Extraction automatique de traductions anglaises de mots composés français

Mathieu Constant, Takuya Nakamura, Stavroula Voyatzi, André Bittar

Neveu F., Muni Toke V., Durand J., Klingler T., Mondada L., Prévost S. (éds.)
Congrès Mondial de Linguistique Française - CMLF 2010
978-2-7598-0534-1, Paris, 2010, Institut de Linguistique Française

Traitement automatique des langues
DOI 10.1051/cmlf/2010255

Extraction automatique de traductions anglaises de mots composés français

Constant, Matthieu

Université Paris-Est, LIGM & CNRS
mconstan@univ-mlv.fr

Nakamura, Takuya

Université Paris-Est, LIGM & CNRS
nakamura@univ-mlv.fr

Voyatzi, Stavroula

Université Paris-Est, LIGM & CNRS
voyatzi@univ-mlv.fr

Bittar, André

Université Paris-Diderot, ALPAGE
andre.bittar@linguist.jussieu.fr

1 Introduction

La traduction des expressions multi-mots pose de sérieux problèmes du fait de leurs contraintes syntaxiques et sémantiques. Par ailleurs, bien qu'elles soient très présentes dans les textes, la fréquence des expressions multi-mots prises individuellement est relativement faible (Sag et al. 2002) ce qui cause des difficultés statistiques pour extraire les traductions.

De plus en plus d'études ont été réalisées sur ce sujet, expérimentant des méthodes statistiques (entre autres, Smadja et al., 1996 ; Caseli et al., 2007 ; Bai et al., 2009) et/ou des méthodes plus linguistiques (Lü et Zhou, 2004 ; Seratan et Wehrli, 2007). La plupart du temps, elles traitent des collocations et utilisent des corpus parallèles multilingues.

Dans notre article, nous traitons uniquement les mots composés, séquences de mots contigus non-compositionnelles, qui sont présentes dans le dictionnaire DELACF (Courtois et al. 1995). Nous confrontons les méthodes utilisées pour les collocations aux mots composés. Alors que les collocations ont tendance à mettre en relation deux mots pleins (ex. verbe-nom pour les collocations verbe-objet, ex: *prendre l'apéritif* ; nom-adjectif pour les collocations nominales: *pain perdu*), certains types de mots composés comme les prépositions ne possèdent souvent qu'un seul mot plein entouré de mots grammaticaux (*au sein de*), ce qui les rend plus difficile à repérer et traduire que les collocations traditionnelles.

Etant donné un mot composé identifié dans une phrase en français d'un corpus parallèle, le but est d'extraire automatiquement la traduction du mot composé dans la phrase correspondante en anglais, si elle existe, en tenant compte du fait qu'elle n'est pas forcément un mot composé anglais. Ce balisage permet d'extraire du corpus un ensemble de traductions et ainsi initier la création d'une ressource bilingue. Les mots composés que nous traitons appartiennent à quatre catégories : les noms, les adverbes, les conjonctions et les prépositions. Nous nous basons sur les études réalisées sur l'extraction statistique des traductions de collocations. Celles-ci se fondent sur les modèles probabilistes IBM d'alignement (Caseli et al., 2007) ou sur des mesures d'association (Bai et al. 2009).

Dans la section 2, nous décrivons les ressources lexicales et les outils qui sont utilisés pour le repérage des mots composés. La section 3 décrit notre corpus de travail qui est un sous-corpus d'Europarl (Koehn, 2003) et nous montrons ses différentes caractéristiques statistiques (sur les mots composés en particulier).

Neveu F., Muni Toke V., Durand J., Klingler T., Mondada L., Prévost S. (éds.)
Congrès Mondial de Linguistique Française - CMLF 2010
978-2-7598-0534-1, Paris, 2010, Institut de Linguistique Française

Traitement automatique des langues
DOI 10.1051/cmlf/2010255

Dans la section 4, nous expliquons deux méthodes de repérage de traductions de mots composés, exploitant directement les résultats de l'aligneur mot à mot Giza++ (Och et Ney, 2003). Dans la section 5, nous détaillons une méthode fondée sur des mesures d'associativité. Dans une dernière section, nous évaluons ces différentes méthodes en les confrontant à un corpus d'évaluation annoté semi-automatiquement.

2 Les mots composés

2.1 Le dictionnaire DELACF

Les mots composés sont des séquences de mots avec des contraintes sémantiques et syntaxiques. Par exemple, le sens de l'adverbe temporel *tout de suite* et du nom *eau de vie* ne peuvent pas être déduits du sens de leurs composants internes simples. Les mots composés sont souvent considérés comme des unités sémantiques et syntaxiques. Cette propriété rend donc indispensable leur recensement pour tenir compte du phénomène du figement dans le domaine du traitement automatique des langues.

Dans cet article, nous utiliserons un lexique construit par une équipe de linguistes du LADL dans les années 1990, le DELACF qui recense plus de 250000 mots composés fléchis (Courtois et al., 1995). Celui-ci est librement disponible (http://infolingu.univ-mlv.fr). Les mots composés recensés ont la propriété d'être des séquences contiguës de mots, ce qui les distingue entre autres des collocations et des phrases figées qui peuvent mettre en relation des mots de manière discontinue : *le* ***record*** *vieux de dix ans a été* ***battu*** *hier, Luc* ***prend*** *ce problème* ***au sérieux*** *dans la discussion*.

