Application de plusieurs stratégies pour trouver des réponses en anglais à des questions posées en français

Brigitte Grau, Gabriel Illouz, Laura Monceaux, Isabelle Robba, Anne Vilnat, Olivier Ferret, Faïza Elkateb-Gara

Application de plusieurs stratégies pour trouver des réponses en anglais à des questions posées en français

Brigitte Grau[1], Gabriel Illouz[1], Laura Monceaux[2], Isabelle Robba[1], Anne Vilnat[1], Olivier Ferret[3], Faiza El Kateb[1]

[1]*LIMSI, BP 133, 91 403 Orsay Cedex*

Prenom.Nom@limsi.fr

[2]*LINA, Université de Nantes*

Laura.Monceaux@univ-nantes.fr

[3]*LIUM-LIST, CEA*

Olivier.Ferret@cea.fr

Résumé :

Notre système de question-réponse MUSCLEF, qui a participé à l'évaluation CLEF en 2004, a été conçu pour fournir des réponses en anglais à des questions posées en français. Il est fondé sur notre système pour l'anglais, QALC, qui a participé à TREC, et y a obtenu de bons résultats quand nous avons combiné plusieurs stratégies. QALC recherchait des réponses dans la collection donnée et sur le WEB. Nous avons gardé ces deux stratégies pour CLEF, à partir des questions traduites. Nous avons aussi géré le multilinguisme en traduisant les termes significatifs tirés des questions et en adaptant QALC pour construire le système MUSQAT. Nous avons combiné les résultats de ces trois recherches pour produire le résultat final et nous montrons l'apport de cette combinaison par rapport aux résultats de chacune des stratégies seules.

MOTS-CLES : système de question-réponse, bilinguisme, termes.

1. Introduction

La recherche de réponses précises à des questions de types factuels (c'est-à-dire des questions amenant une réponse exprimable par une formulation courte) est un champ de recherche attirant l'intérêt d'un nombre croissant de chercheurs. Les

spécifications varient d'un système à l'autre, ou d'une campagne d'évaluation à l'autre. Néanmoins, le but est toujours de fournir un court extrait du ou des documents contenant la réponse, allant de la réponse exacte uniquement à un ou plusieurs passages. Un challenge supplémentaire, présent dans la campagne d'évaluation CLEF[1], consiste à pouvoir passer d'une langue à l'autre. L'apport de cette fonctionnalité réside dans l'augmentation de l'espace de recherche, tout en permettant de garder sa langue maternelle pour l'interrogation. Cette possibilité est d'autant plus intéressante si on travaille sur le Web et que l'on recherche des réponses en anglais, comme le montre la figure 1.

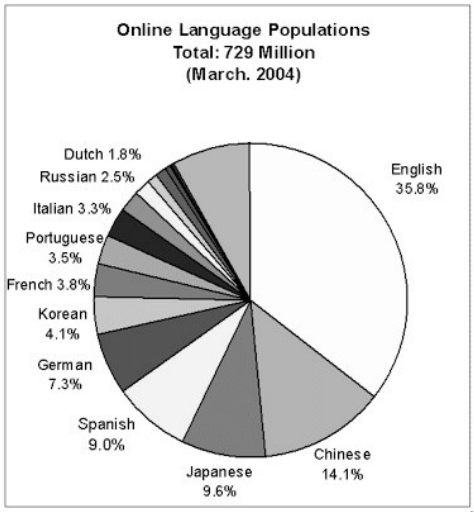

Figure 1 : Répartition des langues sur le Web[2]

Aussi, nous nous sommes intéressés à la recherche de réponses en anglais à partir de questions posées en français. De plus, notre but est de fournir une et une seule réponse à chaque question, sans que des informations non pertinentes soient présentées à l'utilisateur.

Outre le traitement du multilinguisme, le problème consiste à estimer la confiance que le système porte à ses propositions. Lors de TREC11[3], nous avions traité ce problème en recherchant la réponse dans la collection d'une part et sur le Web d'autre part, tout en gardant la même stratégie de résolution, et en combinant les propositions obtenues par le système. Lorsque la même réponse figurait dans les 5 premières propositions provenant de chacune des recherches, cette réponse était fortement privilégiée, considérant que ramener la même réponse de deux sources de connaissances différentes lui confère un fort degré de pertinence, venant supplanter le poids que le système attribue en fonction des processus appliqués. Cette

[1] Cross Language Evaluation Forum
[2] Cette figure provient du « Centre for Public Policy of the University of Melbourne » : http://www.public-policy.unimelb.edu.au/egovernance/papers/33_Skidmore.pdf
[3] Text Retrieval Evaluation Conference

évaluation de la pertinence d'une réponse s'est montrée très efficace, et le système QALC a ainsi réorganisé ses propositions de telle manière que la bonne réponse est remontée en première position dans 21% des cas par rapport aux résultats obtenus à partir d'une seule source de connaissances. QALC était ainsi le système qui ordonnait le mieux ses bonnes réponses [CHA03]. D'autres systèmes ont également montré l'intérêt de combiner de multiples stratégies de résolution [JIJ04], [CHU02], ou de recherches combinées dans différentes sources [BRI01], [CLA01], [HER02] , [MAG02a], [MAG02b].

