

Présentation

Linguistique et accès automatisé à l'information : un bilan

Anne Condamines, CNRS-Université de Toulouse-Le Mirail

Thierry Poibeau, CNRS-Université Paris 13

So we need to be alert. It's not just that we may find ourselves putting the cart before the horse. We can get obsessed with the wheels, and finish up with uncritically reinvented, or square, or over-refined or otherwise unsatisfactory wheels, or even just unicycles. What matters is the way the cart, its load, and the horse, together make a rational journey (...), it's about how you integrate the language and the computation on the one hand, the symbolic and the statistical on the other and to succeed in this we should not forget that mainstream linguistics may have some things to offer us, even if not as many as linguists themselves may suppose.

Donc, nous devons être en alerte. Pas simplement parce que nous pourrions nous retrouver à mettre la charrue avant les bœufs. Nous pouvons devenir obsédés par les roues et finir par réinventer sans fin des roues carrés, ou trop perfectionnées, ou inadéquates, ou juste des monocycles. Ce qui compte, c'est la façon dont la charrue, les bœufs et le chargement font ensemble un trajet cohérent (...). C'est la façon dont on intègre langage et calcul d'une part, symbolique et statistique de l'autre. Et pour réussir en cela, nous ne devons pas oublier que la linguistique générale peut avoir des choses à nous offrir, même si ce n'est pas tout ce que les linguistes eux-mêmes peuvent imaginer.

Karen Spärck Jones, *Computational Linguistics: What about the Linguistics?*¹

Ce numéro de la *Revue Française de Linguistique Appliquée* concerne les rapports entre linguistique et recherche d'information. Au-delà de la recherche d'information au sens strict (recherche de documents pertinents par rapport à des mots-clés), nous souhaitons élargir la problématique à des domaines de recherche connexes :

- L'extraction d'information,
- Les systèmes de question/réponse,
- Le résumé automatique,
- La catégorisation de textes.

La dimension langagière est présente dans toutes ces applications : la plupart exigent une analyse du contenu (on parle aujourd'hui de moteur de recherche « sémantique ») et ne peuvent se contenter de voir le document comme un simple « sac de mots ». La linguistique devrait donc jouer un rôle important dans ces disciplines. Or, ce n'est pas toujours le cas.

¹ Karen Spärck Jones a été une des principales figures de la recherche d'information des années 1960 à aujourd'hui. Ces quelques mots sont extraits de son dernier article, écrit en guise de bilan et publié dans *Computational Linguistics* (Vol. 33-3, 437-441) quelques mois après sa mort en 2007.

Des relations complexes

L'histoire commune de la recherche d'information et de la linguistique a plus de cinquante ans. Si, à l'origine, la convergence d'intérêt a paru évidente, les techniques et les besoins évoluant ont amené à des rapports complexes et parfois antagonistes entre ces disciplines.

L'augmentation exponentielle des données textuelles disponibles sur support numérique – on pense évidemment à l'Internet mais il ne faut pas oublier les immenses bases de données littéraires, scientifiques ou techniques – a favorisé le développement, ces dernières années, d'approches simples et efficaces. Dans ce contexte, on s'est longtemps interrogé sur la plus-value apportée par les connaissances linguistiques : même l'apport de techniques simples comme l'élargissement de requêtes a été discuté, voire contesté. Enfin, les soucis applicatifs ont laissé dans l'ombre la réflexion sur la langue et la modélisation linguistique.

Il se peut que la linguistique n'ait plus sa place dans une grande partie des travaux sur l'accès à l'information. Il nous a toutefois semblé nécessaire de revenir sur cette idée préconçue, afin d'observer la situation de plus près et de dresser un état des lieux des rapports entre linguistique et recherche d'information.

Des liens qui se renouvellent

Cette idée d'état des lieux tient au fait que l'on observe souvent un sentiment de frustration chez les linguistes face à la recherche d'information : le développement de ressources est long, coûteux, et n'apporte pas toujours un gain de performance. Toutefois, plusieurs éléments poussent à nuancer ce constat :

- Certes, la partie visible de l'iceberg (à savoir les moteurs de recherche d'information sur le web) fait largement abstraction de connaissances linguistiques. Mais la nécessité de répondre à des requêtes précises, notamment dans les domaines techniques, oblige à des traitements plus fins, remettant au premier plan la question du sens dans les moteurs de recherche spécialisés.
- Les méthodes robustes, même si elles sont hégémoniques, butent sur les nouveaux modes d'expression et les nouveaux médias – mails, blogs, wikis, *etc.* Il devient nécessaire de mieux prendre en compte les variations du fonctionnement langagier : les travaux qui prennent en compte la variation linguistique, en fonction notamment du genre textuel, sont en plein essor.
- Enfin, et peut-être surtout, la réflexion sur le sens en linguistique peut trouver à se nourrir des évolutions qui ont lieu dans les méthodes d'accès à l'information.

