Évaluation en extraction de lexiques bilingues à partir de corpus comparables

Martin Laville

COLLEGE ED MATHEMATIQUES
DOCTORAL STIC
PAYS DE LA LOIRE

Nantes Université

THÈSE DE DOCTORAT DE

NANTES UNIVERSITÉ
ÉCOLE DOCTORALE N° 641
Mathématiques et Sciences et Technologies du numérique, de l'Information et de la Communication
Spécialité : *Informatique*

Par
Martin LAVILLE

Évaluation en extraction de lexiques bilingues à partir de corpus comparables

Thèse présentée et soutenue à Nantes, le 1er février 2023
Unité de recherche : Laboratoire des Sciences du Numérique de Nantes (LS2N)

Rapporteurs avant soutenance :

Eric GAUSSIER	Professeur des Universités, Université Grenoble Alpes
Marianna APIDIANAKI	Chargée de Recherche, CNRS, Université de Pennsylvanie

Composition du Jury :

Président :	Pierre ZWEIGENBAUM	Directeur de Recherche CNRS, Université de Paris-Saclay
Examinateurs :	Eric GAUSSIER	Professeur des Universités, Université Grenoble Alpes
	Marianna APIDIANAKI	Chargée de Recherche, CNRS, Université de Pennsylvanie
	Pierre ZWEIGENBAUM	Directeur de Recherche CNRS, Université de Paris-Saclay
Dir. de thèse :	Emmanuel MORIN	Professeur des Universités, Nantes Université
Co-encadrant :	Philippe LANGLAIS	Professeur des Universités, Université de Montreal

Remerciements

Merci Emmanuel de m'avoir ouvert la porte de ton bureau pour ce stage de fin de M2. Merci de m'avoir ensuite proposé cette thèse et de m'avoir supporté pendant plus de 4 ans. Merci de m'avoir aidé, lorsque ça n'allait pas personnellement ou plus simplement dans le cadre de la thèse. Merci, vraiment, sans toi, je ne serais jamais allé au bout de ce manuscrit et de ce long travail qu'est une thèse. Je n'ai pas été, loin de là, le doctorant parfait, mais tu as toujours cherché à m'aider à avancer et à ne pas baisser les bras et probablement plus que tu ne le penses. Donc une dernière fois, un grand merci.

Ensuite, merci à Philippe Langlais, mon co-encadrant. Ta vision des choses m'a aidé à aller dans une direction pas forcément évidente au début. Et encore une fois, je m'excuse de tous ses articles envoyés très tard et te remercie de tes relectures toujours très rapides et pertinentes.

Je tiens aussi à remercier les différents membres de mon jury : Eric Gaussier et Marianna Apidianaki, mes rapporteurs, pour leurs retours et questions avisées, et particulièrement, mon examinateur Pierre Zweigenbaum, qui aura aussi été membre de mon CSI et présent tout au long de ma thèse.

Je remercie tous les membres de l'équipe TALN, d'abord les permanents et plus particulièrement Richard et Florian, qui m'ont aidé à aller jusqu'au bout, en partie grâce à leurs piques incessantes. Mais aussi les autres doctorants : Mérième, Victor, Adrien, Ygor, Oumaïma, Fawzi... qui ont tous été plus ou moins présents, mais avec qui ça a toujours été un plaisir de procras... travailler ! Mais aussi Amir, qui m'a beaucoup aidé et apporté lors de mon stage et le début ma thèse.

Merci aussi à tous les membres du LS2N qui travaillent dans l'ombre et aident les doctorants tout au long de leur parcours : Annie Boilot, Haritiana Rasolofoniaina, Virginie Olivier, Virginie Dupont, Alexiane Brard... Et bien évidemment, Charlery Vilar toujours disponible pour résoudre nos problèmes techniques.

Aux personnes extérieures qui m'ont aussi beaucoup apporté, mes amis qui m'ont aidé dans les moments de doute et tout particulièrement Lisa et Hugues, mais bien évidemment tout Seum 41**/SNAM CITY** ou je ne sais quel autre nom vous avez pu avoir, tant qu'on garde les deux étoiles, mais aussi les Ornithos et le SPA pour m'avoir permis de

m'évader lors de ces trois années.

Et finalement, mes parents et ma famille, dont beaucoup étaient présents le jour J. Maman, j'espère que tu sais à peu près quoi dire quand on te demande le sujet de ma thèse maintenant...

COYG

Table des matières

INTRODUCTION

Contexte

À l'heure des réseaux de neurones et des approches de traitement automatique des langues de plus en plus performantes, une ressource simple comme un lexique bilingue reste indispensable pour de nombreux travaux. Les lexiques bilingues peuvent être utilisés en tant que matériau au sein de systèmes de traduction automatique ou de recherche d'information, ou tout simplement en tant que dictionnaires spécialisés pour des traducteurs. Du fait de la difficulté de les réunir manuellement, de nombreux travaux ont été menés pour être capable de les réunir automatiquement. Les lexiques bilingues utilisés en traitement automatique des langues sont des ressources pouvant se trouver sous plusieurs formes. Généralement, on retrouve un mot en relation avec sa traduction, mais certains lexiques proposent aussi la catégorie grammaticale des mots ou leur forme canonique par exemple.

L'extraction de lexique bilingue (*bilingual lexicon induction* : BLI) est la tâche qui a pour but la production de ces ressources. Cette tâche était originellement effectuée à l'aide de corpus réunissant des textes en relation de traduction et a rapidement évolué vers des corpus dit comparables, plus facile à réunir. Les corpus comparables sont composés de documents traitant d'un même registre ou construits à partir d'une même source (Wikipédia par exemple) mais sans relation de traduction entre les textes, à l'inverse des corpus parallèles. Les corpus peuvent aussi être considérés comme généraux (sans sujet particulier), pour lesquels les lexiques extraits servent par exemple dans les systèmes de traduction automatique, ou spécialisés dont les lexiques sont alors souvent dédiés à des traducteurs. L'exécution de la tâche de BLI nécessite aussi généralement l'existence d'un premier lexique, même si les approches les plus récentes tendent à se détacher de ce besoin.

Les approches de BLI sont pour la plupart composées de trois principales étapes. Dans un premier temps, il est nécessaire de créer une représentation sémantique des mots. Cette représentation, sous forme vectorielle, est apprise à l'aide de corpus, souvent de manière monolingue. Comme cette tâche s'effectue sur au moins deux langues, il est ensuite nécessaire de trouver un moyen de comparer les représentations des mots des deux langues.

Cette phase d'alignement utilise habituellement un premier lexique bilingue d'entraînement, même si la nécessité de ce dernier va en diminuant, du fait d'approches de plus en plus performantes. Finalement, une fois les deux vocabulaires rendus comparables, on peut indiquer pour un mot de la langue source quelles sont les traductions correspondantes dans le vocabulaire cible.

Objectifs de recherche

L'extraction de lexique bilingue en domaine spécialisé pose problème du fait de la difficulté de réunir des corpus en qualité et quantité suffisantes pour les approches classiques de BLI. Ces dernières années, les approches étudiées dans la plupart des travaux liés au traitement automatique des langues nécessitent de grandes quantités de données ce qui rend encore plus rares les études en domaine de spécialité. Hazem and Morin (2016, 2018) proposent de combiner les données spécialisées à des données générales pour améliorer les résultats en ayant des corpus d'apprentissage bien plus grands. Mais cet ajout induit des problèmes de polysémie que l'on cherche à éviter dans des données spécialisées. **En continuant dans cette direction, est-il tout de même possible d'améliorer les résultats en domaine spécialisé, mais cette fois sans introduire de polysémie qui dénature les données spécialisées ?**

En domaine général, l'évaluation de la tâche de BLI a rarement été conduite de manière similaire en fonction des travaux menés ces dernières années. Ce manque d'uniformité rend difficile les comparaisons entre les nombreux systèmes qui ont été proposés au cours des années. On note toutefois des tentatives d'uniformiser les approches avec l'arrivée par exemple du jeu de données MUSE (Conneau et al., 2017), même si la quantité semble avoir été favorisée à la qualité. Ce jeu de données comporte de nombreux problèmes que nous étudierons au cours de cette thèse et il peut être difficile de légitimer des avancées en BLI à partir de ces données. **Nous cherchons donc à analyser et comprendre les données utilisées aujourd'hui en BLI et à valider s'il est possible de les utiliser de manière uniforme et cohérente pour évaluer les systèmes.**

Contributions

Dans cette thèse, nous nous intéressons au BLI en domaine de spécialité et général. Dans le cas des domaines spécialisés, nous étudions en profondeur les problèmes apportés

par la technique d'ajout de données proposée par Hazem and Morin (2016, 2018). Nous proposons alors d'utiliser des techniques de sélection de données, pour n'ajouter dans le corpus spécialisé que les données générales qui correspondent au domaine étudié. En comparaison aux approches utilisant de l'augmentation de données, notre système permet d'améliorer les résultats de près de 6 points de MAP et permet la réduction des temps de calcul par un facteur de 10.

Dans le cas du domaine général, des premières expériences nous ont rapidement permis de réaliser que la qualité des listes d'évaluation dans le domaine général était hautement discutable du fait de la présence de paires de mots représentant soit peu d'intérêt en BLI soit carrément incorrectes. Nous nous concentrons alors sur l'évaluation en BLI et conduisons une analyse approfondie des méthodes et listes d'évaluation et proposons des pistes pour obtenir une évaluation plus uniforme entre les systèmes et qui s'intéresse aux réelles difficultés du BLI.

Plan

Dans le Chapitre 1, nous présentons les ressources qui sont souvent utilisées en BLI. Nous présentons dans un premier temps différents types de corpus, qui sont le matériau principal de la majeure partie des approches de BLI. Puis nous discutons rapidement des lexiques utilisés pour créer un pont entre les différentes langues qui peuvent être étudiées. Enfin, nous introduisons les jeux de données qui seront utilisés au cours de nos expérimentations.

Être capable de représenter les mots d'un corpus est la première étape de la plupart des tâches de traitement automatique des langues et le BLI ne déroge pas à cette règle. Dans le Chapitre 2, nous étudions donc différents types de représentations sémantiques. Nous détaillons la méthode historique basée sur les sacs de mots, puis nous intéressons aux plongements de mots (Mikolov et al., 2013a) ainsi qu'aux modèles plus complexes tels que ELMo (Peters et al., 2018) ou BERT (Devlin et al., 2019).

Les méthodes de représentation présentées dans le chapitre précédent sont avant tout adaptées pour des approches monolingues et il est donc nécessaire de pouvoir adapter les représentations ou les méthodes pour obtenir des représentations comparables entre les langues. Le Chapitre 3 présente les méthodes qui permettent la comparaison entre deux langues différentes. Nous présentons deux principaux types de méthodes, d'abord les méthodes d'alignement, qui prennent les représentations monolingues et cherchent à les

projeter dans un espace commun, puis ensuite les méthodes d'entraînement conjoint qui, à l'aide de combinaison des corpus ou de changements dans les objectifs d'apprentissage, permettent la création de représentations bilingues directement lors de l'entraînement à partir des corpus. La plupart de ces méthodes nécessitent un moyen de faire le lien entre les deux langues étudiées, souvent sous la forme d'un lexique d'entraînement.

Dans le Chapitre 4, nous introduisons une étude du domaine spécialisé et des problèmes que l'on peut y rencontrer. Les domaines spécialisés sont souvent peu fournis en données, ce qui rend difficile de faire fonctionner correctement les approches de BLI. L'ajout de données générales est un premier pas qui permet d'obtenir de meilleures représentations des mots, mais qui induit des problèmes comme l'introduction de polysémie et une grande augmentation des temps de calcul. Nous proposons alors d'utiliser des techniques de sélection de données pour n'ajouter au domaine spécialisé que les données venant du corpus général présentant un réel intérêt pour le domaine spécialisé étudié.

Dans le Chapitre 5, nous nous concentrons dans un premier temps sur les méthodes utilisées en BLI et les résultats obtenus en fonction de différents scénarios. Ces études nous permettent de nous questionner sur le processus et les données utilisées pour l'évaluation de la tâche de BLI. Nous conduisons alors une étude approfondie du jeu de données MUSE utilisé dans la majeure partie des derniers travaux étudiant la tâche de BLI, puis nous cherchons à définir un processus unifié d'évaluation en réunissant certaines propositions déjà faites, mais rarement suivies, et formulons nos propres recommandations.

Enfin, nous concluons et proposons différentes perspectives pour continuer ces travaux.

Chapitre 1

Type de ressources et jeux de données

En BLI, le matériau principal des systèmes est le corpus. Un corpus est un ensemble de documents textuels, parfois monolingue, mais dans notre cas d'étude, il sera toujours composé d'au moins deux langues. Il sert à entraîner les différents systèmes utilisés pour accomplir cette tâche. Initialement conduite sur des corpus parallèles (des textes en relations de traduction), la tâche a rapidement évolué pour utiliser des corpus comparables du fait des limitations des premiers nommés (Rapp, 1995; Fung, 1998). Pour pouvoir comparer les différentes langues, les méthodes utilisées vont souvent nécessiter des lexiques bilingues. Ce besoin a tendance à s'effacer avec de nouvelles approches n'en nécessitant pas, tout en étant quasiment aussi efficaces que celles en faisant usage.

Dans ce chapitre, nous présentons les différents types de ressources utilisées en BLI, d'abord les corpus, que nous classons sur deux axes : leur comparabilité et leur spécialisation ; ensuite, les différents types de lexiques et comment ils sont utilisés (ou non) en fonction du type de méthode. Enfin, nous introduisons brièvement différents corpus et lexiques qui seront utilisés au fil de ce manuscrit.

1.1 Corpus

Dans cette section, nous présentons les différents types de corpus utilisés en BLI. Ces corpus servent de données d'entraînement pour créer les représentations sémantiques des mots que les systèmes de BLI utilisent pour obtenir des traductions.

1.1.1 Parallèle & comparable

Dans le cadre du BLI, les corpus utilisés sont composés de deux parties. Une partie appelée *Source*, de la langue d'origine que l'on cherche à traduire et une partie appelée

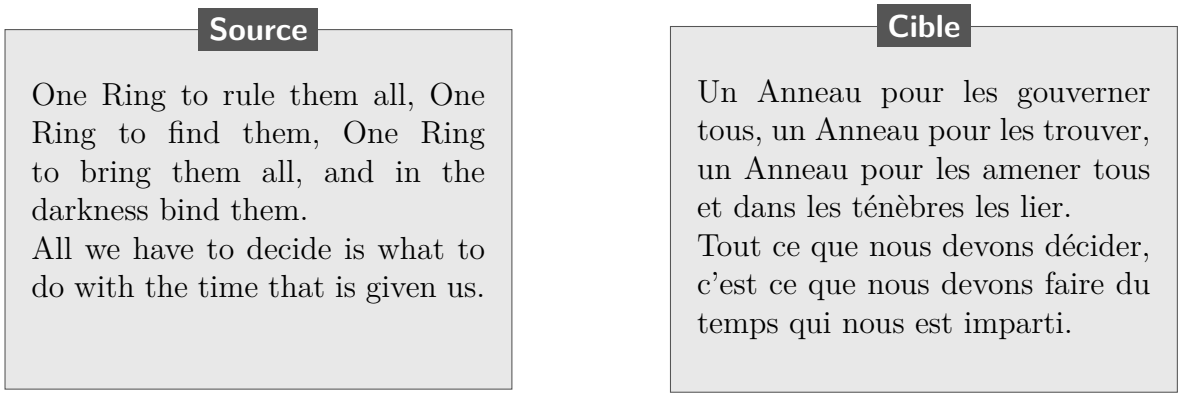

FIGURE 1.1 – Exemple de phrases alignées issues d'un corpus parallèle (*J.R.R. Tolkien, The Fellowship of the Ring / La Communauté de l'Anneau*)

Cible, de la langue d'arrivée.

Différents types de corpus peuvent être utilisés, d'abord, on peut parler des **corpus parallèles**, qui ont été les premiers utilisés dans cette tâche. Les deux parties de ce corpus sont en relation directe de traduction, ainsi comme représenté en Figure 1.1, chaque phrase source est alignée avec sa traduction dans la partie cible. Ces corpus sont assez difficiles à obtenir et à créer, car il est nécessaire d'avoir recours à un traducteur. Cependant, la tâche de BLI est rapidement considérée comme résolue dans ce cadre (Kay et al., 1994). On peut mentionner le corpus parallèle Europarl (Koehn, 2005) qui est une collection des débats du parlement européen alignés dans 11 langues.

Les corpus parallèles étant limités d'accès, ils sont finalement peu adaptés au BLI, car ils sont d'abord quasiment inexistants dans le cas de langues peu dotées. Et même dans le cas des langues plus fournies telles que l'anglais, il reste difficile de réunir beaucoup de données parallèles et le vocabulaire qui pourra être étudié est réduit à celui du corpus parallèle dont on dispose.

Les corpus parallèles présentent aussi un autre problème via le biais des traducteurs. En effet, un traducteur aura tendance à écrire différemment d'un locuteur natif. Un exemple simple est le fait qu'un traducteur donner plus d'informations que ce qui était implicitement évoqué dans le texte source (Frankenberg-Garcia, 2009). De plus, l'usage d'adverbe est différent en fonction de la langue originelle : Koppel and Ordan (2011) utilisent ce phénomène pour entraîner des classifieurs déterminant la langue source d'un texte traduit.

Finalement, la tâche de BLI étant relativement triviale avec ces corpus, les travaux qui ont suivi ont assez logiquement fait usage de **corpus comparables** qui, en plus de

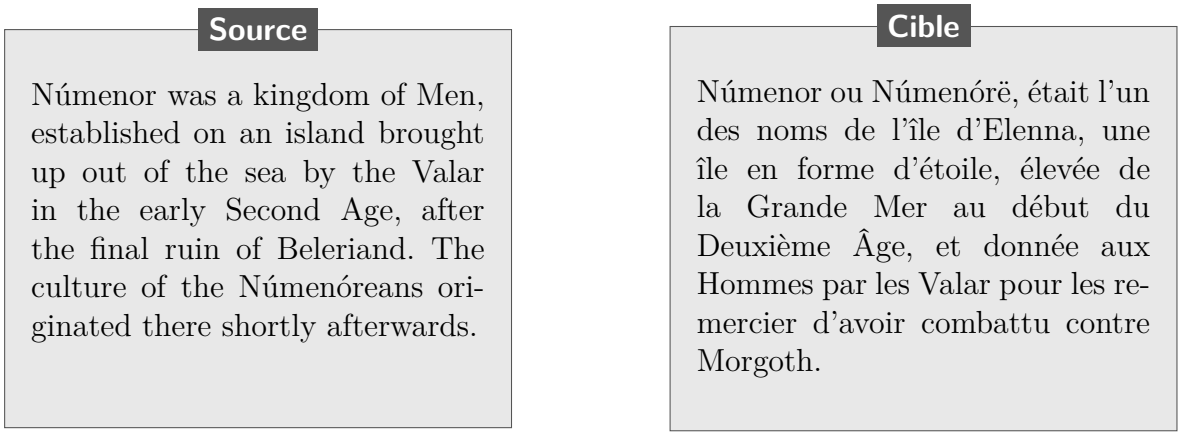

FIGURE 1.2 – Exemple d'un corpus comparable, extrait d'un wiki [1] à propos d'un même sujet.

représenter un défi plus intéressant d'un point de vue recherche, sont, à l'inverse des corpus parallèles, bien plus faciles à construire et à réunir et ne possèdent pas de biais liés à une langue originelle de traduction ou un traducteur.

À l'inverse des corpus parallèles, ces corpus n'ont pas leurs deux parties en relation de traduction, mais sont composés de textes de la même époque, d'un registre similaire et de sujets relativement semblables.

Sharoff et al. (2013) proposent plus de nuance en introduisant différents degrés de comparabilité. D'abord, les corpus parallèles qui, comme indiqués précédemment, sont des ensembles de textes en relation de traduction. Ensuite, les corpus fortement comparables, qui concernent des textes provenant d'une même source, mais écrits dans différentes langues (par exemple, BBC News en anglais ou roumain (Munteanu and Marcu, 2006)) ou alors des articles de Wikipédia sur un même concept. Troisièmement, les corpus faiblement comparables, qui rassemblent des textes sur un même sujet, mais qui décrivent différents évènements : par exemple, des débats parlementaires sur des sujets de santé, mais conduits dans différents pays. Finalement, ils ajoutent aussi les corpus non liés, qui sont composés de textes ne possédant aucune des caractéristiques précédemment évoquées. Par exemple, prendre le contenu de pages aléatoires d'internet dans différentes langues correspondra à un corpus non lié.

1. https://lotr.fandom.com/

1.1.2 Général & spécialisé

On distingue aussi un autre axe de caractérisation des corpus : leur spécialisation. Un corpus général est composé d'un vocabulaire très large et aborde de nombreux sujets. À l'inverse, un corpus spécialisé se concentre sur un domaine bien précis avec un vocabulaire très spécifique (principalement scientifique ou technique).

Dans un domaine général, les mots peuvent avoir plusieurs sens et posséder plusieurs traductions. Dans le domaine spécialisé, il est par contre admis que les mots ne possèdent qu'un seul sens et une seule traduction.

Les corpus spécialisés sont des ressources difficiles à construire, car être capable de réunir des données sur un même sujet scientifique ou technique en grande quantité est une tâche assez difficile. Cette problématique fait que les corpus spécialisés sont souvent de taille très faible en comparaison aux à des corpus généraux. Cependant, obtenir les traductions de termes plus complexes via les corpus spécialisés présente par exemple un grand intérêt pour les traducteurs, étant donné que ces traductions auront tendance à se faire plus rares dans les dictionnaires généraux. La possibilité de les extraire à partir de corpus est donc intéressante, car elle permet de facilement replacer ces mots dans leur contexte.

1.2 Lexiques

L'objectif premier de la tâche de BLI est d'extraire des paires de mots en relation de traduction. Pour cela, les méthodes utilisent elles-mêmes souvent un lexique bilingue d'entraînement qui va servir de liant entre les deux langues étudiées, dans d'autres cas, le lexique sera généré avant l'apprentissage à partir de similarité entre les langues ou de manière non supervisée. Les lexiques seront majoritairement représentés comme des listes de paires (voir Table 1.1), sans définition ni information grammaticale.

