TALN 2011, Montpellier, 27 *juin* –1er *juillet 2011*

Degré de comparabilité, extraction lexicale bilingue et recherche d'information interlingue

Bo Li[1] Eric Gaussier[1] Emmanuel Morin[2] Amir Hazem[2]
(1) Université Grenoble I, LIG UMR 5217
(2) LINA, UMR 6241,Université de Nantes
{bo.li,eric.gaussier}@imag.fr, {emmanuel.morin,amir.hazem}@univ-nantes.fr

Résumé. Nous étudions dans cet article le problème de la comparabilité des documents composant un corpus comparable afin d'améliorer la qualité des lexiques bilingues extraits et les performances des systèmes de recherche d'information interlingue. Nous proposons une nouvelle approche qui permet de garantir un certain degré de comparabilité et d'homogénéité du corpus tout en préservant une grande part du vocabulaire du corpus d'origine. Nos expériences montrent que les lexiques bilingues que nous obtenons sont d'une meilleure qualité que ceux obtenus avec les approches précédentes, et qu'ils peuvent être utilisés pour améliorer significativement les systèmes de recherche d'information interlingue.

Abstract. We study in this paper the problem of enhancing the comparability of bilingual corpora in order to improve the quality of bilingual lexicons extracted from comparable corpora and the performance of cross-language information retrieval (CLIR) systems. We introduce a new method for enhancing corpus comparability which guarantees a certain degree of comparability and homogeneity, and still preserves most of the vocabulary of the original corpus. Our experiments illustrate the well-foundedness of this method and show that the bilingual lexicons obtained are of better quality than the lexicons obtained with previous approaches, and that they can be used to significantly improve CLIR systems

Mots-clés : Corpus comparables, comparabilité, lexiques bilingues, recherche d'information interlingue.

Keywords: Comparable corpora, comparability, bilingual lexicon, cross-language information retrieval.

1 Introduction

Les lexiques bilingues sont une ressource incontournable dans différentes applications multilingues du traitement automatique des langues comme la traduction automatique (Och & Ney, 2003) ou la recherche d'information interlingue (Ballesteros & Croft, 1997). Dans la mesure où la constitution manuelle de lexiques bilingues est une tâche coûteuse et qu'il est difficilement envisageable de développer un lexique pour chaque domaine d'étude, les recherches se sont intéressées à l'extraction automatique de ces lexiques à partir de corpus. Dans la mesure où la plupart des corpus bilingues existants sont par essence comparables, c'est-à-dire qu'ils regroupent des documents dans des langues différentes traitant du même domaine sur la même période sans être en relation de traduction, différents travaux s'intéressent à l'extraction de lexiques bilingues à partir de corpus comparables (Fung & McKeown, 1997; Fung & Yee, 1998; Rapp, 1999; Déjean *et al.*, 2002; Gaussier *et al.*, 2004; Robitaille *et al.*, 2006; Morin *et al.*, 2007; Garera *et al.*, 2009; Yu & Tsujii, 2009; Shezaf & Rappoport, 2010, entre autres). Le

socle commun à ces travaux est de reposer sur une hypothèse de distribution qui postule que les mots qui sont en correspondance de traduction sont susceptibles d'apparaître dans des contextes identiques pour des langues différentes. En s'appuyant sur cette hypothèse fondatrice, les chercheurs ont aussi cherché à identifier de meilleures représentations pour le contexte des mots de même qu'à utiliser différentes méthodes pour mettre en correspondance les mots entre différentes langues toujours en s'appuyant sur cette représentation du contexte. Ces méthodes semblent avoir atteint leur limite en termes de performance et les améliorations les plus récentes concernent plus le cadre d'évaluation des ces approches, plus contraint et limité (Yu & Tsujii, 2009), ou encore le traitement de langues spécifiques (Shezaf & Rappoport, 2010). Plus récemment, et en s'éloignant des approches traditionnelles, Li & Gaussier (2010) ont proposé une approche basée sur l'amélioration de la comparabilité des corpus comme préalable à l'extraction lexicale bilingue. Cette approche postule qu'il ne sert à rien d'essayer d'extraire des lexiques bilingues à partir d'un corpus avec un faible degré de comparabilité puisque la probabilité de trouver des traductions d'un mot donné sera faible dans une telle situation. Notre étude se situe dans la même veine que cette précédente approche et vise dans un premier temps à améliorer la comparabilité d'un corpus donné, tout en préservant une large part de son vocabulaire. Néanmoins, nous nous différencions de ce précédent travail en montrant qu'il est possible de garantir un certain degré d'*homogénéité* du corpus amélioré, et que celle-ci induit une amélioration significative de la qualité du corpus résultant et des lexiques bilingues extraits. En outre, nous montrons que les lexiques extraits avec notre approche améliorent de manière manifeste les résultats d'un système de recherche d'information interlingue, même lorsque ces lexiques sont issus d'un corpus différent de la collection interrogée.

2 Améliorer le degré de comparabilité d'un corpus

Nous commençons par donner dans cette partie la mesure de comparabilité que nous utilisons, avant de décrire un algorithme permettant d'améliorer la comparabilité d'un corpus donné. Nous fournissons également une preuve du bien-fondé de notre algorithme, ainsi qu'une approximation conduisant à une implantation efficace. Pour des raisons pratiques, notre discussion se fera sur la base du couple de langues anglais-français.