Les mots composés appartiennent aux différentes parties-du-discours utilisées dans la langue, comme n'importe quel autre lexème. Dans cet article, nous nous limiterons aux noms, prépositions, conjonctions et adverbes. Les noms composés sont très étudiés car ils sont très nombreux. Ils ont différentes structures nominales de surface: nom+adjectif (*carte bleue*), adjectif+nom (*bon sens*), nom+de+nom (*pomme de terre*), etc. Avec les collocations, ils font l'objet du plus grand nombre d'expériences d'extraction automatique car, en général, ils mettent en relation deux mots pleins, ce qui les rend plus facilement identifiable par les méthodes statistiques. Les prépositions et les conjonctions comprennent, le plus souvent, au plus un seul mot plein entourés d'éléments grammaticaux: par exemple, *au cours de*, *face à* ou *en tant que*, pour les prépositions ; *alors que* ou *pour que* pour les conjonctions. La classe des adverbes est en général plus mixte: *demain matin* est formé de deux noms ; *dès lors* est formé de deux éléments grammaticaux ; *par exemple* est formé d'une préposition et d'un nom.

Notre expérience de traduction des mots composés du DELACF implique de travailler également sur la langue cible qui est l'anglais. Il existe aussi un dictionnaire de mots composés moins conséquent que le DELACF et comprenant quasiment exclusivement des noms (ex. *inland waterways*). Nous l'utilisons également pour améliorer la finesse linguistique de l'analyse.

2.2 Identification des mots composés

Les mots composés du DELACF peuvent être repérés dans des textes à l'aide des fonctionnalités d'Unitex (Paumier, 2003), une plateforme linguistique basée sur des ressources lexicales à grande échelle. Cependant, leur identification est réalisée sans contexte, ce qui cause un bruit non négligeable. Par exemple, le connecteur composé *sur ce* est identifié de manière erronée, alors qu'il appartient juste à un groupe prépositionnel du type *sur Det N*.

> Nous avons travaillé **sur ce** thème pendant un an **au sein de** la commission des **libertés publiques**.

Nous décidons de contextualiser la reconnaissance des mots composés en utilisant le chunker POM (Blanc et al. 2007) intégrant le DELACF notamment. POM identifie les constituants non récursifs simples ou *chunks* (Abney, 1991). Ainsi, *sur ce* dans l'exemple précédent ne pourra pas être considéré

Neveu F., Muni Toke V., Durand J., Klingler T., Mondada L., Prévost S. (éds.)
Congrès Mondial de Linguistique Française - CMLF 2010
978-2-7598-0534-1, Paris, 2010, Institut de Linguistique Française

Traitement automatique des langues
DOI 10.1051/cmlf/2010255

comme un connecteur mais une sous-partie du chunk prépositionnel *sur ce thème*. Les chunks identifiés ont la propriété d'intégrer les mots composés. Ainsi, les séquences *au sein de la commission* et *des libertés publiques* sont deux chunks prépositionnels car *au sein de* est considéré comme une préposition et *libertés publiques* est un nom. POM extrait de chacun des chunks reconnus les têtes et les prépositions.

> {S} {nous avons travaillé,travailler.XV+ind+p+1+ppvnom} {sur ce thème,thème.XP+3+m+s+prep=sur+head=thème} {pendant un an,an.XP+3+m+s+prep=pendant+head=an } {au sein de la commission,commission.XP+3+f+s+prep=**au_sein_de**+head=commission} {des libertés publiques,libertés publiques.XP+3+f+p+prep=du+head=**libertés_publiques**} . {S}

À partir du texte annoté en chunks, il est alors possible d'identifier les mots composés :

> Nous avons travaillé sur ce thème pendant un an **au_sein_de** la commission des **libertés_publiques**.

Le chunker POM ne fonctionnant pas encore pour l'anglais, les mots composés anglais sont reconnus à l'aide des fonctionnalités d'Unitex. Comme le dictionnaire anglais de mots composés comporte quasi exclusivement des noms, le bruit est très limité, comparé au français.

3 Le corpus de travail

3.1 Le corpus Europarl

Pour notre travail, nous avons besoin d'un corpus parallèle multilingue assez large pour permettre d'obtenir des informations statistiques pertinentes. Notre choix s'est donc porté sur le corpus Europarl (Koehn, 2003). Ce corpus parallèle librement disponible sur Internet provient des actes du Parlement Européen et inclut des versions en 11 langues européennes : français, italien, espagnol, portugais, anglais, néerlandais, allemand, danois, suédois, grec et finnois. Chaque langue comprend environ 1 million de phrases, qui contiennent de l'ordre de 28 millions de mots. Europarl est en général considéré comme un corpus spécialisé pour deux raisons : la structuration du discours est très formatée ; le corpus fourmille de termes spécialisés. Malgré cela, il est intéressant pour notre étude car les phrases utilisées ont des structurations syntaxiques très variées et il existe un grand nombre de mots du langage général. Ainsi, l'application du DELACF qui contient dans sa très grande majorité des mots du langage général permet de repérer un nombre tout à fait raisonnable de mots composés: environ 1 mot composé identifié par phrase (cf. sous-section suivante).