On peut distinguer trois catégories de lexiques bilingues, qui seront utilisés pour trois types de méthodes différentes :

1. Les dictionnaires bilingues. Ils peuvent avoir été créés manuellement (projet EurADiC[2]) ou générés de manière supervisée (Conneau et al., 2017). Les méthodes utilisant des dictionnaires bilingues peuvent avoir des besoins totalement différents en termes de taille de lexique. Morin and Prochasson (2011) utilisent par exemple

2. http://www.technolangue.net/imprimer.php3?id_article=203

Anglais	**Français**
the	le
the	les
the	la
known	connues
known	connue
known	connus
known	connu

TABLEAU 1.1 – Quelques paires d'un dictionnaire bilingue anglais-français

des dictionnaires composés de 50 000 paires. Mikolov et al. (2013b) utilisent les 5 000 mots les plus fréquents du vocabulaire source et leur traduction. Artetxe et al. (2017) proposent une méthode nécessitant initialement seulement 25 paires, même si ce dictionnaire ne sert que pour initialiser le système puis est affiné au fur et à mesure de l'apprentissage. Les méthodes utilisant ces ressources sont dites **supervisées**.

2. Les pseudos-dictionnaires. Prenant appui sur les ressemblances entre les langues, ils peuvent être construits par exemple à l'aide des cognats (Smith et al., 2017) ou des nombres (Artetxe et al., 2017) et servent pour l'initialisation du système avant d'être renforcés par d'autres paires lors de l'entraînement. Les méthodes utilisant ces ressources sont dites **semi-supervisées**.
3. Les dictionnaires non-supervisés. Ces lexiques ne sont construits qu'à partir des données utilisées et n'utilisent pas les similarités graphiques entre les langues. Conneau et al. (2017) proposent d'utiliser un discriminateur qui a pour but de prédire si un vecteur est originaire de l'espace source ou cible. Artetxe et al. (2018b) se basent sur la propriété d'isomorphisme (voir Section 3.2) pour initialiser leur dictionnaire. Les méthodes initialisant un dictionnaire de cette manière sont appelées **non-supervisées**.

Originellement, les lexiques utilisés étaient composés de nombreuses paires de mot. Peu à peu, la taille des lexiques a énormément diminué et les méthodes atteignent aujourd'hui des performances quasiment similaires entre les approches supervisées et non-supervisées. Artetxe et al. (2020) questionnent toutefois l'utilité des approches non-supervisées, car obtenir suffisamment de données pour construire des corpus dans deux langues sans avoir accès à un lexique bilingue est hautement improbable. Cependant, la tâche reste d'un intérêt scientifique pour aider dans la compréhension des langues.

1.3 Jeux de données

Nous présentons maintenant dans cette section différents jeux de données utilisés en BLI et dans les travaux présentés dans le cadre de cette thèse. Les jeux de données que nous exposons ici sont composés de corpus, de lexiques ou des deux à la fois.

Breast Cancer dataset (BC) (corpus et lexique) : le jeu de données du cancer du sein est composé d'un corpus comparable et spécialisé ainsi que du lexique d'évaluation correspondant. Le corpus est composé de documents scientifiques publiés entre 2001 et 2015 sur le portail ScienceDirect [3]. Les documents sont sélectionnés si le terme *Cancer du Sein* ou de la traduction dans la langue désirée (*Breast Cancer* pour l'anglais par exemple) apparaissent en titre ou dans les mots clés. Il est comparable, car composé de textes scientifiques issus d'une plateforme commune, et spécialisé parce que centré sur un concept précis (le cancer du sein). Il est disponible en quatre langues : allemand, anglais, espagnol et français. Les différentes parties du corpus sont de taille relativement modeste : environ 500k mots par langue. ScienceDirect étant relativement peu fourni en termes de textes allemands, cette partie du corpus est aussi construite à partir de sites d'information pour les patients atteints du cancer du sein ou leurs proches. Elle peut donc être considérée de moins bonne qualité que les autres, surtout du point de vue de sa spécialisation. On note d'ailleurs dans le Tableau 1.2 que, s'il est composé du plus petit nombre de mots, le corpus possède le vocabulaire le plus grand des quatre parties, ce qui peut confirmer l'affirmation précédente.

Ce corpus étant spécialisé, il requiert des données d'évaluation adaptées. Le lexique d'évaluation est donc composé de six listes spécialisées reliant les quatre langues (allemand, anglais, espagnol et français). À l'instar de nombreuses études dans les domaines spécialisés, les listes étudiées sont de petites tailles : entre 100 et 250 paires de mots par liste. À titre de comparaison, Chiao and Zweigenbaum (2002) utilisent une liste de 95 paires et Bouamor et al. (2013) des listes de 79 et 125 paires.

Wind Energy dataset (WE) (corpus et lexique) : le jeu de données Wind Energy est similaire au jeu de données BC : il est composé d'un corpus comparable et spécialisé, accompagné d'un lexique d'évaluation adapté. Le corpus, disponible en anglais/français, a été créé dans le cadre du projet TTC [4]. Une aspiration du web a permis de réunir les différents documents qui composent le corpus, en utilisant des mots clés tels que *wind*, *energy*, *rotor* en anglais et leur traduction en français. Les deux parties du corpus sont

3. www.sciencedirect.com/
4. http://www.ttc-project.eu/

légèrement déséquilibrées en termes de taille avec la partie anglaise (environ 300 k mots) deux fois plus petite que la partie française (600 k). Sa méthode de construction (une aspiration globale et pas via une plateforme dédiée), fait qu'il est de qualité plus faible que le corpus BC (hormis pour la partie allemande).

Le lexique d'évaluation quant à lui correspond aux deux langues du corpus : anglais-français. La liste est composée de 143 paires de mots.

Wikipédia corpus (Wiki) (corpus) : le corpus Wikipédia est utilisé dans de nombreux travaux en BLI. Du fait de la multitude de sujets qu'il traite et de sa structure, il est considéré comme un corpus général et comparable. Il est disponible dans plusieurs centaines de langues. La taille des différentes parties est énormément variable en fonction de la langue : plus d'un million d'articles différents en anglais et moins de 1 000 pour d'autres langues (par exemple hawaiian, maori, khmer ou cree). Malgré tout, il reste une ressource facile d'utilisation et très accessible. Il est généralement issu de *dumps*[5] nettoyés à l'aide d'outils tels que WikiExtractor (Attardi, 2015).

Common Crawl (CC) (corpus) : le corpus Common Crawl est composé de textes réunis à travers le web depuis 2008[6]. De la même manière que Wikipédia, il est considéré comme général et comparable. S'il est majoritairement composé de données anglaises (près de 50 %), il est néanmoins disponible en plus de 40 langues. Par son processus de création qui contient de nombreuses sources comparées à Wikipédia par exemple, il est considéré comme un corpus à faible comparabilité.

JRC Acquis corpus (JRC) (corpus) : JRC Acquis est un corpus composé de textes législatifs issus de l'Union Européenne qui comprend des documents écrits entre les années 50 et aujourd'hui[7]. Il comprend 22 langues ayant entre 20 M et 70 M de mots par partie. Il est majoritairement comparable, mais chaque paire de langues possède aussi un certain nombre de phrases alignées, créant ainsi un corpus parallèle. Dans le cadre de ce manuscrit, nous avons utilisé la version anglais-français (Tiedemann, 2012). S'il possède une certaine forme de spécialité, car composé de textes législatifs ayant donc un registre de langue particulier, les sujets traités sont suffisamment vastes pour le considérer comme général.

Web-As-Corpus Kool Yinitiative (WaCKy) (corpus) : le corpus WaCKy (Baroni et al., 2009) est un corpus disponible en plusieurs langues. Il a été composé à l'aide d'une aspiration du web, basé sur des paires aléatoires de mots de la langue recherchée. Par exemple, pour la partie anglaise, 1 000 paires aléatoire de mots ont été créées à

5. https://dumps.wikimedia.org/
6. https://commoncrawl.org/
7. https://ec.europa.eu/jrc/en/language-technologies/jrc-acquis

partir du *British National Corpus*. Ces paires sont ensuite utilisées comme mots clés pour l'aspiration.

ELRA (lexiques) : les dictionnaires ELRA proviennent du projet EurADiC [8]. Dans le cadre de nos travaux, nous n'utilisons que le dictionnaire anglais-français, composé de près de 250k paires de mots. Ils servent majoritairement de lexique d'entraînement, mais peuvent être utilisés pour évaluer des méthodes dans le domaine général.

MUSE (lexiques) : le jeu de donnée MUSE (Conneau et al., 2017) est composé de nombreuses paires de langues. D'abord, six langues (allemand, anglais, espagnol, français, italien et portugais) avec toutes les combinaisons de listes possibles (pour un total de 15 lexiques, 30 en comptant les lexiques réciproques). Ensuite, 40 autres langues (par exemple arabe, hindi, indonésien, chinois...) seulement alignées avec l'anglais. Cette ressource met donc à disposition de nombreuses langues, de différentes familles et construites avec différents systèmes d'écriture. Les lexiques sont composés de dizaine de milliers de paires et proposent un pré-découpage pour l'entraînement (5 000 mots sources et leur traduction) et l'évaluation (1 500 mots sources et leur traduction). Il a été construit de manière automatique ("*using an internal tool*") et est aujourd'hui l'un des jeux de données les plus utilisés en BLI.

MORPH (lexiques) : le jeu de données MORPH a été introduit par Czarnowska et al. (2019). Il est composé de 5 langues slaves (polonais, tchèque, russe, slovaque et ukrainien) reliées entre elles et de 5 langues latines (catalan, espagnol, français, italien et portugais) aussi reliées entre elles. De la même manière que MUSE, il possède un pré-découpage entraînement/évaluation.

Corpus	Spé / Gen	Comp / Para	anglais		français	
			#mots	|voca|	#mots	|voca|
BC	Spé	Comp	525,9 k	14,8 k	521,3 k	11,7 k
WE	Spé	Comp	311,9 k	15,3 k	656,2 k	15,8 k
Wiki	Gen	Comp	1 975 M	4,0 M	781 M	3,0 M
CC	Gen	Comp	81,1 M	259 k	91,3 M	251 k
JRC	Gen	Les deux	64,9 M	229,8 k	69,0 M	231,1 k

TABLEAU 1.2 – Quelques caractéristiques de corpus utilisés dans ce manuscrit. On y indique leur type, ainsi que le nombre de mots et la taille du vocabulaire par partie. Si JRC comprend certaines parties parallèles, cette propriété n'est pas utilisée au cours des travaux présentés dans ce manuscrit.

8. http://technolangue.org/imprimer.php3?id_article=306

Lexique	Langues	train	test
BC	en-fr	-	251
	fr-es	-	125
WE	en-fr	-	143
ELRA	en-fr	243, 5 k	
MUSE	fr-it	5, 1 k	1, 5 k
	en-ru	10, 9 k	2, 4 k
MORPH	es-fr	398, 1 k	144, 1 k
	it-es	361, 8 k	130, 3 k

TABLEAU 1.3 – Quelques exemples de lexiques par jeu de données et paire de langues. On y indique la taille des listes en nombre de paires.

Les Tableaux 1.2 et 1.3 présentent des données sur les corpus et lexiques présentés précédemment. On note que les corpus spécialisés sont bien moins fournis que les corpus généraux, ce qui peut poser problème, car les approches de BLI sont majoritairement pensées pour de larges quantités de données.

1.4 Synthèse

Le BLI est une tâche qui s'est rapidement définie comme utilisant des corpus comparables (des textes possédant un sujet, un registre et une époque communs) plutôt que des corpus parallèles (des textes en relation de traduction) (Rapp, 1995; Fung, 1998). Le degré de spécialisation des corpus est aussi une particularité importante et nous verrons plus tard qu'il est nécessaire d'adapter les méthodes aux corpus spécialisés, car ils sont de taille relativement faible comparés aux corpus généraux.

Dans ce chapitre, nous avons aussi présenté les lexiques et comment ils sont utilisés en fonction des différentes approches. Ces lexiques peuvent servir de données d'apprentissage ou d'évaluation. On notera aussi que le lexique MUSE a été obtenu de manière automatique et nous étudions les problèmes que cela peut apporter dans le Chapitre 5.

Chapitre 2

Représentations sémantiques des mots

Obtenir les meilleures représentations sémantiques des mots (ou des phrases) est une tâche complexe en apprentissage automatique de la langue. L'origine des premières représentations de mots peut être ramenée à l'hypothèse distributionnelle (Harris, 1954), qui suppose que des mots qui apparaissent et sont utilisés dans des contextes similaires doivent avoir un sens proche.

Firth (1957) décrit ce phénomène de manière plus contrainte : " *You shall know a word by the company it keeps*". À l'origine, "company" indiquait les mots possédant des relations syntaxiques avec le mot central, au fil du temps, "company" a évolué vers "contexte", c'est-à-dire les mots qui entourent un mot central. Cette hypothèse a plus tard servi comme explication à l'approche dite de "sac de mots", des représentations vectorielles des mots où chaque dimension représente le nombre de cooccurrences entre un mot central et tous ses mots de contexte.

Si les sacs de mots ont longtemps été l'approche dominante, l'arrivée d'approches basées sur des réseaux de neurones profonds (Mikolov et al., 2013a; Peters et al., 2018) a permis de se défaire des principaux défauts des sacs de mots en réduisant la taille des vecteurs et en les densifiant, ces représentations sont appelées plongements de mots. En contrepartie, les plongements de mots sont plus difficilement interprétables que les vecteurs basés sur une représentation en sacs de mots.

Dans ce chapitre, nous introduisons et explicitons différents types de représentations sémantiques des mots, en commençant par les sacs de mots. Puis, nous présentons les deux principales approches utilisant les réseaux de neurones pour créer des plongements de mots : CBOW et Skip-gram ainsi qu'une de leur principale amélioration, *fastText*. Ensuite, nous étudions brièvement les premières approches créant des plongements de mots contextualisés (c'est-à-dire qui varient pour chacune des occurrences d'un mot en fonction de son contexte) avec BERT et ELMo.

2.1 Sac de mots

La création des vecteurs basés sur une représentation en sacs de mots est basée sur le compte des cooccurrences au sein du corpus. Pour calculer les cooccurrences d'un mot, on définit une fenêtre contextuelle autour de laquelle il est considéré qu'un mot c fait partie du contexte d'un mot central w. Formellement, c cooccurre avec w si $c \in w_{\pm 1..f}$, où f représente la taille de la fenêtre contextuelle et $w_{\pm 1..f}$ correspond alors aux mots présents dans la fenêtre de contexte de w. Dans la Figure 2.1, en prenant ***force*** comme mot central, on aura par exemple $cooc(w) = \{to, the, not, leave\}$ pour une fenêtre de taille 2.

bring	balance	*to*	*the*	**force**	*not*	*leave*	it	in	darkness
w_{-4}	w_{-3}	w_{-2}	w_{-1}	w	w_{+1}	w_{+2}	w_{+3}	w_{+4}	w_{+5}

FIGURE 2.1 – En gras, le mot central et en italique les mots du contexte avec une fenêtre de taille 2.

Le calcul, pour toutes les occurrences d'un mot, de ses cooccurrences, nous permet d'obtenir son vecteur de contexte, dans lequel chacune des dimensions correspond à un mot du vocabulaire et au nombre de fois où il cooccurre avec le mot de départ. Appliquer cette procédure à chaque mot du vocabulaire permet d'obtenir une matrice de cooccurrence (voir Figure 2.2) de la taille du vocabulaire. Dans cette matrice, chaque entrée correspond donc au nombre de fois où deux mots se rencontrent dans le texte dans une fenêtre de taille définie.

Cependant, il faut noter que la qualité de la représentation d'un mot est liée à sa fréquence. En effet, un mot apparaissant énormément dans le texte a plus de cooccurrences, ce qui favorise une représentation précise et dense. À l'inverse, la représentation d'un mot peu fréquent est aléatoire, car creuse (c'est-à-dire que de très nombreuses dimensions vaudront 0) et par conséquent, moins fiable. Dans l'idée de réduire l'impact que peut avoir la fréquence des mots, les vecteurs peuvent être normalisés à l'aide de différentes mesures d'association. On peut notamment citer la *Pointwise Mutual Information* (Fano, 1961) ou le *Log-likelihood* (Dunning, 1993). Voir les travaux de Laroche and Langlais (2010) pour une comparaison de ces mesures en pratique.

Ces représentations en sac de mots seront des vecteurs de grandes dimensions, particulièrement dans les corpus généraux qui peuvent être composés de centaines de milliers de mots uniques. Ces vecteurs sont donc également très creux, car de nombreux mots ne cooccurreront pas avec d'autres. Par contre, un des grands avantages de cette méthode

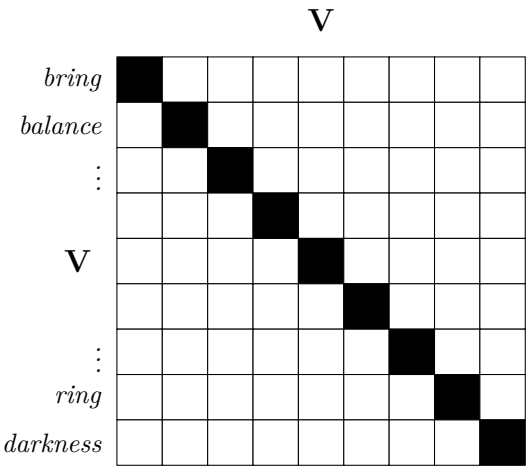

FIGURE 2.2 – Représentation d'une matrice de cooccurrences pour un corpus de taille de vocabulaire **V**.

est que chacune des dimensions d'un vecteur représente un concept précis (un mot du vocabulaire), ce qui n'est pas le cas pour les plongements de mot que nous allons présenter dans la partie suivante.

2.2 Plongements de mots statiques

Les représentations en sac de mots présentent de nombreux avantages : fonctionnement simple à comprendre et à mettre en place, les différentes dimensions qui composent les représentations vectorielles sont compréhensibles et sont la représentation d'un concept clair (chaque dimension représente dans les faits un mot du vocabulaire). Cependant, les sacs de mots peuvent devenir très difficiles à créer et à utiliser lorsque les corpus manipulés dépassent une certaine taille (Jakubina and Langlais, 2016). Les vecteurs étant de la taille du vocabulaire, ils peuvent ainsi atteindre des centaines de milliers de dimensions, rendant les temps de calculs prohibitifs.

Des approches basées sur des réseaux de neurones, qui permettent l'obtention de vecteurs de plus petites dimensions, les rendant ainsi plus facilement manipulables, ont vu le jour. Les deux approches les plus importantes basées sur les réseaux de neurones sont introduites par Mikolov et al. (2013a). Ces nouvelles architectures nommées *Continuous Bag-of-Words* (CBOW) et Skip-gram sont basées sur le principe qu'on peut prédire un mot à l'aide de son contexte ou inversement un contexte à l'aide d'un mot. Dans cette partie, nous décrivons les modèles CBOW et Skip-gram ainsi qu'une de leur principale

extension : *fastText* (Bojanowski et al., 2017).

2.2.1 Plongements de mots CBOW et Skip-gram

Mikolov et al. (2013a) introduisent d'abord l'architecture CBOW. Comme illustré par la Figure 2.3, l'architecture prend en entrée les mots d'un contexte et cherche à prédire le mot central correspondant.

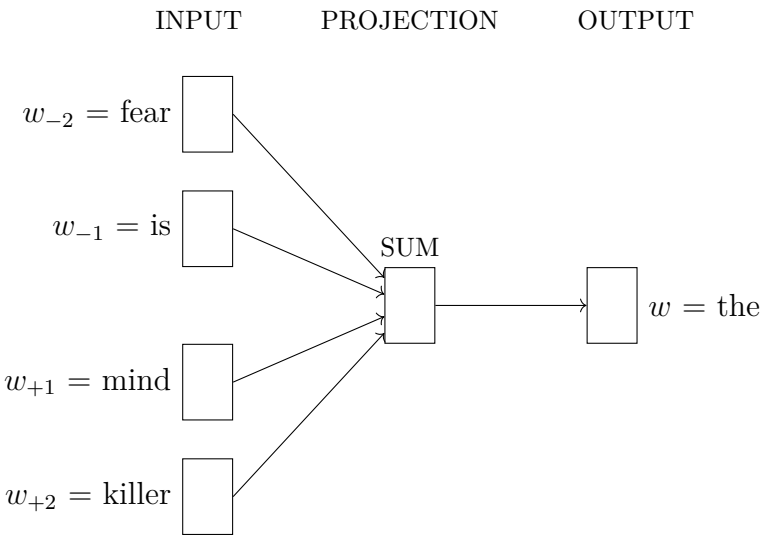

FIGURE 2.3 – Représentation de l'architecture CBOW qui prédit un mot w basé sur son contexte w_c (ici $c \in \{-2, -1, 1, 2\}$). Le mot central "the" va être prédit grâce aux mots de son contexte : "fear", "is", "mind" et "killer". Figure basée sur les travaux de Mikolov et al. (2013a).

Ensuite, Mikolov et al. (2013a) proposent le modèle Skip-gram, qui, à l'inverse de CBOW va se servir du mot central pour prédire les mots de son contexte (voir Figure 2.4).

