Les voies d'amélioration des outils de traitement de l'écrit

Je travaille pour un groupe de recherche appelé le RALI, pour « Recherche appliquée en linguistique informatique ». Le RALI mène des travaux dans plusieurs domaines, mais surtout dans celui de l'écrit. La particularité du laboratoire, c'est qu'il examine la question d'un angle très *ingénieur,* avec la volonté de faire des applications pratiques.

Je parlerai aujourd'hui d'approches et de voies de recherche susceptibles de favoriser l'amélioration des outils de traitement de l'écrit. Avant de commencer, il est cependant utile de parler un peu des techniques employées actuellement.

Dans certains cas, les spécialistes traitent la langue écrite en recourant à des approches symboliques, lesquelles supposent l'expression des connaissances à l'aide de règles. Les approches symboliques présentent un avantage important : elles permettent de travailler sur des bases très solides. En effet, les informaticiens n'ont pas inventé le traitement des langues ; il y a longtemps que les grammairiens ont développé une connaissance fine de celles-ci.

Les approches symboliques comportent cependant un désavantage important : il faut faire de l'adaptation manuelle. Dans bien des cas, cette stratégie est malgré tout la plus efficace qui soit. Par exemple, on sait que les bons correcteurs de fautes d'orthographe fonctionnent essentiellement sur la base d'approches symboliques.

À côté des approches symboliques, il existe des approches probabilistes. Le recours à ces dernières est fructueux quand on manque de connaissances et qu'il n'est pas requis d'articuler des règles pour permettre à l'ordinateur de faire le travail demandé. Par exemple, il n'est pas nécessaire, le plus souvent, d'utiliser une approche symbolique pour aider la machine à établir la langue d'un texte, de manière à déterminer quel correcteur ou traducteur devrait être employé pour le corriger ou le traduire. Il est plus simple de fournir une masse de textes à l'ordinateur et de lui dire : « Voici de l'anglais », « Voici du français » ou « Voici du chinois ».

Dans les années à venir, il faudra perfectionner ces approches pour améliorer les capacités de la machine en matière d'extraction, de synthèse et de traduction de l'information.

Extraction de l'information

On dispose actuellement de bons outils pour retracer l'information dans des bases de données. Cependant, l'information requise *ne se trouve pas toujours* dans les bases de données. Elle réside souvent dans des textes ordinaires. Par exemple, dans les hôpitaux, elle se trouve dans les rapports des médecins ; dans le domaine juridique, elle réside dans les jugements des juges ; etc.

Quand l'information se trouve dans des textes normaux, on est encore relativement mal équipés pour la repérer. Actuellement, ce que l'on a sous la main, ce sont essentiellement des outils de recherche par mot-clé. Ces outils sont performants. Cependant, s'ils permettent fréquemment d'aller chercher le bon document, ils ne permettent pas de dépasser le niveau de la perception et d'en comprendre le sens.

Le RALI travaille actuellement aux questions d'extraction d'information. Plus particulièrement, il s'intéresse de près à l'analyse des renseignements contenus dans les courriels que les clients d'une entreprise envoient à cette dernière. Cela pourra permettre d'y percevoir des tendances.

> « L'extraction d'information consiste à identifier de l'information bien précise d'un texte en langue naturelle et à la représenter sous forme structurée. Par exemple, à partir d'un rapport sur un accident automobile, un système d'extraction d'information sera capable d'identifier la date et le lieu de l'accident, le type d'incident ainsi que les victimes. Ces informations pourront ensuite être stockées dans une base de données pour y effectuer des recherches ultérieures ou être utilisées comme base à la génération automatique de résumés ».
>
> Source : Page Web du RALI

Résumés

La production automatique de résumés présente un grand intérêt commercial. En effet, comme nous sommes submergés d'information, il nous est difficile de faire un tri rapide entre les textes intéressants qui nous tombent sous la

main et les autres. Les résumés peuvent aider les lecteurs à vite se faire une idée du propos d'un document et, par conséquent, à gagner un temps précieux.

