Règles d'association inter-langues au service de la recherche d'information multilingue

Belhaj Rhouma Sourour, Asma Ben Achour, Malek Hajjem, Chiraz Latiri

Laboratoire de recherche LIPAH, Faculté des Sciences de Tunis
Campus Universitaire Tunis El Manar, 1060 Tunis, Tunisie.
sourour.bhr@gmail.com, asmabenachour@gmail.com,
malek.hajjem@gmail.com, chiraz.latiri@gnet.tn

Résumé. Dans cet article, nous proposons de montrer l'intérêt et l'utilité de déploiement des règles d'association inter-langues (RAILs) dans le domaine de la Recherche d'Information Multilingue (RIM). Ces règles sont des connaissances additionnelles résultantes d'un processus de fouille de grands corpus parallèles alignés au niveau de la phrase. En effet, leurs conclusions exprimées dans une langue cible représentent des traductions potentielles de leurs prémisses, exprimées dans une langue source. Nous illus trons l'utilisation des RAILs dans le contexte de la RIM à travers deux propositions, à savoir : (*i*) la traduction des requêtes et (*ii*) la traduction des termes de l'index. L'évaluation expérimentale a été menée sur la collection de documents MUCHMORE. Les résultats ont montré une amélioration significative de la pertinence système.

1 Introduction et motivations

L'abondance des documents, notamment sur le Web, dans de nombreuses langues a rendu nécessaire l'existence d'une Recherche d'Information Multilingue. La RIM consiste ainsi à formuler une requête dans une langue source et à rechercher des documents pertinents dans des langues cibles (RIM) Nie (2010). Dans un contexte multilingue, la requête ainsi que les documents ne sont pas représentés dans un même espace d'indexation étant exprimés dans des langues différentes. Par conséquent, la mise en correspondance de leurs descripteurs sera impossible. Ainsi, pour permettre une recherche multilingue, l'enjeu consiste à représenter les documents et la requête dans un même espace d'indexation.

Dans cet article, notre motivation est double : en premier lieu, il s'agit de déployer les techniques d'ECT qui permettent d'inférer des relations de traduction entre des unités linguistiques pour l'identification de lexiques bilingues à partir des corpus parallèles, et, en deuxième lieu, exploiter ces lexiques bilingues dans le cadre de la RIM Lavecchia et al. (2008); Latiri et al. (2010).

2 RI multilingue : Brève revue de la littérature

La revue de la littérature du domaine de la RIM montre qu'il existe plusieurs types d'approches pour modéliser la tâche de la RI multilingue. Nous pouvons citer *les approches basées sur la traduction automatique* qui s'appuient sur la traduction automatique d'une requête ou des documents de la collection. La traduction automatique de la requête est plus explorée Herbert et al. (2011) malgré qu'elle souffre d'un manque de précision comparée à celle basée sur la traduction d'une collection de documents dans laquelle un contexte d'information nettement plus important est utilisé, diminuant ainsi les risques de mauvaise traduction. De plus, *les approches basées sur des corpus d'apprentissage parallèles ou comparables* Wu et He (2010) qui s'appuient, pour la traduction des requêtes, sur un thésaurus ou des corpus comparables ou parallèles pour trouver des co-occurrences de termes Hazem et al. (2011); Bo et al. (2011). En outre, *les approches basées sur des lexiques bilingues* qui se basent principalement sur l'expansion de requêtes. Elle consiste à reformuler la requête à l'aide de dictionnaires grâce aux lexiques bilingues Levow et al. (2005). Enfin, *les approches à base d'un language pivot (l'interlingua)* qui s'appuient sur un langage unifié (pivot) permettant de représenter la sémantique des différentes langues Hahn et al. (2004). La recherche multilingue se facilite en effectuant la traduction du corpus et des requêtes dans un même langage pivot.
L'ensemble de propositions que nous introduisons dans cet article s'inscrit dans la famille des méthodes qui font appel à des méthodes de traduction, basées sur un lexique bilingue.

3 Génération de règles d'association inter-langues à partir de corpus parallèles

Initialement introduit dans le domaine de la traduction automatique statistique (TAS) par Latiri et al. (2010), les Règles d'Association Inter-langues, notée RAIL, désignent que la prémisse de la règle est exprimée dans la langue source l_S alors que sa conclusion l'est dans la langue cible l_C. Une interprétation intuitive d'une règle inter-langues est que la conclusion de cette dernière est une traduction potentielle de sa prémisse Latiri et al. (2010).

