Utilisation de la similarité sémantique pour l'extraction de lexiques bilingues à partir de corpus comparables

Dhouha Bouamor [1,2,3] Nasredine Semmar[1] Pierre Zweigenbaum[2]
(1) CEA-LIST, LVIC, F91191 Gif sur Yvette Cedex, France
(2) LIMSI-CNRS, F-91403 Orsay, France
(3) Univ. Paris Sud, Orsay, France
dhouha.bouamor@cea.fr, nasredine.semmar@cea.fr, pz@limsi.fr

RÉSUMÉ
Cet article présente une nouvelle méthode visant à améliorer les résultats de l'approche standard utilisée pour l'extraction de lexiques bilingues à partir de corpus comparables spécialisés. Nous tentons de résoudre le problème de la polysémie des mots dans les vecteurs de contexte par l'introduction d'un processus de désambiguïsation sémantique basé sur WordNet. Pour traduire les vecteurs de contexte, au lieu de considérer toutes les traductions proposées par le dictionnaire bilingue, nous n'utilisons que les mots caractérisant au mieux les contextes en langue cible. Les expériences menées sur deux corpus comparables spécialisés français-anglais (financier et médical) montrent que notre méthode améliore les résultats de l'approche standard plus particulièrement lorsque plusieurs mots du contexte sont ambigus.

ABSTRACT
This paper presents a new method that aims to improve the results of the standard approach used for bilingual lexicon extraction from specialized comparable corpora. We attempt to solve the problem of context vector word polysemy. Instead of using all the entries of the dictionary to translate a context vector, we only use the words of the lexicon that are more likely to give the best characterization of context vectors in the target language. On two specialised French-English comparable corpora, empirical experimental results show that our method improves the results obtained by the standard approach especially when many words are ambiguous.

MOTS-CLÉS : lexique bilingue, corpus comparable spécialisé, désambiguïsation sémantique, WordNet.

KEYWORDS: bilingual lexicon, specialized comparable corpora, semantic disambiguation, WordNet.

1 Introduction

Les lexiques bilingues sont des ressources particulièrement utiles pour la Traduction Automatique et la Recherche d'Information Interlingue. Les recherches en extraction lexicale à partir de corpus multilingues se sont largement concentrées sur les corpus parallèles. En effet, la rareté de ces corpus, en particulier pour les domaines spécialisés et pour les couples de langues ne faisant pas intervenir l'anglais, conduit en outre à orienter les recherches en extraction de lexiques bilingues

vers l'utilisation de corpus comparables (Fung, 1995; Rapp, 1995; Chiao et Zweigenbaum, 2003; Gamallo Otero, 2007; Prochasson *et al.*, 2009; Kun et Tsujii, 2009). La plupart de ces travaux héritent de la sémantique distributionnelle (Harris, 1954) et reposent sur la simple observation que si dans une langue source deux mots cooccurrent plus souvent que par hasard, alors dans un texte de langue cible, leurs traductions doivent également cooccurrer plus souvent. Cette approche dite ***standard*** se base sur la caractérisation et la comparaison d'environnements lexicaux des termes sources et cibles, représentés par des *vecteurs de contexte*. Ces vecteurs stockent un ensemble d'unités lexicales représentatif de leur voisinage. Dans la pratique, afin de pouvoir comparer les vecteurs de contexte de langues différentes, le passage d'une langue à une autre est nécessaire et s'effectue généralement par l'intermédiaire d'un dictionnaire bilingue amorce.

Le dictionnaire bilingue est au coeur de l'approche standard. Son utilisation pose des problèmes lorsqu'un mot possède plusieurs traductions, qu'il s'agisse de traductions synonymes ou d'un terme source polysémique. Par exemple, le terme Français *"action"* se traduit en Anglais par les termes *"share, stock, lawsuit"* et *"deed"*. Dans ce cas, il est difficile d'évaluer dans des ressources plates comme les dictionnaires bilingues quelles traductions sont les plus pertinentes, vu qu'elle sont le plus souvent non ordonnées. L'approche standard prend en compte toutes les traductions disponibles et les conserve avec la même priorité dans le vecteur traduit indépendamment du domaine sur lequel porte l'étude. Ainsi, en domaine de la Finance, la prise en compte des termes *"lawsuit"* et *"deed"* ne feront probablement qu'ajouter du bruit dans les vecteurs de contexte.