3.2 Caractéristiques de notre corpus

Pour notre travail, nous nous basons sur un sous-corpus d'Europarl (une partie de l'année 2001) d'un peu moins d'un million de mots par langue[1]. Notre paire de langues est le français et l'anglais. Nous alignons les deux corpus correspondant aux deux langues par phrases à l'aide des outils disponibles sur le site d'Europarl. Nous prétraitons ensuite notre corpus de travail en identifiant les mots composés en français et en anglais (cf section 2.2). Les tableaux 1 et 2 représentent quelques caractéristiques statistiques sur le corpus et en particulier sur les mots composés du DELACF, identifiés automatiquement. On s'aperçoit que les différents mots composés n'ont pas de distribution homogène. La très grande majorité d'entre eux a tendance à apparaître très rarement (plus de la moitié n'apparaissent qu'une seule fois). Ils ont ainsi un comportement assez conforme à la loi de Zipf.

	Français	Anglais

Neveu F., Muni Toke V., Durand J., Klingler T., Mondada L., Prévost S. (éds.)
Congrès Mondial de Linguistique Française - CMLF 2010
978-2-7598-0534-1, Paris, 2010, Institut de Linguistique Française

Traitement automatique des langues
DOI 10.1051/cmlf/2010255

Nombre de phrases	33212	33212
Nombre de token-mots	934139	853731
Nombre de mots	879844	832687
Nombre de mots composés	36565	20126
Nombre de mots composés différents	5885	4081
Nombre de mots composés par phrase	1.1	0.6
Pourcentage de mots composés dans le corpus	4.2%	2.4%

Tableau 1 : quelques chiffres sur le corpus

Fréquence d'un mot composé	1	2	3	4	5-9	10-19	20-99	100+
Pourcentage de mots composés différents	52.6	16.1	7.8	4.6	9.5	4.5	4.2	0.7

Tableau 2 : distribution des mots composés

4 Extraction de traductions au moyen des modèles IBM

4.1 L'aligneur Giza++

Giza++ (Och et Ney., 2003) est un outil statistique très populaire dans la communauté qui permet d'aligner mot à mot les phrases correspondantes dans un corpus parallèle bilingue. Il sert notamment à apprendre des modèles probabilistes pour la traduction automatique. Giza++ se base sur les modèles IBM de 1 à 5 (Brown et al., 1993). Pour notre travail, nous utilisons cet outil pour extraire les traductions des mots composés en exploitant les alignements des mots simples et composés produits par l'outil sur le sous-corpus d'Europarl (section 4). Nous nous servons également des probabilités de traduction d'un mot en français vers un autre en anglais, apprises par l'outil au moyen du modèle IBM-1. Elles seront utilisées par la méthode d'extraction de traductions basée sur des mesures d'associativité (section 5).

4.2 Alignement direct basé sur modèles probabilistes IBM

Une méthode basique consiste à aligner mot à mot les phrases du corpus parallèle, en considérant les mots composés comme des mots simples. Pour cela, il suffit d'utiliser un aligneur mot à mot du type *Giza++* sur un corpus parallèle où les mots composés auront été identifiés au préalable (marqués en gras dans l'exemple ci-dessous) :

> fr**: Ces derniers se retrouvent maintenant **au_sein_de** la convention qui prépare la Charte mais comme un partenaire parmi d autres qui servira de caution à un travail qu il n aura pas maîtrisé
>
> en**: They are now involved **in** the Convention to draft the Charter but only as one partner among others which will guarantee an imperfect job

Neveu F., Muni Toke V., Durand J., Klingler T., Mondada L., Prévost S. (éds.)
Congrès Mondial de Linguistique Française - CMLF 2010
978-2-7598-0534-1, Paris, 2010, Institut de Linguistique Française

Traitement automatique des langues
DOI 10.1051/cmlf/2010255

Théoriquement, un mot composé étant une unité élémentaire, c'est la méthode la plus intuitive. Or, du fait de la distribution des mots composés dans les corpus (cf. section 3.2), l'apprentissage s'avère difficile. En effet, un mot composé donné apparaît peu souvent. Donc les méthodes purement statistiques ont du mal à apprendre leurs comportements.

4.3 Alignement basé sur les modèles IBM et les composants simples

En pratique, une partie des mots composés se traduisent à partir de leurs composants simples. En effet, certains d'entre eux peuvent se traduire mot à mot. Par exemple, le mot *sources d'énergie* est traduit en *sources of energy*.

