Extraction de lexiques bilingues à partir de corpus comparables spécialisés à travers une langue pivot

Alexis Linard Emmanuel Morin Béatrice Daille
LINA UMR 6241, 2 rue de la houssinière, BP 92208, 44322 Nantes Cedex 03, France
alexis.linard@univ-nantes.fr, emmanuel.morin@univ-nantes.fr,
beatrice.daille@univ-nantes.fr

RÉSUMÉ

L'extraction de lexiques bilingues à partir de corpus comparables se réalise traditionnellement en s'appuyant sur deux langues. Des travaux précédents en extraction de lexiques bilingues à partir de corpus parallèles ont démontré que l'utilisation de plus de deux langues peut être utile pour améliorer la qualité des alignements extraits. Nos travaux montrent qu'il est possible d'utiliser la même stratégie pour des corpus comparables. Nous avons défini deux méthodes originales impliquant des langues pivots et nous les avons évaluées sur quatre langues et deux langues pivots en particulier. Nos expérimentations ont montré que lorsque l'alignement entre la langue source et la langue pivot est de bonne qualité, l'extraction du lexique en langue cible s'en trouve améliorée.

ABSTRACT

Bilingual lexicon extraction from specialized comparable corpora using a pivot language

Bilingual lexicon extraction from comparable corpora usually involves two languages. Previous studies on bilingual lexicon extraction from parallel corpora demonstrated that more than two languages can be useful to improve the alignments. Our study shows that it is possible to use the same strategy for comparable corpora. We have defined two original alignment approaches involving pivot languages and we have evaluated them over four languages and two pivot languages in particular. The experiments have shown that when a source/pivot alignment is successful, the quality of the extracted lexicon in target language can be improved.

MOTS-CLÉS : Corpus comparables, lexiques bilingues, domaine spécialisé, langue pivot.

KEYWORDS: Comparable corpora, bilingual lexicon, specialized domain, pivot language.

1 Introduction

L'extraction de lexiques bilingues à partir de corpus comparables se heurte à des difficultés majeures lorsqu'il s'agit de travailler en domaine spécialisé. La première de ces difficultés est liée à la quantité de données mobilisables en domaine spécialisé (autour d'un million de mots) qui est bien moindre que celles mobilisables en domaine de langue générale (autour de dizaines, voire de centaines de millions de mots) (Morin & Hazem, 2014). La seconde difficulté concerne la qualité des ressources disponibles lorsqu'il s'agit en particulier de s'émanciper de l'anglais. Dans cette situation, la constitution d'un corpus comparable peut poser problème et les dictionnaires bilingues nécessaires au processus d'alignement peuvent ne pas exister ou être de qualité médiocre.

Dans le cadre de l'aide à la constitution automatique de lexiques bilingues à partir de corpus comparables spécialisés, nous souhaitons étudier si le passage par une langue pivot est une alternative possible dans le cas où les ressources pour les deux langues considérées manquent. L'intervention d'une troisième langue au processus d'alignement permettrait de bénéficier d'informations lexicales supplémentaires apportées par la langue pivot. Notre hypothèse est que lorsque l'alignement bilingue initial implique des corpus en langue source ou en langue cible de qualité médiocre, ainsi qu'un dictionnaire bilingue, soit de taille réduite, soit de qualité incertaine car obtenu par triangulation, il n'est pas réaliste d'espérer obtenir des lexiques bilingues spécialisés de bonne qualité. L'ajout d'une langue pour laquelle nous aurions à disposition des ressources de qualité satisfaisante devrait améliorer la qualité la traduction de la langue source vers la langue cible. Dans cette optique, nous pensons que l'anglais est un bon candidat en tant que langue pivot, dû à la grande disponibilité de ressources gratuites et, qui plus est, de bonne qualité.

Cet article s'organise de la façon suivante : nous expliquons dans un premier temps les méthodes existantes pour l'extraction de lexiques bilingues en section 2. Nous décrivons ensuite nos méthodes expérimentales à base de corpus pivots en section 3. En section 4, nous présentons les ressources langagières utilisées dans le cadre de cet article. Enfin, nous discutons des bénéfices des méthodes présentées en section 5 et des possibles améliorations et perspectives en sections 6 et 7 [1].

2 Extraction de lexiques bilingues

Initialement conçue pour les corpus parallèles (Chen, 1993) et en raison de la rareté de ce type de ressources (Martin *et al.*, 2005), l'extraction de lexiques bilingues s'est ensuite réalisée à partir de corpus comparables plus accessibles (Fung, 1995; Rapp, 1995). Un algorithme largement répandu réalisant des alignements bilingues à partir de corpus comparables est la *méthode standard* (Fung & McKeown, 1997) basée sur la notion de vecteurs de contexte. De nombreuses implémentations existent afin de la mettre en œuvre (Rapp, 1999; Chiao & Zweigenbaum, 2002; Morin *et al.*, 2007). Un vecteur de contexte $\mathbf{w}$ est, pour un mot donné w, la représentation de ses contextes $ct_1 \dots ct_n$ où à chaque contexte est associé le nombre de cooccurrences $a(ct_i)$ identifiées dans le corpus entier. Dans cette approche, les vecteurs de contexte sont calculés à la fois dans les corpus en langue source et cible. Ils sont également normalisés en utilisant des mesures d'association. Ensuite, à l'aide d'un dictionnaire bilingue, le vecteur de contexte d'un mot à traduire est transféré de la langue source vers la langue cible en traduisant chacun de ses éléments. La similarité entre le vecteur de contexte transféré $\overline{\mathbf{w}}$ et tous les vecteurs de contexte en langue cible $\mathbf{t}$ aboutit à la création d'une liste de traductions candidates classées par ordre de similarité. Le rang de chaque traduction candidate est déterminé en fonction de la similarité entre les vecteurs de contexte. Ainsi, plus deux vecteurs sont similaires, meilleure est classifiée la traduction correspondante.

