

Multilinguisme et question-réponse: adaptation d'un système monolingue

Luc Plamondon (1), George Foster (2)

(1) RALI - Université de Montréal
Montréal, Québec, Canada
plamondl@iro.umontreal.ca

(2) RALI - Université de Montréal
Montréal, Québec, Canada
foster@iro.umontreal.ca

Mots-clefs – Keywords

Question-réponse multilingue, extraction d'information, recherche d'information translinguistique, traduction automatique
Multilingual Question Answering, Information Extraction, CLIR, Machine Translation

Résumé - Abstract

Nous montrons comment il est possible de modifier un système de question-réponse monolingue anglais afin de répondre à des questions posées en français. Nous avons écrit de nouvelles règles afin d'analyser la question en français et nous avons utilisé un moteur de traduction probabiliste pour traduire les termes-clés. Si l'implantation du bilinguisme a entraîné une diminution de performance, cette approche donne de meilleurs résultats qu'un système monolingue couplé à un système de traduction automatique pour traduire les questions.

We describe a method for modifying a monolingual English question-answering system in order to accept French questions. Our method relies on a statistical translation engine to translate keywords, and a set of manually-written rules for analyzing French questions. Although adding bilingualism causes a drop in performance, our method yields better results than a baseline approach of translating French questions automatically before submitting them to the original monolingual system.

1 Introduction

Un système de question-réponse (QR) est un type particulier de moteur de recherche qui permet à un utilisateur de poser une question en langue naturelle, plutôt que de l'obliger à construire une requête en un langage artificiel fait de ET, OU et autres opérateurs. De plus, alors qu'un moteur de recherche classique propose à l'utilisateur une liste des documents les plus pertinents à sa requête, le système de QR extrait lui-même les réponses aux questions. La recherche s'effectue dans un ensemble de documents ou de pages web.

Clarke et ses collaborateurs ont montré que pour des collections de textes totalisant moins de 500 gigaoctets (100 milliards de mots), plus la collection est volumineuse, meilleures sont les performances de leur système de QR (Clarke *et al.*, 2002). Si l'on suppose que les anglophones ont accès à environ 10 fois¹ plus de documents numérisés — pages web, encyclopédies sur cédéroms, etc. — que les francophones, il ne fait aucun doute qu'un système de QR destiné aux francophones mais capable de fouiller des textes en anglais ouvre des perspectives intéressantes à la fois au niveau de la quantité des sujets couverts et de la qualité des réponses trouvées.

Dans le cadre des campagnes d'évaluation TREC, nous avons développé le système de QR Quantum (Plamondon *et al.*, 2002). Ce système fonctionne uniquement en anglais: la question doit être posée en anglais, les documents de la collection sont entièrement en anglais et l'extraction des réponses nécessite des ressources linguistiques en anglais. Nous avons récemment entrepris de rendre Quantum bilingue, de sorte qu'un utilisateur francophone puisse poser sa question en français et obtenir une réponse en français extraite d'une collection de textes en anglais. Nous mettrons en évidence les problèmes que pose le multilinguisme en QR et nous décrirons comment nous avons choisi de les résoudre dans Quantum.

2 Travaux antérieurs

La campagne d'évaluation TREC-8 (*Text Retrieval Conference*, 1999) et les suivantes ont favorisé la création de quantité de systèmes de QR pour l'anglais. Bien que les techniques diffèrent d'un système à l'autre, la plupart suivent ces trois étapes: analyse de la question pour déterminer le type de réponse attendu, filtrage de la collection de textes pour ne conserver que les plus pertinents et identification de la réponse exacte.

Il existe aussi des systèmes monolingues pour l'italien (Magnini *et al.*, 2001), le polonais (Vetulani, 2002), le japonais (Seki & Harada, 2002) et le coréen (Kim & Seo, 2002). Quant aux systèmes pour le français, la campagne d'évaluation EQueR en cours de préparation, dans le cadre du projet EVALDA de Technolanguage/ELDA, donnera sans doute lieu à la même récolte que les campagnes TREC. Il n'existe pas encore, à notre connaissance, de système multilingue qui permette de poser une question dans une langue différente de la collection de textes et d'obtenir la réponse dans la même langue que la question. Le nouveau volet de la campagne d'évaluation annuelle CLEF destiné aux systèmes de QR multilingues favorisera sûrement la conception de systèmes de ce genre (la traduction de la réponse n'étant toutefois pas à l'ordre du jour).

