Alignement de termes de longueurs variables en corpus comparables spécialisés

Jingshu Liu, Emmanuel Morin, Sebastián Peña Saldarriaga

Alignement de termes de longueurs variables en corpus comparables spécialisés

Jingshu Liu[1, 2] Emmanuel Morin[1] Sebastián Peña Saldarriaga[2]
(1) LS2N / Université de Nantes, 2 Chemin de la Houssinière, 44300 Nantes, France
(2) Dictanova, 6 rue René Viviani, 44200 Nantes, France
prénom.nom@ls2n.fr, prénom@dictanova.com

RÉSUMÉ

Nous proposons dans cet article une adaptation de l'approche compositionnelle étendue capable d'aligner des termes de longueurs variables à partir de corpus comparables, en modifiant la représentation des termes complexes. Nous proposons également de nouveaux modes de pondération pour l'approche standard qui améliorent les résultats des approches état de l'art pour les termes simples et complexes en domaine de spécialité.

ABSTRACT

Alignment of variable length terms in specialized comparable corpora

We propose in this paper an adaptation of the extended compositional approach able to align terms of variable lengths from comparable corpora, by modifying the representation of complex terms. We also propose new weighting modes for the standard approach that improve the results of state-of-the-art approaches for simple and complex terms in specialised domains.

MOTS-CLÉS : Multilinguisme, alignement, corpus comparables, vecteur de contexte.

KEYWORDS: Multilinguism, alignment, comparable corpora, context vector.

1 Introduction

L'extraction de lexiques bilingues à partir de corpus comparables a suscité de nombreux travaux depuis le début des années 90 (Fung, 1995; Rapp, 1999; Li & Gaussier, 2010; Morin & Daille, 2012; Mikolov *et al.*, 2013a; Xing *et al.*, 2015; Artetxe *et al.*, 2016; Hazem & Morin, 2016). Deux classes d'approches ont été développées selon la nature du terme à aligner. La première classe s'intéresse à l'alignement de mots et termes simples et repose sur des approches distributionnelles tandis que la seconde classe porte sur l'alignement de termes complexes et repose sur des approches compositionnelles. Peu de travaux se sont intéressés à proposer un cadre unifié permettant de réaliser l'alignement des termes simples et complexes en dehors de Delpech *et al.* (2012) pour l'alignement de termes simples vers des termes complexes. Notre objectif est de proposer un tel cadre unifié permettant l'alignement de termes de longueurs variables en domaine de spécialité.

Nous proposons d'adapter l'approche compositionnelle étendue pour prendre en compte les termes simples et complexes. En outre, nous proposons d'améliorer l'approche standard pour l'alignement de termes simples qui est exploitée dans l'approche compositionnelle étendue.

2 Approches standard et compositionnelle

Nous présentons dans cette section les approches existantes pour l'alignement de termes simples et complexes ainsi que les modifications apportées.

2.1 Approche standard

L'approche par alignement de contextes, appelée également approche standard (AS), est employée pour l'extraction de lexiques bilingues à partir de corpus comparables. Celle-ci repose sur la simple observation qu'un mot et sa traduction ont tendance à apparaître dans les mêmes contextes lexicaux (Fung, 1995; Rapp, 1999). Dans cette approche, il faut commencer par construire une matrice de cooccurrences pour les langues source et cible, où chaque ligne représente un vecteur de contexte dans une fenêtre de *n* mots. Ces vecteurs sont ensuite normalisés par exemple avec l'Information Mutuelle (IM). Il s'agit ensuite de transférer en langue cible le vecteur de contexte d'un mot en traduisant les éléments du vecteur via un dictionnaire bilingue. En ce qui concerne les mots qui ont plusieurs traductions, un poids est distribué en fonction de la fréquence de chaque traduction dans le corpus. Finalement les traductions candidates sont ordonnées en calculant la similarité du vecteur de contexte traduit avec l'ensemble des vecteurs de contexte en langue cible via une mesure de similarité comme le cosinus.

Avec l'AS, nous avons constaté que certains mots dans la fenêtre sont peu liés au terme central, en général, plus ce dernier est éloigné d'un mot du contexte, moins ils sont sémantiquement liés. Après le filtrage des mots outils, un mot à l'origine très éloigné du mot central peut apparaître dans la fenêtre. Cela rend le vecteur de contexte moins pertinent en tant que représentation. Afin de réduire cet effet, il nous faut une fonction de pondération qui satisfait quelques critères :

- La fonction doit être monotone décroissante en $[1, +\infty[$ étant donné que la distance ne peut jamais être inférieure à 1.
- L'image dans $[1, +\infty[$ doit représenter un poids dans $]0, 1]$.
- La fonction ne doit pas pénaliser la cooccurrence lorsque c'est déjà le plus proche du mot central. Autrement dit, la fonction renvoie 1 comme poids si la distance est égale à 1.
- L'écart entre les poids de pénalisation pour les distances longues doit être relativement petit car les mots éloignés du mot central ont une influence comparable en terme de contribution sémantique.