Les travaux actuels dans ce domaine se penchent sur l'amélioration de la qualité du lexique bilingue extrait. Nous pouvons par exemple citer l'utilisation d'un thesaurus bilingue (Déjean *et al.*, 2002), l'implication dans le processus de méthodes prédictives pour le comptage des cooccurrences (Hazem & Morin, 2013) ou encore l'utilisation de corpus déséquilibrés (Morin & Hazem, 2014). Pour pallier le manque de données mobilisables, Hazem & Morin (2013) proposent de mettre en œuvre des techniques de ré-estimation de cooccurrences, tandis que Morin & Hazem (2014) militent pour

1. Ce travail est une version complémentaire d'un article court publié par Linard *et al.* (2015) qui se focalisait sur l'exploitation d'un dictionnaire pivot là où cet article propose en sus l'utilisation d'un corpus pivot.

l'exploitation de corpus comparables déséquilibrés (c'est-à-dire un corpus ne comportant pas la même quantité de données entre les deux langues) de manière à tirer parti de la richesse de la langue la mieux dotée du corpus. De plus, l'utilisation d'une langue pivot a été expérimentée en traduction automatique pour l'amélioration de l'alignement exploitant des corpus parallèles (Dagan & Itai, 1991; Simard, 1999; Och & Ney, 2001; Kwon *et al.*, 2013; Seo *et al.*, 2014; Kim *et al.*, 2015). Nous pouvons aussi mentionner nos premiers travaux sur l'alignement de terminologies à partir de corpus comparables exploitant un dictionnaire pivot (Linard *et al.*, 2015). Cependant, aucune recherche, à notre connaissance, n'a été menée pour l'utilisation du corpus pivot d'une troisième langue, notamment l'utilisation des vecteurs de contexte de la langue pivot. C'est l'objet de cette étude.

3 Alignement avec une troisième langue

Nous présentons deux méthodes originales dérivant de la méthode standard (Fung & McKeown, 1997) et impliquant une troisième langue. Nous présumons que le dictionnaire bilingue nécessaire est indisponible, inexistant ou connu comme étant de piètre qualité. Nous supposons également que les dictionnaires source/pivot et pivot/cible sont de bonne qualité, car impliquant une langue pivot riche en ressources, et nous parions sur la qualité des ressources du troisième corpus. Nous espérons ainsi tirer parti de ces ressources pour améliorer les résultats finaux. Nous rappelons que l'extraction bilingue à partir de corpus comparables fournit en sortie une liste ordonnée de traductions candidates qui peut être très longue. La principale difficulté à résoudre avec l'ajout d'une troisième langue portera sur l'exploitation de ces traductions candidates.

3.1 Traduction successive

Une idée naïve est de réutiliser la méthode standard et de l'appliquer, dans un premier temps, pour effectuer une traduction de la langue source vers la langue pivot, puis dans un second temps depuis la langue pivot vers la langue cible. L'avantage principal de cette méthode effectuant une traduction successive (P_1) est l'utilisation de la terminologie de la langue pivot pour réaliser une méthode standard entre la langue pivot et la langue cible. Ainsi, pour chaque traduction obtenue en langue pivot à partir de l'alignement source/pivot, les vecteurs de contexte des traductions candidates obtenues en langue pivot sont extraits à partir de la terminologie de la langue pivot. Enfin, grâce à cette information obtenue en langue pivot et au dictionnaire pivot/cible, la méthode standard est appliquée pour toutes les traductions candidates en langue pivot.

Au vu de la dépendance des résultats finaux vis-à-vis des résultats intermédiaires obtenus du côté source et du côté pivot, cette méthode peut véhiculer beaucoup de bruit puisque les mauvais résultats obtenus en langue pivot sont répercutés en langue cible. En effet, nous pensons que trouver les traductions via les traductions obtenues par la méthode standard à travers la langue pivot tout en espérant améliorer les résultats n'est pas réaliste. Un moyen de résoudre ce problème serait de prendre en compte, pour chaque traduction candidate obtenue en langue cible, le score de similarité de la traduction pivot correspondante. Ainsi, le bruit véhiculé par les traductions erronées obtenues lors du premier alignement dans la langue pivot serait amenuisé. Nous proposons une telle pondération des traductions candidates finales en fonction des traductions obtenues en langue pivot en section 5.3.

3.2 Calcul de la similarité sur le corpus pivot

Cette méthode expérimentale (P_2) est à mi-chemin entre notre méthode et celle proposée par Linard *et al.* (2015) transposant à la fois les vecteurs de contexte des langues source et cible vers la langue pivot. En effet, nous suggérons une stratégie où nous pourrions à la fois transposer la terminologie cible vers la langue pivot, non seulement en utilisant des dictionnaires bilingues langue source/langue pivot et langue pivot/langue cible, mais aussi en bénéficiant de l'information contenue et apportée par le corpus pivot. Par conséquent, la méthode standard est appliquée dans un premier temps de la langue source à la langue pivot. À ce moment, les traductions candidates en langue pivot sont récupérées. Ensuite, pour chaque traduction en langue pivot obtenue, les vecteurs de contexte correspondants sont obtenus depuis le corpus en langue pivot. Enfin, la similarité entre les vecteurs de contexte obtenus en langue pivot et tous les vecteurs de contexte de la langue cible transférés en langue pivot est calculée.

4 Ressources multilingues

Dans cette étude, nous réalisons l'extraction de traductions candidates à partir de toutes les paires de langues à partir/vers l'anglais, le français, l'allemand et l'espagnol et impliquant l'anglais ou le français comme langue pivot. L'utilisation de l'anglais comme langue pivot est motivée par le fait que celle-ci est la langue *par défaut* pour laquelle un très grand nombre de ressources sont disponibles. D'autre part, l'utilisation du français aussi comme langue pivot est motivée par la qualité des ressources (corpus et dictionnaires) à notre disposition.

