Corpus bilingues comparables et l'extraction automatique de terminologie bilingue français arabe dans les domaines de spécialité

DJEBAILI Farida
Université de Mostaganem

Résumé *:*
On ne tarit pas d'éloges sur le rôle ***des bis textes*** *dans la création d'approches quantitatives d'extraction de ressources traductionnelles. Mais leur manque dans les paires de langue ne faisant pas intervenir l'anglais, ou encore dans les langues présentant une grande divergence de structure a fait appel à l'utilisation des* ***corpus bilingues comparables****. L'inflation textuelle que connait un domaine de spécialité en constante évolution tel que le domaine de l'informatique se traduit par l'apparition de néologismes qui font de plus en plus défaut dans la langue arabe. Nous montrerons dans cet article le rôle d'un corpus comparable bilingue dans la création d'approches quantitatives d'extraction de ressources traductionnelles dans les domaines de spécialité.*

1. Introduction

Tentative après tentative l'automatisation complète de l'activité de traduire s'avère impossible. Un échec dû en grande partie à la particularité des langues naturelles. La compétence linguistique n'est rien pour un système qui ne sait pas de quoi il parle, la machine ne pouvant saisir la dimension illocutoire de l'acte de langage. Pour suppléer le contexte qui faisait défaut on se mit à représenter les connaissances, i.e. coder les connaissances non linguistiques à travers des processus tels que l'étiquetage, la segmentation, la lemmatisation et ré explorer la traduction humaine. La capacité de dire si deux mots ou groupes de mots ont la même signification dans leurs contextes respectifs permet de mettre

fin à des aberrations en traduction automatique tel que ''doctor of women and other diseases [1].

Ainsi naquirent les mémoires de traduction, et les corpus bilingues. Quant aux premières ce sont des bases de données de segments de phrases originales et leurs traductions, traduites par des traducteurs humains, mais qui ne sont pas apparier c'est-à-dire, qui ne sont pas mis en correspondance avec leurs originaux, mais qui sont stockées sous forme de paires langue source-langue cible. Les plus renommées sur le marché Translaror s'Workben, Atril, Transit, et bien d'autres. En ce qui concerne les corpus bilingues on distingue entre les bis textes (notion introduite par ZELLING HARRIS), et les corpus bilingues comparables. A base d'analyse texto- métrique, Les bis-textes sont produits par un logiciel appelé aligneur, qui apparie[2] automatiquement les versions originales et traduites d'un même texte mais sans qu'il soit possible de déterminer lequel a servi de source. La correspondance est réalisée aussi bien au niveau des paragraphes qu'au niveau des phrases ou bien même au niveau des mots. Les corpus comparables bilingues quant à eux ce sont des textes dans deux langues différentes partageant des caractéristiques communes, mais qui ne sont pas en relation de traduction.

En traitement automatique des langues naturelles on ne tarit pas d'éloges sur le bi texte, car il permet à travers son exploitation des régularités statistiques et son comptage des occurrences d'unités lexicales dans différentes parties du corpus de pourvoir la machine de méthodes quantitatives pour l'identification et l'extraction des correspondances. Sauf que, l'absence de corpus parallèles alignés ou alignés correctement

[1] Traduction automatique de l'arabe d'un écrit sur la porte d'un gynécologue " طبيب الأمراض النسائية المختلفة".

[2] Mettre en correspondance

à cause des grandes divergences entre les structures de langues donna une place de choix aux corpus bilingues comparables.

La question qu'on se pose est: Quel est le rôle des corpus comparables bilingues dans la réalisation des équivalences en systèmes de traductions pour la paire de langues français arabe, une paire de langue avec une grande divergence de structures ?

Je vois la réponse à cette question à travers 4 points qui vont construire l'ossature de cette intervention:

1- Qu'est ce qu'un corpus bilingue comparable ? et Quelles sont les critères de comparabilité de corpus dans un contexte bilingue ?

2- Le rôle des corpus bilingues comparables en lexicographie et en reconnaissance sémantique n'est plus à démontrer, qu'en est-il pour leur rôle dans la création d'approches quantitatives pour l'identification et l'extraction de ressources traductionnelles ?

3- Quel rôle jouent-ils dans la construction, la création et l'actualisation de lexique bilingue dans un domaine de spécialité tel que l'informatique, un domaine en constante évolution, qui se traduit par une inflation textuelle dans le domaine, et par la création continuelle de néologismes, qui font souvent défaut dans la langue arabe ?