Le traitement du multilinguisme dans les systèmes de question-réponse relève actuellement de deux approches : la traduction des questions par un traducteur [PER04], [JIJ04], [AHN04], ou la traduction de termes sélectionnés [NEG03], [TAN04], [SUT04]. Comme chacune de ces approches conduit à un nombre de réponses plus restreint que celui obtenu par les systèmes monolingues, nous avons choisi de continuer à combiner les résultats issus de l'application de stratégies différentes, dans la mesure où une même réponse apparaissant dans plusieurs listes a plus de chance d'être correcte, quelle que soit la méthode qui l'a produite. Aussi, notre système MUSCLEF (Multilingual System for CLEF), évalué à CLEF 2004, utilise trois stratégies : la traduction des questions en anglais par la version professionnelle de Systran existant au CEA puis application de QALC sur la collection CLEF d'une part et sur le Web d'autre part, et la traduction des termes de la question, après analyse des questions en français, permettant la recherche dans la collection CLEF, ce qui a conduit au système MUSQAT.

Après une présentation des travaux existant en question-réponse multilingue section 2, nous donnons une vision générale de notre système section 3, pour détailler ensuite les aspects traduction en section 4, la combinaison des listes résultats en section 5 et les résultats obtenus en section 6 avant de conclure.

2. Multilinguisme et question-réponse

Différentes solutions existent pour gérer le multi-linguisme, dans les systèmes de recherche d'information en général et dans les systèmes de question-réponse en particulier. La première consiste à utiliser un traducteur automatique afin de traduire les questions et appliquer un système monolingue par la suite. C'est le choix effectué par [PER04], [JIJ04], [NEU03], [NEU04], [AHN04]. [PER04] et [JIJ04] ont aussi appliqué leur système en version monolingue. Le premier, dans sa version anglais-hollandais, obtient une baisse de 10,5 sur ses résultats : de 91 (45,5 %) à 70 (35 %) réponses correctes, et les résultats du second, dans sa version anglais-français, voit son pourcentage de bonnes réponses diminuer de 13,5 : de 49 (24,5 %) à 22 (11 %) réponses. Pour leur système BiQue, Neumann et Sacaleanu [NEU03], [NEU04] ont eu recours à plusieurs outils de traduction pour obtenir une bonne couverture : 3 en 2003 et 8 en 2004. Les lemmes issus de la fusion des différentes traductions sont réunis et constituent un « sac-de-mots » (bag-of-object) qui est utilisé lors de l'expansion de la requête. L'expansion consiste à compléter le sac-de-mots avec des synonymes mais un module de désambiguïsation est alors nécessaire. Ce module utilise EuroWordNet pour connaître les correspondances entre termes dans les 2 langues (anglais et allemand) et, pour chaque mot ambigu, il regarde

lesquels de ses sens sont exprimés à la fois dans la question d'origine (en allemand) et dans ses traductions (en anglais).

Les deux problèmes majeurs dans l'utilisation de traducteurs pour les questions résident dans la non (ou la mauvaise) résolution de l'ambiguïté des mots des questions et des traductions syntaxiquement fausses, comme nous le verrons section 4.1. Si un mot pertinent pour la recherche de la réponse est mal traduit, cette erreur ne peut en général être rattrapée par la suite par seule compensation des autres mots de la question. En effet les questions sont souvent assez courtes, et la mauvaise traduction d'un mot en change le sens. Si la question produite est agrammaticale, elle est alors mal analysée, les systèmes de question réponse appliquant fréquemment des analyseurs syntaxiques pour extraire les caractéristiques utiles. C'est pour cette raison que Ahn et al. [AHN04] ont choisi de développer un amont et en aval de la traduction de la question produite par Babelfish, un ensemble de règles de transformation, de façon à prévenir ou rectifier des erreurs de traduction assez systématiques sur les formulations syntaxiques des questions[4]. Ainsi, les questions en français « *À quel moment…* » ont été transformées en « *Quand…* » avant d'être soumises au traducteur. De même, les questions avec inversion et reprise pronominale du sujet ont été corrigées après traduction, pour éviter que « *Où X travaille-t-il ?* » ne devienne « *Where X does it work ?* ». Des ensembles de règles similaires ont été développées pour leur système allemand-anglais.

Une deuxième approche consiste à traduire les documents. Dans ce cas, le contexte de traduction est plus grand et donc plus fiable pour gérer l'ambiguïté. Un inconvénient majeur est que la collection résultante est *n* fois plus grande après traduction en *n* langues. Et il n'est pas question de traduire le Web avant interrogation !

La dernière solution consiste à analyser la question dans la langue source et en extraire toutes les caractéristiques utiles à l'extraction des réponses, c'est-à-dire le type de la réponse attendue, les termes importants, les groupes de la phrases (nominaux, verbaux et prépositionnels) ainsi que la structure syntaxique complète (liens de dépendance entre groupes et étiquetage des fonctions grammaticales). La nature de ces informations est la même quelle que soit la langue, hormis les fonctions grammaticales. Aussi, si ces dernières ne sont pas utilisées, ce qui est le cas dans beaucoup de systèmes car elles ne sont pas très fiables, seuls les termes peuvent être traduits, indépendamment de la structure syntaxique complète de la question. Cela ramène au seul problème de la gestion de la polysémie des mots. Nous verrons nos résultats section 4.3 quant aux performances de notre traduction. Cette solution a été choisie par [SUT 04] et [NEG03], [TAN04] dans leur système Diogène qui a participé aux évaluations CLEF avec une tâche monolingue (italien) et 2 tâches bilingues (bulgare/italien vers l'anglais). Tanev et al. ont jugé que les résultats de la traduction automatique, particulièrement pour les questions, n'étaient pas assez encourageants. Ils ont donc eu recours à une traduction des mots clefs de la question : après une étape d'élimination des mots non pertinents, ils traduisent les mots clefs de la question à l'aide d'un dictionnaire bilingue et de MultiWordNet. Puis, afin d'éliminer le bruit inhérent à un tel procédé, ils ne retiennent que les combinaisons de traduction les plus plausibles, i.e. celles qui apparaissent le plus