4.1 Corpus comparables

Pour nos expériences, nous nous appuyons sur deux corpus comparables constitués dans le cadre du projet Européen TTC [2]. Le premier corpus comparable que nous avons utilisé dans nos expériences est le corpus *Énergie Éolienne* (ÉÉ) qui a été créé à partir d'une aspiration de pages web en utilisant des mots-clés en relation avec le domaine des énergies éoliennes et renouvelables. Ce corpus comparable est composé de documents en 7 langues parmi lesquelles figurent l'allemand (DE), l'anglais (EN), l'espagnol (ES) et le français (FR). Le second corpus comparable que nous avons utilisé est le corpus *Technologies Mobiles* (TM) qui a aussi été construit à partir d'une aspiration de pages web. Les deux corpus mentionnés sont composés de 300 k à 470 k mots dans chaque langue.

Langue	Corpus TM	Corpus ÉÉ
FR	438 k	315 k
EN	304 k	314 k
DE	474 k	359 k
ES	475 k	454 k

TABLE 1 – Taille des différents corpus utilisés

Nous avons aussi construit une version quantitativement déséquilibrée du corpus *Énergie Éolienne* pour l'anglais et le français, en comparaison avec les autres corpus, à partir de pages collectées

2. http://www.ttc-project.eu/index.php/releases-publications

sur le web. Cette version déséquilibrée est plus forte en termes de quantité de ressources (corpus plus volumineux) et cela est très utile pour l'utilisation d'un corpus pivot. En effet, nous voulons déterminer si une langue pivot riche en ressources (à la fois corpus plus volumineux et dictionnaire bilingue qualitativement meilleur) peut améliorer *in fine* la qualité des lexiques bilingues extraits en langue cible. La taille du corpus *Énergie Éolienne* est d'environ 700 k et 800 k mots pour le français et l'anglais respectivement.

4.2 Dictionnaires bilingues

Afin d'accomplir l'extraction de lexiques bilingues à partir de corpus comparables, un dictionnaire bilingue est requis. Cependant, nous n'avons à notre disposition que les dictionnaires EURADIC français/anglais, français/espagnol et français/allemand du catalogue ELRA [3]. Ces dictionnaires sont généralistes et contiennent peu ou pas de termes liés aux domaines de l'énergie éolienne et des technologies mobiles. Pour obtenir les dictionnaires manquants, les dictionnaires français/anglais, français/espagnol et français/allemand ont été inversés pour disposer des combinaisons anglais/français, espagnol/français et allemand/français. Pour les combinaisons restantes, elles ont été obtenues par simple triangulation à partir de celles existantes (cf. table 2). En conséquence, ces dictionnaires seront certainement de moins bonne qualité.

EN-DE DE-EN	EN-ES ES-EN	EN-FR FR-EN	FR-ES ES-FR	FR-DE DE-FR	DE-ES ES-DE
600 k	26 k	240 k	100 k	170 k	15 k

TABLE 2 – Nombre d'entrées de chaque dictionnaire

4.3 Listes de référence terminologiques

Afin d'évaluer la sortie des différentes approches, des listes de référence terminologiques sont nécessaires pour chaque corpus et pour chaque langue. Nous nous appuyons à nouveau sur les ressources fournies par le projet TTC. Selon le corpus et la langue en question, les listes sont composées d'entre 48 et 88 termes simples (cf. table 3).

	EN	FR	ES	DE
ÉÉ	48	58	55	55
TM	52	58	60	88

TABLE 3 – Nombre de termes simples des listes de référence

3. http://catalog.elra.info/

5 Expérimentations et résultats

Dans cette section, nous présentons les paramètres expérimentaux mis en place, ainsi que les résultats de nos deux méthodes utilisant un corpus pivot, c'est-à-dire la méthode de traduction successive (P_1) et la méthode effectuant les calculs de similarité sur le corpus pivot (P_2). Nous discutons également les attentes concernant les améliorations apportées par nos techniques, tout particulièrement concernant le nombre de traductions candidates à prendre en compte en langue pivot, ainsi que des différentes pondérations possibles à leur associer.

5.1 Paramètres

Les documents français, anglais, espagnols et allemands ont été prétraités en utilisant l'outil TermSuite (Rocheteau & Daille, 2011)[4]. Les opérations réalisées durant la phase de prétraitements sont les suivantes : tokenisation, étiquetage morpho-syntaxique et lemmatisation. De plus, les mots outils et les hapax ont été filtrés.

Afin de calculer et de normaliser les vecteurs de contexte, la valeur $a(ct_i)$ associée à chaque cooccurrence ct_i d'un mot donné w dans le corpus a été calculée. Pour cela, le *maximum de vraisemblance* (Dunning, 1993), le *rapport des cotes actualisées* (Evert, 2005) et l'*information mutuelle* (Fano, 1961) sont les possibilités les plus communes. Parmi elles, nous avons choisi le maximum de vraisemblance dû à sa bonne représentativité (Bordag, 2008). Les vecteurs de contextes sont ensuite calculés par TermSuite, dont l'un de ses composants est spécialement dédié à cette tâche.