Il existe certainement plusieurs fonctions qui satisfont ces critères, dans ces travaux nous avons employé la fonction de poids g définie ainsi :

$$g(c|w) = \Delta(w, c)^{-\lambda}, \quad \lambda \in [0, 1] \tag{1}$$

où $g(c|w)$ est le poids du mot c dans le contexte de w, Δ la distance entre c et w et λ l'hyperparamètre qui détermine le degré de pénalisation (plus il est élevé, plus les contextes éloignés sont pénalisés). Notons que $\lambda = 0$ correspond à une distribution uniforme. La figure 1 montre le graphe de cette fonction quand λ est fixé à 0,25.

Munis de cette fonction de poids nous pouvons proposer une pondération en fonction de la distance (notée PFD) pour les mots dans la fenêtre :

$$PFD(w, c) = g(c|w) \times cooc(w, c) \tag{2}$$

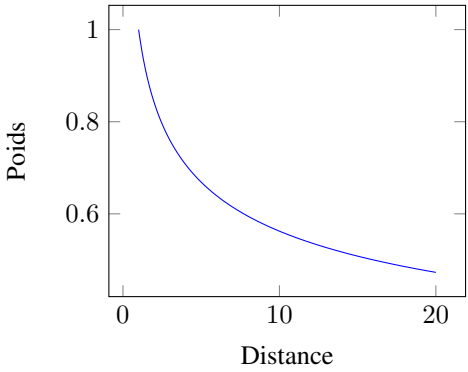

FIGURE 1: Fonction g avec $\lambda = \frac{1}{4}$

De nombreux travaux ont montré que l'IM surestime les faibles occurrences et sous-estime les hautes occurrences. Les travaux de Pennington *et al.* (2014) ont introduit une fonction pour lisser les occurrences sur laquelle nous nous appuyons pour proposer la mesure d'IM pondérée (IMP) pour améliorer l'IM dans nos expériences :

$$IMP(w,c) = f(cooc(w,c)) \times IM(w,c) \tag{3}$$

$$f(x) = \begin{cases} (x/x_{max})^{\alpha}, \alpha \in [0,1], \text{ si } x < x_{max} \\ 1 \text{ sinon} \end{cases} \tag{4}$$

où f est la fonction de Pennington *et al.* (2014), α et x_{max} sont les hyperparamètres qui caractérisent la fonction f. α détermine le degré de réduction pour les faibles occurrences, x_{max} détermine le seuil de réduction, par exemple $x_{max} = 20$ signifie que les occurrences moins de 20 doivent être réduites.

En raison de la petite taille des corpus en domaine spécialisé, les occurrences des mots ou des paires de mots ne sont pas toujours statistiquement fiables. Hazem & Morin (2016) ont montré que l'utilisation d'un corpus de langue générale peut améliorer significativement les résultats de l'AS. Ils proposent deux méthodes d'adaptation pour exploiter des ressources externes. La première adaptation, appelée approche standard globale (ASG), consiste à construire les vecteurs de contexte à partir d'un corpus comparable comportant des documents d'un domaine spécialisé et des documents du domaine général.

Nous avons implémenté la deuxième adaptation appelée approche standard sélective (ASS) qui a donné les meilleurs résultats. Pour chaque mot qui appartient au corpus du domaine spécialisé, s'il apparaît dans le corpus du domaine général, son vecteur de contexte spécialisés et généraux sont fusionnés. Cela permet de filtrer les mots de domaine général qui ne font pas partie du corpus spécialisé et rend l'approche standard sélective beaucoup moins coûteuse en temps calcul que l'approche standard globale. Soient S le vocabulaire du corpus spécialisé, G celui du corpus général, w est un mot à représenter, c est un mot qui apparaît dans la fenêtre autour de w :

$$\forall w \in S \cap G, \forall c \in S \cap G,\ cooc(w,c) = cooc_S(w,c) + cooc_G(w,c) \tag{5}$$

2.2 Approche compositionnelle

L'approche compositionnelle (AC) (Grefenstette, 1999; Tanaka, 2002; Robitaille *et al.*, 2006) est une approche simple et directe qui consiste à traduire chaque élément d'un terme complexe via un dictionnaire et à comparer toutes les permutations possibles par projection dans un corpus. La principale limite de cette approche est son incapacité à traduire un terme lorsqu'un des mots qui le composent n'est pas dans le dictionnaire. Pour résoudre ce problème, Morin & Daille (2012) ont proposé l'approche compositionnelle étendue (ACE), dont l'objectif est de combiner les avantages des approches standard et compositionnelle en substituant les mots hors dictionnaire par leur vecteur de contexte obtenu par l'AS. L'ACE commence par construire la matrice de cooccurrences comme pour l'approche standard. Ensuite il s'agit d'appliquer une traduction directe renforcée par alignement de contexte. Si un mot d'un terme à traduire n'est pas présent dans le dictionnaire, nous utilisons le vecteur de contexte obtenu par l'AS et le traduisons en langue cible, sinon nous récupérons directement le vecteur de contexte de la traduction en langue cible. L'étape suivante est la génération de toutes les combinaisons de la représentation pour un terme en langue source. Finalement les termes candidats sont ordonnés suivant le calcul de similarité avec tous les termes de même longueur en langue cible, le score final pour chaque possibilité étant défini par la moyenne arithmétique ou géométrique de chaque score de similarité.

