Recherche bilingue et multilingue d'information

Vers une sélection des bonnes traductions

Jacques Savoy – Pierre-Yves Berger

Institut interfacultaire d'informatique
Université de Neuchâtel, Pierre-à-Mazel 7, 2000 Neuchâtel (Suisse)
{Jacques.Savoy, Pierre-Yves.Berger}@unine.ch

RESUME. *Afin de pouvoir interroger des corpus écrits dans plusieurs langues, la stratégie la plus simple et la moins onéreuse consiste à traduire la requête soumise dans la (ou les) langue(s) souhaitée(s). Dans ce but, nous nous sommes appuyés sur des ressources disponibles gratuitement sur le Web. En comparant l'efficacité du dépistage entre les requêtes traduites manuellement ou automatiquement, on constate que la machine s'avère moins bonne que l'être humain. Toutefois, cette première conclusion se base sur une moyenne et une analyse plus détaillée indique une forte variabilité, dans le dépistage de l'information, entre les performances des différentes traductions produites par la machine. La question qui se pose est de savoir si l'on peut prédire la performance d'une requête traduite afin de sélectionner seulement la meilleure ou les meilleures traductions. Afin de résoudre ce problème, nous avons conçu un système de prédiction basé sur la régression logistique et capable de prédire les meilleures traductions. L'évaluation de notre approche s'avère supérieure au meilleur système de traduction automatique.*

ABSTRACT. *In order to search within corpora written in two or more languages, the simplest and most effective approach is to translate the submitted request into the required language(s). To achieve this goal, we based our IR model on translation tools freely available on the Web. When comparing the retrieval effectiveness of manually and automatically translated requests, we found that human-based translation outperformed machine-based approaches. However, when we analyzed the query-by-query performance, we found query performances based on machine-based translations to vary a great deal. The question that then arises is whether or not we can predict the retrieval performance of a translated query and as a result we may thus select only the best translation(s). To respond to this, we designed and evaluated a predictive system based on the logistic regression, and used it to select the top most appropriate machine-based translations. An evaluation of this approach shows retrieval performance is better than using the best machine-based translation.*

MOTS-CLES : *Recherche multilingue, sélection automatique, apprentissage automatique.*

KEY WORDS: *Cross-lingual retrieval, automatic selection , automatic learning.*

1. Introduction

Depuis quelques années, des campagnes d'évaluation se sont intéressées aux problèmes spécifiques de la recherche documentaire dans des collections composées de documents écrits dans deux langues (TREC-6 à TREC-8), dans plusieurs langues européennes (forum CLEF qui a débuté en 2000 (Peters *et al.*, 2002 ; Peters *et al.*, 2003a ; Peters *et al.*, 2003b)) ou dans plusieurs langues d'extrême Orient (forum NTCIR dès 1999)[1]. Ces campagnes dont le succès en nombre de participants est indéniable, visent à encourager les chercheurs à aborder de nouveaux thèmes d'une part et, d'autre part, à faciliter le transfert de technologie de pointe des milieux universitaires vers l'industrie. Aborder la recherche d'information bilingue, ou plus généralement multilingue, répond à différents besoins comme ceux liés à la présence de différentes communautés linguistiques sur le Web, à la consultation du droit ou de la jurisprudence européenne[2] voire le dépistage de dépêches et d'annonces rédigées dans diverses langues.

La mise au point de modèles de recherche bilingue ou multilingue impose en premier lieu la conception de système de dépistage efficace dans les diverses langues. Nous ne rencontrons pas de difficultés pour la langue anglaise qui a été étudiée par la communauté scientifique depuis de nombreuses années. Pour les autres, nous devons encore proposer des outils robustes et fiables tels que des listes de mots-outils (mots très fréquents et peu porteurs de sens comme « le », « dans », « lequel » ou « car ») ou des enracineurs (ou *stemmers*) supprimant certaines séquences terminales liées à la flexion (« volcan**s** » et « volcan ») voire certaines dérivations suffixales (« volcan**ique** » et « volcan »). Pour un ensemble de langues, à l'image du français, de l'italien ou de l'espagnol, ces deux outils semblent être suffisants afin de disposer d'un moteur de recherche possédant une performance raisonnable (Peters *et al.*, 2003a). Pour d'autres langues, comme le finnois, la morphologie flexionnelle s'avère plus riche d'une part et, d'autre part, des modifications dans la racine du mot soulèvent des difficultés que nous n'avons pas résolu de manière satisfaisante (par exemple, le mot fleuve au nominatif donne « **joki** » et s'écrit « **joe** » au génitif, ou « **joel**lta » à l'ablatif, ce qui démontre la présence de trois racines distinctes). Les langues germaniques, de même que le finnois, procèdent fréquemment à la concaténation de formes pour générer des mots composés tels que « Versicherungsgesellschaft » (composé de « Versicherung » (assurances) + « S » + « Gesellschaft » (société)) ou « Computersicherheit » (« Computer » + « Sicherheit » (sécurité)). Enfin, certaines langues, à l'exemple du chinois ou du japonais, requièrent plus qu'un octet par caractère et elles n'indiquent pas la séparation explicite des mots par des espaces.

---

[1] Voir les sites http://trec.nist.gov, http://clef.iei.pi.cnr.it ou http://research.nii.ac.jp/ntcir/.