2.1 Mesure de comparabilité

Afin de mesurer le degré de comparabilité d'un corpus bilingue, nous utilisons la mesure développée dans (Li & Gaussier, 2010) : étant donné un corpus comparable $\mathcal{P}$ constitué d'une partie anglaise $\mathcal{P}_e$ et d'une partie française $\mathcal{P}_f$, le degré de comparabilité de $\mathcal{P}$ est défini comme l'espérance de trouver la traduction d'un mot du vocabulaire source (respectivement cible) dans le vocabulaire cible (respectivement source). Soit σ une fonction indiquant si une traduction de l'ensemble des traductions possibles $\mathcal{T}_w$ du mot w se trouve dans le vocabulaire $\mathcal{P}^v$ du corpus $\mathcal{P}$, c'est-à-dire :

$$\sigma(w, \mathcal{P}) = \begin{cases} 1 & \text{si } \mathcal{T}_w \cap \mathcal{P}^v \neq \emptyset \\ 0 & \text{sinon} \end{cases}$$

et soit $\mathcal{D}$ un dictionnaire bilingue dont le vocabulaire anglais (respectivement français) est noté $\mathcal{D}_e$ (respectivement $\mathcal{D}_f$). La mesure du degré de comparabilité M est définie par :

$$M(\mathcal{P}_e, \mathcal{P}_f) = \frac{\sum_{w \in \mathcal{P}_e \cap \mathcal{D}_e} \sigma(w, \mathcal{P}_f) + \sum_{w \in \mathcal{P}_f \cap \mathcal{D}_f} \sigma(w, \mathcal{P}_e)}{\#_w(\mathcal{P}_e \cap \mathcal{D}_e) + \#_w(\mathcal{P}_f \cap \mathcal{D}_f)}$$

où $\#_w(\mathcal{P})$ représente le nombre de mots différents présents dans $\mathcal{P}$. Comme on peut le voir d'après cette définition, M mesure la proportion de mots source et cible dont une traduction est présente dans le vocabulaire cible et source de $\mathcal{P}$. Pour des raisons qui deviendront claires plus tard, nous utiliserons aussi des mesures partielles où seuls les vocabulaires français ou anglais sont considérés. Ainsi, la proportion de mots anglais traduits sera notée M_{ef}, définie par : $\frac{\sum_{w \in \mathcal{P}_e \cap \mathcal{D}_e} \sigma(w, \mathcal{P}_f)}{\#_w(\mathcal{P}_e \cap \mathcal{D}_e)}$. La mesure M_{fe} est définie de la même façon.

2.2 Classer les documents pour une meilleure comparabilité

L'hypothèse distributionnelle sous-tendant l'extraction de lexiques bilingues est d'autant plus valide que les documents dans les différentes langues couvrent des thématiques proches, car les auteurs ont alors tendance à puiser dans le même vocabulaire (voir (Morin *et al.*, 2007) pour une analyse reliée). En d'autres termes, si un corpus couvre un nombre limité de thématiques, il est plus à même de contenir une information répétée et cohérente qui pourra être exploitée au mieux pour l'extraction de lexiques bilingues. Le terme *homogénéité* rend compte de ce phénomène et nous dirons, de façon informelle, qu'un corpus est homogène s'il couvre un nombre limité de thématiques. Nous conjecturons ici que si l'on peut garantir un certain degré d'homogénéité, en plus d'un certain degré de comparabilité, alors les lexiques bilingues extraits seront de meilleure qualité. Comme nous le verrons, cette conjecture sera validée par les expériences menées. De façon à garantir un certain degré d'homogénéité, nous nous appuyons sur des techniques de classification non supervisée (*clustering*). Nous utilisons ici des techniques de classification agglomérative ascendante, mais toute autre technique, pour peu qu'elle dispose d'une procédure de filtrage adaptée, peut être utilisée.

2.2.1 Algorithme de classification bilingue

L'ensemble du processus permettant de construire, à partir d'un corpus donné, un corpus plus homogène et de plus fort degré de comparabilité peut être résumé par les étapes suivantes :

1. À partir de la mesure de similarité, définie en 2.2.2 et fondée sur la mesure de comparabilité présentée ci-dessus, et de l'ensemble des documents anglais et français du corpus originel $\mathcal{P}$, construire les dendrogrammes en suivant les étapes classiques de la classification agglomérative ascendante ;
2. Filtrer les dendrogrammes en ne retenant que les classes les plus profondes (voir ci-dessous) ;
3. Fusionner les classes retenues pour former un nouveau corpus $\mathcal{P}_H$, qui contient une sous-partie homogène et fortement comparable de $\mathcal{P}$;
4. Répéter les étapes ci-dessus pour enrichir la partie restante de $\mathcal{P}$ (partie qui sera notée $\mathcal{P}_L$, $\mathcal{P}_L = \mathcal{P} \setminus \mathcal{P}_H$) avec des documents extraits d'autres corpus.

Les trois premières étapes sont détaillées dans l'algorithme 1, où CAA signifie Classification Agglomérative Ascendante. Comme on peut le remarquer, seul $\mathcal{P}$ est utilisé pour construire $\mathcal{P}_H$, à travers des étapes de classification et de filtrage. Ainsi, l'algorithme 1 vise à extraire de $\mathcal{P}$ une sous-partie fortement comparable et homogène. Une fois cela réalisé, c'est-à-dire une fois que $\mathcal{P}$ a été exploité, il est nécessaire de recourir à des ressources externes si l'on veut construire un corpus fortement comparable à partir de $\mathcal{P}_L$ (qui est la partie restante de $\mathcal{P}$). Pour cela, deux nouveaux corpus comparables sont considérés dans l'étape 4 du processus global : le premier consiste en la partie anglaise de $\mathcal{P}_L$ et la partie française d'un autre corpus $\mathcal{P}_T$; le second consiste en la partie française de $\mathcal{P}_L$ et la partie anglaise de $\mathcal{P}_T$. Les deux sous-parties fortement comparables et homogènes obtenues à partir de ces deux corpus sont alors ajoutées à $\mathcal{P}_H$ pour constituer le corpus final. L'utilisation de la classification agglomérative ascendante et du filtrage associé garantit que le corpus final est homogène. La propriété 1 que nous présentons plus

Algorithm 1: Algorithme de classification bilingue

Entrée :

Ensemble $\mathcal{U}$ de tous les documents anglais et français de $\mathcal{P}$

Réel positif θ (seuil de profondeur)

Sortie :

$\mathcal{P}_H$, sous-partie fortement comparable et homogène de $\mathcal{P}$

```
Initialiser P_H = ∅ ;
CAA(U) → ensemble S de dendrogrammes
for chaque dendrogramme T de S do
    m ← profondeur maximale de T ;
    for tous les nœuds n de T do
        if profondeur(n) ≥ m · θ then
            Ajouter tous les documents sous le nœud n à P_H ;
        end if
    end for
end for
Supprimer les doublons de P_H ;
return P_H ;
```

loin établit que ce corpus est fortement comparable. Mais avant de voir en détail cette propriété, nous introduisons la mesure de similarité utilisée.

