Extraction de lexiques bilingues à partir de corpus comparables par combinaison de représentations contextuelles

Amir HAZEM Emmanuel MORIN
LINA - UMR CNRS 6241, 2 rue de la houssinière, BP 92208, 44322 Nantes Cedex 03
amir.hazem@univ-nantes.fr, emmanuel.morin@univ-nantes.fr

RÉSUMÉ
La caractérisation du contexte des mots constitue le cœur de la plupart des méthodes d'extraction de lexiques bilingues à partir de corpus comparables. Dans cet article, nous revisitons dans un premier temps les deux principales stratégies de représentation contextuelle, à savoir celle par fenêtre ou sac de mots et celle par relations de dépendances syntaxiques. Dans un second temps, nous proposons deux nouvelles approches qui exploitent ces deux représentations de manière conjointe. Nos expériences montrent une amélioration significative des résultats sur deux corpus de langue de spécialité.

ABSTRACT
Bilingual Lexicon Extraction from Comparable Corpora by Combining Contextual Representations

Word's context characterisation constitute the heart of most methods of bilingual lexicon extraction from comparable corpora. In this article, we first revisit the two main strategies of context representation, that is : the window-based and the syntactic based context representation. Secondly, we propose two new methods that exploit jointly these different representations . Our experiments show a significant improvement of the results obtained on two different domain specific comparable corpora.

MOTS-CLÉS : Multilingualisme, corpus comparables, lexique bilingue, vecteurs de contexte, dépendances syntaxiques.

KEYWORDS: Multilingualism, comparable corpora, bilingual lexicon, context vectors, syntactic dependencies.

1 Introduction

Les lexiques bilingues sont une ressource importante pour différentes applications relevant du traitement automatique des langues comme en traduction assistée par ordinateur ou en recherche d'information inter-langue. Bien que les travaux s'appuyant sur des corpus parallèles [1] aient montré de très bons résultats, ce type de corpus reste difficile à collecter (Fung et Yee, 1998) et

1. Un corpus parallèle est un ensemble de textes accompagnés de leurs traductions dans une ou plusieurs langues (Bowker et Pearson, 2002).

plus particulièrement quand il s'agit de traiter des corpus spécialisés ou des couples de langues rares ou moins usitées (Morin *et al.*, 2004). L'exploitation des corpus comparables [2] a marqué un tournant dans la tâche d'extraction de lexiques bilingues, et suscite un intérêt constant depuis le milieu des années 1990 grâce à l'abondance et la disponibilité de tels corpus (Rapp, 1995; Fung, 1995; Rapp, 1999; Déjean *et al.*, 2002; Gaussier *et al.*, 2004; Morin *et al.*, 2004; Laroche et Langlais, 2010). L'essor du Web ayant sensiblement facilité la collecte de grandes quantités de données multilingues, les corpus comparables se sont naturellement imposés comme une alternative aux corpus parallèles. Ils ont donné lieu à plusieurs travaux dont le dénominateur commun est l'hypothèse selon laquelle les mots qui sont en correspondance de traduction, ont de grandes chances d'apparaître dans les mêmes contextes (Rapp, 1999). Cette hypothèse découle directement de la proposition souvent citée de Firth (1957) : *« On reconnaît un mot à ses fréquentations »* [3].

Rapp (1995) et Fung (1995) ont été les premiers à introduire les corpus comparables. Ils se sont appuyés sur l'idée de caractérisation du contexte des mots, contrairement aux travaux s'appuyant sur les corpus parallèles, qui eux se basaient sur des informations positionnelles. En 1998, Fung (1998) a introduit la méthode directe, reprise dans de nombreux travaux, notamment ceux de (Rapp, 1999). Dans cette méthode, la traduction d'un mot comporte plusieurs étapes. Le mot est tout d'abord caractérisé par un vecteur représentatif de son contexte. Puis, ce vecteur est traduit dans la langue cible à l'aide d'un dictionnaire aussi appelé lexique de transfert ou lexique pivot. Enfin, il reste à comparer ce vecteur avec tous les vecteurs de contexte des mots de la langue cible, et en extraire les n plus proches comme traductions candidates. Par la suite, une partie des travaux a porté sur l'adaptation et l'amélioration de cette méthode à différents types de corpus (corpus de langue générale ou de spécialité), et à différentes langues et différents types de termes (termes simples, termes complexes, collocations, etc.) (Déjean et Gaussier, 2002), (Morin et Daille, 2004). De nouvelles méthodes ont également été proposées telles que l'approche par similarité interlangue (Déjean et Gaussier, 2002), l'utilisation de l'Analyse en Composantes Canoniques (CCA) (Haghighi *et al.*, 2008). Récemment, Li et Gaussier (2010) et Li *et al.* (2011) se sont intéressés à l'aspect inverse qui consiste à améliorer la comparabilité des corpus comparables afin d'augmenter l'efficacité des méthodes d'extraction de lexiques bilingues.

La plupart des travaux utilisant les corpus comparables ont comme dénominateur commun le contexte, qui représente le cœur de l'extraction lexicale bilingue. La question principale à se poser est alors la suivante : étant donné un mot quelconque, comment choisir les mots qui caractérisent au mieux son contexte ? Selon l'état de l'art, le contexte d'un mot donné est habituellement représenté par les mots faisant partie de son environnement, c'est-à-dire, les mots qui l'entourent. Ces mots sont extraits, soit à l'aide d'une fenêtre contextuelle (Rapp, 1999; Déjean et Gaussier, 2002), soit à l'aide des relations de dépendances syntaxiques (Gamallo, 2007). L'un des problèmes sous-jacent au contexte extrait à l'aide des fenêtres contextuelles est le choix de la taille des fenêtres. Celle-ci est habituellement fixée empiriquement, et bien que différentes études aient montré une tendance à choisir des fenêtres de petite taille quand il s'agit de caractériser des mots fréquents, et des fenêtres de grande taille quand il s'agit de caractériser des mots peu fréquents (Prochasson et Morin, 2009), cela reste imprécis car il n'y a toujours pas de méthode dite optimale pour le choix de la taille de la fenêtre contextuelle. Quant aux relations de dépendances syntaxiques, leur efficacité est très sensible à la taille des corpus, et bien que cette

2. Un corpus comparable est une collection de documents multilingues produits généralement à la même période et traitant des mêmes sujets.

