Meriama Laib, Nasredine Semmar, Christian Fluhr
Laboratoire d'Ingénierie de la Connaissance Multimédia Multilingue
CEA–LIST–LIC2M- France
meriama.laib@cea.fr,nasredine.semmar@cea.fr, christian.fluhr@cea.fr

Utilisation d'une approche linguistique pour l'indexation et l'interrogation en langage naturel de bases de données textuelles multilingues

Le développement rapide et continu de l'Internet pose de manière aiguë le problème de l'indexation automatique et de la recherche de l'information textuelle multilingue. D'où l'intérêt des techniques d'ingénierie linguistique permettant l'interrogation en langage naturel de textes écrits dans plusieurs langues (Grefenstette, 1998). Le système de recherche d'information interlingue du LIC2M est composé d'un analyseur linguistique, d'un analyseur statistique, d'un reformulateur, d'un comparateur et d'un moteur de recherche (Besançon et al., 2003). L'analyse linguistique traite les documents à indexer ainsi que les requêtes. Cette analyse concerne le français, l'anglais, l'allemand, l'italien, l'espagnol, le chinois ainsi que l'arabe et dispose de ressources linguistiques diverses: dictionnaires, grammaires, règles pour l'extraction des entités nommées, règles pour les relations de dépendance entre les mots et entre les syntagmes, corpus d'apprentissage et matrices de désambiguïsation, etc. C'est essentiellement cette analyse linguistique qui révèle la structure syntaxique des phrases contenues dans les documents et dans les requêtes et qui permet ainsi au système de pouvoir faire le rapprochement entre les deux.

Pour le traitement de l'arabe, le système dispose des ressources suivantes:

- Un dictionnaire de formes qui contient toutes les formes fléchies et dérivées simples des mots en arabe. Ces mots ont deux sortes d'entrées dans le dictionnaires: complètement voyellées ou complètement dévoyellées. Seules les entrées voyellées possèdent des informations linguistiques (catégorie, genre, nombre, etc.). Les entrées non voyellées qui sont ambiguës par nature ne possèdent que des pointeurs vers les entrées voyellées correspondantes.
- Un dictionnaire de proclitiques ainsi qu'un dictionnaire d'enclitiques simples et composés. La même structure est attribuée à ces entrées, c'est à dire une forme voyellée et une forme non voyellée correspondante.
- Des dictionnaires bilingues Arabe-Français et Arabe-Anglais qui permettent la reformulation bilingue dans le cadre de la recherche d'information interlingue.

L'analyse linguistique des documents et des requêtes se fait en plusieurs étapes:

- L'analyse morphologique permet de rechercher tous les mots des documents et des requêtes dans le dictionnaire des formes et, éventuellement de récupérer les informations linguistiques les concernant. Pour les mots semi voyellés ou non voyellés, cette consultation du lexique permet de récupérer les formes voyellées correspondantes, c'est à dire leur alternatives orthographiques lorsqu'elle existent. Lorsque leur forme de surface le permet les mots sont segmentés en proclitique-radical-enclitique ou en proclitique-radical ou en radical-enclitique ou en proclitique-enclitique (Buckwalter, 2002). Les expressions idiomatiques sont ensuite reconnues et regroupées pour être considérées comme un seul mot dans le graphe d'analyse. Si, après ces traitements un mot reste inconnu, le système lui attribue une/des catégorie(s) par défaut en s'appuyant généralement sur des informations révélées par sa forme de surface.
- Après l'analyse morphologique, la majorité des mots restent ambigus notamment à cause du nombre élevé des voyellations possibles. Le rôle du désambiguïseur morpho-syntaxique est donc de réduire les ambiguïtés en utilisant des matrices de n-grammes obtenues à partir d'un corpus étiqueté et désambiguïsé manuellement.
- L'analyse syntaxique utilise des règles pour établir les relations de dépendance entre les mots dans un même syntagme et entre les syntagmes dans une même phrase.
- La reconnaissance des entités nommées utilise des fichiers de listes et de règles pour reconnaître les entités nommées telles que les noms de personnes, d'organisations, de produits et de lieux ainsi que les unités de mesure.

L'analyse statistique attribue des pondérations aux mots simples et aux mots composés normalisés en fonction de leur répartition dans le corpus (Andreewsky et al., 1981).

La reformulation consiste à enrichir les requêtes en utilisant des dictionnaires de synonymie pour une recherche monolingue et des dictionnaires bilingues pour une recherche interlingue (Debili et al., 1988).

Le comparateur calcule la similitude sémantique entre les requêtes et les documents indexés.

Le moteur de recherche permet de trouver dans l'index des documents ceux qui sont les plus proches des requêtes étendues et fusionne les résultats obtenus pour chaque langue.

Cet article présente une approche linguistique pour l'indexation et l'interrogation en langage naturel de bases de données textuelles multilingues. Nous exposerons dans la première partie de cet article le fonctionnement du moteur de recherche d'information interlingue du LIC2M. Ensuite, nous consacrerons la deuxième partie à la description de l'analyseur linguistique de l'arabe et donnerons quelques exemples d'analyse. Dans la troisième partie, nous montrerons le prototype du moteur de recherche interlingue développé dans le cadre du projet ALMA (Semmar et Fluhr, 2004) et nous analyserons les résultats.

Mots clés: Recherche d'information interlingue, analyse morphologique, désambiguïsation morpho-syntaxique, analyse syntaxique, entités nommées.

Références bibliographiques

1. R. Besançon, Gaël de Chalendar, Olivier Ferret, Christian Fluhr, Olivier Mesnard et Hubert Naets, “The LIC2M’s CLEF 2003 system”, Working Notes for the CLEF 2003 Workshop, Trondheim, Norvège, 21-22 Août 2003.
2. T. Buckwalter, “Buckwalter Arabic Morphological Analyzer Version 1.0”, Linguistic Data Consortium, 2002.
3. 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, 1-3 Juillet 1981.
4. F. Debili, C. Fluhr et P. Radasoa, “About reformulation in full text IRS”, Information processing and Management, Royaume-Uni, 1988.
5. G. Grefenstette, “Cross-language information retrieval”, Boston: Kluwer Academic Publishers, 1998.
6. N. Semmar et C. Fluhr, "Multilingual Search Engine implementation", Rapport Technique du projet ALMA, EURO-MED programme, DG XIII, Commission de l'Union Européenne, Systran, France, Juillet 2004.