¹Estimation basée sur la proportion de pages web en chacune de ces langues: selon la firme Global Reach, spécialisée en marketing en ligne international, 40,2% de toutes les pages web sont en anglais contre 3,9% en français (www.global-reach.biz/globstats).

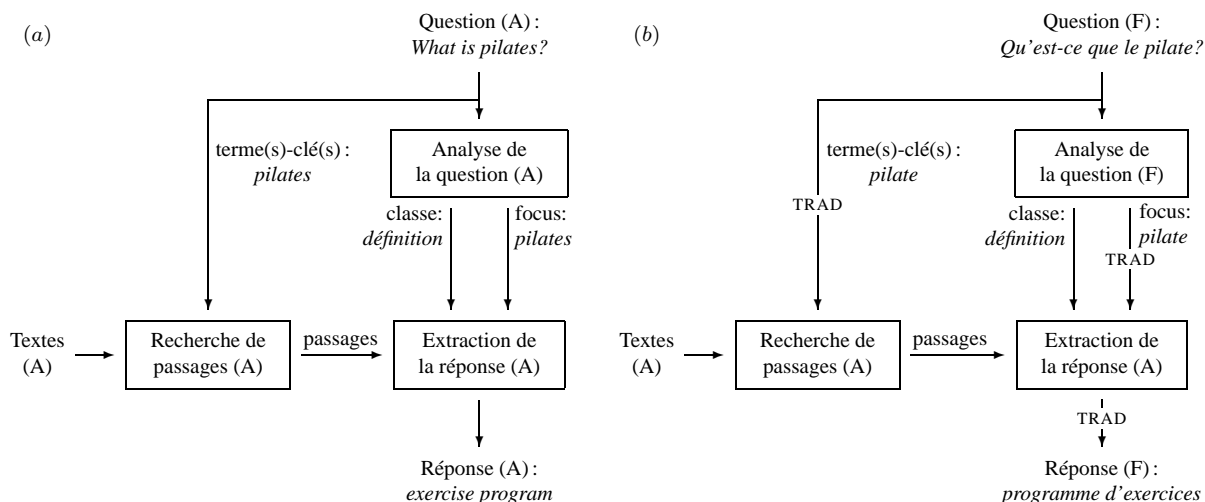


Figure 1: Architecture des deux versions de Quantum. (a) Version monolingue: la question, les textes et la réponse sont en anglais (b) Version bilingue: la question et la réponse sont en français (F). Le module d'analyse des questions a été francisé et les autres modules opèrent en anglais (A). La traduction se fait en 3 points: termes-clés, focus et réponse.

3 Système monolingue

Quantum a d'abord été développé en fonction des campagnes TREC. Il a été conçu pour répondre à des questions en anglais, simples, courtes, factuelles (c'est-à-dire que les réponses sont la plupart du temps des entités nommées) et syntaxiquement bien formées comme *What is pilates? Who was the architect of Central Park? How wide is the Atlantic Ocean? At what speed does the Earth revolve around the sun? Where is the French consulate in New York?* La collection de textes est constituée d'environ un million de dépêches de différentes agences de presse de langue anglaise, pour un total de 3 gigaoctets de données (600 millions de mots). Les réponses sont extraites des documents et doivent être *exactes*, c'est-à-dire que la chaîne de caractères suggérée en guise de réponse ne doit contenir rien de plus ou de moins que la réponse. Pour de plus amples détails, le lecteur est invité à consulter la description de la piste *Question Answering* de TREC-11 (Voorhees, 2002).

L'architecture du système monolingue est exposée à la figure 1a. Nous ne décrivons ici que les éléments pertinents à l'implantation du bilinguisme (figure 1b).