2.2.2 Mesure de similarité

Imaginons deux classes de documents bilingues $\mathcal{C}_1$ et $\mathcal{C}_2$. Pour la tâche d'extraction de lexiques bilingues, ces deux classes sont similaires et devraient être regroupées si leur combinaison permet de compléter le contenu de chacune des classes prise isolément, ou, en d'autres termes, si la partie anglaise $\mathcal{C}_1^e$ de $\mathcal{C}_1$ et la partie française $\mathcal{C}_1^f$ de $\mathcal{C}_1$ sont comparables à leur contrepartie dans l'autre classe (respectivement la partie française $\mathcal{C}_2^f$ de $\mathcal{C}_2$ et la partie anglaise $\mathcal{C}_2^e$ de $\mathcal{C}_2$) [1]. Ceci conduit à la mesure de similarité suivante pour $\mathcal{C}_1$ et $\mathcal{C}_2$:

$$sim(\mathcal{C}_1, \mathcal{C}_2) = \beta M(\mathcal{C}_1^e, \mathcal{C}_2^f) + (1-\beta) M(\mathcal{C}_2^e, \mathcal{C}_1^f) \tag{1}$$

où $\beta\,(0 \leq \beta \leq 1)$ est un poids qui permet de contrôler l'importance de chacune des deux parties ($\mathcal{C}_1^e$, $\mathcal{C}_2^f$) et ($\mathcal{C}_2^e$, $\mathcal{C}_1^f$). De façon intuitive, on aimerait donner plus de poids dans cette combinaison à la partie la plus importante, car elle contient plus d'information. Si nous utilisons le nombre de paires de documents anglais-français pour quantifier cette information, le poids β peut être défini comme la proportion de paires de documents dans ($\mathcal{C}_1^e$, $\mathcal{C}_2^f$) sur l'ensemble des paires de documents dans le corpus fusionné :

$$\beta = \frac{\#_d(\mathcal{C}_1^e) \cdot \#_d(\mathcal{C}_2^f)}{\#_d(\mathcal{C}_1^e) \cdot \#_d(\mathcal{C}_2^f) + \#_d(\mathcal{C}_2^e) \cdot \#_d(\mathcal{C}_1^f)}$$

où $\#_d(\mathcal{C})$ représente le nombre de documents dans $\mathcal{C}$. Dans la mesure où les classes sont tout d'abord formées de documents anglais et français isolés, la mesure de similarité correspond à un score de comparabilité normalisé entre les corpus anglais et français qui forment la nouvelle classe. Cependant, cette mesure ne tient pas compte des

1. Dans la mesure où $\mathcal{C}_1$ et $\mathcal{C}_2$ sont des classes, leurs parties anglaise et française sont comparables par construction.

longueurs relatives des corpus anglais et français, qui ont pourtant un impact sur la qualité des corpus bilingues extraits. Si une contrainte de type 1-1 (c'est-à-dire imposant à chaque classe de contenir le même nombre de documents anglais et français) est trop forte, se reposer sur des classes par trop déséquilibrées n'est pas non plus souhaitable. Nous introduisons donc une nouvelle fonction ϕ qui a pour but de pénaliser les classes pour lesquelles les nombres de documents anglais et français sont trop différents :

$$\phi(\mathcal{C}) = \frac{1}{(1 + log(1 + \gamma \frac{|\#_d(\mathcal{C}^e) - \#_d(\mathcal{C}^f)|}{\min(\#_d(\mathcal{C}^e)), \#_d(\mathcal{C}^f))})} \tag{2}$$

avec $\gamma \in \mathbb{R}^+$. Cette fonction de pénalité fournit une nouvelle mesure de similarité sim_l qui est celle utilisée dans l'algorithme 1 :

$$sim_l(\mathcal{C}_1, \mathcal{C}_2) = sim(\mathcal{C}_1, \mathcal{C}_2) \cdot \phi(\mathcal{C}_1 \cup \mathcal{C}_2) \tag{3}$$

Dans la suite de cette étude, γ est fixé à 1 dans ϕ.

2.2.3 Analyse théorique

Le processus de classification utilisé dans l'algorithme 1 garantit que les documents qui portent sur la *même thématique* seront regroupés avant les documents portant sur des *thématiques différentes*. Le corpus obtenu ($\mathcal{P}_H$) sera ainsi homogène, c'est-à-dire qu'il ne couvrira qu'un nombre restreint de thématiques. De plus, le fait que le corpus comparable (que nous noterons $\mathcal{P}_F$) obtenu au travers des étapes 1 à 4 découle du corpus originel $\mathcal{P}$ indique que la plus grande partie du vocabulaire de $\mathcal{P}$ sera préservée dans $\mathcal{P}_F$. Nous verrons dans la partie expérimentale que c'est bien le cas. Ce qui semble moins évident, c'est le fait que le processus que nous avons défini garantisse un fort degré de comparabilité. La propriété suivante établit que c'est bien le cas.

