Acquisition de lexique bilingue d'expressions polylexicales: une application à la traduction automatique statistique

Dhouha Bouamor
CEA-LIST, LVIC, F91191 Gif sur Yvette Cedex, France
LIMSI-CNRS, F-91403 Orsay, France
Univ. Paris Sud, Orsay, France
dhouha.bouamor@cea.fr

RÉSUMÉ

Cet article décrit une méthode permettant d'acquérir un lexique bilingue d'expressions polylexicales (EPLs) à partir d'un corpus parallèle français-anglais. Nous identifions dans un premier temps les EPLs dans chaque partie du corpus parallèle. Ensuite, nous proposons un algorithme d'alignement assurant la mise en correspondance bilingue d'EPLs. Pour mesurer l'apport du lexique construit, une évaluation basée sur la tâche de Traduction Automatique Statistique (TAS) est menée. Nous étudions les performances de trois stratégies dynamiques et d'une stratégie statique pour intégrer le lexique bilingue d'expressions polylexicales dans un système de TAS. Les expériences menées dans ce cadre montrent que ces unités améliorent significativement la qualité de traduction.

ABSTRACT

Mining a Bilingual Lexicon of MultiWord Expressions : A Statistical Machine Translation Evaluation Perspective

This paper describes a method aiming to construct a bilingual lexicon of MultiWord Expressions (MWEs) from a French-English parallel corpus. We first extract monolingual MWEs from each part of the parallel corpus. The second step consists in acquiring bilingual correspondences of MWEs. In order to assess the quality of the mined lexicon, a Statistical Machine Translation (SMT) task-based evaluation is conducted. We investigate the performance of three dynamic strategies and of one static strategy to integrate the mined bilingual MWEs lexicon in a SMT system. Experimental results show that such a lexicon significantly improves the quality of translation.

MOTS-CLÉS : Expression polylexicale, alignement bilingue, traduction automatique statistique.

KEYWORDS: MultiWord expression, bilingual alignment, statistical machine translation.

1 Introduction

Une expression polylexicale (EPL, en anglais *multiword expression*) peut être définie comme une combinaison de mots pour lesquels les propriétés syntaxiques ou sémantiques de l'expression entière ne peuvent pas être obtenues à partir de ses parties (Sag *et al.*, 2002). Les EPLs regroupent les expressions figées et semi-figées (ex. *cordon bleu*), les collocations (ex. *chemin de fer*), les entités nommées (ex. *New York*), les verbes à particule (ex. *grow up*), les constructions à verbe

support (ex. *faire face à*), etc.(Sag *et al.*, 2002; Constant *et al.*, 2011). Elles sont fréquemment employées dans les textes écrits étant donnée qu'elles constituent une part significative du lexique d'une langue. Jackendoff (1997) estime que la fréquence de leur utilisation est équivalente à celle des mots simples. Bien qu'elles soient facilement employées et reconnues par les humains, leur identification pose un problème majeur pour diverses applications du traitement automatique des langues.

Pour la Traduction Automatique Statistique (TAS), diverses améliorations ont été obtenues avec l'émergence des approches à base de segments (*phrase based approaches* en anglais) (Koehn *et al.*, 2003). Ces segments sont définis comme étant de simples n-grammes systématiquement traduits dans un corpus parallèle sans aucune motivation linguistique. Dans de tels systèmes, le manque d'un traitement adéquat des EPLs pourrait affecter la qualité de la traduction. En effet, la traduction littérale d'une expression non reconnue par le système de traduction comme une EPL constitue une cause principale à une traduction erronée et incompréhensible. Par exemple, un tel système proposera *« way of iron »* comme traduction pour *« chemin de fer »* au lieu de *« railway »*. Il est donc important d'utiliser un lexique dans lequel les EPLs sont prises en compte. Or un des points faibles des lexiques est souvent le manque de couverture pour ces unités (Sagot *et al.*, 2005). Ce point a été abordé dans plusieurs travaux (Fazly et Stevenson, 2007; Caseli *et al.*, 2009).

Cet article porte sur le traitement des EPLs bilingues, allant de l'acquisition automatique à partir de corpus parallèles à leur intégration dans un système de TAS. Nous considérons toute séquence contiguë non compositionnelle, appartenant à l'une des classes définies par (Luka *et al.*, 2006), comme une EPL. Ces unités ont été classées dans trois classes, sur la base de leurs propriétés catégorielles, ainsi que de leur degré de figement syntaxique et sémantique. Les classes sont constituées de *mots composés*, d'*expressions idiomatiques* et de *collocations*. Intuitivement, les EPLs bilingues sont utiles pour améliorer les résultats de la TAS. Cependant, des recherches plus approfondies sont nécessaires pour trouver la meilleure façon d'intégrer ce type d'unités dans ces systèmes. Dans cette étude, nous considérons la TAS comme un mode d'évaluation extrinsèque de l'utilité des EPLs et explorons différentes stratégies d'intégration de ces unités dans un système de TAS. Étant donné un lexique bilingue d'EPLs, nous proposons (1) trois stratégies d'intégration dynamiques où nous cherchons à modifier le modèle de traduction de différentes façons pour une prise en considération des EPLs bilingues et (2) une stratégie d'intégration statique dans laquelle nous incorporons ces unités sans changer le modèle de traduction.

