Traduction des requêtes pour la recherche d'information translinguistique anglais-arabe

Youssef Kadri & Jian-Yun Nie
Laboratoire RALI, Département d'informatique et de recherché opérationnelle
Université de Montréal
{kadriyou,nie} @iro.umontreal.ca

Résumé - Abstract

Nous traitons dans cet article le problème de la traduction des requêtes pour la Recherche d'Information Translinguistique (RIT). Le problème de la RIT consiste à trouver des documents en arabe avec des requêtes en anglais. La traduction des requêtes est une tâche essentielle. Notre approche de traduction de requêtes pour la RIT est basée sur l'entraînement d'un modèle de traduction statistique sur un corpus de textes parallèles extraits du Web. D'autres méthodes de traduction basées sur les textes parallèles et les dictionnaires bilingues sont aussi proposées. Une attention particulière sera mise sur le traitement morphologique de l'arabe pour la lemmatisation. Nos expérimentations montrent que, si on dispose de ressources multiples pour la traduction de requête, leur combinaison améliore grandement la performance de la Recherche d'Information (RI).

This paper deals with the problem of query translation for Cross-Language Information Retrieval (CLIR). Our CLIR system aims to retrieve documents written in Arabic using English queries. Query translation is an essential task. Our approach to query translation is based on statistical translation models trained on a corpus of parallel texts. This corpus is gathered automatically from the Web. Other translation methods based on bilingual dictionaries are also incorporated. A special emphasis is put on morphological processing for stemming Arabic words. Our experiments show that if multiple translation tools are available, a combination of them leads to a higher effectiveness of information retrieval.

Mots clés - Keywords

Recherche d'information translinguistique, traduction des requêtes, modèles de traduction statistique, dictionnaires bilingues, pages Web parallèles.

Cross-language information retrieval, query translation, statistical translation models, bilingual dictionaries, parallel web pages.

1 Introduction

Dans ce travail, nous étudions le problème de la Recherche d'Information Translinguistique (RIT) où les requêtes sont écrites en anglais et la collection de documents est en arabe. Mis à part les problèmes de la Recherche d'Information (RI) monolingue, la RIT doit résoudre en plus le problème de traduction de la requête vers la langue des documents. Diverses techniques ont été explorées dans la littérature pour cette fin : la traduction automatique, les dictionnaires bilingues et les corpus parallèles. Dans cet article, nous étudions les possibilités d'utiliser les méthodes existantes pour la RIT anglais-arabe. Par rapport à d'autres paires de langues étudiées,

le problème de la RIT anglais-arabe est le manque des ressources nécessaires. Même si ces ressources existent, elles ne sont pas suffisantes pour fournir une traduction de qualité qui répond au besoin de la RI. En effet, la plupart des expérimentations (Fraser, 2002) (Darwish et al, 2002) qui ont été réalisées à l'heure actuelle pour la RIT arabe, ont exploité des dictionnaires bilingues ainsi qu'un modèle de traduction statistique entraîné sur un corpus construit manuellement. Vu l'insuffisance des ressources de traduction pour cette paire de langues, notamment un corpus parallèle anglais-arabe, nous avons utilisé un système automatique pour fouiller le Web afin de trouver des pages Web parallèles. Ceci dans l'objectif de construire un corpus bilingue qui sera la base d'entraînement d'un modèle de traduction statistique. Toutefois, l'utilisation d'une seule ressource de traduction est souvent insuffisante du fait du contexte général des thèmes de la requête. Pour consolider la traduction des requêtes, nous avons exploré d'autres techniques de traduction basées sur les dictionnaires bilingues ainsi que les textes parallèles. Dans cet article, nous présentons plusieurs approches de traduction des requêtes. La prochaine section décrit les étapes nécessaires pour la construction d'un corpus parallèle à partir du Web ainsi que l'entraînement du modèle de traduction statistique. Après, nous présentons d'autres ressources de traduction basées sur les dictionnaires bilingues ainsi que les corpus parallèles. La section 4 décrit la manière de combiner toutes ces ressources de traduction. A la fin, nous présentons les résultats des expérimentations que nous avons entrepris sur une collection TREC[1].