Les chercheurs travaillent beaucoup, actuellement, sur les questions de synthèse d'articles de journaux. Il s'agit d'un domaine d'application où le recours aux technologies donne de bons résultats, parce que la machine reçoit un bon coup de main du rédacteur. En effet, les journalistes sont entraînés à résumer leur article dans son premier paragraphe…

Cela dit, les chercheurs commencent à s'intéresser à des domaines beaucoup plus spécialisés et difficiles que celui de l'écrit journalistique. Par exemple, ils travaillent actuellement à produire des résumés automatiques d'articles scientifiques. Cette question présente un intérêt certain, même si les rédacteurs de ce genre de documents en font déjà un résumé. En effet, l'auteur est la pire personne disponible pour faire un bon résumé. Le lecteur d'un texte, en effet, y cherche généralement quelque chose de différent de ce que le rédacteur voulait y mettre. Par ailleurs, la demande pour des résumés de textes juridiques est aussi élevée.

Les chercheurs ont aussi commencé à mettre au point des résumeurs capables de s'attaquer automatiquement à *plusieurs* documents pour en faire la synthèse. Par exemple, un outil de ce type pourrait appréhender plusieurs textes journalistiques publiés sur une même question et produire automatiquement une chronologie.

La génération de manchettes journalistiques est une question qui intéresse également les chercheurs. Ces derniers tentent de produire l'en-tête d'un article à partir de son contenu.

Traduction

La question de la traduction automatique a donné naissance au domaine de la linguistique informatique. Après des années d'hibernation, elle reprend actuellement de la vigueur, surtout aux États-Unis. Pour assurer la sécurité du territoire, les Américains veulent mieux comprendre ce que les non-anglophones disent et trament. Je reviens d'ailleurs de la conférence nord-américaine de l'*Association for Computational Linguistics*. Pendant des années, personne n'y a parlé de traduction ; cette année le quart des publications portaient sur ce sujet.

Les logiciels de traduction automatique fonctionnent plus ou moins bien actuellement. S'ils sont utiles, c'est essentiellement lorsqu'on désire avoir une idée très approximative d'un document produit en langue étrangère.

Les travaux sur la traduction assistée par ordinateur (TAO) paraissent plus prometteurs. De notre côté, nous travaillons à la création de ce genre d'outils en recourant à des approches statistiques. L'idée, quand on emploie cette méthode, est de partir de grands corpus. D'un côté, on a le texte source ; de l'autre, sa traduction. Cette mise en parallèle permet de dégager des règles et des constantes qui permettent de produire de nouvelles traductions ou de créer des outils d'aide à la traduction.

L'illustration 1 présente un système que nous avons développé en laboratoire : TSRali. TSRali permet de trouver la traduction française juste d'expressions en anglais et *vice versa*. Le logiciel vise à pallier les faiblesses de la traduction par chaînes de caractères. Par exemple, on ne peut traduire l'énoncé « tirer à boulet rouge sur quelqu'un » par « *to pull red bullets on someone* ». Cependant, c'est ce que font les logiciels de traduction par chaînes de caractères.

TSRali fonctionne plutôt en s'appuyant sur une mémoire de traduction, une mémoire constituée à l'aide des traductions humaines réalisées pour constituer le *Journal de la Chambre des communes* du Canada. En utilisant TSRali, le traducteur apprend qu'il est par exemple possible de traduire « tirer à boulets rouges » par « fire on ».

Le RALI mène aussi le projet Transtype. Transtype est un aide à la traduction interactif. Il suffit au traducteur de commencer à taper pour se voir proposer des traductions en temps réel.

Éventuellement, il serait intéressant que la machine puisse établir elle-même le niveau de confiance qu'elle a dans une traduction. L'ordinateur suspicieux devant sa propre performance pourrait alors faire part de ses doutes à l'utilisateur. La machine confiante pourrait faire son travail automatiquement, sans demander l'intervention de l'utilisateur. Toutefois, en attendant la création d'un tel mécanisme d'autocritique, le traducteur a le contrôle. Lorsqu'une traduction fait son bonheur, il l'accepte ; sinon il continue à taper.

Dans un domaine connexe, nous travaillons actuellement aux questions de repérage translingue.

Autrement dit, nous essayons de créer des outils qui permettront aux usagers de produire une requête dans une langue – par exemple en français – et de repérer des textes dans une autre – par exemple en anglais. On ne parle pas ici de traduire la requête – souvent celle-ci est trop ambiguë –, mais bien de la jumeler à un texte dans une autre langue. Ce genre d'applications est intéressant, parce qu'il arrive souvent qu'une personne ait seulement une connaissance passive d'une autre langue.