[4] Ils ont constaté que sur la traduction des 200 questions de CLEF03 de l'allemand vers l'anglais par Babelfish, seules 29% étaient jugées acceptables par un juge linguiste.

fréquemment dans 2 corpus de référence (AQUAINT et TIPSTER). L'étape suivante est l'expansion des mots clefs à l'aide de dérivations morphologiques et sémantiques. Ils obtiennent un score de 45 (22,5%) bonnes réponses en version bilingue contre 56 (28%) en monolingue, donc avec une perte de 6% de bonnes réponses seulement. Sutcliffe et al. [SUT04] ont choisi pour leur part de traduire tous les syntagmes issus de l'analyse des questions par un analyseur de surface selon trois méthodes : deux traducteurs (Reverso et WorldLingo) et un dictionnaire (GDT, Grand Dictionnaire Terminologique). Les résultats obtenus sont ensuite combinés, en donnant la préférence au GDT quand le syntagme y figure. L'ensemble des syntagmes traduits sert ultérieurement à la constitution des requêtes.

Les systèmes de question-réponse comportent tous les mêmes grands modules : analyse des questions, traitement des documents ou passages sélectionnés et extraction des réponses. Ils diffèrent dans leur architecture et dans la nature des processus mis en œuvre. On peut classer les systèmes en trois grandes catégories : les systèmes qui opèrent une analyse syntaxico-sémantique des questions et des réponses afin de les apparier [MOL02], [HAR04], [AHN04], les systèmes qui utilisent des processus plus robustes reposant sur des mesures de similarité plus statistiques, même si il y a utilisation de grammaires locales pour extraire les réponses seules, dont notre système fait partie, et les systèmes multi-stratégies [CHA03], [JIJ04], [CHU02] combinant ensuite les différents résultats. [JIJ04] combinent 8 résultats obtenus par des stratégies de résolution différentes (4 au total) ou l'utilisation de sources de connaissances différentes (collection anglaise, collection hollandaise, Web et encyclopédie hollandaise). Les stratégies appliquées reprennent des types de solutions proposées par beaucoup de systèmes, mais leur application en est différente. Par exemple, les patrons d'extraction de la réponse dérivés des questions sont appliqués sur toute la collection et pas seulement sur des passages sélectionnés.

3. Présentation générale de MUSCLEF

L'architecture globale de MUSCLEF est illustrée Figure 2. Elle regroupe l'application de QALC à partir des questions traduites sur la collection CLEF et le Web et la version « bilingue » MUSQAT utilisant les résultats de l'analyse des phrases en français. Ensuite les mêmes modules s'appliquent sur les documents trouvés en anglais.

Le but du module d'analyse de la question est de déduire les caractéristiques qui peuvent contribuer à trouver les réponses possibles dans les passages retenus et de reformuler les questions sous forme déclarative pour le moteur de recherche sur le Web (Google). Ces caractéristiques sont le focus de la question, le verbe principal et les fonctions syntaxiques des modifieurs. Nous avons concentré nos efforts de traduction sur ces éléments, comme cela sera précisé dans la section suivante. Pour la campagne CLEF 04, nous avons développé une nouvelle version de ce module pour analyser les questions en français. La version française de l'analyseur syntaxique XIP [AIT02] sert de base à ce module. Pour l'analyse des questions traduites, nous utilisons IFSP [AIT97].

Les requêtes ne sont pas identiques pour la recherche sur le Web et pour le recherche dans la collection CLEF. Dans le second cas, nous utilisons MG pour

retrouver les passages. Pour interroger sur le Web, nous envoyons une reformulation presque exacte de la réponse, en faisant l'hypothèse que la redondance du Web permettra toujours de sélectionner des documents.

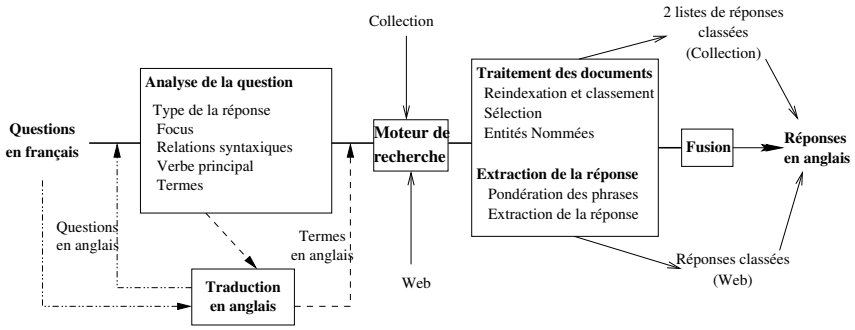

Figure 2 – Architecture de MUSCLEF

Le traitement des documents renvoyés par le moteur commence ensuite. Ils sont ré-indexés par les termes de la question et leurs variantes linguistiques, ré-ordonnés en fonction du nombre et du type des termes qu'ils contiennent, afin de n'en retenir qu'un sous-ensemble. La reconnaissance des différents types d'entités nommées est alors effectuée. L'extraction de la réponse consiste à d'abord attribuer des poids aux phrases avant d'extraire les réponses elles-mêmes. Différents processus sont appliqués suivant le type de réponse attendu, chacun d'eux ayant pour résultat des réponses pondérées. La dernière étape consiste à combiner les réponses trouvées dans la collection, les réponses issues du Web et les résultats issus du système MUSQAT. Un score final est alors calculé ; son principe est de privilégier une réponse qui a été classée dans les cinq premières possibilités par deux chaînes au moins, même si ses scores individuels sont moindres.

