Université de Montréal

Induction de lexiques bilingues à partir de corpus comparables et parallèles

par
Laurent Jakubina

Département d'informatique et de recherche opérationnelle
Faculté des arts et des sciences

Thèse présentée à la Faculté des études supérieures
en vue de l'obtention du grade de Philosophiæ Doctor (Ph.D.)
en informatique

Juillet, 2017

RÉSUMÉ

Les modèles statistiques tentent de généraliser la connaissance à partir de la fréquence des événements probabilistes présents dans les données. Si plus de données sont disponibles, les événements sont plus souvent observés et les modèles sont plus performants. Les approches du Traitement Automatique de la Langue basées sur ces modèles sont donc dépendantes de la disponibilité et de la quantité des ressources à disposition.

Cette dépendance aux données touche en particulier la Traduction Automatique Statistique qui, de surcroît, requiert des ressources de type multilingue. Cette thèse rapporte quatre articles sur deux tâches qui contribuent de près à cette dépendance : l'Alignement de Documents Bilingues (ADB) et l'Induction de Lexiques Bilingues (ILB).

La première publication décrit le système soumis à la tâche partagée d'ADB de la conférence WMT16. Développé sur un moteur de recherche, notre système indexe des sites web bilingues et tente d'identifier les pages anglaises-françaises qui sont en relation de traduction. L'alignement est réalisé grâce à la représentation "sac de mots" et un lexique bilingue. L'outil développé nous a permis d'évaluer plus de 1000 configurations et d'en identifier une qui fournit des performances respectables sur la tâche.

Les trois autres articles concernent la tâche d'ILB. Le premier revient sur l'approche dite "standard" et propose une exploration en largeur des paramètres dans le contexte du Web Sémantique. Le deuxième article compare l'approche standard avec les plus récentes techniques basées sur les représentations interlingues de mots (*embeddings* en anglais) issues de réseaux de neurones. La dernière contribution reporte des performances globales améliorées sur la tâche en combinant, par reclassification supervisée, les sorties des deux types d'approches précédemment étudiées.

Mots clés: corpus parallèle, corpus comparable, alignement, induction lexique bilingue, embedding, représentation de mots, reclassement supervisé

ABSTRACT

Statistical models try to generalize knowledge starting from the frequency of probabilistic events in the data. If more data is available, events are more often observed and models are more efficient. Natural Language Processing approaches based on those models are therefore dependant on the quantity and availability of these resources. Thus, there is a permanent need for generating and updating the learning data.

This dependency touches Statistical Machine Translation, which requires multilingual resources. This thesis refers to four articles tackling two tasks that contribute significantly to this dependency: the Bilingual Documents Alignment (BDA) and the Bilingual Lexicons Induction (BLI).

The first publication describes the system submitted for the BDA shared task of the WMT16 conference. Developed on a search engine, our system indexes bilingual web sites and tries to identify the English-French pages linked by translation. The alignment is realized using a "bag of words" representation and a bilingual lexicon. The tool we have developed allowed us to evaluate more than 1,000 configurations and identify one yielding decent performances on this particular task.

The three other articles are concerned with the BLI task. The first one focuses on the so-called standard approach, and proposes a breadth parameter exploration in the Semantic Web context. The second article compares the standard approach with more recent techniques based on interlingual representation of words, or the so-called embeddings, issued from neural networks. The last contribution reports the enhanced global performances on the task, combining the outputs of the two studied approaches through supervised reclassification.

Keywords: parallel corpus, comparable corpus, alignment, bilingual lexicons induction, embedding, word representation, supervised reclassification

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES SIGLES

ACL	Association for Computational Linguistics
ADB	Alignement de Documents Bilingues
ASE	Analyse Sémantique Explicite
ILB	Induction de Lexiques Bilingues
LTR	Learning To Rank
MSD	Modèle de Sémantique Distributionnelle
TA	Traduction Automatique
TAL	Traitement Automatique de la Langue
TAS	Traduction Automatique Statistique

REMERCIEMENTS

Je tiens tout d'abord à remercier mon directeur de recherche Philippe Langlais pour m'avoir donné l'opportunité d'accomplir un doctorat dans le domaine du traitement automatique de la langue, une spécialité de recherche qui me passionne encore autant. Je tiens aussi à le remercier pour la confiance dont il m'a fait part ainsi que pour sa patience, y compris dans les moments difficiles, durant ces (presque) cinq dernières années. Je lui dis encore merci pour ses conseils, ses corrections et sa vision de la recherche dont je retiens un grand nombre de leçons précieuses.

Je remercie ensuite le "grand" Fabrizio Gotti pour les aides techniques, les discussions stimulantes, le support moral et sa vision ensoleillée[1] de la Belgique.

Mes remerciements vont aussi à l'ensemble des membres du RALI et amis du Département d'Informatique et de Recherche Opérationnelle (DIRO) de l'Université de Montréal pour la bonne ambiance qui m'a entourée pendant cette grande aventure.

Je remercie également le Fond de Recherche du Québec - Nature et Technologies (FRQNT) et le Département d'Informatique et de Recherche Opérationnelle (DIRO) pour leurs soutiens financiers.

Enfin, je veux remercier grandement mes parents, Françoise et Jean-Philippe, pour toujours avoir été derrière moi dans toutes les circonstances, mon frère Florian pour son support moral cynique et sa relecture de ce document, et surtout ma fiancée Teodora Dan qui m'encourage, me stimule intellectuellement et me soutient depuis le premier jour de notre rencontre.

[1] mais erronée :P !

CHAPITRE 1

PLURILINGUISME ET INFORMATIQUE

Selon le dictionnaire Larousse (Larousse, 2016), le **plurilinguisme** décrit la qualité d'une personne ou d'une région d'être plurilingue, c'est-à-dire d'avoir la capacité de comprendre ou de parler plusieurs langues. La même définition est donnée au mot **multilinguisme** – qui est utilisé plus couramment dans la littérature – faisant de ces deux termes, a priori, des synonymes. Cependant, si l'on se réfère au Guide pour l'élaboration des politiques linguistiques éducatives (DPL, 2007) du Conseil de l'Europe (COE, 2016), une distinction existe entre ces deux mots :

> “
>
> - le « **multilinguisme** » renvoie à la présence, dans une zone géographique déterminée – quelle que soit sa taille – à plus d'une «variété de langues», c'est-à-dire de façons de parler d'un groupe social, que celles-ci soient officiellement reconnues en tant que langues ou non. À l'intérieur d'une telle zone géographique, chaque individu peut être monolingue et ne parler que sa propre variété de langue ;
> - le « **plurilinguisme** » se rapporte au répertoire de langues utilisées par un individu ; il est donc, en un sens, le contraire du multilinguisme. Ce répertoire englobe la variété de langue considérée comme «langue maternelle» ou «première langue», ainsi que toute autre langue ou variété de langue, dont le nombre peut être illimité. Ainsi, certaines zones géographiques multilingues peuvent être peuplées à la fois de personnes monolingues et de personnes plurilingues.
>
> ”
>
> *DPL (2007)*

Datant d'un certain nombre d'années, cette distinction s'impose progressivement, des documents gouvernementaux – québécois (CSLF, 2016) – aux blogues personnels (Ca-

feMultilingue, 2016) sur Internet. Cependant, aucune définition ne spécifie le cas des ressources intégrant plusieurs langues telles que les dictionnaires de traduction. Nous avons donc décidé d'employer la forme la plus couramment utilisée – multilinguisme – quand nous nous référerons à ce type de ressources.

Amorcer une discussion autour du plurilinguisme, c'est rappeler la vaste variété de langues créées par les êtres humains afin de communiquer entre eux. Pour une partie d'entre nous, c'est aussi la nécessité d'en apprendre, voire d'en maîtriser plusieurs, et ce, pour des raisons scolaires, professionnelles ou d'héritages familiaux ou culturels. Apprendre une langue est difficile, en comprendre une deuxième peut s'avérer être un grand défi. De plus, l'Histoire nous a déjà maintes fois rappelé l'importance de la communication entre les individus, renforçant cette nécessité d'apprentissage. Pour certains, la solution se trouve dans la création d'une langue unique et commune, comme nous le raconte la légende de la tour de Babel. Plus récemment, d'autres ont vu en l'**informatique** un support possible au langage naturel et au **plurilinguisme** qui l'accompagne.

La science du traitement automatique et rationnel de l'information, ou **Informatique** est née dans les années 1950 avec la création des premiers ordinateurs (Wikipedia, 2016b). Ces derniers servant (et servent toujours) de support aux applications de cette science. Dès ses balbutiements, les chercheurs ont proposé des applications en relation avec le langage humain telles que le Test de Turing[1] et la Traduction Automatique (TA), donnant naissance à la discipline intitulée **Traitement Automatique de la Langue** ou TAL. Notamment, l'expression "traduction automatique" est mentionnée pour la première fois par Warren Weaver dans ses "Memorandum on Translation" (Weaver, 2016) datant de 1949, soit il y a plus de 65 ans.

Depuis, de très nombreuses disciplines et applications de recherche ont rejoint le champ du TAL : le résumé de textes, la reconnaissance de la parole (en lien avec le traitement du signal) ou encore la recherche d'informations et l'analyse de sentiments. Beaucoup de recherches, d'approches et d'améliorations ont été proposées pour chacune de ces

[1] Identification d'êtres humains via des conversations écrites.

disciplines, suivant les idées émises, les périodes d'intérêts, l'évolution des autres disciplines et applications connexes de même que les ressources disponibles. Les premières approches du TAL se sont basées sur l'**utilisation de règles**, chacune d'elles tentant soit de généraliser soit de spécialiser une solution à un problème particulier provenant de la linguistique. Ce type d'approches s'est montré très fastidieux, car il est difficile et long d'énumérer toutes les règles considérant le caractère dynamique du langage naturel. Dans les années 80, l'augmentation de la puissance des ordinateurs a rendu plus aisée l'exploitation massive des données, notamment textuelles, de façon automatique. Plus précisément, il a commencé à être possible d'induire ce que nous appelons des **modèles statistiques** à partir de données et de se servir de ces modèles à travers un nouveau type d'approches. Ces nouvelles approches ont surpassé les systèmes de règles, en évitant notamment le côté fastidieux de leurs identifications, et ont induit une forme de "révolution" dans le domaine du traitement automatique de la langue. Encore aujourd'hui, les plus récentes approches de la traduction automatique, basées sur les réseaux de neurones[2], sont héritières de cette "révolution statistique".

Les modèles statistiques tentent de généraliser la connaissance, par des règles, à partir de la fréquence des événements probabilistes présents dans les données. Si plus de données sont disponibles, les événements sont plus souvent observés et les règles sont mieux généralisées. Autrement dit, les modèles et les performances qui les accompagnent sont meilleurs. Les approches basées sur les statistiques sont donc dépendantes de la disponibilité, en premier lieu, et de la quantité, ensuite, de ressources. En parallèle à leurs exploitations par les approches statistiques est donc apparue la nécessité de **produire plus de ressources d'apprentissage**. Cette nécessité a depuis intéressé de nombreux chercheurs et est aujourd'hui la méta problématique de cette thèse, précisément **pour des données multilingues**.

La recherche autour des données multilingues, notamment leur constitution, a connu un regain d'attention dans les années 1990, parallèlement aux premières approches de la Traduction Automatique basée sur les modèles Statistiques (TAS). Des données sont

[2]Traduction Automatique Neuronale

dites multilingues quand elles incorporent des informations en plus d'une langue. Nous parlons de données bilingues lorsqu'elles couvrent précisément deux langues (exemple : anglais-français), ou trilingues pour exactement trois. Généralement, les données sont regroupées selon une perspective précise (politique, actualité, etc.) sous la forme d'ensembles de documents et forment ainsi un corpus. Différents types de données (texte, image, vidéo, etc.) existent et peuvent, dans certains cas, coexister au sein d'un document ou d'un corpus. Cependant, ce sont les données exclusivement textuelles qui nous intéressent par la suite. Plus précisément, les **corpus textuels bilingues**.

Nous devons distinguer les textes bilingues où les contenus écrits, respectivement dans la langue source et dans la langue cible, sont la traduction l'un de l'autre et les textes bilingues où ce n'est pas le cas. Dans la première situation, on dit aussi que les documents sont en relation de traduction et nous parlons de corpus parallèles quand l'ensemble des documents d'un corpus bilingue sont de ce type. Dans le cas contraire (non parallèle), nous parlons de corpus comparables (où un degré de comparabilité est à déterminer) et de corpus indépendants. Ces derniers sont simplement des ensembles de documents multilingues non reliés au niveau de leurs contenus.

Les corpus parallèles ont été (et sont) la ressource clé nécessaire au développement de la TAS. Une fois le corpus aligné au niveau de la phrase, la nature parallèle de celui-ci est exploitée par les modèles statistiques afin d'apprendre un **alignement** entre les mots de la langue source et ceux de la langue cible. Cet alignement est la pierre angulaire de la traduction automatique et des performances des systèmes de référence d'aujourd'hui, et pour le moment, un alignement aussi précis ne peut être obtenu sans ce parallélisme. Rejoignant la nécessité de construire plus de ressources d'apprentissage en général, il y a un besoin de créer davantage de données parallèles. Le terme "créer" est ici à prendre au sens large, vu qu'il peut s'agir autant d'identifier du contenu parallèle au sein de quantités d'informations plus volumineuses non parallèles que de tenter d'en composer manuellement. Cependant, il va de soi que la nature parallèle des ressources textuelles ne s'atteint pas sans un coût important, notamment en temps. En effet, les corpus parallèles sont généralement constitués par des traducteurs humains. À l'opposé,

l'identification de contenus (déjà) parallèles semble être une tâche plus simple et moins coûteuse. Cette tâche existe et est appelée **Alignement de Documents Bilingues (ADB)**. Très récemment, l'explosion de contenus textuels multilingues sur l'Internet a lancé un regain d'intérêt à l'égard de l'ADB afin de profiter de toutes ces ressources (Buck et Koehn, 2016).

L'ADB permet d'identifier des ressources parallèles. Le caractère dynamique du langage naturel nous oblige cependant à (ré-)actualiser nos ressources. En outre, des domaines ou genres de textes existent pour lesquels les contenus parallèles sont rares ou inexistants. C'est par exemple le cas des domaines de spécialité. Sans ressource pour couvrir le vocabulaire d'un domaine, les modèles statistiques ne gèrent pas ces événements spécifiques et ce faisant, ne peuvent fournir des performances satisfaisantes. C'est la même réalité pour le fameux **problème des mots rares** qui concerne l'ensemble des approches reposant sur des modèles statistiques. Ainsi, la disponibilité des corpus parallèles reste un problème. Considérant le fait que beaucoup plus de textes sont produits de façon monolingue, les chercheurs se sont aussi intéressés à l'utilisation des **corpus dits comparables** pour la TAS.

Les documents sources et cibles d'un corpus comparable ne sont pas des traductions l'un de l'autre, mais leur texte respectif partage des aspects communs tels que le thème, le genre littéraire, le style d'écriture et encore, la période de rédaction. Ainsi, des nouvelles bilingues rapportant des résultats sportifs, mais de deux sports différents peuvent être considérées comme comparables. La définition de (degré de) comparabilité, proposée par Sharoff et al. (2013), est non contraignante au niveau des aspects communs et il conviendra aux chercheurs d'identifier les traits souhaités ou nécessaires afin de répondre aux possibles exigences sur les données (Rapp, 2015).

Un exemple concret de corpus comparable est Wikipédia[3]. Très connu du grand public, Wikipédia est un site web qui tente d'élaborer une encyclopédie universelle collaborative libre. Chaque concept présent y est décrit dans une page web sous forme d'un article dans une langue spécifique. Une fonctionnalité clé de l'encyclopédie est de pou-

[3] `www.wikipedia.com`

voir atteindre l'article concernant le même concept, mais dans une autre langue grâce à un hyperlien spécifique appelé lien inter langue. L'ensemble des articles concernant un même concept reliés par ces liens inter langues sont donc des documents multilingues. La majorité des documents multilingues de l'encyclopédie sont considérés comparables (Patry et Langlais, 2011a) : ils décrivent un même concept et partagent donc une même thématique et un même vocabulaire spécifique. Cependant, le caractère collaboratif et libre de Wikipédia laisse grandement la possibilité que les articles aient été écrits par des auteurs différents, à des moments différents avec des styles différents.

L'absence de relation de traduction entre les documents d'un corpus comparable signifie qu'aucun alignement basé sur la position entre les mots sources et cibles n'est réalisable. Les premiers chercheurs à s'intéresser aux corpus comparables pour la traduction automatique ont donc amené l'idée d'obtenir l'alignement à partir des autres aspects communs disponibles, en particulier depuis le contexte de cooccurrence des mots. Ce choix à amené les chercheurs à travailler sur la tâche d'**Induction de Lexiques Bilingues (ILB)** appliquée aux corpus comparables.

Plus généralement, l'ILB se définit comme la tâche qui tente d'identifier des mots et leurs traductions à travers des corpus multilingues. Cette définition ne contraint pas l'ILB à un type de corpus en particulier. Ainsi, appliquée aux corpus parallèles, la tâche se repose sur l'alignement des mots fourni par la nature parallèle du corpus. Cette même nature garantit aussi que si un mot se trouve dans la partie source du corpus, une traduction est présente dans la partie cible. La probabilité d'identifier avec succès des traductions de mots est donc beaucoup plus élevée. Il faut cependant s'assurer que les corpus contiennent les mots cibles intéressants. Or, comme mentionné précédemment, les corpus parallèles sont plus rares que les corpus comparables. Ainsi, dans cette thèse, nous employons principalement l'ILB sur les corpus comparables[4]. Finalement, nous pouvons aussi voir la tâche comme une tentative d'identifier du contenu parallèle dans du contenu comparable, au niveau des mots.

Dans la recherche sur la traduction automatique statistique, les deux tâches présentées

[4] Après avoir vérifié la couverture de nos ressources parallèles.

ont énormément d'intérêt. L'ADB parce qu'elle répond directement au problème de la disponibilité des données parallèles. L'ILB y répond aussi indirectement, mais surtout, si une solution performante est proposée, la TAS bénéficiera d'alignements en provenance d'un type de ressources beaucoup plus disponibles, les corpus comparables. **Dans cette thèse, nous nous intéressons à ces deux tâches** et nous présentons un article sur la première et trois autres sur l'induction de lexiques bilingues en corpus comparables.

Le premier article décrit le système que nous avons soumis à la tâche partagée d'alignement de documents bilingues de la première conférence sur la traduction automatique (WMT16). Cet article correspond au chapitre 3. Le chapitre 4 rapporte des expériences que nous avons réalisées afin d'évaluer les performances possibles de la tâche d'ILB en corpus parallèles avant de l'expérimenter en corpus comparables. Le deuxième article propose de revenir sur la tâche d'induction de lexiques bilingues en corpus comparables (en utilisant l'approche nommée "standard") dans le cadre du Web Sémantique et constitue le chapitre 5. Le troisième article compare l'efficacité de l'approche standard de l'ILB avec les plus récentes techniques issues de l'exploitation des représentations interlingues de mots (appelées aussi **embeddings** en l'anglais) provenant de réseaux de neurones. Il s'agit du chapitre 6. Le 7^{e} chapitre couvre notre dernière contribution qui propose d'améliorer les performances globales de l'ILB en combinant les résultats de deux types d'approches d'induction de lexiques bilingues. L'ensemble de ces chapitres-articles sont précédés du chapitre 2 qui propose une revue de littérature pour chacune de ces tâches.

CHAPITRE 2

ÉTAT DE L'ART

Cette thèse porte sur deux tâches découlant des problèmes de la Traduction Automatique Statistique (TAS) et de ses ressources. Ces deux tâches sont l'Alignement de Documents Bilingues (ADB) et l'Induction de Lexiques Bilingues (ILB).

Ce chapitre est consacré à la présentation approfondie de ces deux tâches ainsi qu'à la revue de leurs travaux existants. La présentation consiste en une définition précise de la tâche ainsi que de ses notions. L'ADB est traitée dans la section 2.1 et l'ILB dans la section 2.2. L'induction de lexiques bilingues étant explorée par plusieurs branches de recherches, son état de l'art est divisé en trois sous-sections : la section 2.2.1 revoit les travaux qui ont suivi la recherche sur l'adaptation de la procédure d'alignement de mots en corpus parallèles vers les corpus comparables ; la section 2.2.2 tente une revue complète des récentes contributions qui ont employé la tâche d'ILB afin d'évaluer la qualité des représentations interlingues de mots issues de réseaux de neurones ; finalement, la section 2.2.3 présente les articles qui traitent la tâche comme une tâche de reclassement (supervisé) de candidats à la traduction.

2.1 Alignement de documents bilingues

L'Alignement de Documents Bilingues (ADB) est une tâche de recherche en traitement automatique de la langue dont l'objectif est d'identifier, au sein d'une collection de documents bilingues (ou multilingues), des paires de documents {langue source - langue cible} tel que les documents de la paire soient la traduction l'un de l'autre, autrement dit parallèles. Identifier des paires de documents en relation de traduction de manière automatique est une des solutions possibles à la constitution de corpus parallèles, la ressource nécessaire à l'entraînement des systèmes de traduction automatique statistique (chapitre 1). Il est aussi intéressant de mentionner l'existence d'une variante de la tâche tentant d'aligner des documents de type comparable (Sharoff et al., 2015), mais nous n'en discutons pas dans ce travail.

Afin d'apparier avec succès de potentielles paires source-cible de documents en relation de traduction dans une collection, il faut identifier une représentation ou modélisation pour chaque document source qui permette d'évaluer son degré de parallélisme avec chaque (représentation du) document cible. Cette représentation est conçue à partir de plusieurs informations en provenance des documents. Ces informations peuvent être de plusieurs types. Par exemple, pour un document de type page web, on distingue les informations issues du contenu en lui-même (mots, images, liens, etc.) des informations entourant ce contenu (ou méta-informations : structure HTML, paragraphes, etc.). Au sein de ces informations, on peut aussi distinguer des "sous" types d'informations tels que les informations latentes (langue(s) utilisée(s), caractéristiques littéraires, etc.) ou les informations de support (livre, journal, blogue, etc.), parmi d'autres. Les choix des types d'informations considérés pour la modélisation sont souvent ce qui distingue les travaux au sein de l'état de l'art de la tâche d'alignement de documents bilingues.

Ainsi, Resnik et Smith (2003) proposent d'identifier des documents bilingues sur le web en utilisant les conventions de nommage des pages web, autrement dit des URL, avant de les mesurer avec une métrique de distance d'édition. Ce type d'heuristique ne garantissant pas le caractère parallèle, d'autres auteurs ont exploité d'autres indices comme

la longueur des documents, leur structure HTLM ou leur contenu.

Parmi les approches les plus simples, Ma et Liberman (1999) utilisent un lexique bilingue afin de compter le nombre de paires de mots partagées entre le coté source et cible des documents parallèles. De leur côté, Enright et Kondrak (2007) classent les paires candidates en ordre décroissant du nombre d'hapax [1] que partagent les deux documents. Malgré sa simplicité, cette approche affiche des performances intéressantes qui en ont fait la référence à battre pour les approches subséquentes. Shi et al. (2006) utilisent des features résultant de l'alignement des phrases entre les documents d'une paire à évaluer. En plus d'être coûteuse en temps, la performance est dépendante de la qualité de l'alignement phrasique, qui est elle-même dépendante de la paire de langues considérée (notamment de ses ressources).

Quand les ressources en traduction (système de traduction, lexiques bilingues, etc.) sont disponibles, le temps de calcul peut être géré grâce à une approche distribuée. C'est ce que proposent Uszkoreit et al. (2010) qui commencent par traduire toutes les pages non anglaises en anglais (avec un système de traduction existant). Les mots et phrases les plus rares des documents sont ensuite utilisés pour comparer (par distance cosinus) les "vrais" documents anglais aux "faux" et évaluer ainsi leur degré d'alignement. L'approche peut être vue comme un système de détection de documents dupliqués interlangues.

Mais ces travaux dépendent d'informations qui sont spécifiques aux langues traitées. Ainsi, parmi les approches les plus récentes, Patry et Langlais (2011a) proposent d'entraîner un classificateur basé sur des entités nommées et numériques. Suivant le même objectif, Krstovski et Smith (2011) représentent chaque document de la collection bilingue à l'aide de l'ensemble des mots communs aux deux vocabulaires.

Malgré le nombre peu élevé de publications sur la tâche d'ADB, de très bonnes performances (95%) ont été atteintes par certaines d'entre elles (Krstovski et Smith, 2011, Uszkoreit et al., 2010). Mais leur reproduction est délicate en pratique en raison des

[1] Les hapax sont des mots qui n'apparaissent qu'une fois dans un document.

quantités de données manipulées ou du type d'approches proposé ((Uszkoreit et al., 2010) propose une approche distribuée nécessitant beaucoup de ressources machines). Récemment, l'explosion de l'utilisation d'Internet à l'échelle mondiale a contribué à la disponibilité de davantage de contenus multilingues dont nous aimerions profiter dans le cadre de la tâche. Les autres approches mentionnées précédemment (Enright et Kondrak, 2007, Ma et Liberman, 1999, Shi et al., 2006), plus accessibles selon les expériences des publications, sont elles aussi difficilement reproductibles en pratique de fait de la taille de l'Internet actuel. Alors que l'ensemble des publications commencent à dater, de nombreuses améliorations ont été réalisées depuis avec l'utilisation des modèles statistiques, afin de profiter de plus de données et sans nécessairement faire appel à des approches distribuées. Ainsi, il y a un besoin de revenir sur la tâche d'ADB et en 2016, (Buck et Koehn, 2016) a proposé une tâche partagée d'alignement de documents bilingues pour ces raisons.

2.2 Induction de lexiques bilingues

L'Induction de Lexiques Bilingues (ILB) est une tâche de recherche en Traitement Automatique de la Langue dont l'objectif est d'extraire des lexiques bilingues à partir de ressources bilingues, voire plurilingues, de manière automatique. À l'époque de la révolution numérique, c'est la tâche équivalente à la compilation manuelle de lexiques bilingues, ou de traduction, par des lexicographes.

Un lexique bilingue (ou dictionnaire de traduction) est une ressource (le plus souvent textuelle) où chaque ligne est composée d'une entrée (mot, expression...) dans une langue source (ex. : en anglais) et d'une ou plusieurs sortie(s) (mot, expression...) dans une langue cible (ex. : en français) appelée traduction. Un exemple de ce genre de ressources est présenté dans la table 2.I. L'information bilingue (ou plurilingue) à partir de laquelle l'ILB extrait un nouveau lexique est une autre ressource (le plus souvent textuelle) plus dense, potentiellement composée de documents, communément appelée corpus bilingue.

...		...		
certitude	certitude			
patchouli	patchouli			
stickler	pointilleuse	rigoriste		
lewdness	débauche	impudicité	luxure	obscénité
laxness	laxité	mollesse		
nightly	nocturne	nuitamment		
selectee	appelé			
gel	gel	gélifier	prendre	épaissir
foolhardiness	témérité			
astounding	abasourdissant	effarant	renversant	stupéfiant
...		...		

Tableau 2.I: Exemple de contenu d'un lexique bilingue simple.

Nous évaluons la performance de la tâche (et indirectement de son résultat, le lexique bilingue induit) lors d'expérimentations en utilisant un lexique bilingue de référence. Dans ce lexique de référence, les traductions ont été évaluées comme exactes au préalable. Lors d'une exécution de l'ILB, les traductions candidates sont comparées à celles

de référence et un point est attribué pour chaque bonne réponse. Plusieurs traductions candidates par mot peuvent être évaluées durant l'évaluation. Quand x traductions candidates sont vérifiées, nous disons que la performance est évaluée au rang x. Dans cet état de l'art, nous rapportons la performance au rang 1 (quand celle-ci est disponible) de toutes les approches d'induction de lexiques bilingues que nous discutons.

Nous pouvons reformuler l'induction de lexiques bilingues grâce au vocabulaire spécifié jusqu'ici comme la tâche qui tente d'identifier des traductions de mots de manière automatique dans des corpus bilingues. Dans l'introduction (chapitre 1), deux types de corpus bilingues ont été définis, basé sur le niveau de comparabilité[2] entre la partie cible et source du corpus : les corpus parallèles et les corpus comparables. Cette distinction est importante, car elle marque une séparation dans l'état de l'art (et des approches présentées) de la tâche, dépendamment du type de corpus bilingues.

Les premières approches d'induction de lexiques bilingues exploitèrent les corpus parallèles en utilisant la procédure d'alignement des mots de la TAS (Brown et al., 1990) comme point de départ. De fait, les premières publications sur la tâche parlèrent avant tout d'alignement de mots (Gale et Church, 1991) et il fallut attendre Smadja (1992) et Wu et Xia (1994) pour que l'expression "Induction de Lexiques Bilingues" soit attribuée à la tâche et apparaisse dans la littérature.

Les corpus parallèles sont une ressource peu abondante due à leurs coûts de création élevés (Sharoff et al., 2013). A l'opposé, les corpus comparables, résultant de la concaténation de corpus monolingues, sont plus largement disponibles. En conséquence, il n'a pas fallu attendre longtemps pour que l'idée d'une adaptation de l'ILB (sur corpus parallèles) se présente pour les corpus comparables. Plus précisément, l'idée de départ était de dériver la procédure d'alignement de mots en corpus parallèles pour les corpus comparables. C'est le point de départ de cet état de l'art (Section 2.2.1).

Dans la suite de ce document, ainsi que dans la majorité des publications référencées, le caractère comparable des corpus bilingues est implicite et la tâche est simplement

[2] Sharoff et al. (2013) proposent une discussion très détaillée sur le sujet.

intitulée "Induction de Lexiques Bilingues". Alors que la section 2.2.1 revient sur les travaux qui ont suivi la recherche sur l'adaptation de l'alignement de mots en corpus comparables, la section 2.2.2 tente une revue complète des très récents travaux qui ont employé la tâche d'ILB afin d'évaluer la qualité des représentations interlingues de mots (appelées aussi *embeddings* en anglais) issues de réseaux de neurones. Finalement, la section 2.2.3 présente les travaux qui traitent l'ILB comme une tâche de reclassement (supervisé) de candidats à la traduction.

2.2.1 Via l'alignement de mots en corpus comparable

Les premières approches d'ILB en corpus comparables vinrent de l'idée de dériver la procédure d'alignement des mots en corpus parallèles présentée par l'approche statistique de traduction automatique de Brown et al. (1990). En 1995, deux publications sur cette idée, rédigées indépendamment (Sharoff et al., 2013), sont acceptées à la conférence Association for Computational Linguistics (ACL) : Rapp (1995) et Fung (1995).