2 Un modèle de traduction basé sur les pages Web parallèles

Un modèle de traduction statistique a besoin d'un grand nombre de textes parallèles (des textes en 2 langues simultanément pour son entraînement). La mise en application de tels modèles est généralement perturbée par l'absence ou l'insuffisance des textes parallèles pour plusieurs paires de langues, en particulier anglais-arabe. Afin de surmonter cet obstacle, Chen et Nie ont eu l'idée de fouiller le Web pour chercher automatiquement des pages Web parallèles à l'aide du système PTMiner (Nie, 1999) (Nie et al, 2000). Dans ce contexte, nous avons repris PTMiner pour construire un corpus bilingue anglais-arabe. Dans les sections suivantes, nous décrivons les différentes étapes pour construire un corpus parallèle, et pour entraîner un modèle de traduction statistique.

2.1 PTMiner

Nous avons exploité le système PTMiner pour chercher des textes parallèles anglais-arabes sur le Web. Ce système a été utilisé avec succès pour plusieurs paires de langues comme anglais-français ou anglais-chinois. Les principales étapes de ce processus de fouille (Chen et al, 2000) sont les suivantes : la sélection des sites Web candidats, le repérage de tous les documents dans les sites candidats et le repérage des paires de textes parallèles. Après application de ce processus pour plusieurs paires de langues, nous avons constitué un corpus modeste en taille (37 MO) pour la paire anglais-arabe. Le nombre de textes parallèles obtenus en anglais-arabe est beaucoup inférieur à ceux d'autres langues comme la paire anglais-français (300 MO). Ce nombre limité peut être expliqué par la rareté des sites Web pour cette paire de langues (anglais-arabe).
Des textes trouvés sur le Web sont encodés de différentes manières. Afin de les uniformiser, nous avons utilisé le système SILC[2] pour identifier la langue ainsi que le jeu de caractères dans

---

1 TREC (Text REtrieval Conference) est une série de conférences annuelles pour tester les performances des systèmes de RI sur de grands corpus (http://trec.nist.gov/).

2 http://www-rali.iro.umontreal.ca/SILC/

lequel un texte est représenté. Les textes avec un encodage non standard seront transformés en un encodage standard.

2.2 Lemmatisation

Les textes parallèles collectés du Web, subissent 2 procédures de lemmatisation avant leur utilisation pour l'entraînement des modèles de traduction. La lemmatisation consiste à transformer un mot en une forme "standard". Les documents anglais sont lemmatisés avec la méthode de Porter (Porter, 1980). Pour trouver les lemmes des mots anglais, Porter fait juste des troncatures sur les terminaisons de mots pour trouver une forme standard. Par exemple le mot 'needy' est remplacé par 'needi'. La lemmatisation de l'arabe est un traitement plus complexe dû à la morphologie très riche de l'arabe (Kadri, 1992). La forme des mots arabes peut avoir 4 catégories d'affixes : les antéfixes, les préfixes, les suffixes et les postfixes. Ainsi un mot arabe peut avoir une forme plus compliquée s'il y a présence de tous ces affixes attachés à sa forme standard. On peut les catégoriser selon leur rôle syntaxique. Les antéfixes sont généralement des prépositions. Les préfixes représentés par une seule lettre indiquent la personne de la conjugaison des verbes au présent. Les suffixes sont les terminaisons de conjugaison des verbes et les signes du pluriel et du féminin pour les noms. Enfin, les postfixes représentent des pronoms. Pour le besoin de la RI, nous avons opté pour une lemmatisation assouplie (Kadri, 2003) qui fait la troncature d'un ensemble restreint d'affixes. L'idée de cette lemmatisation est de tronquer quelques préfixes qui ne sont rien d'autres que des prépositions attachées aux mots, et quelques suffixes, étant généralement des pronoms accordés à la fin des mots. Pour ce faire, nous avons regroupé ces affixes dans 2 grandes classes : préfixes et suffixes, et nous avons établi des statistiques sur les fréquences d'occurrence de ces affixes sur le lexique des mots de la collection TREC. La liste de ces préfixes est la suivante :
(فب ,وبال ,فل ,ولل ,كال ,فال ,ول ,وب ,يال ,لل ,وال ,ل ,ب ,ال , ا ,و). Les suffixes que nous avons jugé nécessaire de tronquer sont les suivants : (ا ,ه ,ن ,ي ,ت ,ها ,ين ,و ,ان ,ات ,ون ,هم ,نا ,ك ,وا ,هما ,تي).
Notons aussi que les mots outils (les mots non porteurs de sens, comme "of", "the" en anglais et "و" en arabe) ont été éliminés des textes. Nous avons utilisé deux tables de mots outils respectivement pour l'arabe et l'anglais.