La similarité peut être calculée par le *cosinus* (Salton & Lesk, 1968) ou la distance de *Jaccard pondérée* (Grefenstette, 1994). Nous avons décidé de ne présenter que les résultats obtenus par la mesure du cosinus, les différences en termes de MRR (Mean Reciprocal Rank) (Voorhees, 1999) dégagées par l'une ou l'autre distance étant négligeables. Le *cosinus* est défini par l'équation suivante ($\overline{\mathbf{w}}$ représente le vecteur de contexte transféré, et $\mathbf{t}$ l'ensemble des vecteurs de contexte en langue cible) :

$$\text{Cosinus}\ (\overline{\mathbf{w}}, \mathbf{t}) = \frac{\sum_k a(\overline{\mathbf{w}}_k) a(\mathbf{t_k})}{\sqrt{\sum_k a(\overline{\mathbf{w}}_k)^2} \sqrt{\sum_k a(\mathbf{t_k})^2}}$$

Afin d'évaluer nos approches, nous avons utilisé le MRR qui a l'avantage de prendre en compte le rang des traductions candidates. Il est défini comme suit (t réfère aux termes à évaluer et r_t au rang obtenu de la traduction correcte de t) :

$$MRR = \tfrac{1}{|t|} \times \sum_{k=1}^{|t|} (\frac{1}{r_{t_k}})$$

5.2 Limites sur le nombre de traductions candidates en langue pivot

L'utilisation des traductions pivots obtenues dans le cas des méthodes à base de corpus pivots demande de décider sur le nombre de traductions candidates en langue pivot à retenir. En fonction du calcul

4. termsuite.github.io

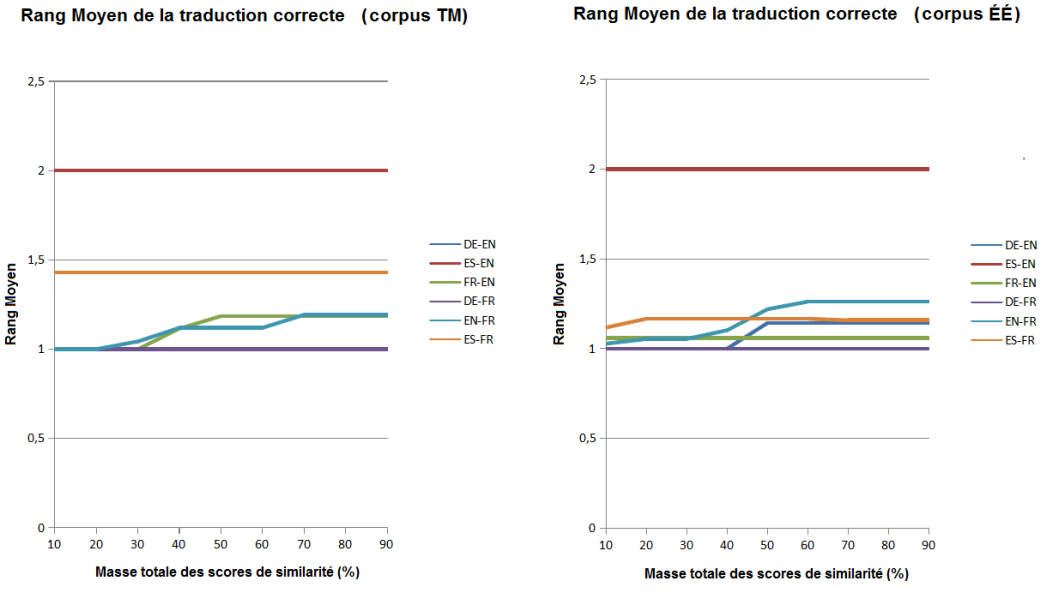

FIGURE 1 – Rang moyen de la traduction correcte en fonction de la masse totale des scores de similarité

de similarité, le nombre de traductions candidates est important en raison du bruit apporté par les mauvaises traductions. Dans le cas où les traductions candidates en langue cible ne sont pas pondérées par les traductions candidates en langue pivot correspondantes, plus il y a de traductions en langue pivot, plus nous nous attendons à une diminution des résultats finaux. Dans le cas contraire, si les traductions cibles sont pondérées par les traductions pivot, les résultats peuvent être améliorés. Bien que de mauvaises traductions en langue pivot soient toujours présentes, le poids des traductions cibles correspondantes ne sera pas trop important, si bien que le bruit apporté conduira à de moindres effets de bord.

Pour avoir une idée du bon nombre de traductions candidates en langue pivot à accepter, nous avons calculé le rang moyen de la bonne traduction obtenue par la méthode standard dans le cadre de toutes les langues sources vers une traduction en français ou en anglais. Afin de calculer cette métrique, nous avons fait varier la masse de la totalité des scores de similarité entre 10 % et 90 %. Cela signifie que sont prises en compte dans le calcul et la présentation des traductions en langue cible, toutes les n premières traductions candidates totalisant $X\%$ du total des scores de similarité de l'ensemble des traductions candidates, chaque traduction candidate étant associée à un score de similarité, et X étant le paramètre à faire varier. En outre, nous avons également utilisé les listes de référence terminologiques à notre disposition.

À partir des graphiques de la figure 1, nous observons que le rang moyen de la bonne traduction en langue pivot est situé entre 1 et 2. Néanmoins, nous ne pouvons pas prendre en compte plus de 2 traductions candidates en langue pivot pour la suite du processus, car plus il y a de traductions candidates pivots impliquées, plus il y a d'ajout de contextes lexicaux non discriminants, ce qui consiste à une introduction de bruit dans le processus. Dans ce cas, nous ne nous attendons pas à améliorer significativement les résultats.