Le reste de l'article est organisé comme suit : dans la section 2, nous passons en revue les principaux travaux en rapport avec la tâche d'extraction de traduction pour les EPLs. Puis, nous décrivons, dans la section 3, l'approche utilisée pour identifier ces unités et présentons, par la suite, l'algorithme d'alignement que nous avons implémenté. Dans la section 4, nous décrivons les stratégies d'intégration proposées. La section 5 est consacrée aux expériences menées ainsi qu'à la présentation des résultats obtenus. Nous concluons notre article par une présentation des principales perspectives en section 6.

2 État de l'art

Au cours des dernières années, de nombreux travaux de recherche ont été menés sur la tâche d'extraction bilingue d'EPLs à partir de corpus parallèles. La plupart d'entre eux commencent

tout d'abord par identifier les EPLs dans chaque partie du corpus parallèle, ensuite, se basent sur différentes techniques d'alignement pour les apparier. Les techniques d'extraction monolingue d'EPLs tournent autour de trois approches : (1) des méthodes symboliques reposant sur des patrons morphosyntaxiques (Okita *et al.*, 2010; Dagan et Church, 1994) ; (2) des méthodes statistiques utilisant des mesures d'association pour classer les EPLs candidates (Vintar et Fisier, 2008) et (3) des méthodes hybrides combinant (1) et (2) (Seretan et Wehrli, 2007; Daille, 2001). Aucune des approches n'est sans limitations. Il est difficile d'appliquer des méthodes symboliques à des données sans annotations morphosyntaxiques. En ce qui concerne les méthodes statistiques, bien qu'elles soient conçues pour des bigrammes, elles exigent la définition d'un seuil à partir duquel un segment extrait peut être considéré comme une EPL.

Pour identifier des correspondances entre expressions dans différentes langues, quelques travaux font appel à des outils d'alignement de mots simples pour guider l'alignement d'EPLs (Dagan et Church, 1994). D'autres se basent sur des algorithmes d'apprentissage statistique comme par exemple l'algorithme itératif de ré-estimation *Expectation Maximization* (Kupiec, 1993; Okita *et al.*, 2010). Une hypothèse largement suivie pour acquérir des EPLs bilingues est qu'une expression dans une langue source garde la même structure syntaxique que son équivalente dans une langue cible donnée (Seretan et Wehrli, 2007; Tufis et Ion, 2007). Or, les EPLs ne se traduisent pas forcément par des expressions ayant la même catégorie grammaticale (i.e *« insulaire en développement »* et *« small island developing »*) ou la même longueur[1] (i.e *« en ce qui concerne »* et *« as regards »*).

Le but principal de la majorité des travaux de recherche menés sur cet objet linguistique était l'acquisition *a priori* de correspondances entre les paires d'unités textuelles pour l'enrichissement de ressources lexicales. En comparaison, peu de travaux ont été réalisés sur l'exploitation de telles ressources, afin de rendre possible leur intégration dans des applications clés, telles que la désambiguïsation sémantique (Finlayson et Kulkarni, 2011) ou la recherche d'information interlingue (Vechtomova, 2005). En TAS, Lambert et Banchs (2005) introduisent une méthode dans laquelle les EPLs sont considérées comme un élément unique dans le corpus d'apprentissage. En exploitant un corpus de petite taille, ils ont montré que la qualité de l'alignement et la précision de traduction ont été améliorées. Cependant, ils ont obtenu, dans des études plus récentes (Lambert et Banchs, 2006), basées sur un corpus de taille importante, un score BLEU (Papineni *et al.*, 2002) plus bas. Nous citons notamment les travaux de (Ren *et al.*, 2009) qui implémentent une méthode permettant d'intégrer des termes multi-mots issus du domaine médical dans MOSES (Koehn *et al.*, 2007). Leur méthode a permis de gagner 0,17 points de score BLEU par rapport au système de référence. La présente étude est une extension de l'approche présentée dans (Bouamor *et al.*, 2011). Nous proposons tout d'abord une méthode ayant pour but d'extraire et aligner des EPLs bilingues. Nous étudions ensuite l'impact de l'utilisation de ces unités dans un système de TAS.

3 Acquisition de lexique bilingue d'EPLs

Dans cette section, nous décrivons l'approche proposée pour extraire un lexique bilingue d'EPLs à partir d'un corpus parallèle français-anglais aligné au niveau de la phrase. Cette approche est réalisée en deux étapes. Dans la première étape, nous identifions les EPLs dans chaque partie

[1] La longueur d'une EPL est calculée en nombre de mots

du corpus parallèle. La deuxième étape consiste en l'acquisition de correspondances bilingues d'EPLs.

3.1 Extraction monolingue d'EPLs

La méthode d'identification monolingue d'EPLs est fondée sur une approche symbolique, très similaire à celle présenté dans (Okita *et al.*, 2010). Là où ils définissent des patrons pour extraire seulement des syntagmes nominaux, notre approche identifie à la fois des syntagmes nominaux, des expressions figées et des entités nommées. La méthode proposée requiert simplement une analyse morphosyntaxique des textes source et cible, comme étape préliminaire à la procédure de construction d'expressions. Nous faisons donc appel à la plate forme d'analyse multilingue LIMA du CEA-LIST (Besançon *et al.*, 2010), qui produit une liste de lemmes étiquetés par leurs catégories grammaticales. Le processus d'identification d'EPLs opère sur des lemmes plutôt que sur des formes de surface.