3. *« You shall know a word by the company it keeps »*

représentation soit plus intéressante d'un point de vue sémantique, elle atteint ses limites lorsqu'il s'agit de traiter des corpus de petite taille. Une proposition, qui vient naturellement à l'esprit consiste à utiliser conjointement ces deux représentations afin de tirer profit de leurs avantages respectifs. Une première approche exploitant les deux représentations proposée par Andrade *et al.* (2011) combine quatre modèles statistiques et compare les dépendances lexicales pour identifier les traductions candidates. Dans cet article, nous proposons une autre manière de combiner les deux précédentes représentations contextuelles, partant de l'intuition que cette combinaison permettrait un lissage du contexte en prenant en compte deux informations complémentaires qui sont : (i) l'information globale véhiculée par la représentation par fenêtre contextuelle et (ii) une information sémantique plus fine apportée par les relations de dépendances syntaxiques. L'objectif étant d'améliorer la représentation contextuelle et les performances de l'extraction de lexiques bilingues à partir de corpus comparables.

Dans la suite de cet article, nous présentons en section 2 les deux principales stratégies de représentations contextuelles. La section 3 décrit ensuite nos deux approches de combinaison de contextes. La section 4 se concentre sur l'évaluation des méthodes mises en œuvre. Nous terminons enfin par une discussion en section 5 et une conclusion en section 6.

2 Construction de contextes

2.1 Cooccurrences graphiques

Le contexte par sac de mots consiste simplement à collecter des mots entourant un mot donné, sans règles précises hormis le choix du nombre de mots à sa gauche et à sa droite, appelé aussi fenêtre contextuelle. Soit la phrase suivante : *«(...) Pour les cas traités pour danger ostéoporotique les densitométries osseuses comparatives ont montré une amélioration sous THS (...)»*.

Pour le terme *ostéoporotique*, si nous choisissons une fenêtre contextuelle de taille 5, c'est-à-dire deux mots à gauche et deux mots à droite de celui-ci. Le contexte de *ostéoporotique* sera : *traités*, *danger*, *densitométries* et *osseuses*. Ce processus est répété autant de fois que le terme *ostéoporotique* apparaît dans un corpus donné. Cette technique de représentation du contexte a montré son efficacité surtout lorsqu'il s'agit de mots très fréquents. Intuitivement, nous pouvons nous dire que tous les mots entourant un mot donné n'ont pas la même importance et qu'il serait parfois utile de ne pas tous les considérer de la même manière. Cependant, toute la difficulté réside dans la prise de décision concernant tel ou tel mot. Brosseau-Villeneuve *et al.* (2010) proposent une méthode de pondération des mots du contexte selon leur position pour la tâche de désambiguïsation du sens des mots. Une autre méthode pour pallier cette difficulté consiste en l'utilisation des relations de dépendances syntaxiques entre les mots que nous présentons dans la section suivante.

2.2 Cooccurrences syntaxiques

Afin de mieux représenter le contexte d'un mot, plusieurs travaux se sont intéressés aux relations de dépendances syntaxiques (Gamallo, 2008a; Garera *et al.*, 2009). L'idée n'est plus de représenter le contexte seulement par les mots avoisinants mais de rajouter une information supplémentaire

qui spécifie le type de relation syntaxique entre les mots. Une relation de dépendance est une relation binaire asymétrique entre un mot appelé tête ou parent (Head or parent) et un modificateur ou dépendant (modifier or dependant). Les relations de dépendances forment un arbre qui inter-connecte tous les mots d'une phrase. Un mot dans une phrase peut avoir plusieurs modificateurs mais chaque mot ne peut modifier qu'au plus un seul mot (Lin, 1998). La racine de l'arbre de dépendance aussi appelée Head, ne modifie aucun mot de la phrase. Une liste de tuples est utilisée pour représenter un arbre de dépendances : ([word], [category], [head], [relationship]) avec :

– word : est le mot représenté dans le nœud de l'arbre ;
– category : constitue la catégorie lexicale du mot (word) ;
– head : spécifie quel mot est modifié par word ;
– relationship : est une étiquette attribuée à la relation de dépendance (subj pour subject, spec pour specifier, etc.).

En outre, le signe « < » signifie précédent et « > » signifie successeur.
Pour la phrase suivante : *« I have a brown dog »*, l'arbre de dépendance serait celui donné en table 1 :

Modificateur	**Catégorie**	**Head**	**Type**
I	Noun	< have	subj
have	Verb	-	-
a	Det	< dog	spec
brown	Adj	< dog	adjn
dog	Noun	> have	comp

TABLE 1 – Exemple de relations de dépendances syntaxiques

Pour plus de détails concernant les dépendances syntaxiques et plus particulièrement pour les tâches de désambiguisation de mots et de résolution des dépendances, se rapporter à Gamallo (2008b). Dans Gamallo (2007), trois notions élémentaires de dénotation sont abordées :

– Les mots lexicaux ;
– Les dépendances syntaxiques (sujet, relation d'objet direct, relation prépositionnelle entre deux noms, relation prépositionnelle entre un verbe et un nom, etc.) ;
– Les modèles lexico-syntaxiques qui consiste à combiner les mots et leurs catégories syntaxiques en terme de dépendance (Noun+ subj + Verb).