Rapp (1995) part du postulat suivant : "Si deux mots cooccurrent plus souvent que par le hasard dans une langue source, alors leur traduction doivent cooccurrer plus souvent que par le hasard dans la langue cible." Autrement dit, les mots en relation de traduction à travers des corpus comparables doivent avoir des motifs de cooccurrence, appelés aussi "contextes", similaires à travers les langues. Cette hypothèse rappelle l'hypothèse de distribution de Harris (1954). En pratique, Rapp (1995) représente ces motifs de cooccurrence par des matrices d'associations, une par langue, et un alignement tente d'être identifié en réorganisant (aléatoirement) les matrices afin que les motifs soient similaires. De son côté, Fung (1995) propose une approche basée sur l'hypothèse que les mots ayant des contextes productifs (respectivement, rigides) dans une langue source se traduisent par des mots dont le contexte est productif (respectivement, rigides) dans la langue cible. La productivité (ou rigidité) d'un mot est estimée selon une mesure d'hétérogénéité du contexte. Cette mesure compte le nombre de mots différents à gauche (resp. à droite) du mot cible. Nous observons que les deux travaux ont mis l'emphase sur la notion de "contexte" proposition de solution à l'alignement de mots en corpus

comparables.

Par la suite, la tâche a attiré l'attention d'un grand nombre de chercheurs et une approche, appelée communément approche "standard" (Bouamor et al., 2013a, Hazem et Morin, 2014) s'est imposée, basée sur l'hypothèse de cooccurrence de Rapp (1995). Dans le reste de cette section, nous commençons par décrire brièvement son fonctionnement et ensuite, nous passons en revue les travaux qui ont participé à l'étude de cette tâche depuis maintenant plus de 20 ans.

Computationnellement, chaque mot d'intérêt[3] est représenté par un vecteur de contexte, c'est-à-dire un vecteur modélisant les mots qui cooccurrent avec le mot d'intérêt. Une valeur de cooccurrence pour chaque mot-entrée du vecteur est calculée ; il peut s'agir d'un booléen (présence/absence), de fréquences ou d'un réel mesurant la force d'association entre le mot-entrée et le mot d'intérêt. Cette force d'association est aussi utilisée pour normaliser le vecteur. Ensuite, le vecteur de contexte du mot que l'on souhaite aligner (ou dont nous souhaitons identifier la traduction) est projeté dans la langue cible en traduisant les mots du vecteur grâce à un lexique bilingue dit "amorce". Enfin, le vecteur de contexte projeté est comparé aux vecteurs de contexte des mots (du vocabulaire) de la langue cible. La comparaison est évaluée par une mesure de similarité, dont le score résultant permet de classer les mots cibles comme potentielle traduction du mot source (ou de son vecteur projeté). De multiples paramètres interviennent tout au long des étapes de l'approche ; outre le type de la valeur de cooccurrence, le lexique bilingue amorce (et ses caractéristiques telles que sa taille, son domaine sémantique, etc.) et la mesure de similarité, il y a aussi la taille de la fenêtre de cooccurrence[4]. S'ajoute aussi une série de paramètres "implicites" : le corpus comparable (son domaine sémantique, sa taille, la fréquence de ses mots, etc.), les caractéristiques des données de tests (fréquence – fréquent vs rare – et nombre de mots – unigramme vs multimots –, etc.). L'ensemble de ces paramètres ont un impact sur les résultats. Pour des descriptions plus détaillées (et imagées) de l'approche standard, le lecteur est invité à consulter Bouamor

[3]Cela dépend de la langue du mot. Généralement, dans le langue source, il s'agit souvent des mots dont une traduction est cherchée. Dans la langue cible, il s'agit alors de tous les mots du vocabulaire.

[4]Le nombre de mots à droite et à gauche du mot d'intérêt qui sont intégrés dans le vecteur de contexte.

(2014) ou Laroche et Langlais (2010), parmi d'autres.

Notre revue des travaux antérieurs repose sur l'impressionnant état de l'art présenté par Sharoff et al. (2013) qui couvre une grande majorité des travaux jusqu'à 2012. Ainsi, nous proposons un résumé condensé qui place ces travaux en fonction de l'axe d'améliorations (de l'approche standard) sur lequel ils ont travaillé. Certains de ces travaux ne sont pas mentionnés dans cette section, car ils traitent, parfois indirectement, du problème de l'ILB comme une tâche de reclassement de candidats à la traduction qui est traitée plus loin dans ce chapitre. Après ce résumé, nous décrivons les travaux publiés depuis 2012[5].

Selon Bouamor (2014), l'élément clé de l'approche standard est le dictionnaire bilingue amorce. Ce dernier sert de pont (de traduction) entre la langue source et cible. L'ensemble de ses caractéristiques (taille, fréquence des mots, domaine sémantique, etc.) influence aussi les performances de la tâche, car les vecteurs de contexte sources projetés dépendent de ces caractéristiques. De fait, de nombreux auteurs se sont intéressés à son cas et aux problèmes hérités de celui-ci selon différentes stratégies. Le dictionnaire amorce peut proposer plusieurs traductions par entrée source. Il convient alors de choisir la plus adaptée – la moins ambiguë– lors de la projection du vecteur de contexte. Ainsi, (Gaussier et al., 2004, Ismail et Manandhar, 2010) proposent d'intégrer un processus de désambiguïsation à l'approche standard. Plus précisément, Gaussier et al. (2004) conçoivent une vue géométrique et tentent de décomposer les vecteurs des mots en fonction de leur(s) sens. Cependant, le contexte local peut ne pas fournir assez d'information pour réussir à choisir le bon sens. Poursuivant le même objectif, Chiao et Zweigenbaum (2004), Morin et Prochasson (2011a), Vulić et Moens (2012) utilisent de l'information provenant d'autres sources (par exemple : un autre dictionnaire de spécialité) afin d'améliorer le processus de désambiguïsation. D'autres tentent de remplacer totalement le dictionnaire bilingue amorce par une autre source pouvant servir de pont de traduction. Ainsi, Hassan et Mihalcea (2009) remplacent le dictionnaire par

[5] `http://www.statmt.org/survey/Topic/DictionariesFromComparableCorpora` énonce une partie des travaux plus récents.

les liens inter-langues de Wikipédia. Enfin, les derniers essaient purement de le supprimer, autrement dit d'adapter l'approche standard afin d'éviter son utilisation. Ainsi, Diab et Finch (2000) utilisent quelques traductions amorces afin d'induire une fonction de correspondance (*mapping* en anglais) ou de traduction entre les termes sources et cibles. Cette fonction est dérivée d'une métrique de distance entre chaque paire de termes source et chaque paire de termes cibles. Cette approche d'amorçage est cependant coûteuse en temps de calcul (Bouamor, 2014, Sharoff et al., 2013). Haghighi et al. (2008) proposent un modèle génératif (basé sur l'analyse de corrélation canonique) qui apprend à partir de traits caractéristiques[6] tels que les cooccurrences et l'orthographe des mots. Ce modèle est ensuite utilisé pour générer de nouvelles paires de mots en relation de traduction.

Outre le dictionnaire bilingue amorce, nous avons présenté précédemment les méta-paramètres qui influencent la performance de l'approche standard. Idéalement, nous aimerions identifier les valeurs de ces paramètres qui maximisent les performances de la tâche. Ce genre d'explorations en largeur des paramètres ont été proposées par Chiao et Zweigenbaum (2002) ainsi que par Laroche et Langlais (2010).

Comme toute approche statistique basée sur la fréquence des événements probabilistes, l'insuffisance d'apparitions d'événements peut produire certains biais. L'approche standard représente les mots par des vecteurs de contexte qui utilisent la fréquence des mots et de leurs cooccurrents. Si ces fréquences sont faibles, les représentations manquent de généralisation et sont biaisées. Cette situation se produit avec les mots rares et est, de fait, un des problèmes difficiles qui touche l'approche standard. Cependant, rares sont les auteurs (Pekar et al., 2006, Prochasson et Fung, 2011) qui se sont intéressés à cette problématique.

Finalement, nous terminons en énonçant des travaux qui se sont concentrés sur une problématique (ou amélioration) précise de l'approche standard. Rapp (1999) propose de prendre en compte l'ordre des mots du contexte dans les vecteurs de contexte et obtient une précision de 72% sur une liste de mots issue du domaine de la psychologie.

[6] *Features*, en anglais.

Li et Gaussier (2010) introduisent une mesure de comparabilité des corpus comparables. L'idée est d'évaluer à quel point des corpus sont comparables et, en conséquence, de les nettoyer afin d'améliorer la performance de l'approche standard telle qu'elle est proposée au départ. Garera et al. (2009) créent des vecteurs de contexte en utilisant des relations de dépendances entre les mots et des équivalences de catégorie grammaticale[7] plutôt que la cooccurrence des mots du contexte.

Rapp et al. (2012) proposent une approche d'alignement de mots en corpus comparable qui ne repose pas sur un lexique bilingue amorce mais qui nécessite un ensemble de documents comparables alignés (ex : Wikipédia). L'approche d'alignement se déroule en deux étapes ; premièrement, des termes clés pour chaque document sont identifiés en mesurant leur "facteur-clé" dans l'article courant et en comparaison avec ses occurrences dans le reste du corpus. Cette mesure (ex : log-likelihood) se base sur les fréquences de chaque mot. Deuxièmement, un poids entre chaque mot de la langue source et de la langue cible est induit par un algorithme de style réseau de neurones (inspiré par un connectionist WINner-Takes-It-All Network). À la fin de ses itérations, les poids offrent une mesure de la probabilité de traduction. Durant les itérations, les poids entre des paires de mots sont incrémentés s'ils apparaissent dans une paire d'articles alignés. Inversement, leurs poids diminuent s'ils occurrent individuellement. Ils évaluent leur approche sur 9 langues, dont la langue cible est l'anglais. Leurs corpus comparables alignés proviennent de Wikipédia et ont été minimalement prétraités. Ils ont préparé une liste d'évaluation de 1000 termes, traduite en 8 langues à l'aide de Google Traduction. Ils se comparent à deux approches de référence, Rapp (1999) et une de ces adaptations sur les données de Wikipédia. Ils affichent des performances améliorées entre 9 et 11% par rapport à ces derniers. Cependant, la comparaison exacte avec les références n'a pu se faire que sur un sous-ensemble de 100 mots. Finalement, une reproduction de cette approche a été présentée pour les termes multi-mots (Rapp et Sharoff, 2014).

Tamura et al. (2012) montrent le problème de la perte d'information de cooccurrence

[7] Part-of-speech, en anglais.

quand certains mots du contexte sont absents du lexique bilingue amorce[8] lors de la projection des vecteurs de contexte vers la langue cible. Ils proposent alors d'intégrer l'information de cooccurrence des cooccurrents (relation indirecte) en plus de l'information des cooccurrents (relation directe) dans les vecteurs de contexte. En pratique, ils représentent l'information de cooccurrence entre deux mots sous forme d'un graphe de relation (plutôt qu'une matrice), et implémentent leur idée grâce à une technique (de graphe) de propagation d'informations (label). Ils proposent aussi une seconde version où l'information de cooccurrence est remplacée par l'information de similarité. Ils évaluent leurs approches sur un corpus comparable prétraité anglais-japonais dans le domaine de la Physique. Ils construisent deux lexiques bilingues initiaux, de taille petite et large, et sélectionnent 1000 mots japonais (sens japonais vers anglais) de genre nominal pour l'évaluation. Ils se comparent à deux approches de référence, Rapp (1999) ainsi que Andrade et al. (2010) et affichent des performances multipliées par 2. Ils terminent en étudiant l'impact de la taille du lexique bilingue et de la fréquence des mots sur les performances.

Hazem et al. (2013b) étudient l'impact de 5 techniques de lissage, appliquées sur les comptes de cooccurrences, dans l'approche standard de Rapp (1995). Ils expérimentent leurs techniques sur deux corpus comparables anglais-français prétraités, un dictionnaire bilingue amorce de 4000 mots et s'évaluent sur une liste de référence de (321 + 100) entrées. Ils observent une augmentation de performance entre 2 et 10% en moyenne (en comparaison avec l'approche standard), en fonction de la technique et certains paramètres de l'approche standard. Ils observent aussi l'efficacité des différentes techniques de lissage en fonction de la fréquence des mots.

Bouamor et al. (2013b) s'intéressent à la polysémie dans le cadre de l'ILB. Plus précisément, ils proposent l'utilisation d'un processus de désambiguïsation (du sens des mots) au sein de l'approche standard afin d'améliorer la représentation des vecteurs de contextes. Techniquement, le processus de désambiguïsation intervient pendant la

[8]La taille/couverture du lexique bilingue amorce a un grand impact sur les performances de l'approche.

projection des vecteurs sources, et ils tentent de choisir la meilleure traduction possible pour les mots polysémiques du vecteur source. Ce choix est basé sur l'information donnée par 3 mesures de similarité sémantique et 2 mesures de parenté, toutes dérivées de `WordNet` à partir des mots et traductions fournis dans le dictionnaire bilingue amorce. Ils expérimentent leur processus sur deux corpus comparables prétraités à domaines spécifiques sur la paire de langues français-anglais. Ils se comparent à l'approche standard, ainsi qu'à une version améliorée (Morin et Prochasson, 2011a), en utilisant 2 listes de référence de 125 entrées et 79 entrées respectivement. Ils montrent des performances améliorées de l'ordre de 15 à 20% avec la métrique F-Measure au rang 20.

Bouamor et al. (2013a) proposent une nouvelle approche d'ILB basée sur Wikipédia, plus précisément ils adaptent l'Analyse Sémantique Explicite (Gabrilovich et Markovitch, 2007) (ASE) afin d'attaquer les deux problèmes inhérents à l'utilisation du dictionnaire bilingue amorce dans l'approche standard : sa couverture et son ambiguïté. L'ASE est une méthode qui utilise Wikipédia comme base de connaissance pour représenter des textes, allant de termes simples à des documents entiers, dans un espace sémantique structuré. L'adaptation proposée représente les termes à traduire dans l'espace des titres de Wikipédia. L'étape de transposition vers la langue cible se réalise alors par l'utilisation des liens inter-langues qui sont disponibles pour un certain nombre d'articles de Wikipédia. Ils comparent leur approche à l'approche standard revisitée par Morin et Prochasson (2011a) et Bouamor et al. (2013b) sur quatre corpus comparables prétraités à domaines spécifiques écrits en deux paires de langues (français-anglais et roumain-anglais). Les 8 listes de référence sont composées de 100 unigrammes. Ils affichent des augmentations moyennes de performances de l'ordre de 35 à 40% avec la métrique F-measure au rang 20, dépendamment de la paire de langues.

Morin et Hazem (2014) remettent en question l'hypothèse que les corpus comparables doivent être équilibrés –c'est à dire que la partie source et cible doivent être de taille similaire – pour la tâche d'ILB. Ils étudient ainsi la performance de l'approche standard Rapp (1995) en faisant varier la taille de la partie source (anglais) du corpus comparable et observent que les corpus peuvent être déséquilibrés. Plus précisément, ils observent

des performances améliorées de l'ordre de 10% quand les corpus le sont car plus de données permet une meilleure modélisation des vecteurs de contexte. Leurs expériences utilisent des données très similaires à Hazem et al. (2013b). Ils proposent aussi d'étudier l'impact d'un certain nombre de fonctions de prédiction (modèle de régression linéaire, non linéaire, logistique, etc.) sur les comptes de cooccurrences. En effet, ces modèles permettent d'ajuster les comptes quand les données d'entraînement sont (trop) faibles en quantité. Ils montrent des performances légèrement supérieures (1%) à l'approche sans prédiction.

Linard et al. (2015) visent à améliorer la qualité de l'ILB quand elle est appliquée entre des paires de langues pauvres en ressources (corpus et lexique bilingue amorce). Pour ce faire, ils proposent d'utiliser une troisième langue, appelée langue pivot, qui a elle des ressources de haute qualité. Deux approches sont présentées : la première applique l'approche standard entre la langue source et la langue pivot, et ensuite applique l'approche standard entre la langue pivot et la langue cible. La deuxième applique l'approche standard des deux langues initiales (source et cible) vers la langue pivot. Autrement dit, les deux langues d'intérêt sont projetées dans le même espace vectoriel que la langue pivot. Ces deux approches sont comparées à l'approche standard sur des corpus comparables de domaines spécifiques et des listes de références composées d'unigrammes et de multi-mots sur 4 langues (anglais, français, espagnol et allemand) dont 2 pivots (anglais, français). Les corpus sont prétraités et les vecteurs de contexte sont créés selon une configuration fixée. Ils affichent des résultats de l'ordre de 10% supérieurs à l'approche standard, selon le niveau de comparabilité du corpus. Plus précisément, les approches montrent des résultats meilleurs quand les corpus ont un faible niveau de comparabilité. Dans le cas contraire, l'approche standard reste la meilleure.

2.2.2 Comme tâche d'évaluation des embeddings interlingues

La tâche d'Induction de Lexiques Bilingues (ILB), définie dans la section 2.2.1 et introduite par Rapp (1995), a attiré l'attention d'un grand nombre de chercheurs. Un certain nombre d'entre eux ont porté une plus grande attention à la notion de vecteur de contexte – la représentation des mots basée sur leurs motifs de cooccurrence – afin de mieux en comprendre les propriétés. Ce faisant, leurs travaux ont aussi abordé un autre domaine de recherche de la linguistique computationnelle : la sémantique distributionnelle.

La sémantique distributionnelle est un champ de recherche qui étudie des modèles et des méthodes afin de représenter les similarités sémantiques entre des objets linguistiques, basé sur leurs propriétés distributionnelles. Ces dernières peuvent être obtenues via d'importants échantillons de données langagières (Wikipedia, 2016a). Cette définition se substitue sans équivoque à l'hypothèse bien connue de distribution[9] de Harris (1954) : "Les objets linguistiques qui ont des distributions similaires ont des significations similaires". Les premiers travaux sur la sémantique distributionnelle prennent leurs racines dans les années 1950. Depuis, de nombreux modèles – appelés Modèles de Sémantique Distributionnelle (MSD) (et de nombreux synonymes : représentations distribuées, espace vectoriel sémantique ou encore vecteurs de mots) – ont été proposés (Gavagai, 2015), souvent directement influencés par les tâches du TAL pour lesquelles ils sont étudiés : reconnaissance d'entités nommées, étiquetage de catégories grammaticales, analyse syntaxique, regroupement de documents, modèles de langues et induction de dictionnaires.

En pratique, un MSD aspire à représenter un terme en transformant son information distributionnelle en un vecteur[10], en prenant en compte un certain nombre de paramètres. Ensuite, grâce à l'algèbre linéaire comme outil de calcul, de nombreuses opérations et manipulations sont possibles sur ces vecteurs. Par exemple, la similarité sémantique entre deux termes peut s'évaluer grâce au calcul de la distance entre les vecteurs

[9] Voir Sahlgren (2008) pour une discussion plus poussée.

[10] D'où la notion de "vecteur" très présente dans le vocabulaire de la sémantique distributionnelle.

de ces termes. Le modèle de "vecteurs de contexte" proposé par Rapp (1995) est un de ces MSD qui utilise l'information contextuelle (précisément : la cooccurrence avec les mots-voisins) d'un mot pour représenter ce dernier sous forme d'un vecteur (nommé : "de contexte"). L'étape d'alignement des vecteurs, dont le résultat induit le lexique bilingue, correspond au calcul de la distance entre chaque vecteur de chaque mot de la langue source avec chaque vecteur de chaque mot de la langue cible.

Récemment, dû à la popularité grandissante des méthodes d'apprentissage machine, une pléthore de nouveaux MSD ont été présentés sous l'expression (anglaise) *Word Embedding*[11]. Derrière ce terme se présentent des représentations de mots[12] issues d'approches basées sur des réseaux de neurones. En pratique, ces dernières tentent de modéliser les mots en induisant un espace vectoriel continu (à valeurs réelles) à peu de dimensions (dense) avec la contrainte que les mots qui ont des distributions similaires sont situés à proximité les uns des autres dans cet espace vectoriel. Parmi les approches publiées, beaucoup d'attention a été portée aux modèles *Continuous Bag-of-Words Model (CBOW)* et *Continuous Skip-gram Model (SKG)* de Mikolov et al. (2013a). En effet, ces derniers ont permis d'atteindre des résultats état-de-l'art sur un certain nombre de tâches monolingues. Et ce, relativement facilement grâce au toolkit fourni conjointement aux modèles : Word2Vec. L'émergence des plongements de mots[13] a contribué à certains progrès de la sémantique distributionnelle ainsi qu'au traitement automatique de la langue en général.

Naturellement, il se pose la question de leurs impacts sur les tâches plurilingues (dont la tâche d'induction de lexiques bilingues fait partie). En grande partie parce qu'il serait intéressant de transférer la connaissance généralisée d'une langue dotée de beaucoup de ressources (telle que l'anglais) vers des langues moins dotées (la majorité des autres langues). Or apprendre à généraliser est l'objectif premier de l'apprentissage machine

[11] Anecdote : en 2015, un certain nombre de personnes ont proposé de renommer la conférence "Empirical Methods in NLP" (EMNLP) en "Embedding Methods in NLP" due à la grande quantité de publications dédiées aux Word Embeddings (Gavagai, 2015).

[12] Donc, toujours des vecteurs.

[13] *Embeddings* en français. Notons que nous utiliserons le terme "embedding" par la suite pour faciliter la rédaction.

(et de sa communauté). Une partie de l'intérêt s'est alors porté vers un nouveau "type" de embeddings : les embeddings bilingues ou interlingues. Ces représentations sont induites en tentant d'intégrer la dimension bilingue ou plurilingue au sein de l'espace vectoriel. Plus précisément, partant de l'hypothèse que des régularités linguistiques peuvent se généraliser en explorant conjointement un grand nombre de langues, le modèle contraint les représentations vectorielles de mots similaires[14] dans plusieurs langues à être similaires. Une grande variété d'embeddings bilingues ont été présentés, qui diffèrent selon le type de signal bilingue (ou plurilingue) et par la façon dont il est utilisé durant l'entraînement. Ruder (2016) présente une revue de littérature très complète sur ce sujet, en classant aussi l'ensemble des représentations selon le type d'apprentissage plurilingue (projection monolingue, pseudo-interlingue, optimisation jointe...) et selon le type de données du signal plurilingue (lexique bilingue, données parallèles alignées au niveau de la phrase, du mot...). Cette classification est très similaire à celle proposée par Vulić et Korhonen (2016).

Dans la recherche sur les embeddings interlingues, la tâche d'ILB a une définition légèrement différente, car elle n'est pas un but en soi. En effet, elle est vue comme une tâche d'évaluation de la qualité des embeddings interlingues induits[15] (Ammar et al., 2016). Plus précisément, elle évalue le degré de proximité de mots similaires (traductions) dans différentes langues dans l'espace (plurilingue) des embeddings (Duong et al., 2016). Cependant, sa finalité[16] n'est pas différente et c'est la raison pour laquelle ces approches sont discutées ici. Nous reprochons quand même à la grosse majorité de ces publications d'avoir ignoré leur lien avec la sémantique distributionnelle, et donc avec la tâche d'ILB (Gavagai, 2015, Rapp, 2015). Plus que d'être une mauvaise pratique scientifique, c'est surtout le risque de réinventer la roue plutôt que de construire sur de la connaissance existante. Aussi, du fait d'avoir utilisé la tâche d'ILB plutôt comme une tâche

[14]Mots qui ont des propriétés syntaxiques et sémantiques semblables dans leur langue respective, p. ex. les traductions.

[15]D'ailleurs, à l'exception de Duong et al. (2016), Vulić et Korhonen (2016), Vulic et Moens (2015), la tâche n'est pas référencée sous l'expression d'ILB.

[16]Identifier des traductions de mots, dans des corpus comparables.

d'évaluation[17] de la qualité et de la performance de leurs représentations, rares sont les auteurs qui ont étudié la tâche en tant que telle (comme les travaux de la section 2.2.1).

Évidemment, tous les modèles d'embedding publiés n'ont pas été évalués sur la tâche d'ILB. Selon l'objectif de leur représentation, d'autres tâches sont utilisées à cet effet (Ruder, 2016). Dans la suite de cette section, nous révisons les publications qui ont porté un intérêt explicite à la tâche d'induction de lexiques bilingues.

Mikolov et al. (2013b) commencent par observer que les embeddings d'une série de mots et de leur traduction respective, une fois projetés dans un espace vectoriel à deux dimensions[18] pour chacune des langues, ont des dispositions géométriques très similaires (Figure2.1). Cette observation suggère qu'il est possible, dans un premier temps, d'apprendre une projection linéaire (aussi appelée : matrice de transformation) d'un espace à un autre à partir d'une liste de paires de $(mot; traduction)$ de référence. Cette liste de paires d'entraînement est habituellement appelée dictionnaire bilingue (amorce). Dans un deuxième temps, cette projection peut être utilisée pour transformer l'embedding d'un mot de la langue source – dont la traduction n'est pas connue a priori – en un embedding "cible"[19] à placer dans l'espace des embeddings de la langue cible. Finalement, la similarité cosinus est utilisée pour trouver la représentation cible qui est la plus proche de la représentation projetée. En théorie, la plus proche représentation cible est celle de la traduction du mot source choisi initialement.

[17] D'ailleurs, Faruqui et al. (2016) critique l'utilisation de ce genre de tâches à des fins d'évaluations des représentations de mots.

[18] Grâce à des méthodes comme PCA (Analyse en composantes principales) ou t-SNE (t-distributed stochastic neighbor embedding).

[19] Que l'on pourrait aussi appelé "embedding traduit".

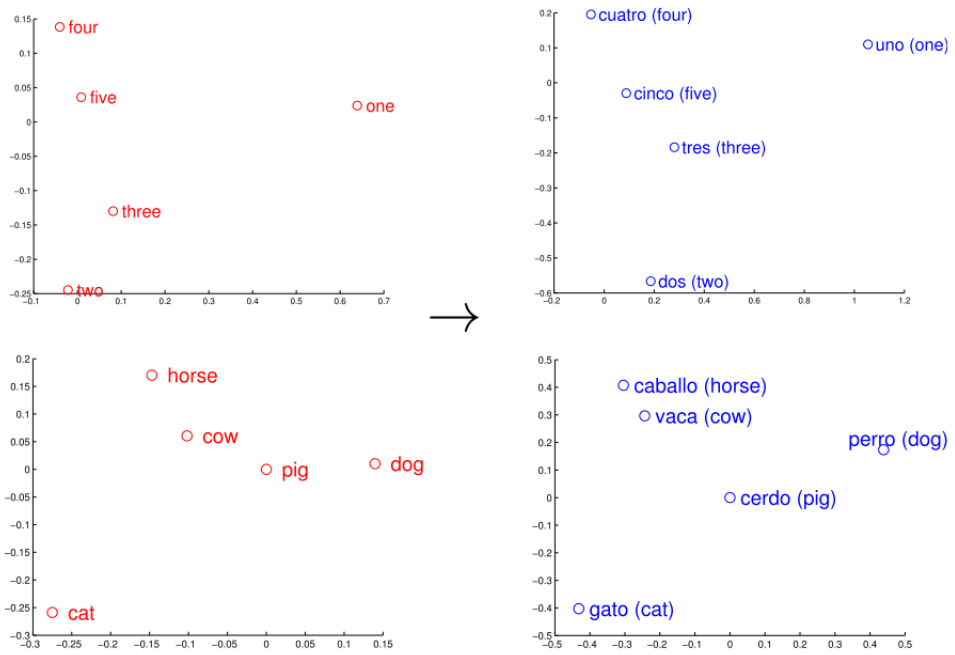

Figure 2.1 : Les embeddings des mots et de leur traduction ont des dispositions géométriques similaires dans l'espace vectoriel (Mikolov et al., 2013b).

Formellement, la matrice de transformation W se calcule en tentant de minimiser la distance entre les représentations monolingues des mots sources $\hat{x}_i$ et celles des traductions cibles $\hat{y}_i$ des paires obtenues par le dictionnaire bilingue. Ce problème d'optimisation (Eq. 2.1) peut être résolu par une descente de gradient stochastique.

$$\min_{W} \sum_{i=1}^{n} ||W\hat{x}_i - \hat{y}_i||^2 \tag{2.1}$$

Mikolov et al. (2013b) comparent leur approche à deux approches de référence : une basée sur la distance d'édition entre les mots et l'autre est l'approche présentée dans la section 2.2.1. Ils construisent 3 corpus monolingues à partir des données textuelles de WMT11[20] en anglais, espagnol et tchèque. Leur dictionnaire bilingue amorce est généré en prenant les 5000 mots les plus fréquents des corpus, et en les traduisant avec

[20] `http://www.statmt.org/wmt11/`

Google Traduction. Les 1000 prochains mots de cette liste sont conservés pour évaluer la performance de l'approche sur la tâche d'induction de lexiques bilingues. Enfin, les embeddings sont générés avec l'architecture Continuous Bag-of-Words (CBOW) du toolkit Word2Vec[21]. Concernant les résultats, leur modèle présente des performances supérieures de 15 à 20% en TOP@1 et TOP@5 en comparaison à l'approche des vecteurs de contexte. Par la suite, Mikolov et al. (2013b) étudient l'impact de la taille du dictionnaire bilingue amorce et de la fréquence des mots sur les performances. Notamment, ils montrent et appuient le besoin de plus de données monolingues afin d'améliorer la traduction des mots rares avec ce genre d'approche.

Xing et al. (2015) avancent que les travaux de Mikolov et al. (2013b) reposent sur un certain nombre d'inconsistances et proposent de les résoudre. Dans un premier temps, ils normalisent les vecteurs (embeddings) afin de corriger le fait d'utiliser le produit scalaire pendant l'entraînement alors que c'est la distance cosinus qui est employée pour vérifier si les vecteurs de mots sont bien proches les uns des autres selon leur sémantique. Dans un second temps, ils présentent une transformation orthogonale (à la place de la transformation – projection – linéaire). En effet, alors que la proximité des mots dans l'espace de projection est mesurée par la distance cosinus, c'est la distance d'Euler qui est utilisée dans la fonction objectif du calcul de la matrice de transformation. Cette deuxième correction permet aussi de respecter la normalisation réalisée en premier lieu. Pour vérifier l'impact de leurs corrections, ils suivent le même protocole que Mikolov et al. (2013b) mais sur la paire de langues anglais-espagnol. Ils observent des augmentations de l'ordre de 10% en fonction de la taille des embeddings.