Comme la plupart des expressions sont constituées de combinaisons de noms,d'adjectifs ou encore de prépositions, nous produisons une liste de n-grammes candidats ($2 \leq n \leq 4$), dont la structure morphosyntaxique respecte une configuration prédéfinie, telle que celles décrites dans le tableau 1. seize configurations ont été manuellement définies. Notons qu'il existe des patrons d'extraction (ou configurations) pour lesquels aucune EPL n'a été produite (c-à-d. Past_Participle-Noun). Un tel type d'analyse permet de ne garder que des n-grammes jugés pertinents et d'écarter ceux constitués de mots vides comme par exemple *« is a, of the, de la »*.

Configuration	**Exemples anglais/français**
Adj-Noun	Plenary meeting/Libre circulation
Noun-Adj	Oil tanker/Parlement européen
Noun-Noun	Member state/Etat membre
Past_Participle-Noun	Developped country/...
Noun-Past_Participle	Parliament adopted/Pays developpé
Adj-Adj-Noun	European public prosecutor/...
Adj-Noun-Adj	Social market economy/Bon conduite administratif
Adj-Noun-Noun	Renewable energy source/...
Noun-Noun-Adj	.../Industrie automobile allemand
Noun-Adj-Adj	.../Ministère public européen
Adj-Noun-Adj	Social fund assistance/Important débat politique
Noun-Prep-Noun	Point of view/Chemin de fer
Noun-Prep-Adj-Noun	Court of first instance/Court de premier instance
Noun-Prep-Noun-Adj	.../Source d'énergie renouvelable
Adj-Noun-Prep-Noun	European court of justice/...
Noun-Adj-Prep-Noun	.../Politique européen de concurrence

TABLE 1 – Configurations morphosyntaxiques permises.

A cette liste de candidats, nous ajoutons des expressions idiomatiques prépositionnelles comme par exemple *« in the light of, with regard to, en ce qui concerne... »* et des entités nommées telles que *« Middle East, South Africa, El-Salvador... »* reconnues par la plate-forme LIMA. Le résultat

ID.PHRASE	PHRASE
2	*...semblerait être* ***à nouveau*** *mis en accusation, le ministère public ...*
55	*...vous demande donc* ***à nouveau*** *de faire le nécessaire ...*
$n-1$	*...aussi de promouvoir* ***à nouveau*** *l'activité des femmes ...*
n	*...que le règlement soit* ***à nouveau*** *modifié en collaboration ...*

⇓

	1	2	3	4	.....	55	.....	$n-1$	n
à nouveau	0	1	0	0	.....	1	.....	1	1

FIGURE 1 – Représentation vectorielle de l'expression *« à nouveau »*. ID.PHRASE correspond à un identifiant unique de la phrase contenant l'expression dans notre corpus.

de l'extraction est représenté par une liste de candidats triée par ordre décroissant selon leur fréquence dans le corpus. Plusieurs candidats parmi ceux produits apparaissent dans d'autres candidats. Afin d'éviter un effet de surgénération, nous proposons les heuristiques de nettoyage suivantes :

- Si une expression est imbriquée dans une autre et qu'elles ont la même fréquence, on ne garde que la plus couvrante (plus longue).
- Si une expression apparaît dans un grand nombre d'autres expressions, nous suivons l'approche proposée par (Frantzi *et al.*, 2000) et éliminons toutes les expressions plus longues.

À la différence de beaucoup d'autres systèmes existants (Daille, 2001; Seretan et Wehrli, 2007; Vintar et Fisier, 2008), notre système n'applique pas de filtre fondé sur des mesures d'association ou sur la fréquence. Nous prenons en considération toutes les expressions extraites, aussi bien fréquentes que non fréquentes et celles dont le degré de corrélation entre ses constituants est élevé ou faible. A notre connaissance, aucune approche n'a pris en considération tout l'ensemble.

3.2 Méthode d'alignement

Dans cette section, nous présentons une méthode qui tente de trouver, pour chaque expression source, la traduction qui lui est adéquate dans l'ensemble d'expressions cibles. Cette tâche pose de sérieux problèmes en l'absence de ressources externes comme les dictionnaires bilingues ou les outils d'alignement de mots simples. Nous proposons une méthode indépendante de toute ressource externe, qui requiert simplement un corpus parallèle et la liste de candidats source et cible à traduire. Notre approche hérite de la sémantique distributionnelle, où nous associons à chaque expression source et cible une représentation spécifique qui servira par la suite de base pour l'établissement d'une relation de traduction entre chaque paire d'expressions (source, cible). Nous faisons appel au modèle vectoriel (Salton *et al.*, 1975), un modèle algébrique souvent utilisé en recherche d'information. Nous représentons chaque expression par un vecteur de dimension n (nombre de phrases dans le corpus) indiquant si elle apparaît ou non dans chaque phrase du corpus. La figure 1 décrit le vecteur représentant l'EPL française *« à nouveau »*.

Pour extraire des paires de traduction d'EPLs, nous proposons un algorithme itératif d'alignement opérant de la façon suivante :

1. Trouver l'expression la plus fréquente dans chaque phrase source.

Français	→	Anglais
parlement européen	→	european parliament
état par état	→	amount of state
coup d'état	→	military coup
zone non fumeur	→	no smoking area
insulaire en développement	→	small island developing
de bonne foi	→	good faith
politique de concurrence	→	competition policy
chemin de fer	→	railway sector
en ce qui concerne	→	in regard to
en ce qui concerne	→	as regards
en ce qui concerne	→	with reference to
en ce qui concerne	→	with respect to
coupe forestier	→	cut in forestation

TABLE 2 – Exemples d'EPLs bilingues alignées par l'algorithme décrit ci-dessus.