2.3 Adaptation à l'alignement de termes de longueurs variables

L'approche compositionnelle étendue ne permet pas de prendre en compte le problème de fertilité, c'est-à-dire l'alignement de termes de longueurs variables. Par exemple, le terme anglais « *wind vane* » peut être traduit par « *girouette* » en français et le terme anglais « *wind energy* » par « *Windenergie* » en allemand. Afin de prendre en compte ces cas, nous nous sommes inspirés des travaux de Blacoe & Lapata (2012) dans lesquels la représentation d'une phrase est la somme des représentations distributionnelles de chacun de ses mots. Cependant dans ce travail la pondération de chaque élément d'un terme complexe n'est pas prise en compte. En effet le vecteur construit par la simple somme de chaque élément est davantage orienté vers les vecteurs qui ont des valeurs plus importantes. Dans ce travail nous souhaitons vérifier l'hypothèse que la moyenne des vecteurs normalisés de tous les éléments représente plus fidèlement un terme complexe. L'intuition derrière cette hypothèse est que ces éléments de longueur uniforme assurent un impact équivalent pour la construction sémantique d'un terme complexe. Nous proposons ainsi de modifier la représentation des termes complexes dans l'ACE ainsi que dans l'AS :

$$vecteur(terme) = \frac{1}{n}\sum_{i}^{n}\frac{vecteur(w_i)}{||vecteur(w_i)||} \tag{6}$$

où $||\vec{x}||$ représente la l^2-$norme$ d'un vecteur $\vec{x}$ et n la longueur du terme.

Dans la figure 2 nous démontrons la logique de la moyenne sur les vecteurs normalisés par un exemple : soit un terme complexe composé de deux mots a et b, et leurs vecteurs de contexte respectifs $\vec{a}$ et $\vec{b}$. Si $\vec{a}$ est plus long que $\vec{b}$ (dans le contexte de l'approche standard cela signifie que les cooccurrences concernant le mot a sont plus importantes), la moyenne des deux vecteurs de contexte sera plus proche de a. Or il n'est pas toujours vrai que le sens d'un terme complexe est déterminé par l'élément plus fréquent. Le vecteur final que nous proposons, illustré en rouge sur la figure 2, forme un angle égal avec $\vec{a}$ et $\vec{b}$. Dans la partie basse du graphe nous illustrons la différence (l'angle) entre les deux

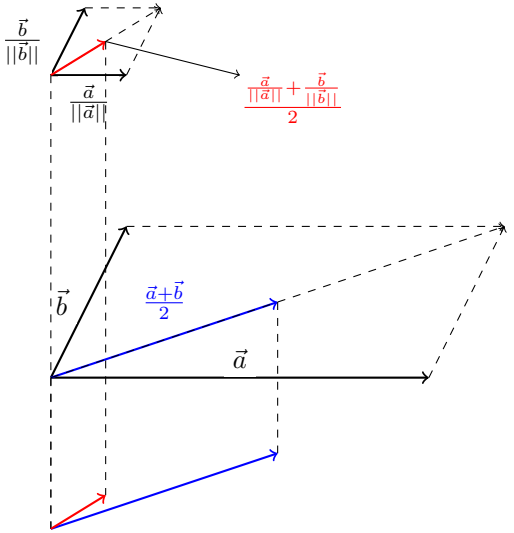

FIGURE 2: Illustration de la différence entre la moyenne des vecteurs non normalisés (le vecteur en bleu) et normalisés (le vecteur en rouge).

vecteurs obtenus par les deux stratégies. Il est par ailleurs clair que si nous calculons la similarité par le Cosinus, la moyenne et la somme des vecteurs sont équivalentes car le Cosinus est une mesure en fonction de la direction seule.

La représentation du terme se fait dans un seul vecteur, donnant la capacité de gérer les traductions de longueurs variables tout en réduisant le temps de calcul. En effet, dans l'ACE originale, aligner un terme complexe demande de calculer toutes les permutations possibles, il faut donc comparer un nombre factoriel de vecteurs, contre un seul vecteur à comparer pour la version adaptée.

2.4 Adaptation de l'approche compositionnelle étendue

Nous illustrons le schéma de notre méthode qui consiste en l'approche compositionnelle avec l'adaptation de la représentation pour les termes complexes.