Vulic et Moens (2015) proposent une approche simple pour apprendre directement des embeddings bilingues. Ils commencent par créer un corpus appelé "pseudo-bilingue" où les documents alignés d'un corpus comparable (ex : Wikipédia) sont concaténés (par paire), pour ensuite mélanger aléatoirement les mots de telle sorte que les mots de la langue source servent de contexte aux mots cibles et vice et versa. Ensuite, les

[21] Architecture et toolkit rendus populaires par les mêmes auteurs, présentés auparavant dans (Mikolov et al., 2013a).

embeddings sont induits de ce corpus pseudo-bilingue avec l'architecture Skip-gram du toolkit Word2Vec (Mikolov et al., 2013a). L'avantage suggéré par leur approche est de ne reposer sur aucune donnée parallèle (ni corpus bilingues alignés au niveau de la phrase, ni dictionnaire amorce) pour entraîner les embeddings, à l'inverse de (Mikolov et al., 2013b). Ils expérimentent leur approche sur la tâche d'induction de lexiques bilingues en se comparant à 3 approches de référence dont l'une est l'approche des vecteurs de contexte (Section 2.2.1). La performance est évaluée en TOP@1 sur 1000 paires $(mot; traduction)$ de référence. Ils utilisent la version 2013 de Wikipédia sur 3 paires de langues (espagnol-anglais, italien-anglais et néerlandais-anglais) et ont filtré le corpus afin de ne conserver que les syntagmes nominaux. Leur approche montre des résultats entre 5 et 10% supérieurs à l'approche des vecteurs de contexte. La taille de la fenêtre de contexte semble avoir un large impact sur les performances.

Gouws et al. (2015) proposent une architecture, appelée BilBOWA (Bilingual Bag-of-Words without Alignements[22]), pour induire directement des embeddings bilingues. Doutant de la capacité d'une projection (linéaire (Mikolov et al., 2013b) ou non Xing et al. (2015)) de capturer toutes les relations entre les mots de la langue source et cible, ils définissent une nouvelle fonction objectif (pour réseau de neurones) à deux objectifs : d'une part, les mots similaires d'une seule langue doivent avoir des représentations similaires (objectif monolingue) ; d'autre part, les mots similaires entre des langues (traductions) doivent avoir des représentations similaires (objectif interlingue). En apprentissage machine, ceci est appelé une optimisation jointe. La formulation proposée permet d'entraîner les embeddings sur des données monolingues, tout en les alignant en utilisant un signal bilingue extrait d'un petit ensemble de données parallèles alignées au niveau de la phrase. Comme Vulic et Moens (2015), l'approche ne requiert donc pas de dictionnaire bilingue amorce. Sur la tâche d'induction de lexiques bilingues, Gouws et al. (2015) reproduit une configuration expérimentale semblable[23] à Mikolov et al. (2013b) pour comparer les performances de leur architecture, mais seulement sur la paire de langues anglais-espagnol. Leurs embeddings, de taille 40, sont induits des don-

[22]En français : sac de mots bilingues sans alignements.

[23]Ils n'utilisent pas les 5000 termes les plus fréquents pour entraîner une matrice de traduction.

nées monolingues de Wikipédia et des données parallèles d'Europarl. Ils affichent des améliorations de l'ordre de 6 à 9% en TOP@1 (par rapport à l'approche de Mikolov et al. (2013b)).

Coulmance et al. (2015) suivent Gouws et al. (2015) et proposent une nouvelle architecture – autrement dit, une nouvelle fonction objectif pour réseau de neurones – appelée Trans-Gram, basée sur une optimisation jointe des objectifs monolingues et interlingues. Plus précisément, parce que la fonction objectif Skip-gram (proposée par Mikolov et al. (2013a)) a montré sa capacité intrinsèque à extraire des informations linguistiques intéressantes, ils souhaitent utiliser une fonction objectif pour l'objectif interlingue qui serait plus proche de Skip-gram que la fonction proposée par BilBOWA (Gouws et al., 2015). De plus, ils ajoutent la possibilité d'entraîner leur architecture sur plus d'une paire de langues à la fois, via une langue pivot (l'anglais dans leur cas). Pour leurs expériences, ils ont entraîné des embeddings de taille 300 sur 21 langues à l'aide de Wikipédia 2013 et d'Europarl (corpus aligné au niveau de la phrase, où toutes les langues sont pivots à l'anglais). Ils testent leur modèle d'embeddings (entraîné sur 21 langues) sur la tâche d'induction de lexiques bilingues, suivant le protocole de (Gouws et al., 2015, Mikolov et al., 2013b) sur la paire de langues anglais-espagnol (et opposé). Ils affichent des performances améliorées d'environ 5% en TOP@1 et de 10% en TOP@5 sur le modèle de BilBOWA. Ils concluent en montrant que leurs embeddings interlingues réussissent à intégrer des propriétés interlingues intéressantes aussi pour d'autres tâches.

Vulić et Korhonen (2016) s'interrogent sur l'impact du dictionnaire bilingue amorce, tel qu'utilisé par les approches de (Dinu et Baroni, 2014, Mikolov et al., 2013b, Vulic et Moens, 2015, Xing et al., 2015). Pour commencer, ils définissent deux propriétés que tout modèle d'embedding bilingue devrait respecter. Ces deux propriétés impliquent de replacer le dictionnaire bilingue amorce au centre des approches d'induction d'embeddings, en opposition avec les approches de (Coulmance et al., 2015, Gouws et al., 2015) par exemple. De plus, les auteurs rappellent que beaucoup de travaux présupposent avoir accès à des lexiques bilingues de bonne qualité mais que la "bonne qualité" d'un

lexique bilingue n'a jamais été clairement définie. Pour contrer cela, ils proposent une nouvelle approche : il s'agit d'apprendre un premier espace d'embeddings bilingues, d'y appliquer l'induction de lexiques bilingues, de récupérer les paires $(mot;traduction)$ apprises afin d'induire un nouvel espace d'embeddings bilingues. En respect avec les propriétés énoncées précédemment, ils proposent aussi que le premier espace d'embedding soit induit d'une de leurs précédentes approches (Vulic et al., 2016), qui ne requiert que l'alignement de documents comparables comme signal bilingue. Les performances de la deuxième induction étant grandement influencées par le lexique bilingue induit, ils analysent l'impact de sa taille et de la fiabilité des paires extraites dans ce lexique lors de leurs expériences. Ils comparent leur approche, via un certain nombre de variantes basées sur leurs analyses, à plusieurs modèles (Gouws et al., 2015, Mikolov et al., 2013b, Vulic et al., 2016) en suivant le protocole de Vulic et Moens (2015). En conclusion, leur modèle offre des performances améliorées de 1 à 2% sur les références, dépendant de certains paramètres et de la paire de langues.

Ammar et al. (2016) adaptent 2 approches de génération d'embeddings bilingues (Mikolov et al., 2013b) et (Faruqui et Dyer, 2014), afin que celles-ci puissent être entraînées sur plus d'une cinquantaine de langues en même temps et ainsi induire des embeddings interlingues (comme Coulmance et al. (2015), Gouws et al. (2015)). Ils comparent ensuite leurs modèles à 2 approches de référence, une extension multilingue de Luong et al. (2015) et une extension multilingue de LSA (Gardner et al.) sur la tâche d'ILB. Leurs embeddings de taille 512 sont entraînés sur un (plus petit) ensemble de 12 langues, pour lesquelles le corpus Europarl est disponible. Leurs données de tests pour les 12 langues (listes de paires $(mot;traduction)$) sont extraites de Wiktionary. Leurs résultats affichent des performances améliorées de 20% en faveur de l'approche adaptée de LSA. Cependant le taux de couverture (des paires de tests) ne sont pas identiques entre les approches, rendant leurs résultats difficiles à comparer. Enfin, à travers leurs expériences, ils proposent de réévaluer les tâches généralement utilisées pour évaluer la qualité des embeddings, en définissant un certain nombre de propriétés que les évaluations devraient intégrer. Pour aider la recherche dans ce domaine, ils proposent un por-

tail web où les utilisateurs peuvent déposer leurs embeddings interlingues afin d'être évalués.

Duong et al. (2016) proposent une adaptation du modèle CBOW[24] (Mikolov et al., 2013a) – modèle d'embeddings monolingues – afin que ce dernier induise directement des embeddings bilingues (Coulmance et al., 2015, Gouws et al., 2015)). Pour ce faire, ils modifient la fonction objectif du modèle CBOW en y intégrant une optimisation jointe ainsi qu'un terme de régularisation (qui a pour objectif de mieux combiner les représentations sources et cibles durant l'entraînement joint). Contrairement aux approches jointes qui s'appuient sur des corpus parallèles (Coulmance et al., 2015, Gouws et al., 2015), ils appuient le fait qu'utiliser 2 corpus monolingues et un dictionnaire bilingue est le choix le moins strict[25] mais aussi le plus sensé[26] (similairement à Vulić et Korhonen (2016)). La particularité de leur travail est l'adaptation bilingue de CBOW : ils utilisent le contexte d'une langue pour prédire la traduction du mot central dans l'autre langue. De plus, leur méthode choisit cette traduction basée sur le contexte grâce à un algorithme (inspiré d'EM[27]) qui traite explicitement la polysémie. Ils expérimentent leur modèle sur la tâche d'induction de lexiques bilingues en suivant le protocole de Vulic et Moens (2015) et se comparent à Gouws et al. (2015) et à Vulic et Moens (2015) (mais avec leurs données) . Les données monolingues proviennent de Wikipédia 2013 et leur dictionnaire bilingue amorce est le Panlex[28]. Ils observent des améliorations moyennes entre 10 et 20% sur les références, jusqu'à plus de 30% sur la paire de langues néerlandais-français.

Artetxe et al. (2016) succèdent aux travaux de Xing et al. (2015) en proposant un ensemble de contraintes sur la transformation linéaire présentée par Mikolov et al. (2013b). Ainsi, ils proposent que la matrice de projection soit orthogonale, que les vecteurs (des

[24]Contextual Bag Of Words.

[25]Les autres signaux bilingues sont plus rares.

[26]C'est commun pour les linguistes de construire un lexique bilingue quand ils apprennent un nouveau langage. C'est une des premières étapes afin de documenter la langue.

[27]Expectation-Maximization

[28]Dictionnaire multi-lingue couvrant plus de 1300 langues, 12 millions d'expressions. Couverture élevée mais des traductions très bruitées.

embeddings) soient normalisés et finalement de s'assurer que les embeddings de deux mots tirés au hasard aient bien leur similarité cosinus égale à 0. Une fois ces spécificités prises en compte dans leur fonction objectif, ils remarquent que leur approche est similaire à Faruqui et Dyer (2014). Cependant, ils jugent le fait que ces contraintes soient implicites au modèle de ces derniers comme étant conceptuellement confus et de fait, qu'elles ont un impact négatif. Ils testent leur approche en reproduisant le protocole d'évaluation d'ILB décrit dans (Mikolov et al., 2013b) mais sur des données différentes : ils entraînements leurs embeddings de taille 300 sur un corpus anglais composé de ukWaC + Wikipédia + BNC, et leurs embeddings italiens sur le corpus itWac. Le dictionnaire bilingue amorce, utilisé pour la projection ainsi que les tests, provient d'Europarl. Enfin, ils comparent leurs résultats avec Faruqui et Dyer (2014), Mikolov et al. (2013b), Xing et al. (2015) et obtiennent des améliorations situées entre 2 et 4% dépendant de la méthode comparée.

2.2.3 Via reclassement supervisé de candidats à la traduction

Le but de la tâche d'induction de lexiques bilingues est d'identifier des traductions de mots à partir de corpus comparables. En application, la tâche retourne une liste de mots candidats (dans la langue cible) pour chaque mot (de la langue source) dont la traduction est recherchée. Chaque mot candidat est trié dans la liste selon un score. Ce score est le résultat d'une fonction de similarité entre la représentation du mot source et celle du mot cible. Le score est donc dépendant de l'ensemble des facteurs de la tâche que sont les paramètres de la représentation des mots et le choix de la fonction de similarité. La performance de la tâche dépend donc de ces facteurs. Choisir une autre fonction de similarité ou une autre représentation des mots[29] ne changent pas le but de la tâche, mais peuvent influencer sa performance. De façon symétrique, changer la valeur des facteurs influence le score de la liste de candidats. Il est alors naturel de générer plusieurs listes de candidats (selon différentes valeurs pour les facteurs) afin d'en combiner les scores. Par la suite, une instance particulière de l'ensemble des facteurs possibles est appelée un **signal**[30].

Dans les sections précédentes (2.2.1 et 2.2.2), deux métamodèles de représentations de mots ont été présentés : les vecteurs de contexte (explicites et légers) ainsi que les embeddings (vecteurs continus et denses). Avec des points de vue différents sur la même tâche, ces modèles découlent d'architectures distinctes, et ont de fait leurs propres paramètres (nous les appelons : "facteurs architecture"). Cependant, ils utilisent la même entrée ("facteur source") : le contexte (éléments situés autour des mots à représenter). Ainsi, ces modèles partagent des paramètres communs ("facteurs modélisation") : taille de la fenêtre de contexte, fréquence des éléments, etc. Enfin, nous avons vu que la fonction de *similarité cosinus* est généralement utilisée pour comparer les représentations (sources et cibles) induites par chacun des modèles. L'ensemble des "facteurs" (architecture, source, modélisation, fonction de similarité) influence la performance de la tâche et un **signal** est une instance de cet ensemble de facteurs pour une expérience précise.

[29] Ce qui peut être fait en changeant la valeur d'un seul paramètre du modèle.

[30] Typologie reprise de (Irvine et Callison-Burch, 2013).

Habituellement, les chercheurs tentent d'identifier un signal qui maximise les performances de la tâche sur un jeu de tests donné (Sharoff et al., 2013). C'est le cas de la majorité des publications présentées dans les sections précédentes qui utilisent principalement l'information de contexte comme première et unique source de modélisation. D'autres sources, telles que l'information temporelle (Schafer et Yarowsky, 2002), thématique (Bouamor et al., 2013a, Hassan et Mihalcea, 2009, Vulić et al., 2011), ou encore orthographique (Gamallo et Garcia, 2012, Rubino et Linarès, 2011), sont aussi de bons signaux pour la tâche d'induction de lexiques bilingues. Par la suite, nous nous intéressons aux travaux qui utilisent les listes de candidats résultant de plusieurs signaux afin de reclasser les candidats et ainsi améliorer la performance de la tâche d'ILB. C'est ce que nous appelons des **approches de reclassement**. Nous distinguons d'abord deux types d'approches de reclassement : le reclassement supervisé et non supervisé.

Peuvent être appelées "non supervisées" les approches qui combinent plusieurs signaux de manière non apprise. Pionnier de l'utilisation de plusieurs signaux, Schafer et Yarowsky (2002) combinent les rangs résultant de quatre signaux (temporel, contexte, fréquence et sporadicité[31]). De cette façon (et avec l'aide d'une langue "pont", ils proposent d'éviter le dictionnaire bilingue amorce utilisé dans l'approche de base de l'ILB (Rapp, 1995). Rubino et Linarès (2011) combinent trois niveaux différents de représentation sur des termes médicaux par la paire de langues anglais-français. Ces trois niveaux sont les suivants : le contexte, le thème et l'orthographe (distance de Levenshtein entre les mots anglais et français). Chaque signal est utilisé comme un juge dans une combinaison de votes. Plusieurs configurations de votes sont expérimentées. Gamallo et Garcia (2012) extraient des traductions de mots qui ont des orthographes équivalentes, dans la paire de langues portugais-espagnol. L'originalité de leur combinaison est qu'ils emploient successivement les signaux pour filtrer leurs données et déterminer des candidats. D'abord des documents comparables de Wikipédia sont identifiés (grâce aux liens inter-langues). Une similarité de Dice est appliquée entre les lemmes anglais et français des documents. La distance d'édition est ensuite utilisée pour déterminer les

[31] État ou qualité de ce qui est rare et irrégulier.

cognats entre les lemmes sources et cibles. Deux signaux sont donc exploités, le signal thématique et celui orthographique (deux fois). Harastani et al. (2013) reclassent (par moyenne géométrique pondérée) les candidats obtenus par l'approche de Rapp (1995) en exploitant un score retourné par une méthode d'alignement phrasique. Cette méthode tente d'identifier les phrases comparables dans lesquelles les mots et leurs candidats à la traduction occurrent. Le signal "contexte" est donc utilisé deux fois. Cependant, d'autres signaux (surtout de positions de mots, donc contextuels) sont manipulés pour rechercher les meilleures phrases comparables. Chu et al. (2014) proposent une combinaison linéaire du signal contextuel (Rapp, 1995) et du signal thématique (Vulić et al., 2011) dans une procédure itérative où le nouveau lexique bilingue généré est réutilisé dans leur méthode d'induction de lexiques.

En apprentissage machine, l'apprentissage supervisé est un type d'apprentissage grâce auquel un système apprend un modèle à partir de données d'entraînements annotées. Autrement dit, des exemples (d'une tâche à réaliser) considérés comme vrais, et leur(s) caractéristique(s)[32], sont utilisés pour inférer les paramètres d'une fonction qui va tenter de modéliser la résolution de ladite tâche. Le modèle est ensuite appliqué sur de nouvelles données (afin de les résoudre). Aujourd'hui, nous savons qu'avec beaucoup de données, un bon niveau de généralisation est possible. Enfin, deux types de problèmes sont solubles par l'apprentissage supervisé, dépendant du type d'ensemble (de nombres) dans lequel les étiquettes prennent leur valeur ; si l'ensemble est continu, nous parlons d'un problème de régression. Si l'ensemble est fini, il s'agit d'un problème de classification.

La tâche d'induction de lexiques bilingues peut être vue comme un problème de classification : pour une paire de mots source-cible – un mot dans une langue source (ex : en anglais) et sa traduction dans une langue cible (ex : en français) – , il est possible d'imaginer un système binaire capable de répondre "Oui, cette paire constitue un mot et sa traduction" ou "Non, cette paire ne constitue pas un mot et sa traduction". Des données

[32]Features, en anglais.

d'entraînement annotées sont disponibles grâce aux lexiques bilingues amorces[33] mais sans caractéristiques. Ces dernières sont générées à l'aide d'un ou plusieurs signaux. Ainsi, l'apprentissage supervisé est identifié comme prometteur par Sharoff et al. (2013) et comme naturel par Irvine et Callison-Burch (2013) pour tenter de résoudre cette tâche.

Il est aussi possible de concevoir l'ILB comme une tâche de (re)classement[34] : les différents signaux nous renvoyant plusieurs listes scorées (classements) de candidats pour un mot source, il est naturel de modéliser un système profitant de ces listes et de les combiner afin de proposer un meilleur classement de ces candidats. Cette analogie est notamment mise en avant par Delpech et al. (2012) qui proposent d'examiner la contribution des algorithmes de Learning to Rank (LTR) aux reclassements de candidats à la traduction. Les algorithmes LTR sont utilisés en Recherche d'Information (RI) afin de reclasser les documents retournés pour répondre à une requête (Li, 2011). En effet, l'ordre des documents est important car on s'attend à ce que les documents les plus pertinents soient placés en premières positions. Il en est de même pour les bons candidats à la traduction. Il existe trois catégories d'approches LTR : pointwise (prédit le score d'une paire Requête-Document ; dépendant du type de score (continu ou fini) un algorithme de régression ou de classification est utilisé), pairwise (choisit quel est le meilleur document dans une paire de documents, selon une requête donnée) et listwise (essaie d'ordonner correctement tous les documents de la liste donnée, selon une requête ; cette dernière catégorie est celle qui s'apparente au problème de Learning to Rank). Remarquons que même avec cette conception différente de l'induction de lexiques bilingues, les approches *pointwise* et *pairwise* nous renvoient à un problème de classification, et donc à de l'apprentissage supervisé.

Au meilleur de nos connaissances, trois contributions ont utilisé de l'apprentissage supervisé dans le cadre de la tâche d'induction de lexiques bilingues. Prochasson et Fung (2011) exploitent le signal contextuel de deux façons, avec lesquelles ils entraînent un

[33] Ressource que l'approche standard (Section 2.2.1) a toujours utilisée.
[34] Ranking, en anglais.

classificateur[35]. L'une d'entre elles reproduit l'approche standard (Rapp, 1995), l'autre propose une métrique[36] qui vise à évaluer la relation de traduction entre deux mots (source et cible) par le nombre de documents comparables alignés dans lesquels ces derniers occurrent. La métrique se veut simple, notamment grâce au fait qu'ils ne s'intéressent qu'aux mots peu fréquents. Quatre langues sont utilisées : l'anglais, le français, l'espagnol et le chinois. L'utilisation de deux signaux montre une précision améliorée jusqu'à 98%[37] sur les paires de langues anglais-espagnol et français-espagnol, sur un jeu de termes médicaux. Irvine et Callison-Burch (2013) utilisent jusqu'à 5 sources d'informations (contexte, temporel, orthographe, thématique et fréquence), dont certaines sont issues de plusieurs corpus. Un total de 13 signaux est ainsi exploitable pour entraîner un classificateur[38]. Les expériences sont réalisées sur 22 langues, de peu dotées (le somalien) à très dotées (l'espagnol). Des gains relatifs jusqu'à 44% sont observés, confirmant l'intérêt de l'apprentissage supervisé sur la tâche d'ILB. Kontonatsios et al. (2014) emploient 2 signaux (contexte et compositionnel) pour entraîner un classificateur[39]. Le signal compositionnel provient de leurs anciens travaux et le signal contextuel suit l'approche standard (Rapp, 1995). Cinq langues sont utilisées, avec l'anglais comme langue source vers les quatre langues cibles suivantes : l'espagnol, le français, le grecque et le japonais. Ils analysent l'impact de la fréquence de leurs termes médicaux (fréquent et rare) autant sur les signaux de manière individuelle que sur leur combinaison. La combinaison montre de bonnes améliorations sur la précision TOP@20 (jusqu'à 20%), et des améliorations mineures en TOP@1.

Enfin, citons également l'existence d'une série de travaux qui utilisent aussi du reclassement de candidats. Il s'agit des travaux sur la traduction compositionnelle, c'est-à-dire qui exploitent des équivalences de traductions, soit au niveau des mots (Baldwin et Tanaka, 2004, Robitaille et al., 2006) soit au niveau des morphèmes (Cartoni, 2009, Delpech

[35] Arbre de décision J48 de l'environnement Weka.

[36] Basée sur la similarité de Jaccard.

[37] Ces performances impressionnantes nous ont amenés à reproduire l'approche dans notre deuxième publication (chapitre 5).

[38] Vowpal Wabbit (Fast Learning).

[39] Linear-SVM du package LIBSVM (avec les valeurs par défaut pour tous les paramètres).

et al., 2012, Garera et Yarowsky, 2008) pour générer des traductions candidates (mots ou multi-mots). La spécificité de ces travaux est due à leur utilisation du principe de compositionnalité[40]. Ainsi, indépendamment du niveau d'équivalence traductionnelle utilisé, les candidats générés peuvent ne pas être des expressions de la langue cible. Une phase de sélection est donc appliquée pour filtrer les vrais des faux candidats. C'est durant cette dernière phase que certains utilisent du reclassement, soit non supervisé (Cartoni, 2009, Garera et Yarowsky, 2008, Robitaille et al., 2006), soit supervisé (Baldwin et Tanaka, 2004, Delpech et al., 2012). Nous retenons (Delpech et al., 2012) qui, en plus de proposer une vue d'ensemble complète des travaux mentionnés ci-dessus, utilisent jusqu'à 4 signaux dans des algorithmes de Learning to rank.

[40] "The meaning of the whole is a function of the meaning of the parts" (Keenan et Faltz, 1985)

CHAPITRE 3

BAD LUC@WMT16 : A BILINGUAL DOCUMENT ALIGNMENT PLATFORM BASED ON LUCENE

Laurent Jakubina et Phillippe Langlais. Bad luc@wmt 2016: a bilingual document alignment platform based on lucene. Dans *Proceedings of the First Conference on Machine Translation*, pages 703–709, Berlin, Germany, August 2016a. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W/W16/W16-2370`

Cet article a été publié à la 1^{ere} Conférence sur la Traduction Automatique[1], dans le cadre de la tâche partagée intitulée Alignements de Documents Bilingues (ADB). Il a aussi été présenté sous forme de *poster* à ladite conférence qui avait lieu en même temps que la conférence ACL (Association of Computational Linguistics) en août 2016 à Berlin.

3.1 Contexte

Ce travail a été réalisé en participant à la tâche partagée d'alignement de documents bilingues de la première conférence sur la traduction automatique et l'article a été la troisième publication de mon doctorat, dans l'ordre chronologique. Les autres publications de cette thèse portant majoritairement sur l'induction de lexiques bilingues, nous signalons que le contexte ne poursuit pas la recherche des articles 1 (chapitre 5) et 2 (chapitre 6) dans l'ordre chronologique. Ainsi, nous avons montré que les deux tâches sont connexes vis-à-vis des problèmes de la traduction automatique dans le chapitre 1 et c'est pour ces raisons que nous nous sommes intéressés à cette tâche. De plus, l'ADB pourrait aussi être vue comme une étape de prétraitement à la tâche d'induction de lexiques bilingues et c'est ce dernier point qui explique que nous commençons par

[1] `http://www.statmt.org/wmt16/` - Est également la 11^{eme} édition de ce qui était précédemment un workshop en traduction automatique.

décrire ce travail.

Les promoteurs de la tâche ont rappelé l'importance des corpus parallèles pour la traduction automatique avant de faire remarquer que les collections de textes issues de la recherche académique ont souvent été *ad hoc* et limitées en taille (Buck et Koehn, 2016). Notre état de l'art de la section 2.1 confirme ce constat. Ils ont donc souhaité recentrer l'attention sur ce problème, et plus particulièrement sur une des étapes clés de prétraitement de l'acquisition de données parallèles : l'identification de documents en relation de traduction. La raison énoncée de ce choix étant que c'est cette dernière qui a reçu le moins d'attention par le passé. Cependant, notre revue de littérature montre qu'un certain nombre de méthodes existent, proposant des solutions selon les différents types de documents. Enfin, il semblait évidemment de concentrer les efforts sur le web, la ressource multilingue grandissant de jour en jour. Ainsi, la tâche a été précisément d'identifier des paires de documents parallèles en anglais-français sur une série de sites web bilingues. De plus amples détails sur la tâche et les données reliées sont fournies dans (Buck et Koehn, 2016).

Notre objectif de participation était de proposer un système d'ADB simple et léger, basé sur un moteur de recherche d'information (inter-langue) pour profiter de la puissance de l'indexation des documents, notamment du contenu textuel, HTML et des méta-informations des pages web. Nous avons choisi Lucene[2], car nous avions déjà de l'expérience avec cet outil et qu'il était possible de réaliser un prototype rapidement.

3.2 Contributions

L'idée de participer à la tâche a été émise par mon directeur de recherche, Philippe Langlais. L'idée d'utiliser Lucene est venue de moi, plus précisément l'utilisation du module de requêtes *MoreLikeThis*. Par la suite, j'ai donc implémenté toutes les idées et variantes à expérimenter en m'y rapportant. L'idée de modéliser les URL des documents à l'aide du modèle de sac de mots est venue de moi également. L'ensemble des

[2] www.lucene.org

variantes explorées est le résultat de l'exploration en largeur des paramètres ainsi que de la comparaison avec certains systèmes de l'état de l'art connus de Philippe Langlais. Les algorithmes de post-traitement utilisés ont été proposés par mon directeur de recherche et implémentés par moi-même. La publication, sous forme d'un papier court, a été rédigée par moi-même dans un premier temps, et ensuite corrigée et réécrite à certains endroits par Philippe Langlais.

3.3 Impacts

Visant la simplicité, nous avons imaginé une approche basée sur la représentation "sacs de mots" pour modéliser les documents et nous souhaitions aussi profiter des deux champs d'informations donnés ; le texte et les URL. Dans Lucene, il existe déjà un module de requêtes, nommé *MoreLikeThis*, permettant d'identifier des documents similaires. Ce module représente les documents par des requêtes "sacs de mots" et (surtout) permet de régler des paramètres tels que la fréquence minimale et maximale des mots à accepter dans la requête, les mots outils à rejeter, la taille de la requête, etc. Pour faire le pont entre les deux langues, en mettant l'accent sur la légèreté, nous avons seulement utilisé un lexique bilingue et paramétré son utilisation. En combinant l'ensemble de ces modules, nous avons construit un cadre de recherche d'information inter-langue permettant d'explorer en profondeur l'ensemble des métaparamètres de tous les modules de l'approche proposée. Grâce à ce cadre, nous avons pu expérimenter plus de mille configurations et parmi les variantes, nous avons en découvert une, basée sur les URL – plus précisément, la découpe de celles-ci en sacs de mots et caractères –, affichant des performances étonnamment bonnes et nous permettant d'atteindre notre objectif, c'est à dire, de soumettre un système simple et léger. Le système final combine les sources texte et URL conjointement, ainsi que d'autres heuristiques (longueur des documents), pour obtenir les meilleurs résultats.

À l'annonce des résultats, notre système s'est placé 9^{eme} sur 12 soumissions avec un rappel à 85%. Une description résumée de chaque système et le classement détaillé est

présenté dans Buck et Koehn (2016). On y apprend que la majorité des systèmes se situent dans un intervalle de résultats entre 80% et 95% de rappel et que le classement est variable selon que les promoteurs de la tâche adaptent leur métrique d'évaluation à certaines erreurs dans les données de test. L'ensemble des systèmes offrent donc des résultats proches les uns des autres mais il est intéressant de savoir quelle est l'approche proposée par le meilleur système.

Deux systèmes arrivent en première position selon la métrique d'évaluation. Il s'agit de (Gomes et Pereira Lopes, 2016) et de (Dara et Lin, 2016). Ces deux systèmes utilisent un moteur de traduction automatique (MTA) complet afin de réaliser la tâche[3]. Gomes et Pereira Lopes (2016) calculent un score de couverture entre les documents sources et cibles en utilisant une table de phrases. Une variante de leur système (la meilleure) emploie aussi les URL pour améliorer l'identification des paires. Dara et Lin (2016) traduisent les documents français en anglais avant d'évaluer les paires en fonction du nombre commun de bigrammes et de 5-grammes. Ils utilisent aussi une heuristique basée sur la longueur des documents.

Concernant les autres systèmes, la majorité des participants ont utilisé une représentation des documents à base de mots (*TF/IDF weighted vector*, *bag of words representation*, *word vector representation*) avec un pont de traduction obtenu par traduction automatique (résultats : 96%, 92%, 91%, 87%, 82%) ou par lexique bilingue (résultats : 93%, 88% et 82%). Les URL ont aussi été utilisées par d'autres participants (résultats : 91%, 89%, 88%, 88%, 82%). Un système a attiré notre attention, basé sur les problèmes de la tâche mentionnés à l'état de l'art (Section 2.1), qui propose des performances autour de 85% en n'utilisant que des éléments non dépendants du langage tels que les liens vers les documents d'un même site web, les URL, les nombres, les noms des fichiers-images ainsi que de la structure HTML.

Des critiques[4] peuvent être émises sur notre travail. Basé sur la simplicité, notre processus de traduction n'a employé qu'un simple dictionnaire bilingue et n'a pas fait usage

[3]Choix qui va à l'encontre de notre objectif de simplicité et légèreté en ce qui nous concerne.
[4]Des améliorations par symétrie.

de données de traductions (alors que certaines étaient distribuées par les organisateurs). Aussi, les données finales ont été significativement plus importantes en taille que les données d'entraînements. Et c'est à la fin de la période de soumissions que nous avons réalisé qu'un prétraitement sur les données aurait pu améliorer le temps d'exécution ainsi que le résultat final[5]. Finalement, une seule version du système (supposément la meilleure selon les données initiales) a été soumise. Cependant, plusieurs versions auraient pu être envoyées afin d'évaluer certaines fonctionnalités séparées du système complet sur les données finales (et potentiellement constater une amélioration du score) ainsi que pour la curiosité scientifique.