4. Analyse des questions

Comme nous l'avons indiqué plus haut (voir Figure 2), deux solutions ont été testées pour représenter les informations contenues dans les questions afin de permettre leur comparaison avec les documents. Dans la première, nous faisons appel à un traducteur automatique pour passer la question du français à l'anglais et ensuite procéder à l'analyse des questions traduites. La seconde solution consiste à analyser les questions dans la langue source (le français dans notre cas) et à traduire en anglais (notre langue cible), ceux des termes qui ont été considérés comme les plus importants, ce qui entraîne aussi la traduction des différentes caractéristiques telles le focus et le verbe principal.

4.1. Traduction automatique des questions

La première solution que nous avons donc testée pour résoudre la différence de langues entre les questions et les documents consiste à faire appel à un traducteur automatique sur les questions. Dans notre cas, cette traduction automatique a été effectuée par l'interface en ligne SYSTRANLinks fournie par Systran (nous tenons à remercier Systran qui nous a donné accès à ce service dans le cadre du projet ALMA). Nous n'avons fait appel à aucun dictionnaire complémentaire. Comme la plupart des questions de l'évaluation CLEF ne présentait pas une grande complexité syntaxique et concernait des sujets généraux, leur traduction peut souvent être considérée comme fiable, comme l'illustre la Figure 3.

0009 - Quand est apparu pour la première fois le virus Ebola ?
0009 - When did the Ebola virus appear for the first time?

0166 - Où se trouve Halifax ?
0166 - Where is Halifax?

Figure 3 Exemples de questions correctement traduites

Cependant, les erreurs de traduction peuvent aussi se produire pour des questions simples, comme l'illustre la figure 4. Ces erreurs peuvent concerner la syntaxe. Dans la question 175 par exemple, « *Quel est* » devrait être traduit par « *Who is* » et non pas par « *Which is* ». De même, dans la question 165, l'expression « *Qu'est-ce que* », spécifique aux questions, n'est que partiellement traduite. En fait cette observation montre que, tout comme pour les étiqueteurs morpho-syntaxiques et les analyseurs syntaxiques, les questions devraient être spécifiquement considérées par les systèmes de traduction, ce qui n'est généralement pas le cas. Les erreurs sont aussi d'ordre sémantique. Dans la question 175 à nouveau, « *réalisateur* » est traduit par « *realizer* » alors qu'il y a plus de chance de trouver une réponse avec une traduction comme « *director* » ou « *film director* ». Enfin la question 165 montre aussi le problème posé par le fait que les dictionnaires sont incomplets, ce qui est inévitable dans un système en domaine ouvert, spécialement pour les acronymes. Ainsi, « *OMC* » (Organisation Mondiale du Commerce) devrait être traduit en « *WTO* » (World Trade Organization), tout comme « *OTAN* » est traduit en « *NATO* » dans la question 143.

0175 - Quel est le réalisateur de "Nikita" ?
0175 - Which is the realizer of "Nikita"?

0165 - Qu'est-ce que L'OMC ?
0165 - What OMC?

0143 - En quelle année a été créée l'OTAN ?
0143 - In which year was created NATO?

Figure 4 - Exemples d'erreurs dans les traductions des questions

4.2. Evaluation des analyses

119 questions sur 200 attendent une entité nommée en réponse. Les types d'entités nommées que nous avons définis sont les types classiques légèrement raffinés : personne, organisation, lieu-endroit, ville, nom propre, nombre, pourcentage, montant financier, quantité physique (surface), vitesse, poids, volume, longueur, température, age, heure, date (différents types de dates tels que jour, jour et mois, etc.), durée et période. Le typage de la réponse sur le français a une précision de 99% et un rappel de 95%. Sur l'anglais, la précision est la même, mais le rappel est de 83%. Il y a 14 types de réponses non reconnus supplémentaires, dus à des formes syntaxiques inhabituelles ou fausses en anglais.

4.3. Traduction de termes

Différentes méthodes peuvent être utilisées pour traduire les termes. Les résultats peuvent être obtenus par une traduction fondée sur des ontologies bilingues ; mais comme nous venons de le voir dans la section précédente, celles-ci n'existent pas en domaine ouvert ou ne sont pas suffisamment complètes. Parmi les autres possibilités de traduction, nous nous sommes intéressés à la plus simple, consistant à utiliser un dictionnaire bilingue pour traduire les termes de la langue source vers la langue cible. Cette méthode présente deux inconvénients : d'une part, il est impossible de lever directement les ambiguïtés sur les différents sens des mots à traduire, d'autre part la richesse lexicale des deux langues doit être comparable. Comme cette dernière contrainte est vérifiée pour le couple français/anglais, nous avons décidé d'utiliser quand même cette méthode. Toutefois, pour avoir une estimation des ambiguïtés que nous risquions de rencontrer dans le contexte Question-Réponse, nous avons étudié le corpus des 1893 questions en anglais de TREC. Après analyse, nous avons conservé 9000 des 15624 mots utilisés dans ce corpus. La moyenne du nombre de leurs sens est de 7,35 dans WordNet. Les valeurs extrêmes sont 1 (pour *neurological* par exemple) et 59 (pour *break* par exemple). Autour de la valeur moyenne, on trouve des mots communs, tels que *prize, blood, organization*. De ce fait, nous ne pouvons pas considérer un dictionnaire donnant seulement un sens par mot.

