Expansion de requêtes pour l'optimisation de la recherche d'information multilingue basée sur la traduction des contenus

Benoît Gaillard, Jean-Léon Bouraoui, Emilie Guimier de Neef, Malek Boualem

Orange Labs
2 avenue Pierre Marzin
22300 Lannion cedex, France

{benoit.gaillard, jeanleon.bouraoui, emilie.guimierdeneef, malek.boualem}@orange-ftgroup.com

1. Recherche d'information multilingue : approche par traduction des contenus

La quantité d'information en ligne croît très rapidement, ainsi que le nombre de langues dans lesquelles ces contenus sont disponibles. En revanche, la complexité des requêtes reste limitée (2 à 3 mots en moyenne). Des traitements spécifiques s'avèrent donc nécessaires pour préciser le sens de certaines requêtes, ou au contraire pour en élargir la portée. L'Expansion de Requêtes (ER), par exemple, permet, en ajoutant des mots aux requêtes d'optimiser leur mise en relation avec les documents. Par ailleurs, la Recherche d'Information Multilingue ou CLIR (Cross Language Information Retrieval), permet d'accéder à des documents dont la langue est différente de celle dans laquelle la requête est formulée. les systèmes de CLIR ont souvent recours à la Traduction Automatique (TA), soit de la requête, soit des contenus. Des études comparatives montrent que l'approche par traduction des contenus est plus performante, mais les principaux systèmes de CLIR privilégient l'approche par traduction des requêtes. En effet, traduire un grand nombre de documents dans un grand nombre de langues serait trop coûteux. Nous proposons ici l'optimisation d'un prototype de CLIR dont les contenus indexés sont en quantité suffisamment faible pour que l'approche par traduction des contenus ait été choisie. Le vocabulaire employé dans les textes traduits automatiquement est réduit par rapport au vocabulaire spontané. Par exemple, *"night club"* peut se traduire en français par *« discothèque »* ou par *« boîte de nuit »*. Le système de traduction du prototype étudié ici traduit systématiquement *"night club"* par *« boîte de nuit »*, un utilisateur n'aura par conséquent accès à aucun résultat s'il saisit la requête *« discothèque »*. Dans un corpus constitué de 51461 requêtes, nous avons mesuré le nombre de mots absents dans le corpus de contenus traduits, constitué des titres et

des descriptions d'environ 57000 vidéos, traduits. Pour filtrer les erreurs d'orthographe, les entités nommées ou les url, nous avons croisé le corpus de requêtes avec deux lexiques du français: *Lexique 3* (www.lexique.org) et le thésaurus conçu dans notre laboratoire, obtenant ainsi un *« corpus utile »* de 34070 mots dont 2800, tels que *« mincir »* ou *« potager »*, n'appartiennent pas au vocabulaire des contenus traduits. Le défaut de couverture lexicale peut se représenter par le ratio: $R_{mismatch}$=2810/34070=8%

2. Expansion de requêtes : un lien avec le vocabulaire des données traduites

L'ER cherche et propose des termes "voisins" à ajouter à la requête initiale. Dans le module d'ER, les termes voisins sont sélectionnés sur la base de la proximité sémantique, lexicale et/ou morphologique, selon 5 modes: flexion, synonymes, hyperonymes, termes dérivés, expansion géographique. Ces modes sont basés sur le thésaurus mentionné ci-dessus et sur la base de données *Geonames* (www.geonames.org). De cette sélection résulte une liste de termes ordonnée par ordre décroissant de proximité avec le sens de la requête initiale. Nous sélectionnons dans cette liste les 20 premiers termes qui font partie du vocabulaire des contenus traduits. Ainsi l'ER est adaptée à l'index des contenus traduits. Notre système se présente donc sous la forme de deux modules complémentaires annexés à un moteur de recherche monolingue: en front office, un module d'ER ciblée, en back office un module de TA de données.

3. Evaluation et interprétation des résultats : réduction du silence sans perte de précision

La comparaison du prototype aux autres solutions de CLIR dépasse le cadre de cet article. L'évaluation consiste donc à comparer le silence et la précision du moteur avec et sans ER ciblée. Le nombre moyen de documents trouvés avec le prototype de CLIR est de 124, alors qu'il est de 193 avec la version optimisée par l'ER. L'amélioration est donc de 50%. Le taux de réduction du silence (pourcentage de requêtes silencieuses qui ne le sont plus après expansion) est de 8%, mesuré sur le corpus de requêtes. Pour vérifier que la réduction du silence ne se fait pas au détriment de la Précision, nous avons comparé la *Précision à 10* (P_{10}) du moteur de recherche de CLIR avec et sans ER filtrée. Pour chaque requête, $P_{10}=N_r/\min(N_t,10)$, où N_r est le nombre de résultats pertinents parmi les 10 premiers, et N_t est le nombre total de résultats. Nous avons mesuré P_{10} pour les 400 requêtes les plus fréquentes du corpus utile. Nous avons observé que la P_{10} moyenne avec ER filtrée est de 55%, et sans ER de 59%. Il n'y a donc pas de changement significatif dans la précision. Ces résultats peuvent être améliorés en mesurant la réduction du silence sur le corpus utile, et en adaptant l'ER plus spécifiquement à l'index des contenus traduits. Des évolutions de ce travail pourraient être orientées vers l'amélioration de la précision.