Dans la figure 3 nous montrons comment un terme complexe en langue source est aligné à un terme de longueur différente en langue cible. w_{sn} est le n-ième mot du terme complexe en langue source, w_{sn} le n-ième mot du terme en langue cible. Dans l'exemple de la figure 3, 3 traductions ont été trouvées pour w_{s1} dans le corpus comparable via le dictionnaire, une pour w_{s2} et aucune traduction n'a été trouvé pour w_{s3} dans le corpus comparable. t_n^m signifie le vecteur de contexte pour m-ième traduction possible en langue cible pour le mot w_{sn}, alors que s^n représente le vecteur de contexte obtenu par l'approche standard pour le mot w_{sn}.

Il est à noter que tous ces vecteurs sont dans l'espace commun de la langue cible. Les opérations par élément telle que l'addition sont donc raisonnables. Lors de l'étape suivante 3 ($\prod m$) traductions sont possibles pour le terme complexe parce que nous ne prenons pas en compte l'ordre des éléments du terme complexe. Nous appliquons notre méthode de la représentation des termes complexes et nous obtenons un seul vecteur pour chaque traduction.

Ensuite pour comparer la similarité entre une possible traduction du terme en langue source $w_{s1}w_{s2}w_{s3}$ avec le terme en langue cible $w_{t1}w_{t2}$, il suffit de calculer le produit scalaire (nous supposons que les vecteurs sont déjà normalisés après le calcul de la moyenne). Finalement la similarité entre les deux termes est $max(similarité1, similarité2, similarité3)$.

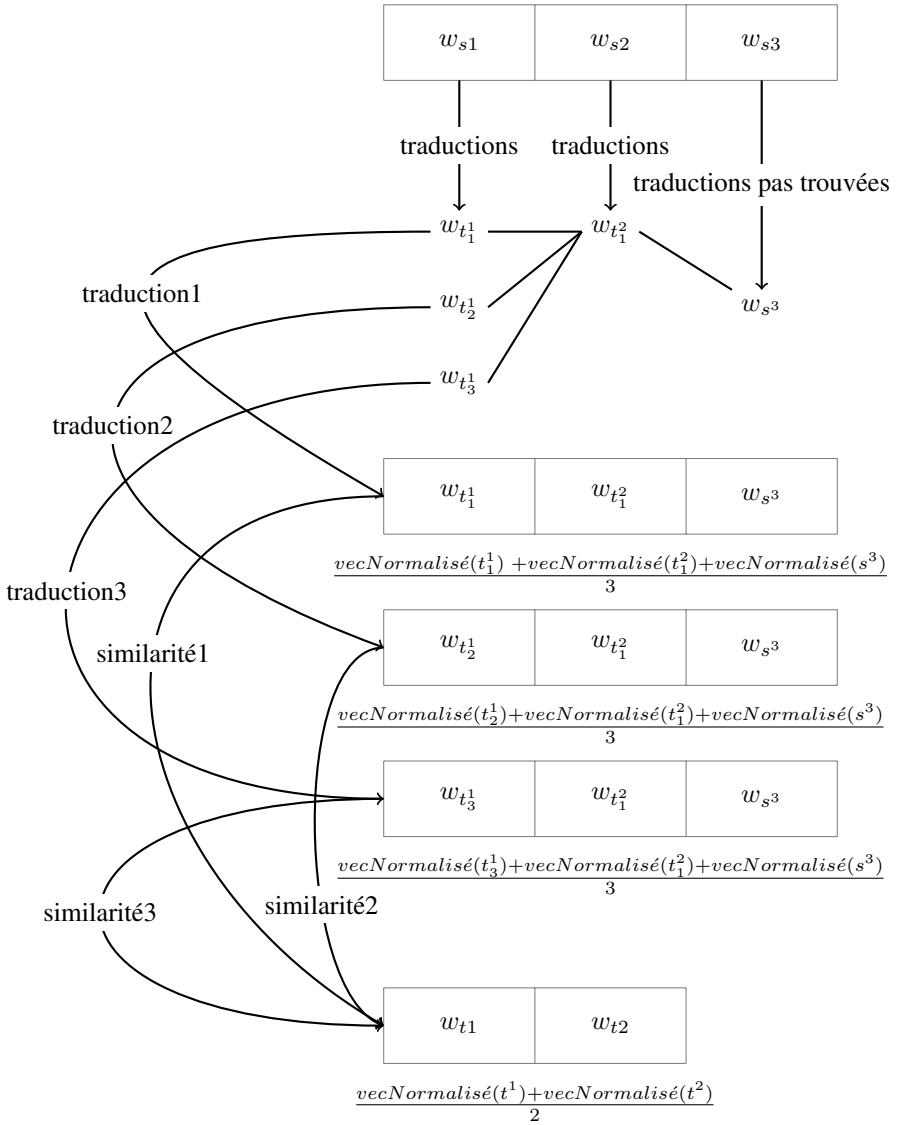

FIGURE 3: Schéma de notre approche pour l'alignement unifié

3 Expériences

Afin de valider notre implémentation des méthodes de l'état de l'art, nous les testons sur les termes simples (TS) avant de les appliquer au sujet qui nous intéresse : les termes complexes (TC). Nous avons aussi expérimenté l'utilisation de l'ASS dans l'approche compositionnelle étendue. À notre connaissance, notre proposition est le premier travail à les combiner.