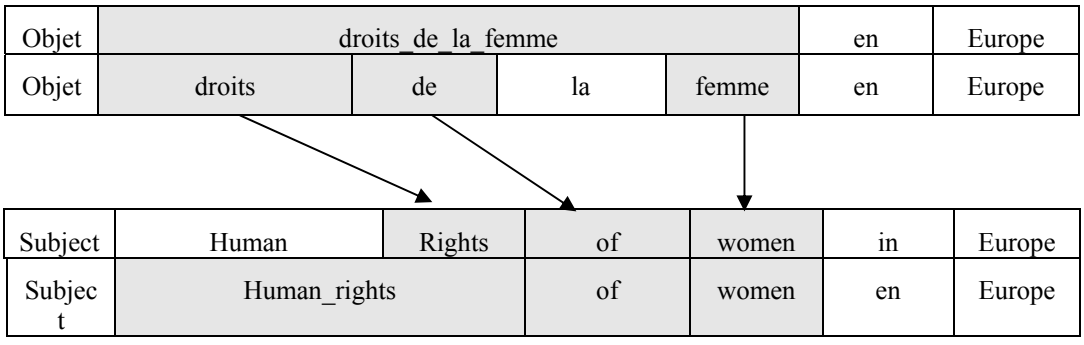
> fr**: Je pense que nous devrions le surmonter et prendre conscience du fait que ces **sources_d_énergie** renouvelable n ont qu un seul ennemi
>
> en**: I think we should put this enmity behind us and acknowledge that these renewable **sources_of_energy** only have one enemy

La traduction de certains mots composés est parfois réalisée moyennant une traduction directe des mots pleins et une restructuration syntaxique standard. Par exemple, pour *systèmes de protection sociale*, les mots pleins sont traduits directement (*système -> systems* ; *protection -> protection* ; *sociale -> social*) ; la structure *nom+adjectif* (respectivement *nom1+de+nom2*) devient *adjectif+nom* (resp. *nom2 nom1*).

> fr**:le document de la présidence insiste également sur le nécessaire renforcement de la convergence sociale et la modernisation de nos **systèmes_de_protection_sociale**
>
> en**:The presidency s document also stresses the necessary reinforcement of social convergence and the modernisation of our **social protection systems**

Étant donné cela, une méthode d'identification des traductions des mots composés est d'aligner les mots simples des phrases parallèles. On considérera alors que la traduction d'un mot composé sera l'union des alignements des mots simples. Cette approche est illustrée dans le graphique 3. Elle a été utilisée dans (Bai et al, 2009) comme *baseline* pour évaluer leur procédé d'extraction de traductions pour les collocations.

En plus, nous décidons de tenir compte des mots composés en langue cible. Ainsi, dans le cas où un composant simple d'un mot composé français est aligné avec un composant simple d'un mot composé anglais, on considérera que le composant simple français est aligné avec le mot composé anglais. Par exemple, dans le graphique 3, le mot simple français *droits* (inclus dans le mot composé *droits de la femme*) est initialement aligné avec le mot simple anglais *rights* lui-même composant simple du nom composé *Human rights*. On considérera que *droits* est aligné avec *Human_rights*



Graphique 3 : alignement du nom composé *droits de la femme*

Neveu F., Muni Toke V., Durand J., Klingler T., Mondada L., Prévost S. (éds.)
Congrès Mondial de Linguistique Française - CMLF 2010
978-2-7598-0534-1, Paris, 2010, Institut de Linguistique Française

Traitement automatique des langues
DOI 10.1051/cmlf/2010255

5 Extraction de traductions avec mesures d'associativité

5.1 Principe

De nombreuses méthodes d'extraction de traductions des collocations utilisent des mesures d'associativité. Ces dernières calculent le degré de corrélation d'un mot ou un groupe de mots en langue source avec un mot ou un groupe de mots en langue cible. La mesure de corrélation la plus populaire car la plus efficace est la mesure de Dice. Elle se base sur la fréquence de cooccurrence d'un mot ou groupe de mots *e* dans une phrase source avec un mot ou groupe de mots *f* dans la phrase cible correspondante. Nous notons *count(e,f)*, cette fréquence de cooccurrence. Celle-ci est normalisée par la somme du nombre d'occurrences de *e* et *f* indépendamment les uns des autres (*count(e)* et *count(f)*).

$$dice(e,f) = \frac{count(e,f)}{count(e) + count(f)}$$

Une méthode classique d'extraction de traductions de collocations est la suivante : pour chaque collocation en langue source, on extrait, dans tout le corpus d'apprentissage, l'ensemble **C** des mots simples en langue cible les plus corrélés à la collocation. Ensuite, pour chaque occurrence de la collocation, on forme une liste de groupes de mots de **C** dans la phrase cible, candidats pour la traduction ; pour chaque collocation, la traduction est alors le groupe candidat de mots le plus corrélé. Dans la suite, nous adaptons cette méthode aux mots composés en nous basant notamment sur Bai et al. (2009).