Dans ce présent travail, nous présentons une nouvelle approche qui tente de résoudre le problème de polysémie des mots non traité par l'approche standard. Un mot polysémique est une unité lexicale ayant plusieurs sens dans une langue ou une fois traduite dans une autre langue. Nous introduisons un processus de désambiguïsation sémantique des vecteurs de contexte construits par l'approche standard. L'intuition qui sous-tend cette méthode est que, pour chaque mot polysémique du vecteur de contexte, au lieu de considérer toutes les traductions proposées par le dictionnaire bilingue, nous n'utilisons que les traductions susceptibles de donner la meilleure représentation du vecteur de contexte en langue cible. Le processus de désambiguïsation repose sur une mesure de similarité sémantique calculée en se basant sur le thésaurus WordNet (Fellbaum, 1998). Nous testons cette méthode sur deux corpus comparables spécialisés pour le couple des langues français-anglais. Une amélioration des résultats de l'approche standard est reportée plus particulièrement lorsque plusieurs mot du corpus sont ambigus.

La suite de l'article est organisée comme suit : dans la section 2, nous présentons l'approche standard et passons en revue les principaux travaux connexes à la tâche d'extraction de lexiques bilingues à partir de corpus comparables. Puis, nous décrivons, dans la section 3, le processus de désambiguïsation sémantique proposé. La section 4 sera consacrée aux expériences menées ainsi qu'à la présentation des résultats obtenus. Notre article se conclura par une présentation des principales perspectives (section 5).

2 Extraction de lexiques bilingues

2.1 Approche standard

La plupart des travaux traitant la tâche d'extraction de lexiques bilingues à partir de corpus comparables se basent sur l'approche standard (Fung, 1998; Chiao et Zweigenbaum, 2002; Laroche et Langlais, 2010). Cette approche se décompose en trois étapes :

– **Constitution des vecteurs de contexte** : Ces vecteurs sont d'abord extraits en repérant les mots qui apparaissent autour d'un terme à traduire S dans une fenêtre contextuelle de n mots. Habituellement, des mesures d'associations comme l'information mutuelle (Morin et Daille, 2006), le taux de vraisemblance (Morin et Prochasson, 2011) ou encore le rapport des chances (odds-Ratio) (Laroche et Langlais, 2010) sont utilisées pour définir les entrées du vecteur de contexte.
– **Transfert des vecteurs de contexte** : Afin de rendre possible la comparaison des vecteurs sources et cibles, les vecteurs des termes sources sont traduits par le biais d'un dictionnaire bilingue amorce. Si le dictionnaire propose plusieurs traductions pour un élément, nous ajoutons l'ensemble des traductions proposées. Les mots ne figurant pas dans le dictionnaire sont tout simplement ignorés.
– **Comparaison des vecteurs sources et cibles** : Les vecteurs traduits sont ensuite comparés à l'ensemble des vecteurs de contexte en langue cible à l'aide d'une mesure de similarité vectorielle. La plus populaire étant le cosinus, mais de nombreux auteurs ont étudiés des métriques alternatives comme la distance du Jaccard pondérée ou encore le city-block. En fonction des valeurs de similarité, nous obtenons une liste ordonnée de traductions candidates pour le terme S.

2.2 Travaux reliés

La couverture du dictionnaire bilingue assurant le transfert des vecteurs de contexte en langue cible demeure le noyau de l'approche standard. Si trop peu de mots sont traduits, la comparaison de vecteurs traduits et de vecteurs cibles ne sera pas significative puisque réalisée sur un échantillon trop faible de vocabulaire. Pour limiter cet effet, des techniques visant à améliorer les résultats de l'approche standard ont vu le jour et ce par l'adjonction de ressources dictionnariques spécialisées supplémentaires préétablies (Déjean *et al.*, 2002; Chiao et Zweigenbaum, 2003), extraites de corpus parallèles (Morin et Prochasson, 2011) ou encore du même corpus d'étude (Vulić et Moens, 2012).

Récemment, des recherches fondées sur l'hypothèse que plus les vecteurs de contexte sont représentatifs, meilleure est la mise en correspondance bilingue ont été menées. (Prochasson *et al.*, 2009) utilisent les translittérations et mots savants comme 'points d'ancrage'. L'objectif est que la comparaison des vecteurs se fonde en priorité sur les points d'ancrage, puis sur le reste d'éléments. Outre les translittérations, (Rubino et Linarès, 2011) combinent la représentation contextuelle avec une représentation thématique de termes médicaux, en émettant l'hypothèse qu'un terme et sa traduction partagent des similarités d'un point de vue thématique. (Hazem et Morin, 2012a) proposent deux critères de filtrage du dictionnaire bilingue dans le but de ne garder que les mots qui donnent la meilleure représentation du vecteur de contexte dans la langue cible. Le premier critère se base sur les catégories grammaticales des mots du contexte

mais aucune amélioration n'a été démontrée. Le deuxième critère étant basé sur une mesure de pertinence d'un mot pour un domaine donné. Contrairement au premier critère, celui ci apporte une petite amélioration (4% en précision) par rapport à la méthode standard.