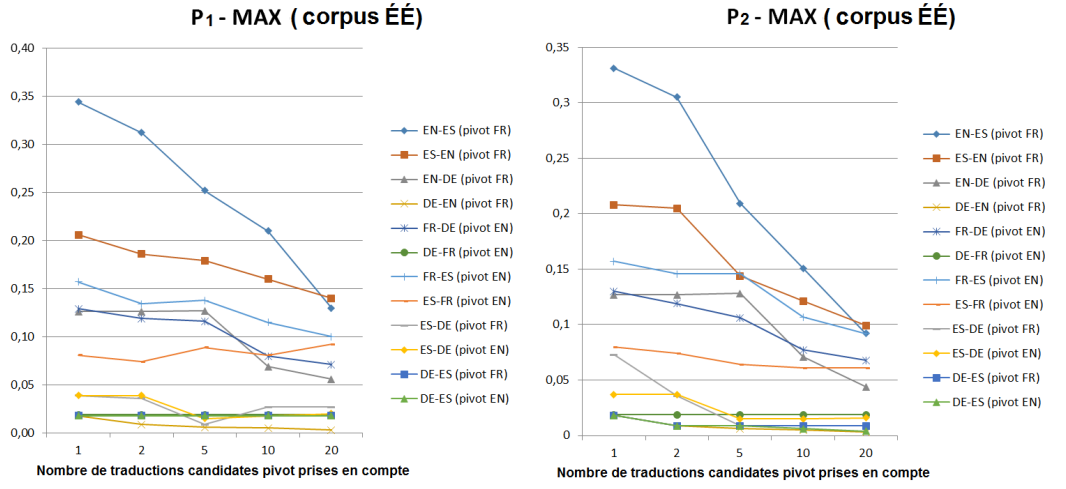

FIGURE 2 – MRR pour P_1 et P_2 en fonction du nombre de traductions candidates (méthode MAX)

5.3 Pondération des traductions candidates en langue cible selon les traductions pivot

Dans la description des deux méthodes à base de corpus pivots décrites ci-dessus (traduction successive et calcul de similarité sur corpus pivot), nous n'avons pas mentionné comment les traductions candidates données en langue cible étaient mises en relation et pondérées en fonction du mot en langue pivot correspondant. Lors de nos premières tentatives, nous aboutissions à des traductions associées à un score de similarité donné en langue cible. Cependant, nous ne faisions pas de rapport avec les traductions pivot.

Voici un court exemple illustrant le problème : étant donné w_s un mot en langue source à traduire, nous obtenons 3 traductions en langue pivot w_{p_1}, w_{p_2} et w_{p_3} avec des scores de similarité de $0,8$, $0,15$ et $0,01$ respectivement. Pour chaque mot en langue pivot, nous obtenons les traductions en langue cible suivantes, avec pour scores de similarité : $w_{t_{11}}(0,2)$, $w_{t_{12}}(0,1)$ et $w_{t_{13}}(0,05)$ pour w_{p_1} ; $w_{t_{21}}(0,7)$, $w_{t_{22}}(0,4)$ et $w_{t_{23}}(0,003)$ pour w_{p_2} ; $w_{t_{31}}(0,15)$, $w_{t_{32}}(0,002)$ et $w_{t_{33}}(0,001)$ pour w_{p_3}. Au final, les traductions finales triées sont : $T = \{w_{t_{21}}, w_{t_{22}}, w_{t_{11}}, w_{t_{31}}, w_{t_{22}} \ldots\}$. Le problème est que le poids des candidats en langue pivot n'est pas pris en compte. En d'autres termes, même si une traduction en langue pivot est la bonne, et a été triée et déterminée avec une bonne espérance, cela n'est pas du tout pris en compte du côté pivot/cible du processus. Ce phénomène est illustré en figure 2, où une prise en compte de trop de traductions candidates en langue pivot entraîne une baisse significative des résultats (méthode MAX, sans pondération).

Afin d'éviter ce phénomène, nous avons décidé de pondérer le score de similarité des traductions en langue cible en fonction du mot en langue pivot associé, grâce aux mesures suivantes :

$$PROD = sim(p) \times sim(t)$$

$$MEAN = \frac{sim(p)+sim(t)}{2}$$

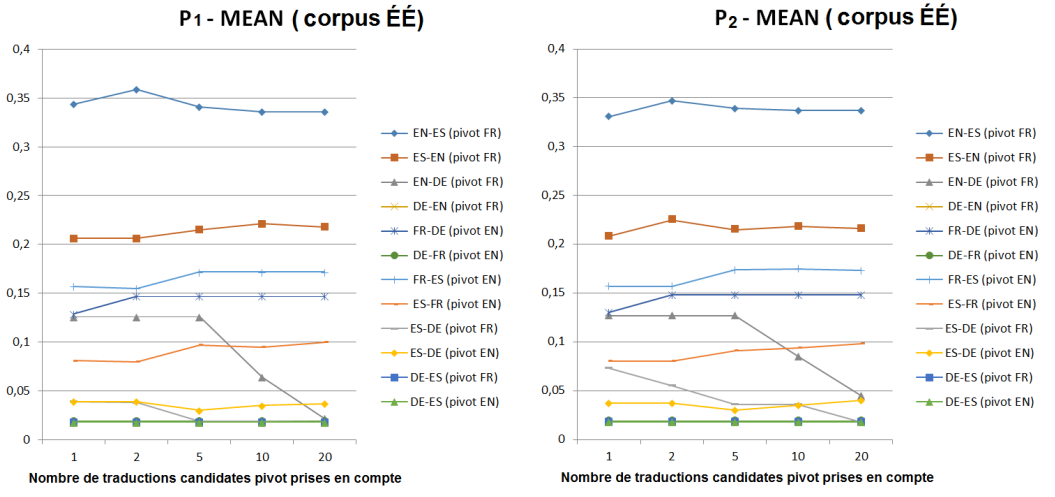

FIGURE 3 – MRR pour P_1 et P_2 en fonction du nombre de traductions candidates (méthode $MEAN$)