« Après avoir confondu bien des sceptiques, le logiciel de traduction Transtype, mis au point au [RALI] passe la vitesse supérieure et devient une affaire internationale. Transtype 2 sera [...] bientôt utilisé non seulement par les traducteurs pour passer de l'anglais au français et inversement, mais aussi par ceux qui travaillent avec l'espagnol ou l'allemand.

« Notre premier logiciel a montré que la traduction assistée par ordinateur était possible ; la seconde phase permet d'associer des partenaires européens à notre aventure et d'étendre les possibilités de l'appareil », a dit le professeur Guy Lapalme à l'occasion du lancement de Transtype 2, le 29 septembre dernier, peu avant la première réunion des partenaires en sol canadien.

C'est grâce à un financement du Conseil de recherches en sciences naturelles et en génie du Canada et à la participation financière d'un consortium européen que ce nouveau départ est possible. Au Canada, ce sont 750 000 $ qui seront investis en trois ans dans ce projet, permettant la création d'une vingtaine d'emplois, alors que les partenaires européens, universitaires et industriels, y injecteront quelque 2,6 M$. Au total, l'équipe canado-européenne recevra 3,4 M$.

Mathieu-Robert Sauvé, « Les logiciels de traduction passent à la deuxième génération », *Forum,* 20 octobre 2003, consulté le 5 novembre 2003 à l'adresse http://www.iro.umontreal.ca/~lapalme.

Besoins

Différentes mesures permettraient de favoriser les travaux francophones dans le domaine de l'écrit. Premièrement, il serait important d'accentuer nos efforts en matière d'évaluation.

On y arrive parfois assez bien. Par exemple, en matière de recherche d'information, il n'est pas si difficile de savoir si les documents retracés par l'ordinateur sont pertinents ou non. Il suffit essentiellement de tenir compte des mots-clés employés à des fins de repérage, puis de compter le nombre de documents pertinents et non pertinents repérés pour avoir des statistiques en matière de précision et de rappel.

Dans d'autres domaines, mener des évaluations sera toutefois plus difficile. À titre d'exemple, les choses se gâtent lorsqu'on tente de mesurer la performance d'un système de traduction. En effet, il est difficile de déterminer avec exactitude ce qu'est une bonne ou, encore, une mauvaise traduction. Deux traductions peuvent être bien différentes l'une de l'autre – sur le plan de la longueur notamment – et être de qualité égale.

Actuellement, les gens qui travaillent au traitement du français écrit n'ont pas, souvent, pour priorité de se comparer les uns aux autres. Il faudrait corriger cette situation. Les résultats seront parfois rassurants. Après tout, comme le dit le dicton : « Quand je me regarde, je me désole ; quand je me compare, je me console. »

Deuxièmement, il faudra mettre au point des matériaux langagiers.

D'abord, il ne fait pas de doute que la question des corpus est importante dans le secteur de l'écrit. Pendant longtemps, il a été difficile de mettre la main sur des corpus représentatifs. Les textes numérisés, en effet, étaient rares, en partie parce qu'il était difficile de saisir les textes papier. Maintenant le problème a changé, puisque les textes sont généralement créés directement à l'aide de la machine. L'arrivée du Web a aussi facilité l'accès à des documents intéressants.

La question du codage des textes demeure toutefois. Ainsi, on commence à s'entendre sur l'utilisation d'Unicode pour représenter les documents écrits, mais tous les documents ne sont pas encodés selon cette norme. De même, XML, standard de balisage, suscite beaucoup d'intérêt, mais il ne suffit pas de mettre des balises dans un document : il faut s'entendre sur leur signification et les former adéquatement.

Sur le plan des dictionnaires électroniques, il existe aussi des problèmes. Ainsi, il est sûr qu'il en existe en français. Cependant, ces outils sont moins disponibles en

français qu'ils ne le sont en anglais. Le problème vient de ce que les organismes qui les ont développés, souvent, refusent de les partager ou les vendent à un prix déraisonnable, comme un dollar ou un euro le mot. Si les gens ne jurent que par *WordNet,* dictionnaire anglophone, c'est parce que *WordNet* est relativement gratuit. *WordNet* a plein de défauts, mais il est bien fait, utile et d'accès très économique.