5.2 Extraction des composants simples candidats

La première étape consiste ainsi, pour chaque mot composé *w*, à sélectionner l'ensemble des mots simples en langue cible les plus corrélés. Cela revient à calculer pour chaque mot *f* en langue cible, son degré de corrélation avec *w* ; puis, à ne garder que les *n* meilleurs qui ont un degré de corrélation supérieur à un certain seuil. La mesure de Dice a souvent démontré sa pertinence pour ce type de tâche. Bai et al. (2009) estiment néanmoins qu'elle a plusieurs défauts. En particulier, les mots composés ont parfois un lien de collocation fort avec leur contexte. Il est donc nécessaire de tenir compte de celui-ci pour calculer le degré de corrélation entre un mot en langue cible et l'expression en langue source. Ainsi, ils ont mis au point le principe de fréquence de corrélation normalisée qui tient compte du contexte dans lequel le mot composé est plongé. En effet, calculer la fréquence de cooccurrence d'un mot *f* en langue cible avec une collocation en langue source, consiste à ajouter 1 lorsque *e* et *f* sont co-occurrents. Or, ce poids de comptage (1) est le même quel que soit le contexte du mot composé. L'idée est que si un mot *f* fait partie de la traduction d'une collocation *w* alors son poids de comptage devrait être plus élevé que pour un autre mot du contexte ne faisant pas partie de la traduction. Le poids de comptage de cooccurrence d'un mot *f* en langue cible avec un mot composé *w* dans le contexte d'une phrase en langue source *E* est le suivant :

$$wcc(f,E,w) = \frac{\sum_{e_i \in w} P(f \mid e_i)}{\sum_{e_i \in E} P(f \mid e_i)}$$

La probabilité *P(f|e)* est la probabilité de traduction de *e* par *f*, apprise par le modèle IBM-1 sur le corpus parallèle d'entraînement avec *Giza++*. Ainsi, la fréquence de corrélation normalisée *NCF* d'un mot *f* par rapport à la collocation *w* est la somme de ses poids de comptage sur l'ensemble du corpus parallèle.

$$NCF(f,w) = \sum_{E \in Corpus} wcc(f,E,w)$$

Neveu F., Muni Toke V., Durand J., Klingler T., Mondada L., Prévost S. (éds.)
Congrès Mondial de Linguistique Française - CMLF 2010
978-2-7598-0534-1, Paris, 2010, Institut de Linguistique Française

Traitement automatique des langues
DOI 10.1051/cmlf/2010255

En appliquant les deux mesures (*Dice* et *NCF*) sur notre corpus, on s'aperçoit rapidement que la fréquence de corrélation normalisée produit de bien meilleurs candidats comme l'avaient montré Bai et al. (2009). Tout d'abord, la mesure de Dice a tendance à écarter les mots grammaticaux, ce qui pose problème pour les prépositions ou les conjonctions qui en sont remplies. Ceci est illustré dans le tableau 4 montrant les meilleurs composants simples, candidats traductions de la préposition composée *au_sujet_de* selon que l'on utilise Dice ou NCF. Par ailleurs, la mesure fonctionne mal pour les mots composés peu fréquents. Elle nécessite en général un nombre minimal d'occurrences du mot dans le corpus afin d'être exploitable. À titre d'exemple, (Smadja et al., 1996) utilisent un seuil minimum de 5 occurrences. Dans le tableau 4, l'exemple de *conflit_armé* est particulièrement frappant.

conflit_armé (#occurrences = 2)			
candidat	dice	candidat	ncf
suppress	0.33	**conflict**	1.73
lebanon	0.31	**armed**	0.96
totalitarian	0.17	ongoing	0.87
rehabilitation	0.10	military	0.60
terrible	0.07		
Armed	0.06		

au_sujet_de (#occurrences = 57)			
Candidat	dice	candidat	ncf
Sahara	0.07	of	22.50
Ethiopia	0.06	on	9.36
macro-	0.06	**about**	9.17
wednesday	0.05	in	7.55
Dioxin	0.05	to	4.21
Offer	0.05	for	3.88

Tableau 4 : exemple de listes de composants candidats

Lors de cette phase, nous devons ajuster deux paramètres afin de limiter la liste des candidats : le nombre maximum de candidats et le score minimal des candidats, que l'on détermine par un ratio du score du meilleur candidat.

5.3 Sélection des traductions candidates

Pour chaque mot composé *w* d'une phrase en français, il s'agit maintenant de sélectionner les possibles traductions dans la phrase correspondante en anglais. La méthode est la suivante : former une liste de traductions candidates à partir de la liste des composants simples candidats puis trier cette liste en fonction de la mesure de Dice calculant le degré de corrélation entre le mot composé et la traduction candidate.

Tout d'abord, pour chaque mot composé français *w*, nous extrayons une première liste de base à l'aide des composants simples candidats calculés dans la phase précédente (section 5.2). Nous nous basons sur l'observation suivante : les traductions des mots composés ont tendance à être des segments de mots contigus. Ainsi, la liste de base contient les plus longs segments de composants simples candidats contigus dans la phrase anglaise. Cette liste est ensuite étendue. Pour tenir compte de la discontinuité des traductions, on fait l'hypothèse que les traductions discontinues concernent en général les noms et sont formées de deux segments assez proches en distance. On rajoute donc les segments formés de deux segments de la première liste, séparés d'une distance maximale fixe (1 ou 2 mots en général). Une fois les segments discontinus ajoutés dans la liste de traductions candidates, on y intègre les différents facteurs des traductions candidates existantes.

