Traitement des questions de type liste dans un système de question-réponse

Domaine : Traitement Automatique des Langues
Début : fin 2010
École Doctorale Informatique Paris-Sud
Laboratoire : LIMSI (Orsay), groupe ILES
Encadrant : Anne VILNAT
Co-encadrant : Véronique MORICEAU
Contact : moriceau@limsi.fr

Pré-requis : Connaissances en Traitement Automatique des Langues (français, anglais)

Description :
Dans le domaine de la recherche d'information, l'un des défis actuels porte sur la détermination de l'information précise recherchée par un utilisateur. Cette focalisation sur la recherche précise d'information s'est concrétisée ces dernières années par un intérêt porté aux systèmes de questions-réponses en domaine ouvert. L'objectif de ces systèmes est de fournir une réponse précise à une question exprimée en langue naturelle en trouvant cette réponse dans un ensemble de documents, ensemble éventuellement très large et pouvant aller jusqu'au Web. Par exemple, à la question "À quelle date est mort Henri IV ?", la réponse renvoyée doit être "1610" (et non pas des pages contenant la réponse, comme c'est le cas des moteurs de recherche classiques).
A l'heure actuelle, les systèmes de question-réponse traitent surtout les questions dites factuelles (qui attendent comme réponse un fait, une entité nommée) mais très peu les questions complexes. Parmi ces questions complexes, on trouve notamment les questions en « pourquoi », « comment » et les questions de type liste. Ces dernières attendent comme réponse une liste d'items.

Les questions de type liste ont été introduites pour la première fois en 2001 lors de la campagne d'évaluation TREC. Pour répondre correctement à ces questions, les systèmes doivent renvoyer une liste complète d'items, ces items devant se trouver dans une même phrase. Il est évident que rechercher tous les items différents répondant à une question liste afin d'avoir une réponse exhaustive implique de ne pas ignorer les réponses les moins fréquentes : il est alors difficile d'appliquer de simples techniques basées sur la redondance des réponses.
Les approches existantes imposent par exemple de connaître le nombre d'items attendus (Harabagiu et al., 2001), par exemple *Donnez le nom de 5 aéroports européens* ; d'autres se contentent de traiter ces questions comme des questions factuelles classiques et renvoient les *N* premières réponses dont le score est supérieur à un seuil fixé (Zhou et al., 2006), (Whittaker et al., 2006). A partir de 2007, la campagne d'évaluation INEX (sur des documents structurés) a mis en place la tâche *Entity Ranking* qui consiste à renvoyer les entités qui traitent d'un thème : dans cette campagne, les questions ne sont pas de réelles questions en langue naturelle mais des *topics* sous forme de mots-clés. D'autres approches utilisant des techniques d'apprentissage extraient les réponses à une question liste soit grâce à une classification automatique préalable des types de réponse (Wang et al., 2008), soit grâce à une classification des documents selon les types d'entités nommées qu'ils contiennent (Yang & Chua, 2004).
Toutes ces approches nécessitent soit un pré-traitement de la collection de documents, soit d'avoir des formes contraintes de questions.

La thèse proposée se situe dans le cadre du système de questions-réponses FIDJI développé pour le français et l'anglais et qui traite des questions factuelles et complexes, en combinant des informations d'ordre syntaxique et des techniques "classiques" du domaine, telles que la reconnaissance des entités nommées et la pondération des termes de la question. FIDJI ne nécessite

pas de pré-traitement de la collection de documents, condition nécessaire quand on s'intéresse au Web notamment. Il s'agira donc d'étudier les approches pour :

- **L'analyse des questions de type liste.** Il s'agit d'identifier ces questions en tant que telles afin de déclencher les stratégies appropriées pour la recherche des réponses . Les questions peuvent avoir des indices explicites (par exemple, "*Quelles sont...*", "*Qui sont...*", "*Donnez les 7 Merveilles du Monde ?*", etc.) ou non (par exemple, *"Qui a découvert la comète Shoemaker-Levy ?*" où seule la découverte des réponses permet de déduire que la question est de type liste).
- **La recherche des réponses.** Les différents éléments de réponse à une question liste peuvent être extraits d'une même phrase, d'un même document ou de documents différents. Un des objectifs est aussi de rechercher les réponses sur le Web où les réponses peuvent être présentées sous des formes variées : listes à puce, tableaux, etc.
- **La présentation des réponses.** Il s'agit d'étudier les différentes façons de présenter les réponses à un utilisateur une fois les éléments de réponse assemblés (agrégation de résultats). Cette présentation pourra se faire par exemple sur des critères spatio-temporels (par exemple, à la question "*Donnez la liste des présidents français ?*", les réponses pourront être présentées dans l'ordre chronologique) ou autres.
- **L'évaluation des réponses.** Plusieurs points sont à évaluer et des méthodologies d'évaluation doivent être définies :
    - Le système a-t-il bien répondu ? : comment juger qu'une réponse liste est correcte ? Exhaustive ? En effet, une réponse liste peut être constituée :
        - de plusieurs éléments corrects et tous différents,
        - de plusieurs éléments exprimés de façon différente mais faisant référence à la même entité ("*Nicolas Sarkozy*" et "*le président français*"),
        - de plusieurs éléments même si la question n'est a priori pas de type liste (par exemple, pour la question "*Quelle est la hauteur de la Tour Eiffel ?*", les réponses peuvent être "*309 mètres sans l'antenne*" *et "324 mètres avec l'antenne*"),
    - Les utilisateurs sont-ils satisfaits des réponses fournies ?

Références :

Y. Zhou, X. Yuan, J. Cao, X. Huang, and L. Wu. FDUQA on TREC2006 QA Track. In Proceedings of the 15th Text Retrieval Conference (TREC-15), Gaithersburg, USA, 2006.

E. Whittaker, J. Novak, P. Chatain, and S. Furui. TREC2006 Question Answering Experiments at Tokyo Institute of Technology. In Proceedings of the 15th Text Retrieval Conference (TREC-15), Gaithersburg, USA, 2006.

S. Harabagiu, D. Moldovan, M. Pasca, M. Surdeanu, R. Mihalcea, R. Girju, V. Rus, V. Finley Lacatusu, P. Morarescu, R. Bunescu: Answering Complex, List and Context Questions with LCC's Question-Answering Server. TREC 2001.

H. Yang, T. Chua. Web-based list question answering. In Proceedings of the 20th international conference on Computational Linguistics, Geneva, Switzerland, 2004

R. Wang, N. Schlaefer, W. Cohen, E. Nyberg. Automatic set expansion for list question answering. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Honolulu, Hawaii, 2008.