

Améliorations de la recherche de correspondances dans TransSearch

TransSearch est un concordancier bilingue commercial, développé depuis plus de 15 ans par l'équipe du **professeur Guy Lapalme**. Il explique ici ses intérêts pour la traduction et comment il travaille à l'amélioration du système.



Pour commencer, pourriez-vous décrire vos intérêts de recherche ? Comment votre expérience passée vous a-t-elle préparé à l'étude de programmes de traduction ?

J'ai commencé à travailler sur le traitement automatique de la langue naturelle (TALN) dans les années 80 en collaboration avec mes étudiants gradués. Nous avons développé le premier programme jouant au Scrabble à un niveau de championnat et nous avons été les premiers à implanter un générateur de texte

basé sur les principes de la théorie Sens-Texte pour laquelle nous avons aussi développé un éditeur de dictionnaire. J'ai été un des pionniers de l'utilisation de la programmation logique au Canada.

En 1997, j'ai créé le groupe de Recherche Appliquée en Linguistique Informatique (RALI) afin d'intégrer les activités d'un groupe de la traduction assistée par ordinateur qui oeuvrait au sein du gouvernement canadien depuis 1984. Avec ses 3 professeurs, 20 étudiants gradués et stagiaires postdoctoraux et un chercheur, le RALI est un des plus grands laboratoires universitaires spécialisés dans le traitement de la langue au Canada.

En plus d'une forte activité en traduction automatique statistique et traduction assistée par ordinateur, le RALI a créé des liens forts avec l'industrie au moyen de subventions de recherche coopérative avec des entreprises canadiennes telles que Terminotix, NLP technologies, Druide Informatique, Lexum, KeaText, Nuance et Yahoo ainsi qu'avec des organismes gouvernementaux tels qu'Environnement Canada. Au fil des ans, j'ai réussi à maintenir un bon équilibre entre la collaboration avec l'industrie et la poursuite de principes académique. En 2010, j'ai d'ailleurs été

invité à présenter une présentation sur le sujet à la conférence annuelle d'intelligence artificielle canadienne.

Quelles autres contributions avez-vous faites dans le domaine du TALN ?

J'essaie de combiner la recherche de pointe et son application pratique au moyen de collaborations à long terme avec différents partenaires tant académiques qu'industriels. J'ai publié plus de 300 articles dans des revues et conférences avec arbitrage. L'Université de Neuchâtel m'a récemment décerné un doctorat honoris causa pour mes contributions au domaine du TALN. Au cours de ma carrière, j'ai dirigé 44 étudiants à la maîtrise, 22 au doctorat et huit au postdoctorat; 18 d'entre eux oeuvrent dans le milieu académique comme professeurs d'université ou chercheurs et trois ont créé des entreprises de TALN.

Comment avez-vous développé l'outil TransCheck ? Quel est le but de ce programme ?

TransCheck cherche à réduire le coût du contrôle de qualité en traduction en automatisant partiellement le travail de révision. Cet outil peut aider un réviseur humain en détectant

Un traducteur interroge TransSearch pour les traductions de "in keeping with"

TransSearch répond avec des phrases contenant la requête...

... ainsi qu'avec leur traductions respectives

TransSearch surligne la traduction de la requête...

... et regroupe toutes les variantes trouvées dans sa base de données.

La traduction de TransSearch

certain types d'erreurs de traduction et en aidant à appliquer certains standards bilingues dans un grand service de traduction. TransCheck a été conçu pour vérifier les correspondances entre un texte source et une ébauche de traduction afin de s'assurer de maintenir certaines propriétés reconnues comme étant souhaitables en traduction. C'est qui différencie un outil comme TransCheck travaillant sur un bitexte par rapport aux outils d'aide à la rédaction unilingue comme un détecteur de fautes d'orthographe ou de grammaire. Comme ces derniers travaillent sur un texte unilingue, ils ne sont pas capables de détecter des fautes de traduction, même les plus flagrantes.

Vous travaillez actuellement sur TransSearch, qu'est-ce qui vous a inspiré ce projet et qu'essayez-vous de réaliser.