Neveu F., Muni Toke V., Durand J., Klingler T., Mondada L., Prévost S. (éds.)
Congrès Mondial de Linguistique Française - CMLF 2010
978-2-7598-0534-1, Paris, 2010, Institut de Linguistique Française

Traitement automatique des langues
DOI 10.1051/cmlf/2010255

Quelques heuristiques simples sont ensuite appliquées afin de filtrer la liste des candidats, à l'aide d'une liste de 34 mots « vides » anglais (*stopwords*). Par exemple, la traduction d'un nom composé ne peut pas commencer ou se terminer par un mot vide ; un mot non vide ne peut apparaître au plus qu'une fois dans le segment candidat[2]. Enfin, on calcule le degré de corrélation de chaque groupe de mots de la liste avec le mot composé *w,* à l'aide de la formule de Dice. La liste est alors triée dans l'ordre décroissant par rapport à Dice. Par ailleurs, les candidats dont cette mesure ne dépasse pas une valeur minimale, sont supprimés de la liste.

L'exemple ci-dessous et le tableau 5 illustrent la procédure utilisée :

> the second principle is that **of equal opportunities** particularly for men and women as well as the european strategy for employment and the context **of** economic and monetary union

Liste primaire	Facteurs	Liste finale filtrée et triée
of equal opportunities	of equal opportunities	equal opportunities (0.78)
Of	of equal	opportunities (0.39)
	equal opportunities	equal (0.31)
	equal	
	opportunities	
	of	

Tableau 5 : procédure d'extraction des traductions candidates de *égalité des chances*

5.4 Sélection de la meilleure traduction

Il s'agit maintenant de trouver pour chaque paire de phrases, la meilleure traduction de chacun des mots composés identifiés dans la phrase en français, avec la contrainte suivante : les traductions sont toutes disjointes dans la phrase anglaise, c'est-à-dire qu'il ne peut y avoir de recouvrement entre les différentes traductions. Pour cela, on cherche et on annote la meilleure traduction parmi les traductions candidates des différents mots composés de la phrase avec la contrainte que cette traduction ne peut recouvrir une traduction déjà annotée. On supprime ensuite de la liste des candidats les traductions possibles du mot composé que l'on vient de traduire. On continue jusqu'à ce qu'il ne reste plus de candidats possibles.

Pour chaque mot composé, il est alors possible d'établir la liste de ses meilleures traductions en les triant selon leur fréquence dans le corpus. Par exemple,

cas particulier => special case, *particular case*

en réponse à => in response to

d'abord => firstly, *first*, *first of all*

Neveu F., Muni Toke V., Durand J., Klingler T., Mondada L., Prévost S. (éds.)
Congrès Mondial de Linguistique Française - CMLF 2010
978-2-7598-0534-1, Paris, 2010, Institut de Linguistique Française

Traitement automatique des langues
DOI 10.1051/cmlf/2010255

6 Evaluation

6.1 Corpus d'évaluation

Pour évaluer les différentes méthodes d'extraction de traductions, nous avons utilisé un corpus de référence (REF) où les traductions des mots composés des phrases en français sont annotées dans les phrases anglaises correspondantes. Le corpus d'évaluation comporte 1002 phrases contenant 26422 mots français et 25082 mots anglais. Il inclut un total de 998 mots composés, dont 532 différents.

Ce corpus de référence a été construit semi-automatiquement de la manière suivante : nous avons d'abord appliqué la deuxième méthode par *Giza++* de repérage des traductions dans les différentes phrases (cf. section 4.2). Des annotateurs[3] ont ensuite vérifié et post-édité manuellement (en cas d'erreur) les occurrences des mots composés français identifiés et pour chacun d'eux, la traduction repérée. Ils ont couvert la même part du corpus et les annotations obtenues concordaient à environ 85%. La principale source de désaccord était la traduction des prépositions composées qui sont souvent traduites de manière détournée, par exemple via des reformulations libres des phrases traduites, ce qui provoque des difficultés dans l'annotation. Les erreurs d'inattention forment l'autre part des erreurs. Les désaccords ont été systématiquement analysés et ont été résolus après discussion.

Ainsi, le corpus contient une séquence de phrases alignées. Chaque paire de phrases a le format suivant :

> <sentence id=31378 lang=fr> je voudrais vous garantir quoi qu il en soit que la Commission assumera ses propres responsabilités dans le développement d une **<1>politique_intérieure__noun</1>** commune de l **<2>Union_européenne__noun</2>** **<3>en_matière_d___prep</3>** immigration et d asile comme le prévoient les conclusions de Tampere</sentence>

> <sentence id=31378 lang=en> Let me assure you anyway that the Commission will take its own responsibilities in the development of a common **<1>internal</1>** **<2>European_Union__O</2>** immigration and asylum **<1>policy</1>** as foreseen in the conclusions of Tampere</sentence>

Dans cet exemple, il y a trois mots composés reconnus en français qui ont chacun un numéro de balise (*politique intérieure* : 1, *Union européenne* : 2, *en matière d'* : 3). Leurs traductions sont balisées avec le même numéro. Ainsi, on observe trois cas :

(1) Il existe une traduction qui est une séquence de mots contigus : *Union européenne -> European Union*