Connaissant ces contraintes, nous avons étudié les différents dictionnaires que nous pouvions utiliser : d'une part les dictionnaires en ligne, tels que Reverso[5], Systran[6], Google[7], Dictionnaire Terminologique[8] ou FreeTranslation[9], et d'autre part les dictionnaires sous licence GPL , tels que Magic-Dic[10] ou Unidic. Les dictionnaires en ligne sont généralement complets. Mais ils résolvent les ambiguïtés (ou tentent de le faire) et ne fournissent qu'une traduction par mot. Un autre inconvénient que nous avons constaté a été l'impossibilité de modifier ces dictionnaires et l'obligation de tenir compte de quelques contraintes techniques telles que le nombre limité de requêtes que nous pouvions faire et les temps de réponse.

[5] http://translation2.paralink.com
[6] http://babel.altavista/translate.dyn
[7] http://www.google.com/language.tools
[8] http://granddictionnaire.com
[9] http://www.freetranslation.com
[10] http://magic-dic.homeunix.net/

En ce qui concerne les dictionnaires GPL, ils sont notoirement moins complets, mais ils peuvent être augmentés. De plus, ils sont très rapides et pour la plupart, ils donnent plusieurs traductions pour une requête. Parmi les dictionnaires GPL, nous avons choisi Magic-dic, du fait de ses capacités d'évolution : des termes peuvent être ajoutés par n'importe quel utilisateur, mais ils sont vérifiés avant d'être intégrés, ce qui n'est pas le cas d'Unidic. Par exemple, le mot porte donne les résultats suivants (nous n'en donnons qu'un extrait) :

– porte bagages - *luggagerack, luggage rack*
– porte cigarettes - *cigarette holder*
– porte clefs - *key-ring*
– porte plume - *fountain pen*
– porte parole, locuteur - *spokesman*
– porte - *door, gate*

4.4. Module multilingue

Nous allons illustrer la stratégie de MUSQAT sur l'exemple suivant : *Quel est le nom de la principale compagnie aérienne allemande ?*, qui est traduite en anglais par : *What is the name of the main German airline company?*.

La première étape est l'analyse de la question en français, qui fournit une liste de tous les mono-termes et de tous les bi-termes (tels que adjectif/nom commun) présents dans la question, et élimine les mots vides. Les bi-termes sont utiles, parce qu'ils permettent de désambiguiser en donnant un (petit) contexte à un mot. Dans notre exemple, les bi-termes (sous leur forme lemmatisée) sont : *principal compagnie, compagnie aérien, aérien allemand* et les mono-termes : *nom, principal, compagnie, aérien, allemand*.

En s'aidant du dictionnaire Magic-dic, nous avons essayé de traduire les bi-termes (quand ils existent), et les mono-termes. Toutes les traductions proposées sont conservées. Tous les termes sont étiquetés grammaticalement. Si un bi-terme ne peut pas être directement traduit, il est reconstitué à partir des traductions des mono-termes qui le composent, en suivant la syntaxe anglaise. Pour notre exemple, nous avons obtenu pour les bi-termes : *principal compagny/main compagny, air compagny, air german* ; et pour les mono-termes : *name/appellation, principal/main, compagny, german*. Quand un mot est absent du dictionnaire, nous le conservons à l'identique en supprimant les signes diacritiques.

Ces termes, avec leurs catégories remplacent alors les mots d'origine en entrée des autres modules de MUSQAT. Le module de traduction n'essaie pas de résoudre les ambiguïtés entre les différentes traductions : la requête envoyée au moteur MG est constituée de l'union de toutes les traductions et la levée d'ambiguïtés a lieu éventuellement lors de la sélection des documents par le moteur ou après ré-indexation. Si les différents termes sont synonymes, des documents pertinents sont trouvés avec ces synonymes, donnant ainsi lieu à une recherche plus large. Si le mot est incohérent dans le contexte du document, nous faisons l'hypothèse que son influence n'est pas suffisante pour créer du bruit.

Nous avons évalué la traduction produite par MUSQAT. Les 200 questions en français contenaient 731 mots, correspondant à 1091 mots anglais, et 932 termes (mono-termes + bi-termes) correspondant à 1464 termes en anglais. En étudiant ces traductions, nous avons observé que :

– 59% des termes traduits étaient corrects, (même si pour 12,63% des termes la traduction pourrait être améliorée)
– 8% des termes traduits étaient corrects mais identiques aux termes dans la langue source
– 33% des termes traduits étaient incorrects

Il est évident que le dictionnaire n'était pas assez complet pour cette campagne. Nous devrions obtenir une meilleure couverture en le complétant manuellement avec les traductions manquantes (il n'y a pas par exemple de traduction de verbe français *jouer* dans son sens *to play*). Pour pallier son incomplétude, et parce qu'il a été prouvé qu'utiliser plusieurs dictionnaires donne de meilleurs résultats que d'en utiliser un seul, nous avons l'intention de l'enrichir aussi avec des dictionnaires en ligne.