[2] Voir, par exemple, le site http://europa.eu.int/eur-lex/

Dans une seconde étape, les modèles de recherche d'information multilingue doivent surmonter les barrières linguistiques. Dans ce but, diverses approches ont été suggérées mais, pour l'essentiel, deux stratégies se dégagent pour surmonter cette difficulté. En premier lieu, on peut traduire automatiquement tous les documents dans une seule et même langue (souvent l'anglais) puis on autorise l'usager à écrire ses requêtes dans cette langue (Chen *et al.*, 2003). Comme alternative, plusieurs auteurs suggèrent de construire un index pour chaque langue retenue et, après réception de la requête, de la traduire automatiquement dans toutes les langues souhaitées. Dans ce dernier cas, le système doit encore fusionner les listes de résultats provenant des recherches effectuées dans les diverses langues afin de présenter une liste unique à l'usager. Dans le cadre de cet article, nous avons choisi cette alternative en recourant à divers outils disponibles gratuitement afin de traduire la requête soumise dans différentes langues.

La suite de cet article se subdivise de la manière suivante. Le prochain chapitre présentera les modèles de recherche d'information utilisés, les collections de documents et notre méthodologie d'évaluation. Dans le troisième chapitre, nous aborderons les différentes possibilités de traduire automatiquement les requêtes dans diverses langues et nous évaluerons ces approches. Le quatrième chapitre décrira notre approche basée sur la régression logistique afin de sélectionner la ou les meilleures traductions et évaluera cette stratégie de sélection à l'aide de notre collection-test.

2. Recherche d'information dans diverses langues

Afin d'obtenir un panorama assez complet de la qualité des réponses fournies par différentes stratégies de dépistage, nous avons utilisé le logiciel SMART pour implanter diverses variantes du modèle vectoriel ainsi que le modèle probabiliste Okapi (Robertson *et al.*, 2000). Ces modèles sont décrits brièvement dans la section 2.1 tandis que la section 2.2 présente une évaluation de ces modèles sur la base de cinq collections comprenant des documents rédigés dans cinq langues différentes.

2.1. *Modèles de recherche d'information*

Afin de pouvoir dépister des documents en réponse à une requête donnée, nous devons au préalable les indexer c'est-à-dire extraire une liste de mots-clés caractérisant au mieux leur contenu sémantique. Dans ce but, l'ordinateur identifie les mots tout en ignorant les mots-outils, pour ensuite éliminer des marques générées par la morphologie flexionnelle voire dérivationnelle. Enfin, une pondération est calculée pour chacun des termes d'indexation T_j issus du document D_i. Cette pondération devrait tenir compte des facteurs suivants :

- du nombre d'occurrences du terme T_j (mot simple ou composé, syntagme nominal) dans le document D_i, fréquence notée tf_{ij} (seuls les mots simples ont été retenus dans cet article);
- de la fréquence documentaire (notée df_j) c'est-à-dire du nombre de documents dans lesquels le terme T_j apparaît, ou plus précisément de l'idf_j, le logarithme de l'inverse de la fréquence documentaire ($idf_j = \ln [n/df_j]$, avec n indiquant le nombre de documents dans la collection) ;
- de la longueur des documents.

Afin de ne pas alourdir le texte, l'annexe 1 présente une liste complète des formules de pondération utilisées dans cet article dont la dénomination retenue est dérivée du système SMART. Ainsi, pour décrire précisément un modèle de dépistage, un premier triplet de lettres décrit la pondération utilisée lors de l'indexation des documents et, un second triplet, celle appliquée aux requêtes. Par exemple, une stratégie « bnn-bnn » indique une indexation binaire (terme présent ou non) tandis que la séquence « nnn-nnn » signifie que seul le nombre d'occurrences est retenu pour pondérer les termes des documents et des requêtes.

Pour le modèle vectoriel classique « ntc-ntc », l'indexation tient compte à la fois de la fréquence d'occurrences dans le document et de l'inverse de la fréquence documentaire (nombre de documents dans lesquels le terme apparaît). De plus, dans cette stratégie « ntc-ntc », les poids sont normalisés selon la formulation du cosinus.

Cependant la fréquence d'occurrences peut être modifiée pour tenir compte du fait que l'apparition de la première occurrence devrait posséder un poids important. De plus, nous devrions accorder une importance décroissante au fil des répétitions d'un même terme dans un document. Ainsi, la différence entre une fréquence d'occurrences de neuf ou de huit n'apporte pas une information très précieuse tandis que la différence entre une fréquence unitaire ou nulle constitue une information très pertinente. Afin de respecter ces deux principes, nous proposons de pondérer un terme selon l'équation $[0{,}5 + 0{,}5 \cdot (tf_{ij}/\max tf_{i.})]$, de prendre le logarithme de la fréquence d'occurrences $(\ln (tf_{ij})+1)$ ou de recourir au double logarithme $(\ln (\ln (tf_{ij})+1)+1)$.

De plus, de nouvelles formules de pondération plus complexes ont été mises au point, en particulier, le modèle probabiliste Okapi (Robertson *et al.*, 2000), le modèle vectoriel « Lnu-ltc » (Buckley *et al.*, 1995) ou la stratégie « dtu-dtn » (Singhal *et al.*, 1998). Ces dernières possèdent l'avantage de tenir compte de la longueur des documents en cherchant à pénaliser les longs documents abordant généralement plusieurs sujets et qui répondent, en moyenne, moins bien aux attentes de l'usager.

Dans toutes nos expériences, le score de chaque document (ou son degré de pertinence jugé par la machine) est obtenu par le calcul du produit interne. Par exemple, pour l'approche « bnn-bnn », ce score indiquera le nombre de termes communs entre le document et la requête. Pour l'approche « nnn-nnn », ce score

tiendra compte de la fréquence d'occurrences des termes communs entre le document et la requête.