2.3 Alignement

Le point d'entrée de l'entraînement des modèles de traduction statistiques est un ensemble de phrases parallèles ou ce que l'on désigne aussi par les bitextes. On peut obtenir des bitextes à partir d'un corpus de textes parallèles des phrases. Pour nos expérimentations, nous avons utilisé un outil d'alignement basé sur les cognates (Simard et al, 1992). L'idée de cet alignement est que 2 phrases en relation de traduction partagent souvent des mots communs ou proches comme les noms propres, les symboles, les chiffres ou tout simplement partagent une forme identique dans les 2 langues (par exemple, linguistique v.s. linguistics en français et en anglais). Deux phrases dans les deux langues respectives contenant de tels éléments ont une forte chance de s'aligner. Ce facteur est combiné avec d'autres critères, notamment l'ordre de phrases et la longueur de phrase, dans un processus d'alignement qui détermine un ensemble d'alignements les plus probables pour chaque paire de textes parallèles.

2.4 Modèles de traduction probabiliste IBM

En se basant sur les phrases alignées, c'est au tour des mots d'être alignés. Ceci revient à estimer les relations de traduction de mots. Cette tâche est l'objet des modèles de traduction statistiques. Brown propose 5 modèles de traduction (Brown et al, 1993). Chaque modèle a sa propre prescription pour calculer la probabilité conditionnelle P(a | e) qu'on appelle la probabilité de

traduction du mot source *e* en mot cible *a*. Au début, chaque mot dans une phrase en langue cible est supposé être une traduction possible de chaque mot dans la phrase parallèle en langue source correspondante. Intuitivement, plus une paire de mots apparaît dans des phrases parallèles, plus la chance est meilleure que les 2 mots de cette paire soient la traduction l'un de l'autre. De cette façon, la probabilité initiale de traduction P(*a* | *e*) est calculée comme $\frac{\#(a,e)}{\#(e)}$ où #(*a*,*e*) est le nombre de paires de phrases contenant respectivement *a* et *e*, et #(*e*) le nombre de phrases contenant *e*. Ces probabilités initiales sont soumises ensuite à un processus de maximisation d'espérance (EM) afin de maximiser les probabilités d'aligner les phrases parallèles du corpus. EM est un algorithme itératif utilisé pour la recherche du paramètre réalisant le maximum de vraisemblance. Cet algorithme est la base d'estimation des paramètres des modèles IBM (Brown et al, 1993). A la fin, nous aurons une fonction de probabilité P(*a* | *e*) exprimant la probabilité du mot *a* qu'il soit la traduction du mot *e*. Avec cette fonction, nous pouvons déterminer un ensemble de traductions probables dans la langue cible pour chaque mot de la langue source. Pour l'entraînement de ces modèles, nous avons utilisé le système GIZA (Al-Onaizan et al, 1999) qui implémente les modèles IBM 1, 2 et 3. Pour le besoin de la RI, le modèle IBM 1 est suffisant car la RI ne tient pas compte de l'ordre des mots, comme dans le modèle IBM 1.

3 Identification d'autres ressources de traduction

La deuxième ressource est un modèle de traduction bâti par l'équipe BBN (Fraser, 2002) à partir d'une large collection de textes bilingues alignés anglais-arabe. Le corpus est construit manuellement à partir des archives des Nations Unies. Il renferme 38 000 paires de documents. Ce modèle est entraîné avec GIZA++ (Och et al, 2000) qui est une extension du GIZA. Les autres sources de traduction que nous avons exploitées sont des systèmes de traduction en ligne sur le Web. Parmi une variété disponible sur le Web, nous avons sélectionné 2 sites : Almisbar[3] et Ajeeb[4] qui offrent une meilleure qualité de traduction. Ces deux sites sont utilisés comme des dictionnaires bilingues, qui nous suggèrent des traductions pour chaque mot source.

