repérage de traductions" (translation spotting), ce qui permettra d'afficher des correspondances entre les mots d'une phrase source et leurs équivalents dans la phrase cible. Cette nouveauté facilitera le travail des traducteurs, mais surtout elle permettra d'élargir la gamme des applications de XXX sur le web et dans l'environnement de bureau des traducteurs. Ce partenariat université et industrie, intégrant un chercheur post-doctoral, un étudiant de doctorat et un étudiant de maîtrise, permettra au Canada de garder sa position de chef de file dans le domaine des outils d'aide à la traduction.

1. Mérite scientifique et faisabilité technique

Le projet proposé confirme la continuité et l'évolution des travaux de recherche du laboratoire XXX de l'Université de Montréal, notamment dans le domaine de l'aide à la traduction. Le projet focalise notamment sur l'exploitation de nouvelles techniques d'alignement de textes au niveau des expressions et des mots pour le repérage des traductions. Il est à noter que contrairement à l'alignement automatique des textes et des paragraphes, l'alignement à des niveaux plus fins (expressions et mots) est encore un sujet moyennement maitrisé. Toutefois, le XXX a une notoriété indiscutable dans ce domaine de recherche. En outre, il s'agit d'un programme de recherche pour lequel les chercheurs ont déjà amorcé des réflexions et des travaux concrets (c.f. communications et

publications sur le sujet). Ce projet ne se situe donc pas dans un registre complètement exploratoire où les risques peuvent être plus ou moins importants.

Afin de valider ses observations (y compris sur le comportement des utilisateurs) et bien mener ses expérimentations dans le cadre de ce projet, le XXX semble pouvoir disposer des corpus des requêtes soumises à XXX (plusieurs millions de requêtes). Cette possibilité d'accès à des corpus de requêtes, facilitée par la collaboration avec la société Terminotix, donne un avantage considérable pour mener à bien les travaux de recherche et de développement envisagés dans le cadre de ce projet.

Outre, les outils d'alignement du XXX et l'outil XXX, le projet envisage de réutiliser un certain nombre de composants et de ressources mis en œuvre par le XXX ou par XXX (logiciel de dépouillement terminologique, étiqueteurs pour l'anglais et le français, corpus alignés d'autres langues, etc.). La disponibilité et la réutilisation des composants existants sont importantes pour garantir la faisabilité technique du projet. En matière de plateforme informatique, le projet part sur des bases rassurantes car il envisage d'utiliser des technologies modernes (Java, Python, etc.).

Par ailleurs, il est intéressant de constater que les travaux de R&D envisagés dans le cadre de ce projet, sur le thème de l'alignement d'éléments sous-phrastiques pour l'aide à la traduction, pourront également apporter des avancées intéressantes pour d'autres domaines en lien avec l'informatique linguistique, comme l'extraction terminologique bilingue, etc.

La durée du projet est prévue sur trois ans. Ce délai me semble raisonnable pour la faisabilité technique des activités du projet.

Concernant l'organisation du projet, je n'ai pas de remarque particulière au sujet du calendrier des activités qui me semble réalisable. Toutefois, je suggère aux auteurs de mieux distinguer (à travers le lotissement du projet) les activités de recherche et les activités de développement. Il serait sans doute plus pratique, pour la suite du projet, pour des adaptations ultérieures ou même pour des actions de communication scientifique, de distinguer les activités scientifiques liées aux méthodes d'alignement et au repérage des traductions et les activités liées aux tests, aux évolutions des scénarios d'utilisation du service web à travers les technologies Ajax, à l'intégration des nouvelles fonctionnalités dans XXX, à la compatibilité avec les éditeurs de textes existants, l'internationalisation et la localisation de XXX, etc.

2. Compétences de l'équipe de recherche

Le projet sera dirigé par deux chercheurs ayant une longue et solide expérience et une notoriété scientifique indiscutable, au niveau mondial, dans le domaine de la linguistique informatique et du traitement automatique du langage naturel : XXX et XXX, professeurs à l'université de Montréal.

Les autres membres de l'équipe du projet seront :

- un linguiste informaticien, également de renommée internationale et ayant une longue expérience sur le système XXX, pour superviser les tests, la terminologie et les corpus.

- un stagiaire postdoctoral pour les travaux sur la base de données, l'intégration des traitements linguistiques et la supervision du serveur web du service.
- un étudiant en thèse de doctorat pour des activités autour de nouveaux modèles d'alignement des mots.
- Un étudiant de maîtrise pour la partie client du service web et pour les interfaces avec les éditeurs de textes.

Je trouve que cette équipe, qui sera également aidée par deux ingénieurs informaticiens ayant aussi travaillé sur le système XXX, réunit toutes les compétences requises pour mener à bien les activités du projet et pour atteindre les objectifs visés.

Par ailleurs, le partenaire industriel XXX a également un rôle important dans la validation, l'intégration et la valorisation des résultats des travaux de recherche.

Remarque :
Habituellement, les stages postdoctoraux ne peuvent pas excéder deux ans. Comme le projet est prévu sur une durée de trois ans, les auteurs devront préciser s'ils envisagent de recourir à deux ou trois stages postdoctoraux sur la durée du projet.

3. Possibilités de formation

Le projet associe au moins plusieurs étudiants (un stagiaire postdoctoral, un étudiant doctoral et un étudiant de maîtrise). Je pense que le projet pourra également associer d'autres étudiants dans le cadre de stages ponctuels ou de courtes durées.

Même si le projet, par sa nature scientifique et technique, n'a pas une vocation pédagogique, je pense que la contribution de ces étudiants permet de considérer que le projet participe à un effort de formation aux domaines de la recherche, à travers, notamment, des participations éventuelles à des conférences scientifiques.