Définition 1 *Une règle d'association inter-langues, notée par RAIL, est une implication de la forme : $R : S_S \Rightarrow S_C$ telles que S_S et S_C sont deux séquences de termes fermées fréquentes en langue source et cible, de tailles respectives n et m mots* Latiri et al. (2010).

Une règle d'association inter-langues est également appréciée par les deux métriques de *support* et de *confiance* Agrawal et Skirant (1994). De plus, une règle d'association inter-langues est dite *valide* si sa confiance est supérieure ou égale au seuil minimal de confiance La génération des règles associatives inter-langues est réalisée à partir d'un corpus parallèle Anglais-Allemand par un parcours de l'espace de recherche qui s'effectue au niveau de la phrase. La phase de génération est précédée par une étape d'extraction de séquences fréquentes. Nous apportons, pour effectuer cette étape, des adaptations à l'algorithme BFSM Chang (2004) liées au contexte de la fouille de données textuelles. L'algorithme qui permet de dériver les règles d'association inter-langues à partir de l'ensemble des séquences inter-langues extraites est décrit dans Latiri et al. (2010).

Nous proposons dans ce qui suit, de déployer l'ensemble des règles d'association inter-langues générées pour étudier leur apport dans la RI multilingue.

4 Déploiement des règles d'association inter-langue pour la RIM

Nous proposons deux réflexions de recherche en RIM à base des RAIL, à savoir : (*i*) Traduction d'une requête pour la RI multilingue par les règles d'association inter-langues ; et, (*ii*) Traduction des termes de l'index par les règles d'association inter-langues.

4.1 Traduction d'une requête pour la RI multilingue par les règles d'association inter-langues

Partant d'une requête exprimée dans une langue source, il s'agit de définir un modèle de RI capable de restituer des documents formulés dans chacune des langues cibles de la collection multilingue. Dans notre contexte, la traduction d'une requête peut être conduite en deux étapes. D'une part, une première étape qui consiste à dériver des règles d'association inter-langues à partir des corpus parallèles (l_S, l_C). Nous exploitons dans ce cadre le processus d'extraction des règles d'association inter-langues défini dans Latiri et al. (2010) et explicité dans la section 3. Ainsi, le résultat de cette étape est un ensemble de lexiques bilingues. D'autre part, une deuxième étape consistant à déployer le lexique bilingue extrait dans la traduction de requête. La traduction de requêtes consiste à remplacer chacun des termes de la requête dans sa langue source avec des termes dans la langue cible. Nous soutenons l'idée que les termes de la requête à traduire, exprimée en langue source l_S, figurent dans les prémisses des règles inter-langues dérivées. De ce fait, leurs traductions potentielles sont représentées par les conclusions de ces règles d'association inter-langues. Le but de cette étape est donc de supprimer certaines traductions jugées inadéquates dans le contexte de la requête Req_S. Nous suggérons d'utiliser la mesure de *confiance* pour ne garder que les règles d'association valides par rapport à un seuil minimal de confiance $minconf$, qui permet de retenir les meilleures traductions avec une confiance assez élevée. Cette étape diminue ainsi l'ambiguïté dans le choix des traductions sans forcément l'éliminer.

4.2 Traduction des termes de l'index des documents par les règles d'association inter-langues

Nous partons d'une requête Req_S formulée dans une langue source l_S et d'une collection bilingue $\{C_{l_S}, C_{l_C}\}$. Pour chaque document d de la collection C_{l_C}, son index dans la langue cible, noté $Index_{l_C}(d)$ est généré par un processus classique d'indexation. Ensuite, le processus d'extraction de règles d'association inter-langues est lancé sur les corpus bilingue $\{C_{l_S}, C_{l_C}\}$. Le but de ce processus est de générer des corrélations entre des unités linguistiques de la langue cible l_C, dans laquelle est représenté l'index d'un document de la collection, et celles de la langue source l_S. Dans ce contexte, les corrélations inter-langues sont générées dans le sens contraire que celui considéré pour la traduction de requêtes, *i.e.,* de la langue cible vers la langue source. La sélection des RAILs les plus pertinentes pour la traduction se fait sur la base du seuil minimal de confiance $minconf$. L'index $Index_{l_C}(d)$ est par la suite traduit en utilisant ces règles inter-langues valides, où chaque terme t_i de l'index est traduit par le terme ou la séquence de termes qui apparaît dans une corrélation inter-langues contenant le

terme t_i. Cette nouvelle représentation de l'index des documents d'une collection multilingue permet ainsi une interrogation monolingue, avec une requête Req_S exprimée en langue source l_S et un corpus de documents en langue cible l_C, dont les index des documents sont traduits moyennant les lexiques bilingues.

