Fusion de ressources hétérogènes pour la recherche d'information multilingue

Frederik Cailliau, Aude Giraudel et Céline Poudat

Sinequa Labs
12, rue d'Athènes
F-75009 Paris
cailliau@sinequa.com
giraudel@sinequa.com
poudat@sinequa.com

RÉSUMÉ. *Afin d'améliorer la recherche multilingue dans le moteur de recherche Sinequa Engine, nous avons intégré les connaissances multilingues du service Sensagent au module de requêtes du moteur de recherche Sinequa Engine. L'interface développée propose une extension de la requête aux choix de l'utilisateur par traduction des différents mots dans les langues sélectionnées. Pour limiter le grand nombre de traductions que peut engendrer une requête complexe, nous avons déployé un filtrage sémantique par calcul vectoriel. L'ensemble de la chaîne de traitement repose fortement sur les ressources linguistiques de Sinequa. L'utilisation d'une ressource extérieure, si elle résout le problème de la seule traduction, pose des problèmes d'exploitation et d'adéquation des ressources entre elles, qui ne pourraient être résolus que par une vraie fusion des ressources.*

ABSTRACT. *This paper presents the integration of Sensagent's multilingual knowledge in the process of query translation to improve Sinequa's search engine capacities in cross-language information retrieval. The developed interface expands the query with the translations of each word in the languages chosen by the user. To deal with the problem of numerous translations generated by complex queries, a semantic filtering using vectorial representation has been implemented. The processing strongly depends on both Sinequa's linguistic resources and the translations provided by Sensagent. This way of associating different sources of knowledge remains nevertheless a major issue in terms of adequacy and compatibility.*

MOTS-CLÉS: *recherche d'information multilingue, recherche interactive.*

KEYWORDS: *Cross Language Information Retrieval, interactive search.*

1. Introduction

Le développement exponentiel de grands ensembles documentaires multilingues entraîne une demande de plus en plus forte de systèmes de Recherche d'Information Multilingue (RIML) opérationnels. Différentes approches permettent à l'utilisateur de formuler une requête dans une langue donnée pour ramener des documents écrits dans une langue différente : les approches mobilisant des dictionnaires bilingues ou multilingues, ou des systèmes de traduction automatique, les méthodes statistiques sur corpus, ou encore les approches conceptuelles permettant de dépasser les problèmes posés par la surface des langues en privilégiant un niveau conceptuel (Volk et Buitelaar, 2003). Afin d'optimiser les résultats, ces méthodes sont souvent combinées aux différents moments du processus de RIML – traduction de la requête, recherche des documents à partir de la requête traduite, fusion des résultats ramenés dans les différentes langues. On recourt plus généralement à la traduction des requêtes qu'à celle des documents, pour des raisons de temps d'indexation – bien que ce choix dépende naturellement de l'application, comme le soulignent (Oard et Ertunc 2002).

Développé en tant qu'application pilote dans le cadre du projet Vodel[1], le système que nous avons développé s'inscrit dans l'approche de traduction de la requête en interaction avec l'utilisateur. L'application finale est hybride, puisqu'elle repose sur l'utilisation de deux technologies existantes et opérationnelles si on les considère isolément : le moteur de recherche Sinequa Engine[2] et l'ensemble des dictionnaires gérés par Memodata et accessibles via Sensagent[3] qui permettent de donner la main à l'utilisateur pour sa recherche multilingue.

L'analyse de la requête et l'interaction entre Sinequa Engine et Sensagent reposent sur des techniques de traitement automatique des langues (TAL) mises en œuvre par les outils de Sinequa. Il s'agit d'une désambiguïsation morpho-syntaxique et d'une lemmatisation par accès à des lexiques morpho-syntaxiques. Le calcul vectoriel prend en compte un thésaurus sémantique. Ces ressources ont une cohérence globale dans la suite de traitements existante. L'utilisation de Sensagent comme ressource extérieure résout le problème de la traduction, mais pose des problèmes d'exploitation et d'adéquation des ressources entre elles qui ne pourraient être résolus que par une vraie fusion des ressources.