Propriété 1 *Soit $\mathcal{C}_1$ et $\mathcal{C}_2$ deux classes de documents qui doivent être regroupées dans le processus de classification. Nous faisons l'hypothèse que le dictionnaire bilingue $\mathcal{D}$ a été construit indépendamment des documents traités, ce qui implique que le degré de comparabilité M_{ef} (respectivement de même pour M_{fe}) est à peu près le même pour différentes parties du corpus*[2]*. Nous faisons de plus l'hypothèse que :*

$$\textit{(I)} \quad \frac{|\mathcal{C}_1^e \cup \mathcal{C}_2^e|}{|\mathcal{C}_2^e|} = \frac{|\mathcal{C}_1^f \cup \mathcal{C}_2^f|}{|\mathcal{C}_2^f|}$$

Alors :

$$M(\mathcal{C}_1^e \cup \mathcal{C}_2^e, \mathcal{C}_1^f \cup \mathcal{C}_2^f) \geq min(M(\mathcal{C}_1^e, \mathcal{C}_1^f), M(\mathcal{C}_2^e, \mathcal{C}_2^f))$$

Démonstration (esquisse) : Soit $V = \mathcal{C}_1^e \cap \mathcal{C}_2^e$. En utilisant le fait que $M_{ef}(\mathcal{C}_i^e, \mathcal{C}_i^f) \leq M_{ef}(\mathcal{C}_i^e, \mathcal{C}_i^{f'})$ pour tout $\mathcal{C}_i^{f'}$ tel que $\mathcal{C}_i^f \subseteq \mathcal{C}_i^{f'}$ (et de même pour la direction français vers anglais), nous avons, pour $i = 1, 2$:

$$\sum_{w \in \mathcal{C}_i^e \setminus V} \sigma(w, \mathcal{C}_1^f \cup \mathcal{C}_2^f)) \geq |\mathcal{C}_i^e \setminus V| M_{ef}(\mathcal{C}_i^e, \mathcal{C}_i^f)$$

et, pour les mots de V :

$$\sum_{w \in V} \sigma(w, \mathcal{C}_1^f \cup \mathcal{C}_2^f)) \geq |V| \max(M_{ef}(\mathcal{C}_1^e, \mathcal{C}_1^f), M_{ef}(\mathcal{C}_2^e, \mathcal{C}_2^f))$$

2. En d'autres termes, la proportion de mots anglais (respectivement français) traduits dans le corpus français (respectivement anglais) est homogène sur l'ensemble du corpus.

Alors, d'après l'hypothèse d'indépendance entre corpus et dictionnaire faite en énonçant la propriété 1 :

$$\sum_{w \in (\mathcal{C}_1^e \cup \mathcal{C}_2^e) \cap D_e} \sigma(w, \mathcal{C}_1^f \cup \mathcal{C}_2^f))$$
$$\geq |(\mathcal{C}_1^e \cup \mathcal{C}_2^e) \cap D_e| \min(M_{ef}(\mathcal{C}_1^e, \mathcal{C}_1^f), M_{ef}(\mathcal{C}_2^e, \mathcal{C}_2^f))$$

Un développement similaire sur M_{fe} et l'utilisation de la condition (I) complètent la démonstration.

La propriété précédente garantit que la classe obtenue en fusionnant deux classes existantes a un degré de comparabilité au moins égal à celui de la classe la moins comparable. Le degré de comparabilité ne peut donc décroître dans le processus de classification agglomérative. Comme l'on commence par fusionner les documents les plus comparables, on ne construit que des classes avec un bon degré de comparabilité. Enfin, la condition (I) a de grandes chances d'être réalisée car tous les corpus sont prétraités de façon à éliminer les documents trop courts ou trop longs, souvent source de bruit, et la pénalité utilisée dans la mesure de similarité fournit des classes comprenant des nombres comparables de documents dans les deux langues. Le processus global que nous avons défini permet donc d'obtenir des corpus homogènes et fortement comparables.

2.3 Considérations informatiques

Dans la mesure où les corpus comparables disponibles à l'heure actuelle comprennent en général un nombre important de documents, la classification agglomérative peut s'avérer trop coûteuse. Nous proposons ici une borne inférieure de la mesure de comparabilité qui peut être calculée efficacement ainsi qu'une mise à jour efficace de la matrice de similarité pendant le processus de classification. Le fait de se reposer sur une borne inférieure de la mesure de similarité garantit que les classes obtenues auront un bon degré de comparabilité, car seules les classes les plus similaires sont regroupées à chaque itération de l'algorithme de classification. La propriété suivante établit une telle borne inférieure, sur la base du degré de comparabilité moyen des paires de documents.

Propriété 2 *Soit $\mathcal{P}$ un corpus comparable comprenant une partie anglaise $\mathcal{P}_e$ et une partie française $\mathcal{P}_f$, et soit $\mathcal{D}$ un dictionnaire bilingue, $\mathcal{D}_e$ dénotant le vocabulaire anglais et $\mathcal{D}_f$ le vocabulaire français. Supposons que le dictionnaire est distribué de façon uniforme sur le corpus, c'est-à-dire que :*

$$\forall d_e \in \mathcal{P}_e, \frac{\#_w(d_e \cap \mathcal{D}_e)}{\#_w(d_e)} = \frac{\#_w(\mathcal{P}_e \cap \mathcal{D}_e)}{\#_w(\mathcal{P}_e)}$$

et de même pour la partie française. Supposons de plus que tous les documents, ainsi que les parties anglaise et française du corpus, ont à peu près la même longueur :

$$\forall d_e \in \mathcal{P}_e \text{ and } d_f \in \mathcal{P}_f, \frac{\#_w(d_e)}{\#_w(\mathcal{P}_e)} \simeq \frac{\#_w(d_f)}{\#_w(\mathcal{P}_f)} (= \lambda)$$

Alors :

$$M(\mathcal{P}_e, \mathcal{P}_f) \geq \frac{1}{\#_d(\mathcal{P}_e) \cdot \#_d(\mathcal{P}_f)} \sum_{d_e \in \mathcal{P}_e, d_f \in \mathcal{P}_f} M(d_e, d_f)$$

Nous ne détaillons pas ici la démonstration de cette propriété, purement technique. La première hypothèse faite semble raisonnable (et rejoint celle faite dans la propriété précédente) en l'absence de toute connaissance *a priori* sur les thématiques couvertes par le corpus et leur lien avec le dictionnaire. La seconde hypothèse est en partie garantie dans notre cas par le processus de construction que nous avons défini et la fonction de pénalité associée.