2. Corpus bilingues comparables

Pour Mc ENRY et WILSON un corpus est une collection de plus d'un texte. Marcus [3] quant à lui réserve le terme corpus à l'ensemble de textes choisis de façon très précise pour répondre à des besoins et des intérêts particuliers. Saint Claire, utilise le terme données, car pour lui un corpus est : ''une collection de données qui sont sélectionnées et organisées

[3] P,Marcus, B.Santorini, and M.A Marcinkiewizc 1994. Building a large annotated corpus of English : The Penn Tree-bank. Computational Linguistics.

selon des critères linguistiques explicites pour servir d'échantillon de langue''.Pour l'industrie de langue, un corpus c'est un certain nombre de données, d'un certain domaine, sans accorder la moindre attention sur comment s'est construit.

Revenons à notre notion de comparabilité : Un corpus comparable monolingue est une collection de textes dans une même langue, ayant des points en commun. Thématique, genre, registre, auteur, lexique partagés, période, support …etc. Le but étant soit de comparer les divers emplois d'un mot, ou le sens d'un même terme, ou observer la fréquence des mots, ou encore identifier les collocations, les définitions, observer les propriétés distributionnelles de certains mots …etc. Les corpus comparables bilingues quant à eux désignent une collection de textes en langues différentes partageant aussi des points en commun, et rassemblés selon des critères de similarité.

Pour Hervé Dejan et Al 2002 [4] «deux corpus en L1 et L2 sont dit comparables s'il existe une sous partie non négligeable du vocabulaire de L1 respectivement L2 dont la traduction se trouve dans L2 respectivement L1».

Mais à quel point deux corpus peuvent-ils être comparables ?comment juger de la comparabilité d'un corpus bilingue ? Quelles sont les mesures de comparabilités d'un corpus bilingues ? KILGARIFF résume le problème de comparabilité en deux questions, « En quoi deux corpus sont-ils similaires ? Et en quoi sont-ils différents[5]?

[4] Dejean,H et Gaussier, Extraction de lexique bilingue à partir de corpus comparables dans Lexicometrica , 2002 .
[5] Kilgariff , A *Comparing Corpora, International Journal of Corpus Linguistics, 2001.*

3. Critères de comparabilité dans un contexte bilingue

Comparer deux corpus de deux langues différentes peut être à travers des critères qualitatifs [BIBER 93] ou bien à travers des critères quantitatifs basés sur la fréquence des mots [KILGARRIFF]. Dans les premières, les critères stylistiques cités par BIBER 93, la comparabilité se fait par rapport au contenu ; le genre, l'auteur, la période, le registre. Les critères quantitatifs quant à eux entre dans le cadre de la linguistique quantitative et traiteront les données au niveau du plan d'expression par opposition au plan du contenu.

L'approche de KILGARIFF, est basée sur la fréquence des mots, et comme toute analyse quantitative elle repose sur des techniques d'inférence. Son argumentation est: que pour une manipulation automatique objective on peut se passer des informations stylistiques, car le corpus est présenté sous forme de liste de fréquences de mots. Une liste qu'il est facile de créer automatiquement, ce qui permettra de juger de la similitude des textes dans beaucoup moins de temps que dans l'analyse qualitative. Il s'agira donc de manipulations statistiques et de scores de comparabilité sur lesquelles nous ne nous attarderons pas vu la brièveté de l'exposé.

Nous pensons que les deux approches se complètent, puisqu'elles répondent à des logiques semblables et que chacune peut être représentative à sa manière. Toutes deux portent sur un échantillon qui est par définition la représentation de certaines données. Ainsi pour mesurer le degré de comparabilité constituant notre corpus nous utiliserons pour un premier tri les critères stylistiques, et nous affinerons par les critères quantitatifs.