4 Combinaison des ressources

L'utilisation de plusieurs sources de traduction donne plus de confiance à la qualité et oriente le processus de traduction positivement surtout dans le cas d'une couverture réduite i.e. une source de traduction prise séparément ne répond pas favorablement à une entrée d'une requête. Plus précisément, nous avons réuni 4 traductions différentes entre l'anglais et l'arabe provenant de 4 sources différentes. La question qui se pose pertinemment à ce stade est comment combiner ces ressources potentielles pour augmenter la performance de la RIT ?

Dans une première étape, nous avons considéré les ressources de traduction séparément pour nos expérimentations. Ainsi pour chaque ressource, nous avons pris 5 traductions pour chaque mot de la requête. Dans la deuxième étape, nous avons croisé les 4 modèles d'une façon linéaire. C'est-à-dire nous avons retenu 4 traductions pour chaque terme de la requête, chacune provenant d'un modèle. Pour les modèles probabilistes, on n'a considéré que la traduction dont la probabilité P(*a* | *e*) est la plus élevée. Cette façon de combiner les modèles ressemble à la technique d'expansion de requête où un terme de la requête est enrichi par plusieurs synonymes. Après application de cette combinaison, on a constaté une nette augmentation de la performance de RIT mesurée en précision moyenne (Tableau 2). La précision moyenne est la moyenne des précisions sur les 11 points du rappel (Salton, 1971). C'est une mesure standard utilisée en RI.

---

[3] http://www.almisbar.com/

[4] http://www.ajeeb.com/

5 Résultats et discussion

Pour nos expérimentations, nous avons utilisé la collection de test du TREC 2001 qui contient 383 872 articles de journaux. D'autre part, nous avons évalué les 25 requêtes provenant de la même collection TREC. Le modèle vectoriel Smart (Salton, 1971) est adapté aux textes arabes sans difficulté pour l'indexation et la recherche. Avant l'indexation de la collection, nous avons appliqué quelques normalisations morphologiques sur les caractères arabes (Kadri, 2003) parce que certains caractères ont plus d'une forme (par exemple la lettre 'ا'). Les documents sont lemmatisés par la technique de lemmatisation assouplie et les mots outils sont éliminés. Les requêtes sont traduites par le biais des 4 ressources de traduction que nous avons identifiées. Les 2 collections de documents et de requêtes sont soumises à Smart pour l'indexation. Enfin, Smart classe les documents pour chaque requête selon la similarité cosinus (Salton, 1971) entre les termes de la requête avec ceux des documents. Le tableau 2 montre les performances de recherche d'information en terme de précision moyenne. Nous remarquons clairement que l'avantage est au coté des dictionnaires bilingues. Ceci est dû à la couverture limitée des termes des requêtes par les modèles de traduction construits à base des textes parallèles. D'autre part, le nombre réduit des paires de textes parallèles recueillis à partir du Web (2 816) comparé à 38 000 paires du corpus des Nations Unies, explique le score faible de précision moyenne (8 %) obtenue dans Test 1 par rapport à 13 % du modèle statistique entraîné sur le corpus des Nations Unies (Test 2).

Sur les 2 tableaux ci-dessous, R.P. représente la rétroaction de pertinence. Cette technique consiste à supposer les 10 premiers documents retrouvés dans la première recherche comme pertinents[5] et sélectionner après les 20 termes de poids forts parmi ces 10 documents présumés pertinents. Ces nouveaux termes sont injectés dans la requête initiale pour construire une requête étendue. La technique de Rocchio (Rocchio, 1971) a été reprise pour calculer les poids des termes des nouvelles requêtes. Nous observons que la prise en considération de ce facteur augmente nettement les performances dans les deux cas monolingue et translinguistique.