Remplacer M par la borne ci-dessus dans l'équation 1 conduit à une mesure de similarité qui peut être vue comme la valeur accumulée de toutes les connexions entre deux classes. Il est alors possible de mettre à jour la matrice de similarité de façon itérative. Supposons en effet que le processus de classification doive, à un instant donné, fusionner les classes $\mathcal{C}_1$ et $\mathcal{C}_2$ en une seule classe $\mathcal{C}_{new}$. Un nouveau score de similarité entre $\mathcal{C}_{new}$ et toutes les autres classes doit alors être calculé. La similarité entre $\mathcal{C}_{new}$ et une autre classe $\mathcal{C}_3$ peut s'écrire, à partir de l'équation 3 et de la formule de similarité :

$$sim_l(\mathcal{C}_{new}, \mathcal{C}_3) = \frac{(N_{\mathcal{C}_1} + N_{\mathcal{C}_2})\phi(\mathcal{C}_1 \cup \mathcal{C}_2)}{\#_d(\mathcal{C}^e_{new}) \cdot \#_d(\mathcal{C}^f_3) + \#_d(\mathcal{C}^e_3) \cdot \#_d(\mathcal{C}^f_{new})}$$

où $(j = 1, 2)$ et :

$$N_{\mathcal{C}_j} = \frac{(\#_d(\mathcal{C}^e_j) \cdot \#_d(\mathcal{C}^f_3) + \#_d(\mathcal{C}^e_3) \cdot \#_d(\mathcal{C}^f_j))sim_l(\mathcal{C}_j, \mathcal{C}_3)}{\phi(\mathcal{C}_j \cup \mathcal{C}_3)}$$

Dans le processus de classification, dans la mesure où $sim_l(\mathcal{C}_1, \mathcal{C}_3)$ et $sim_l(\mathcal{C}_2, \mathcal{C}_3)$ sont déjà connus avant le calcul de $sim_l(\mathcal{C}_{new}, \mathcal{C}_3)$, la matrice de similarité peut directement être mise à jour à chaque itération. En notant N_c le nombre de classes avant fusion, la complexité de cette mise à jour est de l'ordre de $\mathcal{O}(N_c)$, alors qu'elle atteint $\mathcal{O}(N_c \times \bar{C}^2)$ si l'on applique directement les équations 1 et 3 ($\bar{C}$ représentant le nombre moyen de documents par classe).

3 Expériences et résultats

Les différentes expériences que nous avons réalisées ont pour objectif d'évaluer : (i) si l'algorithme que nous avons proposé induit des corpus d'une meilleure qualité en ce qui concerne la comparabilité, (ii) si les lexiques bilingues extraits de ces corpus sont eux aussi d'une qualité plus importante, et (iii) si ces lexiques peuvent être utilisés pour améliorer les performances des systèmes de recherche d'information interlingue.

Dans nos expériences, différents corpus sont utilisés : le corpus anglais TREC [3] de l'*Associated Press* (noté *AP*) et les corpus fournis dans les tâches multilingues des campagnes CLEF [4] dont pour l'anglais le *Los Angeles Times* (*LAT94*) et le *Glasgow Herald* (*GH95*) et pour le français *Le Monde* (*MON94*), le *SDA 94* (*SDA94*) et *95* (*SDA95*). Outre ces corpus existants, deux corpus monolingues ont été extraits à partir de *Wikipédia* : le corpus anglais *Wiki-En* construit en retenant l'ensemble des articles appartenant à la catégorie *Society* pour une profondeur inférieure à 4 (soit 33 000 mots anglais distincts) et le corpus français *Wiki-Fr* toujours pour la catégorie *Société* pour une profondeur inférieure à 7 (soit 28 000 mots français distincts). Le dictionnaire bilingue bd_0 nécessaire pour la tâche d'extraction de lexiques est quant à lui construit à partir de dictionnaires en ligne. Dans toutes nos expériences, nous utilisons la méthode décrite dans le présent article complétée par celle présentée dans (Li & Gaussier, 2010). Cette dernière méthode est à notre connaissance la seule approche alternative pour améliorer la comparabilité des corpus, d'où son importance dans l'évaluation.

3.1 Comparabilité de corpus

L'algorithme de classification décrit en section 2.2.1 est utilisé pour améliorer le degré de comparabilité d'un corpus comparable. Les corpus *GH95* et *SDA95* sont utilisés pour construire le corpus comparable $\mathcal{P}^0$ (56 000

3. http://trec.nist.gov/
4. http://www.clef-campaign.org

mots pour l'anglais et 42 000 le français). En outre, nous exploitons deux corpus comparables supplémentaires pour nous assurer que l'efficacité de notre algorithme n'est pas liée à une ressource externe spécifique : i) $\mathcal{P}_T^1$ composé à partir des corpus *LAT94*, *MON94* et *SDA94* (109 000 mots pour l'anglais et 87 000 pour le français) et ii) $\mathcal{P}_T^2$ composé à partir des corpus *Wiki-En* et *Wiki-Fr* (368 000 mots pour l'anglais et 378 000 pour le français).

Après le processus de classification, nous obtenons les corpus $\mathcal{P}^1$ (pour le corpus externe $\mathcal{P}_T^1$) et $\mathcal{P}^2$ (pour le corpus externe $\mathcal{P}_T^2$). Comme nous l'avons indiqué précédemment, nous utilisons aussi la méthode décrite dans (Li & Gaussier, 2010) sur les mêmes données pour comparer nos résultats et obtenons ainsi le corpus $\mathcal{P}^{1'}$ (pour $\mathcal{P}_T^1$) et $\mathcal{P}^{2'}$ (pour $\mathcal{P}_T^2$) à partir de $\mathcal{P}^0$. Au niveau de la couverture lexicale, $\mathcal{P}^1$ couvre 97,9% du vocabulaire de $\mathcal{P}^0$, tandis que $\mathcal{P}^2$ couvre 99,0% de celui de $\mathcal{P}^0$. Nous pouvons ainsi constater qu'une très grande partie du vocabulaire du corpus d'origine a été conservé, ce qui est l'une des exigences de notre approche. En ce qui concerne les scores de comparabilité, $\mathcal{P}^1$ atteint 0,924 et $\mathcal{P}^2$ 0,939. Les deux corpus comparables ont donc bien un degré de comparabilité supérieur au corpus d'origine qui était de l'ordre de 0,881 comme cela est suggérée par la propriété 1. En outre, les corpus $\mathcal{P}^1$ et $\mathcal{P}^2$ sont plus comparables que le corpus $\mathcal{P}^{1'}$ (comparabilité de 0,912) et $\mathcal{P}^{2'}$ (comparabilité de 0,915) ce qui montre bien que l'homogénéité est un élément crucial pour évaluer la comparabilité.