Le RALI collabore depuis 2001 avec un partenaire industriel pour commercialiser TransSearch, mais aussi pour l'améliorer afin de tirer parti des développements récents en informatique et dans la conception d'applications web. Le projet de recherche s'est concentré sur l'identification des correspondances de traduction, qui permet à TransSearch d'afficher non seulement les phrases correspondantes, mais aussi de surligner les mots correspondant à la requête dans chaque phrase cible. Ceci facilite la tâche du traducteur, mais en plus élargit l'éventail des applications de TransSearch, tant sur le web que sur station de travail.

Quels aspects de la recherche de correspondance avez-vous améliorés dans TransSearch ?

Nous nous sommes concentrés sur l'identification des correspondances. Deux chercheurs postdoctoraux ont développé, implanté et évalué plusieurs algorithmes afin de trouver un bon compromis entre la précision et la vitesse. Un autre défi était la fusion de traductions équivalentes, c.-à-d. des traductions qui sont des variations grammaticales les unes des autres ou qui contiennent des mots grammaticaux qui ne changent pas foncièrement la traduction.

Cette combinaison de méthodes permet de regrouper les traductions identiques parmi plusieurs exemples d'un même texte source et le calcul de statistiques d'occurrence. TransSearch peut ainsi afficher la liste des traductions différentes en ordre décroissant de fréquence, permettant au traducteur de se concentrer sur quelques exemples déjà triés et fusionnés plutôt que d'avoir à se pencher sur plusieurs traductions pêle-mêle.

L'identification de correspondances dans TransSearch est un projet pour améliorer un concordancier bilingue à l'aide de l'alignement de mots. Basé à **l'Université de Montréal**, le projet cherche des traductions d'expressions complexes de la langue tout en fournissant des contextes d'utilisation.

BEAUCOUP DE TEMPS et d'efforts ont été consacrés à l'amélioration de la traduction automatique et pourtant les experts affirment qu'il y a encore beaucoup à explorer et à améliorer dans les outils de traitement automatique de la langue. Une des dernières études dans ce domaine est menée par le laboratoire de Recherche Appliquée en Linguistique Informatique (RALI) qui est un des grands laboratoires universitaires de TALN au Canada. Le RALI réunit une équipe d'informaticiens et de linguistes spécialisés avec une longue expérience dans le domaine. Le groupe cherche à simplifier le travail de l'utilisateur en améliorant la recherche de correspondances au moteur de recherche TransSearch permettant d'identifier les correspondances des mots dans un texte et sa traduction.

Le RALI, dirigé par Guy Lapalme - professeur au Département d'informatique et de recherche opérationnelle à l'Université de Montréal - a développé TransSearch. C'est une application web commerciale solide - un moteur de recherche - semblable à Google - orienté vers les professionnels de la langue, conçue pour trouver des traductions de mots ou de phrases parmi des millions de paires de phrases qui sont des traductions mutuelles.

Essentiellement, l'outil permet aux traducteurs de soumettre des requêtes, typiquement un mot ou une expression, à une mémoire de traduction, pour obtenir une liste de solutions toutes faites pour une multitude de problèmes difficiles de traduction. Le moteur de recherche ne retourne pas seulement les phrases contenant la requête, mais aussi leur traduction dans l'autre langue. Les traductions similaires sont regroupées, permettant une exploration rapide de différentes traductions possibles parmi lesquelles le traducteur choisira la plus appropriée.

Présentement TransSearch cherche des traductions dans une banque de plus de 500 millions de mots dans plus de 15 millions de paires

de phrases anglais-français. Les textes sont les transcriptions bilingues de plus de 25 ans de débats parlementaires de la Chambre des communes et du Sénat canadiens, des décisions de cours de justice canadiennes et des textes de l'Organisation internationale du travail.

Même si TransSearch n'est pas un traducteur automatique, c'est un découvreur de traductions conçu pour améliorer la précision et accélérer le travail des traducteurs. Cet outil n'est ni un simple dictionnaire bilingue ni un thesaurus parce qu'il trouve des traductions de groupes de mots. S'il lui manque l'intelligence d'un traducteur humain (par exemple, lors du traitement de textes complexes qui requiert une analyse et une interprétation précises), il laisse toutefois le contrôle à l'utilisateur pour des résultats plus précis.