5 Évaluation expérimentale

5.1 Cadre d'évaluation

Nous considérons une collection fournie dans le cadre du projet MUCHMORE [1]. Cette collection regroupe des résumés de documents scientifiques dans le domaine médical, obtenus à partir du site web de Springer, rédigés en anglais et en allemand. Dans de nombreux cas toutefois, la version anglaise est une reformulation complète du résumé allemand, ce qui rend difficile un alignement au niveau des phrases. Ce corpus est considéré comme étant un corpus parallèle bruité. Premièrement, les corpus sont étiquetés et lemmatisés afin d'extraire les lemmes des mots pleins (noms, verbes, adjectifs, adverbes). Seuls les noms et les adjectifs sont pris en compte par nos algorithmes de génération des RAILs.

5.2 Scénarios d'évaluation

Nous avons réalisé deux scénarios d'évaluation basés sur la collection MUCHMORE. La base d'évaluation comparative(baseline), notée dans la suite de l'article *MT* est le résultat donné dans Volk et al. (2003) où les auteurs ont utilisé la collection MUCHMORE pour évaluer une approche de RIM à base de traduction automatique des requêtes de l'allemand vers l'anglais.

Scénario 1 : Traduction des requêtes par les RAILs. Le premier scénario, noté dans la suite de l'article *Trad-Req*, consiste à interroger le corpus exprimé en anglais avec des requêtes exprimées en allemand de la collection MUCHMORE. Dans ce cas, les termes de la requête sont traduits de l'allemand (langue source) vers l'anglais (langue cible) en utilisant les règles d'association inter-langues dérivées à partir du corpus parallèle allemand-anglais.

Scénario 2 : Traduction des index par les RAILs. Le deuxième scénario, noté dans la suite *Trad-index*, consiste à interroger le corpus en allemand avec des requêtes exprimées en anglais. Dans ce cas, chaque terme de l'index relatif à chaque document du corpus de la collection sont traduits de l'allemand (langue cible) vers l'anglais (langue source) moyennant les lexiques bilingues illustrés par les RAILs (*cf.,* sous-section 4.2).

5.3 Résultats expérimentaux et discussion

Le Tableau 1 synthétise les résultats de la recherche multilingue utilisant la collection MUCHMORE. À la lecture du Tableau 1, nous constatons que le déploiement des RAILs pour la traduction de requêtes (scénario *Trad-Req*) réalise une amélioration significative de la MAP

1. Multilingual Concept Hierarchies for Medical Information Organization and Retrieval
http ://muchmore.dfki.de/

	MAP	Docs Pert. Retr.	P@10
Baseline *MT*	0.1381	440	0.2920
Trad-Req	0.2187 ($+58,36\%$)	282	0.2920
Trad-Req-GT	0.3322 ($+51.89\%$)	370	0.4560
Trad-index	0.2386	518	0.4880 ($+64,72\%$)

TAB. 1 – *Performance de la RI multilingue en terme de MAP, Docs Pert.Retr et P@10 (Amélioration en $\%$).*

($+58,36\%$) par rapport à la base d'évaluation *MT*. Cependant, ce scénario diminue le taux rappel (282 documents pertinents retrouvés). De l'autre coté, le scénario *Trad-Req-GT* réalise une MAP de 0.3322 plus importante de $51,89\%$ par rapport au scénario *Trad-Req*, aussi pour le taux de rappel (370 documents pertinents retrouvés). Par ailleurs, le deuxième scénario *Trad-doc* qui consiste à traduire les termes d'index des documents par les RAILs, conduit à une amélioration de la pertinence système de la tâche RI en terme de MAP ($+9,09\%$) par rapport à la traduction de requêtes par les RAILs (*Trad-Req*). De plus, le scénario *Trad-doc* effectue le meilleur taux de rappel aves 518 documents pertinents retrouvés. Ceci est également justifié par la variation significative de la précision exacte P@10 qui traduit une augmentation du nombre de documents pertinents restitués et réordonnés parmi les documents les mieux classés. Nous constatons ainsi que ces deux scénarios ont prouvé l'apport en efficacité de la RI, obtenu en utilisant les règles d'association inter-langues dans une contexte de RI multilingue.