3.1 Ressources

Nous avons utilisé le corpus spécialisé *Breast Cancer* (BC) (Hazem & Morin, 2016) pour l'expérience sur les termes simples et le corpus de langue générale *News Commentary* (NC) [1]. En ce qui concerne l'alignement des termes complexes, nous avons utilisé le corpus spécialisé *Wind Energy* (WE) [2] et deux corpus propriétaires spécialisés traitant des domaines du luxe et de la cosmétique.

Afin d'évaluer l'alignement des termes simples, nous avons utilisé la liste de référence fournie par Hazem & Morin (2016) pour le corpus BC. En ce qui concerne les termes complexes, nous avons trois listes de référence, celle du corpus WE est construite à partir des listes terminologiques fournies avec le corpus, et celles des corpus luxe et cosmétique sont construites manuellement par des experts du domaine à partir d'une liste de termes simples et complexes extraits par un système propriétaire symbolique qui utilise des règles morpho-syntaxiques comme *ACABIT* (Daille, 2003) [3] ou *TermSuite* (Daille, 2016) [4]. Les traductions sont validées en 3 itérations par les même experts du domaine qui ont choisi les termes à traduire. De multiples traductions pour un terme source sont possibles dans la liste de référence pour les corpus WE, Luxe et Cosmétique. Nous avons inclus les termes simples car leurs traductions peuvent être des termes complexes, ils s'intègrent donc naturellement dans l'alignement des termes complexes. De plus les termes à traduire dans les corpus pour la tâche des termes complexes sont hors dictionnaire.

Les candidats pour nos expériences sur les termes simples sont tous les mots dans le vocabulaire, pour nos expériences sur les termes complexes les candidats sont tous les mots dans le vocabulaire plus tous les termes complexes extraits par le système d'extraction terminologique qui génère généralement trois fois plus de termes complexes que de mots dans le vocabulaire.

Le tableau 1 présente les principales caractéristiques de ces différentes ressources.

Corpus	Nombre de mots		Taille de vocabulaire		Référence
	FR	EN	FR	EN	
BC	521 262	525 934	6 630	8 821	TS : 248
WE	314 549	313 943	6 038	7 134	TC : 73
Luxe	101 542	139 867	3 064	3 981	TC : 276, TS : 13
Cosmétique	430 106	837 579	3 913	5 592	TC : 185, TS : 14
NC	5,7 M	4,7 M	23 597	29 489	

Tableau 1: Caractéristiques des corpus utilisés

1. opus.lingfil.uu.se
2. ttc.project.eu
3. www.bdaille.com/index.php?option=com_content&task=blogcategory&id=5&Itemid=5
4. termsuite.github.io

3.2 Configuration

Pour toutes nos expériences, nous prétaitons les corpus via la tokenisation, le pos-tagging et la lemmatisation, ensuite nous filtrons les hapax et les mots outils. Dans la PFD, l'hyperparamètre a été empiriquement fixé à $\lambda = 0,25$. La taille de la fenêtre contextuelle d'un terme est de trois mots avant et trois mots après. Dans l'IMP, l'hyperparamètres x_{max} est fixé à 20 car la taille de nos corpus est relativement petite. L'autre hyperparamètre α de l'IMP est fixé à $\frac{3}{4}$ comme Pennington *et al.* (2014) ont décidé dans leur travaux. Il est d'ailleurs intéressant que dans les travaux de Mikolov *et al.* (2013b), une même puissance fractionnelle a été introduite pour obtenir les meilleurs résultats.

Pour la traduction des vecteurs de contexte nous avons utilisé le dictionnaire FRAN-EURADIC de ELRA [5] comportant 243 539 entrées.

Afin d'accélérer le processus pendant la phase de comparaison, tous nos vecteurs sont normés à 1 et nous avons utilisé la similarité cosinus car elle permet d'utiliser la parallélisation sur CPU ou GPU.

3.3 Résultats

Le tableau 2 montre les résultats de l'AS et l'ASS pour les termes simples. Nous voyons que l'IM pondérée améliore les résultats par rapport à ceux obtenus par Hazem & Morin (2016) avec l'IM mais aussi à ceux qu'ils obtiennent avec le *DOR* (*Discounted Odds Ratio*) (Evert, 2005) à la place du cosinus, avec le DOR la MAP atteint 0,270. Nos résultats montrent l'intérêt de pénaliser les petites occurrences pour compenser la surestimation de l'IM originale. Cependant cette pondération est moins efficace quand les données sont enrichies car la surestimation des petites occurrences est lissée par l'ajout de données exogènes. L'IM pondérée avec les données exogènes déchoit quelques bonnes traductions de la position top 1 (P@1 de 50,8 à 48,0), mais elle promeut plus de bonnes traductions dans le top 5 (P@5 de 62,5 à 64,1). Étant donné que certains termes à traduire sont assez peu fréquents ou même inexistants dans le corpus général, il est possible que pénaliser toutes les petites occurrences réduise des traits discriminants du corpus général. Pourtant nous constatons que la PFD améliore systématiquement nos résultats dans tous les cas. Si nous regardons la comparaison entre l'AS et l'AS + PFD, et celle entre l'ASS et l'ASS + PFD, les résultats sont meilleurs pour l'ensemble des quatre mesures.