Plusieurs facteurs ont conduit à la décision d'améliorer ce concordancier bilingue web. Lors de ses recherches, le RALI a constaté que les usagers du système sont majoritairement des traducteurs professionnels. De plus, l'étude des journaux de requêtes a montré que le système est consulté sur des problèmes difficiles de traduction. Parmi environ 7 millions de requêtes soumises au système sur une période de six ans, 87 pour cent comprenaient au moins deux mots. Les traducteurs cherchent le plus souvent des expressions idiomatiques, parmi les plus fréquentes on retrouve "in keeping with" (voir page précédente), "in light of" ou "look forward to". Les verbes et les adjectifs contenant une préposition ont aussi fait l'objet de multiples requêtes comme "consistent with" et "focus on".

ACTUALISATION DE TRANSSEARCH

L'équipe du RALI a cherché à améliorer TransSearch en tirant avantage des développements récents en informatique linguistique. Comme le souligne Guy Lapalme : «Comme la recherche des correspondances impliquait de modifier en

profondeur le système de gestion des textes, nous avons remodelé la conception du système pour prendre en compte les nouvelles possibilités d'interaction des navigateurs internet modernes. Après consultation avec notre partenaire industriel Terminotix, deux chercheurs ont développé une nouvelle interface qui permet à un usager d'interroger le système, à un gestionnaire de groupe de gérer les usagers et à un administrateur de manipuler l'ensemble des composantes du système.» La recherche de correspondances a donc incité l'équipe à changer non seulement le système de gestion de textes sous-jacent, mais les a aussi amenés à restructurer l'architecture du système.

En 2011, le système a été transféré à Terminotix, une entreprise montréalaise spécialisée dans les outils d'aide à la traduction. Terminotix a complété certains éléments essentiels tels que la localisation, la migration des usagers, la documentation et la formation des usagers. Le service est maintenant déployé pour des milliers d'usagers tant au pays qu'à l'étranger. Faisant le pont entre la R et D et l'application au monde réel, cette collaboration fructueuse entre les deux partenaires RALI et Terminotix a donné lieu à une entente pluriannuelle pour poursuivre leur coopération.

ÉQUIPE DE DÉVELOPPEMENT

Les chercheurs du RALI effectuent des études théoriques et pratiques dans le domaine du TALN. Les membres du RALI incluent le professeur Philippe Langlais, qui poursuit des études originales en morphologie et traduction automatique basées sur l'analogie, et le professeur Jian-Yun Nie, un expert mondial dans la recherche d'information translinguistique, domaine sur lequel il vient de publier un livre. Ils ont tous les deux de fortes équipes d'étudiants et de chercheurs.

De plus, le RALI participe régulièrement à des compétitions internationales dans ces domaines. Depuis sa création il y a plus de 15 ans, le RALI organise une série de séminaires hebdomadaires, où des invités et des étudiants présentent leurs recherches.

En plus des outils d'aide à la traduction présentés ici, le RALI a innové en développant un traducteur automatique de grande qualité pour les avertissements météorologiques émis par Environnement Canada.

TRANSCHECK

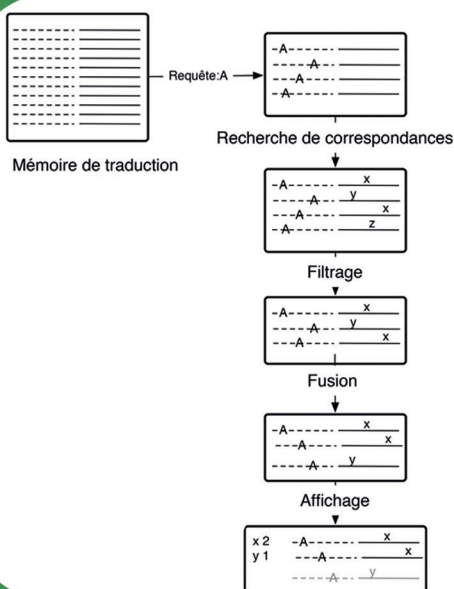
Un autre projet récent du RALI a été le développement du prototype TransCheck, qui comme son nom anglais le suggère, est un outil pour aider les usagers dans la détection des erreurs de traduction et dans la standardisation de la traduction. L'idée sous-jacente était d'aider à réduire le coût du contrôle de qualité en supportant le travail du réviseur humain. Le système aligne d'abord les phrases d'un texte source avec celles d'un brouillon de traduction et il s'assure que les paires de phrases alignées ne contiennent pas de correspondances incorrectes, mais qu'elles contiennent certaines correspondances souhaitables.