3.1 Analyse de la question

Le but de l'analyse de la question est de déterminer le type de la réponse à chercher, donc le mécanisme d'extraction — ou fonction d'extraction — à utiliser. Parfois, il est nécessaire de transmettre à cette fonction un paramètre appelé *focus* de la question: il s'agit d'un mot ou groupe de mots présents dans la question et ayant un rapport étroit avec la réponse. Par exemple, la réponse à la question *What card game uses only 48 cards?* doit être un hyponyme du focus *card game*, c'est pourquoi il est nécessaire de transmettre ce paramètre à la fonction d'extraction

choisie. Quant à la réponse à la question *How many black keys are on the piano?*, elle doit contenir un nombre suivi d'une répétition du focus *black keys*. Cependant, certaines autres questions comme *When was water found on Mars?* ne nécessitent pas l'identification d'un focus puisqu'il s'agit ici de trouver une entité nommée de temps pour compléter le concept désigné par *when*. Peu importe qu'ils soient focus ou non, tous les mots de la question contribuent à l'identification de la réponse par le biais du score de la recherche d'information (section 3.2). Il est important de souligner que dans la classification des questions que nous avons retenue, la nécessité ou non de passer un focus à telle ou telle fonction d'extraction ainsi que les critères de sélection du focus sont motivés par des considérations purement techniques et propres au fonctionnement interne de Quantum. Une classification plus générale basée sur des critères psycho-linguistiques rigoureux a été faite par Graesser (Graesser *et al.*, 1992).

L'analyse de la question s'effectue à l'aide d'un segmenteur de mots, d'un étiqueteur grammatical statistique ainsi que d'un segmenteur de groupes nominaux basé sur les étiquettes grammaticales. Nous avons élaboré une soixantaine de règles d'analyse utilisant à la fois des mots, des étiquettes grammaticales et des étiquettes de groupes nominaux (Plamondon, 2002). Par exemple, la règle d'analyse qui s'applique à la question de la figure 1a est la suivante:

what BE <syntagme nominal SN₁> → type = *définition*, focus = SN₁

3.2 Filtrage des passages les plus pertinents

Les mécanismes d'extraction de réponse sont trop complexes pour être appliqués à la totalité de la collection de textes à chaque question. Afin de réduire l'espace de recherche de 3 milliards de caractères à quelques milliers, nous utilisons un moteur de recherche conventionnel qui ordonne les passages les plus pertinents. Nous utilisons à cet effet le moteur Okapi car il peut raffiner sa recherche jusqu'à suggérer les paragraphes les plus pertinents, à la différence des autres moteurs de recherche qui s'en tiennent à une liste de documents complets. Il nous suffit ensuite de conserver les 20 meilleurs paragraphes et le score de pertinence qu'Okapi leur attribue. La requête est formée de tous les mots de la question; Okapi se charge de les tronquer et d'éliminer les mots vides (*stopwords*) avant d'effectuer la recherche. La requête de l'exemple de la figure 1a est composée du seul mot de la question qui ne soit pas vide de sens: *pilates*.

3.3 Extraction de la réponse

La fonction d'extraction choisie lors de l'analyse de la question, paramétrée ou non avec le focus, est appliquée aux paragraphes les plus pertinents (Plamondon, 2002). Trois techniques ou outils peuvent être utilisés par l'une ou l'autre des 12 fonctions d'extraction: des expressions régulières, le réseau sémantique Wordnet et l'extracteur d'entités nommées Annie de la suite de développement GATE. Par exemple, il est souhaitable que la réponse à une demande de définition contienne un hypéronyme du focus; ainsi, pour répondre à la question *What is ouzo?*, Wordnet nous permet de vérifier que la réponse *liquor* est bien un hypéronyme du focus *ouzo*. Cependant, dans l'exemple à la figure 1a, la réponse doit être trouvée à l'aide d'expressions régulières car Wordnet n'a pas d'entrée *pilates*, le focus de la question.

Chaque groupe nominal dans les paragraphes retenus se voit attribuer un score d'extraction selon qu'il satisfait aux critères de la fonction d'extraction. Ce score d'extraction est combiné

au score de pertinence du paragraphe afin de tenir compte de la densité des mots de la question se trouvant autour du groupe nominal examiné. Le meilleur groupe nominal est sélectionné pour constituer la réponse. Nous avons choisi de considérer les groupes nominaux comme unités de base de réponse car il s'avère que seulement 2 % des questions des campagnes TREC ne peuvent être répondues à l'aide d'un groupe nominal.