[5]Cette remarque est aussi mentionnée par d'autres participants de la tâche.

3.4 Publication

[6]We participated in the Bilingual Document Alignment shared task of WMT 2016 with the intent of testing plain cross-lingual information retrieval platform built on top of the Apache `Lucene` framework. We devised a number of interesting variants, including one that only considers the URLs of the pages, and that offers — without any heuristic — surprisingly high performances. We finally submitted the output of a system that combines two informations (`text` and `url`) from documents and a post-treatment for an accuracy that reaches 92% on the development dataset distributed for the shared task.

3.4.1 Introduction

While many recent efforts within the machine translation community are geared toward exploiting bilingual comparable corpora — see Munteanu et Marcu (2005) for a pioneering work and Sharoff et al. (2013) for an extensive review — there is comparatively much less work devoted to identifying parallel documents in a (potentially huge) collection. See Smith et al. (2013), Uszkoreit et al. (2010) for two notable exceptions. This is due in large part to conventional wisdom that holds that comparable corpora can be found more easily and in larger quantity than parallel data. Still, we believe that parallel data should not be neglected and should even be preferred when available.

The Bilingual Document Alignment shared task of WMT 2016 is designed for precisely identifying parallel data in a (huge) collection of bilingual documents mined over the Web. The collection has been processed by the organizers in such a way that this is easy to test systems : the language of the documents is already detected, and we have access to the content of the Web pages. Although the organizers encouraged participants to test their own way of pre-processing data, we decided (for the sake of simplicity) to use the data as prepared.

[6]Certaines adaptations stylistiques ont été réalisées sur l'article afin de l'adapter au format de la thèse.

We describe the overall architecture of the BadLuc framework as well as its components in Section 3.4.2. We explain in Section 6.4.4 the experiments we conducted and provide some analysis in Section 7.4.5. We conclude in Section 6.4.6.

3.4.2 BadLuc

We built variants of a Cross-Language Information Retrieval (CLIR) platform making use of the popular Apache framework Lucene.[7] We describe here the different components embedded in this platform.

We participated in this task by relying entirely on the pre-processing carried out by the organizers, that is, we used the text of the pages as it was extracted. Figure 3.1 shows an excerpt of the text extracted from a given URL. Sometimes, the conversion to text is noisy and deserves further work. While we could have used the machine translation provided as well, we decided to resort to a bilingual dictionary, mainly for the sake of simplicity : the resulting system is very light and can be deployed without retraining any component.

After some exploration with the platform, we settled for a configuration — named RALI — that we used for treating the official dataset of the shared task. RALI is a combination of variants that delivers good performance both in terms of processing time and accuracy. This system achieves 92.1% TOP@1 on the development dataset, a performance we consider satisfactory considering the simplicity of the approach.

3.4.2.1 Indexes

We built two main indexes. One from the source and one from the target documents of the collection provided. This last was organized into webdomains (49 in the development set) but, to ease implementation, we built the indexes from all, and enforced a posteriori that only target documents of a given web-domain are returned. In each

[7] https://lucene.apache.org/core/

index, documents are indexed (and tokenized) into three Lucene fields, one based on the text itself (text), one based on its url and one with the size of the text content (in number of tokens).

Lucene provides a number of tokenizers, but we felt the need to develop our own in order to properly handle the cases where punctuations is glued to words, and other typical cases one finds in real data. One point worth mentioning is that our tokenizer splits urls into several tokens.[8] This way of handling urls leads us to a simple but efficient url-based baseline. See Figure 3.1 for an illustration of a few bag-of-word queries considered in BadLuc.

3.4.2.2 Query Instantiation

Each field of each (source) document can be treated as a bag-of-word query. We used the *MoreLikeThis* query generator available in Lucene[9] to implement this. The generator uses a variant of *tf.idf* and allows for the adjustment of a number of meta-parameters mainly for finding an application-specific compromise between the retrieval speed and its accuracy. We investigated the following ones[10] :

- minimum frequency of a term in a document (tf) for it to be considered in a query,
- minimum ($mindf$) and maximum ($maxdf$) number of documents in the collection that should contain a candidate query term,
- minimum (*minwl*) and maximum (*maxwl*) word length of a term in a query,
- maximum number of terms in a query (*size*),
- only words absent from a specified stop-list are legitimate query terms (*stop*).

[8] We split urls according to a list of 33 separators, among which : *@, ?,/,<,>,(,),+, !,%,∼*

[9] https://lucene.apache.org/core/4_4_0/queries/index.html

[10] The *MoreLikeThis* mechanism also allows to settle a boost factor per query term, but we did not play with it.

These meta-parameters allow to easily create specific-purpose queries on the fly. For instance, by setting $mindf$ and $maxdf$ to 1, we built a collection-wide hapax query, while setting *minwl* and *maxwl* to 1 allows to build queries containing only punctuations marks.

3.4.2.3 Mono and Bilingual Queries

We tested two main families of queries : monolingual (`mono`) and bilingual (`bili`). The former is a way of easily capturing the tendency of a document and its translation to share a number of specific entities such as named-entities, numbers, or `urls`, for which no translation is required. Obviously, we were not expecting a high accuracy with monolingual queries, but we thought it would provide us with a very simple baseline. Actually the performance of such an engine on a given collection might be a valid metric to report, as a measure of the *difficulty* of the collection.

Bilingual queries involve a translation procedure. We simply translate the terms of the query based on a bilingual lexicon. We could have used the machine translated text provided by the organizers, but we decided early on in our experiments to resort to a simple bilingual lexicon approach, to simplify deployment. As a matter of fact, in a previous work on identifying parallel material in Wikipedia Rebout et Langlais (2014), we observed the inadequacy of the features computed from a generic SMT engine. Arguably our lexicon might not be very good either to deal with the nature of data collected over the Web, but we felt that a *general* bilingual lexicon might be more robust in this situation.

There are two meta-parameters that control our translation procedure :

- *keep* when set to true (which we note κ), will leave untranslated terms (that is, terms unknown from our lexicon) in the query.[11] Hopefully this will leave in named- and numerical-entities that are useful for distinguishing parallel documents Patry et Langlais (2011b).

[11] At least terms that meet the *MoreLikeThis* meta-parameters.

http://creationwiki.org/Earth Earth - CreationWiki, the encyclopedia of creation science [...] 23.439281 0.409rad 26.044grad Physical characteristics Mass 5.9736 * 1024 kg [...] taking 23 hours, 56 minutes, and 4.091 seconds to line up relative to the stars (Sidereal day), and 24 hours plus or minus 20 seconds to line up relative to the sun [...] is closer to the sun at some times of the year than others ; the Earth moves faster [...] Kepler's laws of planetary motion [...] Saturnine - Uranian - Neptunian [...]	
text mono	*texttok :* hours neptunian 1024 tennessee 2008 closer 397 ...
text bili	*texttok :* hours neptunian 1024 penchant théorie visible intensité fois tennessee prononcée prénommé 2008 métrique note closer équateur ...
url mono	*urltok :* earth creationwiki / org http . : ...
url bili	*urltok :* déblai masse tanière terre earth creationwiki / org http . : ...
both bili	*texttok :* hours neptunian 1024 penchant théorie tennessee absorbant prononcée prénommé 2008 métrique 397 équateur ... *urltok :* déblai masse tanière terre earth creationwiki / org http . : ...

Figure 3.1 : Excerpt of bag-of-word queries for `http://creationwiki.org/Earth`.

- *nbTrans* controls the number of translations to keep when there is more than one available for a given source term. We consider two possible values : `all`puts all available translations in the query, while first picks the first one listed in the bilingual lexicon.[12]

3.4.2.4 Queries on Both Fields

`Lucene` allows to combine queries made on different fields. We use this functionality in order to produce queries with terms to be searched in both fields (`text` and `url`) in a single pass. An explicit example of this query is illustrated at the bottom of Figure 3.1.

[12]There is no specific order in the multiple translations listed in our lexicon for a given term, but some lexicons might list more general translations first.

3.4.2.5 Length-based Filter

`Lucene` allows to write queries as filters. It is basically a query that is executed before the main one and that returns an initial list of target documents on which the main query is applied. We implemented one such filter (`size`), using the third indexed field, based on the observation that pairs of parallel documents should have similar lengths (counted in tokens). We assumed the size ratio of source/target documents follows a normal distribution whose variance defines a confidence interval in which the target document size should fall. Unfortunately we estimated the parameters of the normal distribution on all reference pairs of documents provided by the shared task. This could explain our unsatisfactory scores on the official test set of the shared task[13].

3.4.2.6 Post Processors

Query execution produces for each source document a ranked list of target documents. Since each query is carried out independently over the collection, we run the risk of having a given target document associated with more than one source document. As a solution, we tested a few post-processors that select exactly one candidate per source document :

`hungarian` the Hungarian Algorithm Kuhn (1955) is a well-known combinatorial optimization algorithm[14] that solves the assignment problem in polynomial time.

`b-greedy` a *batch* greedy solution which picks the best ranked candidate among all the source documents paired, removes the selected pairs and loops until all source documents get paired with exactly one document.

`o-greedy` an *online* version of the greedy procedure just described, where we select for each source document the top ranked candidate that has not been paired

[13] We noticed this bias after the submission period.

[14] Implementation available here : `https://github.com/KevinStern/software-and-algorithms`

with a previous source document yet. Once selected, the target document is removed from the potential list of candidates for subsequent source documents.

On a task of identifying parallel documents in Wikipedia, Rebout et Langlais (2014) shows that both the `hungarian` and the `b-greedy` algorithms deliver good performance overall.

3.4.3 Experiments

3.4.3.1 Protocol

We conducted these experiments on the `lett.train` webcrawl available on the WMT2016 shared task webpage.[15] This crawl consists of 49 webdomains of various sizes, and the language of each document has been identified.

The test set made available for participants to calibrate their systems contains 1624 English `url`s for which the target (French) parallel counterpart is known. It is noteworthy that the task does not evaluate the ability of a system to detect whether a given source `url` has its parallel counterpart in the collection, which would require to train a classifier. [16] Because of this, we always propose a target `url` for a source one ; and we measure performance with accuracy at rank 1, 5 and 100. Accuracy at rank i (TOP@i) is computed as the percentage of source `url`s for which the reference `url` is identified in the top-i candidates.

On top of our tokenizer which is clearly biased toward space-oriented language scripts, we use two language specific resources : a stop-word list for English which comprises 572 entries,[17] as well as an in-house English-French bilingual lexicon of 107 799 entries. Very roughly, our lexicon could help the translation of only half of the query terms, which is an issue we should look at in the future.

[15] `http://www.statmt.org/wmt16/bilingual-task.html`

[16] We have conducted the training of such a classifier in past experiments Rebout et Langlais (2014), with results we evaluated to be around 85%.

[17] We downloaded it from : `http://www.perseus.tufts.edu/hopper/stopwords`

	Strategies		Top (%)		
	Query	[MLT] + [Trans]	@1	@5	@100
		`text` variants			
default	mono	[2,5,∞,0,25,F]	6.4	15.8	49.5
default+tok		[2,5,∞,0,25,F]	35.4	57.0	83.9
		[1,1,∞,1,200,F]	48.3	78.2	94.7
		[1,1,∞,1,∞,T]	57.2	86.2	96.2
		[1,1,∞,1,200,T]	64.9	87.2	96.8
	+size	[1,1,∞,1,200,T]	76.2	92.1	97.3
	+size	[1,1,∞,1,∞,T]	76.6	92.6	97.2
stop words	+size	[1,1,∞,1,∞,F]	69.2	89.7	96.4
wl = 3	+size	[1,1,∞,3,∞,T]	75.1	92.0	97.1
hapax	+size	[1,1,1,1,∞,T]	49.5	49.8	49.8
	bili	[1,1,∞,1,∞,T] + [κ,first]	74.4	93.5	98.7
		[1,1,∞,1,∞,F] + [κ,first]	71.9	92.8	**98.8**
		[1,1,∞,1,∞,F] + [κ,all]	34.5	53.2	88.4
		[1,1,∞,1,∞,T] + [κ,all]	44.1	64.5	95.0
	+size	[1,1,∞,1,∞,F] + [κ,all]	81.2	**97.1**	98.3
	+size	[1,1,∞,1,∞,T] + [¬κ,all]	81.0	94.8	98.2
best-`text`	+size	[1,1,∞,1,∞,T] + [κ,first]	**83.3**	96.2	98.2
		`url` variants			
wmt 2016			67.9		
	mono	[1,1,∞,1,∞,F]	75.4	84.4	92.9
	+size	[1,1,∞,1,∞,F]	78.4	87.5	95.3
	bili	[1,1,∞,1,∞,F] + [κ,all]	77.0	86.6	93.5
	+size	[1,1,∞,1,∞,F] + [κ,first]	78.8	88.0	91.3
best-`url`	+size	[1,1,∞,1,∞,F] + [κ,all]	**80.1**	**88.6**	**95.6**
rali	bili-size	best-`text`+best-`url`	**88.6**	**97.6**	98.3

Tableau 3.I: Performances of some selected variants we tested. The MLT meta-parameters are [tf, $mindf$, $maxdf$, $minwl$, $maxwl$, $size$, $stop$], while those specific to the translation process are [$keep$, $nbTrans$]. See Section 3.4.2 for more.

3.4.3.2 Results

We tested over a thousand configurations, varying the meta-parameters of the *More-LikeThis* (MLT) query generator, as well as the components described in the previous section. Table 3.I shows a selection of some of the variants we tested. A line in this table indicates the best MLT meta-parameters we found for the configuration specified.

First of all, and without much surprise, we are able to outperform the `url`baseline (line WMT 2016) proposed by the organizers and which relies on some rules for matching `url`s in both language. Our best variant (line best-`url`) relying only on `url`s significantly outperforms this baseline by 12 absolute points in TOP@1. This variant tokenizes the `url`, then translates its words.[18]

Focusing on variants that exploit the text of the documents, we achieve a decent result without involving translation at all : the best monolingual variant we tested performs at 76.6 TOP@1, which also outperforms the WMT baseline. It should be noted that the default `Lucene` configuration (line `default`) does not perform well at all. Clearly, some tuning is necessary. In particular, using our tokenizer instead of the default one (which separates words at spaces) drastically increases performance (line `default+tok`). See Figure 3.1 for the kind of noisy input a tokenizer needs to handle. Unquestionably, using translation increases performance. The best variant we tested (line best-`text`) picks only one translation per source word, and leaves in the query the terms without translation.

Another interesting fact is the positive impact of the length-based filter presented in Section 3.4.2.5. Not only does this filter improve performances (a gain of 2 to 40 absolute points in TOP@1 is observed depending on the configuration tested), it also gives an appreciable speed up (2 to 10, depending on the variants).

Incidentally, we reproduced a proxy to systems that would only consider hapax words, somehow similarly to Enright et Kondrak (2007), Patry et Langlais (2011b). The best variant we obtained lagged far behind other variants exploiting all the available text.

[18]Keeping all translations is in this case preferable to keeping only one translation.

One reason for this bad result might simply be that only collection-wide hapax terms are considered here.

The impact of the post-processor can be observed in Table 3.II. With the exception of the `url` variants, applying a post-processor improves TOP@1, a finding that corroborates the observations made in Rebout et Langlais (2014). We do not observe a huge performance difference between the algorithms. For the final submission, we applied the `o-greedy` algorithm because the others could not handle the size of the data set[19].

	`url`	`text`	`both`
w/o	80.1	83.3	88.6
`o-greedy`	79.7	87.6	91.6
`b-greedy`	80.7	87.9	92.1
`hungarian`	80.4	87.9	**92.1**

Tableau 3.II: TOP@1 of the post-processors we tested.

3.4.4 Analysis

3.4.4.1 Sensitivity to Source Document Size

We explored how our variants behave as a function of (source) document length. Figure 3.2 reports the cumulative accuracy of selected variants as a function of document size. We observe (red curve) the tendency for the RALI variant (the one we submitted) to globally improve as source documents get larger. Comparing the two dotted green curves, we also see that the benefit of embedding translation increases with document size. It is not entirely clear why we observe an increase in performance of the `url` variants as document size increases, since only the `url`s are considered. There are not many documents with a very short size, therefore the very first point of each curve is likely not significant.

[19] Without deep modifications of the algorithms.

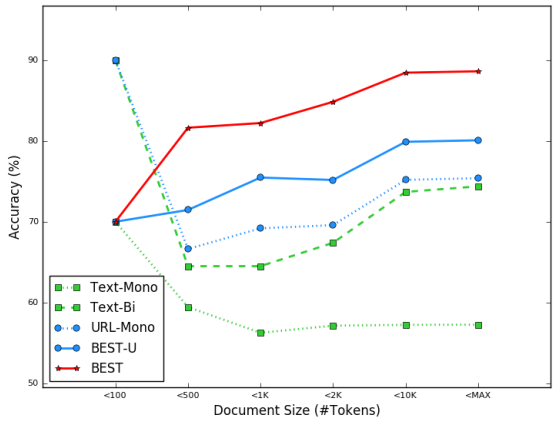

Figure 3.2 : Accuracy (TOP@1) as a function of document size (counted in tokens).

	Almost no text inside
src	`http://rehazenter.lu/en/medical/explorations_fonctionnelles/explorations_posture/laboratoire_de_biomecanique`
trg	`http://rehazenter.lu/fr/medical/explorations_fonctionnelles/explorations_posture/laboratoire_de_biomecanique`
src	`http://www.dakar.com/2009/DAK/RIDERS/us/equipage/57.html`
trg	`http://www.dakar.com/2009/DAK/RIDERS/fr/equipage/57.html`
	Reference problem
src	`http://www.nauticnews.com/en/2009/06/23/burger-boat-company-launches-151-03-fantail-motor-yacht-sycara-iv`
trg	`http://www.nauticnews.com/2009/07/13/ishares-cup-2009-a-bord-dholmatro`

Tableau 3.III: Examples of problematic pairs of `urls` found in the development set.

3.4.4.2 Error Analysis

We conducted a small-scale analysis of the errors made by the RALI configuration. First of all, we observed frequent cases where a French page contains a fair amount of English material (which might explain part of the performance of monolingual variants). We also noticed that a given document has often several associated `urls`. In such a situation, our system will almost invariably pick the largest `url` (more tokens do match), which is not necessarily the case of the reference.

In the 1.7% cases of RALI could not identify the expected target document in the top-100 positions, we observed that many documents contained almost no text. Typical

examples are provided in Table 3.III. In such cases, the `url`-based approach should be more efficient. This means that learning which variant to trust given a source document could be fruitful. We also observed inevitable reference errors (see the bottom line of Table 3.III for an example). Last, we noticed that some documents are rather specific, and our lexicon does not help much the translation process. This is the case for the document shown in Figure 3.1.

3.4.5 Conclusion

Our participation in the shared task has been carried out thanks to the `Lucene` framework. We devised a number of configurations by varying the parameters of the *MoreLikeThis* query mechanism, as well as by exploiting other built-in features. We notably found a simple yet efficient way of matching documents thanks to their `url`s, which outperforms the baseline provided by the organizers. We also observe that querying the target collection with queries built without translation already achieves a decent performance and that involving a translation mechanism as simple as using a bilingual lexicon gives a nice boost in performance. We also propose to filter target documents based on the length of the source document. This not only improves results, but also speeds up retrieval. Last, we measured that applying a post-processor (such as the Hungarian algorithm) further improves performance.

The best system we identified on the development set combines (in a single query) terms translated from the source document as well as terms from its `url`. A length-based filter is applied, as well as a post-processor (Hungarian algorithm). This system achieves a TOP@1of 92.1, and a TOP@100 of 98.6, a respectable performance for such a simple system.

We are currently investigating whether better performance can be obtained by using machine translation instead of the lexicon-based translation approach used here.

CHAPITRE 4

ILB EN CORPUS PARALLÈLES : ÉVALUATION DE LA COUVERTURE DE NOS RESSOURCES PARALLÈLES

Comme mentionné dans l'introduction (chapitre 1), l'induction de lexiques bilingues est applicable sur les corpus parallèles. Dans la littérature des corpus parallèles, la tâche est aussi appelée *transpotting* (Bourdaillet et al., 2010). Les corpus parallèles sont composés de paires de documents en relation de traduction. Par définition, ils garantissent donc que si un mot se trouve dans la partie source du corpus, une traduction est présente dans la partie cible. C'est une caractéristique qui n'est pas assurée dans les corpus comparables. De plus, avec la nature parallèle[1] du corpus, la probabilité d'identifier avec succès des traductions de mots est beaucoup plus élevée. Il est donc naturel de commencer par appliquer l'ILB sur des corpus parallèles. Il faut cependant vérifier que les corpus contiennent les mots cibles intéressants. En effet, à l'inverse des corpus comparables, les corpus parallèles sont moins nombreux et de fait, il est moins probable d'y trouver des termes souhaités ainsi que leurs traductions.

Dans ce chapitre, nous évaluons la couverture du vocabulaire de quatre corpus parallèles afin d'estimer leur utilité à identifier des traductions de mots grâce à l'induction de lexiques bilingues (sur ce type de corpus, dans un deuxième temps). Plus précisément, nous étudions la distribution des fréquences des termes sources et cibles d'une liste de référence à travers les corpus parallèles sous la contrainte d'alignement.

4.1 Approche

Pour évaluer la couverture du vocabulaire, nous avons simplement compté le nombre de fois où un terme de notre liste de référence est observé du coté anglais du corpus parallèle ainsi que le nombre de fois où sa traduction est observée du coté français,

[1] Un alignement basé sur la position est disponible entre les mots sources et cibles.

avec la contrainte que la traduction doit se trouver dans des phrases françaises alignées aux phrases anglaises contenant le terme. Nous appelons cette contrainte, la *contrainte d'alignement.*

Un exemple de cette contrainte d'alignement est illustré dans la figure 4.1 avec le terme *greeting* et sa traduction *salutation.*

GIGAWORD^{en}	GIGAWORD^{fr}
21677278 : *French and English* greeting *translation cards* [...]	21677278 : *Des fiches de* salutation *en anglais et en français* [...]

Figure 4.1 : Exemple de phrases parallèles dans le corpus GIGAWORD respectant la contrainte d'alignement pour le terme *greeting* et sa traduction *salutation.*

4.2 Matériel d'évaluation

Dans le cadre de nos expériences, nous avons sélectionné quatre corpus parallèles. Les caractéristiques en nombres de mots et de phrases de chacun des corpus sont détaillés dans la table 4.I.

HANSARDS est le bitexte utilisé par l'application `www.tsrali.com`. Il contient tous les débats du Parlement Canadien publié entre 1986 et 2007, ce qui représente 8 millions de paires de phrases.

EUROPARL contient l'ensemble des débats du parlement Européen. Il est très utilisé comme corpus d'entraînement par les moteurs de traduction statistique et la paire de textes anglais-français contient plus de 2 millions de paires de phrases.[2]

HEALTH résulte du forage des sites web de *Santé Canada*[3] et de l'*Agence de la Santé Publique Canada*[4], deux sites web particulièrement bien organisés du point de vue bilingue. Les détails de ce corpus peuvent être trouvés dans (Langlais et al., 2006).

[2] `http://www.statmt.org/europarl`
[3] `www.hc-sc.gc.ca`
[4] `http://www.phac-aspc.gc.ca`

GigaWord Comme son nom l'indique, le corpus[5] réunit plus de 10^9 mots provenant de textes institutionnels Canadien et Européen, téléchargés sur le Web (Callison-Burch et al., 2010). Barrière et Isabelle (2011) ont montré que la couverture de ce corpus est propice au forage de termes de domaines spécifiques et de leurs traductions.

Corpus	$\#Phrases^{Paires}$	$\#Mots^{en}$	$\#Mots^{fr}$
	(en milliers)		
Hansards	8 802	242	253
Europarl	2 007	132	142
Health	809	133	142
GigaWord	22 520	3 002	2 775

Tableau 4.I: Caractéristiques en nombre de mots et de phrases des quatre corpus parallèles sélectionnés.

Afin de mesurer ces performances oracle de l'ILB sur nos corpus bilingues, nous avons besoin d'une liste de termes sources et de leur traduction respective. Nous avons construit cette liste de référence à partir des articles de Wikipédia anglais qui sont liés à des articles français. Plus précisément, les titres des articles anglais et français sont respectivement les termes sources et leur traduction de référence. Quelques exemples de cette liste sont présents dans le tableau 4.II.

Terme source (anglais)	Traduction de référence (français)
Activities of daily living	Acte de la vie quotidienne
Boston Port Act	Acte du port de Boston
Layout Versus Schematic	Logiciel de vérification de schéma
Output gap	Écart de production
South Sulawesi languages	Langues sulawesi du Sud

Tableau 4.II: Exemples d'entrées de notre liste de référence.

En inspectant les titres des articles anglais, nous avons remarqué qu'un grand nombre d'entre eux était des expressions multi-mots. D'autres contenaient aussi des marques de ponctuation. Afin d'éviter trop de manipulations de chaînes de caractères durant ces expériences d'évaluation, nous avons simplement filtré ces titre de notre liste de référence.

[5] `http://www.statmt.org/wmt13/translation-task.html`

Nous avons aussi réalisé que beaucoup d'entrées étaient des entités-nommées qui, dans la paire de langue considérée (anglais-français), ne nécessitent pas d'être traduites. De fait, nous avons aussi supprimé ces entrées de notre liste[6]. Ajoutons qu'appliquer ces filtres (ponctuations et entités nommées) a réduit notre liste de référence de l'ordre de 75%, passant d'environ 835.000 entrées à approximativement 200.000 entrées, dont plus de 500.000 d'entités nommées. Cette dernière information laissant suggérer que Wikipedia est fortement orienté comme une encyclopédie d'entités nommées.

4.3 Évaluation

	Hansards		Europarl		Health		GigaWord	
n	EN	FR	EN	FR	EN	FR	EN	FR
0	95.2	97.4	96.0	98.0	96.0	97.6	<u>84.0</u>	89.5
1	0.8	0.5	0.8	0.5	0.9	0.6	2.3	2.0
2	0.5	0.3	0.4	0.2	0.5	0.3	1.3	1.0
3	0.3	0.2	0.2	0.1	0.3	0.2	0.9	0.6
4	0.2	0.1	0.2	0.1	0.2	0.1	0.6	0.4
5-10	0.6	0.4	0.5	0.3	0.6	0.4	2.2	1.5
11-100	1.2	0.7	1.0	0.5	1.0	0.6	4.7	2.9
>100	1.2	0.6	0.9	0.3	0.5	0.2	4.0	<u>2.0</u>

Tableau 4.III: Pourcentage de notre liste de référence où le mot anglais est identifié *n* fois du coté source du corpus parallèle (EN) ainsi que le nombre de fois où la traduction apparaît exactement *n* fois du coté cible (FR).

Les résultats de cette évaluation sont rapportés dans la table 4.III. Les résultats soulignés indiquent, par exemple, qu'il y a (seulement) 16% de nos termes anglais dans la partie source du corpus GigaWord et que (seulement) 2% des traductions françaises de ces termes apparaissent plus de 100 fois dans le coté cible. Ce constat est identique dans les autres corpus, sans oublier qu'ils sont plus petits que le corpus GigaWord. Ce qui est clairement insatisfaisant pour le succès de la tâche, que cela soit en terme de couverture ou en terme de fréquence d'occurrence des mots. En effet, c'est cette dernière information qui permet d'évaluer la capacité à identifier des traductions basée sur

[6]Grâce à un filtre rapide basé sur les catégories de Wikipédia.

l'alignement de mots en corpus parallèles.

4.4 Conclusion

Grâce à cette évaluation, nous avons montré que les corpus parallèles[7] ne nous permettraient pas d'obtenir des performances satisfaisantes si nous y appliquions l'induction de lexiques bilingues. En effet, les corpus parallèles peuvent offrir des performances excellentes grâce à l'alignement des mots mais encore faut-il que leur couverture soit adéquate, ce qui n'est pas le cas avec ces corpus. De fait, par la suite, nous nous sommes intéressés à l'application de l'ILB sur les corpus comparables sur lesquelles l'alignement est plus difficile à réaliser mais où la couverture est beaucoup plus large.

[7] Du moins, ceux en notre possession.

CHAPITRE 5

PROJECTIVE METHODS FOR MINING MISSING TRANSLATIONS IN DBPEDIA

Laurent Jakubina et Phillippe Langlais. Projective methods for mining missing translations in dbpedia. Dans *Proceedings of the Eighth Workshop on Building and Using Comparable Corpora*, pages 23–31, Beijing, China, July 2015. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W15-3404`

Cet article a été publié au 8^{eme} Workshop on Building and Using Comparable Corpora qui s'est tenu en même temps que la conférence ACL (Association of Computational Linguistics) de juillet 2015 à Pékin.

Cet article a été la première publication de mon doctorat, dans l'ordre chronologique. Il contient de fait mes premiers résultats obtenus sur la tâche d'Induction de Lexiques Bilingues (ILB), problème qui constituera la trame principale de mon doctorat[1]. Ainsi, les publications 2 (chapitre 6) et 4 (chapitre 7) utiliseront les résultats présentés dans cet article.

5.1 Contexte

DBpedia (Jens Lehmann, 2014) est le silo de données (dataset) le plus important[2] du Web Sémantique alias Linked Data (Bizer et al., 2009). Le Web Sémantique a été spécifié pour être indépendant de la langue, en ce sens que l'information y est représentée à l'aide de concepts abstraits[3] et l'accès à ces concepts en langage naturel se fait via des labels[4] définit par une relation de labellisation (ex : `rdfs:label` dans DBpedia).

[1] Excepté la troisième publication (chapitre 3).

[2] Décembre 2014 - `http://lod-cloud.net/`

[3] Par exemple : l'idée, l'abstraction que nous nous faisons de ce qu'est un *ordinateur*

[4] et nous nous référons à l'idée de l'*ordinateur* par le terme (label) **ordinateur** en français, **computer** en anglais.

Grâce à cette représentation, le Web Sémantique offre l'opportunité d'avoir un vrai World Wide Web Multilingue (Gracia et al., 2012). Le problème, au jour d'aujourd'hui, est que la majorité des labels sont en anglais (Gómez-Pérez et al., 2013). En effet, la plupart des silos de données résultent de l'extraction de ressources anglophones. DBpedia en fait partie, étant le silo RDF endogène de Wikipédia. De fait, DBpedia ne propose de labels français (`rdfs:label@fr`) que pour un cinquième[5] de ces concepts. Évidemment, ces concepts ont au moins tous un label anglais (`rdfs:label@en`). L'objectif serait donc d'apporter un `rdfs:label@fr` aux concepts qui n'en possèdent pas, par identification d'une traduction du `rdfs:label@en` de ces mêmes concepts. La relation `rdfs:label` retourne une chaîne de caractères "human-readable" décrivant un concept[6]. Cette chaîne de caractères peut être composée de plusieurs mots. Au sein d'un silo de données, ces chaînes de caractères sont isolées du point de vue textuel. En effet, dans un premier temps, nous ne considérons pas la structure du silo pour représenter le contexte des labels. Sous ces hypothèses, ce problème est fortement analogue au problème de l'induction de lexiques bilingues.