TransCheck effectue quatre vérifications principales:

- Terminologie négative: s'assurer que certains termes sont bien traduits par des équivalents

autorisés, leur liste pouvant être redéfinie pour chaque application ou texte

- Terminologie négative: à l'aide d'un dictionnaire de traductions découragées, indiquer les traductions suspectes
- Expressions numériques: une erreur fréquemment citée - par exemple, la mauvaise traduction d'une expression numérique, d'un numéro de pièce, d'une date ou d'un montant d'argent - peut être aussi vérifiée par TransCheck
- Omissions: surligner les paires de phrases quand les longueurs d'un original et de sa traduction diffèrent trop



LES GRANDS MOMENTS DU RALI

Plusieurs réalisations ont marqué l'histoire du RALI. Un exemple est l'outil de traduction interactive TransType, un système qui observe un traducteur qui tape son texte et lui suggère des complétions basées sur le texte précédent. À chaque caractère tapé par le traducteur, TransType révisé sa suggestion et fournit une nouvelle prédiction, comme le souligne Guy Lapalme: «La nouveauté réside dans le mode d'interaction entre l'utilisateur et le système de traduction automatique (TA) sous-jacent. Dans les essais précédents de traduction interactive, l'utilisateur devait aider le système à analyser le texte source afin d'améliorer la qualité de la traduction automatique finale; le fait de baser l'interaction sur le texte cible comme moyen d'interaction fournit un outil plus naturel et utile pour les traducteurs.

Durant le développement du système, l'équipe a rencontré plusieurs défis techniques. Par exemple, il était essentiel que le système limite ses suggestions à celles dans lesquelles il avait suffisamment confiance et qu'il adapte son comportement à l'information qu'il avait accumulée sur le traducteur humain. Traditionnellement, la TA n'analyse que le



Guy Lapalme sait bien qu'il reste beaucoup de chemin à faire, mais il est convaincu du besoin d'aides à la traduction

texte source alors que TransType analyse aussi la traduction déjà entrée. De plus, l'outil devait être capable de travailler en temps réel et fonctionner assez rapidement pour s'arrimer à la vitesse de l'utilisateur. Une autre difficulté était d'assurer que TransType puisse faire ses suggestions et prédictions clairement, mais sans déranger, en plus de fournir un affichage efficace et un moyen pour l'utilisateur de les accepter ou de les refuser.

TransType est le fruit d'efforts collaboratifs en vue d'améliorer la performance des traducteurs et, grâce au financement du Conseil de recherches en sciences naturelles et en génie (CRSNG) du Canada, du Ministère québécois des affaires internationales et de la Communauté européenne, nous espérons qu'il pourra être commercialisé comme un système efficace de suggestions aidantes.

AUTRES PROJETS ET RÉALISATIONS FUTURES

Le RALI a été impliqué dans plusieurs autres projets, en particulier dans le domaine du résumé automatique où il est un chef de file international depuis les travaux, à la fin des années 90 d'Horacio Saggion, alors étudiant au doctorat et actuellement un chercheur à Barcelone. Ils avaient alors développé SumUM qui comme le souligne Guy Lapalme : "produisait de courts résumés d'articles scientifiques longs en deux étapes: d'abord un résumé indicatif qui identifiait les sujets importants du document et ensuite un résumé informatif qui élaborait sur les sujets choisis par l'utilisateur." Depuis trois ans, en collaboration avec son étudiant au doctorat Pierre-Étienne Genest, Guy Lapalme a développé une toute nouvelle méthode de résumé par abstraction qui diffère de la majorité des autres méthodes actuelles qui sont toutes orientées vers l'extraction de phrases.