4 Implantation du bilinguisme

Pour Quantum comme pour bien d'autres systèmes de QR, l'étape d'extraction de la réponse est l'étape la plus complexe. Elle est donc déterminante dans le choix de l'approche à favoriser lorsqu'il s'agit d'adapter un système monolingue à des questions ou à une collection de textes d'une autre langue. Deux facteurs sont à considérer plus particulièrement: la disponibilité de ressources linguistiques nécessaires à l'extraction de la réponse et la facilité d'implantation du multilinguisme.

Dans le cas de Quantum, ces deux facteurs ont joué en faveur de la conservation du module d'extraction en anglais (figure 2a) et, par conséquent, en faveur de la traduction de la question et de la réponse plutôt que des textes. D'une part, les ressources linguistiques sont généralement plus nombreuses et plus poussées en anglais qu'en français, si ce n'est qu'elles sont disponibles gratuitement (c'est le cas de Wordnet et de l'extracteur d'entités nommées Annie auxquels Quantum fait appel). D'autre part, en laissant le module d'extraction dans sa langue d'origine, l'implantation du bilinguisme s'en trouve facilitée. En effet, si nous avons plutôt choisi de transposer le module d'extraction dans la même langue que celle de la question (figure 2b), il aurait fallu trouver de nouvelles ressources linguistiques, adapter le système à leur interface et traduire des textes entiers de l'anglais au français, ce qui pour le moment demeure une entreprise longue et ardue. Par contre, ne traduire que la question et la réponse est plus facile; nous verrons plus loin qu'il n'est pas nécessaire de parvenir à une traduction syntaxiquement correcte de la question et que la traduction de la réponse peut tirer parti du contexte particulier de la QR.

Cependant, dans le cas où la langue d'origine du système s'avère être la plus pauvre en ressources linguistiques, comme ce pourrait être le cas si l'on voulait rendre bilingue un système monolingue français, les deux facteurs à considérer jouent l'un contre l'autre. Il faudrait choisir entre une implantation rapide en laissant le coeur du système en français (figure 2b) et un gain potentiel de performance en transformant son coeur pour qu'il utilise des ressources linguistiques en anglais (figure 2a).

Si l'enjeu était plutôt d'adapter un système monolingue anglais non pas à des questions posées dans une langue différente mais bien à des textes écrits dans une langue différente, des considérations supplémentaires devraient être prises en compte. En effet, laisser le coeur dans sa langue d'origine (figure 3a) faciliterait l'implantation du bilinguisme mais obligerait à traduire de longs textes. Cependant, en utilisant des moteurs de traduction probabilistes, il serait relativement simple d'adapter le système à une collection regroupant des textes en plusieurs langues. L'approche 3b, quant à elle, dépendrait moins du moteur de traduction mais nécessiterait un système différent pour chaque langue rencontrée dans la collection.

Dans les scénarios présentés, le coeur du système est considéré comme une boîte noire hermétique au processus de traduction; ceci suppose une traduction parfaite des questions ou des textes. Les méthodes de traduction actuelles ne le permettent évidemment pas. Il est donc par-

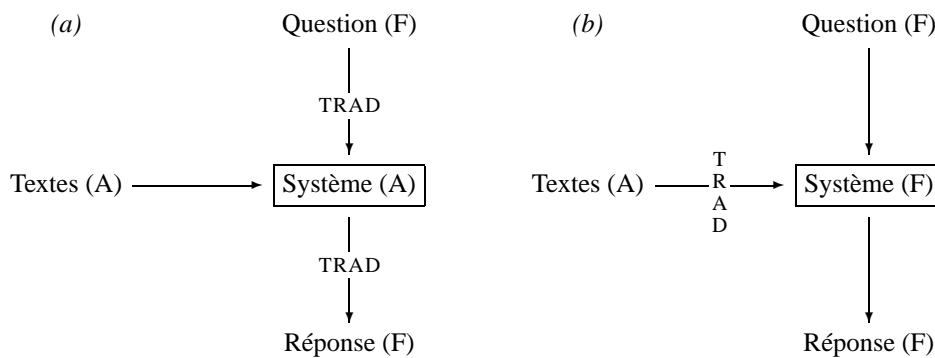


Figure 2: Deux approches pour l'adaptation d'un système monolingue anglais (A) pour répondre à des questions en français (F). En (a), le coeur du système demeure intact et des ressources linguistiques en anglais peuvent être utilisées. En (b), le coeur du système est transposé dans l'autre langue et les textes sont traduits.