(Gaussier *et al.*, 2004) tentent de résoudre le problème d'ambiguïté de mots des vecteurs de contexte en langues source et cible. Ils utilisent une vue géométrique et décomposent le vecteur d'un mot en fonction de ses sens par l'utilisation de plusieurs méthodes comme l'analyse canonique de corrélation et l'analyse sémantique latente. Les meilleurs résultats sont obtenus par l'utilisation d'une approche mixte avec une amélioration de la F-Mesure au $Top20$ de +2% par rapport à l'approche standard. Dans cet article, nous présentons une approche traitant le problème d'ambiguïté des mots des vecteurs de contexte mais qui diffère de celle proposée par (Gaussier *et al.*, 2004). Alors qu'ils mettent l'accent sur l'ambiguïté des mots en langues source et cible, nous jugeons qu'il serait suffisant de lever l'ambiguïté des éléments des vecteurs de contexte en langue source vu que l'ambiguïté parvient lors du transfert des vecteurs de contexte sources

3 Désambiguïsation lexicale des vecteurs de contexte

Nous proposons dans cet article une approche qui tente d'améliorer les résultats de l'approche standard. Nous abordons le problème associé aux mots polysémiques révélés par le dictionnaire bilingue amorce lors du transfert des vecteurs de contexte sources. Comme il a été mentionné dans la section 1, lorsque l'extraction lexicale porte sur un domaine spécialisé, les traductions proposées par le dictionnaire bilingue ne sont pas toutes pertinentes pour la représentation des vecteurs de contexte en langue cibles. Par exemple, dans le domaine juridique, la traduction du mot *action* (Fr) par *share* ou *stock* (An) ne fera qu'introduire du bruit dans les vecteurs traduits. L'intuition derrière notre approche est qu'il conviendrait d'introduire un *processus de désambiguïsation sémantique lexicale* visant à améliorer l'adéquation des vecteurs de contexte traduits et par conséquent améliorer les résultats de l'approche standard. Dans cette section, nous commençons par décrire la ressource sémantique sur laquelle se base notre approche. Ensuite, nous présentons en détail notre méthode de désambiguïsation des vecteurs de contexte.

3.1 Ressource sémantique

Un grand nombre de techniques de désambiguïsation lexicale ont été présentées dans la littérature. Les plus populaires sont celles mesurant une similarité sémantique en se basant sur le thésaurus *WordNet*. Cette ressource est structurée autour de la notion de *synsets*, c'est-à-dire en quelque sorte un ensemble de synonymes qui forment un concept. Chaque synset représente un sens de mot. Les synsets sont reliés entre eux par des relations, soit lexicales (antonymie par exemple) ou taxonomiques (hyperonymie, méronymie, etc). Ce thésaurus est largement utilisé dans des applications reposant sur le calcul de similarité des mots telles que la recherche de documents (Hwang *et al.*, 2011) ou d'images (Cho *et al.*, 2007; Choi *et al.*, 2012). Dans ce travail, nous l'utilisons pour dériver une similarité sémantique entre les éléments de chaque vecteur de contexte permettant de sélectionner les sens des mots les plus saillants à la représentation des termes à traduire. À notre connaissance, c'est une première application de WordNet en extraction de lexiques bilingues à partir de corpus comparables.

Vecteur de contexte	{action}, {dividende}, {liquidité}, ...
Dictionnaire bilingue	{act, stock, action, deed, lawsuit, fact, operation, plot, share} , {dividend} , {liquidity}
Sem_{Sim}	{dividend, act} ; {dividend,stock} ; ... ; {liquidity, act} ; {liquidity,stock} ; ...
Ave_Wup(*action*)	share :0.5236, stock :0.5236, action :0.4256, act :0.2139, operation :0.2045, plot :0.2011, fact :0.1934, deed :0.1594, lawsuit :0.1212

TABLE 1 – Désambiguïsation sémantique du vecteur de contexte du terme *bénéfice*

Parmi les mesures de similarité sémantique utilisant WordNet, nous retrouvons les mesures basées sur la distance taxonomique. Le principe général de ces mesures est de compter le nombre d'arcs qui séparent deux sens dans WordNet. Dans ce cadre, nous choisissons la mesure définie par (Wu et Palmer, 1994). La similarité est définie selon la distance qui sépare deux concepts par rapport à leur sens commun le plus spécifique (LCS) que la racine de la taxonomie. La similarité entre deux sens s_1 et s_2 est :

$$Sim_{wup}(s_1, s_2) = \frac{2 \times depth(LCS)}{depth(s_1) + depth(s_2)} \quad (1)$$

Où depth(LCS) est le nombre d'arcs qui séparent LCS de la racine et depth(s_i) avec i le nombre d'arcs qui séparent s_i de la racine en passant par LCS. Cette mesure a l'avantage d'avoir de meilleures performances par rapport aux autres mesures de similarité (Lin, 1998).