3.2 Extraction de lexiques bilingues

TABLE 1 – Évaluation des lexiques bilingues extraits pour différents corpus comparables

	$\mathcal{P}^0$	$\mathcal{P}^{1'}$	$\mathcal{P}^{2'}$	$\mathcal{P}^1$	$\mathcal{P}^2$	$\mathcal{P}^1 > \mathcal{P}^0$	$\mathcal{P}^2 > \mathcal{P}^0$
Précision	0,226	0,277	0,325	0,295	0,461	0,069 30,5 %	0,235 104,0 %
Rappel	0,103	0,122	0,145	0,133	0,212	0,030 29,1 %	0,109 105,8 %

TABLE 2 – Comparaison de la précision pour différents intervalles de fréquences des mots de la liste d'évaluation

	$\mathcal{P}^0$	$\mathcal{P}^{2'}$	$\mathcal{P}^2$	$\mathcal{P}^{2'} > \mathcal{P}^0$	$\mathcal{P}^2 > \mathcal{P}^0$	$\mathcal{P}^2 > \mathcal{P}^{2'}$
W_l	0,135	0,206	0,304	0,071 52,6 %	0,169 125,2 %	0,098 47,6 %
W_m	0,256	0,390	0,564	0,134 52,3 %	0,308 120,3 %	0,174 44,6 %
W_h	0,434	0,632	0,667	0,198 45,6 %	0,233 53,7 %	0,035 5,5
,**All**	**0,226**	**0,325**	**0,461**	**0,099 43,8 %**	**0,235 104,0 %**	**0,136 41,8 %**

Comme les travaux antérieurs en extraction de lexiques bilingues à partir de corpus comparables exploitent des ressources différentes et opèrent des choix distincts des nôtres, il est relativement difficile de se comparer à ceux-ci (Laroche & Langlais, 2010). En outre, puisque notre approche vise à améliorer la comparabilité de corpus, elle peut être ensuite couplée à une méthode existante d'extraction de lexiques bilingues. Il est donc tout aussi intéressant de directement évaluer si un tel couplage peut conduire à des performances accrues en termes de qualité des lexiques extraits.

L'extraction de lexiques bilingues à partir de corpus comparables repose sur la méthode proposée par Fung & Yee (1998) plus connue maintenant sous le nom d'*approche standard* notamment dans les travaux de (Déjean *et al.*, 2002; Gaussier *et al.*, 2004; Yu & Tsujii, 2009). Dans cette approche, chaque mot est représenté sous la forme d'un vecteur de contexte composé des mots qui co-occurrent avec lui dans une fenêtre donné. Les vecteurs de contexte de la langue source sont ensuite traduits vers la langue cible en s'appuyant sur un dictionnaire bilingue.

Enfin, la traduction d'un mot est obtenue en comparant son vecteur de contexte traduit à l'ensemble des vecteurs de la langue cible à travers une mesure de distance ou similarité vectorielle telle que le cosinus.

3.2.1 Paramètres expérimentaux

Afin d'évaluer la qualité des lexiques bilingues extraits, nous divisons notre dictionnaire bilingue bd_0 en deux parties : 10 % des mots anglais accompagnés de leurs traductions sont choisis aléatoirement et uniquement utilisés comme liste d'évaluation, les 90 % restant sont utilisés pour assurer la traduction des vecteurs de contexte dans l'approche standard. Les mots anglais absents de $\mathcal{P}_e$ ou pour lesquels aucune traduction n'a été trouvée dans $\mathcal{P}_f$ sont retirés de la liste d'évaluation. Pour chaque mot anglais de la liste d'évaluation, tous les mots français de $\mathcal{P}_f$ sont ordonnés suivant leur similarité avec les mots anglais. Les mesures de précision et rappel sont ensuite calculées sur les N premiers candidats. Les valeurs de la précision dans ce cas correspondent à la proportion de listes contenant la traduction correcte (en cas de traductions multiples, une liste est réputée contenir la traduction correcte dès lors que l'une des traductions possibles est présente). Le rappel est quant à lui la proportion de traductions correctes trouvée dans les listes sur toutes les traductions fournies dans le corpus. Cette manière de procéder a été utilisée dans différents travaux antérieurs et peut être maintenant considérée comme un méthode d'évaluation attestée. En outre, plusieurs études ont montré qu'il est plus facile de trouver les traductions correctes pour les mots fréquents que pour les mots rares (Pekar *et al.*, 2006). Afin de prendre en compte ce phénomène, nous distinguons différents intervalles d'effectifs pour évaluer la validité de notre approche. Ainsi, les mots avec un effectif inférieur à 100 sont définis comme étant des mots de faibles fréquence (W_l), ceux avec un effectif supérieur à 400 sont définis comme étant des mots très fréquents (W_h), et enfin les mots dont l'effectif est compris entre ces deux seuils sont considérés comme des mots de fréquence intermédiaire (W_m).

3.2.2 Analyse des résultats

Dans une première série d'expériences, les lexiques bilingues sont extraits à partir des corpus obtenus ii) par notre approche ($\mathcal{P}^1$ et $\mathcal{P}^2$), ii) par la méthode décrite dans (Li & Gaussier, 2010) ($\mathcal{P}^{1'}$ and $\mathcal{P}^{2'}$) et iii) enfin avec le corpus d'origine $\mathcal{P}^0$, avec N fixé à 20. La table 1 présente les résultats obtenus. Les deux dernières colonnes "$\mathcal{P}^1 > \mathcal{P}^0$" et "$\mathcal{P}^2 > \mathcal{P}^0$" indique les différences absolue et relative, exprimées en pourcentage, par rapport à $\mathcal{P}^0$. Comme nous pouvons le constater, les meilleurs résultats sont obtenus à partir des corpus construits avec la méthode que nous avons proposée. Les lexiques extraits à partir du corpus où le degré de comparabilité a été renforcé sont d'une bien meilleure qualité que ceux obtenus à partir du corpus d'origine ou encore du corpus construit avec l'approche de (Li & Gaussier, 2010). La différence de qualité est encore plus notable avec $\mathcal{P}^2$ qui est obtenu à partir d'un corpus externe volumineux $\mathcal{P}^2_T$. Ces résultats semblent confirmer l'intuition qu'il est possible de trouver plus aisément dans des corpus volumineux des documents en relation avec un corpus donné.