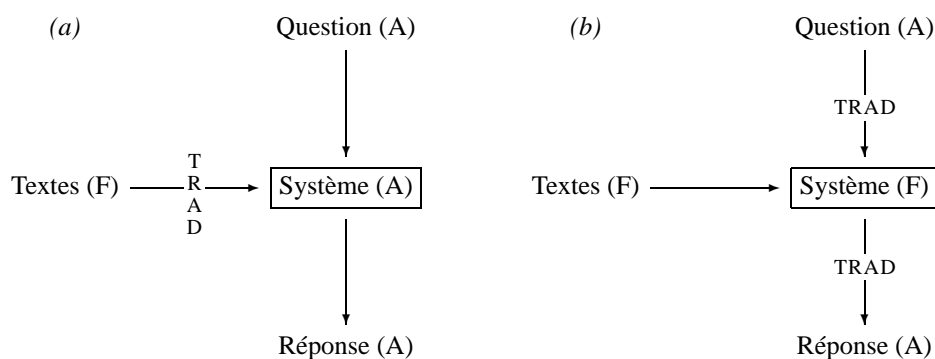


Figure 3: Deux approches pour l'adaptation d'un système monolingue anglais (A) pour fouiller des textes en français (F). En (a), le coeur du système demeure intact mais les textes doivent être traduits. En (b), le coeur du système est transposé dans l'autre langue.

fois avantageux d'*ouvrir* le système afin de tirer parti d'un maximum d'informations pouvant faciliter le processus de traduction. Cela permet entre autres de contourner le problème complexe de traduction de la question en une question anglaise syntaxiquement bien formée. Qui plus est, des techniques de traduction différentes peuvent être employées selon la tâche.

Pour ce faire, nous avons d'abord remplacé le module d'analyse de la question afin que l'analyse se fasse en français (figure 1b) car les méthodes de traduction probabilistes dont nous disposons sont peu performantes en regard de l'ordonnancement syntaxique des constituants de la phrase et les règles du module d'analyse ne pourraient s'appliquer adéquatement. La fonction d'extraction choisie par le module d'analyse peut être transmise directement au module d'extraction; seul le focus, lorsque la fonction d'extraction en requiert un, nécessite d'être traduit vers l'anglais pour les besoins du module d'extraction (entre autres parce que le focus doit être situé dans l'ontologie de Wordnet). Quant au module de recherche de passages, il requiert que les termes-clés de la question soient traduits vers l'anglais. Le module d'extraction de la réponse, lui, ne nécessite aucune modification. Examinons ces considérations en détail.

4.1 Conversion du module d'analyse des questions et traduction du focus

L'analyse des questions par Quantum se fait à l'aide d'expressions régulières combinant mots et étiquettes grammaticales. Le module anglais comprend environ 60 règles; nous en avons écrit à peu près le même nombre pour le français afin d'obtenir une couverture que nous espérons équivalente (nous présenterons une évaluation du module plus loin).

Pour écrire les règles d'analyse, nous avons suivi la même méthodologie pour l'anglais et le français: nous sommes partis du nombre limité de pronoms et adjectifs interrogatifs propres à la langue et nous avons recensé les constructions syntaxiques qu'ils commandent. Le français s'est avéré plus difficile à analyser car nous avons noté une plus grande flexibilité dans la formulation des questions. Par exemple, *How much does one ton of cement cost* peut se dire de deux façons en français: *Combien coûte une tonne de ciment* ou *Combien une tonne de ciment coûte-t-elle*. De plus, les pronoms interrogatifs — à la base de l'analyse — ne se traduisent pas nécessairement de la même façon: c'est le cas de *what* qui peut se traduire par *qu'est-ce que* dans *What is leukemia / Qu'est-ce que la leucémie*, par *que* dans *What does target heart rate mean / Que signifie rythme cardiaque cible* et par *quoi* dans *Italy is the largest producer of what / L'Italie est le plus grand producteur de quoi*. L'adjectif interrogatif *which*, comme dans *which city*, peut quant à lui prendre la marque du féminin et du pluriel: *quel, quels, quelle, quelles*. À ces difficultés s'ajoutent d'autres particularités de la langue française: le *t* euphonique de la forme interrogative de certains verbes à la 3^e personne du singulier (*Combien une tonne coûte-t-elle* mais pas *Combien deux tonnes coûtent-elles*), l'élosion (*Qu'appelle-t-on*) et deux formes de passé (*Quand le téléphone fut-il inventé / Quand le téléphone a-t-il été inventé* alors que seule la forme *When was the telephone invented* est acceptable en anglais). De plus, les questions similaires à *Dans quelle série télévisée Pierce Brosnan a-t-il joué* présentent une difficulté au niveau de la segmentation des groupes nominaux. Il est en effet difficile de segmenter correctement *série télévisée* et *Pierce Brosnan*, alors qu'en anglais ces deux groupes nominaux sont la plupart du temps séparés par un auxiliaire: *What TV series did Pierce Brosnan play in*.