4. Corpus comparables bilingues et extraction automatique de lexique bilingue en domaine de spécialité

Partant du principe de l'observation de la langue en usage, l'utilisation des corpus dans le domaine des études touchant à la traduction connait une popularité grandissante. Le manque de corpus bilingues pour les paire de langues ne faisant pas intervenir l'anglais, et pour certaines langues tel que l'arabe, le manque de corpus bilingues de bonne qualité a fait que de plus en plus de recherches en TAL utilisent les corpus comparables bilingues. Ils sont utilisés en analyse pragmatique du discoure, pour tester les outils d'analyse textuelle, les outils pour aide à la traduction, ils sont utilisés pour la reconnaissance sémantique, ainsi qu'en lexicographie, mais qu'en est il pour leur utilisation dans la création d'approches quantitatives pour l'identification et l'extraction de ressources traductionnelles ? Qu'en est- il pour l'extraction automatique de lexiques dans des domaines de spécialité en pleine évolution, tels que l'intelligence artificielle et l'informatique?

Il est vrai que d'une part la tache est très ambitieuse et, qu'elle est beaucoup plus complexe que lors de l'utilisation des bis textes, dans le sens ou, contrairement au bi texte, dans le corpus bilingue comparable l'espace de recherche de traduction ne peut être réduit aux segments alignés, les deux volets du corpus n'étant pas en relation de traduction. D'autre part la matière première des corpus comparables bilingues est très accessible (données /textes), le web constitue à lui seul une source accessible et inépuisable. Mais là où les corpus bilingues comparables assurent une longueur d'avance certaine sur les bis- textes c'est que les premiers préservent l'usage réel et originel des termes. Puisque les deux (ou plus) volets du corpus sont des productions monolingues. Alors que dans les corpus bilingues lors de la traduction le choix du vocabulaire de la langue source peut influencer le traducteur;

par exemple l'utilisation des cognats peut être privilégiée dans la version.

Alors comment apparier les deux volets d'un corpus bilingue comparable, et réaliser ainsi une correspondance au niveau sémantique, pour en extraire automatiquement des ressources traductionnelles? Comment ça marche d'un point de vue formel ?

4.1. Sémantique distributionnelle

A la base de toute démarche d'extraction de lexiques bilingues de corpus comparables, la sémantique distributionnelle. Dans un contexte monolingue Z. HARRIS postule que le sens d'une unité lexicale peut être défini par ses contextes. Autrement dit, le sens d'un mot est décrit par sa distribution sur un ensemble de contextes ; deux mots ayant la même distribution sont sémantiquement liés. Par distribution ou contexte on entend deux mots ayant les mêmes mots comme voisins. Ce qui induit d'un point de vue statistique, que plus deux vocables ont les mêmes listes de mots avoisinants dans un corpus, plus la liaison sémantique entre eux est forte. Sauf que les relations sémantiques sont diverses. C'est soit une relation de synonymie, d'antonymie, d' hyperonymie, d'hyponymie, de métonymie… Pour départager la nature de la liaison sémantique HARRIS nous renvoie à la nature des contextes, car selon lui la nature des contextes a une incidence sur la nature des liaisons sémantiques.

Transposer dans un cadre bilingue la sémantique distributionnelle postulerait l'hypothèse suivante : deux mots dans deux langues différentes ayant les mêmes distributions sont en relation de traduction. Pour l'ordinateur ''les mêmes distributions ''ou'' un contexte commun'', c'est la liste des mots avoisinants, c'est-à-dire liste des cooccurrences. On dira alors qu'un mot en langue A dont la distribution est similaire à

celle d'un mot en langue B est avec une forte probabilité, traduction de ce mot. Pratiquement on compare les distributions d'un mot X source avec celle de tous les mots cibles candidats à sa traduction pour ne retenir que les plus proches. On dira donc que la distribution d'un mot source X est proche de la distribution d'un mot cible Y par rapport à une mesure de la proximité du mot X à l'ensemble des mots cibles avoisinant le mot Y. Ainsi pour chaque mot on aura deux mesures de proximité la première ; mesure la proximité des mots de la langue source avec ceux de la langue cible, et la deuxième mesure la proximité des mots de la langue cible avec ceux de la langue source.

4.2 Symétrie distributionnelle

Mais ce qui pourrait faire défaut à la méthode de sémantique distributionnelle dans le contexte bilingue, c'est encore et toujours la diversité des relations sémantiques dont elle rend compte. Relations qui risqueraient de déséquilibrer la liste des mots avoisinant, et les vecteurs la représentant, d'une langue à l'autre dans le cas de mots polysémiques. Par exemple dans la paire de termes '' disquettes et son équivalent arabe أقراص, le mot أقراص sera proche des mots associés à ses multiples significations alors que le mot non polysémique **disquette** ne sera proche que des mots associés à son unique signification. Le mot disquette aura certainement une liste de cooccurrences bien plus réduite que le mot أقراص. Les vecteurs qui vont représenter statistiquement les distributions des mots seront très différents. Pour cerner la relation sémantique entre les termes de la langue source et les termes candidats à leur traduction dans la langue cible ZEWEIGNBAUM postule que ''si deux mots sont proches dans une direction de traduction, ainsi que dans l'autre, alors ils ont de plus fortes probabilités d'être en relation de traduction, que s'ils ne sont proches que

dans une seule direction de traduction. A cette relation de traduction transitive il donne le nom de symétrie distributionnelle[6].