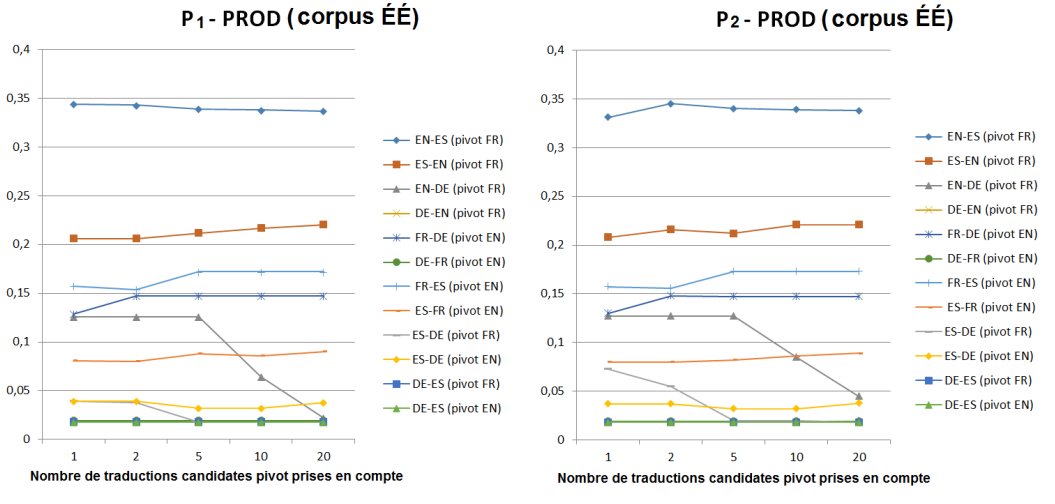

FIGURE 4 – MRR pour P_1 et P_2 en fonction du nombre de traductions candidates (méthode $PROD$).

où $sim(p)$ représente le score de similarité du mot pivot associé, $sim(t)$ le score de similarité de la traduction cible avant l'opération, $PROD$ ou $MEAN$ pour le score final fixé par la traduction selon le calcul de la moyenne ou du produit des scores de similarité. Les résultats obtenus en fonction du nombre de traductions en langue pivot est présenté dans les figures 3 et 4. Nous voyons bien que pour certaines paires de langues, un optimum se dégage, alors que pour d'autres les résultats se dégradent à partir de ce même seuil, ou bien entamment une amélioration à partir de ce seuil.

5.4 Attentes à propos de la qualité des traductions en langue pivot

Le MRR obtenu pour les traductions candidates en langue pivot est un bon indicateur sur comment nous pourrions améliorer l'alignement original source/cible. En effet, si les traductions obtenues en langue pivot sont incorrectes pour le mot source correspondant, il y a peu d'espoir d'obtenir la traduction finale en langue cible.

D'après les résultats de la table 4, nous sommes confiants quant à l'amélioration des résultats de la méthode standard pour l'anglais en langue source vers n'importe quelle langue cible via le français, le français en langue source vers n'importe quelle langue cible via l'anglais, ainsi que l'espagnol vers une langue cible en passant par le français. Enfin, nous n'espérons pas améliorer les résultats dans les autres cas (allemand en langue source, espagnol en langue source et anglais en langue pivot).

Langues	EN-FR	ES-FR	DE-FR	FR-EN	ES-EN	DE-EN
corpus ÉÉ	0,585	0,210	0,038	0,527	0,119	0,018
corpus TM	0,685	0,238	0,034	0,585	0,193	0,074

TABLE 4 – MRR des traductions candidates en langue pivot

5.5 Résultats

Langues	Pivot	Std.	Linard *et al.* (2015)	P_{1E}	P_{1D}	P_{2E}	P_{2D}	R_{MAX}	C
EN-ES	FR	0,268	0,390	**0,339**	**0,441**	**0,340**	**0,456**	*0,646*	65.76%
ES-EN	FR	0,119	0,233	**0,212**	**0,270**	**0,212**	**0,270**	*0,491*	
EN-DE	FR	0,158	0,215	0,126	0,042	0,127	0,068	*0,458*	66.21%
DE-EN	FR	0,018	0,018	0,018	0,018	0,018	0,018	*0,200*	
FR-DE	EN	0,056	0,132	**0,147**	**0,148**	**0,147**	**0,147**	*0,418*	77.63%
DE-FR	EN	0,038	0,028	0,019	0,019	0,019	0,019	*0,151*	
FR-ES	EN	0,366	0,176	0,172	0,160	0,173	0,163	*0,528*	82.36%
ES-FR	EN	0,210	0,117	0,088	0,089	0,082	0,088	*0,357*	
ES-DE	FR	0,000	0,097	**0,018**	**0,003**	**0,020**	**0,018**	*0,273*	44.24%
	EN	0,000	0,045	**0,032**	**0,030**	**0,032**	**0,029**		
DE-ES	FR	0,001	0,018	**0,018**	**0,018**	**0,018**	**0,018**	*0,218*	
	EN	0,001	0,018	**0,018**	**0,018**	**0,018**	**0,018**		

TABLE 5 – MRR obtenu pour les approches à base de corpus pivots (corpus ÉÉ)

Le MRR obtenu pour les deux approches est présenté dans les tables 5 et 6 pour les corpus Énergie Éolienne et Technologies Mobiles respectivement. Les paramètres retenus sont les suivants : l'utilisation du maximum de vraisemblance pour le calcul et la normalisation des vecteurs de contexte, le cosinus pour le calcul de similarité entre vecteurs de contexte, la limitation à 5 traductions candidates en langue pivot et la pondération des traductions candidates en langue cible avec la méthode $PROD$ (produit du score de similarité d'une traduction candidate en langue cible avec celui de la traduction pivot correspondante). Nous présentons également en guise de comparaison, les résultats obtenus par la méthode standard (Std.), la méthode effectuant une traduction successive en utilisant un corpus équilibré (P_{1E}) ou déséquilibré (corpus pivot plus volumineux, P_{1D}) et la méthode ramenant le calcul de similarité à la langue pivot (P_{2E} et P_{2D}). Nous donnons aussi une information additionnelle, comme le meilleur résultat atteignable en raison des listes de référence terminologiques et des mots ayant été filtrés lors des opérations de prétraitements (R_{MAX}), et la comparabilité des corpus C (Li & Gaussier, 2010). Nous rappelons également les résultats obtenus avec la meilleure des méthodes exploitant un seul dictionnaire pivot (Linard *et al.*, 2015).