Les mots lexicaux représentent des ensembles de propriétés {Noun, Verb, Adj, Adv ... } alors que les dépendances et les modèles lexico-syntaxiques sont définis comme des opérations sur ces ensembles. Une dépendance est une relation binaire qui prend en entrée deux ensembles de propriétés et donne en sortie un ensemble plus restreint qui est l'intersection des ensembles données en entrée. Nous retrouvons sept types de relations de dépendances (Gamallo, 2007) résumés dans la table 2.

Par exemple, pour le mot *recurrence*, il existe une relation Lmod avec l'adjectif *local*. Ainsi dans le processus de construction du contexte de *recurrence*, nous comptabiliserons le nombre de fois où l'adjectif *local* apparaît à gauche de *recurrence* dans le corpus. Nous ferons de même pour les autres relations de dépendances syntaxiques.

Relation	type	Exemple
Lmod	modificateur gauche si relation Adj - Noun	local - recurrence
Rmod	modificateur droite si relation Noun - Adj	number - insufficient
modN	modificateur de Nom si relation Noun - Noun	breast - cancer
Lobj	objet à gauche si relation Noun - Verb	study - demonstrate
Robj	objet à droite si relation Verb - Noun	have - effect
PRP	si relation prépositionnelle Noun-PRP-Noun	malignancy - in - woman
iobj	si relation objet indirecte Verb-PRP-Noun	occur - in - portion

TABLE 2 – Liste des relations de dépendances syntaxiques

2.3 Synthèse

Nous venons de voir deux manières de représenter le contexte, à savoir une représentation graphique (par sac de mots) et une représentation syntaxique (par relations de dépendances syntaxiques). L'intérêt de passer d'une coloration graphique à une coloration syntaxique des mots peut être vu selon deux aspects. Le premier consiste à se dire que l'information véhiculée par une coloration graphique n'est principalement qu'une information quantitative très variable et fortement dépendante des corpus utilisés. D'où l'idée d'abandonner ce type de coloration pour passer à une coloration syntaxique porteuse d'informations qualitatives et idéalement indépendante de la taille des corpus. Le deuxième aspect serait de dire que malgré tout, la coloration graphique a un intérêt et qu'au lieu de s'en écarter il vaudrait peut être mieux la combiner avec la coloration syntaxique afin de tirer le meilleur des deux. C'est notre hypothèse de complémentarité entre les informations qualitatives et quantitatives des mots.

3 Combinaison de contextes

Nous nous positionnons ici dans le cadre de l'amélioration de la méthode directe décrite dans plusieurs travaux dont Fung (1998) et Rapp (1999). Notre démarche vise à montrer que l'exploitation des deux principales représentations contextuelles a un intérêt particulier pour la tâche de constitution de lexiques bilingues. Nous proposons donc deux manières de combiner les contextes (graphique et syntaxique) que nous appellerons : la combinaison *a posteriori* des contextes et la combinaison *a priori* des contextes.

Une première manière de combiner les deux représentations contextuelles est une combinaison *a posteriori*, c'est-à-dire la combinaison des scores renvoyés par la méthode directe selon les deux représentations. La seconde manière consiste en une combinaison *a priori* qui utilise les deux informations contextuelles *a priori* dans un même vecteur pour ensuite appliquer la méthode directe une seule fois sur l'ensemble du corpus.

3.1 Combinaison *a posteriori* des contextes

Dans le domaine de la recherche d'information, la combinaison de plusieurs listes renvoyées par différents moteurs de recherche est souvent utilisée pour améliorer les performances d'un

système de questions/réponses (Aslam et Montague, 2001). Nous partons du principe que chaque représentation du contexte correspond à une méthode bien définie. Nous nous retrouvons donc dans le cas d'une combinaison de deux méthodes bien distinctes. La première est la méthode directe basée sur une représentation graphique et la seconde est la méthode directe basée sur une représentation syntaxique. Une manière classique de fusionner les deux méthodes est de prendre, comme entrée, la sortie de chacune des méthodes citées. Dans notre cas, pour chaque mot à traduire, nous prenons comme entrée une liste de scores retournée par chacune des deux méthodes, puis nous fusionnons les deux listes par une simple combinaison arithmétique des scores. Ceci nous donne une nouvelle liste de mots ordonnés (sachant que les scores fusionnés sont compatibles à partir du moment où nous utilisons la même mesure de similarité pour les deux méthodes). En utilisant les scores comme critère de fusion, nous calculons le score de similarité d'un candidat à la traduction, en sommant les scores qui sont renvoyés par chacune des deux méthodes comme suit :

$$S_{comb}(w) = S_{fen}(w) + S_{rel}(w) \quad (1)$$

où $S_{comb}(w)$ est le score final du mot w, $S_{fen}(w)$ est le score retourné par la méthode directe basée sur une représentation graphique et $S_{rel}(w)$ est le score retourné par la méthode directe basée sur une représentation syntaxique.

Cette équation peut aussi s'écrire comme suit :

$$S_{comb}(w) = (\lambda) \times S_{fen}(w) + (1 - \lambda) \times S_{rel}(w) \quad (2)$$

avec λ comme indice de confiance donné à chaque méthode ($\lambda \in [0, 1]$). Dans notre cas, $\lambda = 0,5$, notre but n'étant pas de trouver la valeur optimale de λ pour obtenir les meilleurs résultats. Différentes expériences ont été menées qui indiquent que les meilleurs résultats sont globalement ceux montrés dans la section 4 avec un $lambda \in [0,5, 0,6]$. Par ailleurs, d'autres méthodes de combinaisons de scores ont été testées comme la combinaison harmonique des rangs et des scores (Morin, 2009), mais la méthode que nous avons choisi (combinaison arithmétique des scores) est celle qui donne les meilleures performances.