Les deux approches permettent d'entraîner des plongements de mots très semblables. Dans les faits, CBOW sera plus adapté à des grandes quantités de données alors que Skip-gram sera plus adapté aux petits corpus. Comme l'architecture des deux approches est très similaire, nous ne détaillons que l'approche Skip-gram par la suite. L'objectif d'entraînement du modèle est donc d'apprendre à prédire les mots du contexte à partir d'un mot central. Il peut être défini ainsi pour une séquence de mot $\{w_1, w_2 ... w_T\}$:

$$O = \frac{1}{T} \sum_{i=1}^{T} \sum_{c=-f, c \neq 0}^{f} \log p(w_{i+c} | w_i) \tag{2.1}$$

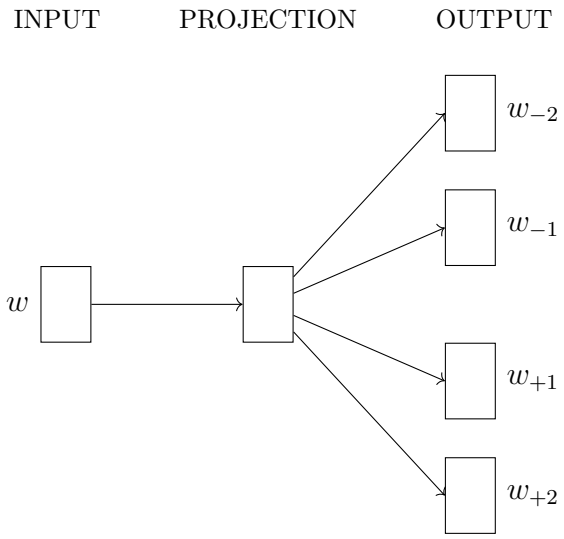

FIGURE 2.4 – Représentation de l'architecture Skip-gram qui prédit les mots du contexte w_c (ici $c \in \{-2, -1, 1, 2\}$) à partir du mot central w. Figure basée sur les travaux de Mikolov et al. (2013a).

où f est la taille de la fenêtre du contexte. Cependant, calculer le gradient de $p(w_{i+c}|w_i)$ est coûteux, car il augmente avec la taille du vocabulaire. Mikolov et al. (2013c) proposent donc l'utilisation du *negative sampling* pour réduire ces temps de calcul. L'idée est de réussir à différencier les bons exemples (le mot central et un des mots de son contexte) des exemples négatifs, le mot central et un mot aléatoire du vocabulaire). La fonction objectif est donc la suivante :

$$O = \log \sigma(u_{w_i}^T v_{w_c}) + \sum_{j \sim P(i)} \log \sigma(-u_{w_j}^T v_{w_c}) \tag{2.2}$$

où u et v sont les représentations "output" et "input" des mots w_i et w_c qui représentent respectivement un mot central et un mot de son contexte (voir Figure 2.4). σ indique la fonction sigmoïde $\sigma(x) = 1/(1 + \exp^{-x})$. Les j exemples négatifs sont tirés d'une distribution de probabilité $P(i)$ basée sur la fréquence des mots dans le corpus.

Une caractéristique intéressante de ces plongements de mots est qu'ils permettent d'obtenir des relations simples entre les mots. Par exemple, si l'on prend les vecteurs des mots *king*, *queen*, *woman* et *man*, il est possible de retrouver la relation suivante : $\overrightarrow{king} + \overrightarrow{woman} - \overrightarrow{man} = \overrightarrow{queen}$. De la même manière, calculer $\overrightarrow{Madrid} + \overrightarrow{Spain} - \overrightarrow{France}$ devrait nous donner un vecteur plus proche de $\overrightarrow{Paris}$ que de tout autre mot (Mikolov et al., 2013c,d).

2.2.2 Plongements de mots *fastText*

Les modèles CBOW et Skip-gram ne prennent pas avantage de la morphologie des mots. En effet, chaque mot est représenté par un vecteur qui n'aura été entraîné qu'en fonction de son contexte. Ainsi, la représentation du mot *luminosité* ne sera lié à celle du mot *lumineux* que par les similitudes liées à leurs contexte alors que les deux mots possèdent une racine commune *lumin-* qui désigne un seul et même concept.

Dans l'idée d'introduire une information morphologique, Bojanowski et al. (2017) proposent l'architecture *fastText* qui se base sur la mécanique des sous-mots. On garde alors le principe des modèles CBOW et Skip-gram, mais chaque mot est alors décomposé en sous-mots. Un sous-mot est une sous-partie d'un mot : par exemple *foun*, *ndat* et *oundatio* sont des sous-mots de *foundation*. Dans *fastText*, chaque mot du vocabulaire est donc composé d'un ensemble de sous-mots qui possèdent chacun leur propre représentation vectorielle. Avant de créer l'ensemble des sous-mots représentatifs du mot d'origine, les caractères $<$ et $>$ sont ajoutés au début et à la fin du mot, pour distinguer les préfixes et les suffixes d'autres chaînes de caractères : ainsi les sous-mots *<son (sonner)*, *son (raisonner)*, *son> (maison)* et *<son> (son)* auront tous une représentation différente.

Dans son implémentation d'origine, *fastText* crée le vecteur final à l'aide de tous les sous-mots de taille minimale 3 et de taille maximale 6 ainsi que le mot complet (toujours accompagné des marqueurs $<$ et $>$). Ainsi, la représentation vectorielle finale du mot est obtenue en sommant les vecteurs de tous les sous-mots qui le composent.

La fonction d'entraînement reste la même, mais pour tout mot w, on obtiendra son vecteur ainsi :

$$u_w = \sum_{g \in G_w} z_g \tag{2.3}$$

où G_w représente l'ensemble des sous-mots g composants le mot w, z quant à lui est la représentation vectorielle des différents sous mots.

Les plongements de mots *fastText* permettent aussi de gérer les mots hors vocabulaire (OOV). De la même manière que pour créer la représentation d'un mot appartenant au vocabulaire du corpus, il suffit de sommer les vecteurs des sous-mots d'un OOV pour obtenir son plongement.

Les auteurs proposent de plus de nombreux ensembles de plongements pré-entraînés sur plusieurs centaines de langues [1].

1. https://fasttext.cc/

2.3 Plongements de mot contextuels

Les plongements de mots présentés précédemment sont appelés statiques, car, le vecteur représentant un mot est commun à toutes ses occurrences, quel que soit le contexte. Ainsi, pour les mots polysémiques, le plongement de mot obtenu sera moins fidèle, étant donné qu'il aura été créé à partir de plusieurs types de contextes différents. Par exemple, le mot *sein* dans un contexte scientifique n'aura pas la même signification qu'*au sein de* la locution sus-citée.

De nouvelles approches cherchent donc à proposer des plongements qui s'adaptent en fonction des contextes de chacune des occurrences d'un même mot. Dans cette partie, nous introduisons deux de ces méthodes, ELMo et BERT.

2.3.1 Plongements de mots ELMo

Les plongements de mots du modèle ELMo (*Embeddings from Language Models*) sont créés à partir de l'entièreté d'une phrase et vont donc varier en fonction du contexte. Peters et al. (2018) proposent d'apprendre les plongements à partir de deux LSTM bidirectionnels et de la combinaison des couches cachées de ces derniers. Un premier modèle de langue calcule la probabilité qu'une séquence de mots soit suivie par tel mot (vue avant), le second est similaire, sauf qu'il lit la séquence de mots en sens inverse (vue arrière). Pour une phrase de taille N $(t_1, t_2, ...t_N)$, on a pour le modèle de langue "vue avant" :

$$p(t_1, t_2, ...t_N) = \prod_{k=1}^{N} p(t_k | t_1, t_2 ... t_{k-1}) \tag{2.4}$$

et pour celui "vue arrière" :

$$p(t_1, t_2, ...t_N) = \prod_{k=1}^{N} p(t_k | t_{k+1}, t_{k+2} ... t_N) \tag{2.5}$$

Pour chacun des mots, une représentation statique x_k est d'abord envoyée à travers les différents biLSTM (ici deux couches). Pour chacune des couches j des biLSTM, on obtient deux représentations $\overrightarrow{h}_{k,j}$ (pour la vue avant) et $\overleftarrow{h}_{k,j}$ (pour la vue arrière). La dernière couche sert à prédire le mot suivant (t_{k+1}) ou précédent (t_{k-1}) à l'aide d'une fonction Softmax. Le biLSTM combine les modèles de langue avant et arrière et cherche à maximiser le logarithme de la probabilité dans les deux directions :

$$O = \sum_{k=1}^{N} (\log p(t_k|t_1, t_2 ... t_{k-1}); \Theta_x, \overrightarrow{\Theta}_{LSTM}, \Theta_s) \\ + \log p(t_k|t_{k+1}, t_{k+2} ... t_N); \Theta_x, \overleftarrow{\Theta}_{LSTM}, \Theta_s)) \quad (2.6)$$

Les paramètres des couches de la représentation statique Θ_x et du Softmax Θ_s sont communs aux vues avant et arrière.

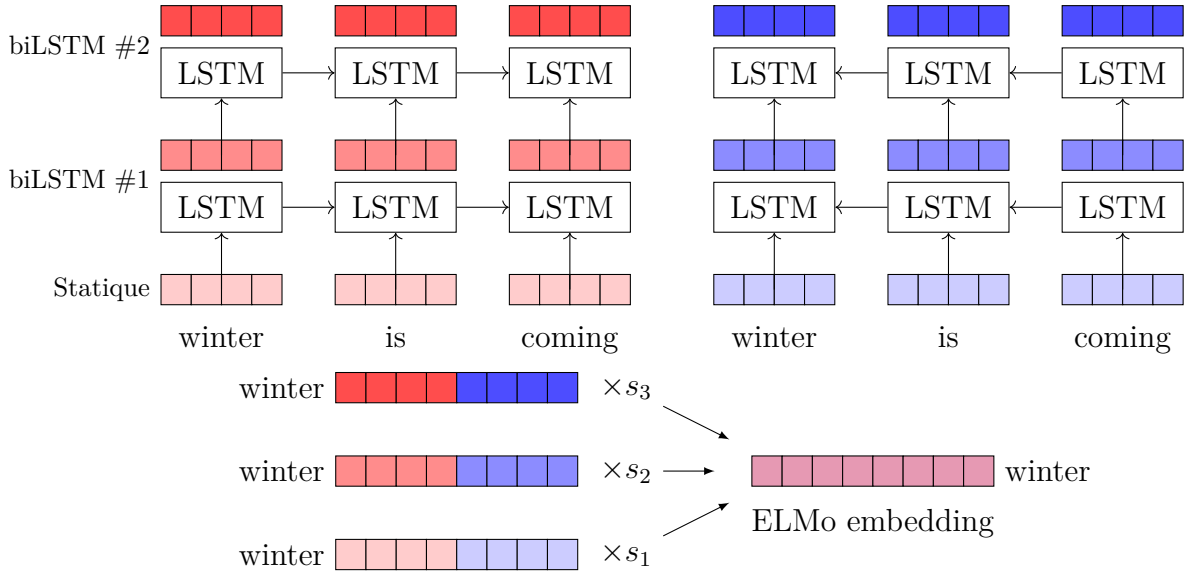

FIGURE 2.5 – Obtention des plongements de mots à partir de l'architecture ELMo.

Pour obtenir la représentation finale d'un mot, Peters et al. (2018) combinent d'abord les plongements d'un mot pour chacune des couches et proposent ensuite de combiner les plongements de chacune des couches en un vecteur final (pondérés par des poids s, voir la Figure 2.5) ou dans une version plus simple de ne prendre que la couche finale (biLSTM #2). Reimers and Gurevych (2019) étudient de nombreux scénarios pour construire des plongements adaptés à une tâche précise et démontrent que la couche finale présente des résultats relativement médiocres pour la plupart des tâches. Dans la majorité des scénarios étudiés, les plongements du premier biLSTM (biLSTM #1) obtiennent les meilleurs résultats et ne sont que rarement améliorés (et par des gains assez minimes) par des combinaisons à base de concaténation ou de moyenne pondérée des différentes couches.

2.3.2 Plongements de mots BERT

Le modèle BERT (Devlin et al., 2019), basé lui sur des modèles de type transformers (Vaswani et al., 2017), permet aussi la création de plongements de mots contextualisés. Alors qu'ELMo utilise deux modèles de langue unidirectionnels (l'un lit la séquence dans le sens normal de lecture, l'autre dans le sens inverse) pour apprendre les plongements de mots, BERT propose un premier objectif d'apprentissage : le *Masked Language Model*. Ce modèle va masquer de manière aléatoire des mots de la séquence et l'objectif du système est de réussir à retrouver ces mots masqués à l'aide du contexte. Le deuxième objectif d'apprentissage de BERT est le *Next Sentence Prediction* qui comme son nom l'indique cherche à prédire si une phrase est réellement la suite d'une autre phrase.

De la même manière qu'ELMo, il est possible d'adapter BERT à des tâches spécifiques. Devlin et al. (2019) améliorent d'ailleurs l'état de l'art sur 11 tâches de traitement du langage. La Figure 2.6 montre l'architecture entre l'entraînement de BERT (*pre-training*) et l'adaptation à une tâche spécifique (*fine-tuning*). La seule différence entre les deux systèmes est la couche de sortie.

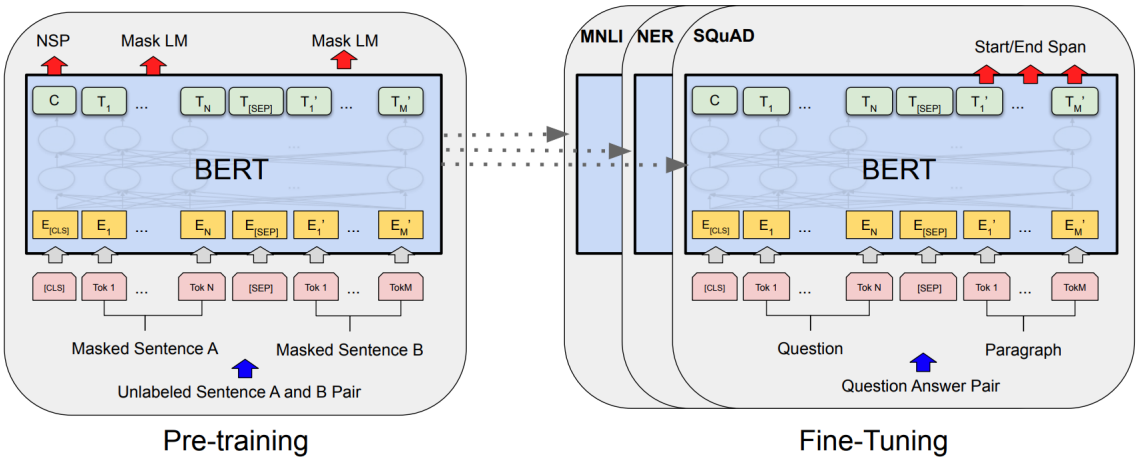

FIGURE 2.6 – Représentation de l'architecture BERT pour l'entraînement et l'adaptation. Figure issue de l'article de Devlin et al. (2019, p. 4173).

De nombreuses variantes de BERT ont été étudiées : Roberta qui est une version très optimisée de BERT (Liu et al., 2019), CamemBERT qui est une version française (Martin et al., 2020), etc.

2.4 Synthèse

Dans ce chapitre, nous avons d'abord défini les représentations en sacs de mots, basées sur l'hypothèse distributionnelle. Les vecteurs créés par cette approche sont de grandes dimensions, très creux et donc difficiles à manipuler. Cependant, chaque dimension représente un concept précis, ce qui les rend facilement explicables.

Ensuite, nous avons étudié les approches de plongements de mots statiques (aussi appelés *word embeddings*), comme Skip-gram et *fastText*, qui, contrairement aux sacs de mots, proposent des vecteurs denses et de petites dimensions. Par contre, la réduction des dimensions rend difficile d'expliquer les significations de chacune d'entre elles.

Finalement, nous avons présenté les plongements de mots contextualisés qui, à l'inverse des deux types de représentations précédents, crée des représentations différentes pour chacune des occurrences d'un même mot, propre à son contexte.

Chapitre 3

Méthodes d'alignement et entraînement conjoint

S'il est important d'obtenir des représentations sémantiques de qualité, dans le cadre de la tâche de BLI, il est nécessaire de pouvoir les comparer entre les langues. Les méthodes de création de plongements de mots que nous avons évoquées dans le chapitre précédent étant adaptées pour des entraînements monolingues, elles créent donc des représentations faisant partie d'espaces vectoriels différents. Dans ce chapitre, nous présentons deux types de famille d'approches permettant une comparaison entre les langues.

D'abord, les méthodes d'alignement qui utilisent les plongements ou sacs de mots présentés précédemment, et les alignent (ou projettent) dans un espace commun. Cet espace peut être un des deux espaces originel (on projettera ainsi un des espaces dans l'autre) ou un tout nouvel espace. Les méthodes d'alignement avec les plongements de mots sont basées sur le principe d'isomorphisme, qui suppose que des espaces vectoriels vont posséder une répartition structurelle similaire entre les différentes langues.

Ensuite, les méthodes d'entraînement conjoint (*joint-training*) qui ont pour objectif d'entraîner simultanément et à base d'un même système les représentations sémantiques des deux langues. Pour cela, les méthodes vont généralement passer par la création d'un corpus bilingue et utiliser des méthodes classiques de création de plongements de mots, ou alors adapter l'objectif d'apprentissage pour y intégrer une dimension bilingue.

Dans ce chapitre, nous présentons d'abord l'approche standard qui est à l'origine de la tâche de BLI et utilise les représentations basées sur les sacs de mots. Puis, nous décrivons le principe d'isomorphisme qui a servi d'hypothèse de base aux approches d'alignement utilisant des plongements de mots statiques que nous présentons par la suite, ainsi qu'une méthode d'adaptation des plongements de mots contextualisés. Nous nous intéressons ensuite aux approches d'entraînement conjoint et pour conclure, nous discutons et définissons le problème d'*hubness*, phénomène touchant les espaces à haute dimensionnalité et qui se traduit par la création d'agglomérats de points très proches les uns des autres,

ce qui peut poser problème en BLI.

3.1 Approche standard

Rapp (1995) et Fung (1998) introduisent l'approche standard en extraction de lexiques bilingues, basée sur les représentations en sacs de mots. Cette approche est basée sur l'idée qu'un mot dans une langue A possède une traduction dans une langue B avec des cooccurrences similaires. Cette méthode projette les vecteurs de contexte des deux langues dans un espace commun à l'aide d'un lexique bilingue d'entraînement. Cette projection permet la comparaison entre les espaces vectoriels des deux langues.

Comme défini en Section 2.1, chacune des dimensions des vecteurs de contexte correspond au poids établi entre le mot que le vecteur de contexte représente et les mots avec lesquels il cooccurre. Il est donc possible, pour chacune des dimensions, de la "traduire" à l'aide d'un dictionnaire. L'approche standard regarde, pour chacune des dimensions du vecteur de contexte, si le mot correspondant est présent dans un dictionnaire bilingue. Si c'est le cas, le poids correspondant dans le vecteur source et correspondant à un mot source, est projeté dans un nouveau vecteur dans l'espace cible et correspondant à la traduction du mot. Dans le cas où le dictionnaire possède plusieurs traductions pour un même mot, le poids est réparti de manière pondérée en fonction de la fréquence des traductions dans le corpus cible. En cas d'absence dans le dictionnaire, la dimension correspondante est simplement supprimée du vecteur.

Après avoir effectué cette projection de l'espace source vers l'espace cible pour tous les mots du vocabulaire source, il est possible de comparer les vecteurs à l'aide d'une mesure de similarité telle que la similarité cosinus.

La Figure 3.1 représente le processus de projection d'un vecteur d'une langue à une autre. Le dictionnaire français/anglais considéré ici est tel que :

Français	**Anglais**
énergie	power
énergie	energy
désarmement	disarmament
centrale	plant
arme	weapon

Le mot *énergie* possédant plusieurs traductions, son poids est réparti entre elles (ici on considère que les deux mots ont une fréquence d'apparition égale).

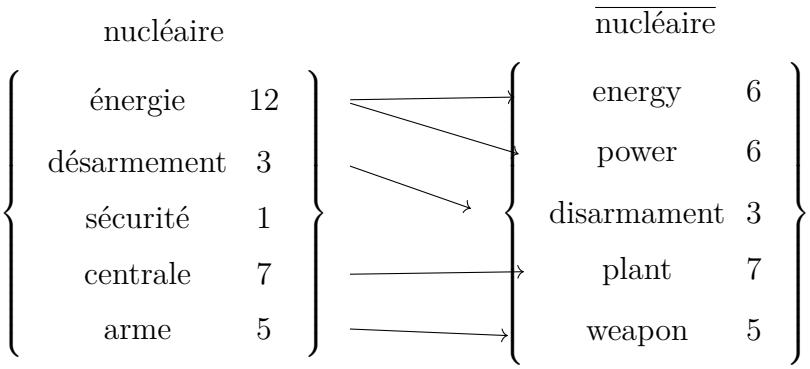

FIGURE 3.1 – Projection du vecteur de *nucléaire* dans l'espace cible anglais. Pour obtenir la traduction du mot, on cherchera dans l'espace cible le vecteur le plus proche de $\overline{\text{nucléaire}}$. Les nombres correspondent au poids attribué à la relation entre le mot de départ (nucléaire) et ses cooccurrences.

3.2 Isomorphisme

Pour les plongements de mots, il n'est pas possible d'aligner chacune des dimensions à l'aide d'un dictionnaire bilingue, car elles n'ont pas d'équivalents à travers les langues. Cependant, il est possible d'utiliser le principe d'isomorphisme.

En théorie des graphes, deux graphes isomorphes sont deux graphes structurellement similaires pour lesquels il existe une bijection qui préserve les arêtes. Plus simplement, deux graphes sont isomorphes s'ils ont le même nombre de sommets et que ces sommets sont connectés de manière semblable. Dans le cadre d'espaces vectoriels de grande taille, une version simplifiée consiste à vérifier si les graphes des k plus proches voisins sont isomorphiques (Søgaard et al., 2018).

Les graphiques de la Figure 3.2 montrent une répartition structurelle similaire entre des mots anglais et leur traduction française. Cette similarité entre les deux espaces permet d'imaginer un transfert d'un espace vers l'autre qui permettrait pour chacun des mots d'avoir en plus proche voisin sa traduction. De nombreux travaux utilisant ce principe seront présentés dans la section suivante.

Cependant, Søgaard et al. (2018) démontrent que l'isomorphisme est en fait loin d'être atteint dans la majeure partie des paires de langues. Ils définissent plusieurs paramètres qui jouent en faveur (ou défaveur) de cette hypothèse et font que les méthodes utilisant l'isomorphisme ont des résultats plus ou moins intéressants.