4. Contribution de l'industrie et pertinence

Le projet proposé s'appuie sur une collaboration déjà existante entre le laboratoire de recherche et une entreprise industrielle d'Ottawa (XXX) qui commercialise des outils d'aide à la traduction. L'outil XXX commercialisé par la société XXX est utilisé depuis plusieurs années par les traducteurs professionnels (près de 2000 utilisateurs) sous la forme d'un service en ligne. Le projet proposé consiste à apporter une évolution significative à ce service par le rajout d'une fonctionnalité de "repérage de traductions". Cette fonctionnalité, basée sur de nouvelles techniques d'alignement (*domaine sur lequel le XXX a une notoriété reconnue*), devrait permettre une identification automatique plus fine des segments correspondants dans les deux langues concernés par la traduction, en affichant les correspondances entre les mots d'une phrase source et leurs équivalents dans la phrase cible. Cette méthode permet notamment de minimiser les risques d'erreurs d'alignement et éviter leur propagation durant le processus de traduction.

Cette nouveauté devrait, en effet, faciliter le travail des traducteurs en augmentant l'éventail des correspondances paraphrastiques entre les deux langues.

Pour le partenaire industriel, ces fonctionnalités additionnelles permettraient aussi d'élargir la gamme des applications de XXX sur le web et dans l'environnement de bureau des traducteurs et accroitre sa clientèle.

Bien entendu, cette croissance économique pour l'entreprise devrait également être financièrement bénéfique pour l'université dans le cadre du contrat de licence établi avec l'entreprise.

Par ailleurs, le partenariat industriel avec la société XXX, permet au XXX d'accéder aux corpus des requêtes soumises à XXX (plusieurs millions de requêtes). Cette possibilité donne un avantage considérable pour mener à bien les travaux de recherche et de développement envisagés dans le cadre de ce projet.

Je suis globalement d'accord avec le postulat des auteurs concernant le bénéfice de ce partenariat entre l'université et l'industrie et qui devrait permettra au Canada de garder sa position de chef de file dans le domaine des outils d'aide à la traduction.

5. Budget du projet

La proposition inclut un plan détaillé des postes budgétaires et de dépenses pour chacune des trois années du projet. Les montants demandés me semblent être dans des proportions raisonnables et en adéquation avec les coûts pratiqués habituellement, en vue de la réalisation des objectifs du projet.

6. Propriété intellectuelle

Malgré le potentiel de ce projet en idées innovantes, la proposition ne mentionne pas des intentions concrètes de dépôt de brevet ou de logiciel. Je suggère de demander aux auteurs de préciser si le projet intègre de tels éléments de propriété intellectuelle et de compléter le budget par les coûts correspondants éventuels.

7. Avantages pour le Canada

Le multilinguisme au Canada a toujours été très favorable au développement de la traduction, qu'elle soit humaine, semi-automatique ou automatique. D'ailleurs, il est connu que le meilleur système de traduction automatique est Canadien, même s'il couvre un domaine d'application très spécifique. Il est connu aussi que le marché de la traduction est proportionnellement très porteur au Canada, que ce soit pour le couple de langues français/anglais ou pour d'autres couples de langues en fonction du développement des partenariats économiques du Canada.

Le projet proposé ici s'inscrit dans le cadre de l'aide à la traduction pour apporter des fonctionnalités nouvelles et innovantes à destination des utilisateurs et des professionnels de la traduction (y compris les traducteurs du service parlementaire qui semblent utiliser TransSearch de manière significative).

Compte tenu de ces éléments, il apparaît clairement que ce projet aura une contribution significative sur le plan économique au Canada.

D'autre part, il me parait également évident que ce projet contribuera à maintenir l'avancée déjà acquise par le Canada dans le domaine des outils d'aide à la traduction.

8. Autres points à considérer

- A mon avis, il serait opportun de demander aux auteurs du projet de compléter les motivations ayant poussé à cette proposition : s'agit-il d'une demande concrète des utilisateurs du service XXX ? ou est-ce qu'il s'agit d'une volonté de faire bénéficier le service d'une avancée scientifique ?

- A mon avis, il n'est pas opportun d'évoquer, dans ce projet, une quelconque similarité ou comparaison avec Google. Le projet ne vise pas la recherche d'informations mais plutôt l'aide à la traduction par l'intermédiaire d'un mécanisme de concordance. Ce mécanisme ne peut pas être assimilé à un mécanisme de recherche d'information où la base d'index est complètement différente d'une mémoire de traduction.

- Je pense que l'analyse de la concurrence n'est pas suffisamment développée dans ce dossier. Même si cette concurrence est "faible" compte tenu des spécificités de la solution du XXX/XXX, je pense qu'une analyse plus détaillée, notamment sur les prévisions des années à venir, serait bénéfique pour l'organisation du projet.

- La description du projet évoque la traduction statistique et son positionnement par rapport à la traduction à base de règles linguistiques. Contrairement à la position des auteurs, je pense que les deux approches ne peuvent pas véritablement être comparées de manière aussi "plate" car une telle comparaison devrait tenir compte de la nature des données traités et surtout des applications visées et des services rendus par les systèmes de traduction. Un débat à ce sujet existe au sein de la communauté scientifique du traitement automatique des langues et les auteurs sont certainement concernés et même impliqués dans ce débat.

9. Recommandation générale

Compte tenu de tous les éléments évoqués dans mon rapport, j'estime que le programme de recherche proposé mérite amplement d'être soutenu.