3.2 Combinaison *a priori* des contextes

Le vecteur de contexte a pour but d'enregistrer un ensemble d'information sur le contexte d'un mot w donné. Dans le cas de la représentation graphique, ces informations sont les mots qui cooccurrent avec le mot w. Dans le cas d'une représentation syntaxique, ce sont les mots en relation avec w qui sont sélectionnés pour faire partie de son vecteur de contexte. Dans un cadre plus générique, nous pourrions imaginer plusieurs autres sources d'informations à exploiter. Cependant si chaque nouvelle information engendre un nouveau vecteur de contexte, nous pourrions vite être dépassés par le nombre de sources à fusionner. Pour remédier à cela, une autre manière serait de représenter dans un seul vecteur de contexte toutes les informations concernant le mot w. C'est la position adoptée avec la combinaison *a priori* des contextes.

Dans cette technique de combinaison, nous considérons le vecteur de contexte d'un mot comme un descripteur qui contient plusieurs informations pour chaque entrée du vecteur. Dans notre cas, nous avons deux types d'information : (i) une information de cooccurrence globale fournie par la

Représentation graphique	Représentation syntaxique	Combinaison
$regional_{13}$	$regional_{Lmod_2}$	$regional_{13}$, $regional_{Lmod_2}$
$local_5$	$local_{Lmod_1}$	$local_5$, $local_{Lmod_1}$
$oestrogen_1$	-	$oestrogen_1$
$rate_{32}$	$rate_{modN_{29}}$,$rate_{PRPV_3}$	$rate_{32}$, $rate_{modN_{29}}$,$rate_{PRPV_3}$

TABLE 3 – Exemple de la représentation du contexte du mot *recurrence* et du nombre de ses cooccurrences, en fonction des représentations graphique et syntaxique ainsi que de leur combinaison

représentation graphique et (ii) une information plus spécifique fournie par la représentation syntaxique. Si nous prenons par exemple le mot *regional* (représenté dans la table 3), nous pouvons voir qu'il apparaît 13 fois avec le mot *recurrence* selon la représentation graphique et 2 fois comme modificateur gauche (Lmod) selon la représentation syntaxique. La combinaison prend en compte les deux informations, en considérant que le mot *regional* apparaît 13 fois avec *recurrence*, dont 2 fois en tant que modificateur gauche. Une information importante à souligner est que la méthode directe se basant sur les relations de dépendances syntaxiques considère $rate_{modN_{29}}$ et $rate_{PRPV_3}$ par exemple, comme étant deux mots distincts. L'un des avantages de la combinaison *a priori* est que si l'une des méthodes manque une information (un mot), comme nous pouvons le constater avec le mot *oestrogen* par exemple, la fusion permet de pallier ce manque (grâce ici à la représentation graphique). Nous considérons les deux représentations contextuelles comme étant complémentaires. Le but de la combinaison *a priori* est de préserver le classement et renforcer les scores des entrées des vecteurs de contexte afin de lisser les contextes et corriger certaines erreurs qui peuvent apparaître.

Nous illustrons dans les tables 4, 5 et 6 les 10 premières entrées du vecteur de contexte du mot *recurrence* extrait du corpus du cancer du sein, en fonction de trois mesures d'association, à savoir : le taux de vraisemblance (Log), le Odds-Ratio (Odds) et l'information mutuelle (Im). La notation (+/-) indique l'apport positif ou négatif de la combinaison *a priori*. L'indice '+' indique qu'un mot classé dans les 10 premières entrées du vecteur de contexte de la méthode par fenêtre ou par relation de dépendance, conserve son classement dans les 10 premières entrées après combinaison. Le signe '-' en revanche, indique l'apparition d'un mot non classé dans les 10 premières entrées du vecteur de contexte.

w=5		RelDep		Combinaison		+/-
local	818,98	$local_{Lmod}$	618,17	$local_{Lmod}$	936,05	+
rate	119,71	$risk_{PRPN}$	96,02	local	791,15	+
distant	72,62	$rate_{modN}$	68,34	$risk_{PRPN}$	153,14	+
risk	61,00	$tumor_{modN}$	62,82	rate	113,96	+
salvage	39,15	$rate_{PRPN}$	40,18	$rate_{modN}$	110,28	+
year	39,08	$time_{PRPN}$	32,85	$tumor_{modN}$	104,71	+
time	31,84	$disease_{modN}$	28,76	distant	70,23	+
tumor	31,04	$isolated_{Lmod}$	24,29	$rate_{PRPN}$	64,69	+
isolate	30,15	$distant_{Lmod}$	24,28	risk	54,89	+
inoperable	28,16	$patient_{PRPN}$	23,64	$time_{PRPN}$	53,13	+

TABLE 4 – Illustration des 10 premières entrées du vecteur de contexte du mot *recurrence* en fonction du taux de vraisemblance (Log) pour les représentations graphique ($w = 5$) et syntaxique ($RelDep$) ainsi que par la combinaison *a priori*

La table 4 montre que la combinaison *a priori* a un apport positif car, elle engendre un vecteur de contexte qui respecte le classement des méthodes $w = 5$ et $RelDep$ et ceci, grâce à la mesure

d'association du taux de vraisemblance.

w=5		RelDep		Combinaison		+/-
isolated	5,10	$freedom_{PRPN}$	7,83	$freedom_{PRPN}$	8,12	+
geographic	4,62	$heat_{Robj}$	6,72	fat_{PRPN}	7,02	+
adjudication	4,44	$operable_{Rmod}$	6,72	$disappointing_{Rmod}$	7,02	+
conspicuous	4,44	fat_{PRPN}	6,72	$operable_{Rmod}$	7,02	+
reconcile	4,44	$disappointing_{Rmod}$	6,72	$threat_{PRPN}$	7,02	+
liberate	4,44	$threat_{PRPN}$	6,72	$heat_{Robj}$	7,02	+
evade	4,44	$local_{Lmod}$	5,89	$local_{Lmod}$	6,02	+
inoperable	4,38	$fear_{PRPN}$	5,63	$fear_{PRPN}$	5,93	+
quarter	4,29	$suspicion_{PRPN}$	5,63	$suspicion_{PRPN}$	5,93	+
local	4,28	$inoperable_{Lmod}$	5,63	$inoperable_{Lmod}$	5,93	+