D'abord, la proximité de la paire de langue étudiée est à prendre en compte (voir aussi (Patra et al., 2019)). En effet, dans la plupart des travaux, les études sont effectuées sur

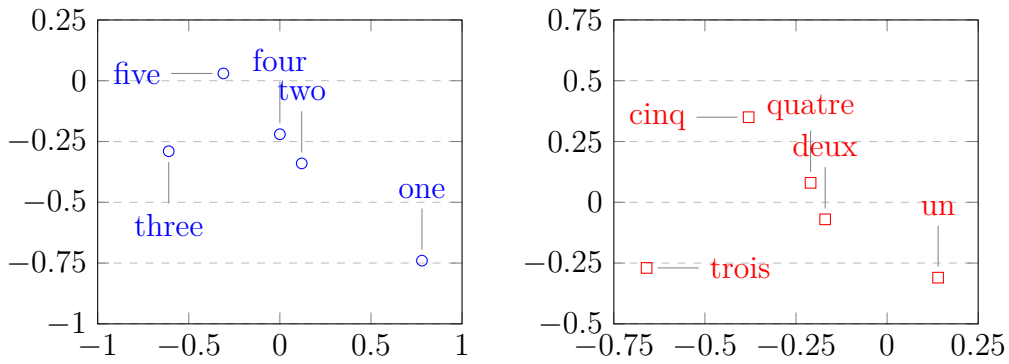

FIGURE 3.2 – Représentation vectorielle créée artificiellement de quelques nombres en anglais (à gauche) et français (à droite). Figure basée sur les travaux de Mikolov et al. (2013b).

des langues comme l'allemand, l'anglais, l'espagnol ou le français, qui, en plus d'être très fournies en données, sont relativement proches. Et si l'isomorphisme de ces espaces est déjà relativement faible, il est néanmoins suffisant pour que les méthodes puissent obtenir des résultats acceptables. Cependant, Søgaard et al. (2018) démontrent que des langues plus éloignées comme l'estonien ou le finnois présentent de très mauvais résultats une fois associées à l'anglais.

Ensuite, la comparabilité des corpus est présentée comme ayant un impact important. Pour trois corpus chacun disponibles en différentes langues, (Søgaard et al., 2018) étudient trois paires de langues différentes : anglais-{espagnol, finnois, hongrois}. Ils démontrent que les résultats sont fondamentalement différents selon que les deux langues utilisent ou non les mêmes corpus.

De la même manière, plus les hyper-paramètres (par exemple, type de plongements de mots, taille de la fenêtre de contexte, etc.) utilisés seront similaires, plus les résultats seront bons. En prenant par exemple deux méthodes d'apprentissage différentes, par exemple CBOW pour une langue et Skip-gram pour l'autre, les résultats obtenus sont très mauvais.

Finalement, pour certaines paires de langues, en jouant sur la taille des vecteurs, Søgaard et al. (2018) montrent que l'augmenter permet d'améliorer les résultats alors que pour d'autres, les performances peuvent grandement diminuer. Les méthodes basées sur l'isomorphisme sont donc grandement dépendantes des scénarios dans lesquels elles sont utilisées, phénomène qui a tendance à s'aggraver avec les méthodes non supervisées.

3.3 Alignement des espaces vectoriels

Les représentations vectorielles de mots à l'aide de réseaux de neurones présentent de nombreux avantages, cependant, il n'est pas possible de traduire les dimensions à l'aide d'un dictionnaire de la même manière que pour une représentation en sacs de mots. Mikolov et al. (2013b) proposent l'utilisation d'une matrice de traduction, qui permet de projeter l'espace vectoriel d'une langue dans celui de l'autre. Dans cette partie, nous définissons différentes méthodes d'alignement et les améliorations qui ont suivi.

3.3.1 Alignement supervisé

Mikolov et al. (2013b) proposent de tirer profit du principe d'isomorphisme pour permettre la comparaison entre deux espaces vectoriels de langues différentes. Ainsi, en apprenant une matrice dite de traduction à l'aide d'un dictionnaire d'apprentissage, il est possible d'obtenir une transformation linéaire (une rotation accompagnée d'un changement d'échelle) qui favorisera le rapprochement des mots en relation de traduction. Par exemple, à l'aide des paires *one-un* et *four-quatre*, on peut apprendre les traductions des autres chiffres.

L'objectif est donc d'apprendre une matrice de traduction W permettant ce transfert entre les espaces vectoriels en minimisant l'erreur quadratique moyenne entre les plongements de mots sources x, transformés à l'aide de W, et cibles y :

$$O = \sum_{i=1}^{n} \|Wx_i - y_i\|^2 \tag{3.1}$$

où x_i et y_i sont les représentations vectorielles d'un mot source et de sa traduction (obtenue à l'aide du dictionnaire d'alignement de taille n). On peut aussi représenter cette équation sous sa forme matricielle :

$$O = \|WX - Y\|_F^2 \tag{3.2}$$

où X et Y sont les matrices des mots sources et cibles et $\|_F$ désigne la norme de Frobenius.

Une fois la matrice de traduction W entraînée, il est possible de projeter un mot source x dans la langue cible en calculant $y = Wx$. Ensuite, il reste alors à trouver le plus proche voisin de y à l'aide d'une mesure de similarité pour obtenir la traduction de x. Ce type de méthode est souvent classée dans les méthodes dites de régression. Dinu et al. (2014) proposent aussi l'ajout d'une phase de normalisation des plongements de mots.

Xing et al. (2015) montrent que cette méthode peut être améliorée en forçant la matrice de traduction W à être orthogonale (soit $WW^T = I$). Avec cette contrainte, on obtient $W = VU^T$ avec $X^TY = U\Sigma V^T$ qui représente la décomposition en valeurs singulières (*singular value decomposition*) de X^TY. Cette contrainte est aussi appuyée par Artetxe et al. (2016) qui démontrent qu'elle permet de garder une cohérence monolingue des différents espaces. Ces méthodes sont dites orthogonales.

On note aussi les méthodes canoniques, basées sur l'analyse canonique des corrélations (*Canonical Correlation Analysis*, CCA). Faruqui and Dyer (2014) proposent dans un premier temps l'utilisation de la CCA pour aligner les deux espaces dans un nouvel espace commun, cette méthode a ensuite été enrichie dans un scénario multilingue (Ammar et al., 2016) utilisant l'anglais comme pivot. Finalement, les méthodes dites de marge cherchent à maximiser la marge entre un mot et sa traduction réelle par rapport aux autres candidats (Lazaridou et al., 2015).

Si la plupart des approches présentées précédemment s'appuient sur des dictionnaires d'alignement comportant plusieurs milliers de paires de mots, Artetxe et al. (2017) réduisent cette contrainte à des dictionnaires de 25 paires ou encore proposent d'utiliser les nombres présents dans les deux corpus en tant que paires (le chiffre 1 en langue source associé avec le chiffre 1 en langue cible par exemple), pour ensuite renforcer ce premier pseudo-dictionnaire grâce à une procédure d'auto-apprentissage. Comme indiqué en Figure 3.3, la méthode reprend les principes de base des méthodes d'alignement, mais, au lieu de s'arrêter après une première étape, elle réutilise le dictionnaire obtenu pour recommencer une procédure d'alignement.

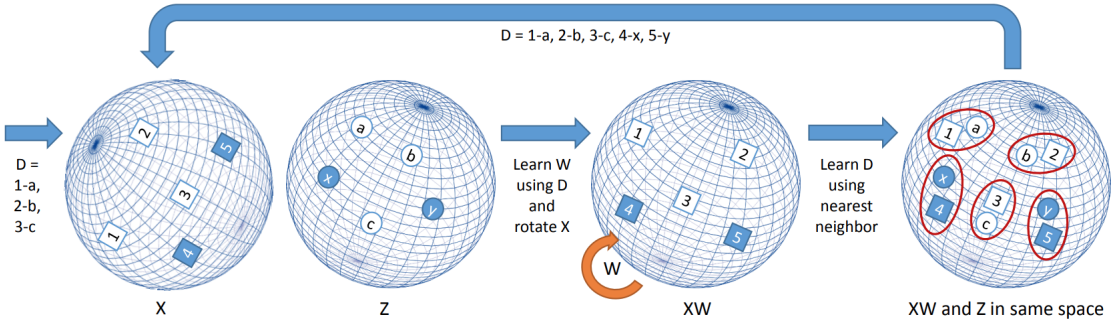

FIGURE 3.3 – Schéma d'une approche d'alignement contenant de l'auto-apprentissage. Les approches ne contenant pas l'auto-apprentissage s'arrêtent après un premier alignement, alors qu'ici, on utilise les nouvelles paires créées pour chercher à améliorer l'alignement. Figure issue de l'article de Artetxe et al. (2017, p. 452).

3.3.2 Alignement non supervisé

L'intérêt envers des approches nécessitant de moins en moins de données pour accomplir l'alignement des espaces a fini par ouvrir la porte à des approches non supervisées n'utilisant aucune donnée bilingue.

Conneau et al. (2017) proposent d'utiliser l'apprentissage contradictoire (*adversarial training*) pour obtenir un dictionnaire bilingue artificiel qui sert d'initialisation à une méthode d'alignement plus classique. Cette approche se base toujours sur l'apprentissage d'une matrice de traduction W. Pour cela, la matrice sera entrainée de manière à rendre WX (X étant la matrice source) et Y (la matrice cible) aussi similaire que possible. Cette matrice va alors prendre les plongements de mots sources et les transformer de manière à réussir à tromper un discriminateur dont l'objectif sera de retrouver l'espace d'origine des plongements de mots qu'on lui présente. Une fois cet apprentissage terminé, un dictionnaire est créé à partir des mots les plus fréquents et qui sont respectivement plus proches voisins. S'ensuit une phase d'auto-apprentissage à la manière de Artetxe et al. (2017). La Figure 3.4 représente cette procédure.

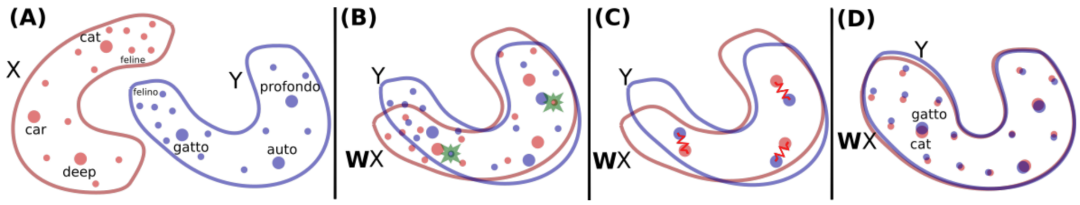

FIGURE 3.4 – (A) représente les deux espaces originels. (B) à l'aide d'adversarial training, on apprend une matrice W qui aligne les deux espaces, les étoiles vertes correspondent à des mots pour lesquels le discriminateur est chargé de déterminer la langue d'origine. (C) une fois un premier alignement effectué, on extrait des paires de mots pour créer un dictionnaire artificiel qui sert à affiner l'alignement. (D) les deux espaces sont maintenant alignés et il est possible de les comparer. Figure issue de l'article de Conneau et al. (2017, p. 3).

Artetxe et al. (2018b) utilisent le concept d'isomorphisme pour obtenir un premier dictionnaire. En supposant que les espaces des plongements de mots soient parfaitement isométriques, les matrices de similarité $M_x = XX^T$ et $M_y = YY^T$ seront équivalentes en trouvant la bonne permutation de lignes et colonnes qui permettrait de définir le dictionnaire.

3.4 Entraînement conjoint

Les approches par alignement ont longtemps été les plus étudiées, du fait de leur maniabilité. En effet, il est possible, à l'aide d'un dictionnaire (ou non) et de plongements de mots pré-entraînés, d'évaluer une approche par alignement très simplement, sans avoir nécessairement à disposition un corpus d'entraînement. Cependant, ces approches souffrent d'autres problèmes. Par exemple, Søgaard et al. (2018); Patra et al. (2019) démontrent que l'hypothèse d'isomorphisme est en fait peu réaliste pour la plupart des paires de langues, ce qui a tendance à diminuer la qualité de l'alignement. Ormazabal et al. (2019) indiquent que ce problème provient du fait que les espaces sont entrainés séparément et que les approches d'entraînement conjoint permettent l'obtention de plongements de mots plus isométriques.

Gouws and Søgaard (2015) proposent de créer un corpus bilingue en concaténant les deux parties et en traduisant aléatoirement certains mots en fonction d'un dictionnaire de départ. Ainsi, la phrase :

> "*It is important, when killing a nun, to ensure that you bring an army of sufficient size.*" - Mark Lawrence, Red Sister.

pourrait devenir :

> "*It is important, quand killing a nonne, to ensure that you emmène an army of suffisante size.*"

Une fois le corpus bilingue créé, n'importe quelle méthode d'apprentissage de plongements de mots peut être utilisée pour obtenir des plongements bilingues.

Duong et al. (2016) proposent d'améliorer le modèle CBOW en concaténant les deux parties du corpus et en rajoutant dans l'objectif d'apprentissage la prédiction de la traduction $\bar{w}_i$ des mots :

$$O = \alpha \log \sigma(u_{w_i}^T v_{h_c}) + (1-\alpha) \log \sigma(u_{\bar{w}_i}^T v_{h_c}) + \sum_{j \sim P(i)} \log \sigma(-u_{w_j}^T v_{h_c}) \tag{3.3}$$

où α est fixé à 0,5 et contrôle la contribution du mot par rapport à sa traduction et h_c représente le vecteur de contexte. La Figure 3.5 montre la prédiction du mot "always" et de sa traduction "toujours" à l'aide du contexte.

Si dans le cadre d'un mot avec de multiples traductions, Gouws and Søgaard (2015) choisissent une traduction de manière aléatoire, Duong et al. (2016) choisissent une traduction basée sur sa similarité avec le contexte à l'aide d'un algorithme d'espérance-

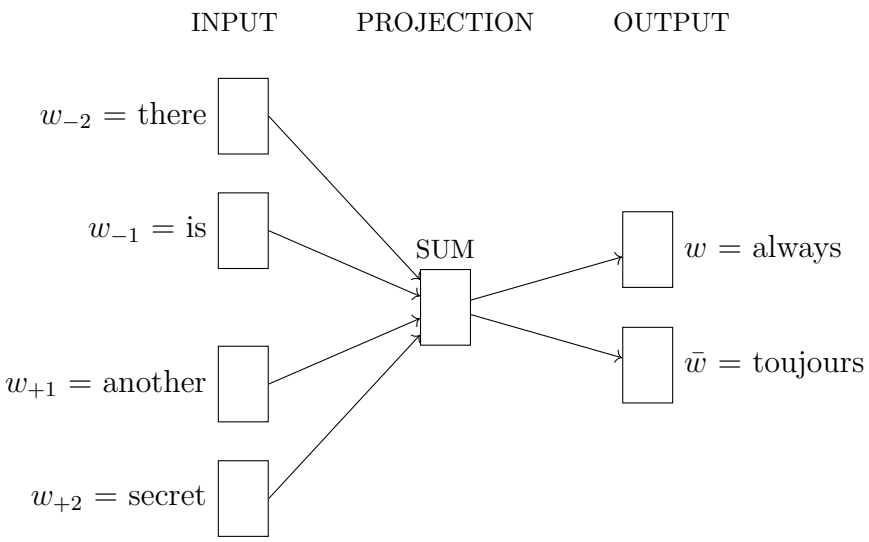

FIGURE 3.5 – Représentation de l'architecture CBOW adaptée par Duong et al. (2016) qui prédit un mot w et sa traduction $\bar{w}$ basé sur son contexte w_c (ici $c \in \{-2, -1, 1, 2\}$). Le mot central "always" et sa traduction "toujours" vont être prédit grâce à ses mots contextes : "there", "is", "another" et "secret".

maximisation. Hakimi Parizi and Cook (2020) ajoutent à cette méthode l'information des sous-mots à la manière de *fastText* (Bojanowski et al., 2017), ce qui, en plus de permettre une amélioration des résultats, permet de traiter les mots hors vocabulaire.

Wang et al. (2020) créent simplement un corpus bilingue en concaténant les deux parties sur lesquelles ils entraînent des plongements de mots *fastText*, puis, proposent d'ajouter une phase d'alignement en procédant à une phase de réallocation du vocabulaire. Pour cela, les mots étant présents dans les deux langues sont répartis dans les vocabulaires en fonction de leur fréquence. Si un mot est bien plus fréquent dans une langue que dans l'autre, il est considéré comme faisant partie du vocabulaire de cette langue. Par contre, si les fréquences d'apparition des deux mots sont similaires, ils sont considérés comme faisant partie des deux vocabulaires et sont donc retirés de la procédure d'alignement pour éviter de dégrader leur représentation. Ormazabal et al. (2021) ajoutent à cette méthode la traduction des mots dans l'objectif d'apprentissage à la manière de Duong et al. (2016) et Hakimi Parizi and Cook (2020).

3.5 Alignement à l'aide de plongements de mots contextuels

Si les plongements de mots contextuels sont porteurs d'une information sémantique plus riche que leurs variantes statiques, les aligner semble plus complexe étant donné qu'il existe de multiples représentations pour un même mot, même monosémique.

Schuster et al. (2019) proposent une solution pour résoudre ce problème et créer un alignement entre les espaces des deux langues. Pour cela, ils construisent des plongements de mots statiques (appelés *anchors* : ancres) à partir des plongements de mots contextuels, qui servent à obtenir un alignement pour l'espace contextuel à partir des méthodes présentées précédemment. Le vecteur ancre d'un mot donné est créé en calculant la moyenne de ses plongements (voir Figure 3.6). Une fois cette étape accomplie, les ancres peuvent être utilisées pour apprendre une matrice d'alignement (n'importe quelle méthode utilisant des plongements de mots statiques pourra faire l'affaire). Cette matrice peut ensuite être appliquée sur les plongements de mots contextuels, car ils sont l'origine des ancres.

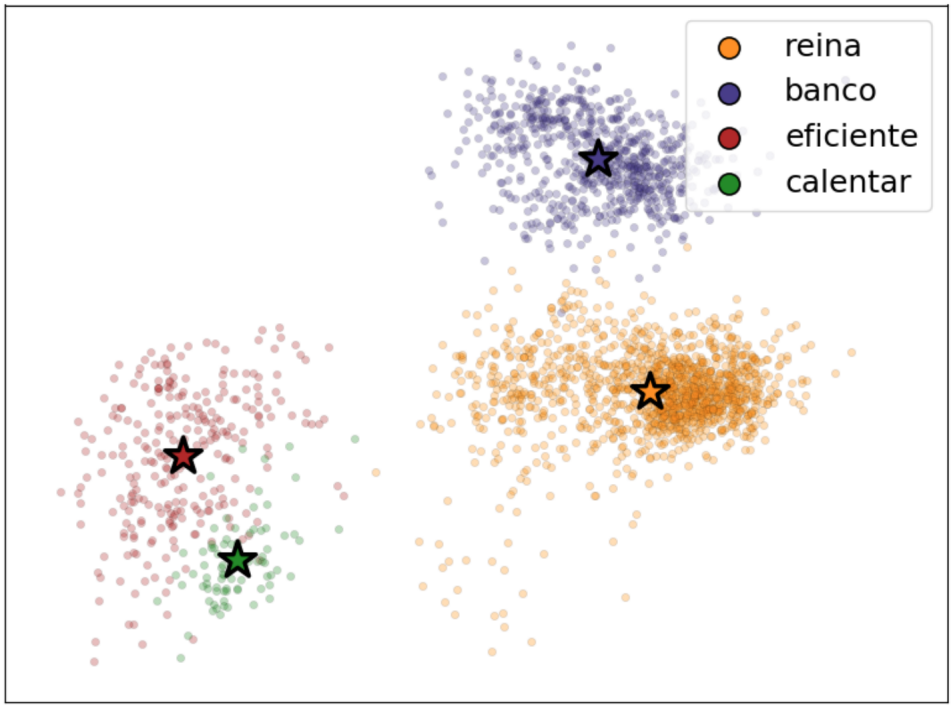

FIGURE 3.6 – Représentation en deux dimensions d'un espace de plongements de mots contextuels. Les étoiles correspondent aux ancres des mots correspondants. Figure issue de l'article de Schuster et al. (2019, p. 1601).

On obtient donc la possibilité de comparer les mots polysémiques à travers les langues. Par exemple, la Figure 3.7 montre que le mot anglais *bear* va avoir deux nuages principaux, correspondants chacun à ses deux traductions espagnoles *oso* et *tener*. Le nuage du mot *bear* correspondant à l'animal va d'ailleurs être proche des ancres d'autres animaux (ici *dog*, *elephant* ou *cat*).

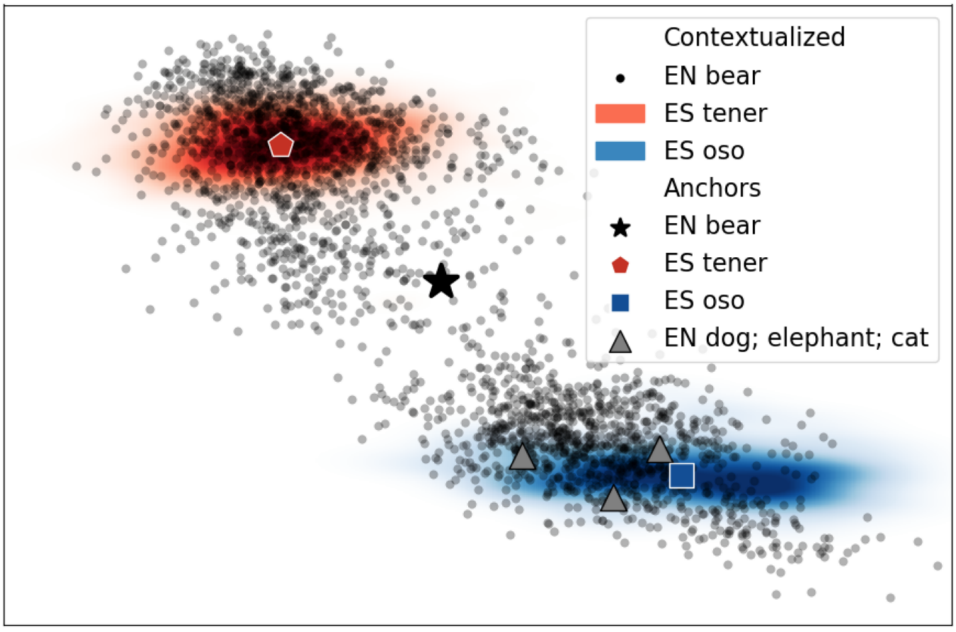

FIGURE 3.7 – Représentation d'un alignement de plongements de mots contextuels créés à partir d'ancres. Figure issue de l'article de Schuster et al. (2019, p. 1602).