Bien que le père des logarithmes soit MAHMOUD IBN MOUSSE EL KHAWARISMI, la modélisation de la langue arabe laisse à désirer. Car l'avènement de l'internet a mis en évidence une réalité : l'informatique en arabe a besoin d'un arsenal linguistique pour faire face à l'évolution constante dans ce domaine. C'est dans ce but que s'inscrit notre réflexion.

5 .Constitution du corpus

Pour la constitution du corpus arabe nous avons puisé dans les sites web en langue arabe, traitant du software dans un but didactique, et nous avons récoltés 40 articles écrits par des informaticiens universitaires arabes, sur le domaine de la recherche en modélisation de la langue arabe, sur des recherches concernant la description formelle de la langue arabe dans le cadre de projet tel que le projet MEDAR[7] . Le corpus français a été constitué aussi de textes didactiques en software, et d'articles d'expression française, sur les recherches en modélisation des langues naturelles. Pour la création du corpus comparables, nous avons donc commencé par un premier tri par les critères qualitatifs ensuite nous avons affiné par des critères quantitatifs, cités plus haut, croyant en leur complémentarité

6. Exploitation du corpus

Après constitution du corpus, nous avons dû d'abord étiqueter les deux parties du corpus pour limiter le problème des polysémies entre les termes homographes, qui auraient

6 -CHIAO , YC- 2004- hal. Archive- ouvertes.fr

7 - www.rdig-eg.com/projects/MEDAR.htm

causé une ambigüité déroutante lors du traitement des données. Nous avons étiqueté (associer un mot à sa fonction grammaticale grâce à leur définition et leur contexte) la partie française à l'aide d'une version d'un étiqueteur en ligne Tree Tager[8]. L'étiquetage de la partie arabe s'est fait manuellement car l'étiquetage automatique de l'arabe reste jusqu'à présent problématique à cause des deux spécificités de la langue arabe, à savoir: l'agglutination et la non voyellation. Ce fut un travail de longue haleine mais nécessaire pour nettoyer le corpus et débroussailler le terrain en limitant l'ambigüité.

En deuxième lieu nous avons procédé à l'extraction monolingue de terminologie informatique, en français puis en arabe. Nous nous sommes fait aidé par des professionnels du domaine, car on n'a pas encore pu accéder à des logiciel d'extraction monolingue tel que ACABIT développés par BEATRICE DAILLE, LEXTER (canada) ou encore XELDA. Des outils à base d'algèbre linéaire et de théorie de terminologie postulant que le terme est un symbole d'une notion[9]. Outils qui ne sont d'ailleurs pas fonctionnels pour la langue arabe.

A l'aide d'un outil de statistique textuelle LEXICO 3[10] qui fonctionne aussi sur l'arabe et un outil de calcule formel nous avons calculé des vecteurs de contexte de l'ensemble des termes du corpus. Pour un mot donné un vecteur de contexte représente les mots qui apparaissent le plus fréquemment dans son entourage. Ensuite par le même outil nous avons pu calculer les vecteurs de similarités de l'ensemble des termes (pour un mot donné un vecteur de similarité représente les mots qui apparaissent avec les mêmes entourages) nous avons utilisé des mesures de similarité tel que Cosinus et Jaccard.

[8] www.ims-uni-stuttgart.de/projekte/CRPLEX/ Tree Tager
[9] www .btl.termiumplus.gc.ca/…/1987_terminologie_avenir_f.htm
[10] www.cavi.univ-paris3.fr/.../ilpga/tal/lexicoWWW/lexico3.htm

Pour chaque terme dont on cherche la traduction, on traduit l'ensemble du vecteur de similarité français vers l'arabe. On obtient ainsi un vecteur de traduction en langue arabe. Finalement l'identification des vecteurs de contexte les plus proches du vecteur de traduction donne un ensemble de termes candidats à la traduction du terme français initial.