2. Extraire les expressions cibles qui apparaissent dans toutes les phrases parallèles à celles où figure l'expression source.
3. Calculer un score de confiance pour chaque couple (source, cible).
4. Considérer l'expression cible qui maximise ce score comme la meilleure traduction.
5. Supprimer la paire de traductions du processus et retourner vers 1.

Le score de confiance est calculé sur la base de la mesure de l'indice de Jaccard (équation 1).

$$\text{Jaccard} = \frac{I_{st}}{V_s + V_t - I_{st}} \tag{1}$$

Cette mesure est fondée principalement sur le calcul de nombre de phrases partagées par chaque expression source et cible nommé ici I_{st} qu'on normalise par la somme des normes des vecteurs V_s et V_t diminué de l'ensemble d'intersection.

En observant certaines paires d'expressions du tableau 2, nous remarquons que notre méthode présente plusieurs avantages. Premièrement, pour trouver la traduction adéquate pour chaque EPL et contrairement à la plupart des travaux antérieurs (Dagan et Church, 1994; Ren *et al.*, 2009) qui reposent sur la traduction mot à mot des composantes d'une EPL, notre méthode capture l'équivalence sémantique entre les EPLs en n'ayant recours à aucune information préalable sur l'alignement des mots. Elle permet aussi d'aligner des expressions à caractère idiomatique tel que *« à nouveau → once more »* ou encore *« état par état → amount of state »* et trouve toutes les correspondances bilingues possibles pour les EPLs source pour lesquelles plusieurs EPLs cible correctes existent. Par exemple, notre méthode fournit pour l'EPL *« en ce qui concerne »* les traductions suivantes : *« in regard to », « with reference to », « with respect to », « as regards »*.

Nous avons pu aussi identifier une classe d'erreurs dont la cause provient essentiellement du choix de la taille des n-grammes. Comme nous ne prenons en considération que des alignements *m-n* avec m≥2 and n≥2, quelques expressions dont la traduction de référence est constituée

d'un seul mot ne sont pas alignées correctement. Par exemple, l'expression française *« chemin de fer »* correspondant normalement au mot simple anglais *« railway »* est traduite par l'expression *« railway sector »*.

4 EPLs dans Moses

Dans la section précédente, nous avons décrit l'approche suivie pour acquérir un lexique bilingue d'EPLs. Pour évaluer sa qualité, nous avons mené dans (Bouamor *et al.*, 2011) une évaluation intrinsèque à petite échelle dans laquelle nous comparons les paires d'EPLs bilingues acquises à un alignement de référence créé manuellement. Sur un corpus de test constitué de 100 paires de phrases parallèles issues du corpus Europarl, nous avons obtenu une précision de 63,93%, un rappel de 62,46% et une F-mesure de 63,19%. Comme il n'existe à ce jour aucun protocole commun permettant d'évaluer les résultats d'alignement d'EPLs, nous conduisons une évaluation extrinsèque dans laquelle nous étudions l'impact de l'utilisation de ces unités dans Moses, un système de TAS à base de segments. Néanmoins, comme mentionné dans la section 1, la difficulté consiste à trouver la meilleure façon d'intégrer les EPLs dans de tels systèmes. À cet effet, nous proposons trois *stratégies d'intégration dynamiques* où le modèle de traduction est modifié de différentes façons et une *stratégie d'intégration statique* dans laquelle nous introduisons les EPLs au décodeur sans changer le modèle de traduction et comparons leurs performances dans la section 5.

4.1 Moses comme système de référence

Le système de traduction de référence (**Réf**) utilisé est Moses, un outil sous licence libre. Dans ce système, l'unité de traduction est le segment, qui correspond à un groupe de mots contigus. Le modèle de traduction sert de pont entre les langues source et cible. Son rôle est de guider la construction, pour chaque phrase source, d'un ensemble d'hypothèses de traduction en langue cible. Lors de la phase de décodage, ces hypothèses de traduction sont sélectionnées à partir d'un inventaire constitué d'un ensemble d'appariements entre des segments de longueur variable. Ces associations et les scores qui les accompagnent constituent la table de traduction *(phrase table)*.

4.2 Stratégies d'intégration dynamiques

4.2.1 Nouveau modèle de traduction

Les tables de traduction constituent la source principale de connaissance pour le décodeur. Le décodeur consulte ces tables pour déterminer comment traduire une phrase source en langue cible. Cependant, en raison d'erreurs d'alignement automatique de certains mots, des segments extraits peuvent être dénués de sens. Pour remédier à ce problème, nous proposons de considérer les EPLs comme des paires de phrases parallèles : nous les ajoutons au corpus d'apprentissage et entraînons un nouveau modèle de traduction. Dans cette méthode (**Train**), nous espérons que par l'augmentation du nombre d'occurrences des paires d'EPLs, considérées comme de bons segments, une modification de l'alignement et de la probabilité de la traduction soit enregistrée.