3.2 Processus de désambiguïsation

Une fois transféré en langue cible, le processus de désambiguïsation des vecteurs de contexte intervient. Ce processus tente de trouver pour chacune des entrées polysémiques dans les vecteurs traduits le sens le plus adéquat. Pour ce faire, nous utilisons les unités non polysémiques pour déduire les sens de celles polysémiques. Nous émettons l'hypothèse qu'un mot est non polysémique s'il ne possède qu'une seule traduction dans le dictionnaire bilingue. Cette hypothèse est vérifiée dans 95% des cas dans WordNet (i.e mots associés à un seul synset).

Précisément, pour chaque entrée polysémique de chaque vecteur, nous mesurons la similarité sémantique entre toutes les traductions qui lui sont associées et toutes les unités non polysémiques du même vecteur. En fonction des valeurs de similarité, nous obtenons une liste ordonnée de sens ou traductions pour chaque mot polysémique.

Plus formellement, puisqu'un mot peut appartenir à plus d'un sens ou synset dans WordNet, nous déterminons la similarité sémantique entre deux mots m_1 et m_2 comme le maximum de Sim_{wup} entre le ou les synsets qui incluent les $synsets(m_1)$ et les $synsets(m_2)$ selon la formule suivante :

$$Sem_{Sim}(m_1, m_2) = \max\{Sim_{wup}(s_1, s_2); (s_1, s_2) \in synsets(m_1) \times synsets(m_2)\} \quad (2)$$

Ensuite, pour identifier le sens le plus approprié pour chaque mot polysémique k dans les vecteurs de contexte, nous mesurons une ***moyenne de similarité*** (Formule 3) pour chacune des

traductions proposées kj.

$$Ave\_Wup(k_j) = \frac{\sum_{i=1}^{N} Sem_{Sim}(m_i, k_j)}{N} \quad (3)$$

où N est le nombre total des mots non polysémique du vecteur traduit et Sem_{Sim} est la valeur de similarité entre k_j et le mot non polysémique m_i. Dans le cas où tous les mots du vecteur de contexte sont polysémiques, il est possible de calculer la similarité sémantique entre toutes les combinaisons de mots. Dans de tels cas, nous choisissons de ne pas toucher au vecteurs de contexte puisque avec le calcul de ce type de similarité une augmentation de la complexité algorithmique et détérioration des résultats d'extraction ont été constatés dans des expérimentations préliminaires.

Un exemple de désambiguïsation de vecteur de contexte du terme "bénéfice" est décrit dans la table 1. Ce vecteur est construit à partir de corpus comparable spécialisé et contient les mots *action*, *dividende*, *liquidité* et d'autres unités. Lors du transfert de ce vecteur de la langue source (Français) à celle cible (Anglais), le dictionnaire bilingues propose les traductions suivantes « *act, stock, action, deed, lawsuit, fact, operation, plot, share* », « *dividend* » et « *liquidity* » pour traduire respectivement les mots « *action* », « *dividende* » et « *liquidité* ». Nous utilisons les unités lexicales non polysémiques « *dividende* » et « *liquidité* » pour désambiguiser le mot « *action* ». En observant la valeur de Ave_Wup, nous remarquons que dans ce contexte, les mots *share* et *stock* sont les traductions les plus appropriées au mot *action*. Nous remarquons aussi que les mots issus du domaine général se placent après pour retrouver à la fin les unités les moins proches (*deed et lawsuit*).

4 Expérimentations et résultats

4.1 Ressources linguistiques

Dans le cadre de cette étude, nous avons construit deux corpus comparables spécialisés français-anglais à partir de l'encyclopédie libre Wikipédia[1]. Nous exploitons l'aspect multilingue cette ressource pour en extraire de la terminologie spécialisée qui pourra créer ou enrichir des ressources linguistiques existantes. Nous nous intéressons particulièrement au domaine de la « *finance des entreprises* » et à la thématique du « *cancer du sein* » relevant du domaine médical. Notre approche repose en premier lieu sur l'extraction de pages de Wikipédia en langue source. Ensuite, les liens interlingues sont utilisés afin de chercher l'information translinguistique et donc construire la partie du corpus en langue cible (Sadat et Terrasa, 2010).