Zhang et al. (2019) proposent d'utiliser les plongements de mots contextualisés pour pouvoir traiter les mots polysémiques à l'aide d'algorithmes de clustering.

3.6 *Hubness* & CSLS

Radovanovic et al. (2010) définissent le phénomène d'*hubness*. Dans les espaces vectoriels à haute dimensionalité, certains vecteurs, appelés *hubs*, tendent à être les plus proches voisins de nombreux vecteurs. À l'inverse, d'autres vecteurs apparaissent très rarement dans les plus proches voisins des autres vecteurs, les *antihubs*. Les plus proches voisins étant asymétriques (le fait que x soit dans les k plus proches voisins de y ne veut pas dire que y est dans les k plus proches voisins de x), ce phénomène est encore aggravé. Ce problème a déjà été observé dans de nombreux domaines, comme le traitement de l'image

(Hicklin et al., 2005) ou du son (Aucouturier and Pachet, 2008).

Cependant, dans le cadre de la tâche de BLI, l'espace vectoriel est en fait constitué de deux ensembles (les deux langues) projetés dans un espace commun. Dinu et al. (2014) proposent dans un premier temps de définir la traduction d'un mot source w_s non pas comme le mot cible w_t le plus proche, mais comme le mot cible qui classe w_s au plus haut rang possible. Ensuite, Smith et al. (2017) utilisent un softmax inversé, en mesurant la probabilité qu'un mot cible se traduise en notre mot source de départ. Finalement, Conneau et al. (2017) introduisent la *Cross-Domain Similarity Local Scaling* (CSLS). On considère alors l'espace commun comme un graphe biparti où chaque mot d'une partie du corpus est relié à ses K plus proches voisins dans l'autre langue. On va noter $\mathcal{N}_T(w_s)$ (respectivement $\mathcal{N}_S(w_t)$) l'ensemble contenant les plus proches voisins de w_s (w_t) dans l'espace cible (source). On peut donc représenter la similarité moyenne d'un mot source w_s à ses plus proches voisins dans l'espace cible par :

$$\text{knn}_T(w_s) = \frac{1}{K} \sum_{y_t \in \mathcal{N}_T(w_s)} \cos(w_s, w_t) \tag{3.4}$$

Et de la même manière $\text{knn}_S(w_t)$ défini la similarité moyenne d'un mot cible w_t à ses plus proches voisins dans l'espace source. On peut ensuite utiliser ces fonctions pour définir la mesure de similarité CSLS entre un mot source w_s et un mot cible w_t comme suit :

$$CSLS(w_s, w_t) = 2\cos(w_s, w_t) - \text{knn}_T(w_s) - \text{knn}_S(w_t) \tag{3.5}$$

Cette mesure de similarité a pour effet d'augmenter la similarité des *antihubs*, car $\text{knn}(w)$ sera relativement faible et a l'effet inverse pour les *hubs*. Dans leurs différents tests, Conneau et al. (2017) utilisent une valeur de $k \in \{5, 10, 50\}$ sans noter de différence dans les résultats finaux et proposent donc de fixer $k = 10$.

3.7 Synthèse

Dans ce chapitre, nous avons présenté différents types de méthodes pour obtenir des représentations sémantiques bilingues. Originellement, l'approche standard se basait sur le fait que chacune des dimensions des vecteurs représentaient un mot précis qu'il était donc possible de traduire entre les langues pour permettre la comparaison. Si l'arrivée des plongements de mots a supprimé cette possibilité, des méthodes basées sur le concept

d'isomorphisme et utilisant des matrices de traduction ont vu le jour pour permettre la création de plongements de mots bilingues. S'il a été démontré que l'isomorphisme n'était pas toujours présent pour les paires de langues les plus éloignées pour les approches par alignement, les approches d'entraînement conjoint cherchent à créer des espaces plus isomorphes. Ruder et al. (2019) proposent une analyse poussée de ses différentes méthodes et Glavaš et al. (2019) conduisent une large étude pour comparer les résultats de ces approches.

Nous avons aussi présenté pour les plongements de mots contextualisés une méthode simple pour adapter les approches d'alignement.

Finalement, nous avons introduit la CSLS, une mesure de similarité qui cherche à corriger le problème d'*hubness* : le fait que dans les espaces à haute dimensionnalité, certains mots sont les plus proches voisins de nombreux autres.

Chapitre 4

Augmentation et sélection de données

Dans le Chapitre 1, nous avons recensé les différents types de corpus utilisables en BLI ainsi que les corpus utilisés dans les différents travaux de ce manuscrit. Et, bien que les travaux en BLI se concentrent aujourd'hui sur des corpus comparables, qui sont bien plus faciles à réunir que les corpus parallèles, il reste coûteux de réunir de grandes quantités de documents spécialisés. La réalisation de la tâche de BLI en domaine de spécialité souffre de ce manque de données. En effet, avoir peu de données rend les approches basées sur les cooccurrences des mots moins performantes, car les mots apparaissant peu ont une représentation moins fiable. Cette problématique rend le BLI en domaine de spécialité plus complexe qu'en domaine général, où l'on peut facilement obtenir des corpus composés de centaines de millions de mots. Nous soulignons également que pour certaines langues, il est particulièrement difficile de produire des ressources spécialisées en quantité suffisante (voir à ce sujet la construction du corpus BC en Section 1.3).

Dans l'idée de réduire ce problème, Hazem and Morin (2016, 2018) proposent d'ajouter des données générales au corpus spécialisé de départ pour augmenter le nombre d'occurrences (et donc de cooccurrences) des mots du vocabulaire, dans l'idée d'obtenir de meilleures représentations vectorielles. Si cette solution permet effectivement une amélioration des résultats, l'ajout de données générales a pour conséquence l'introduction de la polysémie dans les données d'entraînement spécialisées, ainsi qu'une augmentation importante des temps de calcul. Dans l'idée de réduire ces deux problèmes, nous cherchons à sélectionner les données intéressantes dans les corpus généraux afin de mettre de côté les données qui risquent de dégrader la qualité des représentations de mots spécialisés.

Dans ce chapitre, nous expliquons dans un premier temps le fonctionnement de l'ajout de données générales (avec l'approche standard, puis avec les plongements de mots) proposé par Hazem and Morin (2016, 2018). Ensuite, nous décrivons les problèmes engendrés par cet ajout. Enfin, nous introduisons notre système de sélection de données spéciali-

sées, basé sur des techniques de sélection de données simples comme la *Cross Entropy* ou le TF-IDF ou plus complexe comme BERT. Ce travail a été l'objet d'une publication à COLING (Laville et al., 2020a).

4.1 Augmentation de données avec l'approche standard

L'idée d'ajouter des données générales en domaine de spécialité a déjà été appliquée en traduction automatique (Moore and Lewis, 2010; Axelrod et al., 2011). En BLI, Li and Gaussier (2010) cherchent à améliorer la qualité des corpus comparables à l'aide de données extérieures, mais sans travailler sur des corpus de spécialité, cette amélioration a pour conséquence d'augmenter la taille des corpus. Pour améliorer les résultats en domaine spécialisé à l'aide de corpus général, Hazem and Morin (2016) reprennent l'approche standard basée sur les représentations en sac de mots et que nous avons présentée en Section 3.1 et proposent deux adaptations utilisant des données générales en plus des corpus spécialisés : la Global Standard Approach (GSA) et la Selective Standard Approach (SSA), que nous décrivons dans les sections suivantes.

4.1.1 Global Standard Approach

La GSA est une adaptation simple à mettre en place de l'approche standard, où l'idée est de concaténer le corpus spécialisé avec le corpus général, puis de calculer les nouveaux vecteurs à partir de ce corpus.

La Figure 4.1 représente la matrice de cooccurrences de la GSA. L'ajout du corpus général induit l'ajout de nombreux nouveaux mots dans le vocabulaire, ce qui a pour conséquence d'avoir bien plus de vecteurs avec une haute dimensionnalité. Par exemple, dans Hazem and Morin (2016) le corpus spécialisé BC est composé de 6 630 mots distincts alors que le corpus général ajouté en comprend 250 999. La matrice de cooccurrences étant bien plus grande, les temps de calcul augmentent de manière significative. Pour les étapes suivantes, la méthodologie reste la même que pour l'approche standard.

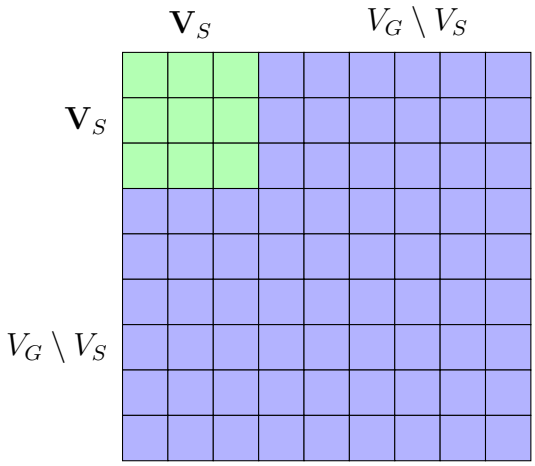

FIGURE 4.1 – Représentation de la matrice de cooccurrences de la GSA. S (respectivement G) représente le domaine spécialisé (général), V_S (V_G) étant la taille du vocabulaire spécialisé (général).

4.1.2 Selective Standard Approach

Pour cette deuxième approche, on construit dans un premier temps les vecteurs de contexte séparément pour les corpus spécialisé et général. Ensuite, avant la phase de normalisation des vecteurs, et pour chacun des mots appartenant au corpus spécialisé apparaissant dans le corpus général, on fusionne les deux vecteurs de contexte. Cette approche peut être formellement décrite comme suit :

$$\forall w \in S, \forall c \in S \cup G, cooc_{dsa}(w,c) = \begin{cases} cooc_S(w,c) + cooc_G(w,c) & \text{si } w \in G \text{ et } c \in G \cap S \\ cooc_S(w,c) & \text{si } w \notin G \text{ ou } c \notin G \\ cooc_G(w,c) & \text{sinon} \end{cases} \tag{4.1}$$

où $cooc(w,c)$ représente le nombre de cooccurrences entre un mot w et un mot de son contexte c, S et G sont les vocabulaires spécialisé et général respectivement. Plus simplement, l'approche additionne les cooccurrences entre w et c dans le corpus spécialisé avec les cooccurrences du corpus général. Dans le cas où un des deux mots n'apparaît pas dans un des corpus, on prend les cooccurrences de l'autre corpus. Cela permet de filtrer les mots n'apparaissant pas dans le corpus spécialisé tout en améliorant la qualité de leur vecteur avec l'ajout de nouvelles cooccurrences. La Figure 4.2 représente les différentes matrices de la SSA et leur combinaison.

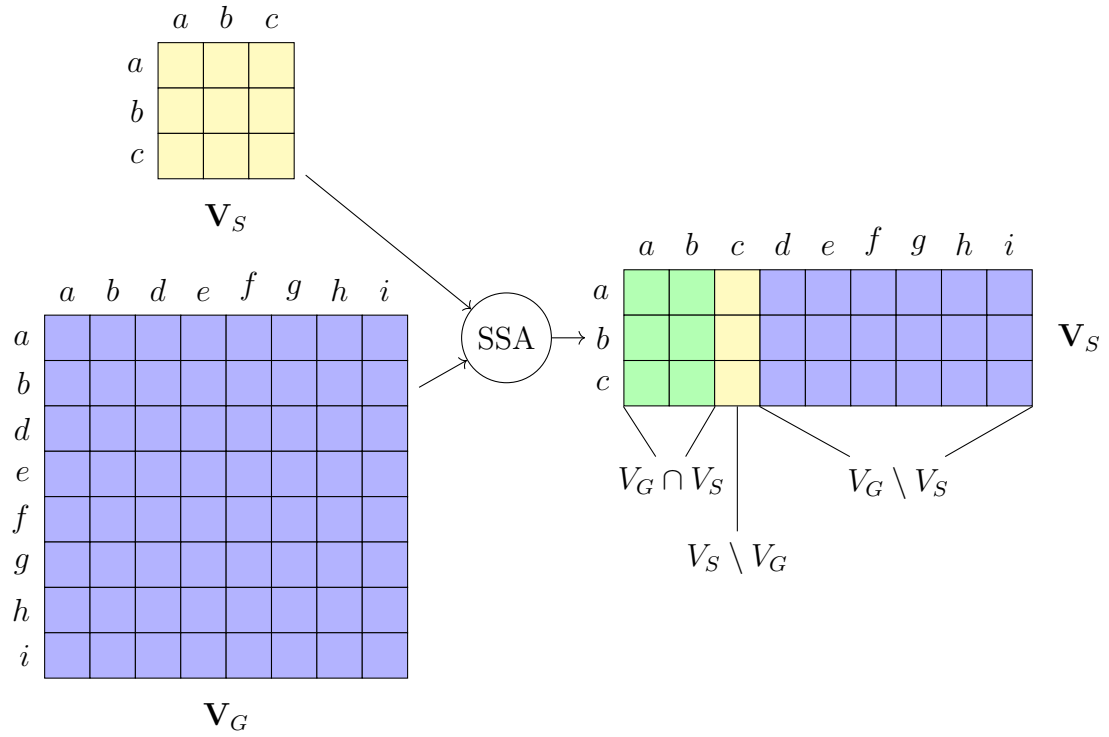

FIGURE 4.2 – Représentation de la combinaison des matrices spécialisée et générale pour créer la matrice finale de la méthode SSA. Dans la matrice finale, les mots a et b sont présents dans les deux vocabulaires et c n'est présent que dans le vocabulaire spécialisé. Ces trois mots sont donc présents sur les deux dimensions. Le reste du vocabulaire ne fait partie que du vocabulaire général et n'est donc présent qu'en une seule des dimensions.

Pour cette méthode, si la matrice de cooccurrences reste plus grande qu'avec le corpus spécialisé seul, on augmentera ici seulement les dimensions des vecteurs et pas leur nombre, ce qui aura pour conséquence, si l'on prend le corpus BC combiné au corpus JRC, d'obtenir une réduction de la matrice de cooccurrences par un facteur d'environ 15 entre la GSA et la SSA (voir le Tableau 1.2). Comme pour la GSA, la procédure pourra ensuite s'effectuer de la même manière que l'approche standard.

4.1.3 Résultats

Nous présentons les résultats des auteurs dans le Tableau 4.1. On note une tendance assez claire : avoir plus de données permet l'obtention de meilleurs résultats. On peut même souligner que les corpus généraux JRC ou CC seuls sont plus intéressants que le corpus spécialisé seul correspondant (ligne SA). Les corpus seuls sont dépassés par la combinai-

	Spécialisé	Général		Spécialisé	Général	
Méthode	BC	JRC	CC	WE	JRC	CC
SA	25,9	53,2	75,8	15,6	63,4	72,1
GSA (Spec. + Gen.)	-	63,3	80,7	-	65,3	73,2
SSA (Spec. + Gen.)	-	66,8	**81,6**	-	67,8	**74,9**

TABLEAU 4.1 – Tableau des résultats (% MAP) de la GSA et la SSA sur la paire de langue anglais-français par rapport à l'approche standard (SA) pour deux corpus spécialisés et leur combinaison avec deux corpus généraux. La ligne SA représente les résultats avec le corpus correspondant seul, les lignes GSA et SSA représentent les résultats avec les corpus combinés comme présenté précédemment. Les approches sont évaluées sur le lexique spécialisé correspondant au corpus. Voir les travaux de Hazem and Morin (2016) pour les résultats complets.

son des corpus généraux et spécialisés (GSA et SSA), méthodes pour lesquelles on note de meilleurs résultats pour la SSA qui obtient en moyenne 2 % de MAP supplémentaire que la GSA. La raison étant que la SSA réduit les candidats possibles en se restreignant au vocabulaire spécialisé, alors que la GSA prend aussi en compte le vocabulaire du corpus général.

4.2 Augmentation de données avec les plongements de mots

Il est aussi possible d'utiliser des données générales pour améliorer les résultats de BLI en domaine spécialisé à l'aide d'approches basées sur les plongements de mot. Ces travaux sont dans la continuité de Hazem and Morin (2018). En concaténant les corpus spécialisé et général, on obtient une manière simple et efficace d'obtenir des plongements de meilleure qualité tout en gardant suffisamment d'informations spécialisées. À partir de ces corpus concaténés, on entraîne des représentations *fastText* séparément pour les deux langues d'intérêt. Ensuite, on peut les aligner à l'aide d'une matrice de traduction, on utilise pour cela l'outil VecMap (Artetxe et al., 2018a).

4.2.1 Résultats

Le Tableau 4.2 présente les résultats que nous avons obtenus en augmentation de données pour une approche par plongements. La ligne *fastText* correspond aux résultats

avec le corpus seul, la ligne Augmentation indique les résultats obtenus avec le corpus général en plus du corpus spécialisé.

	Spécialisé	Général		Spécialisé	Général	
Méthode	BC	JRC	Wiki	WE	JRC	Wiki
fastText	50,6	59,8	82,7	53,4	66,4	69,7
SSA (Spec. + Gen.)	-	66,8	-	-	67,8	-
Augmentation (Spec. + Gen.)	-	81,0	**83,9**	-	68,8	**75,9**

TABLEAU 4.2 – Tableau des résultats (% MAP) de deux approches par plongement de mots sur la paire de langue anglais-français. *fastText* correspond aux résultats avec le corpus correspondant seul, Augmentation représente les résultats en concaténant les corpus spécialisé et général. On ajoute aussi la meilleure ligne des résultats des approches basés sur les sacs de mots (SSA).

On note que les meilleurs résultats sont obtenus grâce à des plongements de mots et l'augmentation de données et qu'une fois de plus, les plus grands corpus présentent les meilleurs résultats. La taille des vecteurs des plongements de mots comparés aux représentations en sacs de mots étant aussi en faveur des approches par plongements de mots, les expériences suivantes seront conduites sur les plongements de mots seuls.

4.3 Augmentation de données : quelles conséquences ?

Si l'ajout de données générales améliore la représentation des mots à l'aide de nouvelles occurrences, ce qui permet d'obtenir de meilleurs résultats, il introduit aussi le problème de polysémie. Par exemple, dans la paire de langue anglais/français dans le corpus BC, un système ne se basant que sur les données spécialisées n'aura pas de soucis à relier entre eux les mots *breast / sein*. Par contre, en ajoutant des données générales, on peut se rendre compte que pour le terme français *sein*, on va ajouter de nombreuses occurrences liées à l'expression *au sein de* (qui peut être traduite par *within*), ce qui aura pour effet d'éloigner les représentations vectorielles des mots de la paire. De la même manière, le terme anglais *breast* pourra être associé à de la nourriture via des expressions comme *chicken breast*.

Une analyse plus précise sur la polysémie par l'ajout progressif de données générales sera conduite en Section 4.4.4.

4.4 Sélection de données

Apporter des données générales dans un contexte spécialisé est, comme nous venons de le décrire, une bonne chose, mais a pour conséquence l'introduction de bruit avec l'ajout de polysémie sur les mots à aligner. Nous proposons pour corriger cela de sélectionner les données générales à ajouter au corpus spécialisé en mesurant leur proximité à ce domaine. Être capable de sélectionner un sous ensemble de données générales permettra par la même occasion de réduire la taille des données d'entraînement et donc les temps de calcul.

Dans cette section, nous présentons dans un premier temps le protocole expérimental. Ensuite, nous décrivons les différentes techniques de sélection de données utilisées dans ce travail. Pour finir, nous présentons les résultats pour lesquels nous produisons une analyse détaillée.

4.4.1 Protocole

Nous utilisons deux corpus spécialisés : BC et WE avec les listes d'évaluation correspondantes, ainsi que deux corpus généraux JRC et Wiki (pour simplifier les calculs, le corpus Wiki a été réduit à une taille de 300 M de mots pour chacune de ses parties) sur lesquels nous effectuons la sélection de données. Les expériences sont menées sur la paire de langue français/anglais. Les corpus sont présentés en détail dans le Chapitre 1 en Section 1.3.

Le protocole expérimental est le suivant :

— Pour chaque corpus spécialisé, nous utilisons de manière monolingue différentes méthodes de sélection de données pour trouver les parties les plus proches du domaine spécialisé pour les parties sources et cibles du corpus. En fonction de la technique de sélection de données utilisées, les corpus sont considérés comme étant découpés par document (créés selon le mode de construction du corpus) ou par phrase. Cette sélection nous permet d'ordonner les différents corpus par la proximité des unités (documents ou phrases en fonction de la technique de sélection) avec le domaine spécialisé étudié.

— Ensuite, nous créons les données d'entraînement pour les deux langues en concaténant l'entièreté de notre corpus spécialisé avec des sous parties du corpus général ordonné. Chacune des sous parties est créée en ajoutant 10 % du corpus général, de la plus similaire au domaine spécialisé à la plus éloignée, mais aussi dans l'ordre inverse, de manière à pouvoir comparer plus clairement les différentes approches.

— Pour entraîner les plongements de mots, nous utilisons *fastText* (Bojanowski et al., 2017) sur les corpus sources et cibles précédemment créés [1].
— Pour l'alignement des deux espaces de plongements de mots dans un espace commun, nous utilisons une technique d'alignement supervisé à l'aide de *VecMap* (Artetxe et al., 2018a) couplé au dictionnaire ELRA.
— Finalement, pour tous les mots sources de la liste d'évaluation, nous mesurons leur similarité avec tous les mots du vocabulaire cible à l'aide de la CSLS (Conneau et al., 2017). Les résultats sont mesurés à l'aide de la MAP (Manning et al., 2008).