Nous avons limité la taille du corpus français, car les textes informatiques écrits en arabe ne sont pas faciles à trouver, et beaucoup d'outils nous aidant au traitement des données ne fonctionnent pas sur l'arabe .Nous avons voulu cette expérience qui est toujours en cours et dont les résultats ne sont pas encore finaux, pour tester la validité de la démarche, en attendant d'outillages propres pour l'arabe, en cour d'acquisition.

Des recherches dans la même visée ont été faites par des chercheurs sur d'autres paires de langues et sur des corpus de langue générale. Les approches furent basées sur des analyses purement statistiques. Reinhard Rapp en 1995, sur un corpus journalistique comparable constitué de quotidiens anglais et allemands, obtint jusqu'à 72 % de score[11] . HERVE DEJAN travaillant sur un corpus de spécialité'' le domaine médical'' anglais allemands; des résumés d'articles de chercheurs dans le domaine, et associant une analyse linguistique à l'analyse statistique, obtint un score de 63 %[12].

Pour notre part nous avons obtenu un score de 43 %, en introduisant dans la démarche la notion de symétrie distributionnelle, en plus de l'analyse statistique et la notion de sous langage, i.e. les termes de spécialité qui ont une sémantique « gérable », c'est-à-dire que les mots sont

[11] RAPP Reinhard, Automatic Identification of Word,Translations from Unrelated Corpora, In Proceedings of ACL , 1999 .
[12] Herve DEJEAN & Eric Gaussier, Une nouvelle approche à l'extraction de lexiques bilingues à partir de corpus comparables dans Lexicométrica/Thema 6 2002

relativement univoque. Nous pensons que si notre score est relativement faible ce n'est pas à défaut de démarche, mais plutôt à cause du volume relativement réduit du corpus, et parce que les termes informatiques arabes sont généralement des mots composés, alors que ceux de la version française sont des termes simples. Ce qui posa problème lors du traitement statistique. CURSUER/مؤشر الكتابة CLAVIER لوحة المفاتيح , police / نوع الخط.....60 % des termes candidats à la traduction sont des mots français arabisés, (juste transcrit en arabe) انترنت اكسبلولر،تويتر، الشات ... Même si, il existe des termes proprement arabes, utilisés par peu d'informaticiens. Des mots tel que ألقوريتمية, الخوارزمية, المرتيب, محادثة, les locuteurs préfèrent utiliser الحاسوب, الشات .....

Conclusion

Extraire un lexique bilingue à partir d'un corpus bilingue dont les deux volets ne sont pas en relation de traduction peut sembler trop ambitieux. Une approche à la fois linguistique et statistique basée sur la symétrie distributionnelle, et les mesures de similarité rend ceci très possible. Ceci est d'autant plus faisable quand il s'agit de textes de spécialités ou les termes sont généralement univoques. Sauf que la qualité du corpus est un facteur à ne pas négliger, il faut qu'il soit soumis à une expertise humaine, et qu'il soit de taille suffisante.

Nous pensons que cette approche peut constituer en premier lieu un moyen de créer des ressources traductionnels en langue arabe, dans le domaine de l'informatique. En second lieu elle peut être le terrain de création d'outil d'aide à la traduction de et vers l'arabe, avec peut être une dose d'intelligence artificielle. Un outil qui fera ressortir le terme arabe dés que son homologue français survient dans le texte.

Quant au fait que les termes informatiques arabes soient pour la plupart juste arabisés de l'anglais ou du français, nous pensons que la tache incombe aux informaticiens arabes de créer des termes informatiques proprement arabes.

Bibliographie

- CHIAO YC, (2004) *Extraction lexicale bilingue de texte médicaux comparables : application à la recherche*
- *d'information* .Ttranslangue dans hal. archive-ouverte.
- DAILLE, B, (2005) *Découverte et exploitation de corpus comparables pour l'accès à l'information multilingue* dans DECO. Université de Nantes. France.
- MARCUS,P. SANTORINI,B and
MARCINKIEWIZC, M .A (1994). *Building a large annotated corpus of English* : The Penn
Tree-bank, Computational Linguistics.United Kingdom .
- *KILGARIFF Adam,(2001) Comparing Corpora, International Journal of Corpus Linguistic,* Liverpool University. United Kingdom.