Nous nous sommes donc intéressés à ce problème en revisitant la tâche d'ILB sur le corpus comparable Wikipédia et, plus précisément, en explorant l'approche standard proposé par Rapp (1995). En effet, malgré l'abondance de travaux (Section 2.2.1), certaines considérations sont inexplorées ou délaissées, car les auteurs l'utilisent dans des cadres d'expérimentations qui visent à identifier des améliorations spécifiques[7]. Ainsi, nous assistons souvent à des échantillonnages non justifiés[8] sur les données des expériences. Ces échantillonnages se réalisent de plusieurs façons à travers l'ensemble des travaux existants, comme avec des filtres sur les mots en fonction de leur catégorie ou de leur fréquence (exemple : les mots ayant une fréquence inférieure à 50 ou 100 sont supprimés des corpus en prétraitement). Ce dernier type de filtre est très couramment utilisé (Gaussier et al., 2004, Morin et Prochasson, 2011b, Tamura et al., 2012,

[5] `http://wiki.dbpedia.org/Datasets/DatasetStatistics`
[6] `http://www.w3.org/TR/rdf-schema/`
[7] Ce qui n'est pas notre cas.
[8] Non justifiés vis à vis des objectifs visés par les publications, mais justifiés pour le côté computationnel.

Vulić et al., 2011) et cependant, son impact est rarement évalué. Ce dernier point est d'autant plus critique que nous savons que l'approche standard fonctionne très mal sur les mots rares (Bouamor et al., 2013a, Sharoff et al., 2013) et que peu d'études ont été réalisées sur ceux-ci. Suivant la logique de l'échantillonnage, personne n'avait encore évalué les performances de l'approche dans le contexte du Big Data, c'est à dire en utilisant d'énormes quantités de données non prétraitées directement. Enfin, à l'exception de (Laroche et Langlais, 2010) et de (Chiao et Zweigenbaum, 2002), rares sont ceux qui ont proposé des explorations systématiques et conjointes des paramètres, notamment en considérant les remarques précédentes.

Nos objectifs de recherche ont donc été, pour commencer, de reproduire l'approche standard d'induction de lexiques bilingues afin d'évaluer ses performances (en largeur) pour résoudre le problème du manque de labels français dans le Web Sémantique (DBpedia en particulier). En particulier, nous souhaitions utiliser un dump de Wikipédia complet, sans filtre sur le genre des mots ni sur les fréquences d'occurrences, afin de profiter de toute l'information contextuelle disponible. Exploitant la structure de Wikipédia (catégories, classements, etc.), nous avions émis l'hypothèse que des fonctions de voisinage (ensemble d'articles sémantiquement proches d'un mot via les catégories) pourraient aider à améliorer les performances de l'approche standard. Ensuite, afin de nous comparer à l'état de l'art, nous tenions à reproduire l'approche basée sur l'Analyse Sémantique Explicite (ASE) de Bouamor (2014). Finalement, nous voulions analyser les performances de l'ensemble de ces approches sur une liste de référence contenant des gammes de fréquences afin, notamment, d'étudier leur comportement sur les mots rares.

5.2 Contributions

Considérant le contexte "web sémantique", l'idée d'utiliser l'alignement de mots en corpus comparables pour induire des nouvelles traductions à intégrer dans DBpedia ainsi que l'idée des fonctions de voisinages sur Wikipédia vinrent de Philippe

Langlais. L'intégration de l'approche d'Analyse Sémantique Explicite a été choisie par moi-même. J'ai réalisé l'ensemble des expériences, en ayant préalablement implémenté l'approche de base d'alignement des mots en corpus comparables[9] (Rapp, 1995), la variante basée sur les fonctions de voisinages ainsi que l'approche d'Analyse Sémantique Explicite (Bouamor, 2014, Gabrilovich et Markovitch, 2007). Philippe Langlais m'a beaucoup aidé pour la rédaction de ce premier article.

5.3 Impacts

Dans cette publication, nous reproduisons deux approches utilisées pour induire des lexiques bilingues sur des corpus comparables. Une des plus vieilles, communément appelée approche standard (Rapp, 1995), et une des plus récentes, basée sur de l'Analyse Sémantique Explicite et proposée par Bouamor et al. (2013a). Avec ces deux approches, nous avons réalisé une des études les plus complètes à ce jour, notamment au niveau des méta-paramètres de chacune des approches ainsi que sur différentes gammes de fréquences de mots, ce qui nous permet de mettre en évidence le biais de ces approches en faveur des mots fréquents et d'appuyer la nécessité de penser à des méthodes pour les termes rares. Ainsi, malgré l'absence d'améliorations pour l'approche standard dans cette publication, notre travail vient confirmer l'état de l'art et améliorer la compréhension globale de la tâche sur plusieurs points (Bullinaria et Levy, 2007, Langlais et al., 2006, Prochasson et Fung, 2011, Sharoff et al., 2013).

Grâce à l'énorme corpus (Wikipédia) utilisé pour créer les vecteurs de contexte, nous montrons qu'il est irréalisable, computationnellement parlant, de profiter de toute l'information contextuelle disponible et que de l'échantillonnage est nécessaire à certains niveaux (notamment pour atteindre un temps de calcul raisonnable). Ainsi, dans notre étude et à l'opposé des autres travaux (section 5.1), nous avons adapté ces limites sur, par exemple, le nombre d'occurrences des mots très fréquents à considérer, et non pas en réduisant le vocabulaire disponible pour la tâche. En conclusion, une étape de cali-

[9] Sur le moteur de recherche open source Lucene[10] notamment.

bration conséquente est nécessaire quand nous souhaitons utiliser l'approche standard.

En outre, à partir de cette implémentation et exploration soigneusement conçues de l'approche standard, nous avons tenté de l'améliorer en utilisant des fonctions de voisinages disponibles grâce à Wikipédia mais ce fut sans succès. Des expérimentations avec l'approche de Bouamor et al. (2013a) nous ont permis d'identifier un biais encore plus fort vis-à-vis des mots fréquents, la rendant moins performante que l'approche standard sur des données de tests plus équilibrées (en fréquence des mots). Nous avons également exploré brièvement la possibilité de combiner leurs résultats pour améliorer les performances de la tâche en général.

5.4 Publication

[11]Each entry (concept) in DBpedia comes along a set of surface strings (RDFS :LABEL) which are possible realizations of the concept being described. Currently, only a fifth of the English DBpedia entries have a surface string in French, which severely limits the deployment of Semantic Web Annotation for this language. In this paper, we investigate the task of identifying missing translations, contrasting two projective approaches. We show that the problem is actually challenging, and that a carefully engineered baseline is not easy to outperform.

5.4.1 Introduction

The LOD (Linked Open Data) (Bizer et al., 2009) is conceived as a language independent resource in the sense that the information is represented by abstract concepts to which "human-readable" strings — possibly in different languages — are attached, *e.g.* the RDFS :LABEL property in DBpedia. For instance, we can access the abstract concept of computer by natural language queries such as ordinateur (RDFS :LABEL@FR) in French or computer (RDFS :LABEL@EN) in English. Thanks to this, Semantic Web offers the advantage of having a truly multilingual World Wide Web (Gracia et al., 2012).

At the core of LOD, lies DBpedia (Jens Lehmann, 2014), the largest dataset that constitutes a hub to which most other LOD datasets are linked [12]. Since DBpedia is (automatically) generated from Wikipedia, which is multilingual, one would expect that each concept in DBpedia is labeled with a French surface string. This is for instance the case of the concept House of Commons of Canada[13] which is labeled in French as Chambre des communes du Canada. One problem, however, is that most labels are currently in English (Gómez-Pérez et al., 2013).

Indeed, the majority of datasets in LOD are primarily generated from the extraction

[11]Certaines adaptations stylistiques ont été réalisées sur l'article afin de l'adapter au format de la thèse.
[12]December 2014 - http://lod-cloud.net/
[13]http://dbpedia.org/page/House_of_Commons_of_Canada

of anglophone resources. DBpedia, the endogenous RDF dataset of Wikipedia is no exception here, since it proposes labels in French (RDFS :LABEL@FR) for only one fifth[14] of the concepts. Of course, all concepts in English Wikipedia have at least one English label. For instance, the concept School life expectancy[15] has — at least at the time of writing — no label in French, while for instance, durée moyenne de scolarité appears in the (French) article Indice_de_développement_humain,[16] and is a good translation of the English term.

This situation comes from the fact that currently, a concept in DBpedia receives as its RDFS :LABEL property in a given language the title of the Wikipedia article which is inter-language linked to the (English) Wikipedia article associated to the DBpedia concept.

The lack of surface strings in a foreign language does not only reduce the usefulness of RDF indexing engines such as SIG.MA,[17] but also limits the deployment of Semantic Web Annotator (SWA) systems ; *e.g.* (Mihalcea et Csomai, 2007, Milne et Witten, 2008). This motivates the present study, which aims at automatically mining French labels for the concepts in DBpedia that do not possess one yet.

Identifying the translations of (English) Wikipedia article titles is partially solved in the BabelNet project (Navigli et Ponzetto, 2012). In this project, the translation of concepts in Wikipedia that are not inter-language linked are taken care of by applying machine translation on (minimum 3 and maximum 10) sentences extracted from Wikipedia that contain a link to the article whose title they seek to translate. The most frequent translation is finally selected. There are on the order of 500k articles in English Wikipedia that do not link to an article in French and which are not named entities (which typically do not require translation). BabelNet[18] provides a translation (not necessarily a good one) for 13% of them. This suggests that the projection of a resource such as

[14] http://wiki.dbpedia.org/Datasets/DatasetStatistics
[15] http://dbpedia.org/page/School_life_expectancy
[16] http://fr.wikipedia.org/wiki/Indice_de_developpement_humain
[17] http://sig.ma
[18] Version 2.0.1 - March 2014

DBpedia into French is not yet a solved problem.

In the remainder, we describe the approaches we tested in Section 5.4.2. Our experimental protocol is presented in Section 5.4.3. Section 5.4.4 reports the results we obtained. We conclude in Section 5.4.5.

5.4.2 Approaches

Identifying the translations of a term in a comparable corpus — two texts (one in each language of interest) that share similar topics without being in translation relation — is a challenge that has attracted many researchers. See (Sharoff et al., 2013) for a recent overview of the state-of-the-art in this field. In this work, we investigated several variants of two approaches for extracting translations from a comparable corpus : the seminal approach described in (Rapp, 1995) which uses a seed bilingual lexicon to induces new translations, and the approach of Bouamor et al. (2013a) which instead exploits the Wikipedia structure. The latter approach has been shown to outperform the former significantly on a task of translating 110 terms in 4 different domains, making use of medium-sized corpora.[19]

5.4.2.1 Standard Approach (Stand)

The idea that the context of a term and the one of its translation share similarities that can be used to rank translation candidates has been previously investigated in (Fung, 1998, Rapp, 1995). Since then, many variants of this idea have been tested ; see (Sharoff et al., 2013) for a recent discussion.

We reproduced this approach in this work. In a nutshell, each term to be translated is represented by a so-called *context vector* ; that is, the set of words that co-occur with this term in the source part of the corpus. An *association measure* is typically used to score the strength of the correlation between the term and the context words. Each transla-

[19] 400k words on the English side, 260k words on the French side.

tion candidate (typically each word of the target vocabulary) is similarly represented in the target language. Thanks to a *bilingual seed lexicon*, the source context vector is projected into a target one.[20] This projected target language vector is then compared to the vector of each of the target language candidates by the means of a *similarity measure*.

There are several parameters to the approach among which the size of the window used to collect co-occurrent words, the association and the similarity measures, as well as the seed lexicon.

We investigate the impact of the window size in section 5.4.4. We also compare two different association measures, namely the discontinuous odds-ratio (Evert, 2005, p. 86) named ORD hereafter, and the log-likelihood ratio (Dunning, 1993), named LLR, the most popular measures used in this line of work. Both measures (Eq. 5.1 and 5.2) are computed directly from the (monolingual) contingency table depicted in Table 5.I for two words w_1 and w_2 where, for instance, O_{12} stands for the number of times w_1 occurs in a window, while w_2 does not.

	w_1	$\neg w_1$	
w_2	O_{11}	O_{12}	R_1
$\neg w_2$	O_{21}	O_{22}	R_2
	C_1	C_2	N

Tableau 5.I: Contingency table

$$\text{ORD}(w_1, w_2) = \log \frac{(O_{11} + \frac{1}{2})(O_{22} + \frac{1}{2})}{(O_{12} + \frac{1}{2})(O_{21} + \frac{1}{2})} \tag{5.1}$$

$$\text{LLR}(w_1, w_2) = 2 \sum_{ij} O_{ij} \log \frac{N \times O_{ij}}{R_i \times C_j} \tag{5.2}$$

We did not investigate the impact of the nature and size of the bilingual seed lexicon, but decided to use one large lexicon comprising 116 354 word pairs populated from

[20] In our implementation, when no translation is found for a source word, the word is left as such in the target context vector. On the contrary, multiple translations are all added to the target context vector.

several available resources as well as an in-house bilingual lexicon.[21] A similar choice is made in (Bouamor et al., 2013a) where a seed lexicon of approximately 120 000 entries is being used, and in (Hazem et al., 2013a), where the authors use a lexicon of 200 000 entries (before preprocessing).

Since in (Laroche et Langlais, 2010) the best performing variant uses the cosine similarity measure (Eq. 5.3), we used it in our experiments.[22]

$$cos(v_{src}, v_{trg}) = \frac{v_{src} \cdot v_{trg}}{\|v_{src}\| \cdot \|v_{trg}\|} \tag{5.3}$$

In the standard approach, the co-occurrent words are extracted from all the source documents of the comparable corpus in which the term to translate appears. We name this variant STAND hereafter.

5.4.2.2 Neighbourhood variants (LKI, LKO, CMP and RA)

Since we are interested in translating Wikipedia titles, a natural way of populating the context vector of a term is to consider the occurrences of this term in the article whose title we seek to translate. This avoids populating the context vector with words co-occurring with different senses of the word to translate. We implemented such a variant which is inherently facing the issue that too few occurrences of the term of interest may appear in a single article, especially in our case where the average length of a Wikipedia article is approximatively 1 400 words. Therefore we considered a variant which involves a *neighbourhood function*, that is, a function that returns a set of Wikipedia articles related to the one under consideration for translation. We investigated three such functions (as well as many combinations of them) :

LKI(a) returns the set of articles that have a link pointing to the article a under conside-

[21] Ergane (12 914 entries - `http://download.travlang.com`), Freelang (38 869 entries - `http://www.freelang.net`), as well as an in-house lexicon (99 747 entries).

[22] Actually, the authors reported that with the LLR association measure, the Dice similarity was a better choice, but we kept along with the cosine measure for simplicity.

ration (in links). For instance, both COMPUTER_SCIENCE and ART are two articles pointing to ENTERTAINMENT.

LKO(a) returns the set of articles to which *a* points (out links). For instance the article ENTERTAINMENT points to PARTY and FUN.

CMP(a) returns the set of articles that are the most similar to *a*. We used the *MoreLikeThis* method of the search engine Lucene[23] for this. For instance, DANCE and DANCE IN INDONESIA are the top-2 documents returned by this function for the article ENTERTAINMENT.

For sanity check purposes, we also considered the RND function which randomly returns articles. Note that the LKI() and LKO() functions were obtained with the Wikipedia Miner toolkit (Milne et Witten, 2013).

5.4.2.3 Explicit Semantic Analysis (ESA-B)

We also implemented the approach described in (Bouamor, 2014) which has been shown by the author to be more accurate than the aforementioned standard approach. The proposed method is an adaptation of the Explicit Semantic Analysis approach described in (Gabrilovich et Markovitch, 2007).

A term to translate is represented by the titles of the Wikipedia articles in which it appears. The projection of the resulting context vector into the target language is obtained by following the available inter-language links.[24] The words of the articles reached this way are candidates to the translation and are further ranked by a tf-idf schema. This approach avoids the need for a seed bilingual lexicon, but uses instead the structure of Wikipedia, and its multilingualism more particularly.

One meta-parameter of this approach is the maximum size of the context vector, that is, the maximum number of article titles to keep for describing a term. One might think

[23] `http://www.lucene.org`
[24] Articles with no inter-language links are simply ignored.

that considering all the articles in which a term to translate is found is a good idea, but this strategy faces some sort of *semantic drift*. For instance, while translating the term tears, the context vector is populated with articles related to music albums that contain this term in their text content, while the associated French article (when available) almost never contains the translation. We investigate this meta-parameter in section 5.4.4. The other parameters were set as recommended in (Bouamor, 2014).

5.4.3 Experimental Protocol

5.4.3.1 Comparable corpus

DBpedia is extracted from Wikipedia (Jens Lehmann, 2014). Thus, we downloaded the Wikipedia dump of June 2013 in both English and French. The English dump contains 4 262 946 articles, and the French one contains 1 398 932. Although some articles that share an inter-language link are parallel (Patry et Langlais, 2011a), most article pairs are actually only comparable (Hovy et al., 2013).

5.4.3.2 English terms without translation

The vast majority (82,3%) of articles in the English Wikipedia do not have a link to an article in the French Wikipedia. We are interested to identify the translation of their title. Yet, we noticed that many of them are actually describing named entities (persons, geographic places, etc.), which typically do not require translation.[25] In order to filter named entities, we applied the BabelNet filter.[26] We ended up with a list of 521 895 (18,5%) terms we ultimately seek to translate. In this study, we further narrowed down our interest on unigrams.[27] This represents roughly 30% of those English terms.

[25] Some languages do involve transliteration, but this is definitely beyond the scope of this paper.

[26] We used the BabelSynset.getSynsetType() function of the BabelNet API for this purpose.

[27] Methods that handle multi-word expressions typically embed single word translation (Morin et Daille, 2009); therefore our choice.

5.4.3.3 Reference List

To evaluate our different approaches, we build a test set — a list of English source terms and their reference (French) translation. For this, we randomly sampled pairs of articles in Wikipedia that are inter-language linked. It is accepted that the titles of a pair of articles inter-language linked often constitute good translations (Hovy et al., 2013). Therefore, for each term (title) of our test set, we collected the associated title as a reference translation.

The sampling was done without considering named entities. For this purpose, we only considered article pairs which English title belongs to the bilingual lexicon we used as a seed lexicon for the Stand approach. Since the frequency of a source term is a key parameter of projective approaches, we also paid attention to vary the frequency range of the English terms we considered in our test set. More precisely, we gathered terms in those different ranges : infrequent [1-25], moderate [26-100], large [101-1000] and huge [1001+], where the frequency is the one in (English) Wikipedia. Some examples of pairs in each range are displayed in Table 5.II.

[1-25]	74 (8.5%)	
	myringotomy	paracentèse
[26-100]	267 (30.7%)	
	syllabification	césure
[101-1000]	259 (29.8%)	
	numerology	numérologie
[1001+]	269 (30.9%)	
	entertainment	divertissement
Total	869 (100%)	

Tableau 5.II: Distribution of the number of test forms at a given frequency range along with an example of an English term and its reference (French) translation.

We measured that using a large parallel corpus,[28] we could only identify the translation

[28]We gathered 32 millions of sentence pairs from different available parallel corpora, in-

of roughly 1% of those terms, which indicates that parallel data might be of little interest in identifying the translations of Wikipedia article titles.

5.4.3.4 Evaluation

Our approaches have been configured to produce a ranked list of (at most) 20 candidates for each source (English) term. We compute two metrics to compare them : precision at rank 1 (P@1) which indicates the percentage of terms for which the best ranked candidate is the reference one, and Mean Average Precision at rank 20 (MAP-20), a measure commonly used in information retrieval (Manning et al., 2008) which averages precision at various recall rates.

5.4.3.5 Technical considerations

The standard approach (STAND) can be rather computation and time consuming, since any target word in Wikipedia is a potential candidate for a given source term, and we are dealing with a rather large comparable corpus. Just as an illustration, the word france occurs more than 1 million times in the French Wikipedia, and its context vector potentially contains as much as 136 514 words (considering a context window of 6 words). Therefore, in our experiments, we only consider the first 50 000 occurrences of each term while populating the context vectors. Also, comparing source and target vectors can be time consuming, especially with context vectors of very high dimension. To save some time (and memory), we only represent a context vector (source or target) by (at most) the 1000 top-ranked terms according to the association measure being used.

cluding the GIGAWORD corpus we downloaded from `http://www.statmt.org/wmt13/translation-task.html`.

5.4.4 Results

5.4.4.1 Stand

In some calibration experiments,[29] we observed that increasing the size of the window in which we collect the context words leads to noise (see Table 5.III). The optimal window size was 6 (3 words on each side of the word under consideration, excluding function words), which means that the co-occurrent words should be taken in the immediate vicinity of the term to translate. This corroborates the study in (Bullinaria et Levy, 2007). Therefore, we set the value of this meta-parameter to 6 in the remainder.

|window|	MAP-20
2	0.72
6	0.75
14	0.62
30	0.55

Tableau 5.III: MAP-20 of Stand (ord) measured on a development set, as a function of the window size (counted in word).

	[1-25]		[26-100]		[101-1000]		[1001+]		[Total]	
	P@1	MP	P@1	MP	P@1	MP	P@1	MP	P@1	MP
Stand (llr)	.00	.00	.01	.01	.01	.02	.13	.15	.05	.06
Stand (ord)	.02	.05	.21	.28	.42	.47	.46	.50	.33	.38
Stand (o-100)	.02	.05	.14	.20	.15	.21	.10	.16	.12	.18
lki-1000	.00	.00	.06	.08	.12	.15	.12	.15	.09	.11
lko-1000	.00	.00	.01	.02	.08	.11	.03	.04	.04	.05
cmp-1000	.01	.02	.07	.09	.13	.17	.09	.12	.09	.12
rnd-1000	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
Esa-B	.01	.08	.05	.12	.20	.30	.42	.51	.21	.29

Tableau 5.IV: Precision (at rank 1) and MAP-20 (MP) of some variants we tested. Each neighbourhood function was asked to return (at most) 1000 English articles. The Esa-B variant is making use of context vectors of (at most) 30 titles.

The results of two variants of the standard approach are reported in Table 5.IV (line 1 and 2). Clearly, using ord as an association measure drastically improves perfor-

[29] We used a development set of 125 (unigram) terms, considering a candidate list of 50k words randomly selected to which we added the reference translations.

mance. This definitely corroborates the findings of Laroche et Langlais (2010). Still, the differences between both variants is surprisingly high : ORD delivers over six time higher performance than LLR does on average, while in the aforementioned work, the difference was much less marked.[30] Therefore, we use this association measure in the neighbourhood variants we tested.

We observed in practice the tendency of ORD to reward word pairs that appear often together even though the frequency of each word is very low. Thus, the context vector gathered with ORD tend to contain rare words that only appear in the context of the article under consideration. Those words offer a good discriminative power in our task, thus leading to much higher performance than the context vectors computed by LLR, which tend to gather more general related terms. This tendency can be observed in Figure 5.V where ORD leads to a context vector with much more specific words. This observation deserves further investigations.

ORD	LLR
myringoplasty (16.32)	**tube** (147.6)
myringa (16.14)	**laser** (44.90)
laryngotracheal (15.13)	**procedure** (40.83)
tympanostomy (14.60)	usually (31.86)
laryngomalacia (14.19)	**knife** (30.13)
patency (13.43)	**myringoplasty** (29.85)
equalized (11.75)	**ear** (28.19)
grommet (11.58)	**laryngotracheal** (27.45)
obstructive (11.09)	**tympanostomy** (26.39)
incision (10.37)	cold (24.09)

Tableau 5.V: Top words in the context vector computed with ORD and LLR for the source term MYRINGOTOMY. Words in bold appear in both context vectors.

A second observation that can be made is the strong correlation between the frequency of the term to translate and the performance of the approach. As a matter of fact, the performance for very frequent terms ([1001+]) is more than ten times the one measured on infrequent ones ([1–25]). This is a well-know fact that has been analyzed for instance

[30] In Table 3 of their article, the authors measured on a testset of 500 terms a MAP of 0.536 for ORD, and 0.413 for LLR.

in (Prochasson et Fung, 2011) where the authors report a precision of 60% for frequent test words (words seen at least 400 times), but only 5% for rare words (seen less than 15 times).

Overall, and even if a close comparison is difficult, the results we obtained for STAND are in-line with those reported in (Laroche et Langlais, 2010) that also focused on Wikipedia, but mining translations of medical terms. The authors reported a precision at rank one ranging from 20.7% up to 42.3% depending on test sets and configurations considered.

As we discussed in Section 5.4.3.5, due to computational issues, we cut the context vectors of the STAND approach after 1 000 terms. In order to measure how sensitive this cut-off is, we computed a variant where the top-100 terms only are kept (considering the association measure). The results of this variant are reported in line 3 of Table 5.IV. As expected, the performance of the STAND approach drops significantly on average, and especially for very frequent terms ([1001+]).

5.4.4.2 Neighbourhood variants

We tested our neighbourhood functions as well as several combinations of them. One meta-parameter we investigated is the maximum number of articles returned by a function. We early observed that the more the better, something we explain shortly. Thereafter, each function was asked to return at most 1 000 articles. The results obtained by the 3 neighbourhood functions we described in section 5.4.2 are reported in lines 4 to 6 of Table 5.IV.

Clearly, all the neighbourhood variants we considered yielded a significant drop in performance, which is disappointing from a practical point of view. This suggests that there is no obvious way to reduce the number of source documents to consider while populating the context vector of the term to translate. One explanation for this is that in our implementation, the context vector of each target candidate is computed by considering the full (French) Wikipedia collection. This dissymmetry introduces a mismatch

between the source and target context vectors, leading to poor performances. A solution to this problem consists in computing target context vectors online from a subset of target documents of interest.[31] A drawback of this solution is (of course) that the computation must take place for each term to translate. This is left as a future work.

At least, the neighbourhood variants we experimented outperform the one where random documents are sampled (RND). This latter variant could not translate a single term of the test set.

5.4.4.3 ESA-B

In the default configuration of the approach described in (Bouamor et al., 2013a), the authors limit the size of the context vector to 100, which we found suboptimal in our case. We varied the dimension of the context vectors and observed the best value to be 30 (see Table 5.VI). This is the value used in the sequel.

$\lvert context \rvert$	MAP-20
10	0.248
20	0.287
30	0.293
50	0.291
100	0.271

Tableau 5.VI: MAP-20 of ESA-B measured on the test set, as a function of the context vector dimension.

Somehow contrary to what has been observed in (Bouamor et al., 2013a), we observe that ESA-B (P@1 = 0.211) under-performs the STAND approach with the ORD association measure (P@1 = 0.338). One explanation for the difference is that, in (Bouamor et al., 2013a), the authors filter in words such as nouns, verbs and adjectives when populating the context vectors, while we do not. This filter might interfere with the observation made in section 5.4.4.1 that, with ORD, rare words (which might be filtered out, such as

[31] This subset could, for instance, be defined by following the inter-language links of the source documents returned by the neighbourhood function.

URLs or even spelling mistakes) tend to appear in the context vectors, and happen to help in discriminating translations.

5.4.4.4 Analysis

If we consider the 528 test terms that appear over a hundred times in Wikipedia ([101+]), a test case where both approaches perform well, STAND (ORD) translates correctly 362 of them (considering the top-20 solutions), while ESA-B translates 351. If we had an oracle telling us which variant to trust for a given term, we could translate correctly 431 terms (81.6%), which indicates the complementarity of both approaches.

We analyzed the 97 terms for which our two approaches failed to propose the reference translation in the top-20 candidates and we identified a number of recurrent cases we describe hereafter.

First, English terms do appear in the French Wikipedia material that eventually get selected by the STAND approach. This is, for instance the case for the term barber (oracle translation : coiffeur) for which STAND proposed the translation barber.

Second, we observed that STAND (and perhaps ESA-B in a less systematic way) often proposes morphological variants of the reference translation. For instance, coudre(a verbal form) is the first proposed translation for sewing, while the reference translation is the noun couture.

Third, it happens in a few cases that the reference translation, although correct is very specific. Of course this penalizes equally both approaches we tested. For instance, the reference translation of veneration is dulie, while the first translation produced by STAND is vénération (a correct translation).

Also, and by far the most frequent case, we observed a *thesaurus effect* of both approaches where terms related to the source one are proposed. This effect can be observed in Figure 5.1 in which top candidates proposed by several variants we tested are reported for the terms exemplified in Table 5.II.

Finally, it happens that the top-20 candidates proposed are just noise (e.g. noun translated as spora).

5.4.5 Discussion

In this study, we implemented and compared two projective approaches for identifying the translation of terms that correspond to articles in English Wikipedia that do not have an inter-language link to an article in the French Wikipedia. Doing so would potentially help in enriching the RDFS :LABEL property attached to concepts in DBpedia, thus easing semantic annotation in French. One method is a variant of the popular approach pioneered by (Rapp, 1995) which uses a bilingual seed lexicon for mapping source and target context vectors, and the other one has been proposed in (Bouamor et al., 2013a) for which the authors shown to deliver state-of-the-art performance.

Among other things, our experiments suggest that the STAND approach performs as well or better than the ESA-B approach and combining both approaches, especially for high frequency terms might improve our results.

We also observed the well-known bias of those approaches toward frequent terms, which urges the need for methods adapted to less frequent terms. As a future work, we will investigate the solution proposed in (Prochasson et Fung, 2011) which is one step in this direction.