Une autre évaluation concerne les bi-termes, que nous avons présentés comme très important dans la perspective de désambiguiser les mono-termes ambigus. Pour cela, nous avons déterminé la fréquence documentaire de chaque traduction des différents bi-termes dans le corpus CLEF. Si la fréquence est élevée, alors le bi-terme peut être considéré comme une traduction adéquate. Suivant cette étude, 47,5% des bi-termes ont été trouvés dans le corpus. Une approche prometteuse pourra être de valider les traductions, en les notant en fonction de leurs fréquences, à la fois dans un corpus bilingue aligné, et dans un corpus monolingue (dans la langue cible).

Nous avons également noté qu'un travail important restait à faire sur les noms propres, spécialement les noms géographiques, les noms d'organisation et les acronymes. Il faudra pour cela développer des listes bilingues des noms les plus fréquents.

5. Combinaison des listes résultats

Comme cela a été dit dans la section 3, nos résultats sont obtenus en comparant 3 ensembles de réponses : le premier de ces ensembles est retourné par MUSQAT, le second par QALC utilisant le Web comme collection et le troisième par QALC appliqué à la collection CLEF. Le Web constitue une source de connaissances beaucoup plus large que la collection CLEF, le fait d'utiliser une telle source nous permet de confirmer les réponses trouvées dans la collection et donc de renforcer leur score de confiance. Notons que les réponses provenant du Web doivent aussi appartenir à la collection, sinon elles ne sont pas acceptées comme réponses correctes : en effet toutes les réponses doivent être accompagnées d'un document les justifiant.

Chacun de ces 3 ensembles contient pour chaque question un ensemble de réponses ordonnées selon leur score de confiance, score mis à jour tout au long du processus d'extraction de la réponse. Avant de décrire l'algorithme écrit pour la sélection finale, nous décrivons la façon dont ce score est attribué à chaque réponse candidate.

5.1. Pondération des réponses par chaque système

Toutes les phrases issues du traitement des documents sont examinées dans le but de leur attribuer un poids reflétant à la fois la possibilité qu'elle contienne la réponse et la possibilité que le système y localise la réponse. Les critères retenus pour calculer ce poids sont liés aux informations contenues dans la question. Les traits suivants sont ainsi recherchés dans la phrase candidate :

– les lemmes de la question, chacun possédant un poids qui représente son degré de spécificité[11]
– les variantes de ces lemmes
– les mots exacts de la question (seulement dans le cas de la version « tout en anglais »)
– la proximité mutuelle des mots de la question
– la présence des entités nommées attendues

Un premier poids est calculé tout d'abord en fonction de la présence des lemmes et de leurs variantes (les 2 premiers critères). Ensuite, à ce poids est ajouté un poids additionnel pour chaque autre critère satisfait, les poids additionnels ne pouvant dépasser 10 % du poids d'origine.

Au cours de l'extraction de la réponse, un poids est attribué à chaque réponse potentielle. Pour ordonner les réponses de type Entité Nommée, MUSCLEF (resp. QALC) calcule donc des poids additionnels prenant en compte :

– l'entité nommée précise ou généralisée de la réponse
– la place de la réponse par rapport à celles des mots de la question dans la phrase
– la redondance de la réponse dans les 10 premières phrases candidates

Quand le type attendu de la réponse n'est pas une entité nommée, nous utilisons des patrons d'extraction. Chaque phrase candidate est analysée en utilisant les patrons d'extraction associés au type de la question (déterminé lors de l'analyse de la question). Ces patrons d'extraction sont écrits avec des expressions régulières qui utilisent le focus comme pivot et sont pondérés en fonction de leur spécificité. Davantage de détails seront trouvés dans [FER02]. A la fin de l'extraction, les 5 réponses de meilleur score sont retenues pour la sélection finale.

5.2. Algorithme de sélection de la réponse

L'idée ici est de comparer des résultats provenant de différentes sources de connaissance afin de renforcer le score des réponses qui appartiennent à plusieurs ensembles, permettant ainsi à un certain nombre de bonnes réponse d'atteindre le premier rang. Le tableau 1 contient un exemple de ces ensembles de réponses, pour la question « *En quelle année Thomas Mann a-t-il obtenu le Prix Nobel ?* ».

Les 3 ensembles de réponses sont comparés 2 à 2 en utilisant un algorithme écrit pour les évaluations TREC. Cet algorithme examine chaque couple ($réponse_i$, $réponse_j$), i et j sont compris entre 0 et 4 et représentent la position de la réponse dans son ensemble. Quand les 2 réponses sont égales ou incluses l'une dans l'autre, l'algorithme attribue un bonus au meilleur score du couple. Ce bonus a été choisi

[11] Le degré de spécificité d'un lemme est calculé en fonction de l'inverse de sa fréquence relative calculée sur un grand corpus

afin de permettre aux réponses confirmées de passer devant les réponses non confirmées ; il est calculé en fonction des positions i et j : (10 – (i + j)) * 100. De cette façon, l'algorithme construit un ensemble de couples ordonnés en fonction de leur nouveau score. Comme dans CLEF il y avait non pas 2 mais 3 ensembles de réponses, nous avons appliqué cet algorithme sur les trois ensembles de couples de réponses c'est-à-dire 3 fois, la réponse finalement retournée étant celle qui obtient le meilleur score.

En consultant le tableau 1, on voit que 2 dates sont présentes dans les 3 ensembles : « *en 1929* » et « *en 1976* ». Le couple qui apparaît en gras est celui qui obtient le meilleur score final : il a reçu un bonus de 900 points pour un score d'origine de 1082 points. La réponse « *en 1929* » est donc retournée avec un score final de 1982. C'est effectivement la bonne réponse.