2.2. *Performance des collections individuelles*

Les corpus de documents utilisés correspondent aux campagnes d'évaluation CLEF-2001 (Peters *et al.*, 2002) et CLEF-2002 (Peters *et al.*, 2003a) et contiennent différents journaux tels que le *Los Angeles Times* (Etats-Unis), *Le Monde* (France), *La Stampa* (Italie), *Der Spiegel* et *Frankfurter Rundschau* (Allemagne), des dépêches d'agences de presse comme *EFE* (Espagne) ou celles de l'agence télégraphique suisse (disponibles en allemand, français et italien). Les documents de ces corpus couvrent approximativement les mêmes thèmes et sont tous extraits de l'année 1994.

Quelques statistiques sur ces corpus sont reprises dans l'annexe 2. La taille des corpus varie fortement entre les langues avec des volumes plus restreints pour le français et l'italien. Le nombre de termes d'indexation par article reste assez similaire (environ 130) avec une moyenne un peu plus élevée pour la collection anglaise (167,33). Avec ces documents, nous disposons de cent requêtes (50 pour CLEF-2001 soit les requêtes n^0 41 à n^0 90 et 50 requêtes pour CLEF-2002 (n^0 91 à n^0 140)). En consultant le contenu de ces requêtes, nous constatons que ces dernières ne s'adressent pas à un domaine précis mais couvrent différents besoins d'information comme « Des pesticides dans la nourriture pour bébés », « Embargo sur l'Iraq » ou « Coupe du monde de football ». Suivant le modèle proposé par les campagnes d'évaluation de TREC, chaque requête se subdivise en trois parties logiques comprenant un titre bref (« T »), une phrase de description (« D ») et une partie narrative spécifiant les critères de pertinence (« N »).

	Précision moyenne (% de changement)				
Langue modèle	Anglais 89 requêtes	Français 99 requêtes	Italien 96 requêtes	Allemand 99 requêtes	Espagnol 99 requêtes
Okapi - npn	**53,21**	**50,74**	**43,93**	**38,97**	**54,78**
Lnu - ltc	51,51 (-3%)	48,01 (-5%)	41,78 (-5%)	37,13 (-5%)	51,69 (-6%)
dtu - dtn	48,28 (-9%)	48,05 (-5%)	41,24 (-6%)	37,01 (-5%)	49,92 (-9%)
atn - ntc	47,24 (-11%)	46,07 (-9%)	40,51 (-8%)	35,72 (-8%)	49,63 (-9%)
ltn - ntc	41,21 (-23%)	45,31 (-11%)	39,08 (-11%)	35,06 (-10%)	49,04 (-10%)
lnc - ltc	33,58 (-37%)	34,99 (-31%)	32,27 (-27%)	29,76 (-24%)	41,68 (-24%)
ltc - ltc	31,87 (-40%)	33,38 (-34%)	31,23 (-29%)	29,02 (-26%)	39,06 (-29%)
ntc - ntc	30,47 (-43%)	32,54 (-36%)	29,53 (-33%)	29,17 (-25%)	35,66 (-35%)
bnn - bnn	22,73 (-57%)	18,73 (-63%)	22,10 (-50%)	19,81 (-49%)	27,37 (-50%)
nnn - nnn	10,44 (-80%)	14,46 (-72%)	14,81 (-66%)	14,98 (-62%)	23,79 (-57%)

Tableau 1. *Précision moyenne des différents modèles selon les cinq langues (requêtes « TD »)*

Comme méthodologie d'évaluation, nous avons retenu la précision moyenne (Salton *et al.*, 1983) mesurant la qualité de la réponse fournie par l'ordinateur, mesure utilisée par les campagnes d'évaluation TREC ou CLEF. Finalement, pour décider si un système de dépistage est meilleur qu'un autre, on admet comme règle d'usage qu'une différence de 5 % dans la précision moyenne peut être considérée comme significative.

Le tableau 1 indique la précision moyenne des différentes stratégies retenues selon les cinq langues et en utilisant les titres et descriptions des requêtes (« TD »). Les résultats démontrent que le modèle Okapi présente la meilleure performance dans chacune des langues (ou collections). La performance de cette approche nous servira de système de référence à partir duquel les pourcentages de différence seront calculés. De plus, la performance proposée par ce modèle situe parmi les meilleures approches de CLEF 2001 (Peters *et al.*, 2002) et CLEF 2002 (Peters *et al.*, 2003a). On notera que les modèles vectoriels « Lnu-ltc » ou « dtu-dtn » voire « atn-ntc » possèdent une évaluation globale qui les place dans un deuxième groupe au niveau de la performance. Finalement, les différences de performance sont plus marquées pour les langues italiennes et allemandes.