Also, the projective methods we considered embed several meta-parameters which values are sensible. It is therefore difficult to know a priori which configuration to chose for a given task, without conducting costly calibration experiments. Having at our disposal a number of different test cases would help in developing expertise in doing so. With the hope that this might help, the code and resources used in this work will be available at this url : http://rali.iro.umontreal.ca/rali/?q=fr/Ressources

myringotomy [1-25]

ESA-B	–	laryngologie (0.209) oto (0.191) rhino (0.180) traitement (0.125) otite (0.080)
STAND (ORD)	–	permette (0.0489) devra (0.0473) scopie (0.0471) nécessitait (0.046) pût (0.045)
STAND (LLR)	–	melanosporum (0.274) neural (0.272) séminifère (0.269) ncathodique (0.269)

syllabification [26-100]

ESA-B	–	langues (0.517) consonne (0.420) langue (0.353) lettre (0.223) phonétique (0.166)
STAND (ORD)	–	modifier (0.079) suffit (0.074) vouloir (0.074) syllabique (0.074) intonation (0.072)
STAND (LLR)	–	édicté (0.106) exécutoire (0.097) syllabique (0.096) irrévocable (0.092)

numerology [101-1000]

ESA-B	20	œuvre (0.053) gematria (0.037) angels (0.031) nombres (0.029) chiffre (0.027)
STAND (ORD)	1	numérologie (0.095) occultisme (0.062) ésotérisme (0.062) divinatoire (0.058)
STAND (LLR)	5	jyotish (0.415) conditionaliste (0.412) karmique (0.364) domification (0.358)

entertainment [1001+]

ESA-B	2	entertainment (0.392) divertissement (0.151) vidéo (0.121) sony (0.111) jeu (0.073)
STAND (ORD)	–	beatmakers (0.012) manglobe (0.011) spycraft (0.011) déduplication (0.010)
STAND (LLR)	–	dsi (0.299) eshop (0.294) cocoto (0.231) ead (0.225) imagesoft (0.210)

Figure 5.1 : Top candidates produced by several variants of interest for some test terms. The second column indicates the rank of the oracle translation when present in the top-20 returned list (or – if absent).

CHAPITRE 6

A COMPARISON OF METHODS FOR IDENTIFYING THE TRANSLATION OF WORDS IN A COMPARABLE CORPUS : RECIPES AND LIMITS

Laurent Jakubina et Phillippe Langlais. A comparison of methods for identifying the translation of words in a comparable corpus: Recipes and limits. *Computación y Sistemas*, 20(3):449–458, 2016b. URL `http://www.cys.cic.ipn.mx/ojs/index.php/CyS/article/viewFile/2465/2169`

Cet article a été accepté à la 17^{eme} International Conference on Intelligent Text Processing and Computational Linguistics (CICLing). Il aurait dû être présenté sous forme de présentation à ladite conférence qui avait lieu en Turquie en avril 2016, mais cela n'a pas été le cas pour des raisons personnelles. Il a finalement été publié dans le journal Computación y Sistemas[1].

Cet article a été la deuxième publication de mon doctorat, dans l'ordre chronologique. D'un côté, il poursuit (en partie) les expériences débutées avec la première publication (chapitre 5), notamment avec une orientation sur les mots rares. D'un autre coté, il raccroche des travaux initiés par la communauté d'apprentissage machine sur la tâche d'ILB. Ce deuxième coté influencera énormément la suite de la thèse, vu que les résultats de cet article seront utilisés comme source de recherche pour le quatrième et dernier article (chapitre 7) de cette thèse.

6.1 Contexte

Le contexte de recherche plutôt "Web Sémantique" initié par notre première publication s'est vu profondément altéré par ma participation à la conférence Association of Computational Linguistics (ACL) de 2015. Plus précisément, le coté "Web Sémantique"

[1] http ://www.cys.cic.ipn.mx/ojs/index.php/CyS

a été "abandonné" pour se recentrer pleinement sur la tâche d'induction de lexiques bilingues en elle-même, et ce, dû à ma rencontre avec les "fameux" *word embeddings*. En effet, l'énorme intérêt autour de ceux-ci étant à son apogée lors de l'ACL 2015 (comme mentionné dans l'état de l'art, section 2.2.2), une grande majorité des résultats de recherche présentés à la conférence faisaient référence aux embeddings. C'était notamment le cas de la tâche d'ILB. Curiosité oblige, notre intérêt s'est aussi alors porté vers l'utilisation des embeddings pour nos recherches mais sans oublier les futurs travaux présentés dans le premier article, notamment sur les mots rares.

Nous nous sommes donc intéressés à la tâche d'ILB réalisée à l'aide des embeddings. L'état de l'art sur le sujet est présenté dans la section 2.2.2. Une consultation rapide des premières publications sur ce sujet fait ressortir la remarque suivante : peu, voir aucune liaison ni comparaison n'est réalisée avec l'état de l'art plus "ancien" de la tâche, c'est-à-dire avec l'approche standard discutée plus tôt (chapitre 2). De plus, il en ressort le sentiment que les approches à base d'embeddings présentées (jusqu'alors) surpassent (inexorablement) les approches plus anciennes, en partie parce qu'elles sont basées sur de l'apprentissage machine. Sans comparaison, nous pouvons nous demander pourquoi des approches qui utilisent la même source d'information de modélisation (le contexte de cooccurrence) afficheraient des performances significativement différentes. Ensuite, la section 2.2.2 montre qu'il y a une explosion de la quantité des approches basées sur les embeddings mais il est difficile de les comparer car les conditions d'expérimentations varient d'une étude à une autre. Ajoutons à cela qu'une partie de nos commentaires émis sur l'état de l'art précédent (chapitre 5) restent vrais pour le "nouvel" état de l'art : échantillonnages expérimentaux (basé sur la fréquence des mots, notamment) et pas de résultats ni d'analyses sur les mots rares.

Nos objectifs de recherche ont été, entre autres, de reproduire une approche d'ILB basée sur les embeddings afin de comprendre ses rouages et de les analyser pour ensuite comparer l'ensemble (et les performances) avec l'approche standard, notamment sur les protocoles d'expérimentation définis lors du premier article. Ensuite, nous voulions poursuivre notre recherche sur les mots rares en suivant nos indications de travaux

futurs de la publication précédente : reproduction d'une approche de l'état de l'art (approche standard) s'étant intéressée au problème des mots rares. Finalement, nous profitions de l'ensemble nous permettant de mettre en perspective 3 approches aux modélisations différentes afin de fournir des "recettes" pour le choix des méta-paramètres ainsi que des analyses sur ces 3 approches. Potentiellement, nous souhaitions aussi évaluer la possibilité de réaliser une combinaison entre les approches sur le problème de l'induction de lexiques bilingues.

6.2 Contributions

L'idée d'investiguer les approches à base d'embeddings est venue de moi, ainsi que le choix de l'approche orientée pour les mots rares. J'ai réalisé l'ensemble des expériences, notamment en utilisant Word2Vec. La notion de "recettes" présentée dans l'article, ainsi que les analyses basées sur la propriété de similarité des chaînes de caractères et des termes médicaux sont des points proposés par mon directeur de recherche. Il m'a beaucoup aidé pour la rédaction de l'article encore une fois.

6.3 Impacts

Avec cette publication, nous avons poursuivi la méthodologie débutée dans la publication précédente (chapitre 5) c'est-à-dire de comparer des approches (Mikolov et al., 2013b, Prochasson et Fung, 2011, Rapp, 1995) entre elles tout en contribuant à deux problématiques importantes de la tâche d'induction de lexiques bilingues : le problème des mots rares ainsi qu'à la compréhension – autant le fonctionnement que les résultats proposés – des toutes récentes approches d'ILB à base d'embeddings provenant de la communauté d'apprentissage machine.

Ainsi, nous ne présentons pas de nouveaux modèles d'embeddings [2] (fonction objectif) ni d'amélioration pour l'approche standard mais nous proposons une étude soigneu-

[2] Comme c'était à la mode à cette période.

sement menée de chaque approche sur différents jeux de données. Ultimement, ces ensembles de données nous ont permis de mieux comprendre l'impact des différents méta-paramètres de chaque approche ainsi que de leur fonctionnement global (surtout de la plus récente approche à base de réseaux de neurones). Grâce à cette compréhension, nous avons identifié les meilleures valeurs pour les méta-paramètres de chaque approche sur chaque jeu de données. C'est ce que nous rapportons dans l'article via l'expression : "recettes".

En outre, cette étude méticuleuse nous a aussi permis de conclure qu'il est possible d'obtenir les mêmes performances entre les (plus vieilles) approches statistiques à base de comptages et les approches (plus récentes) à base d'apprentissage machine. Notre travail montre que les premières demandent une phase d'ingénierie (notamment, des méta-paramètres) plus poussée alors que les méthodes qui généralisent à partir des données proposent naturellement de bonnes performances sans cette phase. Ce constat fut confirmé par d'autres auteurs durant la même période (Levy et al., 2015).

Cependant, cette généralisation à partir de données n'est effective que si les données sont fréquentes[3]. Ainsi, nous montrons que les approches à base de plongement (du moins celles étudiées) possèdent le même biais envers les mots fréquents que l'approche classique. Pour les mots les moins fréquents que nous avons testés, la meilleure approche (celle à base de plongement) enregistre une précision de seulement 2.2% en TOP@1. En outre, nous n'avons pas réussi à reproduire les résultats de l'approche de Prochasson et Fung (2011) et en conséquence, le problème des mots rares reste un problème ouvert. De façon complémentaire à l'impact de la fréquence, nous étudions aussi l'impact du genre sémantique (précisément, des termes médicaux qui ont souvent été utilisés à des fins de tests dans les travaux mentionnés précédemment) ainsi que de la similarité des chaînes de caractères sur les performances des approches.

À l'instar de l'article précédent (chapitre 5), nous avons évalué la complémentarité des approches entre elles mais cette fois, les résultats nous permettent d'affirmer qu'il est possible de les combiner. Cette combinaison est devenue dès lors le sujet principal de

[3]Comme toute approche statistique.

l'article présenté au chapitre 7.

6.4 Publication

[4]Identifying translations in comparable corpora is a challenge that has attracted many researchers since a long time. It has applications in several applications including Machine Translation and Cross-lingual Information Retrieval. In this study we compare three state-of-the-art approaches for these tasks : the so-called context-based projection method, the projection of monolingual word embeddings, as well as a method dedicated to identify translations of rare words. We carefully explore the hyper-parameters of each method and measure their impact on the task of identifying the translation of English words in Wikipedia into French. Contrary to the standard practice, we designed a test case where we do not resort to heuristics in order to pre-select the target vocabulary among which to find translations, therefore pushing each method to its limit. We show that all the approaches we tested have a clear bias toward frequent words. In fact, the best approach we tested could identify the translation of a third of a set of frequent test words, while it could only translate around 10% of rare words.

6.4.1 Introduction

Extracting bilingual lexicons from comparable corpora has received a massive interest in the NLP community, mainly because parallel data is a scare resource, especially for specific domains. More than twenty years ago, (Fung, 1995) and (Rapp, 1995) described two methods for how this can be accomplished. While those studies differ in the nature of their underlying assumption, they both assume that the context of a word shares some properties with the context of its translation. In the case of (Rapp, 1995) the assumption is that words in translation relation show similar co-occurrence patterns.

Many variants of (Rapp, 1995) have been proposed since then. Some studies have for instance reported gains by considering syntactically motivated co-occurrences, either by the use of a parser (Yu et Tsujii, 2009) or by simple POS-based patterns (Otero, 2007). Extensions of the approach in order to account for multiword expressions have also been

[4]Certaines adaptations stylistiques ont été réalisées sur l'article afin de l'adapter au format de la thèse.

proposed (*e.g.* (Daille et Morin, 2008)). Also, many have studied variants for extracting domain specific lexicons ; the medical domain being vastly studied, see for instance (Chiao et al., 2004) and (Morin et Prochasson, 2011b). We refer the reader to (Sharoff et al., 2013) for an extensive overview of works conducted in this vein.

There is a trend of works on understanding the limits of this approach. See for instance the study of (Laroche et Langlais, 2010) in which a number of variants are being compared, or more recently the work of (Jakubina et Langlais, 2015). One important limitation of the context-based approach is its vulnerability to rare words, which has been demonstrated in (Pekar et al., 2006). The authors reported gains by predicting missing co-occurences thanks to co-occurences observed for similar words in the same language. In (Prochasson et Fung, 2011), the authors adapt the alignment technique of (Gale et Church, 1991) initially proposed for parallel corpora by exploiting the document pairing structure of Wikipedia. The authors show that coupling this alignment technique with a classifier trained to recognize translation pairs, yield an impressive gain in performance for rare words.

Of course, many other approaches have been reported for mining translations in comparable corpora ; again see (Sharoff et al., 2013) for an overview. A very attractive approach these days is to rely on so-called word embeddings trained with neural-networks thanks to gradient descent on large quantities of texts. In (Mikolov et al., 2013a) the authors describe two efficient models of such embeddings which can be computed with the popular `word2vec` toolkit. In (Mikolov et al., 2013b) it is further shown that a mapping between word embeddings learnt independently for each language can be trained by making use of a seed bilingual lexicon. Since then, many refinements of this approach have been proposed (see (Levy et al., 2016) for a recent comparison), but (Mikolov et al., 2013b) remains a popular solution, due to its speed and performance.[5]

[5] As a side note, we observed on a Spanish-English task similar to the one studied in this paper that the approach of (Mikolov et al., 2013b) performs on par (but much faster) with `Bilbowa` (Gouws et al., 2015), provided large enough embedding dimensions (400 or more).

In this study we compare three of the aforementioned approaches : the context-based approach (Rapp, 1995), the word embedding approach of (Mikolov et al., 2013b) and the approach of (Prochasson et Fung, 2011) dedicated to rare words. We investigate their hyper-parameters and test them on words with various properties. This comparison has been conducted at a large scale, making use of the full English-French Wikipedia collection.

In the remainder of this article, we describe in Section 6.4.2 the approaches we tested. Our experimental protocol is presented in Section 6.4.3. Section 6.4.4 summarizes the best results we obtain with each approach, and report on the impact of their hyper-parameters on performances. In Section 6.4.5, we analyze the bias those approaches have toward some word properties. We conclude in Section 6.4.6.

6.4.2 Approaches

We implemented and tested three state-of-the-art alignment techniques. Two of them (`context` and `embedding`) makes use of a so-called bilingual seed lexicon, that is, pairs of words in translation relation ; the other (`document`) is exploiting instead the structure of the comparable collection.

6.4.2.1 Context-Based Projection (`context`)

In (Rapp, 1995), each word of interest is represented by a so-called context vector (the words it co-occurs with). Source vectors are projected (or "translated") thanks to a seed bilingual lexicon. Candidate translations are then identified by comparing projected contexts with target ones, thanks to a similarity measure such as cosine.

We conducted our experiments using the `eCVa` Toolkit[6] (Jakubina et Langlais, 2015) which implements several association measures (for building up the context vectors) and similarities (for comparing vectors).

[6] `http://rali.iro.umontreal.ca/rali/en/ecva-toolkit`

6.4.2.2 Document-based Alignment (`document`)

This approach relies on a pre-established pairing of comparable documents in the collection. Given such a collection, word translations are identified based on the assumption that source and target words should appear collection-wise in similar pairs of (comparable) documents.

The approach was initially proposed for handling the case of rare words for which co-occurrence vectors are very sparse. On a task of identifying the translation of medical terms, the authors showed the clear superiority of this approach over the context-based projection approach. In their paper, (Prochasson et Fung, 2011) show that coupling this approach to a classifier trained to recognize translation pairs is fruitful. We did not implement this second stage however, since the performance of the first stage was not judged high enough, as discussed later on.

6.4.2.3 Word Embedding Alignment (`embedding`)

Word Embeddings (continuous representations of words) has attracted many NLP researchers recently. In (Mikolov et al., 2013b), the authors report on an approach where word embeddings trained mono-lingually are linearly mapped thanks to a projection matrix W which coefficients are determined by gradient descent in order to minimize the distance between projected embeddings and target ones, thanks to a seed bilingual lexicon $\{(x_i, y_i)\}_{i\in[1,n]}$:

$$\min_{W} \sum_{i=1}^{n} ||W\hat{x_i} - \hat{y_i}||^2 \tag{6.1}$$

where $\hat{x_i}$ and $\hat{y_i}$ are the (monolingual) embeddings of the words x_i and y_i respectively. Given a term x absent from the seed lexicon but for which an embedding $\hat{x}$ has been trained, translations are determined by selecting the words whose target embeddings are the closest to the projected one, $W\hat{x}$, thanks to a distance (cosine in their case).

We trained monolingual embeddings (source and target) thanks to the `Word2Vec`[7] toolkit. The linear mapping was learnt by the implementation described in (Dinu et Baroni, 2014).

6.4.3 Experimental Protocol

The approaches described in the previous section have been configured to produce a ranked list of (at most) 20 candidate French translations for a set of English test words. We measure their performance with accuracy at rank 1, 5 and 20; where accuracy at rank i (TOP@i) is computed as the percentage of test words for which a reference translation is identified in the first i candidates proposed.

Each approach used the exact same comparable collection, and possibly a seed lexicon. In the remainder, we describe all the resources we used.

6.4.3.1 Comparable Corpus

We downloaded the Wikipedia dump of June 2013 in both English and French. The English dump contains 3 539 093 articles, and the French one contains 1 334 116. The total number of documents paired by an inter-language link is 757 287. While some pairs of documents are likely parallel (Patry et Langlais, 2011a), most ones are only comparable (Hovy et al., 2013). The English vocabulary totalizes 7.3 million words (1.2 billion tokens); while the French vocabulary counts 3.6 million ones (330 million tokens).

We used all the collection without any particular cleaning, which departs from similar studies where heuristics are being used either to reduce the size of the collection or the list of candidate terms among which a translation is searched for. For instance, in (Prochasson et Fung, 2011) the authors built a comparable corpus of 20 169 document pairs and a target vocabulary of 128 831 words. Also, they concentrated on nouns only. This is by far a smaller setting than the one studied here. While our choice brings some

[7] `https://code.google.com/p/word2vec/`

technical issues (computing context vectors for more than 3M words is for instance rather challenging), we feel it gives a better picture of the merit of the approaches we tested.

6.4.3.2 Test Sets

We built two test sets for evaluating our approaches ; one named **1k-low** gathering 1 000 *rare* English words and their translations, where we defined rare words as those occurring at most 25 times in English Wikipedia ; and **1k-high** gathering 1 000 words occurring more than 25 times. For the record, 6.8 million words (92%) in English Wikipedia occur less than 26 times.

The reference translations were collected by crossing the vocabulary of French Wikipedia with a large in-house bilingual lexicon. Half of the test words have only one reference translation, the remainder having an average of 3 translations. It should be clear that each approach we tested could potentially identify the translations of each test word, and therefore have a perfect recall.

6.4.3.3 Seed Bilingual Lexicon

The `context` and `embedding` approaches both require a seed bilingual lexicon. We used the part of our in-house lexicon not used for compiling the test sets aforementioned. For the `embedding` approach, we followed the advice of (Dinu et Baroni, 2014) and compiled lexicons of size up to 5 000 entries.[8] More precisely, we prepared three seed lexicons : **2k-low** which gathers 2 000 entries involving rare English words (words occurring at most 25 times) ; **5k-high** gathering 5 000 entries whose English words are not rare, and **5k-rand** which gathers 5 000 entries randomly picked. For the `context` approach, we used 107 799 words of our in-house dictionary not belonging to the test material.

[8]The authors tried larger lexicons without success.

6.4.4 Results and Recipes

In this section, we report on the best performance we obtained with the approaches we experimented with. We further explore the impact of their hyper-parameters on performance.

6.4.4.1 Overall performance

	1k-low			**1k-high**		
	TOP@1	TOP@5	TOP@20	TOP@1	TOP@5	TOP@20
`embedding`	**2.2**	**6.1**	**11.9**	**21.7**	**34.2**	**44.9**
`context`	2.0	4.3	7.6	19.0	32.7	44.3
`document`	0.7	2.3	5.0	10.0	19.0	24.0
oracle	4.6	10.5	19.0	31.8	46.8	57.6

Tableau 6.I: Performance of the different approaches discussed in Section 6.4.2 on our two test sets. The best variant according to TOP@1 is reported for each approach. On **1k-high**, we only computed the performance of the `document` approach on a subset of 100 entries.

The performance of each approach on the two test sets are reported in Table 6.I, where we only report the best variant of each approach according to TOP@1. This table calls for some comments. First, we observe a huge performance drop of the approaches when asked to translate rare words. While `context` and `embedding` perform (roughly) equally well on frequent words, with an accuracy at rank 1 of around 20%, and 45% at rank 20, both approaches on the **1k-low** test set could translate correctly only 2% of the test words in the first place ; the `embedding` approach being the less impacted at rank 20 with 12% of test words being correctly translated. What comes to a disappointment is the poor performance of the `document` approach which was specifically designed to handle rare words. We come back to this issue later on.

It is interesting to note that approaches are complementary as evaluated by an oracle which picks one of the three candidate lists produced for each term. This is the figure reported in the last line of Table 6.I. On rare words, we observe almost twice the performance of individual approaches, while on **1k-high**, we observe an absolute gain of 10

to 15 points in TOP@1, TOP@5, and TOP@20. This said, we observe that no more than 57% (resp. 19%) of the test words in **1k-high** (resp. **1k-low**) could be translated in the top-20 positions, which is disappointing.

Examples of outputs produced by our best configurations are reported in Figure 6.1. While it is rather difficult to pinpoint why many test words were not translated, we observe tendencies. First, we notice a "thesaurus effect", that is, candidates are often related to the words being translated, without being translations, as aromatisé (aromatized) proposed for the English word donut. Some errors are simply due to morphological variations and could have been counted correct, as pathologique produced instead of pathologiquement. We also observe a few cases where candidates are acceptable, but simply not appear in our reference.

donut		beigne		
context	-	aromatisé (0.05)	donut (0.05)	beignet (0.04)
embedding	-	liper (0.54)	babalous (0.53)	savonnettes (0.52)
brilliantly		brillamment		
context	-	imaginatif (0.05)	captivant (0.05)	rusé (0.05)
embedding	-	éclatant (0.69)	pathétique (0.67)	émouvant (0.66)
gentle		doucet,	doux,	délicat,
context	-	enjoué (0.05)	serviable (0.05)	affable (0.04)
embedding	-	colérique (0.76)	enjoué (0.75)	espiègle (0.75)
pathologically		pathologiquement		
context	-	cordonale (0.05)	pathologique (0.05)	diagnostiqué (0.05)
embedding	-	psychosexuel (0.60)	psychoaffectif (0.60)	piloérection (0.59)

Figure 6.1 : Top-3 candidates produced by the two best approaches for a few test words.

We compared a number of variants of each approach in order to have a better understanding of their merits. We summarize the main outcomes of our investigations in the following subsections.

6.4.4.2 Recipe for the `context` approach

We ran over 50 variants of the context-based approach, varying some meta-parameters, the influence of which is summarized in the sequel.

Each word in a context vector can be weighted by the strength of its relationship with the word being translated. We tested 4 main association measures and found PMI (point-wise mutual information) and discontinuous odds-ratio (see (Evert, 2005)) to be the best ones. Other popular association measures such as log-likelihood ratios drastically underperforms on **1k-high** (TOP@20 of 7.8 compared to 44.3 for PMI).

Context words are typically picked within a window centered on the word to translate. While the optimal window size somehow varies with the association measure considered, we found the best results for a window size of 7 (3 words before and after) for the **1k-high** test set, and a much larger window size (31) for the **1k-low** test set. For rare words, context vectors are very sparse, therefore, increasing the window size leads to better performance.

Last, we observed a huge boost in performance at projection time by including in the projected vector source words unknown from our seed lexicon. Without doing this, the best configuration we tested decreased in TOP@20 from 44.3 to 31.7. Our explanation for this unexpected gain is that some of those words are proper names or acronyms which presence in the context vector might help to discriminate translations.

6.4.4.3 About the `document` approach

Since this approach does not deliver competitive results (see Table 6.I) we did not investigate many configurations as we did for the other two approaches we tested. One reason for the disappointing results we observed compared to the gains reported in (Prochasson et Fung, 2011) might be the very different nature of the datasets used in our experiments. As a matter of fact, we used the entire English-French Wikipedia collection while in (Prochasson et Fung, 2011) they only selected a set of 20 000 document

pairs. Also our target vocabulary contains almost 3 millions words while theirs is gathering 120k nouns.

We conducted a sanity check where we randomly selected from our target vocabulary a subset of 120k words to which we added the reference translations of our test words (so that the aligner could identify them). On the **1k-low** test set, this led to an increased performance with a TOP@1 of 4.9 (compared to 0.7) and a TOP@20 of 20.2 (compared to 5.0). Those results are actually very much in line with the ones reported by the authors, suggesting that the approach does not scale well to large datasets.

6.4.4.4 Recipe for the `embedding` approach

We trained 130 configurations varying meta-parameters of the approach. In the following, we summarize our main observations.

First of all, the configurations which perform the best on **1k-high** and **1k-low** are different. On the former test set, our best configuration consists in training embeddings with the *cbow* model and the *negative sampling* (10 samples). The best window size we observed is 11, and increasing the dimensionality of the embeddings increases performance steadily. The largest dimensionality for which we managed to train a model is 200.[9] Those findings confirm observations made by (Dinu et Baroni, 2014) for frequent terms. In this work the authors managed to trained a model with embeddings of size 300. On their task, the best model trained performs a TOP@1 of 30 while on our task, the configuration described achieved a TOP@1 of 22. In (Mikolov et al., 2013a) the author also report a TOP@1 of around 30 for embeddings of size 1000. Recall that in our case, the target vocabulary size at test time is around 3 million words, while for instance in (Mikolov et al., 2013a) it is in the order of a hundred thousands (depending of the language pair considered), which might account for some differences in performance.

For the **1k-low** test set, the best configuration we found consists in training a *skip-gram* model with *hierarchical softmax*, and a window size of 21 (10 words on both sides of

[9] We ran our training on nodes of a cluster that can accommodate up to 64 Gb of memory.

each word), and an embedding dimensionality of 250 (the largest value we could afford on our computer). Again this confirms tendencies observed in (Dinu et Baroni, 2014) for the case of translating unfrequent words. Also, (Mikolov et al., 2013b) observed that the *skip-gram* model and the *hierarchical softmax* training algorithm are both preferable when translating unfrequent words.

Regarding the influence of the seed lexicon, we observed that using the largest one is preferable. This confirms the findings in (Dinu et Baroni, 2014) that a lexicon of 5k is optimal for training the mapping of embeddings. For the low frequency test set, we further observed that using a seed lexicon of random words (**5k-rand**) is better. By doing so, we could improve TOP@1 of 1 absolute point and TOP@20 of 3 points. On the **1k-high** test set, the best performance is obtained with **5k-high** (TOP@1 of 44.9%), then **5k-rand** (TOP@1 of 40.5%) and **2k-low** (TOP@1 of 10.3%).

6.4.5 Analysis

In the previous sections, we analyzed the impact of the hyper-parameters of each approach on performance. In this section, we analyze more precisely the results of the best configuration of each aligner in terms of a few properties of our test set. We believe such an analysis useful for comparison purposes. Also, in order to foster reproducibility, we are happy to share the test sets as well as the seed lexicons we used in this study.[10]

6.4.5.1 Frequency

We already observed a clear bias of the approaches we tested toward frequency. Figure 6.2 reports the performance of the best configuration of each approach when translating test words which frequency in English Wikipedia do not exceed a given threshold. For instance, we observe that on the subset of test words which frequency is 10 or less, the best approach according to TOP@1 (`embedding`) achieves a score of 8.76%.

[10]Downloadable at `http://rali.iro.umontreal.ca/rali/en/ecva-toolkit`.

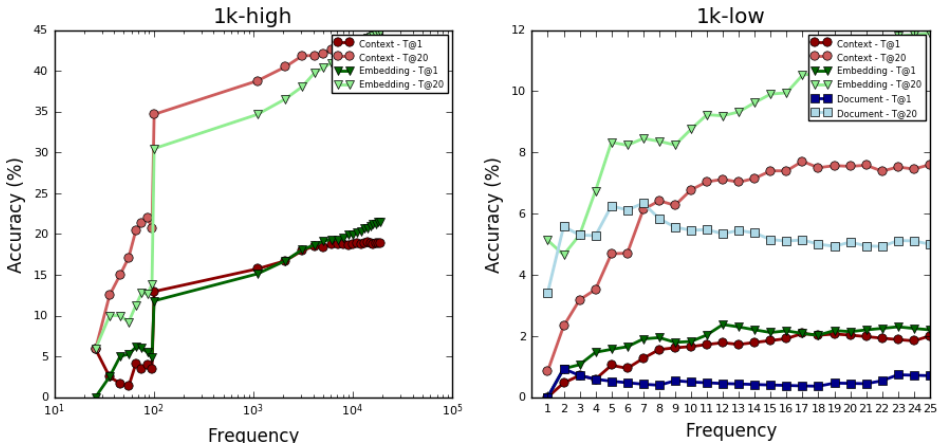

Figure 6.2 : TOP@1 and TOP@20 of the best configurations of each approach as a function of the frequency of the test words in English Wikipedia.

The frequency biais is clearly observable, and even for rather large frequency thresholds. Both `context` and `embedding` compete across frequencies, but if not too frequent test words have to be translated, and if a shortlist is what matters (TOP@20), then `context` might be the good approach to go with.

6.4.5.2 String Similarity

It is rather difficult in the relevant literature to get a clear sense of the intrinsic difficulty of the translation task being tackled. In particular, the similarity of test words and their reference translation is almost never reported, while for some language pairs, it is a relevant information that is even used as a feature in some approaches (*e.g.*, (Laroche et Langlais, 2010)). Figure 6.3 illustrates the performance of our best configurations as a function of the edit-distance between test words and their reference translation.[11]

As we expected, we observe overall a decrease of performance for words which reference translation is dissimilar. On words translated verbatim, TOP@1 performance is as high as 79.3% for `context` and 61.8% for `embedding`, while both approaches

[11]For easing the interpretation, we only considered test words having only one reference translation.

compare as the edit-distance augments. On rare words, `embedding` seems to be less sensitive to the edit-distance between the source term and its reference translation.

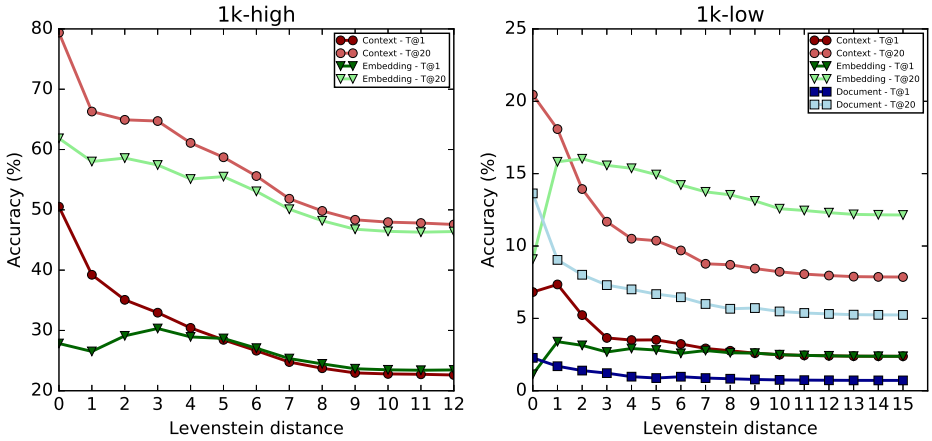

Figure 6.3 : Accuracy TOP@1 and TOP@20 of the best configurations as a function of the edit-distance between test words and their reference translation.