Nous considérons que le domaine d'étude constitue une catégorie dans Wikipédia. Les catégories sont un système de classement thématique des articles de Wikipédia. Une requête composée du domaine d'étude en langue source (par exemple *finance des entreprises*) est donc construite pour extraire une arborescence de catégories ou de thèmes ayant pour catégorie mère le domaine de spécialité. Un exemple d'arborescence est présenté dans la figure 1.

Ensuite, Nous collectons tous les articles associés à chacune des catégories de l'arborescence pour construire un corpus spécialisé monolingue (en langue source). Afin de collecter les articles

[1] http://dumps.wikimedia.org/

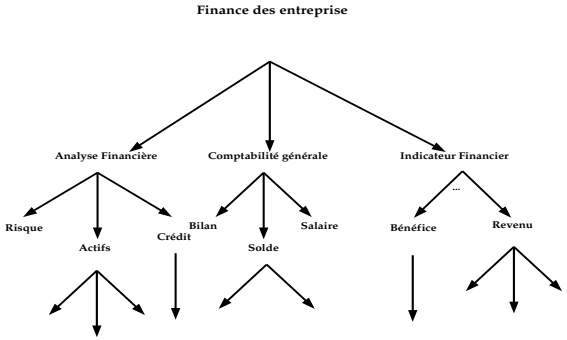

FIGURE 1 – Arborescence de catégories de la thématique Finance des entreprises

en langue cible, les liens interlingues au sein de chaque article du corpus monolingue sont utilisés. Un étiquetage morpho-syntaxtique et une lemmatisation ont été appliqués sur les articles collectés. Nous avons aussi retiré les mots fonctionnels et ceux apparaissant moins de deux fois dans les deux parties du corpus comparable. Nous avons ainsi construit deux corpus comparables de taille réduite. La taille en nombre de mots des corpus résultants est dans la table 2

Corpus	Français	Anglais
Finance des entreprises	402.486	756.840
Cancer du sein	396.524	524.805

TABLE 2 – Taille des corpus comparables. La taille est exprimée en nombre de *mots*

Le dictionnaire bilingue Français-Anglais assurant le transfert des vecteurs de contexte comporte environ 120000 entrées avec en moyenne 7 traductions par entrée. Il s'agit d'un dictionnaire du domaine général comportant quelques mots en rapport avec le domaine financier et médical.

Pour évaluer la qualité de l'approche standard et celle introduisant la désambiguïsation lexicale des vecteurs de contexte, nous avons construit une liste de traductions de référence pour chaque domaine. Habituellement, la taille de ces listes est autour de 100 mots (Hazem et Morin, 2012a; Chiao et Zweigenbaum, 2002). Précisons que nous nous intéressons dans cet article uniquement à l'extraction bilingue de termes simples. D'autres recherches se sont portées sur l'extraction de termes complexes (Morin et Daille, 2004; Laroche et Langlais, 2010). Pour le domaine de la finance des entreprises, une liste composée de 125 mots simples est extraite du *glossaire bilingue de la micro-finance* [2]. En ce qui concerne le domaine du cancer du sein, 79 termes issus du méta-thésaurus *UMLS*[3] et du *MESH*[4] sont extraits. Ces deux listes sont composées de paires de termes français-anglais apparaissant au moins cinq fois dans chaque partie des corpus comparables.

[2] http://www.microfinance.lu/la-microfinance-cest-quoi/glossaire.html
[3] http://www.nlm.nih.gov/research/umls/
[4] http://mesh.inserm.fr/mesh/

4.2 Expérimentations

Afin de mener à bien nos expériences, nous avons besoin de régler trois principaux paramètres : (1) la taille de la fenêtre contextuelle, (2) la mesure d'association et (3) la mesure de similarité. Comme dans la plupart des travaux antérieurs (Hazem et Morin, 2012b; Chiao et Zweigenbaum, 2002), nous fixons la taille de la fenêtre contextuelle à 7, partant de l'idée qu'elle approxime les dépendances syntaxiques. Une étude de différentes combinaisons entre les mesures d'association et les métriques de similarité a été présentée dans (Laroche et Langlais, 2010). Pour le domaine médical, la configuration la plus efficace étant de combiner le rapport des chances [Odds-Ratio] avec le cosinus. Nous avons suivi ces travaux pour la définition de ces paramètres. La formule du rapport des chances est définie dans l'équation ci-dessous :

$$OddsRatio_{disc} = \log \frac{(O_{11} + \frac{1}{2})(O_{22} + \frac{1}{2})}{(O_{12} + \frac{1}{2})(O_{21} + \frac{1}{2})} \tag{4}$$