TABLE 5 – Illustration des 10 premières entrées du vecteur de contexte du mot *recurrence* en fonction du Odds-Ratio (Odds) pour les représentations graphique ($w = 5$) et syntaxique ($RelDep$) ainsi que par la combinaison *a priori*

La table 5 montre aussi que la combinaison *a priori* a un apport positif en utilisant la mesure d'association du Odds-Ratio. Nous remarquons néanmoins que la combinaison a avantagé la méthode $relDep$, car il n'y a que ses entrées qui sont présentes dans les 10 premières entrées du vecteur de contexte de la méthode de combinaison *a priori*.

w=5		RelDep		Combinaison		+/-
$isolated$	8,73	$local_{Lmod}$	14,77	$local$	16,17	+
$geographic$	8,15	$tumor_{modN}$	13,84	$local_{Lmod}$	15,83	+
$inoperable$	8,00	$risk_{PRPN}$	12,84	$breast$	14,64	-
$local$	7,82	$time_{PRPN}$	12,44	$rate$	14,39	-
$adjudication$	7,73	$distant_{Lmod}$	12,09	$tumor$	14,15	-
$conspicuous$	7,73	$rate_{modN}$	11,91	$cancer$	14,04	-
$reconcile$	7,73	$year_{modN}$	11,80	$risk_{PRPN}$	13,90	+
$liberate$	7,73	$rate_{PRPN}$	11,63	$patient$	13,75	-
$quarter$	7,73	$tumour_{modN}$	11,63	$cancer_{modN}$	13,15	+-
	⋮		⋮		⋮	⋮
$rate$	5,59	$cancer_{modN}$	10,51			
$survival$	4,12		⋮			
$tumor$	3,69					
$patient$	3,21					
$breast$	2,92					
$cancer$	2,28					

TABLE 6 – Illustration des 10 premières entrées du vecteur de contexte du mot *recurrence* en fonction de l'information mutuelle (IM) pour les représentations graphique ($w = 5$) et syntaxique ($RelDep$) ainsi que par la combinaison *a priori*

La table 6 montre que la combinaison *a priori* a un apport négatif pour au moins 5 mots. Ces mots n'étaient pas classés dans les 10 premières entrées des méthodes $w = 5$ et $RelDep$, et le sont devenus grâce à la combinaison *a priori*. Ce constat indique que la mesure d'association de l'information mutuelle n'est pas appropriée car elle ne préserve pas le classement des entrées de $w = 5$ et $RelDep$. Elle affecte des scores élevés à des mots qui avaient des scores faibles comme pour $rate$ ou $cancer$ par exemple, qui passent respectivement de $5,59$ à $14,39$ et de $2,28$ à $14,04$.

Les tables 4, 5 et 6 ont montré que l'utilisation du taux de vraisemblance et du Odds-Ratio dans la méthode de combinaison *a priori* avait un apport positif contrairement à l'utilisation de l'information mutuelle. Ce constat se confirme par les résultats des expériences que nous présentons dans la section suivante.

4 Évaluation

4.1 Ressources linguistiques

Nous avons utilisé deux corpus spécialisés français-anglais, à savoir un corpus du « cancer du sein » d'un million de mots et un corpus « énergies renouvelables » de 600 000 mots. Le corpus du cancer du sein a été extrait à partir du portail Elsevier [4] tel que décrit dans l'article Morin (2009). Concernant le corpus des énergies renouvelables, il a été construit avec le crawler nommé Babook (Groc, 2011). Les deux corpus ont été pré-traités (tokenisés, étiquetés, et lemmatisés). Pour évaluer les différentes approches utilisées dans cet article, nous avons sélectionné 122 couples de mots simples pour le corpus du cancer du sein (à partir du meta-thesaurus UMLS [5] et du *Grand dictionnaire terminologique* [6]) et 100 couples de mots simples pour le corpus des énergies renouvelables (à partir du dictionnaire en ligne *WordReference* [7]). Comme dictionnaire bilingue nous avons utilisé le dictionnaire ELRA-M0033. Concernant l'extraction des relations de dépendances syntaxiques, nous avons utilisé l'outil fournit par Gamallo (2008a) [8].

4.2 Résultats

Nous présentons les résultats des expériences menées sur les deux corpus de langue de spécialité. Nous évaluons la méthode directe basée sur une représentation graphique notée $w = k$, où k correspond à la taille de la fenêtre (k prend les valeurs : 5, 9 et 15). La méthode directe basée sur une représentation syntaxique notée $RelDep$, et nos deux nouvelles approches, c'est-à-dire la combinaison *a posteriori* des contextes notée $Comb_{post}$ (qui combine les scores de $w = k$ et de $RelDep$) et la combinaison *a priori* des contextes notée $Comb_{apri}$ (qui exploite les contextes fournis par une fenêtre contextuelle $w = k$ et les relations de dépendances $RelDep$ conjointement dans un même vecteur, pour ensuite appliquer la méthode directe). La comparaison des quatre méthodes se fait en fonction de la précision pour les tops 1 et 10. Ainsi une précision au top 10 notée $P10$, veut dire que la bonne traduction est présente parmi les 10 candidats renvoyés par la méthode. Nous utilisons aussi la mesure MAP qui renvoie une vision plus globale sur le comportement de chaque méthode (Laroche et Langlais, 2010). Comme la méthode directe est très sensible aux mesures d'association et de similarité utilisées, nous avons choisi les 3 couples de mesures les plus connus dans l'état de l'art, à savoir : le taux de vraisemblance et le Jaccard noté (Log-Jac) (Morin, 2009), le Odds-Ratio et le cosinus noté (Odds-Cos) (Laroche et Langlais, 2010) ainsi que l'information mutuelle et le cosinus noté (Im-Cos) (Gamallo, 2008a). Ainsi, chaque case de la table 7 correspond à une mesure d'association et à une mesure de similarité pour les 4 méthodes testées sur le corpus du cancer du sein. La table 8 concerne le corpus des énergies renouvelables et respecte la même configuration que la première table.