(2) Il existe une traduction qui est une séquence discontinue de mots : *politique intérieure -> internal ... policy*. Dans ce cas-là, le balisage se fait de la manière suivante : *<1>internal</1> .... <1>policy</1>*

(3) Il n'y a pas de traduction dans le segment de texte : *en matière d'*

6.2 Expérience

L'expérience que nous avons réalisée a consisté à comparer les deux types de méthodes décrites dans les sections 4 et 5. Nous les avons implémentées en *Python*. Pour les méthodes utilisant les résultats des alignements de *Giza++,* nous avons mis en place deux modules : (1) un module d'alignement direct des mots composés [GIZA-C], (2) un module d'alignement des mots composés au moyen des alignements de leurs composants simples [GIZA-S]. Pour la méthode utilisant les mesures d'associativité [ASSOC], il a été nécessaire d'ajuster, par la pratique, les quatre paramètres vus dans la section 5, en appliquant l'outil sur un corpus de développement de 400 phrases (disjoint du corpus d'évaluation). Le paramétrage optimal trouvé est le suivant : un nombre maximal de composants simples candidats de 20, un seuil minimal de 10% du score du meilleur composant candidat, une distance maximale entre deux segments d'une même

Neveu F., Muni Toke V., Durand J., Klingler T., Mondada L., Prévost S. (éds.)
Congrès Mondial de Linguistique Française - CMLF 2010
978-2-7598-0534-1, Paris, 2010, Institut de Linguistique Française

Traitement automatique des langues
DOI 10.1051/cmlf/2010255

traduction de 2 mots, un score minimal des traductions de 0.1. Chaque module est appliqué sur le corpus de travail et le corpus d'évaluation non annoté. Il produit un nouveau corpus annoté correspondant au corpus d'évaluation. Nous le notons HYP (pour corpus d'hypothèses). Il génère également, pour les différents mots composés, la liste des traductions triées selon leur fréquence dans le corpus de travail.

Nous avons ensuite mis en place une série de mesures :

- le **taux de précision** d'identification des traductions des mots composés dans le corpus : nombre de traductions sur le nombre de mots composés identifiés automatiquement qui se trouvent aussi dans le corpus de référence [PREC]
- le **taux de recouvrement** entre les traductions balisées dans REF et HYP : la proportion de composants simples en commun dans les traductions de REF et HYP par rapport au nombre de mots simples [REC]
- le **taux de rappel** des traductions des différents mots composés : la proportion de traductions des différents mots composés extraits de REF qui se trouvent aussi dans les listes de traductions extraites automatiquement depuis le corpus de travail. [COUV]
- Une variante du taux de rappel qui consiste à mesurer la proportion de traductions des différents mots composés extraits de REF qui se trouvent aussi dans les listes des *n* meilleures traductions extraites automatiquement (n=1,2,3) [COUV-n][4].

Nous avons aussi calculé la proportion de mots composés français mal identifiés automatiquement. Elle est relativement modérée : environ 4%. Désormais, nous ne travaillons qu'avec les mots composés français en commun entre REF et HYP.

6.3 Résultats et discussion

Les résultats obtenus sont synthétisés, en pourcentage, dans les tableaux 6 et 7.

Méthode	PREC	REC	COUV	COUV-1	COUV-2	COUV-3
GIZA-C	55.7	67.7	59	45	54	56
GIZA-S	**67.7**	**82.1**	**71**	**55**	**68**	**69**
ASSOC	62.8	75.3	66	52	62	64

Tableau 6 : résultats globaux

Neveu F., Muni Toke V., Durand J., Klingler T., Mondada L., Prévost S. (éds.)
Congrès Mondial de Linguistique Française - CMLF 2010
978-2-7598-0534-1, Paris, 2010, Institut de Linguistique Française

Traitement automatique des langues
DOI 10.1051/cmlf/2010255

	Répartition (%)	PREC (ASSOC)	PREC (GIZA-S)
Adverbe	8	**65**	63
Conjonction	5	45	**63**
Nom	65	**74**	73
Préposition	22	32	**56**

Tableau 7 : résultats par catégorie grammaticale

La première remarque qui vient à l'esprit est la qualité relativement moyenne du repérage et de l'extraction des traductions des mots composés, lorsque l'on compare avec ce qui se fait pour les collocations. Les méthodes utilisées étant, à peu de chose près, les mêmes, les problèmes semblent avant tout provenir de la nature des expressions à traduire. L'abondante présence des mots grammaticaux au sein des mots composés est la principale source d'erreurs, en particulier pour les prépositions (ex. *in **spite of***, ***construction** of **Europe***).

On constate également que GIZA-S est, de loin, la meilleure méthode d'extraction des trois. Cependant, ce résultat doit être nuancé car on s'aperçoit que pour les adverbes et les noms, la méthode ASSOC dépasse GIZA-S. Les performances de ASSOC s'écroulent pour les prépositions et conjonctions. Ce n'est d'ailleurs pas une surprise car il est bien connu que les mesures associatives capturent mal le comportement des mots grammaticaux. Du coup, les méthodes utilisant ces mesures sont grandement désavantagées dans les expressions contenant une majorité de tels mots, comme c'est le cas pour les conjonctions et les prépositions. Une méthode comme GIZA-S limite les dégâts grâce à l'utilisation du contexte de la phrase entière. Cette analyse des résultats par catégorie grammaticale est extrêmement intéressante : elle montre que l'on pourrait améliorer les résultats si l'on utilisait la méthode GIZA-S pour les prépositions et les conjonctions et la méthode ASSOC pour les noms et les adverbes.