4.2.2 Extension de la table de traduction

Dans cette méthode, nous étendons la table de traduction du système de référence RÉF en y incorporant les paires d'EPLs bilingues acquises. Nous utilisons la valeur de l'indice de Jaccard proposée pour chaque paire d'EPLs pour définir la probabilité de traduction dans les deux directions et fixons les probabilités lexicales à 1 pour des raisons de simplicité. Ainsi, le décodeur prendra en considération des EPLs bilingues lors de la recherche de segments candidats pour traduire une phrase source. Cette méthode est notée **TABLE** dans le reste de cet article.

4.2.3 Trait additionnel pour les EPLs

(Lopez et Resnik, 2006) ont souligné qu'une meilleure définition des traits utilisés peut conduire à un gain substantiel dans la qualité des traductions. Nous suivons cette hypothèse et étendons la méthode TABLE. Nous définissons un nouveau *trait binaire* indiquant pour chaque entrée de la table de traduction s'il s'agit d'une EPL ou pas. Le but de cette méthode notée **TRAIT** est de guider le système pour choisir les EPLs bilingues plutôt que les hypothèses proposées par RÉF.

4.3 Stratégie d'intégration statique

Dans cette méthode, notée **FORCÉ**, nous voulons que le décodeur prenne en considération des EPLs bilingues tout en gardant le modèle de traduction de RÉF. À cet égard, nous utilisons le *mode de décodage forcé* du décodeur de MOSES. Ce dernier comporte un schéma de balisage XML permettant de spécifier des traductions pour des parties des phrases à traduire. Nous pouvons ainsi indiquer au décodeur ce qu'il faut utiliser pour traduire certains mots ou segments dans les phrases à traduire. Dans le cadre de notre étude, nous représentons chaque EPL apparaissant dans le corpus de test par la balise XML adéquate en se basant sur les traductions produites par notre aligneur. Un exemple de représentation de l'EPL *« à nouveau »* est présenté ci-dessous :

. . .sembler être **à nouveau** mis en accusation, le ministère public . . .

⇓

. . .sembler être < ***mwe* translation=*"once more***" >**à nouveau**< /***mwe***> mis en accusation, le ministère public . . .

5 Expériences et résultats

5.1 Corpus et outils

Les données d'apprentissage et de test proviennent du corpus Europarl pour la paire de langues français-anglais. Ce corpus regroupe un ensemble de phrases parallèles extraites des actes du parlement européen. Pour estimer le modèle de traduction du système de référence RÉF, nous avons construit un corpus d'apprentissage contenant après normalisation 100 000 paires de phrases. La normalisation est établie à travers les traitements suivants : tokenisation, suppression de phrases de plus de 50 mots et lemmatisation à l'aide de l'outil TreeTagger. Nous utilisons

Method	BLEU		TER	
	Tous_Test	EPLs_*Test*	*Tous_Test*	EPLs_*Test*
RÉF	28,85	30,83	55,44	53,59
Dynamiques				
TRAIN	**28,87**	**31,06**	**55,38**	**53,32**
TABLE	28,82	**30,88**	**55,42**	**53,46**
TRAIT	**28,95**	**31,06**	55,48	**53,56**
Statique				
FORCÉ	28,20	29,19	56,01	55,05

TABLE 3 – Résultats de traduction des corpus de test *Tous_Test* et EPLs_*Test* en termes de scores BLEU et TER

ce même corpus pour construire un lexique bilingue d'EPLs[2]. Comme les entrées de ce lexique sont sous forme de lemmes et que le mode de décodage forcé de MOSES n'est actuellement pas compatible avec les modèles à base de facteurs, le modèle de traduction a été estimé sur des lemmes plutôt que sur des formes de surface. Outre le modèle de traduction, nous avons entraîné un modèle de langue (trigramme) sur une version lemmatisée de la totalité du corpus Europarl (1,8M) en utilisant la boite à outils IRSTLM[3].

Les stratégies dynamiques et statique décrites précédemment sont ensuite appliquées. Dans TRAIN, les EPLs bilingues sont ajoutées au corpus d'apprentissage pour estimer un nouveau modèle de traduction. En ce qui concerne TABLE, la table de traduction de RÉF est enrichie par les EPLs bilingues. Dans TRAIT, un trait additionnel 1/0 est introduit dans la table de traduction de TABLE. Finalement, FORCÉ maintient le modèle de traduction de RÉF. Tous les modèles obtenus sont optimisés par minimisation du taux d'erreur (*MERT : Minimum Error Rate Training*) (Och, 2003) sur un corpus de développement constitué de 4 000 paires de phrases issues du même corpus.

5.2 Résultats et discussion

Deux séries d'expériences ont été menées : *Tous_Test* et EPLs_*Test*. Le premier corpus de test *Tous_Test* est constitué de 1 000 paires de phrases parallèles extraites aléatoirement du corpus Europarl. Pour mesurer l'apport réel du lexique bilingue d'EPLs, nous avons constitué un corpus de test noté EPLs_*Test* où nous ne conservons que les phrases du corpus *Tous_Test* contenant au moins une EPL. Ce corpus contient 323 paires de phrases parallèles. La qualité de traduction du système RÉF et des différentes stratégies dynamiques et statique d'intégration est évaluée sur les deux corpus de test sur la base des mesures BLEU et TER (Snover *et al.*, 2006). Nous considérons qu'à chaque phrase source correspond une seule phrase de référence en langue cible. Les résultats de traduction pour les différentes configurations sont rassemblés dans le tableau 3.