Dans la table 7, nous constatons que pour la configuration Log-Jac et $w = 5$, les deux méthodes de combinaisons proposées obtiennent de meilleurs résultats que $w = 5$ et $RelDep$, avec une MAP de 0,485 pour $Comb_{post}$ et de 0,488 pour $Comb_{apri}$ alors que $RelDep$ et $w = 5$ n'obtiennent

4. www.elsevier.com
5. www.nlm.nih.gov/research/umls
6. www.granddictionnaire.com
7. www.wordreference.com
8. http://gramatica.usc.es/pln/tools/deppattern.html

	Log-Jac			Odds-Cos			Im-Cos		
	P1	P10	MAP	P1	P10	MAP	P1	P10	MAP
$RelDep$	19,67	49,18	0,297	12,29	46,72	0,237	27,05	48,36	0,332
$w = 5$	31,15	63,93	0,416	26,23	59,84	0,380	34,43	57,38	0,431
$Comb_{post}$	36,88	**68,85**	0,485	38,52	63,93	0,473	**41,80**	**61,48**	**0,482**
$Comb_{apri}$	**38,52**	**68,85**	**0,488**	**40,16**	**71,31**	**0,497**	28,69	52,46	0,373
$RelDep$	19,67	49,18	0,297	12,29	46,72	0,237	27,05	48,36	0,332
$w = 9$	31,97	66,39	0,435	21,31	60,66	0,343	20,49	51,64	0,305
$Comb_{post}$	36,07	75,41	0,494	35,25	68,03	0,460	**40,98**	**59,84**	**0,464**
$Comb_{apri}$	**41,80**	**77,05**	**0,536**	**38,52**	**75,41**	**0,492**	16,39	40,16	0,252
$RelDep$	19,67	49,18	0,297	12,29	46,72	0,237	27,05	48,36	0,332
$w = 15$	27,87	62,30	0,387	17,21	53,28	0,302	13,12	40,16	0,226
$Comb_{post}$	**34,43**	70,49	**0,475**	**37,70**	64,75	0,472	**31,97**	**59,02**	**0,412**
$Comb_{apri}$	**34,43**	**72,95**	0,473	**37,70**	**70,49**	**0,482**	13,12	33,61	0,202

TABLE 7 – Précision (%) pour les tops 1 et 10 ainsi que la MAP pour le corpus « Cancer du sein ». Comparaison de l'approche directe par représentation graphique et de celle par représentation syntaxique ainsi que des deux méthodes de combinaisons (les améliorations indiquent une significativité avec un indice de confiance de 0,05 utilisant le test de Student).

que 0,297 et 0,416. Ce même constat peut être fait pour les autres valeurs de w (9 et 15). Ainsi concernant la configuration Log-Jac, les deux méthodes de combinaison proposées obtiennent de meilleurs résultats que les deux représentations contextuelles prises séparément, avec un avantage pour la méthode $Comb_{apri}$ qui obtient une MAP de 0,536 en combinant $RelDep$ avec $w = 9$. Nous pouvons constater que, pour la configuration Odds-Cos, c'est la méthode $Comb_{apri}$ qui obtient les meilleurs résultats avec une MAP de 0,497 pour un $w = 5$. Concernant la configuration Im-Cos, c'est $Comb_{post}$ qui obtient les meilleurs résultats, et $Comb_{apri}$ n'apporte aucune amélioration et dégrade même les résultats dans certains cas. Pour résumer, nous pouvons dire que les deux méthodes proposées améliorent les performances de la méthode directe, avec une efficacité variable étroitement liée aux mesures d'association et de similarité utilisées.

Pour la table 8 concernant le corpus des énergies renouvelables, nous pouvons aussi constater que pour la configuration Log-Jac et $w = 5$, les deux méthodes de combinaisons proposées obtiennent de meilleurs résultats que $w = 5$ et $RelDep$, avec une MAP de 0,365 pour $Comb_{post}$ et de 0,354 pour $Comb_{apri}$ alors que $RelDep$ et $w = 5$ n'obtiennent que 0,257 et 0,272. Globalement, c'est la méthode $Comb_{post}$ qui obtient les meilleurs résultats. Ce que l'on peut retenir des deux tables c'est que $Comb_{post}$ et $Comb_{apri}$ améliorent les résultats pour toutes les combinaisons de mesures sauf pour $Comb_{apri}$ qui ne fonctionne pas avec le couple (Im-Cos).