En plus de déterminer la fonction d'extraction à utiliser, les règles d'analyse servent aussi à identifier le focus. La nature sémantique du focus a parfois une influence sur le type de réponse attendu. C'est le cas de la question *Quel auteur a écrit sous le nom de plume "Boz"*, dont le focus *auteur* indique que la réponse doit être un nom de personne. Pour effectuer ce lien, nous consultons l'ontologie de Wordnet. Le focus doit donc être traduit en anglais avant que l'analyse de la question puisse être définitive. Pour ce faire, nous utilisons un moteur de traduction probabiliste IBM2 entraîné sur un ensemble de textes parallèles composés de débats de la Chambre des communes du Canada, de bulletins d'information de l'*Union européenne en ligne* et d'un échantillon de questions de TREC. Le modèle IBM2 tient compte de la position des mots dans la phrase source, ce qui permet de déterminer la traduction la plus probable compte tenu du focus et, dans une moindre mesure, des autres mots de la phrase source. Nous ne conservons que la meilleure traduction qui soit un nom.

Pour comparer les modules monolingue et bilingue, nous les testons sur un ensemble de questions tirées des campagnes TREC. Les 4 campagnes TREC-8 à TREC-11 ont donné lieu à la création d'un corpus de 1893 questions, chacune accompagnée des réponses trouvées dans la collection de textes par les systèmes participants et jugées correctes par des juges. Nous avons fait traduire ces 1893 questions en français afin de tester la version bilingue du système. Ce corpus de questions bilingue est disponible sur notre site web². Pour bâtir notre ensemble de test,

²www-rali.iro.umontreal.ca/ProjetLUB.fr.html

nous avons éliminé les 200 questions de TREC-8 à cause de la collection de textes différente utilisée lors de cette campagne. Des 1693 questions restantes, 114 n'ont pas de réponses, soit parce que les questions ont été éliminées des campagnes pour causes d'irrégularités, soit parce qu'elles n'ont pas de réponse connue dans la collection de textes. Nous avons sélectionné au hasard 789 questions parmi les 1579 restantes, c'est-à-dire la moitié.

Nous avons mesuré que les expressions régulières (utilisées en conjonction avec l'ontologie de Wordnet) permettent de choisir la bonne fonction d'extraction pour 96 % des questions en anglais. Cette performance diminue à 77 % pour les questions en français. La dégradation de 20 % par rapport au système monolingue est due autant à la couverture des expressions régulières qu'à la traduction du focus. Le focus est traduit de façon satisfaisante une fois sur deux; la cause d'erreur principale est l'absence du mot à traduire dans le corpus d'entraînement. Notamment, les demandes de définition portent sur des expressions spécialisées: *Qu'est-ce que la thalassémie, l'amoxicilline, un shaman, etc.* Ces expressions étant inconnues du moteur de traduction et la recherche de définitions s'appuyant principalement sur la recherche d'hypéronymes du focus, il s'avère impossible d'utiliser Wordnet pour déterminer que les hypéronymes *anemia, antibiotic* et *spiritual leader* rencontrés dans le texte constituent des réponses correctes.