La comparabilité d'un corpus consiste en l'espérance de trouver la traduction en langue cible de chaque mot en langue source du corpus. En outre, c'est un bon moyen de mesurer la symétrie distributionnelle entre deux corpus étant donné un dictionnaire bilingue. Nous pouvons aussi remarquer que le Rappel Maximal R_{MAX} est plutôt bas pour certaines paires de langues, cela étant dû au nombre élevé d'hapax appartenant aux listes de référence terminologiques ayant été filtrés durant la phase de prétraitements.

D'après les résultats obtenus, nous remarquons qu'il existe une forte corrélation entre les améliorations obtenues et la comparabilité des corpus. Nous avons amélioré la qualité du lexique bilingue extrait seulement dans le cas de corpus peu comparables. C'est le cas des paires ES-DE et EN-ES (et vice versa), qui, dans les cas des deux corpus, sont les paires ayant les scores de comparabilité les plus faibles. Nous ajouterons que l'utilisation d'un corpus déséquilibré peut significativement améliorer les résultats de base de nos approches. Enfin, les résultats de nos méthodes impliquant un corpus pivot surpassent dans la majorité des cas ceux des méthodes employant un dictionaire pivot (Linard *et al.*, 2015).

Langues	Pivot	Std.	Linard *et al.* (2015)	P_{1E}	P_{2E}	R_{MAX}	C
EN-ES	FR	0,445	0,523	**0,528**	**0,530**	*0,882*	66.52%
ES-EN	FR	0,193	0,321	**0,239**	**0,239**	*0,533*	
EN-DE	FR	0,622	0,570	0,210	0,191	*0,896*	68.95%
DE-EN	FR	0,074	0,070	0,046	0,045	*0,455*	
FR-DE	EN	0,053	0,063	**0,088**	**0,088**	*0,597*	80,06%
DE-FR	EN	0,034	0,026	0,034	0,034	*0,432*	
FR-ES	EN	0,514	0,280	0,319	0,322	*0,807*	82.02%
ES-FR	EN	0,238	0,207	0,166	0,165	*0,552*	
ES-DE	FR	0,001	0,067	**0,050**	**0,050**	*0,500*	44.02%
	EN	0,001	0,035	**0,042**	**0,041**		
DE-ES	FR	0,126	0,355	**0,376**	**0,376**	*0,585*	
	EN	0,126	0,179	**0,180**	**0,180**		

TABLE 6 – MRR obtenu pour les approches à base de corpus pivots (corpus TM)

6 Discussion

Les résultats présentés dans les tables 5 et 6 et en figures 2 à 4 ont montré que l'utilisation d'un corpus pivot comme langue intermédiaire peut améliorer la qualité du lexique bilingue final extrait. Premièrement, la façon de prendre en compte le poids des traductions en langue pivot à d'importantes répercussions sur les résultats finaux. En calculant la moyenne ou le produit des scores de similarité des traductions pivot et ceux des traductions cibles correspondantes, nous avons significativement réussi à mettre de côté les effets de bruit véhiculés par les traductions pivot incorrectes. Le nombre de traductions candidates en langue pivot à prendre en compte est aussi un paramètre important : dans le cas de la pondération des traductions candidates en langue cible par les candidats pivot correspondants, nous avons mis en exergue un seuil optimal de 5 traductions pivot candidates. Ce seuil est en effet le plus adapté étant donné un début de dégradation au-delà de ce seuil dans certains cas, et un début d'amélioration déjà amorcé en deçà de ce seuil dans d'autres cas.

De plus, comme attendu en section 5.4, nous avons réussi à améliorer les résultats dans le cas de l'anglais vers l'espagnol (et inversement) et du français vers l'allemand. Finalement, le bien-fondé de l'utilisation d'un corpus pivot déséquilibré a été démontré, en particulier dans le cas de l'anglais vers l'espagnol et vice versa, où la qualité du lexique bilingue extrait a été nettement améliorée.

Nous pouvons aussi mentionner d'autres expérimentations menées – mais non présentées dans cet article – avec un corpus pivot :

— Calcul du vecteur de contexte moyen de tous les vecteurs de contextes des traductions candidates obtenues en langue pivot.
— Application du calcul inverse des traductions obtenues en langue pivot, afin de vérifier que les traductions candidates en langue pivot sont correctes.
— Recalcul des vecteurs de contexte source et cible en appliquant un modèle de régression linéaire basé sur des corpus déséquilibrés (à l'instar des travaux menés par Hazem & Morin (2013)). Les paramètres de la régression linéaire sont appris à partir d'un corpus généraliste d'à peu près 10 millions de mots.

Néanmoins, toutes ces idées se sont avérées infructueuses, dans le sens où elles aboutissent à des dégradations ou des améliorations peu ou pas significatives des résultats.

7 Conclusion

Nous avons fait l'hypothèse qu'une troisième langue pouvait améliorer les résultats de l'alignement de terminologies à partir de corpus comparables lorsque les ressources n'étaient pas de bonne qualité. Nous avons présenté une méthode utilisant un corpus pivot et montré que le problème de la dépendance envers la qualité des ressources peut être surmontée en passant *via* une troisième langue techniquement plus riche, telle l'anglais par exemple. En outre, il ressort également que les améliorations sont plus importantes lorsqu'un corpus pivot déséquilibré – c'est-à-dire plus volumineux en termes de ressources – est utilisé. Enfin, nous avons remarqué que selon les paires de langues, le nombre de traductions candidates obtenues en langue pivot doit être limité pour optimiser les améliorations.