3. Traduction automatique des requêtes

Au chapitre précédent, nous disposions de requêtes écrites dans la même langue que les documents. Dans cette partie, nous nous intéressons à la recherche bilingue d'information, contexte dans lequel la requête soumise est rédigée dans une langue (l'anglais dans notre cas) afin de dépister des documents écrits dans une autre langue, soit le français, l'italien, l'allemand ou l'espagnol. Pour franchir cette barrière linguistique, notre approche se base sur une traduction automatique de la requête écrite en anglais dans la langue correspondante aux documents. Sur la base de cette requête traduite, notre système de dépistage applique les algorithmes décrits dans le chapitre précédent. Afin d'obtenir diverses traductions de nos requêtes, nous avons utilisé cinq systèmes disponibles gratuitement, à savoir :

- REVERSO ONLINE™ translation2.paralink.com
- SYSTRAN™ babel.altavista.com/translate.dyn
- GOOGLE™ www.google.com/language_tools
- FREETRANSLATION™ www.freetranslation.com
- INTERTRAN™ www.tranexp.com:2000/InterTran

De plus, nous pouvons aussi recourir à un dictionnaire bilingue et dans le cadre de cet article, nous avons sélectionné le site BABYLON™ (www.babylon.com). Dans ce dernier cas, nous effectuons une traduction mot à mot de la requête en ne retenant que le premier terme retourné par le dictionnaire (évaluation notée « Babylon 1 »), les deux premières traductions (« Babylon 2 ») ou les trois premiers termes proposés par le dictionnaire (« Babylon 3 »).

La précision moyenne obtenue par le modèle Okapi selon les quatre langues et selon les différents systèmes de traduction est indiquée dans le tableau 2. Comme première ligne de ce tableau, nous avons repris la traduction manuelle et cette valeur va nous servir de valeur de référence afin de calculer les pourcentages de différence. On notera que cette traduction manuelle présente, quelque soit la langue, la meilleure approche. En effet, aucun des systèmes de traduction automatique ne présente une performance moyenne supérieure à celle obtenue par les traductions manuelles.

Au niveau des outils de traduction automatique, aucune des solutions étudiées apporte toujours la meilleure performance pour toutes les langues. On peut toutefois signaler que le dictionnaire bilingue Babylon propose le meilleur outil pour le français et l'italien (en prenant uniquement la première traduction proposée) ou pour l'allemand (en tenant compte des deux premiers termes suggérés mais la différence avec le système Reverso reste très faible). Pour la langue espagnole, le système Reverso occupe la première place au niveau de la performance.

	Précision moyenne (% de changement)			
collection modèle	Français 99 requêtes	Italien 96 requêtes	Allemand 99 requêtes	Espagnol 99 requêtes
Okapi - npn	50,74	43,93	38,97	54,78
Reverso	45,25 (-10,8%)	—	30,24 (-22,4%)	**46,13** (-15,8%)
Systran	44,64 (-12,0%)	31,62 (-28,0%)	29,41 (-24,5%)	40,46 (-26,1%)
Google	44,80 (-11,7%)	31,63 (-28,0%)	29,12 (-25,3%)	40,33 (-26,4%)
FreeTrans	41,05 (-19,1%)	31,80 (-27,6%)	26,34 (-32,4%)	40,98 (-25,2%)
InterTran	38,76 (-23,6%)	29,81 (-32,1%)	23,22 (-40,4%)	38,21 (-30,2%)
Babylon 1	**47,65** (-6,1%)	**32,21** (-26,7%)	29,87 (-23,4%)	38,97 (-28,9%)
Babylon 2	43,45 (-14,4%)	28,84 (-34,4%)	**30,41** (-22,0%)	34,37 (-37,3%)
Babylon 3	42,26 (-16,7%)	27,50 (-37,4%)	29,69 (-23,8%)	32,34 (-41,0%)
Meilleur	54,48 (+7,4%)	42,26 (-3,8%)	41,43 (+6,3%)	52,93 (-3,4%)

Tableau 2. *Précision moyenne des différents outils de traduction selon les langues (modèle Okapi, requête « TD »)*

Or la mesure de performance indiquée dans le tableau 2 est une moyenne qui peut cacher une forte variabilité entre les divers outils de traduction. Afin de quantifier ces variations, nous avons indiqué dans le tableau 3 deux valeurs par langue et outil de traduction. La première indique le nombre de requêtes pour lesquelles le système correspondant propose la meilleure alternative. Cependant, lorsque deux ou plusieurs outils possèdent la même performance optimale, la valeur accordée équivaut à l'inverse du nombre d'outils proposant cette meilleure précision. Ainsi, si deux systèmes de traduction obtiennent la même performance maximale, on leur attribue un score de 1/2, et si quatre systèmes arrivent au même résultat optimal, le score attribué à chacun sera de 1/4. Avec ce mode de calcul, le système

Reverso propose 20,3 fois la meilleure traduction en langue allemande, tandis que « Babylon 2 » apporte la meilleure traduction pour 12,8 requêtes (et bien que ce système soit en moyenne le plus performant, voir tableau 2).

Comme seconde valeur dans le tableau 3, nous avons indiqué le nombre de requêtes pour lesquelles le système de traduction propose l'unique meilleure traduction. En effet, face à une requête courte, les systèmes de traduction automatique peuvent proposer la même solution et obtiennent donc la même performance. Ainsi, pour la langue française, le système Reverso propose 14 fois la meilleure traduction et il s'avère être le seul, dans ces cas, à atteindre cette performance. Pour la langue espagnole, et uniquement pour cette langue, le système Reverso offre la meilleure performance moyenne (voir tableau 2) et obtient la meilleure performance pour 23,2 requêtes, nombre le plus élevé pour cette langue.