Modèle	P@1	P@5	P@20	MAP
AS (Hazem & Morin, 2016)	18,5	35,5	46,0	25,9
AS + IMP	21,4	**37,1**	49,6	28,9
AS + PFD	19,4	36,3	47,2	27,4
AS + IMP + PFD	**22,6**	**37,1**	**50,4**	**29,5**
ASS	50,8	62,5	72,6	56,5
ASS + IMP	48,0	**64,1**	71,8	55,0
ASS + PFD	**51,6**	62,9	**73,4**	**57,3**
ASS + IMP + PFD	48,4	**64,1**	**73,4**	55,8

Tableau 2: Précision@k et MAP (%) pour l'alignement des termes simples sur le corpus BC.

En ce qui concerne les termes complexes, nous effectuons trois ensembles d'expériences reposant sur

5. www.elra.info

l'approche standard (AS), l'approche standard sélective (ASS), l'approche compositionnelle (AC).

Pour vérifier si l'IM pondérée est toujours favorable dans différents scénarios, nous l'appliquons à plusieurs configurations.

Afin d'illustrer la capacité de notre algorithme à aligner des termes de longueurs variables, le tableau 3 montre quelques exemples de traductions ou quasi-traductions trouvées par notre méthode avec l'adaptation de la représentation pour les termes complexes. Ces traductions sont dans le top@5 des candidats pour les différents corpus. Parmi toutes les traductions dans la liste top 5, nous montrons uniquement celles dont la longueur est différente du terme de départ.

Corpus	Alignements (anglais → français)
Luxe	sneaker shoe → sneaker invoice → facture d'achat
Cosmétique	wide variety of choice → nombreux choix very small tight store → petit magasin
WE	greenhouse gas → gaz à effet de serre power system → système éolien de puissance

Tableau 3: Exemples des traductions trouvées dans le top@5

Le tableau 4 montre les résultats des différentes méthodes avec l'adaptation pour les termes complexes. Nous voyons qu'elles se comportent de façon homogène à travers les différents domaines, tant que l'approche utilise des vecteurs de contexte unifiés. Nous ne montrons pas les résultats de l'AS sans l'adaptation de la représentation de TC car ceux-ci sont négligeables. En ce qui concerne le temps de calcul, l'alignement d'un terme complexe dans le corpus Luxe prend en moyenne 10 minutes avec l'ACE alors que la version adaptée prend 20 secondes. De plus la version originale ne permet pas d'aligner les termes de longueurs variables, alors que la traduction pour les termes complexes de la même longueur n'existe pas toujours dans le corpus comparable. Nous avons décidé de ne pas reproduire les résultats de l'ACE originale sur les autres corpus étant donné les performances supérieures de la version adaptée et le temps de calcul beaucoup plus rapide.

Dans le tableau 4 nous observons que nos propositions, l'IMP et la PFD, améliorent significativement (de 3 à 10 points en MAP) les résultats par rapport à l'approche standard avec ou sans données exogènes. L'ASS ne fonctionne pas aussi bien que pour les termes simples. En général, l'approche compositionnelle étendue avec l'adaptation pour les TC présente les meilleurs résultats à travers les trois corpus de différents domaines de spécialité. Sur le corpus Luxe nous avons de plus expérimenté l'approche compositionnelle étendue sans l'adaptation : son amélioration sur l'approche compositionnelle est de 4,2 points en MAP, ce qui est relativement faible par rapport à la version avec l'adaptation qui est de 15,4 points.

3.4 Discussion

Contrairement aux résultats sur les termes simples, l'IM pondérée améliore toujours les résultats avec l'ASS. Nous expliquons cela par le fait que le vecteur pour un terme complexe est une moyenne, par conséquent le risque que chaque élément soit peu fréquent est beaucoup moins élevé. La pénalisation pour les faibles occurrences devient donc une pénalisation partielle.