	QALC + Web		**MUSQAT**		**QALC + Collection**	
Rang	Réponse	Score	Réponse	Score	Réponse	Score
0	**en 1929**	**1082**	*en 1976*	721	11 Octobre 1994	878
1	1875-1955	1005	en 1976	721	**en 1929**	**853**
2	8 Mars 1879	903	en 1929	664	*en 1976*	798
3	en 1903	877	2	640	12 Octobre 1994	703
4	en 1929	849	1964	561	en 1979	696

Tableau 1 : Un exemple des 3 ensembles de réponses

Comme on vient de le constater, cet algorithme qui effectue les comparaisons sur 2 ensembles est facilement applicable à plus de 2 ensembles, puisqu'il suffit de répéter les comparaisons autant que nécessaire. Néanmoins, nous avons observé qu'une comparaison menée d'emblée sur les 3 ensembles aurait donné des résultats différents. En effet, en menant la comparaison 2 à 2, nous ne prenons pas en compte de la même façon les réponses présentes dans les 3 ensembles.

6. Résultats

Le tableau 2 présente une évaluation comparative de MUSQAT et de QALC. Cette évaluation a été faite de façon automatique en recherchant dans les phrases réponses les expressions régulières correspondant aux patrons de réponses. Ces résultats ont été calculés pour les 178 questions pour lesquelles nous avions un patron de réponse[12].

[12] Les autres sont supposées être des questions n'ayant pas de réponse dans le corpus. Comme nous n'en avons reconnu aucune dans nos propositions, nous supposerons que les 178 patrons sont les réponses aux 200 questions afin de ramener le nombre de réponses au même total

La première ligne indique le nombre de réponses correctes trouvées dans les 5 premières phrases retournées par MUSQAT et par les 2 applications de QALC (sur le Web et sur la collection). La deuxième ligne « Réponses à EN » donne le nombre de réponses correctes pour les questions attendant une EN, et la troisième ligne concerne les autres questions. Les résultats sont donnés pour la réponse au rang 1 et pour les 5 premières réponses. La dernière colonne indique le meilleur résultat de notre système, obtenu en utilisant l'algorithme de fusion décrit section 5.2. Le score officiel de MUSQAT étant de 22 bonnes réponses (11 %), nous observons qu'en fusionnant les réponses des différentes stratégies nous avons un gain de 17 réponses (77 %). Ce dernier résultat nous place à égalité avec les systèmes français-anglais de CLEF04.

		MUSQAT	**QALC + Collection**	**QALC + Web**	**Fusion** 200 questions
		178 patrons (200 questions)			
Phrases	5 premiers rangs	56	65	61	
Réponses à EN	rang 1 5 premiers rangs	17 32	26 37	24 43	
Réponses non EN	rang 1 5 premiers rangs	7 12	3 8	0 0	
Total	rang 1 5 premiers rangs	24 (12%) 44	29 (14,5%) 45	24 (12%) 43	39 (19,5 %)

Tableau 2 : Evaluation comparative des différentes stratégies

Pour l'évaluation TREC 2002, le système devait aussi produire pour chaque question une unique réponse. Le tableau 3 donne le nombre de bonnes réponses obtenues. On peut constater que la fusion apporte un gain de 29 réponses, soit 21% seulement en comparaison du précédent. Cela peut s'expliquer par le meilleur classement intrinsèque des bonnes réponses par le système en monolingue. Lorsque l'on évalue QALC sur les 5 premières réponses, 30% environ des bonnes réponses se situent après le premier rang, alors que dans MUSQAT ou QALC-CLEF, il y en a plutôt 35,5 %. L'autre argument réside dans le fait que l'on fusionne trois sources de résultats au lieu de deux ; de plus, les résultats sont issus de stratégies différentes et pas seulement de sources de connaissances différentes.

Résultats TREC 11 500 questions	QALC (Collection Trec)	Fusion Collection + Web
Résultats officiels		139 (27,8 %)
Evaluation automatique	136 (27,2 %)	165 (33 %)

Tableau 3 : Résultats obtenus à TREC 11 par QALC

Les résultats de la dernière ligne du tableau 3 peuvent être comparés à ceux de la dernière ligne du tableau 2 pour les systèmes seuls : en ce qui concerne les bonnes réponses au rang 1, on constate que le problème du multilinguisme a entraîné une

baisse de 13 dans les pourcentages de bonnes réponses, ce qui est comparable aux systèmes présentés section 2.

On peut aussi remarquer que les 3 stratégies sont équivalentes en nombre de réponses. Mais si on compare l'ensemble des 5 premières réponses données pour chaque question par MUSQAT et par QALC appliqué à la collection, on voit qu'il y a seulement 21 réponses communes, et donc 22 trouvées uniquement par MUSQAT et 24 par QALC. Nous obtenons les mêmes chiffres en comparant les autres résultats deux à deux, avec toutefois un peu plus de réponses en commun entre les deux applications de QALC, ce qui s'explique par le fait de partir de la même formulation des questions. Aussi, même si certaines réponses sont proposées car de meilleur poids, il y a lieu de chercher à améliorer le choix de la réponse lorsqu'elle est présente dans une seule liste, puisque nous disposons au total de 65 bonnes réponses parmi les 5 meilleures par couple de résultats. Un point faible de notre système reste l'extraction des réponses dans le cas des questions n'attendant pas une entité nommée comme en témoigne la 3ème ligne du tableau 2.