À première vue, nous remarquons que le score BLEU varie en fonction du type du jeu de test. Concernant le corpus de test *Tous_test*, la meilleure amélioration est obtenue par la stratégie dynamique TRAIT. Cette méthode rapporte un faible gain de +0,1 points BLEU par rapport au

[2]EPLs bilingues extraites par l'algorithme décrit dans la section 3
[3]`http://hlt.fbk.eu/en/irstlm`

SRC	*je entendre en effet lancer un initiative communautaire pour le afrique en étendre le ligne nepad ...*
RÉFERENCE	*indeed , i intend to launch a community initiative* **for africa** *, develop the nepad line...*
RÉF	*i hear be indeed launch an* initiative **for the eu africa** *by extend the nepad line ...*
TRAIT	*i hear in fact launch a community initiative* **for africa** *by extend the nepad line ...*
SRC	*le deuxième groupe de problème relever de le aide international et du prochain engagement de johannesburg.*
RÉFERENCE	*another series of problem* **mention be a matter of** international aid *and the forthcoming johannesburg summit.*
RÉF	*the second group of the problem* **be a matter of** *international aid and the forthcoming johannesburg commitment.*
FORCÉ	*the second group of the problem* **relate to the** *international aid and the forthcoming johannesburg commitment.*

TABLE 4 – Exemples de traduction. Notons que le texte est lemmatisé. Nous soulignons les EPLs et mettons en gras différentes suggestions pour le contexte immédiat gauche ou droite.

système RÉF. Le premier exemple de traduction présenté dans le tableau 4 souligne la contribution du trait introduit à l'amélioration de la qualité de traduction. Contrairement à RÉF, traduisant l'EPL *« initiative communautaire »* par simplement le mot simple *« initiative »*, la stratégie TRAIT mène à bien la traduction de l'EPL *« initiative communautaire »* par *« community initiative »* et de son contexte immédiat (*« for africa »*). Des scores BLEU plus faibles que ceux rapportés par RÉF sont obtenus par les stratégies TABLE et FORCÉ.

Pour le corpus de test EPLs_*Test*, qui ne considère que les phrases contenant des EPLs du lexique bilingue, nous constatons que toutes les stratégies d'intégration dynamiques rapportent des scores BLEU plus élevés que ceux obtenus par RÉF et la stratégie statique FORCÉ. Un gain de $+0,23$ points BLEU est obtenu par TRAIT et TRAIN. La stratégie TABLE rapporte un score légèrement amélioré montrant un gain de $+0,05$ points BLEU. Contrairement aux stratégies d'intégration dynamiques, la méthode FORCÉ obtient de faibles scores sur les deux corpus de test. Ceci peut être expliqué de la manière suivante : nous supposons qu'en forçant le décodeur à traduire une EPL par une unité donnée, ce dernier échoue à bien traduire le contexte immédiat gauche ou droit de l'EPL induisant ainsi une diminution de la valeur du score BLEU. Ainsi, dans le second exemple du tableau 4, les deux systèmes produisent une bonne traduction pour l'EPL *« aide international »*. Cependant, FORCÉ échoue dans la traduction du segment *« relever de »*. Il est important de noter que cette traduction pourrait être soutenue dans le cas où nous associons à chaque phrase source de multiples traductions de référence.

Dans une étude antérieure, (Ren *et al.*, 2009) ont proposé une stratégie similaire à la stratégie TRAIT dans laquelle ils indiquent pour chaque entrée de la table de traduction si un segment contient une paire d'EPL bilingue spécialisée. Pour le domaine médical, leur méthode rapporte un gain de $+0,17$ points BLEU par rapport à MOSES, un gain plus faible que celui obtenu par la stratégie TRAIT. La question que l'on peut se poser en observant les différents résultats obtenus est : est-il possible de prétendre que le système ayant les meilleurs scores est vraiment

le meilleur système ? En d'autres termes, les résultats obtenus par différentes stratégies sont-ils *statistiquement significatifs* ?

Pour évaluer la significativité statistique des résultats obtenus, nous utilisons la *méthode par ré-échantillonage par amorce* décrite par (Koehn, 2004). Cette méthode estime la probabilité (*p-valeur*) qu'une différence mesurée entre les scores BLEU surgisse par hasard, par la création à plusieurs reprises (10 fois) d'échantillons uniformes avec remise à partir des corpus de tests. Nous nous appuyons sur cette méthode pour comparer les méthodes TRAIN, TABLE et TRAIT apportant des gains dans le score BLEU (Tableau 3) par rapport à RÉF. Les résultats obtenus sont présentés dans le tableau ci-dessous.

Méthode	*p-valeur* (95 % *IC*)	
	Tous_Test	EPLs_*Test*
RÉF	-	-
TRAIN	0,1	0,05
TABLE	-	0,3
TRAIT	0,01	0,01

TABLE 5 – Test de significativité statistique des résultats en termes de *p-valeur*

Sur un intervalle de confiance (IC) de 95%, les résultats varient de non significatifs (quand $p > 0,05$) à hautement significatifs. Sur les deux corpus de test, nous remarquons que les améliorations apportées par la stratégie TRAIT ayant une p-valeur de 0,05 sont significatifs. Cependant, le faible gain en score BLEU obtenu par TABLE (0,3 de p-valeur) est non significatif. La cause est que nous utilisons la valeur de l'indice de Jaccard, une mesure utilisée pour comparer la similarité et la diversité entre des échantillons, pour définir la probabilité de traduction. Ceci peut être ajusté par la transformation des valeurs de Jaccard obtenus pour chaque paire d'EPL en une probabilité de traduction assurant ainsi l'uniformité et la cohérence des probabilités dans la table de traduction.