Corpus	Modèle	P@1	P@5	P@20	MAP
Luxe	AS	4,2	11,8	24,2	8,9
	AS + IMP	14,2	21,4	33,2	18,6
	AS + IMP + PFD	**14,5**	**21,5**	**33,9**	**18,9**
	AC	23,8	25,6	25,6	24,7
	ACE[a] + IMP + PFD	24,6	32,2	39,1	28,9
	ACE + IMP + PFD	**34,6**	**44,6**	**57,1**	**40,1**
	ACE[b] + IMP + PFD	24,2	37,0	51,2	31,2
	ASS	1,7	3,1	8,7	3,0
	ASS + IMP	2,8	8,0	13,8	5,8
	ASS + IMP + PFD	**3,1**	**8,3**	**14,5**	**5,9**
Cosmétique	AS	0,5	4,0	8,0	2,9
	AS + IMP	**5,6**	11,1	21,6	**10,3**
	AS + IMP + PFD	4,5	**12,6**	**22,6**	**10,3**
	AC	**12,5**	**19,1**	19,6	15,4
	ACE + IMP + PFD	11,6	**19,1**	**28,1**	**16,8**
	ACE[b] + IMP + PFD	7,0	16,6	24,6	12,9
	ASS	0	1,0	5,6	1,0
	ASS + IMP	**3,5**	9,5	17,1	7,1
	ASS + IMP + PFD	3,0	**10,0**	**17,6**	**7,3**
WE	AS	1,4	24,7	43,8	12,2
	AS + IMP	**12,3**	28,8	**50,7**	21,7
	AS + IMP + PFD	**12,3**	**31,5**	**50,7**	**21,9**
	AC	**59,0**	68,5	68,5	61,5
	ACE + IMP + PFD	42,5	**80,8**	**89,0**	60,0
	ACE[b] + IMP + PFD	53,4	**87,7**	**90,4**	**66,3**
	ASS	**11,0**	20,5	31,5	14,8
	ASS + IMP	9,6	**27,4**	37,0	**17,9**
	ASS + IMP + PFD	8,2	**27,4**	**40,0**	17,8

Tableau 4: Précision@k et MAP (%) pour l'alignement des termes complexes ([a] sans adaptation pour les TC de longueurs variables et [b] avec ASS pour l'alignement des mots hors dictionnaire).

Il est d'ailleurs surprenant que l'enrichissement n'apporte pas d'amélioration substantielle pour les corpus Luxe et Cosmétique, qui sont très bruités et contiennent beaucoup de mots hors dictionnaire (des argots des internautes). Notre hypothèse est qu'avec l'introduction des ressources externes, les mots de langue générale submergent le sens des mots spécifiques au corpus original. Dans nos expériences, le style du corpus général NC diffère beaucoup des corpus extraits des utilisateurs de l'internet. En effet, les mots spécifiques sont absents du corpus NC ou ne se comportent pas de la même manière dans celui-ci. Sur les deux corpus propriétaires extraits de l'internet les meilleurs résultats sont obtenus avec l'approche ACE adaptée sans ASS. Une solution potentielle est d'utiliser davantage de corpus généraux contenant plus de mots spécifiques. Pour le corpus WE qui est un corpus plus propre, et dont la plupart des composants des termes se trouvent dans le dictionnaire, la partie AS apporte peu par rapport à l'approche compositionnelle, et le classement par fréquence de l'AC est meilleur en P@1 et en MAP. Dans ce même corpus l'ASS dans l'ACE améliore sensiblement la MAP par rapport aux approches sans données exogènes, cela corrobore notre intuition sur l'enrichissement des données qui permet d'améliorer l'AS pour les mots qui existent dans les deux corpus et possèdent

des distributions similaires.

Parmi les termes complexes pour lesquels nous n'avons pas de traductions, nous identifions trois catégories de causes d'erreur :

— Faible compositionnalité. La traduction de l'ensemble n'est pas une combinaison de toutes les traductions de chaque élément. Par exemple « *pitch angle* » est traduit par « *angle d'inclinaison* » par le système alors que la bonne traduction est « *angle de calage* » où le mot « *calage* » n'est pas la traduction d'un mot dans le dictionnaire bilingue. Ce problème peut être encore plus grave entre deux autres langues linguistiquement lointaines, Tanaka (2002) rapporte qu'au moins 50% des composés japonais *NN* ne sont pas traduits en anglais par le même modèle syntaxique de composition. C'est d'ailleurs une piste intéressante à poursuivre pour expérimenter si notre système est capable de maintenir la performance sur d'autres langues.

— Ambiguïté. Plusieurs traductions possibles ont la même similarité dans la liste top@k car chaque mot composant a plusieurs traductions dans le dictionnaire et certaines combinaisons existent dans la liste des candidats terminologiques. Par exemple « *collier dior* » est traduit par « *chain* » alors que la bonne traduction est « *dior necklace* » ou « *necklace* » si nous permettons les quasi-traductions. Les mots « *chain* » et « *necklace* » sont tous deux traductions dans le dictionnaire or dans le contexte spécifique, « *chain* » est une fausse traduction. C'est une des causes qui engendrent la randomisation de l'ordre des 5 premiers candidats pour certains termes car plusieurs traductions candidates ont le même score. Cela peut expliquer le fait que l'ACE présente un résultat P@1 pour le corpus Cosmétique et WE inférieur à l'AC. Une couche de plus de désambiguïsation selon le contexte pour les mots qui ont plusieurs traductions dans le corpus pourrait être une solution intéressante.