Au-delà des questions qui se posent sur les relations entre sens et information, se trouve réactualisée la question de la contextualisation du sens. Le rôle de la situation dans la construction du sens est revendiqué par les analystes de discours, les ethnologues de la communication, les sociolinguistes... Par situation, on entend le contexte dans lequel se déroule la communication, aussi bien orale qu'écrite (connaissances supposées partagées des interlocuteurs, référence à la situation de communication...). Mais ne pourrait-on imaginer que l'objectif de la construction des ressources (pour le Traitement Automatique des Langues en l'occurrence), c'est-à-dire le point de vue que l'on pose sur les ressources en fonction des besoins et non seulement par rapport au fonctionnement intra-linguistique, puisse interpeller la linguistique et l'amener à ouvrir son champ d'étude ?

Il semblerait qu'après une dizaine d'années où la recherche d'information au sens classique (recherche de documents) a été au premier plan, les applications se focalisent davantage maintenant sur l'analyse du contenu et des conditions d'énonciation. Les systèmes entièrement automatiques, utilisant peu de ressources linguistiques, montrent rapidement leurs

limites dans ce cadre. On peut donc faire le pari d'un « retour du linguistique » pour l'accès à l'information, mais celui-ci ne pourra pas se faire sans une profonde réévaluation, voire une adaptation des méthodes de travail des linguistes eux-mêmes.

Contenu de ce numéro

Ce numéro vise à présenter des travaux collaboratifs ayant été menés récemment et une réflexion sur le rôle qu'ont pu y jouer les études linguistiques. Il s'agit de dresser un panorama de la situation du point de vue de la linguistique afin d'identifier les domaines dans lesquels l'apport de la linguistique est patent, ceux dans lesquels la linguistique n'a pas (ou plus) sa place, mais aussi les nouveaux domaines où l'apport de la linguistique est réel mais plus inattendu. En effet, les travaux collaboratifs, même s'ils paraissent décevants du seul point de vue de l'efficacité, peuvent ouvrir des pistes de réflexion nouvelles pour les linguistes, renouvelant le rapport avec les techniques de traitement automatique.

En d'autres termes, nous avons souhaité mettre en évidence en quoi la linguistique se trouve interrogée et renouvelée par les besoins et méthodes qui ont récemment émergé dans la recherche d'information. Il s'agit donc aussi de déplacer la problématique de l'évaluation en recherche d'information en ne l'examinant plus seulement du point de vue de l'adéquation des résultats avec la demande mais en interrogeant ces résultats pour leur donner un sens du point de vue du fonctionnement de la langue.

Si on les examine d'un point de vue linguistique, les articles s'organisent autour d'interrogations que l'on peut présenter ainsi.

- Comment caractériser la nature des connaissances mises en œuvre dans les projets d'accès à l'information ? Le niveau du mot est-il suffisant ou faut-il passer au syntagme, voire au texte (contribution de Valette et Slodzian) ? Des notions plus sémantiques comme celles de schémas prédicatifs ou de paraphrase ont-elles leur place ? La modélisation de séquences de mots sous forme de trigrammes permet ainsi d'améliorer la recherche d'information (contribution de Lioma et Van Rijsbergen), tandis que la notion de paraphrase est primordiale pour les systèmes de question-réponse (contribution de Zweigenbaum & *al.*) ou de résumé (contribution de Saggion).
- On peut aussi s'interroger sur le type de ressources linguistiques à utiliser en fonction de la tâche.
 - Les ressources construites *a priori* (lexiques, terminologies, ontologies) sont omniprésentes dans les articles du numéro, à l'image de *Wordnet* pour l'anglais (voir la contribution de Vossen par exemple). A l'inverse, les ressources par domaines de spécialité sont peu nombreuses et peu riches (L'Homme étudie ainsi trois ressources différentes concernant le domaine informatique).
 - Les ressources endogènes construites automatiquement à partir de corpus et éventuellement révisées par des linguistes sont intéressantes même si leur productivité semble insuffisante pour réellement améliorer la recherche d'information (Picton & *al.* présentent l'intérêt mais aussi les limites d'une telle approche).
 - La mise au point des ressources peut enfin exiger une étude minutieuse pour obtenir une catégorisation très précise et non thématique (l'article de Valette et Slodzian présente une expérience de ce type pour le filtrage de textes racistes, qui utilise le même vocabulaire thématique que les textes anti-racistes).
- Comment évaluer les résultats et comment évaluer l'apport de la linguistique ? La question est parfois difficile, surtout quand la tâche comporte une part de subjectivité. Ainsi, Saggion présente différents types de résumés et la grande variation dans leur appréciation