6.4.5.3 Medical terms

	TOP@1		TOP@20	
		1k-low		
`embedding`	4.5	(+2.7)	13.6	(+1.7)
`context`	0.0	(-2.0)	4.5	(-3.1)
`document`	4.5	(+3.8)	22.7	(+17.7)
		1k-high		
`embedding`	27.5	(+5.8)	53.7	(+8.8)
`context`	48.7	(+29.7)	72.5	(+28.3)

Tableau 6.II: Performance for medical test words. Figures in parenthesis are absolute gains over the performance measured over the full test set.

Medical term translation is the subject of many papers (*e.g.* (Hazem et Morin, 2012, Kontonatsios et al., 2014, Morin et Prochasson, 2011b) to mention just a few). In order to measure if this very task is easier than translating any kind of word, we filtered our test

words with an in-house list of 22 773 medical terms. We found only 22 medical terms in **1k-low** and 80 in **1k-high**. Although those figures are definitely not representative, we computed the performance of our best configurations on those subsets. The results are reported in Table 6.II. On frequent words, the gain in performance is especially marked for the `context` approach. We also note that the `document` approach seems to perform rather well on unfrequent medical terms, which is exactly the setting studied in (Prochasson et Fung, 2011). One possible explanation for this positive difference is that medical terms, at least in English Wikipedia tend to be rather frequent and their translation into French have an average edit-distance which is lower than for other types of words, two factors we have shown to impact performance positively.

6.4.6 Conclusion

In this study, we compared three approaches for identifying translations in a comparable corpus, and studied extensively how their hyper-parameters impact performance. We tested those approaches without reducing (somehow arbitrarily) the size of the target vocabulary among which to choose candidate translations. We also analyzed a number of properties of the test sets that we feel are worth reporting on when conducting such a task; among which the distribution of test words according to their frequency in the comparable collection and the distribution of their string similarity to their reference translation.

Among the observations we made, we noticed that the good old context-based projection approach (Rapp, 1995) when appropriately configured competes with the more recent neural-network one (especially on frequent words). This observation echoes the observation made in (Levy et al., 2015) that carefully tuned count-based distributional methods are no worse than trained word-embeddings. This said, in our experiments, the embedding approach revealed itself as the method of choice overall. We also observed that the approach of (Prochasson et Fung, 2011) designed specifically for handling rare words, while being good at translating medical terms had a harder time translating other types of (unfrequent) words. Definitely, translating rare words is a challenge that

deserves further investigations, especially since unfrequent words are pervasive.

We considered only one translation direction in this work. Since mining word translations is likely more useful for language pairs for which limited data is available, we plan to investigate such a setting in future investigations. Also, we would like to compare other approaches to acquire cross-lingual embeddings.

We also provide evidence that the approaches we tested are complementary and that combining their outputs should be fruitful. Since a given approach typically shows different performance depending on the properties of test words (their frequency, etc), it is also likely that combining different variants of the same approach should lead to better performance. This is left as a future work.

CHAPITRE 7

RERANKING TRANSLATION CANDIDATES PRODUCED BY SEVERAL BILINGUAL WORD SIMILARITY SOURCES

Laurent Jakubina et Phillippe Langlais. Reranking translation candidates produced by several bilingual word similarity sources. Dans *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 605–611, Valencia, Spain, April 2017. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/E17-2096`

Cet article a été accepté à la 15^{e} European Chapter of the Association for Computational Linguistics (EACL), au format court et présenté sous forme de poster à ladite conférence, qui avait lieu en Espagne (Valence) en avril 2017.

Cet article a été le quatrième et dernier de mon doctorat, dans l'ordre chronologique. Il relate les expériences réalisées sur la combinaison de résultats obtenus par différentes approches d'ILB, notamment ceux de l'approche standard étudiés dans le premier article (chapitre 5) et ceux des approches basées sur les embeddings, rapportés dans la troisième publication (chapitre 6).

7.1 Contexte

Le contexte de recherche de cette publication résulte des travaux futurs annoncés lors des précédentes publications, plus précisément de la combinaison de résultats de plusieurs approches d'induction de lexiques bilingues. Nous avons énoncé la possibilité d'une combinaison dans la première publication et nous avons montré que les approches étaient complémentaires dans le second article et qu'une combinaison devrait améliorer les performances. De plus, plusieurs auteurs de l'état de l'art de l'ILB (Irvine et Callison-Burch, 2017, Sharoff et al., 2013) mentionnent que c'est un axe de recherche

qui requiert plus d'attention, surtout avec un aspect d'apprentissage supervisé.

L'état de l'art de la tâche d'ILB via reclassement supervisé de candidats à la traduction (section 2.2.3) montre que peu de travaux ont porté attention à ce sujet. À cela, ajoutons que personne ne semble avoir exploré la possibilité de combiner les résultats de plusieurs signaux provenant de la même source[1]. Autrement dit, par exemple, de combiner les résultats de multiples applications de l'approche standard sur le même jeu de données de test, où les différentes applications le sont par des choix divers de paramètres. C'est encore plus vrai pour le cas où les résultats à combiner sont issus de l'ILB basée sur les embeddings[2].

Nos objectifs de recherche ont été donc, entre autres, de combiner les résultats provenant de plusieurs approches d'Induction de Lexiques Bilingues, notamment de l'approche standard ainsi que de plusieurs approches dépendant des embeddings. Ensuite, d'explorer en largeur et d'étudier l'impact de chaque approche sur le processus de combinaison, spécialement sur les mots rares[3].

7.2 Contributions

L'idée de combiner les résultats de plusieurs approches a été proposée par Philippe Langlais[4]. J'ai réalisé l'ensemble des expériences (implémentation et gestion), en ayant préalablement reproduit une nouvelle approche d'ILB basée sur les embeddings et appris à maîtriser un outil pour exécuter des reclassements de candidats. Les analyses subséquentes relèvent d'un effort conjoint entre mon directeur de recherche et moi-même, de même que la rédaction de l'article.

[1] Similairement, (Bouamor et al., 2013b) font voter les résultats de plusieurs mesures de similarité lors de la désambiguïsation de vecteurs de contexte.

[2] Qui n'est qu'une autre façon de modéliser l'approche standard (chapitre 5).

[3] En effet, cette approche pourrait amener une solution au problème des mots rares.

[4] Depuis le début de mon doctorat, aussi cette idée est déjà mentionnée dans mon rapport prédoc.

7.3 Impacts

Avec cet article, nous montrons la puissance (en termes d'améliorations des performances de l'ILB) de la reclassification et de la combinaison de multiples résultats[5] issus de la tâche d'induction de lexiques bilingues.

Suivant les travaux futurs annoncés dans l'article 2 (chapitre 6) sur la combinaison, nous avons pris nos résultats d'ILB en provenance de l'approche standard (Rapp, 1995) et de l'approche d'embeddings (Mikolov et al., 2013b) afin de les combiner. Cependant, nous avons décidé en premier lieu de reproduire une autre approche de production d'embeddings bilingues (Faruqui et Dyer, 2014) pour, d'une part, proposer une comparaison entre deux approches d'embeddings sur la tâche d'induction de lexiques bilingues et, d'une autre part, d'utiliser aussi les résultats de cette dernière dans la combinaison dans l'intention d'en évaluer des potentiels impacts.

Ainsi, nous reprenons une nouvelle fois notre méthodologie (article 1 et 2, chronologiquement) avec les mêmes données pour explorer la nouvelle approche et nous reportons les meilleures valeurs des métaparamètres sur chaque jeu de données. Ceci nous permet de comparer les résultats de plusieurs approches d'embeddings à la manière de (Levy et al., 2016, Upadhyay et al., 2016) mais sur la tâche d'ILB. Notre ensemble d'évaluation sur les mots rares[6] nous permet de démontrer (encore une fois) le biais des approches envers les mots fréquents.

Par la suite, nous révélons l'impact (inattendu) de la reclassification des candidats à la traduction des mots d'évaluation. Malgré le faible nombre de *features* utilisées, nous rapportons de belles améliorations de performances, notamment sur les mots rares, tout en sachant qu'il reste beaucoup de place pour plus d'ingénierie sur les *features*.

Ensuite, nous affichons les résultats de la combinaison proposée et nous constatons de nettes améliorations additionnelles à la reclassification, notamment (encore une fois) sur les mots rares. Nous ne sommes pas les premiers à démontrer l'impact de la com-

[5] Nous pourrions les appeler "Lexiques Bilingues" mais leur qualité ne nous le permet pas.
[6] Mots apparaissant moins de 25 fois dans Wikipédia (il y en a 6,8 millions (92%)).

binaison sur la tâche (Irvine et Callison-Burch, 2013, 2017) mais nous sommes les premiers à utiliser différents signaux (approches) d'une même source (contexte) à notre connaissance. Ces résultats nous laissent suggérer que les meilleures performances seront atteignables en combinant, à la fois, différents signaux d'une même source ainsi que différents signaux de plusieurs sources (Irvine et Callison-Burch, 2013, 2017). Autrement dit, que les machines apprendront mieux quand nous leur offrirons différents points de vue sur différentes sources d'informations[7].

Les analyses proposées à la fin de l'article permettent d'entrevoir un certain nombre d'améliorations possibles. Notamment, l'analyse de l'impact du classificateur sur la position des traductions de référence permet d'observer que les bonnes traductions se sont, en moyenne, toujours rapprochées de la première position dans la liste des candidats à la traduction. En outre, dans notre cas, les bonnes traductions sont en moyenne situées dans les 5 premières traductions retournées, ce qui laisse la possibilité de concevoir un système permettant à un utilisateur externe de nous aider à identifier les bonnes traductions, à l'aide d'une interface web par exemple. Aussi, l'analyse des mauvaises traductions situées en première place des candidats suggère que nous pouvons améliorer nos résultats en prétraitant un minimum nos données, avec par exemple un filtre des mots anglais ou en prenant le lemme de l'ensemble du vocabulaire.

Finalement, il serait très intéressant de reproduire ces expériences (et améliorations) sur d'autres paires de langues, spécialement avec des langues plus éloignées ainsi qu'avec des langues qui possèdent moins de ressources disponibles (Irvine et Callison-Burch, 2017). D'une certaine façon, cette critique nous rappelle que nous nous sommes écartés du sujet de recherche initial, sur le Web Sémantique, où la paire de langues choisie était totalement justifiée.

[7] Similairement aux humains.

7.4 Publication

[8]We investigate the reranking of the output of several distributional approaches on the Bilingual Lexicon Induction task. We show that reranking an *n*-best list produced by any of those approaches leads to very substantial improvements. We further demonstrate that combining several *n*-best lists by reranking is an effective way of further boosting performance.

7.4.1 Introduction

Identifying translations in bilingual material — the Bilingual Lexicon Induction (BLI) task — is a challenge that has long attracted the attention of many researchers. One of the earliest approach to BLI (Rapp, 1995) is based on the assumption that words that are translations of one another show similar co-occurrence patterns. Many variants have been investigated. For instance, some authors reported gains by considering syntactically motivated co-occurrences, either with the use of a parser (Yu et Tsujii, 2009) or by relying on simpler POS patterns (Otero, 2007). Extensions to multiword expressions have also been proposed (Daille et Morin, 2008). See (Sharoff et al., 2013) for an extensive overview.

Recently, vast efforts have been dedicated to identify translations thanks to so-called word embeddings. The seminal work of Mikolov et al. (2013b) shows that learning a mapping between word embeddings learnt monolingually by the popular `word2vec` toolkit (Mikolov et al., 2013a) is an efficient solution. Since then, many practitioners have studied the BLI task as a mean to evaluate continuous word-representations (Coulmance et al., 2015, Duong et al., 2016, Gouws et al., 2015, Luong et al., 2015, Vulic et Moens, 2015). Those approaches differ in the type of data they can process (monolingual data, word-aligned parallel data, parallel sentence pairs, comparable documents). Nevertheless, learning to map individually trained word embeddings remains an extremely efficient solution that performs well on several BLI benchmarks. Read (Levy et al.,

[8]Certaines adaptations stylistiques ont été réalisées sur l'article afin de l'adapter au format de la thèse.

2016, Upadhyay et al., 2016) for two recent comparisons of several of those techniques.

Reranking the output of several BLI approaches has been investigated, mostly for translating terms of the medical domain, where dedicated approaches can be designed to capture correspondences at the morphemic level (Delpech et al., 2012, Harastani et al., 2013, Kontonatsios et al., 2014). A similar idea (generating candidate translations, then filtering them by rescoring) has been proposed in (Baldwin et Tanaka, 2004) for translating noun-noun compounds in English and Japanese. Also, Irvine et Callison-Burch (2013) show that monolingual signals (orthographic, temporal, etc.) can be used to train a classifier to distinghish good translations from erroneous ones.

In this paper, we investigate the reranking of n-best lists of translations produced by two embedding approaches (Faruqui et Dyer, 2014, Mikolov et al., 2013b) as well as a plain distributional approach (Rapp, 1995). We tested a large number of variants of those approaches, for the English-to-French translation direction. The investigation of other language pairs and other BLI approaches is left as future work. To the best of our knowledge, this is the first time reranking embedding-based BLI approaches is reported.

We present our reranking framework in Section 7.4.2, our experimental protocol in Section 7.4.3, and report experiments in Section 7.4.4. We analyze our results in Section 7.4.5 and summarize our contributions in Section 7.4.6.

7.4.2 Reranking

The RankLib[9] library offers the implementation of 8 Learning to Rank Algorithms. We trained each one in a supervised way to optimize precision at rank 1. We used a 3-fold cross-validation procedure where in each fold, 700 terms of the test set were used for training, and the remaining 300 ones served as a test set. For a source term s and a candidate translation t, we compute 3 sets of straightforward and easily extensible features :

[9] `https://sourceforge.net/p/lemur/wiki/RankLib/`

Frequency features Four features recording the frequency of s (resp. t) in the source (resp. target) corpus, the difference between those two frequencies as well as their ratio.

String features Five features recording the length (counted in chars) of s and t, their difference, their ratio, and the edit-distance between the two. Edit-distance has been consistently reported to be a useful hint for matching terms.

Rank features For each n-best list considered, we compute 2 features : t's score in the list, as well as its rank. Whenever several n-best lists are reranked, we also add a feature that records the number of n-best lists t appears in as a candidate translation of s.

7.4.3 Experimental Protocol

7.4.3.1 Data sets

We trained each word's representation on the English and French versions of the Wikipedia dumps from June 2013. The English vocabulary contains 7.3M words forms (1.2G tokens) while the French vocabulary contains 3.6M forms (330M tokens).

One research avenue we explored in this study consisted in assessing the impact of word's frequency on the BLI performance. For this, we gathered two reference lists of words and their translations. One list, named $\text{Rare}_{\leq 25}$, is populated with English words occurring 25 times or less in Wikipedia (English edition). There are 6.8M (92%) such words. Thus, this test set is more representative of a real-life setting. The other list, named $\text{Freq}_{>25}$ contains words whose frequencies exceed 25. Both lists contain 1 000 words that we randomly picked from an in-house bilingual lexicon. Each one of those words had to have at least one of its approved translations belong to the French Wikipedia vocabulary.

Most recent studies on BLI focus on translating very frequent words, in keeping with the protocol described in (Mikolov et al., 2013b), which basically consists in transla-

ting 1 000 terms from the WMT11 dataset. Those term's rank are between 5000 and 6000 when the terms are sorted in decreasing order of frequency (the most frequent 5k words are put aside in order to train the projection). We reproduced this setting for comparison purposes (list Euro_{5-6k}). Only 87.3% of the resulting pairs have both their source term in the English Wikipedia vocabulary and their approved translation in the French counterpart. For the sake of fairness, we report results of the embedding-based approaches on those terms only.

The main characteristics of our test sets are presented in Table 7.I. As an illustration of the difficulty of each test set, we measure the accuracy (TOP@1) of a baseline that ranks candidates in increasing order of edit-distance with the source term. For some reasons, the Wikipedia test sets are easier than Euro_{5-6k} for such an approach,.

	Frequency				
	min	max	avg	*Cov* (%)	TOP@1
$\text{Freq}_{>25}$	27	19.4k	2.8k	100.0	19.3
$\text{Rare}_{\leq 25}$	1	25	10	100.0	17.6
Euro_{5-6k}	1	2.6M	33.6k	87.3	8.0

Tableau 7.I: Characteristics of our test sets. *Cov.* is the percentage of source terms for which the reference translation is part of the French edition of Wikipedia.

7.4.3.2 Metrics

Each approach (see Section 7.4.4) has been configured to produce a ranked list of (at most) 100 candidate translations (in French). We measure their performance with accuracy at rank 1, 5, and 20; where accuracy at rank i (TOP@i or T@i) is computed as the percentage of test words for which a reference translation is identified in the first i candidates proposed.

	Individual			1-Reranked				*n*-Reranked		
	T@1	T@5	T@20	T@1	T@5	T@20		T@1	T@5	T@20
$\text{Freq}_{>25}$										oracle : 69.3
Rapp	20.0	33.0	43.0	$36.3^{2.5}$	$48.8^{1.9}$	$53.8^{1.9}$	base	$34.3^{1.9}$	$47.6^{1.4}$	$58.8^{0.8}$
Miko	17.0	32.6	41.6	$38.1^{1.9}$	$49.0^{1.5}$	$54.3^{1.3}$	R+M	$43.3^{2.9}$	$58.4^{1.4}$	$62.4^{3.1}$
Faru	13.3	26.0	33.3	$34.3^{1.5}$	$44.0^{2.6}$	$47.9^{2.1}$	R+M+F	$\mathbf{45.6}^{2.2}$	$\mathbf{59.6}^{1.1}$	$\mathbf{64.0}^{1.8}$
$\text{Rare}_{\leq 25}$										oracle : 28.6
Rapp	2.6	4.3	7.3	$8.6^{1.2}$	$9.4^{0.8}$	$10.2^{1.0}$	base	$10.7^{0.6}$	$15.9^{1.2}$	$21.8^{0.7}$
Miko	1.6	4.6	10.6	$16.6^{2.2}$	$19.0^{1.5}$	$20.1^{1.4}$	R+M	$18.9^{2.01}$	$22.0^{1.3}$	$23.6^{2.2}$
Faru	1.6	2.6	5.0	$7.9^{2.2}$	$8.7^{2.5}$	$8.9^{2.7}$	R+M+F	$\mathbf{21.3}^{1.86}$	$\mathbf{24.4}^{1.7}$	$\mathbf{25.7}^{1.9}$
Euro_{5-6k}										oracle : 84.4
Rapp	16.6	31.8	41.2	$34.6^{5.7}$	$48.6^{1.2}$	$51.9^{1.2}$	base	$33.6^{1.2}$	$59.3^{1.4}$	$71.7^{2.5}$
Miko	42.0	59.0	67.8	$47.0^{2.3}$	$68.1^{2.7}$	$73.0^{1.7}$	R+M	$\mathbf{49.5}^{3.7}$	$\mathbf{68.7}^{1.5}$	$76.1^{1.0}$
Faru	30.6	47.7	59.8	$41.2^{3.9}$	$58.0^{3.5}$	$66.0^{3.5}$	R+M+F	$47.6^{2.3}$	$68.5^{2.0}$	$\mathbf{76.2}^{1.2}$

Tableau 7.II: Performance of each approach (left-hand side column) and their reranking (middle column), as well as the best reranking of 2 and 3 native *n*-best lists (right-hand side column). The reranked results are averaged over a 3-fold cross-validation procedure, the superscript indicates the standard deviation. `oracle` picks the reference translation among the 3 individual *n*-best lists.

7.4.4 Experiments

7.4.4.1 Individual Approaches

We ran variants of an in-house implementation of (Rapp, 1995) exploring a number of meta-parameters (window size, association measure, seed lexicon, etc.). We refer to this approach as `Rapp` hereafter. We studied a similar number of variants of (Mikolov et al., 2013b) — hereafter named `Miko`— training monolingual embeddings with `Word2Vec` (Mikolov et al., 2013b), varying among other things the model's architecture (skip-gram versus continuous bag-of-words), the optimization algorithm (negative sampling (5 or 10 samples) versus hierarchical softmax), and the context window size (6, 10, 20, 30). The largest embedding dimension for which we managed to train a model is 200 for the *cbow* architecture, and 250 for the *skg* architecture. We learnt the projection matrix with the implementation described in (Dinu et Baroni, 2014). We reproduced the approach of Faruqui et Dyer (2014) — henceforth `Faru` — thanks to the toolkit provided by

the authors. We kept the embeddings that yielded the best performance for the Miko approach, and ran several configurations, varying the bilingual lexicon used, and tuning the *ratio* parameter over the values 0.5, 0.8 and 1.0.

The best performance for the variants of each strategy we tested is reported in the first column of Table 7.II. On $\text{Freq}_{>25}$, the Rapp approach delivers the best performance at rank 1, slightly outperforming the edit-distance baseline (TOP@1 of 19.3). The drop in performance of all approaches on $\text{Rare}_{\leq 25}$ is striking : the best one could only identify the translation of 2.6% of the test terms at rank 1. This clearly demonstrates the bias of the approaches tested in favor of frequent words. On the Euro_{5-6k} test set, the two embedding approaches are rather good (TOP@1 of Miko reaches 42%) and clearly outperform Rapp. This suggests that embeddings are very apt at capturing information for very frequent terms (test terms on Euro_{5-6k} appear roughly 10 times more in Wikipedia than those in $\text{Freq}_{>25}$). Our results are in line with those reported in (Mikolov et al., 2013b). We were more surprised by the lower performance yielded by Faru. It should be noted however that this model's gains, as reported in (Faruqui et Dyer, 2014), have been measured on monolingual tasks. The authors also built on top of embeddings learnt with the *skg* architecture, while we found it to be less accurate for our task.

7.4.4.2 Reranking Individual Approaches

The middle column in Table 7.II reports the reranking of the *n*-best list produced by each individual approach. During calibration experiments, we found better rescoring performances with the *Random Forest* algorithm. We report results for this algorithm only.[10] We observe that reranking is highly beneficial to each approach. For instance, when reranking the *n*-best list produced by Miko, TOP@1 nearly doubles on $\text{Freq}_{>25}$, and is 10 times higher on $\text{Rare}_{\leq 25}$. It is also noteworthy that on $\text{Freq}_{>25}$ all approaches, once reranked perform equally overall (TOP@1 between 34 and 38) — Miko enjoying a slight advantage here — far better than the edit distance baseline.

[10]Results were close with *LambaMart* (2 TOP@1 points lost) and *Mart* (1.5 TOP@1 points lost).

7.4.4.3 Combining by Reranking

We conducted experiments aiming at combining several *n*-best lists with reranking. For comparison purposes, we implemented a naive combination approach that ranks a candidate translation higher if it is proposed in more *n*-best lists. Tied candidates are further sorted in increasing order of edit distance. The results of a few combinations are reported in the right column of Table 7.II.

Combining the *n*-best lists produced by the 3 native approaches leads to the best performance overall, except on Euro_{5-6k} where not considering `Faru` leads to slight improvements in TOP@1 and TOP@5 metrics. This indicates that the reranker puts good use of multiple models. The gains over each reranked approach are impressive on $\text{Freq}_{>25}$ (increase from 38.1% to 45.6%) and $\text{Rare}_{\leq 25}$ (increase fom 16.6 to 21.3) and minor on Euro_{5-6k}(from 47.0% to 47.6%). We also observe that TOP@20 obtained by the reranker is not very far from the oracle performance.

7.4.5 Analysis

In this section, we analyze the characteristics of the reranker we used to combine the 3 aforementioned approaches.

7.4.5.1 Training Size

Figure 7.1 shows the impact of the quantity of material used for learning the reranker, varying from 100 word pairs to 700. In this experiment, we always use the same 300 test words per test set. Increasing the training material increases performance for all test sets,[11] but even a small training set is enough to improve upon native approaches. In particular, using 200 training instances already yields a TOP@1 of 36.6 on $\text{Freq}_{>25}$, while the best native tops at 20.

[11] On $\text{Rare}_{\leq 25}$ however, the gains are very small.

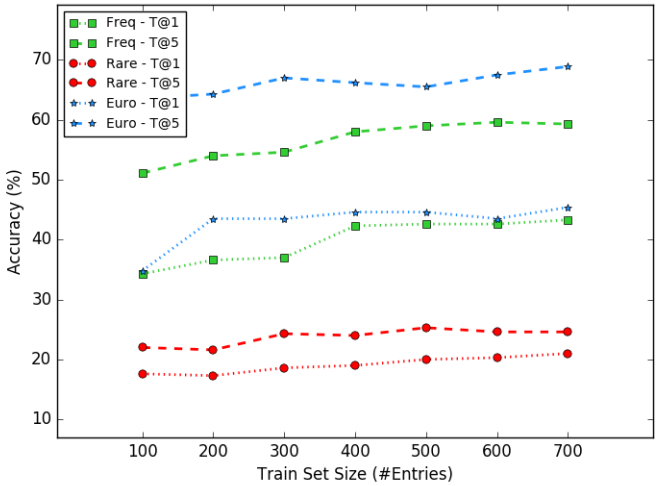

Figure 7.1 : Influence of the training size (number of examples) on the performance of the reranker on $\text{Freq}_{>25}$, $\text{Rare}_{\leq 25}$ and Euro_{5-6k}.

7.4.5.2 Feature Selection

$\text{Freq}_{>25}$	*Sing.*	*Cumul*		$\text{Rare}_{\leq 25}$	*Sing.*	*Cumul*		Euro_{5-6k}	*Sing.*	*Cumul*	
feat.	T@1	T@1	T@100	feat	T@1	T@1	T@100	feat	T@1	T@1	T@100
Rank	33.0	33.0	66.0	String	16.6	16.6	26.6	Rank	46.2	46.2	81.3
+String	32.0	42.0	67.0	+Rank	6.6	20.3	26.3	+String	18.9	43.9	80.3
+ Freq	0.3	43.0	67.3	+Freq	0.0	20.3	26.6	+Freq	2.2	48.8	82.5

Tableau 7.III: Influence of the features used to train the reranker when combining `Rapp`, `Miko`, and `Faru`. Performances are averaged over a 3-fold cross-validation procedure. Each fold uses 700 pairs for training and 300 for testing. *Sing.* indicates the performance of individual features, while *Cumulative* indicates their cumulative performance. Features are listed in decreasing order of gains.

Table 7.IIIshows the influence of the features used for training the reranker. On frequent terms ($\text{Freq}_{>25}$ and Euro_{5-6k}) the rank-based features are the most useful ones, followed by the string-based features. The frequency-based features only help marginally. On $\text{Rare}_{\leq 25}$, the string-based features are more useful. The performance of the reranker using only those features (16.6TOP@1) is close to that of the baseline edit distance approach (17.6TOP@1). Adding the rank-based features increases the performance slightly

(20.3TOP@1).

7.4.5.3 Ranker Analysis

	$\text{Freq}_{>25}$	$\text{Rare}_{\leq 25}$	Euro_{5-6k}
Rapp	12.7	19.6	16.2
Miko	16.3	30.0	7.5
Faru	20.4	35.5	11.3
list-oracle	12.3	9.1	7.1
reranker	5.6	4.0	4.9

Tableau 7.IV: Average rank of the reference translation. Terms for which the reference translation is not found in the first 100 positions are discarded.

With a few exceptions, we observe that whenever at least 2 native approaches propose the reference translation first, the reranker keeps at the first position as well. When only one native approach is accurate at position 1, the results differ from one test set to another. It is only occasionally that the reranker will prefer the reference translation when none of the native approaches does. On $\text{Freq}_{>25}$, this happens 130 times out of 300 cases, but on Euro_{5-6k}, it happened only 4 times over 132 cases, which is disappointing. Still, the average position of the reference translation in the reranker's output is clearly improving for all test sets, as shown in Table 7.IV. The average number of positions gained by reranking is rather high, and outdoes an oracle that picks the n-best list in which the reference translation is best positioned. We note that the average rank of the Rapp approach is lower than that of the embedding approaches, for both Wikipedia test sets.

7.4.5.4 Error Analysis

We manually inspected the first candidate produced by our best reranker (the one combining the 3 native approaches) for the first 100 test forms for which the candidate translation differs from the reference one. We encountered the following representative

	$\text{Freq}_{>25}$	$\text{Rare}_{\leq 25}$	Euro_{5-6k}
MORPHO	18	3	26
RELATED	16	4	23
synonyms	*15*	*1*	*19*
antonyms	*1*	*2*	*2*
hyponym			*1*
cohyponym		*1*	*1*
POLYSEMY	4	0	5
LOOSLY	14	15	20
ENGLISH	21	6	7
JUNK	27	72	19

Tableau 7.V: Annotation of 100 translations produced (at rank 1) for each test set by the reranked output of the 3 native approaches.

cases : morphological variants of the reference translation (*e.g.* trompeur / trompeuse, *litt.* misleading) — MORPHO ; directly related translation, such as synonyms, antonyms, and cohyponyms — RELATED ; loosely related to the reference (*e.g.* gunman / poignardé, *litt.* stabbed) — LOOSLY ; English words – ENGLISH ; translations that apparently have nothing to do with the source term (*e.g.* judged /méritant, *litt.* worthy) – JUNK ; and translations that correspond to another sense of a polysemic term (*e.g.* grizzly / grizzli, while the reference translation is grisonnant, *litt.* gray haired) – POLYSEMY. The counts of each class for each test set are reported in Table 7.V.

We observe that the percentage of JUNK errors is much higher on $\text{Rare}_{\leq 25}$, yet another illustration of the bias the approaches we tested have in favor of frequent terms. If we consider synonyms, morphological variants as well as polysemic cases to be correct, then the percentages of test forms that are redeemed reach 37% for $\text{Freq}_{>25}$ and 50% for Euro_{5-6k} of test forms that were counted wrong are indeed acceptable translations. On $\text{Rare}_{\leq 25}$ however, this percentage is much lower (4%).

7.4.6 Discussion

We have studied the reranking of three approaches to BLI. We reported significant improvements for all approaches, on all test sets. We also show that combining several n-best lists by reranking is a simple yet effective solution leading to even better performance. The gains were obtained by a random forest model learnt on a set of straightforward features, which leaves ample room for better feature engineering. While extra data must be used to train the reranker, we show that as few as 200 training examples often suffice to provide an appreciable boost in performance. As a future work we want to investigate whether similar gains can be obtained for other language pairs.

CHAPITRE 8

CONCLUSION

Cette thèse réunit les contributions accomplies durant mon doctorat. Ces contributions tentent toutes d'améliorer, d'une manière ou d'une autre, les approches qui constituent ou actualisent des ressources de type multilingues (bilingues dans notre cas). Ce sont des ressources nécessaires et cependant jamais suffisantes pour améliorer les performances de la traduction automatique. Plus exactement, nous rapportons notre travail sur la tâche d'Alignement de Documents Bilingues ainsi que nos trois articles sur la tâche d'Induction de Lexiques Bilingues.