Où O_{ij} sont les cellules d'une table de contingence 2×2 regroupant les fréquences d'observation de deux termes dans une fenêtre donnée. Le cosinus de l'angle formé par deux vecteurs source v_s et cible v_c est défini dans l'équation 5.

$$Cos(v_s, v_c) = \frac{\sum_j OddsRatio_j^s \times OddsRatio_j^c}{\sqrt{\sum_j OddsRatio_j^{s\,2}} \times \sqrt{\sum_j OddsRatio_j^{c\,2}}} \tag{5}$$

4.3 Résultats et discussion

Il est difficile de comparer les résultats de différents travaux en extraction de lexiques bilingues à partir de corpus comparables, en raison de différences entre les corpus, les domaines d'études ou encore les ressources linguistiques utilisées (Prochasson et Morin, 2009). À ce jour, aucun jeu de données pouvant servir de référence n'a été mis en place. C'est pour cette raison que nous utilisons les résultats de l'approche standard (AS) comme référence. Nous évaluons les performances de cette approche et de celle présentée en section 3 en utilisant les métriques de précision (P_N), rappel (R_N) au $TopN$ et de MAP (Mean Average Precision) (Manning *et al.*, 2008). La précision est le nombre de traductions correctes divisé par le nombre de termes pour lesquels le système propose au moins une traduction. Le rappel est égal au rapport entre les traductions correctes et le nombre total des termes. La MAP représente la qualité d'un système en fonction de différents niveaux de rappel :

$$MAP(Q) = \frac{1}{Q} \sum_{|Q|}^{j=1} \frac{1}{m_j} \sum_{m_j}^{k=1} Précision(R_{jk}) \tag{6}$$

Où Q constitue le nombre de termes à traduire, m_j est le nombre de traductions de référence pour le $j^{ème}$ terme et $Précision(R_{jk})$ est égale à 0 si la traduction de référence n'est pas trouvée pour le $j^{ème}$ terme ou $\frac{1}{r}$ s'il y figure (r est le rang de la traduction de référence dans les traductions candidates).

Méthode	P1	P10	P20	R1	R10	R20	MAP
AS	4.6	14	18.6	4	12	16	6.4
WN-S_1	6.5	19.6	26.1	5.6	16.8	22.4	8.9
WN-S_2	10.2	**25.2**	30.8	8	**21.6**	26.4	12.2
WN-S_3	10.2	24.2	**32.7**	8.8	20.8	**28**	12.2
WN-S_4	**11.2**	22.4	29.9	**9**	19	25	**12.4**
WN-S_5	9.3	20.5	28	8	17.6	24	11
WN-S_6	8.4	20.5	23.3	7.2	17.6	20	9.41
WN-S_7	7.4	17.7	24.2	6.4	15.2	20.8	9

TABLE 3 – Corpus de « finance des entreprises » : Précision et Rappel au $TopN$ ($N = 1, 10, 20$) et MAP (%)

Rappelons que l'AS utilise toutes les traductions proposées par le dictionnaire bilingue pour le transfert des vecteurs de contexte. Notre méthode de désambiguïsation des contextes fournit pour chaque unité polysémique, un vecteur de sens ordonné en fonction des valeurs de similarité. A cet égard, il convient de s'interroger sur le nombre de sens à considérer pour chaque mot polysémique. Devrions nous considérer que l'élément maximisant la similarité sémantique dans le vecteur de contexte ou envisager un plus grand nombre de sens notamment quand un vecteur de sens contient des synonymes (*share* (An) et *stock* (An) dans la table 1). C'est précisément pour cette raison que nous prenons en considération pour chaque unité polysémique différents nombre de sens dans nos expérimentations allant du sens le plus similaire jusqu'au septième sens. L'arrêt au septième sens ou traduction s'explique par le fait qu'en moyenne, un mot du corpus comparable possède 7 traductions dans le lexique bilingue. Ces méthodes sont notées WN-S_i où i est le nombre de sens associé à chaque unité polysémique. La table 3 présente les résultats obtenus pour le corpus de la finance des entreprises.

Nous constatons que notre méthode qui consiste en une désambiguïsation des vecteurs de contexte dépasse les performances de la méthode de référence AS pour toutes les configurations. La meilleure MAP est atteinte par (WN-S_4), lorsque pour chaque mot polysémique, nous gardons les quatre traductions les plus similaires aux éléments non polysémiques des vecteurs de contexte. La précision au Top20 la plus élevée est obtenue par WN-S_3. L'utilisation des trois premiers sens de mots dans le vecteur fait passer la précision au Top20 de 18.6% à 32.7%. Une dégradation de la MAP, précision et rappel est constatée à partir de WN-S_5. L'ajout progressif des traductions rapproche les résultats obtenus de ceux de l'AS. Nous estimons par conséquent que à partir de WN-S_5, les traductions ajoutées ne font qu'introduire du bruit dans les vecteurs de contextes.