	Nombre de requêtes			
collection	Français	Italien	Allemand	Espagnol
Reverso	**17,1** / 14	–	**20,3** / **19**	**23,2** / **20**
Systran	8,3 / 0	13,1 / 0	12,7 / 8	8,8 / 1
Google	11,3 / 3	14,1 / 3	11,2 / 6	10,3 / 2
FreeTransla	11,8 / 10	**21,6** / 10	14,8 / 14	18,3 / 16
InterTran	13,8 / 13	20,1 / 13	7,3 / 7	13,3 / 12
Babylon 1	17,0 / **15**	19,8 / **15**	9,3 / 9	13,8 / 13
Babylon 2	9,8 / 4	3,8 / 4	12,8 / 5	6,1 / 6
Babylon 3	9,8 / 4	3,3 / 4	10,8 / 3	5,1 / 5

Tableau 3. *Nombre de requêtes pour lesquelles le système de traduction propose la meilleure performance (modèle Okapi, requête « TD »)*

Finalement, si l'on retient pour chaque requête la meilleure traduction automatique, nous obtiendrons la précision moyenne indiquée dans la dernière ligne du tableau 2 (ligne débutant par « Meilleur »). La performance d'un tel système automatique permet d'atteindre celle d'une traduction manuelle pour les langues italienne et espagnole et propose une précision moyenne légèrement meilleure pour les langues française et allemande (puisque nous sommes légèrement au-dessus des 5 % de différence). En se basant sur des connaissances a posteriori, nous pouvons donc atteindre, lors d'une interrogation bilingue, une performance similaire à celle d'une interrogation unilingue, pour les langues française, italienne, allemande ou espagnole pour le moins.

4. Sélection automatique des bonnes traductions

Après réception de la requête en langue anglaise, nous la traduisons automatiquement dans la langue souhaitée au moyen de nos huit stratégies. Mais parmi ces

huit requêtes traduites, laquelle permettra d'obtenir la meilleure performance en termes de rappel et de précision ? Comment pouvons-nous sélectionner automatiquement la (ou les) meilleure(s) traduction(s) fournie(s) ?

Quelques travaux antérieurs ont tenté de prévoir un traitement en fonction des caractéristiques de la requête. Ainsi Savoy *et al.* (1995) suggèrent une stratégie basée sur la méthode des plus proches voisins afin de savoir comment sélectionner le meilleur système de dépistage parmi sept en fonction d'informations statistiques sur la requête soumise. Dans la piste « robuste » de TREC (dépistage de l'information avec des requêtes difficiles) Kwok *et al.* (2003) proposent de concevoir une stratégie permettant de prédire quand il faut prévoir une phase de pseudo-expansion de la requête et quand l'application d'une telle technique va diminuer la performance moyenne.

Dans le cas présent de sélection de la meilleure (ou des) meilleures traductions, nous avons conçu une stratégie automatique de sélection basée sur la régression logistique (Hosmer *et al.*, 2000). Cette approche statistique permet de calculer une probabilité de réalisation d'une variable expliquée binaire selon les valeurs prises par un ensemble de variables prédictives, variables pouvant être réelles, catégorielles ou binaires. Cette méthode statistique a déjà été utilisée dans différents contextes de prédiction en bibliothéconomie (Bookstein *et al.*, 1992), comme stratégie de dépistage (Gey, 1994) ou pour la fusion de listes de résultats dans le cadre de métamoteurs de recherche (Le Calvé *et al.*, 2000).

Dans le cas présent, nous pouvons obtenir diverses informations statistiques sur nos requêtes. Ainsi, pour chacune des huit traductions, nous connaissons le système ayant produit cette traduction de même que le nombre de mots de la requête. De plus, pour chaque terme, nous connaissons la fréquence de ce mot dans la requête ou, plus important, son idf (soit le rapport ln[n/df] avec df indiquant la fréquence documentaire ou le nombre de documents possédant ce mot). On notera que la valeur moyenne de l'idf s'avère être un bon prédicteur de la performance d'une requête selon Cronen-Townsend *et al.* (2002), hypothèse que nous désirons vérifier dans notre contexte bilingue. De plus, nous pouvons, sur la base d'une liste prédéfinie, reconnaître les noms propres de personne (par exemple, Kim Il Sung, Clinton), ceux liés à la géographie (e.g., Nice, France) ou autres noms propres (e.g., Nirvana).

En résumé, nous allons estimer, pour chaque traduction, la probabilité que cette dernière soit la meilleure en recourant aux variables prédictives suivantes :

- le système de traduction automatique ayant traduit la requête, noté *source*;
- le nombre de termes indexés inclus dans la requête, noté *concepts* ;
- la valeur minimale de l'idf de tous les termes de la requête, notée *minidf* ;
- la valeur maximale de l'idf de tous les termes de la requête, notée *maxidf* ;
- la moyenne des valeurs idf des termes de la requête, notée *avgidf* ;
- la présence ou l'absence de noms de personnes, variable binaire notée *person* ;
- la présence ou l'absence de noms liés à la géographie, variable binaire *geo* ;
- la présence ou l'absence d'autres noms propres, variable binaire notée *other* .

Dès lors et pour chaque traduction, nous regroupons ces variables dans un vecteur $\mathbf{X} = [x_1, x_2, \ldots x_k]$ que nous utilisons pour estimer la probabilité que cette traduction soit excellente (voir équation 1).

$$\text{Prob}\,[\,\text{excellente traduction} \mid \mathbf{X}\,] = \frac{e^{\alpha+\beta_1 \cdot x_1+\beta_2 \cdot x_2+ \cdots+\beta_k \cdot x_k}}{1+e^{\alpha+\beta_1 \cdot x_1+\beta_2 \cdot x_2+ \cdots+\beta_k \cdot x_k}} \quad (1)$$

dans laquelle les paramètres α, β_1, β_2, … β_k doivent être estimés sur la base d'un ensemble d'observations (échantillon d'entraînement). Pour cet article, nos estimations ont été réalisées à l'aide du logiciel R (Venaples *et al.*, 1999).