Notre première contribution relate notre participation à la tâche partagée d'alignement de documents bilingues de la première conférence sur la traduction automatique. Nous avons soumis une application simple et légère construite sur un moteur de recherche libre très répandu[1] implémentant la tâche d'alignement à travers de multiples sites web bilingues anglais-français. Cette implémentation nous a permis de proposer une exploration en largeur des méta-paramètres du modèle nommé "sac de mots" utilisé pour représenter les documents, autant par notre application que par la majorité des participants. Nous avons évalué l'impact de plus de mille configurations durant cette exploration ainsi qu'identifié une approche modélisant les documents par "sac de mots d'URL" qui s'est révélée être aussi légère qu'efficace. Malgré tout, notre système s'est retrouvé dans une position jugée décevante en bas de classement. Sur la base des retours des organisateurs de la tâche, il ressort que les meilleurs résultats ont été obtenus par les participants ayant utilisé des ressources de traduction issues de moteurs de traduction automatique, et pas par de simples lexiques bilingues. Ce constat vient confirmer une fois de plus la force implicite des modèles basés sur les statistiques : plus des données sont disponibles, meilleurs sont les résultats.

[1] Lucene - `http://lucene.apache.org/core/`

Nos trois autres contributions portent sur la tâche d'induction de lexiques bilingues et assurent une continuité du contexte de recherche entre elles. En effet, les deux premières publications proposent chacune une analyse complète d'une application d'ILB, chacune d'elle utilisant un modèle de sémantique distributionnelle différent pour représenter les mots. Le premier modèle, appelé "vecteur de contexte", a été présenté avec l'idée d'adapter la procédure d'alignement des mots en corpus parallèles pour les corpus comparables. Le deuxième modèle, plus récent, est basé sur la prédiction du contexte des mots en utilisant des réseaux de neurones et il est appelé *embeddings*. Le dernier article combine et reclasse les candidats à la traduction résultants de multiples applications d'induction de lexiques bilingues, dont celles discutées avec les deux précédents articles, afin d'améliorer les performances de la tâche de manière globale.

L'ensemble de ces travaux ont été réalisés avec un objectif de "conditions réelles", c'est-à-dire en poussant les approches à leurs limites, autant computationnelles que vis-à-vis (des caractéristiques) des données sur lesquelles les approches devraient agir. Ainsi, les applications de l'induction de lexiques bilingues sont testées sans réduire (de façon arbitraire) la taille du vocabulaire cible dans lequel les traductions candidates sont cherchées[2]. De plus, l'ensemble des expériences sont aussi réalisées avec des compilations de mots rares, ce qui permet à chaque fois de mettre en évidence le biais touchant les mots fréquents ainsi que de mieux comprendre les approches.

Plus spécifiquement, notre deuxième contribution propose une meilleure compréhension de l'approche dite "standard" en la plaçant dans des conditions réelles d'application, via le contexte du Web Sémantique. Nous la comparons à une approche récente à l'état de l'art et nous montrons que des valeurs soigneusement choisies pour les méta-paramètres font d'elle l'approche de référence si une étape de calibration est réalisée. Notre troisième contribution tente de fournir une meilleure compréhension des récents embeddings, tant au niveau de leur constitution que de leur(s) impact(s) et utilisés dans le cadre de l'induction de lexiques bilingues. Pour ce faire, nous confrontons l'approche

[2]C'est actuellement un choix qui influence grandement les performances de la tâche, mais que beaucoup ignorent lors de leurs expériences.

"standard" à une approche basée sur les embeddings et nous montrons que des performances similaires sont obtenues par les deux approches. Cependant, comme mentionné précédemment, l'approche standard requiert une étape de calibration qui est réalisée implicitement par l'approche à base d'embeddings. Notre quatrième contribution met en avant l'impact des techniques de reclassement (notamment, supervisées) ainsi que celui de la combinaison à partir de plusieurs signaux d'induction de lexiques bilingues. Nous révélons l'impact (inattendu) de la reclassification des candidats à la traduction avec peu de *features* et ensuite, nous affichons de nettes améliorations grâce à la combinaison, et ce, surtout sur les mots rares.

Malgré la profondeur des expérimentations réalisées dans les travaux de cette thèse, il reste de la place pour de nombreuses perspectives de recherches futures, notamment sur des tâches connexes à celles discutées.

Ainsi, notre application d'alignement de documents bilingues peut être améliorée, principalement sur deux points. Le premier point propose d'inclure la possibilité d'utiliser des données de textes traduits afin de réaliser le pont de traduction entre la langue source et cible des sites web considérés. Concernant les langues traitées, il serait intéressant d'expérimenter l'application sur d'autres paires de langues afin d'en évaluer l'impact sur notre modélisation de référence. Le deuxième point suggère qu'il faut revoir nos étapes de prétraitements et de post-traitements afin d'être plus flexible sur la taille des sites web visés.

Nos analyses méticuleuses de la tâche d'induction de lexiques bilingues implémentée à l'aide de deux métamodèles différents de représentations de mots prouvent l'inaptitude des approches et des modèles à gérer les mots non fréquents. Pour la tâche d'ILB, des solutions à ce problème sont possibles sur deux points. Le premier concerne les modèles de représentation des mots. Il nous faut proposer des modèles qui tiennent compte de la faible occurrence des mots (notamment, celle des mots rares) et qui tentent de l'incorporer dans les modèles de représentation. Deuxièmement, nous suggérons qu'il est possible d'améliorer les performances de la tâche en incorporant des prétraitements ainsi que des post-traitements à la phase d'alignement des vecteurs/embeddings. Un

exemple de prétraitement serait d'échantillonner le vocabulaire cible en choisissant un sous-ensemble de mots qui auraient une plus grande probabilité d'être la traduction du mot source considéré. La combinaison et le reclassement des candidats à la traduction, que notre dernière publication amorce, sont déjà une forme de post-traitement à l'induction de lexiques bilingues.

Concernant ce post-traitement, nos résultats obtenus lors de la dernière publication sont encourageants et peuvent être améliorés de différentes façons. Notamment, il reste de la place pour de larges explorations à mener avec plus de *features* pour le reclassement et l'exploitation de plus de sources de modélisations ainsi que de multiples signaux pour la combinaison. Surtout, nous envisageons de tester cette approche sur d'autres paires de langues afin de potentiellement transférer l'apprentissage à des langues moins dotées pour lesquelles le besoin en ressources multilingues est encore plus nécessaire pour la traduction automatique.

BIBLIOGRAPHIE

Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer et Noah A. Smith. Massively multilingual word embeddings. *arXiv preprint arXiv :1602.01925*, 2016.

Daniel Andrade, Tetsuya Nasukawa et Jun'ichi Tsujii. Robust measurement and comparison of context similarity for finding translation pairs. Dans *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 19–27. Association for Computational Linguistics, 2010.

Mikel Artetxe, Gorka Labaka et Eneko Agirre. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. 2016.

Timothy Baldwin et Takaaki Tanaka. Translation by machine of complex nominals : Getting it right. Dans *Proceedings of the Workshop on Multiword Expressions : Integrating Processing*, pages 24–31, 2004.

Caroline Barrière et Pierre Isabelle. Searching parallel corpora for contextually equivalent terms. Dans *Proceedings of the 15th Conference of the European Association for Machine Translation*, pages 105–112, 2011.

Christian Bizer, Tom Heath et Tim Berners-Lee. Linked data-the story so far. *International journal on semantic web and information systems*, 5(3):1–22, 2009.

Dhouha Bouamor. *Constitution de ressources linguistiques multilingues à partir de corpus de textes parallèles et comparables.* PhD thesis, Université Paris Sud - Paris XI, 2014.

Dhouha Bouamor, Adrian Popescu, Nasredine Semmar et Pierre Zweigenbaum. Building specialized bilingual lexicons using large scale background knowledge. Dans *EMNLP*, pages 479–489, 2013a.

Dhouha Bouamor, Nasredine Semmar et Pierre Zweigenbaum. Context vector disambiguation for bilingual lexicon extraction from comparable corpora. Dans *ACL*, pages 759–764, 2013b.

Julien Bourdaillet, Stéphane Huet, Philippe Langlais et Guy Lapalme. TransSearch : from a bilingual concordancer to a translation finder. *Machine Translation*, 24(3-4): 241–271, 2010.

Peter F Brown, John Cocke, Stephen A Della Pietra, Vincent J Della Pietra, Fredrick Jelinek, John D Lafferty, Robert L Mercer et Paul S Roossin. A statistical approach to machine translation. *Computational linguistics*, 16(2):79–85, 1990.

Christian Buck et Philipp Koehn. Findings of the WMT 2016 Bilingual Document Alignment Shared Task. Dans *Proceedings of the First Conference on Machine Translation, Berlin, Germany. Association for Computational Linguistics*, 2016.

John A. Bullinaria et Joseph P. Levy. Extracting semantic representations from word co-occurrence statistics : A computational study. *Behavior Research Methods*, 39:510–526, 2007.

CafeMultilingue. Plurilinguisme vs Multilinguisme. `http://cafemultilingue.blogspot.ca/2015/03/plurilinguisme-vs-multilinguisme.html`, 2016. [En Ligne; Accéder le 28 Octobre 2016].

Chris Callison-Burch, Philipp Koehn, Christof Monz, Kay Peterson, Mark Przybocki et Omar F. Zaidan. Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation. Dans *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, WMT '10, pages 17–53, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. ISBN 978-1-932432-71-8. URL `http://dl.acm.org/citation.cfm?id=1868850.1868853`.

Bruno Cartoni. Lexical morphology in machine translation : A feasibility study. Dans *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, pages 130–138, 2009.

Yun-Chuang Chiao, Jean-David Sta et Pierre Zweigenbaum. A novel approach to improve word translations extraction from non-parallel, comparable corpora. Dans *Proceedings of the International Joint Conference on Natural Language Processing, Hainan, China. AFNLP*, 2004.

Yun-Chuang Chiao et Pierre Zweigenbaum. Looking for candidate translational equivalents in specialized, comparable corpora. Dans *Proceedings of the 19th international conference on Computational linguistics-Volume 2*, pages 1–5. Association for Computational Linguistics, 2002.

Yun-Chuang Chiao et Pierre Zweigenbaum. Aligning Words in French-English Non-Parallel Medical Texts : Effect of Term Frequency Distributions. Dans *Medinfo 2004 : Proceedings of the 11th World Congress on Medical Informatics*, page 23. Ios Pr Inc, 2004.

Chenhui Chu, Toshiaki Nakazawa et Sadao Kurohashi. Iterative bilingual lexicon extraction from comparable corpora with topical and contextual knowledge. Dans *Computational Linguistics and Intelligent Text Processing*, pages 296–309. 2014.

COE. COE : Accueil. `http://www.coe.int`, 2016. [En Ligne ; Accéder le 28 Octobre 2016].

Jocelyn Coulmance, Jean-Marc Marty, Guillaume Wenzek et Amine Benhalloum. Transgram, Fast Cross-lingual Word-embeddings. Dans *Proceedings of EMNLP*, pages 1109–1113, 2015.

CSLF. La Socialisation Languagière comme Processus Dynamique. `http://www.cslf.gouv.qc.ca/publications/pubf329/f329.pdf`, 2016. [En Ligne ; Accéder le 28 Octobre 2016].

Béatrice Daille et Emmanuel Morin. Effective Compositional Model for Lexical Alignment. Dans *Proceedings of the 3rd IJCNLP*, pages 95–102, 2008.

Aswarth Abhilash Dara et Yiu-Chang Lin. Yoda system for wmt16 shared task : Bilingual document alignment. Dans *Proceedings of the First Conference on Machine Translation, Berlin, Germany. Association for Computational Linguistics*, 2016.

Estelle Delpech, Béatrice Daille, Emmanuel Morin et Claire Lemaire. Extraction of domain-specific bilingual lexicon from comparable corpora : compositional translation and ranking. *Proceedings of COLING*, 2012.

Mona Diab et Steve Finch. A statistical word-level translation model for comparable corpora. Dans *Content-Based Multimedia Information Access-Volume 2*, pages 1500–1508. LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE, 2000.

Georgiana Dinu et Marco Baroni. Improving zero-shot learning by mitigating the hubness problem. *CoRR*, 2014.

DPL. Guide pour l'élaboration des politiques linguistiques éducatives en europe. Rapport technique, Conseil de l'Europe, 2007. Version de synthèse.

Ted Dunning. Accurate Methods for the Statistics of Surprise and Coincidence. *Comput. Linguist.*, 19(1):61–74, 1993.

Long Duong, Hiroshi Kanayama, Tengfei Ma, Steven Bird et Trevor Cohn. Learning Crosslingual Word Embeddings without Bilingual Corpora. *arXiv preprint arXiv :1606.09403*, 2016.

Jessica Enright et Grzegorz Kondrak. A fast method for parallel document identification. Dans *Human Language Technologies 2007 : The Conference of the North American Chapter of the Association for Computational Linguistics ; Companion Volume, Short Papers*, pages 29–32, 2007.

Stefan Evert. *The statistics of word cooccurrences.* Thèse de doctorat, Dissertation, Stuttgart University, 2005.

Manaal Faruqui et Chris Dyer. Improving Vector Space Word Representations Using Multilingual Correlation. Dans *Proceedings of EACL*, 2014.

Manaal Faruqui, Yulia Tsvetkov, Pushpendre Rastogi et Chris Dyer. Problems with evaluation of word embeddings using word similarity tasks. *arXiv preprint arXiv :1605.02276*, 2016.

Pascale Fung. Compiling bilingual lexicon entries from a non-parallel english-chinese corpus. Dans *Third Workshop on Very Large Corpora*, pages 173–183, 1995.

Pascale Fung. A Statistical View on Bilingual Lexicon Extraction : From Parallel Corpora to Non-parallel Corpora. Dans *Proceedings of the 3rd Conference of the AMTA on Machine Translation and the Information Soup*, pages 1–17, 1998.

Evgeniy Gabrilovich et Shaul Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. IJCAI'07, pages 1606–1611, 2007.

William A. Gale et Kenneth Ward Church. Identifying Word Correspondences in Parallel Texts. Dans *HLT*, pages 152–157, 1991.

Pablo Gamallo et Marcos Garcia. Extraction of bilingual cognates from wikipedia. Dans *Proceedings of the 10th International Conference on Computational Processing of the Portuguese Language*, PROPOR'12, pages 63–72, 2012.

Matt Gardner, Kejun Huang, Evangelos Papalexakis, Xiao Fu, Partha Talukdar, Christos Faloutsos, Nicholas Sidiropoulos et Tom Mitchell. Translation invariant word embeddings.

Nikesh Garera, Chris Callison-Burch et David Yarowsky. Improving translation lexicon induction from monolingual corpora via dependency contexts and part-of-speech equivalences. Dans *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, pages 129–137. Association for Computational Linguistics, 2009.

Nikesh Garera et David Yarowsky. Translating compounds by learning component gloss translation models via multiple languages. Dans *IJCNLP*, pages 403–410, 2008.

Eric Gaussier, J-M Renders, Irina Matveeva, Cyril Goutte et Hervé Déjean. A geometric view on bilingual lexicon extraction from comparable corpora. Dans *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 526. Association for Computational Linguistics, 2004.

Gavagai. A brief history of word embeddings (and some clarifications). `https://www.gavagai.se/blog/2015/09/30/a-brief-history-of-word-embeddings/`, 2015. [En Ligne; Accédé le 25 Janvier 2017].

Luıs Gomes et G. Pereira Lopes. First steps towards coverage-based document alignment. Dans *Proceedings of the First Conference on Machine Translation, Berlin, Germany. Association for Computational Linguistics*, 2016.

Stephan Gouws, Yoshua Bengio et Greg Corrado. BilBOWA : Fast Bilingual Distributed Representations without Word Alignments. Dans *Proceedings of the 32nd ICML*, pages 748–756, 2015.

Jorge Gracia, Elena Montiel-Ponsoda, Philipp Cimiano, Asunción Gómez-Pérez, Paul Buitelaar et John McCrae. Challenges for the multilingual web of data. *Web Semantics : Science, Services and Agents on the World Wide Web*, 11:63–71, 2012.

Asunción Gómez-Pérez, Daniel Vila-Suero, Elena Montiel-Ponsoda, Jorge Gracia et Guadalupe Aguado-de Cea. Guidelines for multilingual linked data. WIMS '13, pages 3 :1–3 :12, 2013.

Aria Haghighi, Percy Liang, Taylor Berg-Kirkpatrick et Dan Klein. Learning bilingual lexicons from monolingual corpora. Dans *ACL*, volume 2008, pages 771–779, 2008.

Rima Harastani, Béatrice Daille et Emmanuel Morin. Ranking Translation Candidates Acquired from Comparable Corpora. Dans *Proceedings of the 6th IJCNL*, pages 401–409, 2013.

Zellig S. Harris. Distributional structure. *Word*, pages 146–162, 1954.

Samer Hassan et Rada Mihalcea. Cross-lingual semantic relatedness using encyclopedic knowledge. Dans *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing : Volume 3-Volume 3*, pages 1192–1201. Association for Computational Linguistics, 2009.

Amir Hazem, Morin Emmanuel et others. Word co-occurrence counts prediction for bilingual terminology extraction from comparable corpora. *IJCNLP 2013.*, 2013a.

Amir Hazem et Emmanuel Morin. Adaptive Dictionary for Bilingual Lexicon Extraction from Comparable Corpora. Dans *LREC*, pages 288–292, 2012.

Amir Hazem et Emmanuel Morin. Improving bilingual lexicon extraction from comparable corpora using window-based and syntax-based models. Dans *Proceedings of the 15th International Conference on Computational Linguistics and Intelligent Text Processing - Volume 8404*, CICLing 2014, pages 310–323, 2014.

Amir Hazem, Emmanuel Morin et Laboratore d'Informatique de Nantes-Atlantique. A comparison of smoothing techniques for bilingual lexicon extraction from comparable corpora. Dans *Proceedings of the Workshop on Building and Using Comparable Corpora (BUCC'13)*, pages 24–33, 2013b.

Eduard Hovy, Roberto Navigli et Simone Paolo Ponzetto. Collaboratively built semi-structured content and artificial intelligence : The story so far. *Artificial Intelligence*, 194:2–27, 2013.

Ann Irvine et Chris Callison-Burch. Supervised Bilingual Lexicon Induction with Multiple Monolingual Signals. Dans *Proceedings of NAACL-HLT*, pages 518–523, 2013.

Ann Irvine et Chris Callison-Burch. A comprehensive analysis of bilingual lexicon induction. *Comput. Linguist.*, 43(2):273–310, juin 2017.

Azniah Ismail et Suresh Manandhar. Bilingual lexicon extraction from comparable corpora using in-domain terms. Dans *Proceedings of the 23rd International Conference*

on Computational Linguistics : Posters, pages 481–489. Association for Computational Linguistics, 2010.

Laurent Jakubina et Phillippe Langlais. Projective methods for mining missing translations in dbpedia. Dans *Proceedings of the Eighth Workshop on Building and Using Comparable Corpora*, pages 23–31, Beijing, China, July 2015. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W15-3404.

Laurent Jakubina et Phillippe Langlais. Bad luc@wmt 2016 : a bilingual document alignment platform based on lucene. Dans *Proceedings of the First Conference on Machine Translation*, pages 703–709, Berlin, Germany, August 2016a. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W/W16/W16-2370.

Laurent Jakubina et Phillippe Langlais. A comparison of methods for identifying the translation of words in a comparable corpus : Recipes and limits. *Computación y Sistemas*, 20(3):449–458, 2016b. URL http://www.cys.cic.ipn.mx/ojs/index.php/CyS/article/viewFile/2465/2169.

Laurent Jakubina et Phillippe Langlais. Reranking translation candidates produced by several bilingual word similarity sources. Dans *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics : Volume 2, Short Papers*, pages 605–611, Valencia, Spain, April 2017. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/E17-2096.

Robert Isele Jens Lehmann. DBpedia - a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web Journal*, 2014.

Georgios Kontonatsios, Ioannis Korkontzelos, Jun'ichi Tsujii et Sophia Ananiadou. Combining String and Context Similarity for Bilingual Term Alignment from Comparable Corpora. Dans *Proceedings of EMNLP*, pages 1701–1712, 2014.

Kriste Krstovski et David A. Smith. A minimally supervised approach for detecting and ranking document translation pairs. Dans *Proceedings of the Sixth Workshop on Statis-*

tical Machine Translation, pages 207–216. Association for Computational Linguistics, 2011.

H. W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97, 1955.

Philippe Langlais, Fabrizio Gotti et Alexandre Patry. De la chambre des communes à la chambre d'isolement : adaptabilité d'un système de traduction basé sur les segments de phrases. *Généreux—Corpus de weblogs annotés pour leur humeur T. van de Cruys—An Overview of Noun Clustering in Dutch*, 2006.

Audrey Laroche et Philippe Langlais. Revisiting context-based projection methods for term-translation spotting in comparable corpora. COLING '10, pages 617–625, 2010.

Larousse. Définition : Plurilingue. `http://www.larousse.fr/dictionnaires/francais/plurilingue/61801`, 2016. [En Ligne; Accéder le 28 Octobre 2016].

Omer Levy, Yoav Goldberg et Ido Dagan. Improving distributional similarity with lessons learned from word embeddings. *Transactions of ACL*, pages 211–225, 2015.

Omer Levy, Anders Søgaard et Yoav Goldberg. Reconsidering cross-lingual word embeddings. *arXiv preprint arXiv :1608.05426*, 2016.

Bo Li et Eric Gaussier. Improving corpus comparability for bilingual lexicon extraction from comparable corpora. Dans *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 644–652, 2010.

Hang Li. A Short Introduction to Learning to Rank. *Transactions of IEICE*, pages 1854–1862, 2011.

Alexis Linard, Béatrice Daille et Emmanuel Morin. Attempting to Bypass Alignment from Comparable Corpora via Pivot Language. *ACL-IJCNLP 2015*, page 32, 2015.

Thang Luong, Hieu Pham et Christopher D. Manning. Bilingual word representations with monolingual quality in mind. Dans *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 151–159, 2015.

Xiaoyi Ma et Mark Liberman. Bits : A method for bilingual text search over the web. Dans *Machine Translation Summit VII*, pages 538–542. Citeseer, 1999.

Christopher D. Manning, Prabhakar Raghavan et Hinrich Schütze. *Introduction to Information Retrieval*. 2008.

Rada Mihalcea et Andras Csomai. Wikify ! : linking documents to encyclopedic knowledge. Dans *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 233–242, 2007.

Tomas Mikolov, Kai Chen, Greg Corrado et Jeffrey Dean. Efficient estimation of word representations in vector space. 2013a.

Tomas Mikolov, Quoc V. Le et Ilya Sutskever. Exploiting similarities among languages for machine translation. *CoRR*, 2013b.

David Milne et Ian H. Witten. Learning to link with wikipedia. Dans *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 509–518, 2008.

David Milne et Ian H. Witten. An open-source toolkit for mining wikipedia. *Artif. Intell.*, 194:222–239, 2013.

Emmanuel Morin et Béatrice Daille. Compositionality and lexical alignment of multiword terms. page 0, 2009.

Emmanuel Morin et Amir Hazem. Looking at Unbalanced Specialized Comparable Corpora for Bilingual Lexicon Extraction. 2014.

Emmanuel Morin et Emmanuel Prochasson. Bilingual lexicon extraction from comparable corpora enhanced with parallel corpora. Dans *Proceedings of the 4th workshop*

on building and using comparable corpora : comparable corpora and the web, pages 27–34. Association for Computational Linguistics, 2011a.

Emmanuel Morin et Emmanuel Prochasson. Bilingual lexicon extraction from comparable corpora enhanced with parallel corpora. Dans *Proceedings of the 4th workshop on building and using comparable corpora : comparable corpora and the web*, pages 27–34, 2011b.

Dragos Stefan Munteanu et Daniel Marcu. Improving machine translation performance by exploiting non-parallel corpora. *Comput. Linguist.*, 31(4):477–504, 2005.

Roberto Navigli et Simone Paolo Ponzetto. BabelNet : The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250, 2012.

Pablo Gamallo Otero. Learning bilingual lexicons from comparable english and spanish corpora. *Proceedings of MT Summit xI*, pages 191–198, 2007.

Alexandre Patry et Philippe Langlais. Identifying Parallel Documents from a Large Bilingual Collection of Texts : Application to Parallel Article Extraction in Wikipedia. BUCC '11, pages 87–95, 2011a.

Alexandre Patry et Philippe Langlais. Identifying parallel documents from a large bilingual collection of texts : Application to parallel article extraction in wikipedia. Dans *Proceedings of the 4th Workshop on BUCC*, pages 87–95, 2011b.

Viktor Pekar, Ruslan Mitkov, Dimitar Blagoev et Andrea Mulloni. Finding translations for low-frequency words in comparable corpora. *Machine Translation*, (4):247–266, 2006.

Emmanuel Prochasson et Pascale Fung. Rare word translation extraction from aligned comparable documents. HLT '11, pages 1327–1335, 2011.

Reinhard Rapp. Identifying Word Translations in Non-parallel Texts. Dans *Proceedings of the 33rd ACL*, pages 320–322, 1995.

Reinhard Rapp. Automatic identification of word translations from unrelated English and German corpora. Dans *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 519–526. Association for Computational Linguistics, 1999.

Reinhard Rapp. A methodology for bilingual lexicon extraction from comparable corpora. *ACL-IJCNLP 2015*, page 46, 2015.

Reinhard Rapp et Serge Sharoff. Extracting multiword translations from aligned comparable documents. Dans *Proceedings of the 3rd Workshop on Hybrid Approaches to Translation (HyTra)@ EACL*, pages 83–91, 2014.

Reinhard Rapp, Serge Sharoff et Bogdan Babych. Identifying Word Translations from Comparable Documents Without a Seed Lexicon. Dans *LREC*, pages 460–466, 2012.

Lise Rebout et Philippe Langlais. An iterative approach for mining parallel sentences in a comparable corpus. Dans *LREC*, pages 648–655, 2014.

Philip Resnik et Noah A. Smith. The web as a parallel corpus. *Computational Linguistics*, 29(3):349–380, 2003.

Xavier Robitaille, Yasuhiro Sasaki, Masatsugu Tonoike, Satoshi Sato et Takehito Utsuro. Compiling french-japanese terminologies from the web. Dans *EACL*, 2006.

Raphaël Rubino et Georges Linarès. A multi-view approach for term translation spotting. Dans *Proceedings of the 12th International Conference on Computational Linguistics and Intelligent Text Processing - Volume Part II*, CICLing'11, pages 29–40, 2011.

Sebastian Ruder. A survey of cross-lingual embedding models. `http://sebastianruder.com/cross-lingual-embeddings/index.html`, 2016. [En Ligne ; Accédé le 25 Janvier 2017].

Magnus Sahlgren. The distributional hypothesis. *Italian Journal of Linguistics*, 20:33–54, 2008.

Charles Schafer et David Yarowsky. Inducing translation lexicons via diverse similarity measures and bridge languages. Dans *proceedings of the 6th conference on Natural language learning-Volume 20*, pages 1–7. Association for Computational Linguistics, 2002.

Serge Sharoff, Reinhard Rapp et Pierre Zweigenbaum. Overviewing Important Aspects of the Last Twenty Years of Research in Comparable Corpora. Dans *Building and Using Comparable Corpora*, pages 1–17. Springer Berlin Heidelberg, 2013.

Serge Sharoff, Pierre Zweigenbaum et Reinhard Rapp. BUCC shared task : Cross-language document similarity. Dans *Proceedings of the Eighth Workshop on Building and Using Comparable Corpora*, pages 74–78, Beijing, China, juillet 2015. Association for Computational Linguistics. URL `http://aclweb.org/anthology/W/W15/W15-3411.pdf`.

Lei Shi, Cheng Niu, Ming Zhou et Jianfeng Gao. A dom tree alignment model for mining parallel data from the web. Dans *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 489–496. Association for Computational Linguistics, 2006.

Frank Smadja. How to compile a bilingual collocational lexicon automatically. Dans *Proceedings of the AAAI Workshop on Statistically-Based NLP Techniques*, pages 65–71, 1992.

Jason R. Smith, Herve Saint-Amand, Magdalena Plamada, Philipp Koehn, Chris Callison-Burch et Adam Lopez. Dirt cheap web-scale parallel text from the common crawl. Dans *Proceedings of the 51st ACL*, pages 1374–1383, 2013.

Akihiro Tamura, Taro Watanabe et Eiichiro Sumita. Bilingual lexicon extraction from comparable corpora using label propagation. Dans *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 24–36. Association for Computational Linguistics, 2012.

Shyam Upadhyay, Manaal Faruqui, Chris Dyer et Dan Roth. Cross-lingual Models of Word Embeddings : An Empirical Comparison. Dans *Proceedings of ACL*, 2016.

Jakob Uszkoreit, Jay M. Ponte, Ashok C. Popat et Moshe Dubiner. Large scale parallel document mining for machine translation. Dans *Proceedings of the 23rd COLING*, pages 1101–1109, 2010.

Ivan Vulić, Wim De Smet et Marie-Francine Moens. Identifying word translations from comparable corpora using latent topic models. Dans *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies : short papers-Volume 2*, pages 479–484. Association for Computational Linguistics, 2011.

Ivan Vulic, Douwe Kiela, Stephen Clark et Marie-Francine Moens. Multi-modal representations for improved bilingual lexicon learning. Dans *The 54th Annual Meeting of the Association for Computational Linguistics*, page 188, 2016.

Ivan Vulić et Anna Korhonen. On the role of seed lexicons in learning bilingual word embeddings. Dans *Proceedings of ACL*, pages 247–257, 2016.

Ivan Vulic et Marie-Francine Moens. Bilingual word embeddings from non-parallel document-aligned data applied to bilingual lexicon induction. Dans *Proceedings of the 53rd ACL*, 2015.

Ivan Vulić et Marie-Francine Moens. Detecting highly confident word translations from comparable corpora without any prior knowledge. Dans *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 449–459. Association for Computational Linguistics, 2012.

Warren Weaver. TRANSLATION. `http://www.mt-archive.info/Weaver-1949.pdf`, 2016. [En Ligne ; Accéder le 28 Octobre 2016].

Wikipedia. Distributional Semantics. `https://en.wikipedia.org/wiki/Distributional_semantics`, 2016a. [En Ligne ; Accéder le 30 Janvier 2017].

Wikipedia. Informatique. https://fr.wikipedia.org/wiki/Informatique, 2016b. [En Ligne ; Accéder le 28 Octobre 2016].

Dekai Wu et Xuanyin Xia. Learning an english-chinese lexicon from a parallel corpus. Dans *Proceedings of the First Conference of the Association for Machine Translation in the Americas*, pages 206–213, 1994.

Chao Xing, Dong Wang, Chao Liu et Yiye Lin. Normalized Word Embedding and Orthogonal Transform for Bilingual Word Translation. Dans *HLT-NAACL*, pages 1006–1011, 2015.

Kun Yu et Junichi Tsujii. Extracting bilingual dictionary from comparable corpora with dependency heterogeneity. Dans *Annual Conference of the NAACL, Companion Volume : Short Papers*, pages 121–124, 2009.