Les moteurs de recherche multilingues

La multiplication des documents, notamment sur le Web, dans de nombreuses langues a rendu nécessaire l'existence d'une recherche documentaire cross-lingue (appelée en anglais CLIR : *Cross Language Information Retrieval*). La recherche cross-lingue consiste à formuler une requête dans une langue source et à rechercher des documents pertinents dans des langues cibles. Le défi consiste donc à créer un trait d'union entre une requête dans une langue et les documents qui y répondent dans une ou plusieurs autres langues. Il existe plusieurs types de moteurs de recherche cross-lingue :

A. Les moteurs de recherche basés sur la traduction automatique

Cette approche (en anglais *Machine Translation method* ou *MT-based method*) s'appuie sur la traduction automatique de la requête ou du corpus de document.
Il s'avère que la traduction de la requête présente moins de précision que celle de la collection de documents (Oard et Hackett, 1997), qui contiennent un contexte d'information nettement plus important, ce qui diminue les risques de mauvaise traduction. Cependant, l'application de cette méthode consistant à traduire tous les documents dans toutes les langues désirées est trop compliquée à mettre en œuvre pour des corpus de taille importante. C'est donc la traduction de la requête qui est retenue lorsqu'on parle de recherche d'information par traduction automatique. Cette méthode est en cours de consolidation et les résultats qu'elle donne ne sont pas très satisfaisants, car elle se base sur la traduction automatique dont les résultats ne sont pas toujours probants, surtout lorsque les requêtes sont courtes. En effet, plus les requêtes sont courtes, moins les résultats sont bons.

B. Les moteurs de recherche basés sur des corpus d'apprentissage

Cette méthode basée sur les statistiques est appelée en anglais *Corpus-based* ou *Example-based thesauri method*. Elle s'appuie, pour la traduction des requêtes, sur un thésaurus créé (grâce à un algorithme) à partir d'un corpus d'apprentissage aligné au niveau de la phrase pour trouver des cooccurrences de termes en contextes (Huthins, 2005). Les résultats de cette méthode sont généralement assez satisfaisants mais ont l'inconvénient d'être limités : comme pour toute méthode statistique, on constate un arrêt de la progression des résultats à un certain seuil.

C. Les moteurs de recherche basés sur des dictionnaires de reformulation

Cette méthode, appelée en anglais *«machine-readable dictionaries-based method»* se base sur l'expansion de la requête. Cela consiste à formuler la requête autrement en remplaçant les mots qui la composent par des variantes afin de récupérer des documents pertinents dans lesquels les termes saisis ne sont pas toujours présents. Cette reformulation se fait à l'aide de dictionnaires monolingues qui permettent la reformulation dans une même langue (synonymes, antonymes, etc.) et les dictionnaires bilingues qui permettent la reformulation dans des langues différentes (Aljlayl et *al.*, 2001).

D. Les moteurs utilisant une langue pivot (l'interlingua)

Cette méthode consiste à extraire la sémantique des textes de la langue source grâce à un langage pivot. Il s'agit d'un langage unifié permettant de représenter la sémantique des différentes langues et basé sur le concept des graphes représentant les phrases. Cette méthode est la moins consolidée de toutes celles que nous avons citées et peu de résultats satisfaisants ont été constatés (Hahn et *al.*, 2004).