d'un individu à l'autre. Picton & *al.* s'interrogent sur la façon de déterminer les documents pertinents dans une perspective de recherche d'information.

L'article de Christina Lioma et Keith van Rijsbergen vise à améliorer la recherche d'information, au sens classique du terme. Les auteurs utilisent l'analyse morphosyntaxique associée à des indications statistiques pour repérer des séquences pertinentes pour l'indexation. Ces recherches montrent que l'intérêt de la linguistique pour la recherche d'information n'est pas limitée à l'analyse de textes appartenant à des domaines spécialisés.

L'article de Piek Vossen décrit quant à lui trois expériences intégrant des ressources linguistiques pour améliorer la recherche d'information. L'auteur montre que la nature des ressources utiles dépend de la complexité des applications visées (recherche d'information, extraction de terme ou interfaces dialogiques). Ces expériences sont menées dans un cadre industriel et permettent d'entrevoir les applications de demain.

L'article de Pierre Zweigenbaum & *al.* donne un aperçu des recherches en question-réponse et des ressources linguistiques mises en œuvre dans ce cadre. Ce type de système nécessitant une véritable analyse du contenu textuel (la représentation du document comme un « sac de mots » étant ici impossible), des informations de nature morphologique, syntaxique et sémantique, voire discursive ou pragmatique, sont nécessaires. L'article décrit la complexité de ces différents éléments, ainsi que leur interaction au sein d'un système dédié.

Horacio Saggion propose un panorama des travaux réalisés en résumé automatique. Il s'interroge sur la nature des informations à inclure dans un résumé et sur la façon d'évaluer les résumés produits. Il présente les outils et les méthodes mis en œuvre en mettant l'accent sur la nature des connaissances sémantico-syntaxiques utilisées. Son diagnostic est optimiste quant à l'avenir des recherches en linguistique pour le résumé automatique.

Aurélien Picton, Cécile Fabre et Didier Bourigault s'intéressent à la possibilité de prendre en compte des ressources endogènes (construites par analyse distributionnelle à partir de gros corpus) pour améliorer la recherche d'information. La méthode d'analyse permet de prendre en compte un plus grand nombre de relations sémantiques. Sans être aussi bons que ce qui était espéré par les auteurs, les résultats s'avèrent toutefois prometteurs.

L'article de Marie-Claude l'Homme évalue comment trois ensembles de ressources (base de données lexicales, ontologie de domaine et ressources terminologiques) rendent compte du fonctionnement d'un choix de termes issus du domaine de l'informatique (sélectionnés grâce au logiciel *TermoStat*). Il s'intéresse plus particulièrement à deux candidats : *file* et *store*. Les analyses portant sur la présence des termes, seuls ou en syntagmes, mais aussi sur les liens avec les autres termes, montrent qu'aucun des trois types de ressources ne propose un traitement satisfaisant.

Enfin, Mathieu Valette et Monique Slodzian s'intéressent à la catégorisation de textes. Après le constat des insuffisances dans l'apport des connaissances lexicales ou même phrastiques dans ce type de perspective, les auteurs proposent de mettre en œuvre une sémantique textuelle ; ils s'appuient sur l'approche en trois niveaux de Rastier (microsémantique, mésosémantique et macrosémantique). L'exemple présenté est celui du repérage des textes racistes.

Anne Condamines, <anne.condamines@univ-tlse2.fr>

Laboratoire CCLE, 5 allées Machado, 31058 Toulouse

Thierry Poibeau, <Thierry.Poibeau@lipn.univ-paris13.fr>

Laboratoire d'Informatique de Paris-Nord, 99, avenue Jean-Baptiste Clément, 93430 Villetaneuse