Afin d'évaluer la relation entre la qualité de ces méthodes et la fréquence des mots à traduire, nous nous concentrons sur les meilleurs résultats sur $\mathcal{P}^{2'}$ pour l'approche précédente et sur ceux de $\mathcal{P}^2$ pour notre approche. La table 2 résume les résultats obtenus. On remarquera, sans véritablement de surprise, que les résultats obtenus pour les mots ayant une haute fréquence sont meilleurs que ceux obtenus pour les mots de faible fréquence. En outre, notre approche est la meilleure quel que soit l'intervalle de fréquence pris en compte. La précision globale peut être augmentée en relatif de 41,8 % (de 0,325 à 0,461). En comparant $\mathcal{P}^2$ avec le corpus d'origine $\mathcal{P}^0$, nous pouvons noter pour la précision globale, une augmentation relative de 104,0 % (de 0,226 à 0,461), ce qui est très satisfaisant dans ce contexte d'évaluation. Enfin, l'amélioration pour les mots de faible et moyenne fréquence est plus importante pour $\mathcal{P}^2$, ce qui démontre que notre approche se comporte bien mieux sur ce qui est généralement

considéré comme un problème difficile (Pekar *et al.*, 2006).

3.3 Expériences en recherche d'information interlingue

TABLE 3 – Score MAP pour la tâche de recherche d'information interlingue suivant différents dictionnaires bilingues

	mon	bd_1	bd_1+cc_0	bd_1+cc_1	bd_1+cc_2	bd_2	bd_2+cc_0	bd_2+cc_1	bd_2+cc_2
MAP	$0,422$	$0,313$	$0,327^{\bullet}$	$0,328^{\bullet}$	$0,338^{\bullet}$	$0,375$	$0,382$	$0,377$	$0,391^{\bullet}$

Dans la dernière série d'expériences, nous cherchons à évaluer l'apport des différents lexiques extraits à partir de corpus comparables pour une tâche de recherche d'information interlingue. Pour ce faire, nous exploitons les sujets des campagnes CLEF de 2001 et 2002, rassemblant environ 100 sujets distincts, comme requêtes sur une collection de 113 000 documents issus du *Los Angeles Times*. Les sujets anglais correspondants sont utilisés pour interroger la même collection (référence *mon*). Seul le titre et la partie description des sujets CLEF sont utilisés pour construire des requêtes. En outre, les mots outils et les phrases non pertinentes telles que *find documents which report about* sont supprimés des requêtes. La recherche est réalisée avec le modèle Indri du système de recherche d'information Lemur (http ://www.lemurproject.org). Une variante de l'approche introduite dans (Pirkola, 1998) et (Talvensaari *et al.*, 2007) est aussi utilisée pour transformer les sujets français en requêtes en anglais. L'idée est de borner toutes les possibilités de traduction d'un mot français dans le sujet du texte avec un opérateur WSYN. Ensuite, toutes les traductions candidates dans l'opérateur WSYN sont traitées comme des synonymes avec des poids différents.

Dans nos expériences, nous combinons deux dictionnaires bilingues de langue générale bd_1 (68 0000 traductions) et bd_2 (116 000 traductions) avec les lexiques bilingues obtenus automatiquement dans la précédente section. Nous utilisons ici les lexiques cc_0 (extrait de $\mathcal{P}^0$), cc_1 (extrait de $\mathcal{P}^{2'}$) et cc_2 (extrait de $\mathcal{P}^2$). Différentes combinaisons de ces ressources sont réalisées, y compris $bd_{1/2}$, $bd_{1/2}$+cc_0, $bd_{1/2}$+cc_1, $bd_{1/2}$+cc_2. Lorsque qu'un dictionnaire de langue générale et un lexique extrait sont combinés, plus de poids est attribué aux traductions candidat du dictionnaire de langue générale. Le poids des différents mots traduits à partir de $cc_{0/1/2}$ est quant à lui le cosinus entre les vecteurs de contexte de chaque mot (c'est-à-dire le score donné par l'approche standard précédemment évoquée). Le poids pour les traductions trouvées dans le dictionnaire bilingue est fixé empiriquement à 25. Comme il est d'usage en recherche d'information, nous utilisons la mesure MAP (Mean Average Precision) afin d'évaluer les performances des différents systèmes. L'importance des différences entre les différents systèmes est estimée par un t-test apparié de Student (p-value fixée à 0,1). Les résultats obtenus sont indiqués dans la table 3. Pour le dictionnaire de langue générale bd_1, on note toujours une amélioration significative des résultats (identifiée par la marque $\bullet$) du score MAP lorsque l'un des lexiques bilingues extraits du corpus comparables est utilisé. Lorsque bd_2, qui est beaucoup plus riche que bd_1, est utilisé, seulement le lexique bilingue cc_2 extrait avec notre méthode à partir $\mathcal{P}^2$ conduit à une amélioration significative des résultats. Cela montre que cc_2 est supérieure à cc_1 et cc_0 dans la tâche de recherche d'information interlingue, en particulier lorsque le dictionnaire de langue générale utilisé est d'une taille importante. Ces résultats semblent confirmer que notre approche basée sur de la classification est plus adaptée que l'approche gloutonne des travaux précédents de (Li & Gaussier, 2010). Enfin, la combinaison actuelle des lexiques extraits avec le système de recherche d'information interlingue est relativement simple et pourrait être certainement améliorée en exploitant d'autres modèles de combinaison.