6 Conclusion

Cet article met en évidence le déploiement des règles d'association inter-langues dans le cadre de la RI multilingue. L'ensemble des propositions introduites se basent sur une fouille efficace d'un corpus parallèle aligné au niveau des phrases à partir d'une collection bilingue. L'objectif est d'extraire des corrélations inter-langues entre les unités linguistiques utilisées ultérieurement dans la RI multilingue. L'évaluation expérimentale menée sur la collection de documents MUCHMORE a montré une amélioration significative de la pertinence système. Par ailleurs, ces propositions restent extensibles, dans le sens où les résultats peuvent être améliorés si nous utilisons d'autres métriques statistiques lors de l'extraction des RAILs.

Références

Agrawal, R. et R. Skirant (1994). Fast algorithms for mining association rules. In *Proceedings of the* 20^{th} *International Conference on Very Large Databases, VLDB 1994*, Santiago, Chile, pp. 478–499.

Bo, L., E. Gaussier, E. Morin, et A. Hazem (2011). Degré de comparabilité, extraction lexicale bilingue et recherche d'information interlingue. In *Proccedings de la* $18^{ème}$ *Conférence sur le Traitement Automatique des Langues Naturelles, TALN 2011*, Montpellier, France, pp. 211–222.

Chang, K. Y. (2004). Efficient sequential pattern mining by breadth-first approach. Master degree, National Taiwan University.

Hahn, U., K. G. Markó, M. Poprat, S. Schulz, J.Wermter, et P. Nohama (2004). Crossing languages in text retrieval via an interlingua. In *Proceedings of* 7^{th} *International Conference on Computer-Assisted Information Retrieval, RIAO 2004*, Avignon, France, pp. 100–115. CID.

Hazem, A., E. Morin, et S. P. Saldarriaga (2011). Métarecherche pour l'extraction lexicale bilingue à partir de corpus comparables. In *Proceedings de la* 18^{th} *Conférence sur le Traitement Automatique des Langues Naturelles, TALN 2011*, pp. 283–293.

Herbert, B., G. Szarvas, et I. Gurevych (2011). Combining query translation techniques to improve cross-language information retrieval. In *Proceedings of* 33^{rd} *European Conference on IR Research, ECIR 2011*, Volume 6611 of *LNCS*, Dublin, Ireland, pp. 712–715. Springer-Verlag.

Latiri, C., Y. Slimani, C. Nasri, et K. Smaïli (2010). Extraction des séquences fermées fréquentes à partir de corpus parallèles : application à la traduction automatique. In *Actes des dixièmes journées francophones en Extraction et gestion des connaissances, EGC'2010*, Volume RNTI-E-19 of *Revue des Nouvelles Technologies de l'Information*, Hammamet, Tunisie, pp. 55–60. Cépaduès-Éditions.

Lavecchia, C., K. Smaïli, et D. Langlois (2008). Discovering phrases in machine translation by simulated annealing. In *Proceedings of the* 9^{th} *Annual Conference of the International Speech Communication Association, INTERSPEECH'08*, Brisbane, Australia, pp. 2354–2357. ISCA.

Levow, G., D. W. Oard, et P. Resnik (2005). Dictionary-based techniques for cross-language information retrieval. *Information Processing and Management 41*(3), 523 – 547.

Nie, J. Y. (2010). *Cross-Language Information Retrieval*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Volk, M., S. Vintar, et P. Buitelaar (2003). Ontologies in cross-language information retrieval. In *WOW2003 (Workshop Ontologie-basiertes Wissensmanagement)*, Luzern, Switzerland, pp. 47–50.

Wu, D. et D. He (2010). A study of query translation using google machine translation system. In *Proceedings of the International Conference on Computational Intelligence and Software Engineering, CiSE*, Wuhan, China, pp. 1–4. IEEE.

Summary

In this paper, we propose to disclose how that can be achieved when inter-lingual association rules (ILARs) are used in Cross Langage Information Retrivel (CLIR). The basic idea of ILARs is that the potential translations of a linguistic unit which appears in a premise of an association rule are obtained by selecting linguistic units which are present in conclusions of the same ILAR. We propose two different manners to use ILARs in the CLIR field, namely: (*i*) queries translation and (*ii*) terms indexing translation. The experiments carried out on the bilingual document collection MUCHMORE showed that our ILARs could considerably enhance cross langage information retrieval effectiveness.