Plus récemment, Lapalme a consacré beaucoup de temps et d'intérêt à l'analyse d'opinion et de sentiments dans les textes trouvés sur Internet. Avec l'aide de sa stagiaire postdoctorale Marina Sokolova, maintenant professeure à l'Université d'Ottawa, Guy Lapalme a développé une nouvelle méthode pour analyser les opinions en étudiant les patrons d'utilisation des mots non émotionnels. Les études précédentes avaient surtout porté sur la distribution des mots, en général les adjectifs, associés à des sentiments positifs ou négatifs. Cette nouvelle méthode est intéressante parce qu'elle permettrait de traiter des documents de plusieurs domaines différents comme la médecine ou l'assurance, domaines où on ne retrouve que peu d'expressions d'émotion.

Jusqu'ici, les travaux du RALI ont surtout visé les traducteurs professionnels. Toutefois, le groupe voudrait aussi explorer d'autres types d'utilisateurs, tels les travailleurs du savoir et les interprètes qui doivent composer des articles et des documents en deux langues. Les études et les expériences ont été couronnées de succès, même si des recherches supplémentaires devront être consacrées à établir comment les divers besoins d'autres types d'utilisateurs pourraient être intégrés aux systèmes existants.

Guy Lapalme sait bien qu'il reste beaucoup de chemin à faire, mais il est convaincu du besoin d'aides à la traduction: "À l'avenir, nous voulons continuer à améliorer nos outils informatiques en combinant l'expérience acquise lors de nos différents projets. Par exemple, maintenant que TransSearch peut identifier des traductions cibles, nous aimerions l'intégrer dans une interface à la TransType qui l'interrogerait automatiquement pour trouver des traductions intéressantes et les afficherait comme suggestions au traducteur".

FAITS SAILLANTS

RECHERCHE DE CORRESPONDANCES DANS TRANSSEARCH

OBJECTIFS

- cibler l'identification de traductions pour afficher la correspondance entre les mots de la phrase source et leurs équivalents dans la phrase cible
- faciliter le travail des traducteurs et élargir la gamme d'applications de TransSearch sur le web et sur la station de travail
- permettre au Canada de garder sa position de chef de file dans le domaine des outils pour la traduction

COLLABORATEURS PRINCIPAUX

Les professeurs Philippe Langlais; Jian-Yun Nie; le chercheur Fabrizio Gotti; Université de Montréal

FINANCEMENT

Conseil de recherches en sciences naturelles et en génie du Canada - Fonds de recherche du Québec - Nature et technologies

CONTACT

Guy Lapalme, professeur
Chercheur principal

Département d'informatique
et de recherche opérationnelle
Université de Montréal
CP 6128, Succ Centre-Ville
Montréal
Québec, H3C 3J7
Canada

T +1 514 343 6111 X 47493
E lapalme@iro.umontreal.ca

<http://rali.iro.umontreal.ca>

GUY LAPALME est professeur en informatique à l'Université de Montréal depuis 1980. Il est un expert mondial du traitement automatique de la langue naturelle. Il a publié sur plusieurs aspects de ce domaine incluant la correction d'orthographe, l'édition de dictionnaires, la génération de texte, le résumé automatique, l'extraction d'information, la fouille d'opinion et les outils d'aide à la traduction. Il a aussi travaillé en recherche opérationnelle, les compilateurs et la bioinformatique. Sa carrière est un heureux mélange de recherche innovatrice et de liaison avec le monde pratique par des collaborations à long terme avec des partenaires de diverses provenances, tant académiques qu'industrielles. Il a récemment reçu un doctorat honoris causa de l'Université de Neuchâtel (Suisse) et un Prix d'excellence pour l'ensemble de ses réalisations de la part de CAIAC, l'association de l'intelligence artificielle au Canada.

TERMINOTIX

Université 
de Montréal


rali