En ce qui concerne le corpus traitant la thématique du cancer du sein, des résultats différents ont été obtenus. Comme le montre la table 4, lorsque les vecteurs de contexte sont totalement non ambigus (i.e. chaque unité source est traduite par au plus un mot), une diminution de la précision, rappel et MAP est notée par rapport à l'AS. Néanmoins, dans la plupart des autres cas, des améliorations plus au moins petites sont obtenues. Dans la méthode WN-S_5, nous reportons le meilleur score avec un gain de +3.4% en MAP par rapport à AS. Par contre les meilleurs rappel et précision au Top 10 et 20 sont atteints par WN-S_2 et WN-S_3.

En observant les résultats (table 3 et 4) des domaines de la finance des entreprises et celui du cancer du sein, nous remarquons que dans la plupart des cas l'approche de désambiguïsation des

Méthode	P1	P10	P20	R1	R10	R20	MAP
AS	34.2	54.2	58.5	25	39.5	42.7	31.4
WN-S_1	25.7	50	57.1	18.7	36.4	41.6	25.7
WN-S_2	31.4	61.4	**67.1**	22.9	44.7	**48.9**	31.3
WN-S_3	34.2	**62.8**	**67.1**	25	**45.8**	**48.9**	34.2
WN-S_4	34.2	57.1	64.2	25	41.6	46.8	33.2
WN-S_5	**35.7**	57.1	65.7	**26**	41.6	47.9	**34.8**
WN-S_6	35.7	57.1	65.2	26	41.6	46.8	34.7
WN-S_7	35.7	58.5	65.7	26	42.7	47.9	33.9

TABLE 4 – Corpus du « cancer du sein » : Précision et Rappel au $TopN$ ($N = 1, 10, 20$) et MAP (%)

vecteurs de contexte par l'utilisation de la similarité sémantique de WordNet donne de meilleurs résultats que l'approche de référence AS mais à des degrés différents. Les améliorations reportées en domaine de la finance des entreprises dépassent de loin celles du cancer du sein. Ceci peut-être dû au fait que le vocabulaire utilisé dans le domaine du cancer du sein est plus spécifique et donc moins ambigu que celui utilisé dans les textes de la finance des entreprises. Dans ce cas, les améliorations restent trouvé dans de larges valeurs de N au Top$N$ (la désambiguïsation des contextes aide à apporter des traductions plus éloignées au Top20).

5 Conclusion

Nous avons présenté dans cet article une nouvelle méthode qui tente d'améliorer les résultats de l'approche standard utilisée en extraction lexicale bilingue. Cette méthode a pour but de lever l'ambiguïté des mots polysémiques dans les vecteurs de contexte en sélectionnant uniquement les traductions susceptibles de représenter au mieux les termes à traduire. La technique proposée repose sur le calcul d'une similarité sémantique faisant appel au réseau sémantique WordNet. Les expériences menées sur deux corpus comparables spécialisés montrent que les performances de cette technique sont dans la plupart des cas supérieures à celles obtenues par l'approche standard.

Nous considérons que nos expériences initiales sont positives et peuvent être améliorées de diverses façons. Nous avons d'abord l'intention d'agrandir la taille des corpus comparables utilisés. De plus, dans ce travail, nous considérons que les corpus construit sont de bonne qualité, nous tenterons donc d'agir sur leur qualité en utilisant par exemple la mesure proposée par (Li et Gaussier, 2010). Outre la métrique définie par (Wu et Palmer, 1994), nous comptons utiliser d'autres mesures de similarité sémantique et comparer leurs performances. Nous prévoyons également d'appliquer notre méthode à l'extraction de lexiques bilingues à partir d'autre corpus très spécialisés pour valider nos hypothèses.

Références

CHIAO, Y.-C. et ZWEIGENBAUM, P. (2002). Looking for candidate translational equivalents in specia-

lized, comparable corpora. *In Proceedings of the 19th international conference on Computational linguistics - Volume 2*, COLING '02, pages 1–5. Association for Computational Linguistics.

CHIAO, Y.-C. et ZWEIGENBAUM, P. (2003). The effect of a general lexicon in corpus-based identification of French-English medical word translations. *In Proceedings Medical Informatics Europe, volume 95 of Studies in Health Technology and Informatics*, pages 397–402, Amsterdam.

CHO, M., CHOI, C., KIM, H., SHIN, J. et KIM, P. (2007). Efficient image retrieval using conceptualization of annotated images. Lecture Notes in Computer Science, pages 426–433. Springer.