Le score BLEU relève seulement les améliorations globales et ne montre aucune différence pouvant être révélée par une évaluation humaine. Cette observation nous a motivé à mener une évaluation lexicale fine des EPLs du corpus EPLs_*Test*. Nous avons construit un corpus de test constitué uniquement d'EPLs et avons manuellement créé une liste de références à partir du corpus de référence. Ce corpus a été traduit par RÉF,TRAIN, TABLE, TRAIT et FORCÉ. Les résultats obtenus évalués par les mesures du BLEU et TER sont présentés dans la figure 2. Nous constatons qu'un gain de $+9,8$ et de $-0,2$ respectivement de points BLEU et TER est relevé par la stratégie FORCÉ. Cela vient confirmer que l'obtention d'un faible score BLEU avec la stratégie FORCÉ dans les expériences précédentes n'est pas due à une mauvaise qualité du lexique bilingue d'EPLs. Nous remarquons aussi que les stratégies TRAIN et TRAIT obtiennent des scores plus élevés (respectivement 24,67 et 28,06 points BLEU) par rapport à ceux obtenus par RÉF ayant un score BLEU de 21,84.

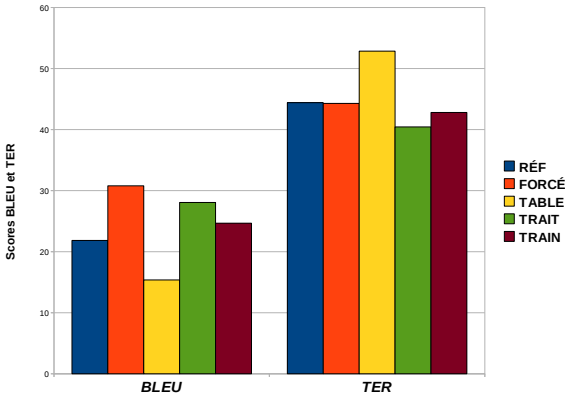

FIGURE 2 – Évaluation lexicale des EPLs en terme de scores BLEU et TER

6 Conclusion et travaux futurs

Dans cet article, nous avons décrit une méthode permettant d'extraire et d'aligner des EPLs dans un corpus parallèle français-anglais. L'algorithme d'alignement proposé effectue des alignements de type *m-n* et prend en considération des EPLs quel que soit le degré de corrélation entre leurs constituants. Pour mesurer l'apport de ces unités pour MOSES, nous avons présenté trois stratégies d'intégration dynamiques où nous avons modifié le modèle de traduction de différentes façons pour une prise en considération des EPLs bilingues et une stratégie d'intégration statique dans laquelle nous avons incorporé ces unités sans changer le modèle de traduction. Les expériences menées dans ce cadre montrent que la stratégie dynamique TRAIT, où un trait additionnel indiquant pour chaque entrée de la table de traduction s'il s'agit d'une EPL ou pas, peut améliorer significativement les résultats obtenus par MOSES avec un gain allant jusqu'à $+0,23$ points BLEU.

Nous considérons que nos expériences initiales sont positives et peuvent être améliorées de diverses façons. Dans ce présent travail, le modèle de traduction est estimé sur des lemmes plutôt que sur des formes de surface. Nous avons d'abord l'intention d'utiliser un modèle de génération pour produire les formes de surfaces adéquates à partir des résultats de traduction, présentés ici en lemmes. Nous comptons aussi entrainer notre système de traduction sur un corpus de taille plus importante afin d'évaluer l'impact du volume des données sur les résultats obtenus. En plus de leur application dans un système de TAS, nous tenterons d'étudier l'impact de ces EPLs sur la pertinence des résultats du moteur de recherche interlingue du CEA LIST.

Références

BESANÇON, R., DE CHALENDAR, G., FERRET, O., GARA, F., LAIB, M., MESNARD, O. et SEMMAR, N. (2010). Lima : A multilingual framework for linguistic analysis and linguistic resources development and evaluation. *In Proceedings of LREC*, Malta.

BOUAMOR, D., SEMMAR, N. et ZWEIGENBAUM, P. (2011). Improved statistical machine translation using multi-word expressions. *In Proceedings of MT-LIHMT*, Barcelona, Spain.

CASELI, H., VILLAVICENCIO, A., MACHADO, A. et FINATTO, M. J. (2009). Statistically-driven alignment-based multiword expression identification for technical domains. *In Proceedings of the Workshop on Multiword Expressions : Identification, Interpretation, Disambiguation and Applications*, Singapore.

CONSTANT, M., TELLIER, I., DUCHIER, D., DUPONT, Y., SIGOGNE, A., BILLOT, S. *et al.* (2011). Intégrer des connaissances linguistiques dans un crf : application à l'apprentissage d'un segmenteur-étiqueteur du français. *In Actes de TALN*, Montpellier, France.

DAGAN, I. et CHURCH, K. (1994). Termight : Identifying and translating technical terminology. *In Proceedings of the 4th Conference on ANLP*, pages 34–40, Stuttgart, Germany.