Dans nos travaux futurs, nous essaierons d'adapter le corpus et le dictionnaire pivots à l'alignement de termes complexes à partir de corpus comparables. Nous pensons en effet que les méthodes actuelles pour l'alignement de termes complexes de deux mots ou plus, les méthodes compositionnelles et semi-distributionnelles, sont adaptables à l'utilisation de ressources provenant d'une troisième langue.

Remerciements

Ce travail a bénéficié d'une aide de l'Agence Nationale de la Recherche portant la référence (ANR-12-CORD-0029).

Références

BORDAG S. (2008). A comparison of co-occurrence and similarity measures as simulations of context. In *Proceedings of the 9th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing'08)*, p. 52–63. Haifa, Israel.

CHEN S. F. (1993). Aligning Sentences in Bilingual Corpora Using Lexical Information. In *Proceedings of the 31st Annual Meeting on Association for Computational Linguistics (ACL'93)*, p. 9–16, Columbus, OH, USA.

CHIAO Y.-C. & ZWEIGENBAUM P. (2002). Looking for candidate translational equivalents in specialized, comparable corpora. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING'02)*, p. 1208–1212, Tapei, Taiwan.

DAGAN I. & ITAI A. (1991). Two languages are more informative than one. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics (ACL'91)*, p. 130–137, Berkeley, CA, USA.

DÉJEAN H., GAUSSIER É. & 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 (COLING'02)*, p. 1–7, Taipei, Taiwan.

DUNNING T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, **19**(1), 61–74.

EVERT S. (2005). *The statistics of word cooccurrences : word pairs and collocations*. PhD thesis, University of Stuttgart.

FANO R. M. (1961). *Transmission of Information : A Statistical Theory of Communications*. Cambridge, MA, USA : MIT Press.

FUNG P. (1995). Compiling Bilingual Lexicon Entries From a non-Parallel English-Chinese Corpus. In *Proceedings of the 3rd Annual Workshop on Very Large Corpora (VLC'95)*, p. 173–183, Cambridge, MA, USA.

FUNG P. & MCKEOWN K. (1997). Finding Terminology Translations from Non-parallel Corpora. In *Proceedings of the 5th Annual Workshop on Very Large Corpora (VLC'97)*, p. 192–202, Hong Kong.

GREFENSTETTE G. (1994). *Explorations in Automatic Thesaurus Discovery*. Boston, MA, USA : Kluwer Academic Publisher.

HAZEM A. & MORIN E. (2013). Word Co-occurrence Counts Prediction for Bilingual Terminology Extraction from Comparable Corpora. In *Proceedings of the 6th International Joint Conference on Natural Language Processing (IJCNLP'13)*, p. 1392–1400, Nagoya, Japan.

KIM J.-H., KWON H.-S. & SEO H.-W. (2015). Evaluating a pivot-based approach for bilingual lexicon extraction. *Computational Intelligence and Neuroscience*, **2015**.

KWON H.-S., SEO H.-W. & KIM J.-H. (2013). Bilingual lexicon extraction via pivot language and word alignment tool. In *Proceedings of the 6th Workshop on Building and Using Comparable Corpora (BUCC'13)*, p. 11–15, Sofia, Bulgaria.

LI B. & GAUSSIER É. (2010). Improving corpus comparability for bilingual lexicon extraction from comparable corpora. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING'10)*, p. 644–652, Beijing, China.

LINARD A., DAILLE B. & EMMANUEL M. (2015). Attempting to Bypass Alignment from Comparable Corpora via Pivot Language. In *Proceedings of the Eighth Workshop on Building and Using Comparable Corpora*, p. 32–37, Beijing, China.

MARTIN J., MIHALCCA R. & PEDERSEN T. (2005). Word alignment for languages with scarce resources. In *Proceedings of the ACL Workshop on Building and Using Parallel Text*, p. 65–74, Ann Arbor, MI, USA.

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 (ACL'07)*, p. 664–671, Prague, Czech Republic.

MORIN E. & HAZEM A. (2014). Looking at Unbalanced Specialized Comparable Corpora for Bilingual Lexicon Extraction. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL'14)*, p. 1284–1293, Baltimore, Maryland.

OCH F. J. & NEY H. (2001). Statistical multi-source translation. In *Proceedings of the 8th Conference on Machine Translation Summit (MT Summit VIII)*, p. 253–258, Santiago de Compostela, Spain.

RAPP R. (1995). Identify Word Translations in Non-Parallel Texts. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL'95)*, p. 320–322, Boston, MA, USA.

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 (ACL'99)*, p. 519–526, College Park, MD, USA.

ROCHETEAU J. & DAILLE B. (2011). TTC TermSuite : A UIMA Application for Multilingual Terminology Extraction from Comparable Corpora. In *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP'11)*, p. 9–12, Chiang Mai, Thailand.

SALTON G. & LESK M. E. (1968). Computer evaluation of indexing and text processing. *Journal of the Association for Computational Machinery*, **15**(1), 8–36.

SEO H.-W., KWON H.-S. & KIM J.-H. (2014). Extended pivot-based approach for bilingual lexicon extraction. *Journal of the Korean Society of Marine Engineering*, **38**(5), 557–565.

SIMARD M. (1999). Text-Translation Alignment : Three Languages Are Better Than Two. In *Proceedings of Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC'99)*, p. 2–11, College Park, ML, USA.

VOORHEES E. M. (1999). The TREC-8 Question Answering Track Report. In *Proceedings of the 8th Text REtrieval Conference (TREC-8)*, volume 99, p. 77–82, Gaithersburg, MA, USA.