CHOI, D., KIM, J., KIM, H., HWANG, M. et KIM, P. (2012). A method for enhancing image retrieval based on annotation using modified wup similarity in wordnet. *In Proceedings of the 11th WSEAS international conference on Artificial Intelligence, Knowledge Engineering and Data Bases*, AIKED'12, pages 83–87, Stevens Point, Wisconsin, USA. World Scientific and Engineering Academy and Society (WSEAS).

DÉJEAN, H., GAUSSIER, E. 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 - Volume 1*, COLING '02, pages 1–7. Association for Computational Linguistics.

FELLBAUM, C. (1998). *WordNet : An Electronic Lexical Database*. Bradford Books.

FUNG, P. (1995). A pattern matching method for finding noun and proper noun translations from noisy parallel corpora. *In Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, pages 236–243. Association for Computational Linguistics.

FUNG, P. (1998). A statistical view on bilingual lexicon extraction : From parallel corpora to non-parallel corpora. *In Parallel Text Processing*, pages 1–17. Springer.

GAMALLO OTERO, P. (2007). Learning bilingual lexicons from comparable English and Spanish corpora. *In Proceedings of MT SUMMIT*, pages 191–198.

GAUSSIER, É., RENDERS, J.-M., MATVEEVA, I., GOUTTE, C. et DÉJEAN, H. (2004). A geometric view on bilingual lexicon extraction from comparable corpora. *In ACL*, pages 526–533.

HARRIS, Z. (1954). Distributional structure. *Word*, pages 146–162.

HAZEM, A. et MORIN, E. (2012a). Adaptive dictionary for bilingual lexicon extraction from comparable corpora. *In Proceedings, 8th international conference on Language Resources and Evaluation (LREC)*, Istanbul, Turkey.

HAZEM, A. et MORIN, E. (2012b). Qalign :a new method for bilingual lexicon extraction from comparable corpora. *In Proceedings of CICLING*, India.

HWANG, M., CHOI, C. et KIM, P. (2011). Automatic enrichment of semantic relation network and its application to word sense disambiguation. *IEEE Transactions on Knowledge and Data Engineering*, 23:845–858.

KUN, Y. et TSUJII, J. (2009). Bilingual dictionary extraction from Wikipedia. *In Proceedings of MT SUMMIT*.

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

LI, B. et GAUSSIER, Ë. (2010). Improving corpus comparability for bilingual lexicon extraction from comparable corpora. *In 23rd International Conference on Computational Linguistics (Coling 2010)*, Beijing, China.

LIN, D. (1998). An information-theoretic definition of similarity. *In Proceedings of the Fifteenth International Conference on Machine Learning*, ICML '98, pages 296–304, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

MANNING, C. D., RAGHAVAN, P. et SCHTZE, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.

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

MORIN, E. et DAILLE, B. (2006). Comparabilité de corpus et fouille terminologique multilingue. *In Traitement Automatique des Langues (TAL)*.

MORIN, E. et PROCHASSON, E. (2011). Bilingual lexicon extraction from comparable corpora enhanced with parallel corpora. *In Proceedings, 4th Workshop on Building and Using Comparable Corpora (BUCC)*, page 27–34, Portland, Oregon, USA.

PROCHASSON, E. et MORIN, E. (2009). Points d'ancrage pour l'extraction lexicale bilingue à partir de petits corpus comparables spécialisés. *Traitement Automatique des Langues*, page 22.

PROCHASSON, E., MORIN, E. et KAGEURA, K. (2009). Anchor points for bilingual lexicon extraction from small comparable corpora. *In Proceedings, 12th Conference on Machine Translation Summit (MT Summit XII)*, page 284–291, Ottawa, Ontario, Canada.

RAPP, R. (1995). Identifying word translations in non-parallel texts. *In Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, ACL '95, pages 320–322. Association for Computational Linguistics.

RUBINO, R. et LINARÈS, G. (2011). Une approche multi-vue pour l'extraction terminologique bilingue. *In CORIA*, pages 97–111.

SADAT, F. et TERRASA, A. (2010). Exploitation de wikipédia pour l'enrichissement et la construction des ressources linguistiques. *In Proceedings of TALN*, Montréal, Canada.

VULIĆ, I. et MOENS, M.-F. (2012). Detecting highly confident word translations from comparable corpora without any prior knowledge. *In Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 449–459, Avignon, France. Association for Computational Linguistics.

WU, Z. et PALMER, M. (1994). Verbs semantics and lexical selection. *In Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, ACL '94, pages 133–138. Association for Computational Linguistics.