5 Discussion

Le but de ce travail était dans un premier temps, de comparer les deux principales représentations contextuelles utilisées dans la méthode directe, et dans un second temps de proposer deux

	Log-Jac			Odds-Cos			Im-Cos		
	P1	P10	MAP	P1	P10	MAP	P1	P10	MAP
$RelDep$	18,00	40,00	0,257	09,00	32,00	0,163	11,00	41,00	0,191
$w = 5$	18,00	47,00	0,272	13,00	41,00	0,217	14,00	44,00	0,221
$Comb_{post}$	**28,00**	55,00	**0,365**	**22,00**	**59,00**	**0,335**	**21,00**	**55,00**	**0,321**
$Comb_{apri}$	**28,00**	**56,00**	0,354	20,00	55,00	0,317	09,00	38,00	0,177
$RelDep$	18,00	40,00	0,257	09,00	32,00	0,163	11,00	41,00	0,191
$w = 9$	21,00	42,00	0,270	11,00	36,00	0,194	08,00	31,00	0,152
$Comb_{post}$	**28,00**	**54,00**	**0,358**	**23,00**	**55,00**	**0,334**	**20,00**	**49,00**	**0,289**
$Comb_{apri}$	26,00	52,00	0,350	19,00	**55,00**	0,318	07,00	29,00	0,137
$RelDep$	18,00	40,00	0,257	09,00	32,00	0,163	11,00	41,00	0,191
$w = 15$	12,00	34,00	0,207	06,00	35,00	0,143	03,00	22,00	0,093
$Comb_{post}$	**22,00**	50,00	**0,316**	**20,00**	**52,00**	**0,316**	**13,00**	**42,00**	**0,234**
$Comb_{apri}$	**22,00**	**52,00**	0,311	**20,00**	49,00	0,314	06,00	24,00	0,118

TABLE 8 – Précision (%) pour les tops 1 et 10 ainsi que la MAP pour le corpus « énergies renouvelables ». Comparaison de l'approche directe par représentation graphique et de celle par représentation syntaxique ainsi que des deux méthodes de combinaisons (les améliorations indiquent une significativité avec un indice de confiance de 0,05 utilisant le test de Student).

nouvelles manières de les combiner pour augmenter les performances. La première remarque concerne l'utilisation de la représentation graphique $w = k$. Il est évident que le choix de la taille de la fenêtre joue un rôle important, comme nous avons pu le constater dans les différentes expériences montrées dans les tables 7 et 8. Dans la plupart des cas, ce sont des fenêtres de taille 5 et 9 qui donnent les meilleurs résultats. Ceci montre que la caractérisation du contexte des mots par ceux qui leurs sont très proches semble être la manière la plus adéquate, si l'on se base sur une caractérisation par fenêtre contextuelle. Le fait de choisir des fenêtres de taille plus grande n'améliore pas significativement les résultats dans nos expériences.

La deuxième remarque concerne la méthode par représentation syntaxique $RelDep$. Cette méthode utilisée par Gamallo (2008a) donne dans ses expériences de meilleurs résultats que la méthode par représentation graphique. Cependant dans nos expériences, la méthode $RelDep$ reste globalement en deçà de $w = k$. Ceci s'explique par deux facteurs. Le premier concerne la taille des corpus. Gamallo (2008a) avait utilisé des corpus de très grande taille (10 millions de mots environs) contrairement à nos corpus spécialisés qui sont de petite taille (600 000 et 1 million de mots). Le deuxième facteur, qui est directement lié au premier, concerne la manière de considérer les entrées des vecteurs de contexte de la méthode $RelDep$. Si dans le vecteur de contexte d'un mot X, il existe un mot Y avec une relation $Lmod$ de X avec un score $S_{Y_{Lmod}}$ et une autre relation $Robj$ avec un score $S_{Y_{Robj}}$, alors dans ce vecteur de contexte Y_{Lmod} et Y_{Robj} sont considérés comme étant deux mots différents, bien que ce soit le même mot avec deux relations de dépendances distinctes, ce qui rend la méthode $RelDep$ plus sensible aux petits corpus que $w = k$. Ceci explique les performances de la méthode de combinaison *a priori* des contextes. En effet, la méthode $Comb_{apri}$ comble le manque de la méthode $RelDep$, car elle considère les deux informations véhiculées par les deux représentations contextuelles. Ainsi, le fait d'exploiter une fenêtre de taille k va permettre d'avoir une information sur le nombre de fois qu'un mot apparaît

dans le contexte d'un autre et, comme deuxième information plus fine la nature des relations qui existent entre deux mots.

Par ailleurs, nous avons pu constater que la méthode $Comb_{apri}$ était plus sensible aux modifications des mesures d'association et de similarité par rapport à la méthode $Comb_{post}$. Ceci s'explique par le fait que $Comb_{post}$ agit sur les scores *a posteriori* alors que $Comb_{apri}$ agit directement sur le contenu des vecteurs de contexte. Les moins bons résultats sur le corpus des énergies renouvelables s'expliquent par la moins bonne qualité de ce corpus en comparaison avec celui du cancer du sein, ainsi que sa plus petite taille. Son utilisation a néanmoins permis de montrer que, même avec un corpus de très petite taille, les deux méthodes proposées restent plus performantes que les deux représentations contextuelles prises séparément.

6 Conclusion

Nous nous sommes intéressés dans cet article aux deux principales manières de représenter le contexte des mots, à savoir : une représentation graphique ainsi qu'une représentation syntaxique. Nous avons ensuite introduit deux nouvelles techniques de combinaison de ces représentations. Les deux approches de combinaisons contextuelles proposées ont montré des résultats supérieurs à l'utilisation de chaque représentation séparément, pour la plupart des paramètres de configurations. Nous espérons que ce travail ouvrira la voie à une recherche plus approfondie concernant l'enrichissement du contenu des vecteurs de contexte par des informations multiples sur les mots les composant. Si les travaux de cet article se sont limités à deux types d'informations contextuelles, d'autres informations sont envisageables comme l'utilisation de thesaurus ou d'autres informations comme les cognats, les translittérations, les collocations, etc.

Remerciements

Ce travail qui s'inscrit dans le cadre du projet CRISTAL `www.projet-cristal.org` a bénéficié d'une aide de l'Agence National de la Recherche portant la référence ANR-12-CORD-0020.