4.4.2 Techniques de sélection de données

Pour ces expériences, nous avons sélectionné trois techniques de sélection de données, qui sont chacune appliquées sur les corpus de façon monolingue, ainsi qu'une sélection aléatoire pour mieux comprendre l'impact des différentes techniques de sélection. Nous les présentons ci-dessous :

TF-IDF (term frequency-inverse document frequency) est une méthode statistique qui permet d'évaluer l'importance d'un mot au sein d'un document. Le TF-IDF nous sert donc à obtenir une représentation vectorielle pour chacun des documents des corpus spécialisé et général, de manière monolingue où chaque poids du vecteur correspond à l'importance d'un mot au sein du document. L'objectif ici étant d'obtenir une similarité entre un document général par rapport au corpus spécialisé complet. Pour cela, on calcule la similarité moyenne d'un document général avec l'ensemble des documents spécialisés. La similarité entre un document général D_G et le corpus spécialisé C_S de taille N peut être calculée ainsi :

$$sim(D_G, C_S) = \frac{1}{N} \sum_{\forall D_S \in C_S} cos(D_G, D_S) \tag{4.2}$$

Une fois la similarité calculée pour tous les documents du corpus général (source et cible séparément), on peut les ordonner en fonction de ce critère et facilement sélectionner les documents généraux les plus proches du domaine spécialisé.

L'Entropie Croisée (*CrossEntropy*) est utilisée comme définie par Moore and Lewis (2010) pour sélectionner les données générales les plus proches du domaine

1. méthode : Skip-gram ; minCount : 5 ; dim : 300 ; ws : 5 ; minn : 3 ; maxn : 6 ; le reste est laissé par défaut.

spécialisé de façon monolingue. Alors que la technique de sélection TF-IDF s'effectue à l'échelle des documents, on cherche ici à donner un score de similarité à chacune des phrases des corpus généraux pour pouvoir les sélectionner. On utilise l'outil *xenC* (Rousseau, 2013) pour calculer l'entropie croisée.

Pour la partie source des corpus, on calcule deux modèles de langue, un premier pour le corpus spécialisé ($LM_{S,s}$) et un second pour le corpus général ($LM_{G,s}$). De la même manière, on calcule un modèle de langue pour les parties cibles des corpus spécialisé et général ($LM_{S,c}$ et $LM_{G,c}$). Les modèles de langues du corpus général sont calculés à partir d'une sous partie du corpus de la taille du corpus spécialisé (Moore and Lewis, 2010). Ensuite, on calcule l'entropie croisée d'une phrase W par rapport à un modèle de langue LM comme suit :

$$H_{LM}(W) = -\frac{1}{n}\sum_{i=1}^{n}\log(P_{LM}(w_i|w_1, ..., w_{i-1})) \tag{4.3}$$

où P_{LM} est la probabilité donnée par un modèle de langue à une séquence de mot W et $w_1, ..., w_{i-1}$ représente les mots précédents w. $H_{LM_{G,s}}(W)$ représente l'entropie croisée d'une phrase W par rapport au modèle de langue $LM_{G,s}$. L'entropie croisée est calculée pour toutes les phrases du corpus général à partir des modèles de langue des corpus général et spécialisé pour les parties sources et cibles séparément. Les phrases sont ensuite évaluées par $H_{LM_{S,s}}(W_G) - H_{LM_{G,s}}(W_G)$ pour la partie source du corpus (ou $H_{LM_{S,c}}(W_G) - H_{LM_{G,c}}(W_G)$ pour la partie cible) et classées en fonction de cette valeur.

BERT est un modèle qui a prouvé son efficacité sur de nombreuses taches de traitement automatique des langues (Devlin et al., 2019). Nous utilisons le modèle bert-base-cased pour l'anglais (Devlin et al., 2019) et Camembert (Martin et al., 2020) pour le français avec les paramètres par défaut. Nous utilisons BERT pour classifier les phrases du corpus général et indiquer si elles font partie du domaine spécialisé étudié. Nous adaptons les modèles sur nos données spécialisées, chacune des phrases du corpus spécialisés sert d'exemple positif et nous sélectionnons aléatoirement pour chacun de ces exemples une phrase du corpus News Commentary (Tiedemann, 2012) en tant qu'exemple négatif. Une fois les modèles entraînés pour chaque langue, nous pouvons obtenir pour chacune des phrases des corpus généraux un score d'appartenance au domaine spécialisé. C'est ce score qui est utilisé pour sélectionner les phrases du corpus général à ajouter à nos corpus spécialisés.

Sélection aléatoire de données nous appliquons aussi une sélection de données aléatoire à l'échelle de la phrase pour étudier plus précisément l'impact des sélections. Pour cela, nous mélangeons de manière aléatoire le corpus général par phrase et sélectionnons les données en fonction de cet ordre.

4.4.3 Résultats

La Figure 4.3 présente les résultats obtenus pour les quatre méthodes de sélection de données (*CrossEntropy*, *Tf-Idf*, *BERT* and *Random*) pour les deux corpus spécialisés étudiés et les différentes combinaisons avec les données issues des corpus généraux. Les courbes marquées + indiquent que les documents (ou phrases) sont sélectionnés des plus au moins similaires, alors que les courbes marquées − indiquent l'ordre de sélection inverse. Ajouter cette distinction nous permet d'étudier le réel impact de chaque méthode de sélection de données. Le premier point des graphes (où l'axe $x = 0$) correspond aux résultats obtenus avec seulement les données spécialisées et tous les points suivants correspondent aux résultats avec le corpus spécialisé combiné à un certain pourcentage de données sélectionnées, jusqu'à avoir le corpus général complet.

On note d'abord que pour tous les corpus et configurations, l'ajout de données générales améliore grandement les résultats. En regardant les tendances des courbes *CrossEntropy+*, on note une grande amélioration dès 10 % puis une baisse des résultats dans le cas du corpus Wiki alors que pour JRC, le pic est atteint bien plus tard (aux alentours de 70 %. Ces résultats sont particulièrement intéressants, car ils démontrent l'importance d'utiliser une technique de sélection de données suffisamment adaptée et précise, mais aussi que le corpus général utilisé doit être suffisamment vaste pour que cette technique de sélection puisse avoir un impact. Le corpus Wiki étant plus grand que JRC, 10 % correspondent à 30 M de mots, quand c'est seulement 6 M pour JRC. De plus, à 50 % de JRC (qui correspondent donc à 30 M de mots), les résultats de 10 % de Wiki sont toujours plus intéressants. La différence de taille des corpus, ainsi que leur origine, fait que Wiki a plus de chance d'obtenir des données (et donc des contextes) adaptés au domaine spécialisé étudié. Par contre, les courbes − présentent des améliorations lentes et bien plus faibles, les meilleurs résultats sont d'ailleurs obtenus avec l'intégralité des corpus généraux dans la plupart des scénarios. Les courbes + des approches *Tf-Idf* et *BERT* sont bien moins intéressantes que celles de la *CrossEntropy*, car on ne retrouve pas cette amélioration immédiate à 10 % suivies d'une lente dégradation des résultats. Finalement, on note que la courbe *Random* suit une tendance très similaire à celle des courbes −.

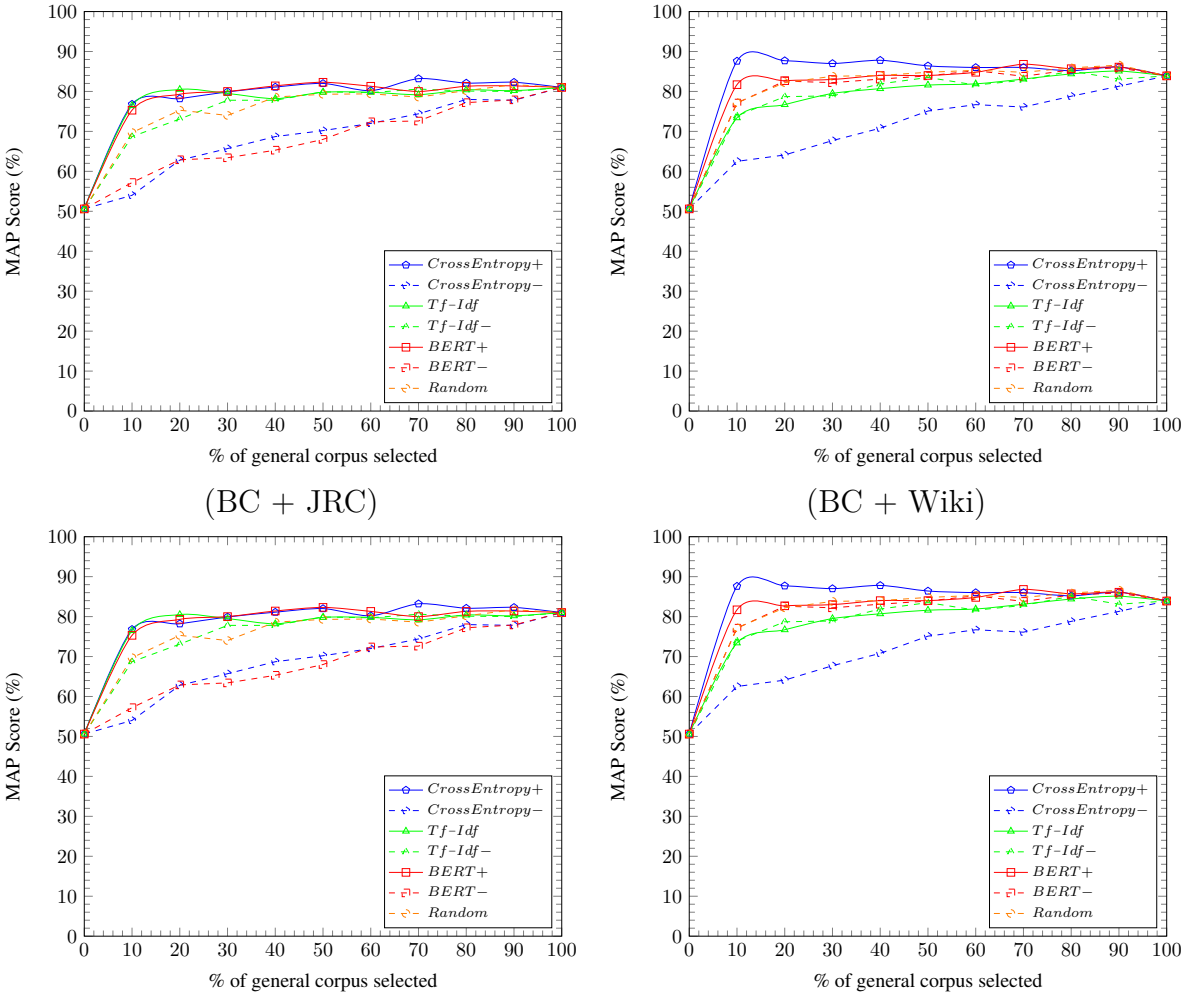

FIGURE 4.3 – Résultats (% MAP) des approches de sélection de données sur différentes combinaison de corpus spécialisés et généraux.

Le Tableau 4.3 présente les résultats sur certains pourcentages de données les plus importants pour la technique de sélection *CrossEntropy*. D'abord, nous avons les résultats pour chacun des corpus seuls (spécialisé ou général), puis les résultats de trois combinaisons. La première colonne (100 %) représente la concaténation simple du corpus spécialisé avec le corpus général : ce que nous avons introduit précédemment comme de l'augmentation de données. Les deux colonnes restantes présentent (pour la colonne +) le pourcentage de données générales sélectionné pour obtenir les meilleurs résultats possibles, ainsi que les résultats obtenus pour le pourcentage équivalent des données les moins intéressantes (la colonne −).

	Spec. *Corpus*	*JRC*	*Spec.* + *n*% *JRC*			*Wiki*	*Spec.* + *n*% *Wiki*		
			100 %	70 %+	70 %-		100 %	10 %+	10 %-
BC	50,6	59,8	81,0	83,2	74,4	82,7	83,9	**87,6**	62,5
WE	53,4	66,4	68,8	72,0	65,6	69,7	75,9	**80,9**	55,5

TABLEAU 4.3 – Résultats (% MAP) sur les corpus et les différentes combinaisons possibles pour la technique de sélection *CrossEntropy*.

On note clairement une augmentation des résultats avec le corpus Wiki en passant d'une technique d'augmentation des données (100 %) à de la sélection (10 %) : le gain en MAP est de 3,7 points pour le corpus BC et de 5 points pour le corpus WE. Pour le corpus JRC, si l'on a toujours un gain en termes de résultats (2, 2 % pour BC et 3, 2 % pour WE), il est nécessaire d'utiliser plus de données (70 %). Ces observations confirment l'utilité de la sélection de données, sous la condition de posséder suffisamment de données générales.

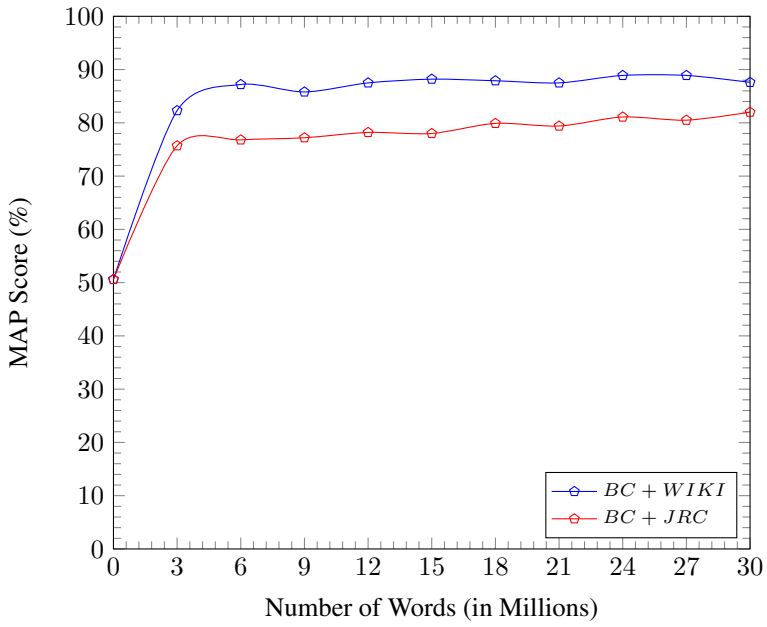

FIGURE 4.4 – Résultats avec la *CrossEntropy*, sur des pourcentages plus faibles des corpus généraux. (3 M de mots représentent 1 % de Wiki et 5 % de JRC)

Compte tenu des résultats de la *CrossEntropy* à 10 % de données générales pour le corpus Wiki, nous avons mené une expérience supplémentaire sur les plus petits pourcen-

tages pour mieux comprendre l'impact immédiat de l'ajout de données. Cette expérience est illustrée en Figure 4.4 où l'évolution de la quantité de données est représentée en termes de million de mots pour permettre une comparaison directe entre les deux corpus (3 M de mots représentent 1 % de Wiki et 5 % de JRC). On note que les deux courbes possèdent la même tendance, avec de meilleurs scores pour la courbe Wiki.

Le détail obtenu par cette sélection plus fine permet de montrer qu'à seulement 1 % des données générales, on atteint déjà 82,3 % de MAP, ce qui est presque autant que le corpus complet (83,9 %). Ce résultat est d'ailleurs surpassé dès 2 % de données, avec 87,2 % de MAP. Les meilleurs résultats sont finalement obtenus à 8 et 9 % de données avec un score de 88,9 %. Pour JRC, l'ajout de peu de données reste intéressant, mais ne permet pas de dépasser les résultats maximums obtenus précédemment avec 70 % de données générales ajoutées.

Type de sélection de données		Temps de sélection (s)	Pourcentage optimal	Temps d'entraînement (s)	MAP
Pas de sélection	BC	-	0	**180**	50,6
Augmentation	BC + Wiki	-	100	65 521	83,9
Sélection	*CrossEntropy*	420	8	5 241	**88,9**
	Tf-Idf	339	90	58 968	85,1
	BERT	147 467†	70	45 865	86,8

Tableau 4.4 – Temps de calcul (en secondes) pour la sélection de données et l'entraînement des plongements de mots pour les corpus BC et Wiki (†Temps GPU).

Le Tableau 4.4 montre le temps de calcul nécessaire pour l'étape de sélection de données ainsi que le temps pour entraîner les plongements de mots pour les corpus BC et Wiki au pourcentage optimal des données (Les deux premières approches n'ont pas à proprement parler de pourcentage optimal, soit on ne sélectionne pas le corpus général, soit on en prend son intégralité). Les expériences ont été menées sur un Intel Core i9-9900K et une GeForce RTX 2080. Le temps d'entraînement des plongements de mots correspond à l'entraînement des espaces sources et cibles cumulés.

On remarque qu'utiliser des données spécialisées seules est très rapide, mais présente des résultats relativement faibles (50,6 % de MAP), alors qu'une approche d'augmentation des données va considérablement augmenter les résultats, mais aussi le temps de calcul (presque 400 fois plus long). Finalement, avec une bonne technique de sélection de

données telle que la $CrossEntropy$, l'augmentation du temps de calcul reste relativement raisonnable (5 061s), tout en augmentant la MAP (88,9 %). Les techniques de sélection de données $BERT$ et $Tf\text{-}Idf$ présentent aussi de meilleurs résultats, mais moins intéressants que pour la $CrossEntropy$. De plus, le temps de sélection de l'approche $BERT$ est immense comparé à une méthode simple comme la $CrossEntropy$.

4.4.4 Analyse

La section précédente a permis d'établir un lien entre la sélection de données et l'amélioration des résultats de la tâche de BLI. Mieux encore, en plus d'améliorer les résultats, nous avons pu remarquer que la sélection de données permettait de réduire considérablement les temps de calcul. S'il est important dans le cadre d'un domaine spécialisé d'être capable de mieux représenter les termes propres à ce domaine, nous supposons qu'il est tout aussi important d'être capable d'améliorer la représentation des mots plus communs. Dans l'idée de mesurer l'impact de ces différents points, nous avons sélectionné différentes paires affectées positivement ou négativement par l'ajout de données.

Le Tableau 4.5 présente les résultats de 6 paires, en indiquant, pour différents pourcentages de données générales sélectionnées, l'évolution du nombre d'occurrences des deux mots dans leur langue respective ainsi que le rang auquel la traduction a été classée. À chacun des rangs est assigné une couleur indiquant l'impact de l'ajout de la nouvelle tranche de données générales. Le vert signifie que l'ajout a permis d'obtenir un résultat au moins aussi bon alors que le rouge signifie une dégradation dans le rang de traduction. Ce tableau présente les résultats pour la $CrossEntropy$, la technique de sélection présentant les meilleurs résultats parmi celles étudiées.

Pour la première paire (*breast-sein*), on note que les données spécialisées seules suffisent amplement à obtenir un résultat satisfaisant (rang 2). Les deux mots de cette paire sont en effet très présents dans le corpus et possèdent donc une représentation précise. Cependant, l'ajout de données brut (colonne 100 %) dégrade grandement ce rang (≥ 1000). En regardant plus précisément quels types de contextes étaient ajoutés par le corpus général, on note que l'expression *au sein de* (qui peut être traduite par *within* ou *at the heart of*) apparaît énormément (47 025 fois) dans les données générales. Comme cette expression peut apparaître *au sein de* n'importe quel contexte, l'ajout de données générales va progressivement éloigner le sens du mot *sein* du domaine spécialisé, ce qui se conclue par l'impossibilité de traduire correctement le mot *breast*. On note tout de même que, si sélectionner 10 % de données spécialisées dégrade légèrement le classement (rang 3 au lieu

BC+n % Wiki	0 %	10 %	20 %	30 %	40 %	50 %	60 %	70 %	80 %	90 %	100 %
En Occ (breast)	$3,6k$	$6,5k$	$7,3k$	$7,8k$	$8,2k$	$8,5k$	$8,7k$	$8,8k$	$8,9k$	$8,9k$	$9,0k$
Fr Occ (sein)	$2,9k$	$7,3k$	$12,3k$	$17,8k$	$23,8k$	$30,2k$	$37,1k$	$44,3k$	$51,5k$	$57,0k$	$59,4k$
Rank	2	3	12	604	399	933	≥ 1000	≥ 1000	≥ 1000	≥ 1000	≥ 1000
En Occ (calcium)	23	$1,8k$	$2,2k$	$2,3k$	$2,4k$	$2,5k$	$2,6k$	$2,6k$	$2,7k$	$2,7k$	$2,7k$
Fr Occ (calcium)	14	983	$1,3k$	$1,5k$	$1,7k$	$1,8k$	$1,8k$	$1,9k$	$1,9k$	$2,1k$	$2,2k$
Rank	140	1	1	1	1	1	1	1	1	1	1
En Occ (back)	27	$7,3k$	$19,5k$	$33,8k$	$49,7k$	$66,6k$	$84,0k$	$102,1k$	$120,2k$	$139,8k$	$153,9k$
Fr Occ (dos)	7	883	$2,0k$	$3,3k$	$4,6k$	$6,0k$	$7,4k$	$8,8k$	$10,3k$	$11,9k$	$14,0k$
Rank	≥ 1000	37	12	4	5	9	14	50	43	27	99
En Occ (lymphoscintigraphy)	20	20	20	20	20	20	20	20	20	20	20
Fr Occ (lymphoscintigraphie)	27	28	28	28	28	28	28	28	28	28	28
Rank	4	1	1	1	1	1	1	1	1	1	1

TABLEAU 4.5 – 4 paires intéressantes pour différents pourcentages de sélection de donnée à l'aide de la *CrossEntropy*.

de 2), la perte reste raisonnable dans les premiers pourcentages.

Inversement, le mot source *calcium* bénéficie de l'ajout de contextes dans les deux parties du corpus. Initialement, sa traduction cible *calcium* était trouvée au rang 140 avec seulement les données spécialisées. L'ajout de nombreuses occurrences des mots dans les deux langues à l'aide des données générales améliore le résultat au rang 1.