DAILLE, B. (2001). Extraction de collocation à partir de textes. *In* MAUREL, D., éditeur : *Actes de TALN 2001 (Traitement automatique des langues naturelles)*, Tours. ATALA, Université de Tours.

FAZLY, A. et STEVENSON, S. (2007). Distinguishing subtypes of multiword expressions using linguistically-motivated statistical measures. *In Proceedings of the Workshop on A Broader Perspective on Multiword Expressions*, pages 9–16.

FINLAYSON, M. et KULKARNI, N. (2011). Detecting multi-word expressions improves word sense disambiguation. *In Proceedings of the Workshop on Multiword Expressions : from Parsing and Generation to the Real World*, pages 20–24, Portland, Oregon, USA.

FRANTZI, C., ANANIADOU, S. et MIMA, H. (2000). Automatic recognition of multi-word terms : the c-value/nc-value method. *In Int. J. on Digital Libraries 3(2)*, pages 115–130.

JACKENDOFF, R. (1997). The architecture of the language faculty. *MIT Press*.

KOEHN, P. (2004). Statistical significance tests for machine translation evaluation. *In Proceedings of EMNLP*.

KOEHN, P., HOANG, H., BIRCH, A., CALLISON-BURCH, C., FEDERICO, M., BERTOLDI, N., COWAN, B., SHEN, W., MORAN, C., ZENS, R., DYER, C., BOJAR, O., CONSTANTIN, A. et HERBST, E. (2007). Moses : Open source toolkit for statistical machine translation. *In Proceedings of ACL, demo session*, Prague, Czech Republic.

KOEHN, P., OCH, F. et MARCU, D. (2003). Statistical phrase-based translation. *In Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 115–124, Edmonton, Canada.

KUPIEC, J. (1993). An algorithm for finding noun phrases correspondences in bilingual corpora. *In Proceedings of the 31st annual Meeting of the Association for Computational Linguistics*, pages 17–22, Columbus, Ohio, USA.

LAMBERT, P. et BANCHS, R. (2005). Data inferred multi-word expressions for statistical machine translation. *In Proceedings of MT SUMMIT*.

LAMBERT, P. et BANCHS, R. (2006). Grouping multi-word expressions according to part-of-speech in statistical machine translation. *In Proceedings of the Workshop on Multi-word Expressions in a multilingual context*.

LOPEZ, A. et RESNIK, P. (2006). Word-based alignment, phrase based translation :what's the link ? *In Proceedings of the association for machine translation in the Americas : visions for the future of machine translation*, pages 90–99.

LUKA, N., SERETAN, V. et WEHRLI, E. (2006). Le problème de collocation en tal. *In Nouveaux cahiers de linguistiques Française*, pages 95–115.

OCH, F. (2003). Minimum error rate training in statistical machine translation. *In Proceedings of ACL*.

OKITA, T., GUERRA, M., ALFREDO GRAHAM, Y. et WAY, A. (2010). Multi-word expression sensitive word alignment. *In Proceedings of the 4th International Workshop on Cross Lingual Information Access at COLING 2010*, pages 26–34, Beijing.

PAPINENI, k., ROUKOS, S., WARD, T. et ZHU, W. J. (2002). Bleu : a method for automatic evaluation of machine translation. *In 40th Annual meeting of the Association for Computational Linguistics*.

REN, Z., LU, Y., LIU, Q. et HUANG, Y. (2009). Improving statistical machine translation using domain bilingual multiword expressions. *In Proceedings of the Workshop on Multiword Expressions : Identification, Interpretation, Disambiguation and Applications*, pages 47–57.

SAG, I., BALDWIN, T., FRANCIS BOND, F., COPESTAKE, A. et FLICKINGER, D. (2002). Multiword expressions :a pain in the neck for nlp. *In CICLing 2002*, Mexico City, Mexico.

SAGOT, B., CLÉMENT, L., DE LA CLERGERIE, É., BOULLIER, P. *et al.* (2005). Vers un méta-lexique pour le français : architecture, acquisition, utilisation. *In Actes de TALN*.

SALTON, G., WONG, A. et YANG, C. (1975). A vector space model for automatic indexing. *In Communications of the ACM*, pages 61–620.

SERETAN, V. et WEHRLI, E. (2007). Collocation translation based on sentence alignment and parsing. *In* BENARMARA, F., HATOUT, N., MULLER, P. et OZDOWSKA, S., éditeurs : *Actes de TALN 2007 (Traitement automatique des langues naturelles)*, Toulouse. ATALA, IRIT.

SNOVER, M., DORR, B., SCHWARTZ, R., MICCIULLA, L. et MAKHOUL, J. (2006). A study of translation edit rate with targeted human annotation. *In Proceedings of Association for Machine Translation in the Americas*.

TUFIS, I. et ION, R. (2007). Parallel corpora, alignment technologies and further prospects in multilingual resources and technology infrastructure. *In Proceedings of the 4th International Conference on Speech and Dialogue Systems*, pages 183–195.

VECHTOMOVA, O. (2005). The role of multi-word units in interactive information retrieval. *In ECIR2005*, pages 403–420, Berlin.

VINTAR, S. et FISIER, D. (2008). Harvesting multi-word expressions from parallel corpora. *In Proceedings of LREC*, Marrakech, Morocco.