Afin de savoir si une sélection du (ou des) meilleure(s) traduction(s) s'avère possible, nous avons décidé de travailler dans le contexte qui nous semblait le plus difficile, à savoir la langue espagnole. En effet, en consultant le tableau 2, on remarque que le système Reverso propose, en moyenne, la meilleure traduction automatique et, selon les informations du tableau 3, cet outil propose également la meilleure solution pour le plus grand nombre de requêtes (23,2 / 20). Sur la base de ces informations, une décision simple et raisonnable serait de toujours sélectionner la traduction proposée par le système Reverso. Afin de dépasser la performance de cette stratégie de référence, nous avons imaginé trois modèles basés sur la régression logistique.

Modèle A. Dans ce premier modèle, nous avons tenu compte de toute l'information disponible et les variables retenues sont : *source*, *minidf*, *person*, *geo* et *other*. La sélection de cet ensemble de variables a été faite selon la procédure `stepAIC` de S (minimisation du critère AIC (Venaples *et al.*, 1999)). Selon nos attentes, l'origine de la traduction est incluse dans cet ensemble (variable *source*), et le fait de connaître que telle ou telle traduction a été produite par Reverso s'avèrera un élément important dans la décision de retenir cette traduction comme la meilleure. Une mesure liée à l'idf a également été retenue (le *minidf* dans ce cas) ainsi que les informations binaires notant la présence ou non de noms propres de personne (*person*), de géographie (*geo*) ou autres (*other*).

Modèle B. Dans notre deuxième modèle, nous avons tenu à masquer l'origine de la traduction afin de permettre une sélection uniquement sur les autres variables. Dans ce cas de figure, et avec l'aide de la procédure `stepAIC`, nous proposons de retenir les variables *concepts*, *minidf*, *person*, *geo* et *other*. On notera que dans ce

deuxième modèle, la variable *source* est remplacée par la variable *concepts* (soit la longueur en nombre de mots indexés de la requête traduite).

Modèle C. Dans ce dernier modèle, nous désirons mettre à l'épreuve le modèle de prédiction de Cronen-Townsend *et al*. (2002) qui indique que l'idf moyen d'une requête s'avère être un bon prédicteur de la précision moyenne. Les variables que nous avons sélectionnées sont : *concepts*, *minidf*, *avgidf*, et *maxidf*. Un modèle basé uniquement l'idf moyen (ou *avgidf*) ne propose pas une bonne performance ; cette dernière s'élève à 47,31 % soit légèrement supérieure à la performance de 46,13 % obtenue en utilisant toujours la traduction proposée par le système Reverso.

Afin d'évaluer les trois modèles retenus, nous pouvons utiliser les 99 requêtes disponibles pour estimer les coefficients α, β_1, β_2, … β_k de la régression logistique et reprendre le même jeu de requêtes pour l'évaluation. Cette méthode d'évaluation, nommée rétrospective, est certes biaisée mais elle nous permet d'avoir une idée de la performance sous-jacente du modèle. Comme méthode alternative d'évaluation, nous avons opté pour le *leaving-one-out* proposant une estimation sans biais. Cette stratégie autorise un entraînement sur l'ensemble des observations moins une et cette dernière est utilisée pour évaluer le modèle. En itérant sur les requêtes disponibles, nous pouvons ainsi obtenir une évaluation faite sur les 99 requêtes avec un apprentissage fait sur 98 requêtes.

Dans une première série d'expériences, nous avons exigé de la machine qu'elle retourne pour chaque requête, la meilleure des traductions parmi les huit disponibles. La performance moyenne obtenue sous ces conditions est indiquée dans le tableau 4 sous la ligne notée « tolérance 0 % ». Pour le modèle A, la décision est relativement simple ; il faut toujours prendre la traduction obtenue par le modèle Reverso (évaluation rétrospective ou *leaving-one-out*). Disposant de moins d'information, les modèles B et C présentent une précision moyenne nettement inférieure au modèle A ou à une stratégie simple retournant toujours la solution proposée par Reverso.

	Précision moyenne (% changement)		
modèle	modèle A	modèle B	modèle C
Reverso	46,13	46,13	46,13
rétrospective			
tolérance 0%	46,13 (0,0%)	42,91 (- 7,0%)	41,29 (-10,5%)
tolérance 5%	46,13 (0,0%)	46,29 (+0,3%)	45,35 (- 1,7%)
tolérance 15%	46,20 (+0,2%)	**48,84** (+5,9%)	47,53 (+3,0%)
tolérance 25%	**47,65** (+3,3%)	48,64 (+5,4%)	**48,21** (+4,5%)
leaving-one-out			
tolérance 0%	46,13 (0,0%)	40,85 (-11,4%)	40,16 (-12,9%)
tolérance 5%	46,13 (0,0%)	44,06 (- 4,5%)	43,39 (- 5,9%)
tolérance 15%	46,09 (- 0,1%)	**48,75** (+5,7%)	47,13 (+2,2%)
tolérance 25%	**46,22** (+0,2%)	48,58 (+5,3%)	**48,24** (+4,6%)

Tableau 4. *Précision moyenne de nos trois modèles de prédiction*

Au lieu d'être stricte et de retenir uniquement la traduction maximisant l'équation 1 (régression logistique), nous pouvons admettre que si la probabilité obtenue s'écartait de, par exemple, 5 % de la meilleure traduction, nous pouvions également considérer la traduction sous-jacente comme une bonne traduction de la requête courante. En admettant cette tolérance, nous ne limitons pas notre procédure de sélection à la recherche de l'unique traducteur optimum mais notre système vise à trouver les bonnes traductions d'une requête écrite en anglais.