7. Conclusion

Même si ces résultats sont encourageants, MUSCLEF, notre premier système multilingue peut encore être amélioré. Son architecture, organisée en plusieurs modules indépendants, a été choisie de façon à permettre aisément ces améliorations. En outre, nous avons vu que les 2 stratégies adoptées : la traduction des termes importants et la traduction des questions, étaient pertinentes et devaient être maintenues dans des expériences futures. Elles sont assez complémentaires et une amélioration consisterait à fusionner les propositions des deux traductions dans chaque système, afin de fournir des synonymes au moment de la recherche des documents. Bien entendu de meilleures ressources sémantiques seront indispensables, mais comme de telles ressources ne sont pas facilement disponibles, utiliser des traductions attestées en corpus pour contrôler les traductions pourrait être une piste intéressante à étudier.

8. Références bibliographiques

[AHN04] K. Ahn, B. Alex, J. Bos, T. Dalmas, J.L. Leidner et M.B. Smillie , 2004, Cross-lingual Question Answering with QED, Working Notes de CLEF Cross-Language Evaluation Forum, Bath UK, pp. 335-342

[AIT97] S. Aït-Mokhtar et J.-P. Chanod. 1997, Incremental finite-state parsing. In Proceedings of the 5th Conference on Applied Natural Language, Processing (ANLP-97), Washington, DC, USA

[AIT02] S. Aït -Mokhtar, J.-P. Chanod et C. Roux, 2002, Robustness beyond shallowness: incremental deep parsing. Natural Language Engineering Vol. 8 (2/3), pp. 121-144

[BRI01] E. Brill, J. Lin, M. Banko, S. Dumais et A. Ng, 2001. Data-Intensive Question Answering. TREC 10 Notebook, Gaithersburg, USA

[CHA03] G. de Chalendar, F. El Kateb, O. Ferret, B. Grau, M. Hurault-Plantet, L. Monceaux, I. Robba, A. Vilnat, Confronter des sources de connaissances différentes pour obtenir une réponse plus fiable, TALN 03, Batz sur Mer

[CHU02] J. Chu-Carroll, J. Prager, C. Welty, K. Czuba et D. Ferruci, 2002. A Multi-Strategy and multi-source Approach to Question Answering. TREC 11 Notebook, Gaithersburg, USA pp. 124-133

[CLA01] C. L. Clarke, G. V. Cormack, T. R. Lynam, C. M. Li et G. L. McLearn, 2001, Web Reinforced Question Answering (MultiText Experiments for Trec 2001), TREC 10 Notebook, Gaithersburg, USA

[FER02] O. Ferret, B. Grau, M. Hurault-Plantet, G. Illouz, C. Jacquemin, L. Monceaux, I. Robba, et A.Vilnat, 2002, How NLP Can Improve Question Answering Knowledge Organization, Vol. 29, N°3-4, pp. 135-155

[HAR04] S. Hartrumpf, 2004, Question answering using sentence Parsing and Semantic Network Matching, Working Notes de CLEF Cross-Language Evaluation Forum, Bath UK, pp.385-393

[HER02] U. Hermjakob, A. Echihabi et D. Marcu. 2002, Natural Language Based Reformulation Resource and Web Exploitation for Question Answering, TREC 11 Notebook, Gaithersburg, USA

[JIJ04] V. Jijkoun, G. Mishne, M. de Rijke, S. Schlobach, D. Ahn et K. Muller, 2004, The University of Amsterdam at QA@CLEF2004, Working Notes de CLEF Cross-Language Evaluation Forum, Bath UK, pp. 321-325

[MAG02a] B. Magnini, M. Negri, R. Prevete et H. Tanev. 2002. Is It the Right Answer? Exploiting Web redundancy for Answer Validation, Proceedings of the 40 th ACL, pp. 425-432

[MAG02b] B. Magnini, M. Negri, R. Prevete et H. Tanev, 2002, Mining Knowledge from Repeated Co-occurrences: DIOGENE at TREC-2002, TREC 11 Notebook, Gaithersburg, USA

[MOL02] D. Moldovan, S. Harabagiu, R. Girju, P. Morarescu, F. Lacatusu, A. Novischi, A. Badalescu et O. Bolohan, 2002, LCC Tools for Question Answering, TREC 11 Notebook, Gaithersburg, USA

[NEG03] M. Negri, H. Tanev et B. Magnini, 2003, Bridging Languages for Question_Answering/ DIOGENE at CLEF2003, Working Notes de CLEF Cross-Language Evaluation Forum, Trondheim, Norvège

[NEU03] G. Neumann et B. Sacaleanu, 2003, A Cross-Language Question / Answering-System for German and English, Working Notes de CLEF Cross-Language Evaluation Forum, Trondheim, Norvège

[NEU04] G. Neumann et B. Sacaleanu, 2004, Experiments on Robust NL Question Interpretation and Multi-layered Document Annotation for a Cross-Language Question / Answering System, Working Notes de CLEF Cross-Language Evaluation Forum, Bath UK, pp.311-320

[PER04] L. Perret, 2004, Question Answering System for the French Language, Working Notes de CLEF Cross-Language Evaluation Forum, Bath UK, pp. 295-305

[SUT04] R. Sutcliffe, I. Gabbay, M. Mulcahy et A. O'Gorman, 2004, Cross-Language French-English Question Answering using the DLT System at CLEF-2004, Working Notes de CLEF Cross-Language Evaluation Forum, Bath UK, pp.305-309

[TAN04] H. Tanev, M. Negri, B. Magnini et M. Kouylekov, 2004, The DIOGENE Question Answering System at CLEF-2004, Working Notes de CLEF Cross-Language Evaluation Forum, Bath UK, pp.325-333