Bibliographie

- L. Alamarguy, R. Dieng-Kuntz, « Acquisition de relations sémantiques pour un Web biomédical », Projet ACACIA, INRIA Sophia Antipolis, 2004.
- M. Aljlayl, O. Frieder, « Effective Arabic-English Cross-Lingual Information Retrieval via Machine-Readable Dictionaries and Machine Translation», 2001.
- A. Andreewsky, J-P Binquet, F. Debili, C. Fluhr et B. Pouderoux, « le traitement linguistique et statistique des textes et son application dans la documentation juridique », Sixième Symposium sur l'Informatique Juridique en Europe, Thessaloniki, Grèce, 1981.
- Y. Betsgen, « Analyse sémantique latente et segmentation automatique des textes, in 7e JATD, 2004.
- M. Braschler, C. Peters, « The CLEF Campaign», 2005.
- T. Buckwalter, « Buckwalter Arabic Morphological Analyzer Version 1.0 », Academic Publishers, 1998.
- J-P Chevallet, M-F Bruander, « Impact de l'utilisation de multi termes sur la qualité des réponses d'un système de recherche d'information à indexation automatique. »
- T. Dkaki, C. Mhamedi, J. Mothe, « Restituer l'information utile à l'utilisateur: visualisation de la pertinence et de la nouveauté dans les textes»,2003.
- F. Douzidia, « Résumé automatique de textes arabes », mémoire de Master d'informatique», Université de Montréal, 2003-2004.
- M. Dunlop, « Reflections on MIRA : Interactive evaluation in information Retrieval », 2000.
- M. Duval, « Le mot-clé », http://www.dsi-info.ca/mot-cle.html, page consultée le 15 octobre 2006.
- O. Ferret, « Segmenter et structurer thématiquement des textes par l'utilisation conjointe de collocations et de la récurrence lexicale » in Nancy, TALN 2002.
- C. Fluhr, P-A Möellic, P. Hede, « Usage-Oriented Multimedia Information Retrieval Technological Evaluation », proceedings of MIR 2006, Santa Barbara.
- A. Fujii, T. Ishikawa, « Cross-Language Information Retrieval Information Retrieval using Compound Word Translation », 2001.
- G. Grefenstette, N. Semmar, F. Gara, « Modifying a Natural Language Processing System for European Languages to Treat Arabic in Information in Processing and Information Retrieval Application », in Proceedings of the ACL workshops, pp 31-38, Ann Arbor, 2005.
- U. Hahn, K. Markó, M. Poprat, S. Schulz, J. Wermter, & P. Nohama : Crossing Languages in Text Retrieval via an Interlingua RIAO 2004 - Conference Proceedings: Coupling Approaches, Coupling Media and Coupling Languages for Information Retrieval, pp. 100-115, 2004.
- John Huthins, Towards a definition of example-based machine translation, 2005.
- Y. Kadri J. Nie « Effective Stemming for Arabic Information Retrieval » in proceedings of the Challenge of Arabic for NLP/ MT Conference, Londres, Royaume-Uni, 2006.
- H. Kammoun, J-L Lamirel, M. Ben Ahmed «Machine-Learning applied to Queries an Documents for an adaptive and Evolutive Information Retrieval» in proceedings of the International Conference on Machine Intelligence, Tozeur, 2005.
- K. Kishida, K. Chen, « Overview of the Fourth NTCIR Workshop », 2004.
- J. Koenemann, N. J. Belkin, « A case for interactive information retrieval behaviour and effectiveness ». In proceedings of CHI 96, New York, 1996.
- M. Laïb-Boukarri, « Constitution d'un corpus multilingue et reconnaissance des entités nommées », mémoire de DESS Ingénierie Multilingue (CRIM/INALCO), 2001-2002.
- M. Laïb, N. Semmar, C. Fluhr, « Utilisation d'une approche linguistique pour l'indexation et l'interrogation en langage naturel de bases de données textuelles multilingues », 2006
- K. Lespinasse, P. Kremer, D. Schilber, L. Schmitt, « évaluation des outils d'accès à l'information textuelle, les expériences américaine TREC et française AMARYLLIS », 2000.

- C. de Loupy, « Evaluation de l'Apport de connaissances Linguistiques en Désambiguïsation sémantique et Recherche Documentaire », thèse de Doctorat en Informatique (Université d'Avignon et des pays de Vaucluse), 2000.
- D. Nakache, E. Metais, « Evaluation : nouvelle approche avec juges », CEDRIC/CNAM INFORSID, 2005.
- D. Oard, « The TREC-2002 Arabic-English CLIR Track », in proceedings of TREC-2002.
- D. Oard, P. Hackett. « Document translation for cross-language text retrieval at the University of Maryland ». In TREC-6, 1997.
- B. Pouliquen, Thèse : « Indexation de textes médicaux par extraction de concepts, et ses utilisations, Université de Rennes I, Faculté de Médecine, 2002.
- L. Rachi, F. Lair, Rapport de mini-projet sur les moteurs de recherche, ENSICAEN, 2006-2006.
- G. Salton, C. Buckley, « On the use of spreading activation methods in automatic information retrieval » in Proceedings of the 11th ACM-SIGIR Conference, 1988.
- N. Semmar, M. Laïb, C. Fluhr, « Using Cross-language Information Retrieval for Sentence Alignment » in proceedings of the Challenge of Arabic for NLP/ MT Conference, Londres, Royaume-Uni, 2006.
- N. Semmar, F. Gara, C. Fluhr, « Using e Stemmer in Natural Languge Processing system to treat Arabic for Cross-language Retrieval » in International conference on Machine Intelligence, Tozeur, Tunisie, 2005.
- A. Serres, Supports de cours de l'URFIST, Université Rennes II Bretagne-Pays de Loire, 2002, http://www.uhb.fr/urfist/Supports/RechInfoInit/RechInfo3Problematique.html, site consulté le 15 octobre 2006.
- M. Steckel, An Introduction to the Thought of S.R. Ranganathan for Information Architects, 2002.
- J. Véronis, « Etude comparative des six moteurs de recherche », 2006.
- E. Voorhees, « Overview of TREC-2002 », 2002.
- Wikipedia, http://fr.wikipedia.org/wiki/Alphabet_arabe, page consultée le 15 octobre 2006.
- K. Williams, C. Hamel, L. Shrestha, « CAI evaluation handbook: Guidelines for user interface design for computer-aided instruction », Naval Training Systems Center, Orlando, Florida, USA, 1987.
- Y. Yang, X. Liu, « A re-examination of text categorization methods. In SIGIR '99», in Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, pages 42 - 49, New York, NY, USA. ACM Press, 1999.
- Y. Yang, J. Carbonell, R. Brown, R. Frederking, « Translingual Information Retrieval: Learning from Bilingual Corpora », in AO Journal special issue, Best of IJCA-97, 2000.
- R. Zghibi, « Le codage informatique de l'écriture arabe : d'ASMO 449 à Unicode et ISO/CEI 10646 », 2002.