Précision moyenne	
Sans R.P.	Avec R.P.
0.20	0.35

Tableau 1 : Performances de la RI monolingue arabe

Run-id	Modèle de traduction	Précision moyenne	
		Sans R.P.	Avec R.P.
Test 1	Pages Web parallèles	0.08	0.22
Test 2	Corpus parallèle des nations unies	0.13	0.25
Test 3	Almisbar	0.16	0.25
Test 4	Ajeeb	0.16	0.28
Test 5	Combinaison linéaire	0.16	0.35

Tableau 2 : Performances de la RIT anglais-arabe selon plusieurs modèles de traduction

Le Test 5 du tableau 2 montre bien que la combinaison des modèles de traduction augmente les performances de la précision moyenne surtout quand la rétroaction de pertinence est ajoutée. La précision moyenne de recherche de la RIT a atteint 80 % de la performance de celle de la RI monolingue (Tableau 1). Si la rétroaction de pertinence est appliquée, les performances sont presque les mêmes (35 % de précision moyenne).

[5] Un document est pertinent pour une requête s'ils ont la même topicalité i.e. ils traitent le même sujet.

6 Conclusion

Dans cet article, nous nous sommes intéressés à la problématique de recherche des documents dans une collection représentée en langue arabe en utilisant des requêtes en anglais. Nous sommes partis à investiguer ce domaine encore en stade d'exploration vu les ressources langagières limitées pour cette paire de langues. Nous avons implanté un modèle de traduction statistique basé sur les pages Web parallèles pour la traduction des requêtes. En outre, d'autres sources de traduction ont été identifiées. Pour nos expérimentations, chaque méthode de traduction a été exploitée séparément ensuite une combinaison linéaire de toutes les sources de traduction identifiées a été proposée. Nous avons remarqué que cette combinaison augmente les performances de recherche des documents pertinents. Dans le futur, nous étudierons d'autres méthodes de combinaison des sources de traduction de requêtes, notamment la façon d'attribuer un facteur de confiance à chaque ressource de traduction utilisée.

Bibliographie

Chen J., Nie J.N. (2000), "Parrallel Web text mining for cross-language", RIAO, Paris, pp. 62--77.

Fraser A., Xu J. Weischedel R. (2002), "Trec 2002 Cross-lingual Retrieval at BBN", Trec11 conference.

Nie J. Y. (1999), "CLIR using a Probabilistic Translation Model based on Web Documents", Trec8. http://trec.nist.gov/pubs/trec8/t8_proceedings.html

Nie J. Y., Simard M., Foster G. (2000), "Multilingual information retrieval based on parallel texts from the web", CLEF2000, Lisbon, ed. C. Peters, LNCS 2069, Springer, pp. 188-201.

Porter M. (1980), "An algorithm for suffix stripping", Automated Library and Information Systems, 14(3), pp. 130--137.

Kadri Y., Benyamina A. (1992), "Un système d'analyse syntaxico-sémantique du langage arabe non voyellé", Mémoire d'ingénieur, Université d'Oran.

Kadri Y. (2003), "Recherche d'information translinguistique sur les documents en arabe", Rapport de prédoc oral, DIRO, Université de Montréal.

Simard M, Foster G., Isabelle P. (1992), "Using Cognates to Align Sentences in Bilingual Corpora", Proceedings of TMI, Montréal, Québec.

Brown P. F., Pietra S. A., Pietra V. J., Mercer R. L. (1993), "The mathematics of statistical machine translation : Parameter estimation". Computational Linguistics, 19(2), 263–311.

Al-Onaizan Y., Curin J., Jahr M., Knight K., Lafferty J., Melamed D., Och F., Purdy D., Smith N., Yarowsky D. (1999), "Statistical Machine Translation", Report, JHU 99 Workshop, Baltimore, MD.

Och F. J., Ney H. (2000), "Improved statistical alignment models", In ACL'00, pages 440--447.

Darwish K, Oard D. W. (2002), "CLIR experiments at Maryland for Trec-2002 : Evidence combination for Arabic-English retrieval", Trec11 conference.

Salton G (1971), "The SMART Retrieval System - Experiment in Automatic Document Processing", Prentice-Hall, Englewood Cliffs, New Jersey.

Rocchio J. (1971), "Relevance feedback in information retrieval", Prentice-Hall, Inc.