Après avoir présenté l'application développée (2.) et les ressources lexicales utilisées (3.), nous décrirons la suite de traitements que nous avons développée pour adapter les deux ressources (4.) ainsi que son application à la recherche d'information (5.). Une évaluation critique du système hybride obtenu sera ensuite proposée (6.).

---

[1] Valorisation Ontologique des Dictionnaires Electroniques (http://vodel.insa-rouen.fr/).
[2] Le moteur de recherche commercialisé par Sinequa (http://www.sinequa.com/).
[3] http://www.sensagent.com/, http://www.memodata.com/.

2. Présentation générale de l'application

Le module que nous avons intégré au moteur de recherche Sinequa Engine propose des traductions pour chaque mot de la requête en utilisant le service Sensagent. L'utilisateur intervient ensuite et décide de prendre en compte ou non ces traductions dans sa recherche. En cas de requête complexe, un filtrage sémantique intervient pour limiter le nombre de traductions. La liste des résultats de la recherche est unique pour toutes les langues et ordonnée selon le score de pertinence des documents. Un pavé de navigation permet de filtrer selon la langue des documents. Notre application cible donc des utilisateurs avertis, qui maîtrisent plus ou moins les langues dans lesquelles ils cherchent.

Le corpus que nous avons indexé pour tester l'application est constitué de 817 201 dépêches de l'AFP produites en 4 langues différentes : français (43,7%), anglais (29,8%), espagnol (18,3%) et allemand (8,1%). Le corpus est multilingue, mais n'est pas aligné – il ne s'agit pas de textes et de leurs traductions. Les textes ont été produits en 2005, ce qui garantit une certaine homogénéité des textes sans en faire pour autant un corpus domanialement homogène.

3. Ressources lexicales utilisées

3.1. *Lexiques monolingues de Sinequa Engine*

Pour les opérations de désambiguïsation morpho-syntaxique, de lemmatisation et de détection d'entités nommées, Sinequa Engine dispose d'un ensemble cohérent de lexiques morpho-syntaxiques, de règles et de corpus pour chaque langue traitée. Pour six langues (dont les quatre visées par notre application), il existe un thésaurus sémantique, organisé en 800 descripteurs, chaque descripteur correspondant à une dimension de l'espace vectoriel. A chaque descripteur est associé un ensemble de mots qui vont permettre de calculer un vecteur sémantique pour chaque document et de le situer dans l'espace vectoriel.

3.2. *Dictionnaires multilingues de Sensagent*

La ressource dictionnairique utilisée dans le cadre du projet Vodel est issue d'un partenariat général entre plusieurs institutions initié par Memodata. Cette ressource présente l'avantage de fournir un contenu très volumineux et d'une très grande diversité. En effet, Sensagent met à disposition depuis janvier 2005 un ensemble de dictionnaires en 22 langues et selon les cas et les contenus disponibles, des définitions, des synonymes, des expressions, des traductions, etc. Le service proposé par Sensagent aide à la compréhension en utilisant des dictionnaires monolingues et multilingues.

4. Extraction et filtrage des traductions de la requête

Le travail réalisé dans le cadre du projet Vodel s'inscrit dans la problématique plus générale de recherche d'information multilingue. L'approche adoptée privilégie l'utilisation de ressources dictionnairiques pour la traduction de requêtes. L'originalité des travaux réalisés repose sur une combinaison de traitements linguistiques intervenant à plusieurs niveaux. Les traitements s'articulent en trois étapes : analyse de la requête, accès aux ressources linguistiques de Memodata via Sensagent, filtrage sémantique. Le schéma suivant présente la chaîne de traitement développée.