La paire *back-dos* est intéressante, car elle permet une grande amélioration dans les premiers pourcentages de données générales ajoutés grâce à l'arrivée de nouveaux contextes. S'en suivent de légères variations à tendance négative, car *back* est utilisé dans des contextes très variés en anglais.

Finalement, notre dernière paire *lymphoscintigraphy-lymphoscintigrapie* montre l'intérêt de l'ajout de données, même lorsqu'il ne se conclut pas par l'ajout de nouvelles occurrences des mots que l'on cherche à traduire. En effet, alors qu'il n'y a (quasiment) aucune nouvelle occurrence d'un des deux mots (seulement une nouvelle occurrence en français), le classement de la paire passe de top 4 à top 1. On peut en conclure qu'être capable de mieux représenter les autres mots est aussi de grande importance.

Ces quatre exemples montrent qu'il est important d'être capable de mieux représenter les mots que l'on cherche à traduire, tout en faisant attention aux contextes ajoutés.

Dans le Tableau 4.6, nous exposons quelques exemples de phrases sélectionnées par l'approche *CrossEntropy* sur les corpus général Wiki et spécialisé BC en anglais.

Il est intéressant de noter que les premières "phrases" sélectionnées (*CrossEntropy+*)

Phrase	*CrossEntropy+*	*CrossEntropy-*
1	tumor suppressor gene	marginal worker household industry worker
2	receiver operate characteristic	notable former player
3	single photon emission compute tomography	main worker household industry worker
4	fine needle aspiration	id bass value blue legend bass
...	...	...
115	breast density is positively associated with breast cancer	-

TABLEAU 4.6 – Exemples de phrases du corpus général Wiki anglais, sélectionnées pour le corpus spécialisé BC par la *CrossEntropy*.

sont actuellement plus des groupes de mots liés au domaine étudié (*tumor*, *gene*, *operate*, *needle*...). On note tout de même la présence de réelles phrases (voir ligne 115) mais la *CrossEntropy* semble favoriser les groupes de mots. Comme espéré, les phrases les moins intéressantes (*CrossEntropy−*) ne présentent aucune relation avec le domaine du cancer du sein.

Paire	0 %	10 %	100 %	Paire	0 %	10 %	100 %
	année	cancer	poitrine		demi-vie	**pression**	**pression**
Breast	**sein**	ovaire	cuisse	Pressure	résorber	tension	tension
-	der	**sein**	cou	-	millimètre	compression	poussée
Sein	mammaire	prostate	épaule	Pression	résorption	poussée	aspiration
	cancer	mammaire	utérus		gradient	dilatation	instabilité

TABLEAU 4.7 – Exemples de l'évolution des traductions proposées pour deux paires : *pressure-pression* et *breast-sein* pour la technique de sélection *CrossEntropy* sur les corpus général Wiki et spécialisé BC.

Finalement, le Tableau 4.7 illustre l'évolution des traductions candidates à trois étapes de la sélection de données (0, 10 et 100 %). La première paire *pressure-pression* montre le problème d'un trop petit corpus et du manque d'occurrences des mots qui le composent. Le mot anglais *pressure* n'est présent que 21 fois dans le corpus spécialisé, alors qu'il apparaît respectivement 1 655 et 8 203 fois avec 10 et 100 % du corpus général ajouté. La seconde paire *breast-sein* comprend deux mots déjà présents de nombreuses fois dans le corpus spécialisé qui devraient donc être correctement représenté, mais il est relativement surprenant d'obtenir au rang 1 *année*. L'ajout de 100 % de données générales montre

clairement l'introduction de la polysémie, si les traductions proposées sont toujours des termes relativement proches comme *poitrine*, *sein* a totalement disparu à cause de l'expression *au sein de*. Ce problème est résolu en partie par la colonne 10 % de données, avec *sein* toujours au rang 3.

4.5 Synthèse

Dans ce chapitre, nous avons étudié un scénario de la tâche de BLI en domaine de spécialité. Nous avons dans un premier temps présenté des approches d'augmentation de données en utilisant l'approche standard ou les plongements de mots. Au vu des problèmes apportés par ces approches, nous avons proposé l'utilisation de techniques de sélection de données et étudié leur impact.

Les expériences misent en œuvre sur deux corpus généraux et deux corpus spécialisés montrent qu'il est parfaitement possible de sélectionner une petite quantité de données adaptée au domaine pour grandement améliorer les résultats (gain de 6 points de MAP sur le corpus BC avec seulement 8 % de Wiki par exemple) au lieu d'ajouter de manière brute des corpus généraux entiers. Cette sélection de données, en plus d'améliorer les résultats, permet aussi de réduire le temps de calcul des approches (par un facteur de 10 dans notre cas).

La sélection de données nous a aussi permis d'étudier plus précisément et de mieux comprendre la présence de polysémie et l'augmentation des résultats est aussi due à une réduction de ce phénomène. Même s'il reste présent dès l'ajout de trop de données générales, il a été grandement réduit par l'ajout de moins de données plus intéressantes. Les différentes analyses ont permis de montrer que les pourcentages optimaux de données générales ne sont pas forcément les mêmes en fonction des différents mots du vocabulaire. Il pourrait alors par exemple être intéressant d'être capable de sélectionner de manière indépendante pour chacun des mots les données générales intéressantes à ajouter.

Chapitre 5

Évaluation en BLI

Nous avons présenté dans ce manuscrit de nombreuses méthodes pour produire des représentations sémantiques des mots. Par exemple, les représentations utilisées historiquement peuvent être basées sur les sacs de mots, de grands vecteurs creux, mais qui sont compréhensibles, ou des plongements de mots, qui sont à plus petite dimensionnalité et donc denses, mais pour lesquels il est compliqué d'interpréter la réelle signification des dimensions. Les plongements de mots peuvent être statiques, c'est-à-dire que pour un mot, il n'existe qu'une seule représentation, quel que soit son contexte, ou contextualisé, pour lesquels il n'y a qu'une représentation par occurrence d'un mot, basée sur son contexte.

Nous avons aussi présenté différentes techniques pour aligner les représentations sémantiques, d'abord l'approche standard pour les représentations basées sur les sacs de mots où l'on utilise le fait que chaque dimension d'un vecteur correspond à un mot du vocabulaire, ce qui permet de traduire les dimensions d'une langue vers l'autre à l'aide d'un lexique d'entraînement. Ensuite, pour les plongements de mots, des techniques souvent basées sur une matrice de traduction qui a pour objectif de projeter les espaces des deux langues dans un espace commun, avec des variations propres à chaque approche. Mais aussi des approches d'entraînement conjoint, où l'on entraîne simultanément et à l'aide d'un même système les représentations sémantiques des deux langues que l'on cherche à étudier.

Toutes ces méthodes et techniques nécessitent d'être comparées, cependant, elles ont rarement été évaluées de manière consistante à travers les différentes études. Par exemple, Mikolov et al. (2013b) s'évaluent sur les paires anglais-espagnol et anglais-tchèque à l'aide de la précision aux Top 1 et 5. Dinu et al. (2014) s'évaluent sur la paire anglais-italien et sont repris par Artetxe et al. (2017, 2018b) en utilisant la Précision. Duong et al. (2016) utilisent des dictionnaires PanLex (Kamholz et al., 2014) et Wiktionnary [1] et utilisent le Rappel. Gouws and Søgaard (2015) s'évaluent quant à eux sur une tâche de POS-tagging.

De manière à unifier les processus d'évaluation, Conneau et al. (2017) proposent le

1. https://www.wiktionary.org/

jeu de données MUSE pour le domaine général (que nous avons présenté en Section 1.3). MUSE a ensuite été repris dans de nombreux travaux tels que Joulin et al. (2018); Patra et al. (2019); Wang et al. (2020) ou encore Rapp et al. (2020) qui proposent d'évaluer les résultats avec la F-mesure. Cependant, ce jeu de données généré automatiquement souffre de nombreux problèmes, comme la sur-représentation des mots à haute fréquence ou de paires graphiquement identiques.

Les jeux de données spécialisés peuvent aussi souffrir de problèmes : souvent de petites tailles (quelques centaines de paires de mots), ils peuvent être difficiles à diffuser, car liés à des corpus spécialisés construits avec des données non diffusables (comme le corpus du cancer du sein que nous utilisons). Cependant, leur petite taille assure une certaine qualité, car ils sont souvent créés manuellement et il est aussi plus facile d'aller regarder paire par paire leur contenu.

Privilégier la quantité de données, comme peuvent le faire les études sur les données générales, peut parfois être un choix mal avisé, car, comme on pourra le démontrer lors de ce chapitre, on risque des pertes de qualité sur des listes de grande taille.

Plusieurs travaux ont étudié le problème de l'évaluation en BLI (Czarnowska et al., 2019; Kementchedjhieva et al., 2019; Glavaš et al., 2019), en se concentrant essentiellement sur les données générales. Cependant, les directions et améliorations proposées par ces travaux ont rarement été suivies.

Dans ce chapitre, nous étudions différents scénarios de BLI, en faisant varier les méthodes utilisées par exemple. Les résultats que nous présentons nous permettent aussi de nous questionner sur le bien-fondé des listes d'évaluation utilisées et de découvrir leurs défauts. Nous cherchons alors à mieux comprendre les problèmes qui subsistent en BLI et les solutions que l'on peut y apporter. Pour cela, nous réunissons les différentes recommandations déjà proposées et ajoutons les nôtres, dans l'idée de proposer un processus d'évaluation unifié.

5.1 Différences dans les méthodes utilisées

De nombreuses approches et méthodes ont été proposées au fil du temps pour accomplir la tâche de BLI. Si la plupart s'évaluent en comparaison avec les meilleures approches précédentes, les comparaisons entre les familles d'approches (méthodes d'alignement face à de l'entraînement conjoint ou des plongements de mots contextualisés face aux plongements de mots statiques par exemple) sont plutôt rares. De plus, la plupart des nouvelles

méthodes présentent de meilleurs résultats dans le scénario étudié dans le travail correspondant. En réalité, toutes les approches présentées ces dernières années présentent des résultats plus ou moins équivalents.

De fait, dans cette section, nous ne cherchons pas à comprendre quelle est l'approche la plus performante, mais plutôt les points où aujourd'hui les approches de BLI ont des problèmes pour obtenir les meilleurs résultats possibles. Pour cela, chaque expérience présentée est suivie d'une analyse détaillée des résultats. Nous montrons aussi qu'il est très facile d'abuser les méthodes d'évaluation à l'aide de techniques très simples basées sur la proximité graphique des paires de mots des listes d'évaluation.

5.1.1 Étude de différents types de représentations

Nous avons présenté dans le Chapitre 2 trois types de représentations sémantiques des mots. Les plus anciennes, basées sur les sacs de mots (BoW), sont à hautes dimensionnalités, mais explicables : chaque dimension correspond à un mot du vocabulaire. Cette grande taille peut toutefois les rendre difficilement manipulables. Ensuite, Mikolov et al. (2013a) ont introduit les modèles CBOW et Skip-gram, qui marquent l'arrivée des plongements de mots statiques qui, à l'inverse des BoW, sont composés généralement de quelques centaines de dimensions, facilitant les calculs, mais les rendant difficilement explicables. Pour ces deux premiers types de représentations, chaque mot possède une représentation commune pour toutes ses occurrences. Finalement, nous avons présenté les plongements de mots contextualisés avec ELMo et BERT. Ces modèles vont créer une représentation différente pour chacune des occurrences d'un mot, en se basant sur le contexte de la phrase dont l'occurrence est extraite.

Dans cette section, nous cherchons à étudier les différences entre ces trois types de représentation, en fonction de différents scénarios. Les résultats que nous présentons par la suite ont été publiés à Canadian AI (Laville et al., 2020c).

Pour étudier ces différents types de représentations, nous les appliquons à deux types de données. D'abord, nous utilisons un jeu de données spécialisé avec le corpus du cancer du sein (BC) et la liste d'évaluation qui va avec, mais aussi un jeu de données général, avec le corpus Wikipédia (Wiki) associé aux listes du jeu de données MUSE. Le corpus Wiki a été réduit à 100 millions de mots dans cette étude, pour simplifier les calculs utilisant des plongements de mots contextualisés. Ces jeux de données ont été présentés en section 1.3.

Finalement, nous testons aussi différentes méthodes d'alignement des représentations sémantiques, que nous classons en deux catégories principales. D'abord, les approches

supervisées, pour lesquelles nous testons deux dictionnaires : MUSE et ELRA, avec l'approche standard pour BoW et une matrice d'alignement pour les deux représentations utilisant des plongements de mots. Ensuite, une approche non-supervisée, utilisée seulement pour les plongements de mots.

Nous évaluons ces différents scénarios sur la paire de langue anglais-français. Avec l'idée de mieux comprendre les résultats obtenus, nous avons décidé de séparer les listes d'évaluation en différentes sous-listes, mais aussi, car nous avons remarqué quelques incohérences au sein des listes MUSE.

D'abord, du fait de la présence de mots qui semblent ne pas appartenir à la langue d'origine dans les listes (*garrison* ou *enjoy* étant présents dans la partie française alors que ce sont des mots anglais), nous créons une première sous-liste (In-dictionary / Full) pour laquelle nous filtrons les mots à l'aide de dictionnaires monolingues [2] [3]. Ce premier filtre réduit la liste du domaine général par près d'un quart de ses paires (de 1 446 paires à 1 139 paires). Nous appliquons aussi ce filtre à la liste spécialisée, ce qui entraîne la suppression d'une trentaine de paires (13 % de la liste initiale). Cependant, en étudiant les mots retirés de la liste spécialisée on se rend compte que ce sont par exemple des sigles (par exemple *AIDS* / *SIDA* ou *DNA* / *ADN*) ou des mots absents du dictionnaire général mais dont la traduction reste intéressante à obtenir (par exemple un type d'opération *lumpectomy* / *zonectomie* ou un nom de médicament *tamoxifen* / *tamoxifène*). Ce plus faible pourcentage et l'intérêt des mots supprimés montrent tout de même une meilleure qualité au sein du lexique spécialisé. Nous retirons tout de même ces mots pour étudier l'évolution des résultats.

Ensuite, en repartant de la sous-liste créée à partir des dictionnaires monolingues, on note la présence de nombreux mots graphiquement proches voire même identiques. Nous créons alors une autre sous-liste (Lev. ≥ 3) à l'aide de la distance de Levenshtein (le nombre de suppressions, insertions ou substitutions requises pour passer d'une chaîne de caractères à une autre) pour étudier les mots qui ne partagent pas la même morphologie.

Enfin, nous créons une dernière sous-liste (Freq. ≤ 100), encore une fois à partir de la sous-liste des mots présents dans les dictionnaires monolingues, pour étudier les mots peu fréquents en ne gardant que ceux apparaissant 100 fois ou moins. Cette dernière sous-liste réduit la liste spécialisée à seulement 18 paires, ce qui rend donc difficile de tirer des conclusions sur cet échantillon. Le détail des différentes listes est présenté en Tableau 5.1.

2. Dictionnaire anglais : `github.com/dwyl/english-words` (466k mots)
3. Dictionnaire français : `infolingu.univ-mlv.fr/` (684k mots)

Domaine	Original	In-dictionary		
		Full	Lev. ≥ 3	Freq. ≤ 100
Général	1 446	1 139 (79 %)	783 (54 %)	146 (10 %)
Spécialisé	248	216 (87 %)	85 (34 %)	18 (8 %)

TABLEAU 5.1 – Taille des listes d'évaluation et de leurs sous-listes pour les domaines général et spécialisé.

Les représentations des mots sont entraînées sur le corpus général complet dans le scénario général et sur le corpus spécialisé concaténé au corpus général (augmentation de données) dans le scénario spécialisé, comme présenté dans le Chapitre 4. Les BoW sont entraînés et alignés selon l'approche standard présentée dans les chapitres précédents. Les plongements de mots statiques sont entraînés à l'aide de l'outil *fastText* (Bojanowski et al., 2017). Pour les plongements de mots contextualisés, ils sont extraits à partir du modèle ELMo proposé par Schuster et al. (2019) (sans adaptation du modèle de notre part). On crée ensuite des ancres pour chacun des mots du vocabulaire, comme présenté dans la Section 3.5 du Chapitre 3.

Pour aligner les méthodes par plongement de mots, nous utilisons une approche supervisée (Artetxe et al., 2018a) et une non-supervisée (Artetxe et al., 2018b) à l'aide de l'outil VecMap. L'approche supervisée est utilisée avec deux dictionnaires différents : ELRA (243 539 paires de mots) et MUSE (10 972 paires).

Nous présentons dans le Tableau 5.2 les résultats obtenus sur les différentes listes présentées précédemment en les combinant à différents types de représentations et approches d'alignement pour les domaines général et spécialisé.

Nous pouvons d'abord observer que, même si les méthodes non-supervisées présentent des résultats intéressants, les méthodes supervisées restent supérieures de quelques points dans la majeure partie des configurations. Aussi, pour les approches utilisant des plongements de mots, l'utilisation d'un dictionnaire de plus grande taille et de supposément meilleure qualité (ELRA) n'implique pas l'obtention de meilleurs résultats qu'en utilisant un dictionnaire généré automatiquement comme MUSE. Par contre, on note que BoW obtient de meilleurs résultats en domaine spécialisé avec le dictionnaire ELRA.

Sur les listes originales, *fastText* obtient toujours les résultats les plus intéressants. Cependant, en regardant les sous-listes, les résultats se dégradent rapidement alors que pour ELMo la dégradation est plus faible, certains scénarios obtiennent même une légère hausse. On peut expliquer ces résultats de plusieurs manières, d'abord, les mots des sous-

listes sont plus rares dans les corpus originels et ELMo à l'avantage d'être basé sur un modèle pré-entrainé sur un autre corpus de taille conséquente, ce qui favorise l'obtention de meilleures représentations et biaise ainsi cette comparaison.

Domaine	Alignement	Méthode	Original	In-dictionary		
				Full	Lev. ≥ 3	Freq. ≤ 100
General	Non-Supervisé	fastText	**68,9**	60,2	38,4	30,8
		ELMo	62,1	**72,2**	**57,2**	**68,0**
	Supervisé (MUSE)	fastText	**70,4**	64,6	44,1	44,9
		ELMo	63,4	**72,7**	**59,0**	**70,1**
		BoW	53,4	49,9	35,7	4,3
	Supervisé (ELRA)	fastText	63,8	63,2	44,2	41,7
		ELMo	57,4	70,1	55,9	58,5
		BoW	43,8	46,2	34,7	3,8
Spécialisé	Non-Supervisé	fastText	**80,6**	**81,4**	60,0	**94,4**
		ELMo	70,4	77,7	**61,2**	61,1
	Supervisé (MUSE)	fastText	**81,8**	**82,3**	**63,5**	**83,3**
		ELMo	68,4	75,3	62,4	50,0
		BoW	59,5	65,3	53,8	16,7
	Supervisé (ELRA)	fastText	80,2	81,9	62,4	77,8
		ELMo	68,8	75,8	61,2	50,0
		BoW	67,6	73,5	61,2	27,8

TABLEAU 5.2 – P@1 (%) sur différents scénarios de BLI. En gras, on note la meilleure approche pour les domaines général et spécialisé parmi les approches non-supervisées d'un côté et supervisées de l'autre. On souligne l'approche la plus efficace pour chacun des domaines (sans prendre en compte si l'approche est supervisée ou non).

Nous proposons maintenant une analyse plus qualitative des résultats obtenus dans les domaines spécialisé et général en les illustrant à partir de quelques paires de mots sélectionnées. Pour cela, nous présentons dans le Tableau 5.3 trois paires du domaine général et trois paires du domaine spécialisé, en indiquant la fréquence des différents mots et les 3 premières traductions proposées par les différents systèmes.

En étudiant d'abord le domaine général, on peut observer que *fastText* trouve majoritairement des mots proches graphiquement, sans réellement saisir le concept du mot

(par exemple, "napoléone" est une plante et "wrestlers" n'est pas un mot français). À l'inverse, ELMo semble mieux comprendre le sens des mots. Par exemple, pour "napoléon", le système propose des termes relatifs à la guerre ou encore, des formes géométriques pour "rings". Les représentations BoW semblent être grandement affectées par la fréquence des mots.

Dans le domaine spécialisé, comme les mots sont supposés n'avoir qu'un seul sens spécifique, il est moins probable de les trouver dans des contextes très variés qui compliqueraient l'obtention d'une représentation précise. Ainsi, *fastText* et BoW réussissent plus facilement à trouver des mots proches, même si une paire peu fréquente comme *vincristine-vincristine* peut toujours poser des problèmes à l'approche BoW (où *dominique* et *monique* sont des prénoms).

Domaine	Méthode	Mot *Traduction*	Top 1	Top 2	Top 3
General	fastText	napoleon : 2,1 k *napoléon* : 5,2 k	**napoléon**	napoléone	napoléonienne
	ELMo		bélisaire	**napoléon**	guerry
	BoW		**napoléon**	bonaparte	xiv
	fastText	rings : 710 *anneaux* : 117	**anneaux**	rings	ring
	ELMo		**anneaux**	ceintures	sphères
	BoW		**anneaux**	rouhault	penon
	fastText	wrestlers : 27 *lutteurs* : 10	catches	catchers	wrestlers
	ELMo		**lutteurs**	joueurs	joueuses
	BoW		grandidieri	bergroth	committeer
Spécialisé	fastText	birth : 9 k *naissance* : 14 k	**naissance**	décès	âge
	ELMo		**naissance**	baptême	éclosion
	BoW		**naissance**	enfant	mère
	fastText	keratin : 66 *kératine* : 52	**kératine**	fibroblaste	adipocyte
	ELMo		**kératine**	collagène	mélanine
	BoW		**kératine**	luminales	fibrine
	fastText	vincristine : 23 *vincristine* : 15	**vincristine**	dominique	monique
	ELMo		**vincristine**	raloxifène	fusarium
	BoW		vinorelbine	herceptin	rechuter

Tableau 5.3 – Détail des 4 meilleures traductions pour différentes paires dans le scénario Supervised (MUSE). Nous indiquons pour chacun des mots de la liste d'évaluation leur fréquence dans les corpus d'entraînement.