Sur cette base d'une marge de tolérance de 5 % (de 15 % ou de 25 %), la machine peut indiquer que deux ou plusieurs traductions automatiques peuvent être considérées comme excellentes et donc être retenues dans le dépistage final. Dans ce cas, la machine concatène les termes provenant de toutes les requêtes sélectionnées.

Les évaluations rétrospectives ou *leaving-one-out* reprises dans le tableau 4 indiquent que c'est seulement en combinant plusieurs traductions automatiques que nous arrivons à dépasser la performance moyenne du meilleur outil de traduction automatique, soit la valeur de 46,13 % obtenue par le système Reverso. La meilleure performance est obtenue avec le modèle B et avec une marge de tolérance de 15 %. La précision moyenne obtenue s'élève à 48,75 % en concaténant, pour ce cas, en moyenne 5,16 traductions. Pour le modèle C, la meilleure performance s'obtient avec une marge de tolérance de 25 % pour une précision moyenne de 48,24 % (avec, en moyenne, 5,88 traductions par requête).

Il est un peu surprenant de constater que notre modèle A n'arrive pas au même niveau de performance bien que disposant, a priori, d'information plus importante. Mais dans ce modèle, le choix est fortement influencé par le système Reverso, bien qu'il soit en moyenne le meilleur, il ne fournit la meilleure traduction « que » pour 23,2 requêtes sur 99.

5. Conclusion

Sur la base de nos expériences, les conclusions suivantes peuvent être tirées :

1. Le modèle probabiliste Okapi présente une performance très attractive dans les interrogations unilingues selon les cinq langues étudiées (voir tableau 1) ;

2. Lors d'interrogations bilingues, la traduction manuelle s'avère meilleure que les systèmes de traduction automatique (voir tableau 2). Par contre, la différence de performance entre la machine et l'homme varie d'une langue à l'autre, de 6,1 % pour le français à 26,7 % pour l'italien ;

3. La qualité de la traduction fournie par la machine varie fortement d'une requête à l'autre, phénomène que l'on rencontre dans les quatre langues étudiées (voir tableau 3) ;

4. Si nous disposons d'un oracle retournant toujours la meilleure traduction automatique pour chaque requête, la performance obtenue s'avère aussi bonne qu'une traduction manuelle pour l'italien ou l'espagnol, voire légèrement supérieure pour le français et l'allemand (dernière ligne du tableau 2) ;

5. En recourant à quelques informations statistiques sur les requêtes traduites, notre approche basée sur la régression logistique permet de prédire les bonnes traductions à combiner pour obtenir une performance supérieure, en moyenne, au meilleur outil de traduction automatique.

Dans le contexte de la recherche bilingue ou multilingue d'information, nos expériences démontrent donc l'intérêt de pouvoir sélectionner les bonnes traductions et de les combiner pour permettre à l'ordinateur de mieux franchir la barrière des langues. Toutefois, la démarche que nous proposons ne s'avère guère utile dans les cas où le nombre de traducteurs est limité. Ainsi, pour les langues moins répandues, à l'image du suédois ou du finnois, le nombre d'outils de traduction automatique disponible reste limité, de même que pour des langues comme l'arabe ou le coréen, qui n'ont pas encore rencontré un grand intérêt de la part des services de traduction en-ligne.

Remerciements

Cette recherche a été subventionnée en partie par le Fonds National Suisse pour la Recherche Scientifique à l'aide du subside 21-66 742.01.

6. Bibliographie

Bookstein A., O'Neil E., Dillon M., Stephen D., « Applications of loglinear models for informetric phenomena », *Information Processing & Management*, vol. 28, n° 1, 1992, p. 75-88.

Buckley C., Singhal A., Mitra M., Salton G., « New retrieval approaches using SMART», *Proceedings of the TREC'4*, Gaithersburg, 1-3 November 1995, p. 25-48.

Chen A., Gey F.C., « Combining query translation and document translation in cross-language retrieval », *Notebook CLEF-2003*, Trondheim, 21-22 August 2003, p. 39-48.

Cronen-Townsend S., Zhou Y., Croft W.B., « Predicting query performance », *Proceedings of the ACM-SIGIR'2002*, Tampere, 11–15 August 2002, The ACM Press, New York, p. 299-306.

Gey F.C., « Inferring probability of relevance using the method of logistic regression », *Proceedings of the ACM-SIGIR'94*, Dublin, 3-6 July 1994, The ACM Press, New York, p. 222-231.

Hosmer D.W., Lemeshow S., *Applied Logistic Regression*, 2nd Ed., John Wiley, New York, 2000.

Kwok K.L., Grunfeld L., Dinstl N., Deng P., « TREC2003 robust, HARD and QA track experiments using PIRCS », *Notebook TREC 2003*, Gaithersburg, 11-15 November 2003, p. 201-209.

Le Calvé A., Savoy J., « Database merging strategy based on logistic regression », *Information Processing & Management*, vol. 36, n° 3, 2000, p. 341-359.

Peters C., Braschler M., Gonzalo J., Kluck M., *Evaluation of Cross-Language Information Retrieval*, Lecture Notes in Computer Science, vol. 2406, Springer-Verlag, Berlin, 2002.