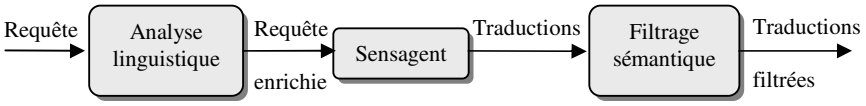

Figure 1. *Chaîne de traitement de la requête.*

4.1. *Analyse linguistique de la requête*

L'objectif de l'analyse est d'éviter au maximum la traduction « mot à mot », ce qui impose un repérage performant des mots composés et des entités nommées (noms de personnes, noms d'entreprises et lieux géographiques). La segmentation de la requête en mots se fait en utilisant les lexiques de Sinequa et la détection des entités repose sur l'application d'un ensemble de règles linguistiques codé sous forme d'automates accompagnés de lexiques spécialisés.

4.2. *Récupération du contenu de la ressource dictionnairique Sensagent*

L'accès à la ressource Sensagent est réalisé par un service Web dont les paramètres d'entrée sont le mot à traduire et la langue source. Le contenu renvoyé est composé de l'ensemble des traductions. Le nombre de traductions renvoyées pour chaque mot d'entrée étant très volumineux en général, un traitement supplémentaire a été nécessaire afin de limiter la quantité d'informations tout en privilégiant la pertinence des traductions conservées : un filtrage sémantique, décrit en détail dans la section suivante.

4.3. *Filtrage sémantique des traductions*

Le filtrage sémantique vise à réduire la liste des traductions renvoyées par Sensagent en éliminant les termes non pertinents. A chaque terme renvoyé par Sensagent est associé un vecteur sémantique qui est mis en correspondance avec celui de la requête. La distance euclidienne entre ces deux vecteurs est alors calculée. Un terme est considéré comme pertinent si cette distance est inférieure à la distance moyenne.

4.4. *Illustration par l'exemple*

Prenons l'exemple de la requête « barreau palais de justice » dans le contexte d'une traduction du français vers l'anglais : l'analyse linguistique de la requête va permettre de reconnaître le mot composé « palais de justice ». L'accès à Sensagent fournit ensuite une liste de traductions pour chaque mot. Les traductions renvoyées pour le mot « barreau » sont les suivantes : *advocacy*, *attorneyship*, *baluster*, *banister*, *bar*, *legal community*, *legal profession*, *rod*, *rung*, *spindle*. On remarque qu'on obtient de manière logique des traductions faisant référence aux deux sens du terme « barreau ». Il est cependant non pertinent de conserver ces deux sens pour une tâche de recherche d'information.

Nous partons de la représentation sémantique de la requête : 123:1,492:1,625:10,787:3. Les composantes du vecteur sémantique sont séparées par des virgules. Chaque composante contient un code sémantique auquel est affecté un poids (après les « : ») qui dépend du nombre de mots ainsi que de la proximité entre ces mots. Chaque axe sémantique définit une thématique particulière ; la combinaison des axes permet alors de représenter la diversité sémantique d'un terme. Par exemple, le vecteur sémantique de la requête est composé de quatre axes sémantiques : *123-ligne,linéaire,droite* (barreau) ; *492-obstacle,qui gêne* (barreau) ; *625-tribunal,procès* (barreau, palais de justice) ; *787-bâtiment,édifice,monument* (palais de justice).

L'axe sémantique principal de la requête concerne logiquement la thématique « tribunal, procès » commune aux deux termes de la requête. La représentation sémantique de chaque traduction ainsi que la distance avec la requête sont données dans le tableau suivant (les axes communs avec ceux de la requête sont mis en évidence).