De la même façon, la recherche d'un hyponyme du focus pour répondre à une question de spécialisation du genre *Quelle fleur Vincent Van Gogh a-t-il peint? → le tournesol* est sérieusement compromise si le focus *fleur* est faussement traduit par *horse* ou si le focus est tout simplement inconnu du moteur de traduction. Mais dans ce cas-ci, c'est l'analyse de la question qui a fait échouer le repérage de la réponse car le focus a été mal identifié dès le départ, à cause d'une erreur de segmentation de *fleur Vincent Van Gogh* en deux groupes nominaux.

4.2 Traduction des termes-clés pour la recherche de passages

Indépendamment du domaine de la QR, beaucoup de chercheurs se sont penchés sur le problème de la recherche d'information translinguistique. La tendance est actuellement de fondre le modèle de traduction avec le modèle de recherche d'information (Kraaij *et al.*, 2003). Cependant, étant donné que notre moteur Okapi ne nous permet pas de modifier le modèle de recherche, nous avons opté pour l'approche classique: d'abord obtenir des termes-clés par traduction et ensuite s'en servir comme requête pour la recherche de passages.

Premièrement, nous utilisons un moteur de traduction IBM1 pour obtenir les termes ayant la plus forte probabilité de faire partie d'une traduction de la question. À la différence du modèle IBM2 utilisé pour traduire le focus, le modèle IBM1 ne tient pas compte de la position des mots de la question: chacun des mots de la question contribue de façon égale à produire les termes en français les plus probables. Les termes générés et qui ne sont pas vides de sens constituent les termes-clés de la requête. Nos expériences ont montré que la recherche translinguistique est plus fructueuse lorsque la requête en français contient autant de termes-clés que la question en anglais en contient, c'est-à-dire en moyenne 5.

Nous avons testé le module de recherche d'information translinguistique avec le même ensemble de questions que pour le module d'analyse des questions. La recherche s'est effectuée dans la collection de textes propre à TREC-9/TREC-10 ou à TREC-11, selon la question. La performance est mesurée à l'aide de la précision moyenne (Kraaij *et al.*, 2003). Le module de recherche monolingue obtient une précision moyenne de 0,570 avec les questions originales en

anglais et le module translinguistique obtient une précision moyenne de 0,467 avec la version française des questions, soit une dégradation de 18 %. À la différence de la traduction du focus, la traduction erronée d'un terme-clé ne compromet pas la recherche de la réponse, surtout lorsque la question est longue et que les termes-clés sont nombreux, comme dans *Quel était le nom de la comédie télévisée dans laquelle Alyssa Milano jouait auprès de Tony Danza?*

4.3 Traduction de la réponse

Une fois la réponse repérée dans un texte, elle peut ou non être traduite, selon l'usage auquel le système est destiné. La traduction de la réponse de l'anglais vers le français fait partie de nos objectifs bien que nous ne nous y soyons pas encore attaqués. Le contexte particulier de la QR joue ici en notre faveur: en effet, beaucoup de réponses sont des entités nommées qui ne requièrent pas de traduction. Par exemple, sur un échantillon de 200 questions prises au hasard, 25 % ont pour réponse un nom de personne ou de lieu identique dans les deux langues, un nombre, une date, une raison sociale ou un titre d'ouvrage. Pour les autres types de réponse, il serait intéressant d'utiliser la question pour désambiguïser les termes avant de les traduire.

4.4 Performance du système complet

Puisque les réponses données par la version bilingue de Quantum sont pour l'instant en anglais, il est possible d'utiliser la procédure d'évaluation automatique fournie lors des campagnes TREC. Le système bilingue, testé sur la version française des questions de notre ensemble de test, obtient 44 % moins de bonnes réponses que le système monolingue testé sur les questions originales. Nous avons comparé avec une alternative plus simple consistant à traduire les questions du français vers l'anglais à l'aide d'un moteur de traduction tel que Babelfish³ puis à utiliser le système monolingue, mais cette approche a entraîné une perte de performance de 53 %, donc pire que celle du système bilingue que nous avons développé.