— Différents ordres avec les même mots. Notre approche ne prend pas compte l'ordre des mots dans les termes complexes. De ce fait plusieurs termes complexes ont le même vecteur de représentation. Par exemple « *power installation* » est traduit par « *puissance d'installation* » alors que la bonne traduction est « *installation de puissance* ». C'est la deuxième cause qui rend l'ordre des 5 premiers candidats aléatoire pour certains termes. C'est aussi pourquoi l'ACE n'est pas meilleure en top@1 par rapport à l'AC pour le corpus Cosmétique et WE, mais elle rattrape et même surpasse l'AC en top@5. Cependant prendre en compte l'ordre provoquerait une augmentation significative du temps de calcul et l'alignement de termes de longueurs variables serait alors plus difficilement réalisable.

4 Conclusion

Nous avons proposé dans cet article une adaptation de l'ACE capable d'aligner des termes de longueurs variables, en modifiant la représentation des termes complexes. Nous avons aussi proposé de nouveaux modes de pondération pour l'AS qui améliorent les résultats des approches état de l'art pour les termes simples et complexes en domaine de spécialité. Nous espérons que les contributions de cet article aideront à approfondir la compréhension des approches par vecteurs de contexte.

L'enrichissement avec des données externes n'a pas montré des résultats homogènes pour l'alignement de termes complexes, cela nous donne des pistes pour notre futur travail. D'un côté nous prévoyons de travailler sur la représentation unifiée, qui est aujourd'hui simpliste car elle considère que chaque composant d'un terme complexe a la même importance dans la constitution du sens global. De l'autre côté, nous envisageons de profiter des système neuronaux, bien que ces méthodes ne s'accordent pas

naturellement avec notre contexte spécifique, cela nous semble intéressant de réfléchir à de nouvelles architectures correspondant mieux à notre besoin.

Remerciements

Les auteurs tiennent à remercier M. Joseph Lark pour ses commentaires et propositions qui ont permis d'étoffer le texte, ainsi que les 3 relecteurs anonymes pour leurs remarques pertinentes.

Références

ARTETXE M., LABAKA G. & AGIRRE E. (2016). Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP'16)*, p. 2289–2294, Austin, TX, USA.

BLACOE W. & LAPATA M. (2012). A comparison of vector-based representations for semantic composition. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP'12)*, p. 546–556, Jeju Island, Korea.

DAILLE B. (2003). Conceptual structuring through term variations. In *Proceedings ACL 2003 Workshop on Multiword Expressions : Analysis, Acquisition and Treatment*, p. 9–16, Sapporo, Japan.

DAILLE B. (2016). Terminology extraction with term variant detection. In *Proceedings of ACL-2016 System Demonstrations*, p. 13–18, Berlin, Germany.

DELPECH E., DAILLE B., MORIN E. & LEMAIRE C. (2012). Extraction of domain-specific bilingual lexicon from comparable corpora : compositional translation and ranking. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING'12)*, p. 745–762, Mumbai, India.

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

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.

GREFENSTETTE G. (1999). The world wide web as a resource for example-based machine translation tasks. In *Proceedings of the ASLIB Conference on Translating and the Computer 21*, London, UK.

HAZEM A. & MORIN E. (2016). Efficient Data Selection for Bilingual Terminology Extraction from Comparable Corpora. In *Proceedings of the 26th International Conference on Computational Linguistics (COLING'16)*, p. 3401–3411, Osaka, Japan.

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 (COLING'10)*, p. 644–652, Beijing, China.

MIKOLOV T., LE Q. V. & SUTSKEVER I. (2013a). Exploiting similarities among languages for machine translation. *CoRR*, **abs/1309.4168**.

MIKOLOV T., SUTSKEVER I., CHEN K., CORRADO G. & DEAN J. (2013b). Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems (NIPS'13)*.

MORIN E. & DAILLE B. (2012). Revising the Compositional Method for Terminology Acquisition from Comparable Corpora. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING'12)*, p. 1797–1810, Mumbai, India.

PENNINGTON J., SOCHER R. & MANNING C. (2014). Glove : Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP'14)*, p. 1532–1543, Doha, Qatar.

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.

ROBITAILLE X., SASAKI Y., TONOIKE M., SATO S. & UTSURO T. (2006). Compiling French-Japanese Terminologies from the Web. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL'06)*, p. 225–232, Trento, Italy.

TANAKA T. (2002). Measuring the Similarity Between Compound Nouns in Different Languages Using Non-parallel Corpora. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING'02)*, p. 1–7, Taipei, Taiwan.

XING C., WANG D., LIU C. & LIN Y. (2015). Normalized Word Embedding and Orthogonal Transform for Bilingual Word Translation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies (NAACL'15)*, p. 1006–1011, Denver, CO, USA.