	Vecteur sémantique	Distance sémantique
Advocacy	468:1,**625:1**	9.64
Attorneyship	624:1,**625:1**	9.64
Baluster	184:1,682:1	10.63
Banister	184:1,**492:1**,682:1,**787:1**	10.34
Bar	56:1,**123:1**,492:1,**625:1**,685:1,725:1,743:1	9.70
Legal community	5:1,19:1,38:1,52:1,378:1,399:1,453:1,584:1,588:1,624:1,**625:1**	10.15
Legal profession	378:1,387:1,624:1,**625:1**,786:1	9.80
Rod	**123:1**,127:1,137:1,238:1,260:1,581:1,710:1	10.77
Rung	120:1,184:1,188:1,337:1,339:1,673:1,758:1,762:1	10.91
Spindle	676 :1	10.58

Tableau 2. *Distance sémantique des traductions (distance moyenne = 10.22) ; vecteur de la requête (123:1,492:1,625:10,787:3)*

Après filtrage sémantique, seules les traductions du terme « barreau » étant dans la thématique « tribunal » sont conservées. Le traitement réalisé est ainsi pertinent. La partie suivante présente une évaluation détaillée des performances et de l'efficacité de la chaîne de traitements.

5. Application à la recherche d'information

5.1. *Formulation de la requête à partir des traductions et recherche interlingue*

A partir de la liste de traductions obtenue après filtrage sémantique, une requête est formulée dans chaque langue par concaténation des différents termes conservés. Chaque requête ainsi produite après traduction comporte en général un nombre de mots plus important que la requête exprimée dans la langue d'origine, ce qui augmente les possibilités de recherche.

Les différentes requêtes obtenues sont alors appliquées aux documents du corpus écrits dans la langue concernée. Cela est rendu possible par une indexation séparée des documents dans les différentes langues.

5.2. *Classement et présentation des résultats*

Pour chaque langue cible sélectionnée, le moteur renvoie une liste contenant les documents pertinents. Les résultats sont ensuite classés par pertinence indépendamment des langues, ce qui nécessite une fusion des différentes listes. On présuppose en effet que l'utilisateur privilégie la pertinence du résultat renvoyé à la langue de départ. Nous avons cependant choisi de donner la possibilité d'une visualisation des documents par langue.

Ce choix de représentation des résultats par pertinence va néanmoins avoir tendance à privilégier un positionnement des documents de la langue source en tête de liste. En effet, le nombre de mots de la requête traduite est généralement supérieur à celui de la requête initiale, ce qui a pour effet immédiat de faire baisser la pertinence globale des documents ramenés dans la langue cible. Une solution envisagée pour pallier ce problème serait de proposer à l'utilisateur les n résultats les plus pertinents dans chaque langue sur la première page de résultats.

6. Evaluation critique

Malgré diverses lacunes, le système proposé a tenté d'exploiter au mieux la richesse des traductions de Sensagent, et des dictionnaires morpho-syntaxique et sémantique de Sinequa. Nous reviendrons de manière critique sur le travail présenté, en soulignant les points forts et les insuffisances du système, et en proposant diverses voies d'amélioration.

6.1. *Filtrage sémantique*

Outre le fait que le choix de la distance moyenne comme seuil de décision permette une réduction du nombre de traductions de moitié, ce traitement présente un double avantage : il est particulièrement efficace dans le cas de requêtes isotopiques qui contiennent un terme ambigu dans la langue source. Par ailleurs, lorsqu'une traduction est ambiguë dans la langue cible (ex : *quote* en anglais), le filtrage sémantique, en éliminant les traductions ambiguës, va avoir pour effet d'augmenter le silence (documents pertinents non renvoyés), ceci étant largement contrebalancé par la réduction du bruit dans les résultats.

Bien qu'il présente l'avantage de privilégier une représentation sémantique pivot indépendante des langues et permettant une réduction significative des traductions sémantiquement éloignées des termes les plus représentés, ce filtrage présente un triple inconvénient :

1/ Si le filtrage des synonymes est efficace dans le cas de requêtes isotopiques de type « récolte de pommes de terre », il a peu d'incidence sur les requêtes plus longues et plus complexes, d'autant que (Oard, 1998) a souligné que l'approche dictionnairique était plus adaptée aux requêtes courtes.

2/ Le seuil de filtrage adopté gagnerait à être affiné. En effet, rappelons que les traductions ramenées sont ensuite concaténées, ce qui réduit leur pertinence dans les langues autres que celles de la requête initiale en augmentant le nombre de mots à retrouver.