5 Conclusion

Nous avons montré comment il est possible de transformer un système de question-réponse monolingue en un système bilingue capable de répondre à des questions posées dans une langue différente de celle de la collection de textes. Théoriquement, les modifications peuvent se faire sans toucher au système lui-même par la simple traduction des entrées/sorties. Cependant, tant que les techniques de traduction ne permettront pas d'obtenir des traductions parfaites, il peut s'avérer plus profitable de scinder le problème et de sélectionner différents points de traduction à l'intérieur du système. Dans le cas de notre système Quantum, nous avons utilisé le modèle IBM1 pour obtenir, à partir de la question, des termes-clés pour la recherche de passages pertinents. Nous avons utilisé le modèle IBM2 pour traduire le focus des questions et nous avons écrit de nouvelles règles d'analyse en français dans le but d'éviter d'avoir à produire une traduction complète et bien formée de la question, sur laquelle les règles d'analyse en anglais auraient pu s'appliquer. Le tout a entraîné une dégradation de performance de 44 % par rapport au système monolingue d'origine.

³world.altavista.com

Nous espérons qu'un système bilingue anglais/français donnera aux francophones l'accès à un plus large éventail de sources d'information. Nous estimons qu'actuellement, un francophone qui utilise la version bilingue de Quantum sur une collection de textes en anglais ou qui utilise un système monolingue français sur une collection en français réduite est confronté au même manque à gagner. En effet, dans le premier cas, la bilinguisation du système monolingue a entraîné une dégradation de performance de 44 %. Dans le deuxième cas, pour évaluer le manque à gagner entraîné par la documentation réduite disponible en français, nous avons testé la version monolingue anglaise de Quantum sur une collection de textes dix fois plus petite que la collection originale et nous avons observé une perte de réponses correctes de l'ordre de 43 %. Cependant, à mesure que les moteurs de traduction s'amélioreront, nous sommes confiants que la QR multilingue offrira des performances de plus en plus intéressantes et ce, sans compter l'effet combiné de l'accès à des collections de textes en français *et* en anglais.

Remerciements

Ce projet a été soutenu financièrement par les Laboratoires Universitaires Bell (LUB), le Conseil de recherche en sciences naturelles et en génie du Canada (CRSNG) et le Conseil national de recherches du Canada (CNRC). Nous désirons remercier plus particulièrement Guy Lapalme et les instigateurs du projet: Joel Martin, du CNRC, et Elliott Macklovitch, du RALI.

Références

- CLARKE C. L. A., CORMACK G. V., LASZLO M., LYNAM T. R. & TERRA E. L. (2002). The Impact of Corpus Size on Question Answering Performance. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research Information Retrieval (SIGIR 02)*, Tampere, Finlande.
- GRAESSER A., PERSON N. & HUBER J. (1992). *Mechanisms that Generate Questions*. Lawrence Erlbaum Associates: Hillsdale, New Jersey.
- KIM H. & SEO J. (2002). A Reliable Indexing Method for a Practical QA System. In *Proceedings of the COLING 2002 Post-conference Workshops*, Taipei, Taïwan. Présenté au "Workshop on Multilingual Summarization and Question Answering 2002".
- KRAAIJ W., NIE J.-Y. & SIMARD M. (2003). Embedding Web-based Statistical Translation Models in CLIR. *Computational Linguistics*, **29**(2). À paraître.
- MAGNINI B., NEGRI M., PREVETE R. & TANEV H. (2001). Multilingual Question/Answering: the DIOGENE System. In *Proceedings of TREC-2001*, Gaithersburg, Maryland.
- PLAMONDON L. (2002). Le système de question-réponse QUANTUM. Mémoire de maîtrise, Université de Montréal. www.iro.umontreal.ca/~plamondl.
- PLAMONDON L., LAPALME G. & KOSSEIM L. (2002). The QUANTUM Question Answering System at TREC-11. In *Notebook Proceedings of TREC-11*, Gaithersburg, Maryland.
- SEKI Y. & HARADA K. (2002). Summarization-Based Japanese Question and Answering System from Newspaper Articles. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)*, Las Palmas de Gran Canaria, Espagne: ELRA.
- VETULANI Z. (2002). Question Answering System for POLISH (POLINT) and its Language Resources. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)*, Las Palmas de Gran Canaria, Espagne: ELRA.
- VOORHEES E. M. (2002). Overview of the TREC 2002 Question Answering Track. In *Notebook Proceedings of The Eleventh Text Retrieval Conference (TREC-11)*, Gaithersburg, Maryland.