Références

ANDRADE, D., MATSUZAKI, T. et TSUJII, J. (2011). Effective use of dependency structure for bilingual lexicon creation. *In Proceedings of the 12th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing'11)*, pages 80–92, Tokyo, Japan.

ASLAM, J. A. et MONTAGUE, M. (2001). Models for Metasearch. *In Proceedings of the 24th Annual SIGIR Conference (SIGIR'01,)*, pages 275–284, New Orleans, Louisiana.

BOWKER, L. et PEARSON, J. (2002). *Working with Specialized Language : A Practical Guide to Using Corpora*. Routeledge, New York, USA.

BROSSEAU-VILLENEUVE, B., NIE, J.-Y. et KANDO, N. (2010). Towards an optimal weighting of context words based on distance. *In Proceedings of the 23rd International Conference on Computational Linguistics (COLING'10)*, pages 107–115, Beijing, China.

DÉJEAN, H., GAUSSIER, É. et SADAT, F. (2002). An approach based on multilingual thesauri and model combination for bilingual lexicon extraction. *In Proceedings of the 19th International Conference on Computational Linguistics (COLING'02)*, pages 1–7, Taipei, Taiwan.

DÉJEAN, H. et GAUSSIER, E. (2002). Une nouvelle approche à l'extraction de lexiques bilingues à partir de corpus comparables. *Lexicometrica, Alignement Lexical dans les Corpus Multilingues*, pages 1–22.

FIRTH, J. R. (1957). A synopsis of linguistic theory 1930–1955. *In Studies in Linguistic Analysis (special volume of the Philological Society)*, pages 1–32. Blackwell, Oxford.

FUNG, P. (1995). Compiling Bilingual Lexicon Entries From a non-Parallel English-Chinese Corpus. *In* FARWELL, D., GERBER, L. et HOVY, E., éditeurs : *Proceedings of the 3rd Conference of the Association for Machine Translation in the Americas (AMTA'95)*, pages 1–16, Langhorne, PA, USA.

FUNG, P. (1998). A statistical view on bilingual lexicon extraction : From parallel corpora to non-parallel corpora. *In Proceedings of Machine Translation and the Information Soup, Third Conference of the Association for Machine Translation in the Americas (AMTA'98)*, pages 1–17, Langhorne, PA, USA.

FUNG, P. et YEE, L. Y. (1998). An ir approach for translating new words from non parallel, comparable texts. *In Proceedings of the 17th international conference on Computational linguistics (COLING'98)*, pages 414–420, Quebec, Canada.

GAMALLO, O. (2007). Learning bilingual lexicons from comparable english and spanish corpora. *In Proceedings of Machine Translation Summit XI*, pages 191–198, Copenhagen, Denmark.

GAMALLO, O. (2008a). Evaluating two different methods for the task of extracting bilingual lexicons from comparable corpora. *In Proceedings of LREC 2008 Workshop on Comparable Corpora (LREC'08)*, pages 19–26, Marrakech, Marroco.

GAMALLO, O. (2008b). The meaning of syntactic dependencies. *Linguistik Online*.

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

GAUSSIER, E., RENDERS, J.-M., MATVEEVA, I., GOUTTE, C. et DÉJEAN, H. (2004). A geometric view on bilingual lexicon extraction from comparable corpora. *In Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL'04)*, pages 526–533, Barcelona, Spain.

GROC, C. D. (2011). Babouk : Focused Web Crawling for Corpus Compilation and Automatic Terminology Extraction. *In Proceedings of The IEEE WICACM International Conferences on Web Intelligence*, pages 497–498, Lyon, France.

HAGHIGHI, A., LIANG, P., BERG-KIRKPATRICK, T. et KLEIN, D. (2008). Learning bilingual lexicons from monolingual corpora. *In Proceedings of the 46nd Annual Meeting of the Association for Computational Linguistics (ACL'08)*, pages 771–779, Columbus, Ohio.

LAROCHE, A. et LANGLAIS, P. (2010). Revisiting context-based projection methods for term-translation spotting in comparable corpora. *In Proceedings of the 23rd International Conference on Computational Linguistics (COLING'10)*, pages 617–625, Beijing, China.

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

LI, B., GAUSSIER, E., MORIN, E. et HAZEM, A. (2011). Degré de comparabilité, extraction lexicale bilingue et recherche d'information interlingue. *In Actes de la 18ème Conférence Traitement Automatique des Langues Naturelles (TALN'11)*, pages 283–293, Montpellier, France.

LIN, D. (1998). Dependency-based evaluation of minipar. *In Proceedings of the Workshop on the Evaluation of Parsing Systems, First International Conference on Language Resources and Evaluation (LREC'98)*, Granada, Spain.

MORIN, E. (2009). Apport d'un corpus comparable déséquilibré à l'extraction de lexiques bilingues. *In Actes de la 16ème Conférence Traitement Automatique des Langues Naturelles (TALN'09)*, Senlis, France.

MORIN, E. et DAILLE, B. (2004). Extraction terminologique bilingue à partir de corpus comparables d'un domaine spécialisé. *Traitement Automatique des Langues. TAL*, 45(3):103–122.

MORIN, E., DUFOUR-KOWALSKI, S. et DAILLE, B. (2004). Extraction de terminologies bilingues à partir de corpus comparables. *In Actes de la 11ème Conférence Traitement Automatique des Langues Naturelles (TALN'04)*, pages 309–318, Fès, Maroc.

PROCHASSON, E. et MORIN, E. (2009). Influence des points d'ancrage pour l'extraction lexicale bilingue à partir de corpus comparables spécialisés. *In Actes de la 16ème Conférence Traitement Automatique des Langues Naturelles (TALN'09)*, Senlis, France.

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

RAPP, R. (1999). Automatic Identification of Word Translations from Unrelated English and German Corpora. *In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, pages 519–526, College Park, MD, USA.