7 Conclusion et perspectives

Dans cet article, nous avons implémenté deux types de méthodes statistiques pour identifier et extraire des traductions de mots composés du dictionnaire DELACF : (1) utilisation du résultat d'alignements mot à mot et (2) utilisation de mesures associatives. Les mots composés ont la propriété de contenir une quantité non négligeable de mots grammaticaux, ce qui altère les performances des outils implémentés. Nous avons néanmoins vu qu'une combinaison des deux approches pouvait être une piste d'étude pour la suite. La relative petitesse du corpus de travail pourrait être une autre raison de la qualité moyenne obtenue. Nous pensons donc augmenter de manière conséquente le corpus de travail. Nous souhaitons également utilisé un lexique de mots composés plus adapté à Europarl pour permettre d'avoir une meilleure couverture de ce type d'unités dans le corpus. Par ailleurs, nous n'avons pas abordé le cas d'approches plus linguistiques utilisant des analyseurs syntaxiques. Nous pourrions nous inspirer en partie de (Seretan et Wehrli, 2007), en faisant l'hypothèse que la structure syntaxique d'un mot composé et celle de sa traduction ne sont pas très éloignées dans les langues européennes, ce qui limiterait les candidats de traductions.

Neveu F., Muni Toke V., Durand J., Klingler T., Mondada L., Prévost S. (éds.)
Congrès Mondial de Linguistique Française - CMLF 2010
978-2-7598-0534-1, Paris, 2010, Institut de Linguistique Française

Traitement automatique des langues
DOI 10.1051/cmlf/2010255

8 Remerciements

Nous souhaitons remercier ardemment les relecteurs anonymes pour leurs remarques pertinentes ayant contribué à améliorer l'article.

9 Bibliographie

Abney S. 1991. Parsing by Chunks, *In* Robert Berwick, Steven Abney and Carol Tenny (eds.), *Principle-Based Parsing*, Kluwer Academic Publishers, Dordrecht

Bai, M. H., J. M. You, K. J. Chen et J.S. Chang. 2009. Acquiring Translation Equivalences of Multiword Expressions by Normalized Correlation Frequencies, Proceedings of the Conference on Empirical Methods in Natural Language Processing, p. 478-486, Singapour

Blanc, O., M. Constant et P. Watrin. 2007. Segmentation in super-chunks with a finite-state approach. Dans *Proceedings of the Workshop on Finite-State Methods and Natural Language Processing*, Potsdam

Brown P. F., A. Stephen, A. Della Pietra, V. J. Della Pietra, R. L. Mercer. 1993. The Mathematics of Statistical Machine Translation: Parameter estimation. Computational Linguistics, 19(2), p. 263-311

Courtois, B., M. Garrigues, G. Gross, M. Gross,R. Jung, M. Mathieu-Colas, A. Monceaux, A. Poncet-Montange, M. Silberztein et R. Vivès. 1997. *Dictionnaire électronique DELAC : les mots composés binaires*. Rapport technique 56, LADL, Université Paris-7

Caseli, H., C. Ramisch, M. Nunes et A. Villavicencio. 2009. Alignment-based extraction of multiword expressions. *Language resources and evaluation*, Springer

Koehn, P. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the Tenth machine Translation Summit* (MT Summit X), p. 79-86, Phuket, Thailand

Lü Y. et M. Zhou. 2004. Collocation translation acquisition using monolingual corpora. Dans *Proceedings of the 42nd Meeting of the Association for Computational Linguistics* (ACL'04), p. 167-174, Barcelona, Spain

Och, F. J. et H. Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1), p. 19-51

Paumier, S. 2008. *Unitex documentation*. http://igm.univ-mlv.fr/~unitex

Sag, I. A., T. Baldwin, F. Bond, A. Copestake et D. Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. Dans *Proceedings of the third International Conference on Intelligent Text Processing and Computational Linguistics* (CICLING 2002), p. 1-15, Mexico City

Seretan, V. et E. Wehrli. 2007. Collocation translation based on sentence alignment and parsing. Dans *actes de* la *conférence sur le Traitement Automatique des Langues Naturelles*, Toulouse

Smadja F., McKeown K et Hatzivassiloglou 1996. Translating collocations for bilingual lexicons: a statistical approach. *Computational Linguistics*, 22(1), p. 1–38.

[1] Notre corpus est limité en taille du fait de contraintes matérielles au moment des expériences.

[2] Ces heuristiques ne fonctionnent clairement pas pour les mots composés *de mieux en mieux* et *crème de la crème*.

[3] Les annotateurs sont au nombre de trois (deux docteurs et un doctorant en linguistique) et font partie des auteurs.

[4] On notera que COUV-n est équivalent à COUV lorsque *n* est égal au nombre maximum de traductions par mot composé.