Je n'ai pas toutes les réponses aux problèmes de matériaux langagiers, mais, selon moi, une avenue intéressante pour les francophones serait de travailler de manière distribuée à la création de dictionnaires ou d'ontologies. Il faut absolument profiter de ce qu'Internet nous permet de collaborer les uns avec les autres. Il existe un précédent à cela: le projet *Papillon* vise en effet la création d'un dictionnaire orienté traduction. Chaque personne peut contribuer à la création de l'outil en fournissant certaines entrées. Ces dernières sont ensuite validées à l'aide d'un procédé de validation bien robuste.

Conclusion

En conclusion, des progrès majeurs ont été enregistrés dans le domaine de l'écrit ces dernières années. Cependant, beaucoup de travail reste encore à faire, particulièrement en langue française.

Guy Lapalme,
professeur-chercheur, centre RALI, Université de Montréal, Québec.

Illustration 1
Le système TSRALI

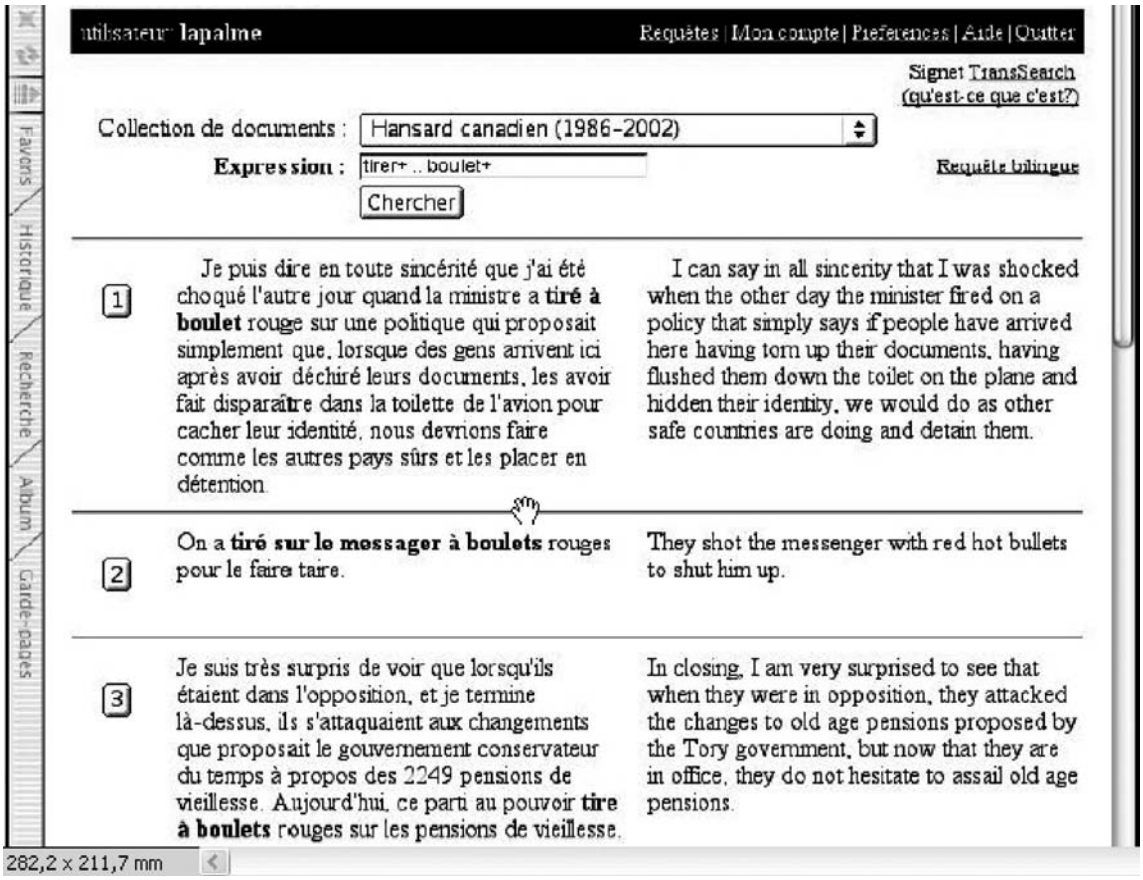

Illustration 2
TransType