4 Conclusion

Dans cet article, nous avons proposé une nouvelle approche pour augmenter le degré de comparabilité des documents constituant un corpus comparable afin d'améliorer la qualité des lexiques bilingues extraits de corpus comparables et les performances des systèmes de recherche d'information interlingue. Nous avons démontré théoriquement puis empiriquement que notre approche permet de garantir un certain degré de comparabilité et l'homogénéité du corpus tout en préservant une large part du vocabulaire du corpus d'origine. Enfin, nos expériences montrent que les lexiques bilingues que nous obtenons sont d'une meilleure qualité que ceux obtenus avec les approches précédentes, et que ces lexiques peuvent être utilisés pour améliorer significativement les résultats des systèmes de recherche d'information interlingue.

Les deux étapes cruciales de notre approche sont d'une part l'extraction d'un noyau fortement comparable du corpus original, et, d'autre part, l'alignement des parties du corpus orignal, non présentes dans ce noyau, avec un corpus externe. Le seuil introduit au niveau du degré de comparabilité permet de contrôler la taille et la qualité du noyau extrait dans la première étape. Si le corpus original n'est que très faiblement comparable, il est alors possible que ce noyau soit vide (ce qui est un résultat souhaitable dans ce cas). Dans tous les cas, excepté celui où le noyau correspond au corpus original, le corpus final dépend de la proximité du corpus original (en fait de la partie restante après extraction du noyau) et du corpus externe utilisé. Bien évidemment, si le corpus externe est trop différent du corpus original, l'on ne pourra pas compléter correctement le noyau. Considérer des corpus externes les plus larges possibles permet ici d'augmenter les chances de trouver des documents comparables [5] L'idéal serait bien sûr d'avoir accès à la collection la plus large possible, et le web constitue ici un excellent candidat. Il est cependant nécessaire de pouvoir, à partir d'un document donné dans une langue source, extraire du web un ensemble de documents comparables en langue cible (on peut ensuite directement utiliser notre méthode sur l'union de ces ensembles). Or nous n'avons pas réussi jusqu'à présent à réaliser correctement cette extraction. La constitution entièrement automatique de collections comparables à partir du web nous semble être un problème difficile, qui requiert d'autres attributs que ceux utilisés pour les corpus parallèles. C'est un point que nous comptons développer dans le futur.

Remerciements

Ce travail qui s'inscrit dans le cadre du projet METRICC (`www.metricc.com`) a bénéficié d'une aide de l'Agence Nationale de la Recherche portant la référence ANR-08-CORD-009. Enfin, nous tenons à remercier les relecteurs pour leurs commentaires précieux.

Références

BALLESTEROS L. & CROFT W. B. (1997). Phrasal translation and query expansion techniques for cross-language information retrieval. In *Proceedings of the 20th ACM SIGIR*, p. 84–91, Philadelphia, Pennsylvania, USA.

5. C'est ce qui distingue les corpus $\mathcal{P}^1$ et $\mathcal{P}^1$ dans nos expériences, le deuxième étant obtenu à partir d'un corpus externe à plus large couverture.

DÉJEAN H., GAUSSIER E. & SADAT F. (2002). An approach based on multilingual thesauri and model combination for bilingual lexicon extraction. In *Proceedings of the 19th International Conference on Computational Linguistics*, p. 1–7, Taipei, Taiwan.

FUNG P. & MCKEOWN K. (1997). Finding terminology translations from non-parallel corpora. In *Proceedings of the 5th Annual Workshop on Very Large Corpora*, p. 192–202, Hong Kong.

FUNG P. & YEE L. Y. (1998). An IR approach for translating new words from nonparallel, comparable texts. In *Proceedings of the 17th international conference on Computational linguistics*, p. 414–420, Montreal, Quebec, Canada.

GARERA N., CALLISON-BURCH C. & YAROWSKY D. (2009). Improving translation lexicon induction from monolingual corpora via dependency contexts and part-of-speech equivalences. In *CoNLL 09 : Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, p. 129–137, Boulder, Colorado.

GAUSSIER E., RENDERS J.-M., MATVEEVA I., GOUTTE C. & DÉJEAN H. (2004). A geometric view on bilingual lexicon extraction from comparable corpora. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, p. 526–533, Barcelona, Spain.

LAROCHE A. & LANGLAIS P. (2010). Revisiting context-based projection methods for term-translation spotting in comparable corpora. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, p. 617–625, Beijing, China.

LI B. & GAUSSIER E. (2010). Improving corpus comparability for bilingual lexicon extraction from comparable corpora. In *Proceedings of the 23rd International Conference on Computational Linguistics*, p. 644–652, Beijing, China.

MORIN E., DAILLE B., TAKEUCHI K. & KAGEURA K. (2007). Bilingual terminology mining - using brain, not brawn comparable corpora. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, p. 664–671, Prague, Czech Republic.

OCH F. J. & NEY H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, **29**(1), 19–51.

PEKAR V., MITKOV R., BLAGOEV D. & MULLONI A. (2006). Finding translations for low-frequency words in comparable corpora. *Machine Translation*, **20**(4), 247–266.

PIRKOLA A. (1998). The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, p. 55–63, Melbourne, Australia.

RAPP R. (1999). Automatic identification of word translations from unrelated English and German corpora. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, p. 519–526, College Park, Maryland, USA.

ROBITAILLE X., SASAKI Y., TONOIKE M., SATO S. & UTSURO T. (2006). Compiling French-Japanese terminologies from the web. In *Proceedings of the 11st Conference of the European Chapter of the Association for Computational Linguistics*, p. 225–232, Trento, Italy.

SHEZAF D. & RAPPOPORT A. (2010). Bilingual lexicon generation using non-aligned signatures. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, p. 98–107, Uppsala, Sweden.

TALVENSAARI T., LAURIKKALA J., JÄRVELIN K., JUHOLA M. & KESKUSTALO H. (2007). Creating and exploiting a comparable corpus in cross-language information retrieval. *ACM Trans. Inf. Syst.*, **25**(1), 4.

YU K. & TSUJII J. (2009). Extracting bilingual dictionary from comparable corpora with dependency heterogeneity. In *Proceedings of HLT-NAACL 2009*, p. 121–124, Boulder, Colorado, USA.