3/ On reste au niveau de la requête sans prendre en compte les spécificités du corpus. On pourrait ainsi envisager à la suite du filtrage de pondérer les mots de la requête posée au moyen d'une mesure de type TF*IDF. D'autres mesures, issues globalement du domaine de l'apprentissage, sont néanmoins envisageables : ainsi, (Gao *et al.*, 2001) proposent d'entraîner un modèle a priori domanial sur le corpus de documents afin de sélectionner les mots traduits appropriés.

6.2. *Problèmes posés par la combinaison de ressources hétérogènes*

Comme il n'y a pas eu de vraie fusion des traitements linguistiques et de leurs ressources lexicales, il est difficile d'évaluer l'apport de chacune des ressources dans l'application finale. Le problème se pose par exemple pour le traitement des mots composés non reconnus par l'une ou l'autre des deux ressources. Une entité nommée qui comporte des mots communs (ex. : *air france*) sera ainsi traduite mot à mot et ramènera du bruit. Inversement il se peut qu'un mot composé (ex. : *énergie solaire*) soit bien reconnu, mais qu'il n'existe pas de traduction dans les dictionnaires, ou seulement pour certaines langues. La seule façon de résoudre ces incohérences est de fusionner toutes les ressources, ce qui est impossible en pratique. Cela demanderait une mise à jour constante des ressources en concertation avec chaque partenaire à chaque modification par une des parties.

7. Conclusion et perspectives

Les différents types de ressources linguistiques que nous avons exploitées partagent un grand nombre de connaissances. Ils ne forment pas pour autant un ensemble de ressources cohérent, car leur objectif d'exploitation n'est pas le même. La mutualisation des ressources hétérogènes pour la recherche d'information est possible, mais provoque des incohérences qui peuvent paraître incompréhensibles pour l'utilisateur, dont la satisfaction reste un des critères d'évaluation clés pour les moteurs de recherche. Différentes améliorations peuvent néanmoins être proposées concernant la prise en compte de la structure de la requête et l'ajustement du seuil de filtrage, qu'on pourrait préciser en considérant les spécificités du corpus avec une mesure de type TF*IDF.

8. Bibliographie

Ballesteros, L. et Sanderson, M., « Addressing the lack of direct translation resources for cross-language retrieval », *Actes de Twelfth International Conference on Information and Knowledge Management*, New Orleans, LA, USA, 2003, p. 147–152.

Gao, J., Nie, J.Y, Xun, E., Zhang, J., Zhou, M., Huang, C., « Improving Query Translation for CLIR using Statistical Models », *24th ACM-SIGIR*, New Orleans, 2001, pp. 96-104.

Lehtokangas, R., Keskustalo, H et Järvelin, K., « Highly relevant documents lost in CLIR : experiments with dictionary translation and pseudo-relevance feedback », *Information Retrieval*, volume 9, n° 4, septembre 2006, p. 395-397.

Nie, J., « Towards a unified approach to CLIR and multilingual IR », *Actes de SIGIR 2002 Workshop I, Crosslanguage information retrieval: a research map*, University of Tampere, Finland, 2002, p. 8–14.

Oard, D.W., « A comparative study of query and document translation for cross-lingual information retrieval », *Actes de AMTA,* Philadelphia, PA, 1998.

Oard, D.W. et Ertunc, F., « Translation-based indexing for cross-language retrieval », *Actes de 24th BCS-IRSG European Colloquium on IR Research: Advances in Information Retrieval*, 2002, LNCS, p. 324-333.

Pirkola, A, « The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval », *Actes de 21st ACM/SIGIR Conference,* 1998, pp. 55-63.

Volk, M., Vintar, S. et Buitelaar, P., « Ontologies in cross-language information retrieval », *Actes de 2nd Conference on Professional Knowledge Management*, Lucerne, 2003.