Peters C., Braschler M., Gonzalo J., Kluck M., *Advances in Cross-Language Information Retrieval*, Lecture Notes in Computer Science, vol. 2785, Springer-Verlag, Berlin, 2003

Peters C., Borri F., *Results of the CLEF 2003 Cross Language System Evaluation Campaign*, Notebook of CLEF-2003, Trondheim, 2003.

Robertson S.E., Walker S., Beaulieu M., « Experimentation as a way of life: OKAPI at TREC », *Information Processing & Management*, vol. 36, n° 1, 2000, p. 95-108.

Salton G., McGill M.J., *Introduction to Modern Information Retrieval*, McGraw-Hill, New York, 1983.

Savoy J., Ndarugendamwo M., Vrajitoru D., « Report on the TREC-4 experiment: Combining probabilistic and vector-space schemes », *Proceedings of the TREC'4*, Gaithersburg, 1-3 November 1995, p. 537-547.

Singhal A., Choi J., Hindle D., Lewis D.D., Pereira F., « AT&T at TREC-7 », *Proceedings of the TREC-7*, Gaithersburg, 9-11 November 1998, p. 239-251.

Venaples W.N., Ripley B.D., *Modern Applied Statistics with S-PLUS*, Spinger-Verlag, New York, 1999.

Annexe 1. Formules de pondération

Afin d'attribuer un poids w_{ij} reflétant l'importance de chaque terme d'indexation T_j, j = 1, 2, ..., t, dans un document D_i, nous pouvons recourir à l'une des formules décrites dans le tableau ci-dessous. Dans cette dernière, tf_{ij} indique la fréquence d'occurrences du terme T_j dans le document D_i (ou dans la requête), n représente le nombre de documents D_i dans la collection, df_j le nombre de documents dans lesquels le terme T_j apparaît (fréquence documentaire), et idf_j l'inverse de la fréquence documentaire ($idf_j = \ln[n/df_j]$). Les constantes ont été fixées aux valeurs suivantes : slope = 0,2, pivot = 150, b = 0,75, k = 2, k_1 = 1,2, avdl = 900. De plus, la longueur du document D_i (ou le nombre de termes d'indexation associé à ce document) est notée par nt_i, la somme de valeurs tf_{ij} par l_i et $K = k \cdot [(1 - b) + b \cdot (l_i/avdl)]$.

bnn	$w_{ij} = 1$	nnn	$w_{ij} = tf_{ij}$
ltn	$w_{ij} = [\ln(tf_{ij}) + 1] \cdot idf_j$	atn	$w_{ij} = \left[0,5 + 0,5 \cdot \frac{tf_{ij}}{\max tf_i}\right] \cdot idf_j$
dtn	$w_{ij} = \ln[[\ln(tf_{ij}) + 1] + 1] \cdot idf_j$	npn	$w_{ij} = tf_{ij} \cdot \ln\left[\left(n - df_j\right) \Big/ df_j\right]$
Lnu	$w_{ij} = \frac{\left(1 + \ln(tf_{ij}) \Big/ \ln(\text{mean tf}) + 1\right)}{(1 - \text{slope}) \cdot \text{pivot} + \text{slope} \cdot nt_i}$	Okapi	$w_{ij} = \left((k_1 + 1) \cdot tf_{ij}\right) \Big/ \left(K + tf_{ij}\right)$
lnc	$w_{ij} = \frac{\ln(tf_{ij}) + 1}{\sqrt{\sum_{k=1}^{t} \left(\ln(tf_{ik}) + 1\right)^2}}$	ntc	$w_{ij} = \frac{tf_{ij} \cdot idf_j}{\sqrt{\sum_{k=1}^{t} \left(tf_{ik} \cdot idf_k\right)^2}}$
ltc	$w_{ij} = \frac{\left(\ln(tf_{ij}) + 1\right) \cdot idf_j}{\sqrt{\sum_{k=1}^{t} \left(\left(\ln(tf_{ik}) + 1\right) \cdot idf_k\right)^2}}$		
Lnu	$w_{ij} = \frac{\left(1 + \ln(tf_{ij}) \Big/ \ln(\text{mean tf}) + 1\right)}{(1 - \text{slope}) \cdot \text{pivot} + \text{slope} \cdot nt_i}$		

Tableau A.1. *Formules de pondération*

Annexe 2. Statistiques sur les collections

Corpus	Anglais	Français	Italien	Allemand	Espagnol
Taille en MB	425	243	278	527	509
nb de doc.	113'005	87'191	108'578	225'371	215'738
nb de formes	330'753	320'526	503'550	1'507'806	528'382
Nombre de formes par document					
moyenne	167,33	130,21	129,91	119,07	111,80
écart-type	126,31	109,15	97,60	109,73	55,40
médiane	138	95	92	89	99
maximum	1'812	1'622	1'394	2'420	642
minimum	2	3	1	1	5
Requêtes					
nb requêtes	89	99	96	99	99
nb doc. pert.	1'677	2'595	2'318	4'068	5'548
nb doc./requ.	18,84	26,21	24,15	41,09	56,04
nb max doc.	107 (n^0 50)	177 (n^0 95)	95 (n^0 50)	212 (n^0 42)	321 (n^0 95)
nb min doc.	1 (n^0 59)	1 (n^0 43)	2 (n^0 44)	1 (n^0 64)	1 (n^0 64)

Tableau A.2. *Quelques statistiques sur les collections utilisées.*