5.1.2 Abuser des similarités entre les langues

Rapp et al. (2020) proposent une campagne d'évaluation de BLI en domaine général.

L'idée est de comparer différents systèmes selon un protocole d'évaluation et des données communes. En effet, les systèmes de BLI sont rarement évalués de manière similaire, comme nous le verrons au cours de ce chapitre.

Contexte

Les organisateurs proposent 6 langues pour effectuer cette tâche : allemand, anglais, chinois, espagnol, français et russe. Les corpus utilisés sont les corpus WaCKy (Baroni et al., 2009) et Wiki que nous avons présentés en Section 1.3. Nous indiquons dans le Tableau 5.4 le détail des paires de langues et les corpus correspondants sur lesquels nous nous sommes concentrés lors de notre participation (Laville et al., 2020b). Les organisateurs de la tâche considéraient que WaCKy était plus adapté à la tâche de BLI, mais il n'est pas disponible en espagnol, d'où l'utilisation de Wiki dans ce cas d'étude.

Langue	*de*	*es*	*fr*	*ru*
en	WaCKy	Wiki	WaCKy	WaCKy
de	-	-	WaCKy	-

Tableau 5.4 – Corpus utilisé pour chacune des paires de langues.

Dans cette tâche, MUSE est utilisé pour entraîner puis évaluer les différents systèmes. Les organisateurs proposent comme données d'entraînement trois listes, découpées en fonction de la fréquence des mots sources : high (mots parmi les 5 000 les plus fréquents), mid (entre 5 001 et 20 000) et low (entre 20 001 et 50 000). Les résultats seront ensuite évalués sur d'autres données fournies quelques jours avant la date de fin. Nous utilisons donc les listes originales en gardant 20 % de chacune d'entre elles pour la validation alors que le reste est regroupé pour servir de lexique d'entraînement.

À l'inverse de la majeure partie des travaux de BLI qui utilisent la Précision au rang 1 (P@1) ou le MAP score, les résultats sont ici mesurés à l'aide de la Précision, du Rappel et du F1-Score. Au lieu de présenter une liste ordonnée des mots du vocabulaire cible pour chacun des mots sources de la liste d'évaluation, il faut proposer une liste finie de candidats traductions. Pour ces travaux, nous avons étudié deux systèmes et leurs combinaisons, un premier basé sur les plongements de mots et un second basé sur les similarités graphiques entre les langues. Nous présentons ces systèmes dans les paragraphes suivants.

Plongements de mots (Skip-gram, CBOW, Concaténation)

Ce système consiste dans un premier temps à entraîner des plongements de mots CBOW et Skip-gram (à l'aide de l'architecture *fastText*) séparément sur les corpus. Nous utilisons ensuite l'outil VecMap pour projeter les espaces vectoriels dans un espace commun à l'aide du lexique d'entraînement. Pour la validation, nous utilisons la séparation 80/20 (entraînement/validation) comme indiqué précédemment. Pour les résultats finaux, nous utilisons tout le lexique d'entraînement pour l'alignement. La similarité entre les mots sources et les mots cibles est mesurée avec la CSLS que nous avons présentée dans le Chapitre 3 en Section 3.6. La validation nous sert à définir comment nous sélectionnons les mots cibles que nous proposons comme traduction des mots sources, nous définissons alors deux critères pour chacune des paires de langue : 1) un nombre maximal de candidats que nous gardons dans la liste finale et 2) une similarité CSLS minimale.

Nous rajoutons aussi une approche en concaténant les plongements de mots CBOW et Skip-gram après alignement. Les plongements originaux étant composés de 300 dimensions, nous obtenons alors des plongements de 600 dimensions.

Le Tableau 5.5 présente les différents paramètres utilisés pour sélectionner les candidats pour chacune des paires de langues. Les valeurs sont communes aux trois approches de plongements de mots (CBOW, Skip-gram et Concaténation) et fixées de manière empirique.

Language pair	*Cand.* $\leq$	*Sim.* $\geq$
en-es	4	0,1
es-en	2	0,08
en-de	5	0,06
de-en	5	0,04
en-fr	3	0,08
fr-en	2	0,04
en-ru	4	0,05
ru-en	2	0,03
de-fr	2	0,08
fr-de	2	0,06

TABLEAU 5.5 – Paramètres pour sélectionner les candidats pour chacune des paires de langue. Nous sélectionnons un nombre maximal de candidats (Cand. $\leq$) et une similarité minimale (Sim. $\geq$). Dans le cas où plus de mots cibles ont une similarité plus élevée que le seuil de candidat maximal que nous avons fixé, nous sélectionnons les mots avec la plus haute similarité en premier.

Cognats parfaits (Perfect Cognates)

Une analyse des lexiques utilisés nous a permis de nous rendre compte que de nombreuses paires de mots étaient graphiquement identiques, particulièrement dans les mots à basse fréquence. Nous décidons alors d'utiliser un second système très simple qui sélectionne comme candidat valide pour un mot source, le mot cible graphiquement identique s'il existe. Nous ajoutons aussi la contrainte que le mot cible doit avoir une fréquence suffisamment proche du mot source. Pour un mot source w_s et son équivalent dans le vocabulaire cible w_c, et selon la fréquence de w_s ($freq(w_s)$) et w_t ($freq(w_t)$), la contrainte peut être formulée ainsi :

$$\frac{1}{n} \leq \frac{freq(w_s)}{freq(w_t)} \leq n \tag{5.1}$$

avec n empiriquement fixé à 100.

Combinaison des approches (Mix (Conc + Cognates))

Pour obtenir de meilleurs résultats, plusieurs systèmes sont souvent combinés. Comme le démontrera le Tableau 5.6, les approches de plongements de mots obtiennent les meilleurs résultats sur les mots à haute fréquence, alors que l'approche basée sur les mots graphiquement identiques est plus intéressante pour les mots à basse fréquence. Ainsi, il est assez naturel de combiner les deux approches pour tirer profit des avantages des deux stratégies. Nous proposons alors une combinaison des listes des deux approches (plus exactement des listes obtenues via la concaténation des plongements de mots avec l'éventuel mot obtenu via les cognats parfaits).

Résultats

Dans cette section, nous présentons dans le Tableau 5.6 les résultats que nous avons obtenus lors de cette campagne d'évaluation (Laville et al., 2020b) et comment il est possible d'abuser des similarités entre les différentes parties des listes d'évaluation.

On note que les meilleurs résultats sont obtenus via l'approche mélangeant les candidats des plongements de mots concaténés et des cognats parfaits (Mix (Conc + Cognates)), et ce pour la majeure partie des paires de langue et marges de fréquence, à l'exception de *en-es* pour les mots à moyenne fréquence (mid) et évidemment les paires *en-ru* dans les deux sens. En effet, le russe et l'anglais possèdent un alphabet différent, il est donc

	en-es				*es-en*			
Fréquence	high	mid	low	all	high	mid	low	all
Skip-gram	60,1	62,8	57,2	60,4	62,5	64,2	65,9	63,9
CBOW	57,1	56,8	54,1	56,4	59,7	60,2	56,2	59,0
Concaténation	60,9	**64,5**	62,8	62,4	62,6	65,5	65,3	64,3
Perfect Cognates	23,3	37,5	63,3	38,3	22,8	37,8	65,4	40,9
Mix (Conc + Cognates)	**61,0**	61,8	**74,4**	**64,3**	**63,5**	**68,6**	**79,1**	**69,5**
	en-de				*de-en*			
Fréquence	high	mid	low	all	high	mid	low	all
Skip-gram	47,6	43,6	29,8	43,4	50,6	47,6	33,7	45,8
CBOW	43,4	41,4	23,0	39,6	45,5	43,9	31,6	41,8
Concaténation	47,9	45,2	30,8	44,3	50,8	50,0	34,0	46,7
Perfect Cognates	21,1	35,6	67,8	37,2	24,1	35,7	69,9	41,2
Mix (Conc + Cognates)	**50,9**	**55,0**	**71,8**	**56,4**	**57,2**	**62,3**	**72,9**	**63,1**
	en-fr				*fr-en*			
Fréquence	high	mid	low	all	high	mid	low	all
Skip-gram	56,5	45,7	31,8	48,0	60,2	49,1	30,3	49,7
CBOW	51,4	42,0	31,1	44,1	58,5	48,7	29,4	48,4
Concaténation	57,8	45,8	34,6	49,3	62,8	55,4	36,2	54,0
Perfect Cognates	27,2	42,7	74,6	45,6	32,5	51,9	75,0	52,0
Mix (Conc + Cognates)	**60,6**	**60,4**	**80,3**	**65,2**	**66,5**	**68,1**	**78,5**	**70,4**
	en-ru				*ru-en*			
Fréquence	high	mid	low	all	high	mid	low	all
Skip-gram	41,3	31,7	13,2	34,0	53,8	40,6	20,7	41,9
CBOW	40,6	28,2	13,7	32,8	49,5	39,5	19,1	39,3
Concaténation	**42,6**	**32,6**	14,4	**35,3**	**55,5**	**44,3**	**22,8**	**44,4**
Perfect Cognates	7,4	6,6	13,2	8,6	0,0	0,0	0,0	0,0
Mix (Conc + Cognates)	42,3	29,9	**21,0**	34,5	-	-	-	-
	de-fr				*fr-de*			
Fréquence	high	mid	low	all	high	mid	low	all
Skip-gram	58,3	41,9	17,4	43,1	56,2	44,0	12,3	42,4
CBOW	52,7	32,7	14,4	36,6	51,2	39,9	11,7	38,5
Concaténation	60,2	44,2	17,9	44,6	56,8	46,9	14,9	44,2
Perfect Cognates	43,4	72,2	82,9	67,4	41,5	68,3	86,9	67,4
Mix (Conc + Cognates)	**67,9**	**78,8**	**85,5**	**77,0**	**62,9**	**74,7**	**87,7**	**74,0**

TABLEAU 5.6 – F1-score pour les différentes approches présentées ainsi que les paires de langues et marges de fréquence étudiées.

normal que l'approche ne fonctionne pas. Cependant, on note un F1-score non nul dans l'approche Perfect Cognates pour la paire *en-ru* (mais pas pour la paire *ru-en*), ce qui

indique que certains mots russes ne sont pas écrits en cyrilliques.

Les meilleurs résultats obtenus par l'approche Mix indiquent une bonne complémentarité entre les deux méthodes, et, comme indiqué précédemment, on peut facilement confirmer cette tendance en regardant les résultats en fonction de la fréquence. Les approches par plongements de mots fonctionnent mieux à haute fréquence et se dégradent d'autant plus que la fréquence des mots baisse. A l'inverse, l'approche Perfect Cognates est très bonne pour les mots basses fréquence, mais la performance diminue sur les mots à plus haute fréquence. Ce déclin est aussi en grande partie dû au fait que les mots à haute fréquence possèdent dans les listes plus de traductions possibles et l'approche Perfect Cognates ne peut par définition prédire au maximum qu'une traduction par mot. Le Tableau 5.7 indique le ratio de traductions cibles par mot source dans les listes que nous utilisons pour la validation.

Language pair	high	mid	low	all
en-es	2,34	1,58	1,10	1,67
en-de	2,83	1,81	1,14	1,93
fr-en	1,64	1,42	1,15	1,40
de-fr	1,08	1,02	1,00	1,03

TABLEAU 5.7 – Ratio du nombre de traductions cibles par mot source dans les listes utilisées pour la validation.

En conclusion, ces expériences nous ont permis de réaliser qu'il est facile d'obtenir de très bons résultats avec des approches très simples qui ne prennent en compte que la proximité graphique des mots. Ce résultat démontre un problème réel dans le jeu de données MUSE, principalement dans les mots les moins fréquents, où l'approche basée sur les cognats parfaits surpasse la meilleure approche basée sur les plongements de mots pour toutes les paires de langues, à l'exception de la paire anglais-russe.

Severini et al. (2022) démontrent l'importance d'utiliser ce type de similarité entre les langues s'il est possible de le faire et proposent même d'étendre ces travaux en utilisant de la translittération pour les paires de langues ayant des alphabets différents dans un contexte non-supervisé.

5.2 Les jeux de données MUSE et MORPH

Dans la section précédente, nous avons présenté différentes expériences en BLI utilisant pour la plupart le jeu de données MUSE mais aussi des scénarios d'évaluation différents.

Nous avons pu remarquer les incohérences et problèmes de ces listes et du processus d'évaluation. Dans cette section, nous étudions de manière approfondie le jeu de données MUSE, mais aussi MORPH (Czarnowska et al., 2019), moins connu, mais qui peut aider à corriger certains problèmes de MUSE et permettre une évaluation plus intéressante.

5.2.1 Analyse du lexique d'évaluation MUSE

Nous présentons d'abord les différents problèmes qu'il est possible de trouver dans les données d'évaluation du jeu de données MUSE. Si ces problèmes sont étendus à l'intégralité du jeu de données, nous nous concentrons ici seulement sur la partie évaluation.

Paires de mots graphiquement identiques

Comme démontré dans les résultats de la Section 5.1.2, les listes d'évaluation MUSE sont composées d'énormément de paires graphiquement identiques. On note même la présence de paires identiques dans la paire anglais-russe (étant donné que les résultats sont supérieurs à zéro pour l'approche se basant sur les cognats parfaits), alors que les deux langues sont écrites avec des alphabets différents. Nous indiquons dans le Tableau 5.8 le pourcentage de mots identiques dans différentes paires de langue des lexiques MUSE, incluant l'allemand, l'anglais, l'espagnol, le français, l'italien et le portugais. Ainsi que des langues seulement liées avec l'anglais, telles que le tchèque, le norvégien et le russe.

	de	en	es	fr	it	pt	avg
de	-	18,5	29,4	49,2	49,8	46,1	38,6
en	16,0	-	16,5	21,0	21,1	18,4	18,6
es	20,3	18,4	-	30,3	31,3	47,9	29,6
fr	41,8	27,5	30,7	-	29,2	24,8	30,8
it	45,8	24,1	32,1	30,8	-	38,0	34,2
pt	40,9	21,6	47,5	27,4	41,2	-	35,7
avg	33,0	22,0	31,2	31,7	34,5	35,0	**31,3**

	en-cs		en-no		en-ru		-
	→	←	→	←	→	←	-
	16,1	17,6	26,1	36,8	2,4	0,0	-

TABLEAU 5.8 – Pourcentage de paires de mots graphiquement identiques entre différentes paires de langues du jeu de données MUSE.

Parmi toutes les paires que nous considérons ici, beaucoup ont plus de 30 % de mots graphiquement identiques. En particulier, les paires allemand-français et allemand-italien

en ont près de la moitié (respectivement 49,2 % et 49,8 %). Par contre, on peut noter que les paires incluant l'anglais ont les pourcentages de paires identiques les plus faibles, ce qui suggère qu'un meilleur contrôle ait pu être fait sur les lexiques anglais ou que la plus grande quantité ou qualité des données anglaises utilisées a permis d'obtenir une meilleure qualité de lexiques générés.

Pour essayer de comprendre pourquoi tant de paires de mots étaient graphiquement identiques malgré des langues plus ou moins éloignées, nous avons étudié précisément les lexiques d'évaluation allemand-français et français-espagnol.

Nous avons extrait les paires graphiquement identiques et les avons manuellement séparées en 4 catégories : Prénoms (P), Entités Nommées (EN), Douteuses (D) (qui sont des paires de mots étant de langues différentes, des acronymes ou tout simplement des erreurs). On précise aussi une sous-catégorie de paires composées de deux mots anglais (AN) (alors que l'anglais ne fait pas partie des langues étudiées) et finalement les paires Correctes (C). Nous reportons le détail de ces annotations dans le Tableau 5.9.

	P	EN	D (AN)	C	Total
de-fr	17,1	28,8	48,9 (21,0)	5,2	767
fr-es	19,6	33,5	40,9 (20,9)	6,0	465

TABLEAU 5.9 – Détail des paires de mots identiques dans les lexiques allemand-français et français-espagnol pour chacune des catégories.

Les catégories Prénoms et Entités Nommées sont toutes les deux des sous-parties des noms propres, mais nous avons préféré les séparer étant donné qu'elles représentent deux concepts différents. Les prénoms (par exemple *Federico* ou *Bryan*) ne sont pas réellement porteurs d'un sens étant donné qu'ils peuvent référencer à des personnes totalement différentes (Pierini, 2008). Par contre, si les EN (des marques, des entités géographiques, des noms de personnes célèbres...) peuvent présenter un intérêt, nous considérons qu'elles ne sont pas l'objectif premier du BLI et être capable d'extraire ces informations correspondraient probablement plus à de la reconnaissance d'entités nommées bilingues. La majeure partie de ces paires est composée de villes ou régions (*orléans, lugano, nebraska*).

Les paires de mots classées comme Douteuses sont majoritairement composées de paires d'autres langues (par exemple *freedom* ou *musica*) mais aussi d'acronymes tels que *nva* (un parti politique belge) ou *avr* (des micro-contrôleurs) qui ne présentent aucun intérêt dans l'évaluation du BLI. On note aussi la présence de paires sans véritables sens comme par exemple *#ffffff* qui représente la couleur blanche en hexadécimal. Le

Tableau 5.10 présente quelques exemples pour la paire de langue français-espagnol.

	Exemples
P	*Pauline, Bryan*
EN	*Orléans, Marx*
D	*musica, #ffffff, avr*
(AN)	*freedom, blood*
C	*terminal, rival*

TABLEAU 5.10 – Quelques exemples de mots composant une paire graphiquement identiques au sein de la liste d'évaluation français-espagnol.

La part restante des paires de mots graphiquement identiques est composée de mots qui sont réellement des cognats parfaits. Par exemple, *terminal* est présent dans les deux paires de langues étudiées. Cette part reste toutefois très faible avec 6 % ou moins de paires correctes.

Kementchedjhieva et al. (2019) ont mis à disposition des sous listes du jeu de données MUSE en les découpant par étiquettes morpho-syntaxique (PoSTag). Les listes comprenant les noms propres nous permettent de nous rendre compte que les noms propres, en plus d'avoir une utilité discutable en BLI comme indiqué précédemment, sont composés à 86 % de paires graphiquement identiques.

Paires de mots graphiquement proches

On note aussi que, même en mettant de côté les paires graphiquement identiques, les lexiques sont composés d'énormément de paires graphiquement proches. En effet, 40,1 % des paires ont une distance de levenshtein inférieure ou égale à 3. Si ces proportions sont assez logiquement plus élevées entre langues romanes (portugais-espagnol 69,8 % ou français-espagnol 57,2 %), il est assez surprenant de voir des paires telles que français-anglais (44,4 %) et italien-anglais (46,5 %) posséder autant de paires de mots similaires alors que le français et l'italien sont des langues romanes et l'anglais est une langue germanique.

Fréquence des mots

Historiquement, la majeure partie des travaux de BLI s'est concentrée sur les mots à haute fréquence. Par exemple, Mikolov et al. (2013b) utilisent les 6 000 mots les plus fréquents pour construire leurs lexiques. De la même manière, Czarnowska et al. (2019)

reportent que les paires de mots des lexiques d'évaluation des jeux de données MUSE font essentiellement partie des 10 000 mots les plus fréquents. Si Jakubina and Langlais (2017) montrent déjà de manière empirique qu'il est bien plus difficile d'obtenir la traduction des mots les moins fréquents, nous rajoutons qu'obtenir la traduction des mots les plus fréquents d'un corpus semble moins intéressant que pour les mots les moins fréquents, car ils seront probablement déjà présents dans des dictionnaires existants.

Fuite sémantique

Czarnowska et al. (2019) indiquent que le jeu de données MUSE souffre de fuite sémantique (*semantic leakage*). C'est-à-dire que certains mots vont apparaître dans les lexiques d'apprentissage alors que des variations morphologiques peuvent apparaître dans la partie évaluation (par exemple, le verbe *tourner* dans une liste d'apprentissage et sa forme conjuguée à la première personne du présent *tourne* dans la partie évaluation). Si Czarnowska et al. (2019) considèrent cette caractéristique comme problématique, nous avançons que dans le cas de traduction humaine, un traducteur est capable de reconnaître les formes conjuguées plus complexes de verbe grâce à sa connaissance de la forme infinitive.

Incohérence morphologique

En plus de posséder très peu de variantes morphologiques pour la plupart de ses mots, le peu de variantes que le jeu de données MUSE contient souffre souvent d'incohérences. Par exemple, dans la liste allemand-français, *allein* n'a comme traduction proposée que la forme masculine (*seul*) et féminine (*seule*) et pas les formes plurielles (*seuls* et *seules*). À l'inverse, *ausgebildet* n'est traduit que par *formé* et *formés*, laissant de côté ses formes féminines. Ces oublis sont parfois accompagnés d'erreurs, par exemple, *christian* est traduit en *chrétien*, *chrétienne* et *chrétiens* (mais aussi par le nom propre *Christian*) alors que *chrétiens* devrait être la traduction de *christians*. On note que la forme féminine plurielle *chrétiennes* n'est pas proposée. La tendance est la même pour les verbes avec, par exemple, *believe* dont les traductions proposées sont *croyez*, *croire*, *croient* et *crois*, mais *croyons* n'est par exemple pas présent.

Nous étudions en Section 5.4 l'impact sur les résultats qu'ont ces différents problèmes que nous venons de présenter.

5.2.2 Analyse du lexique d'évaluation MORPH

Le jeu de données MORPH a été proposé par Czarnowska et al. (2019) dans l'idée de résoudre les trois problèmes principaux du jeu de données MUSE que pointent les auteurs : les faibles variétés fréquentielle (MUSE est composé principalement de mots très fréquents) et morphologique ainsi que la fuite sémantique. Si nous avons établi plus tôt que la fuite sémantique n'était pas un réel problème, car proche du fonctionnement qu'aura un traducteur humain, les deux autres points restent importants.

Avec leurs nouveaux lexiques, Czarnowska et al. (2019) démontrent clairement un lien entre la facilité en BLI et la fréquence des mots (ce que nous avions aussi démontré en Section 5.1.2). La Figure 5.1 illustre ce lien tout en démontrant le problème de MUSE de n'avoir que des paires de mots composés des mots les plus fréquents du vocabulaire.

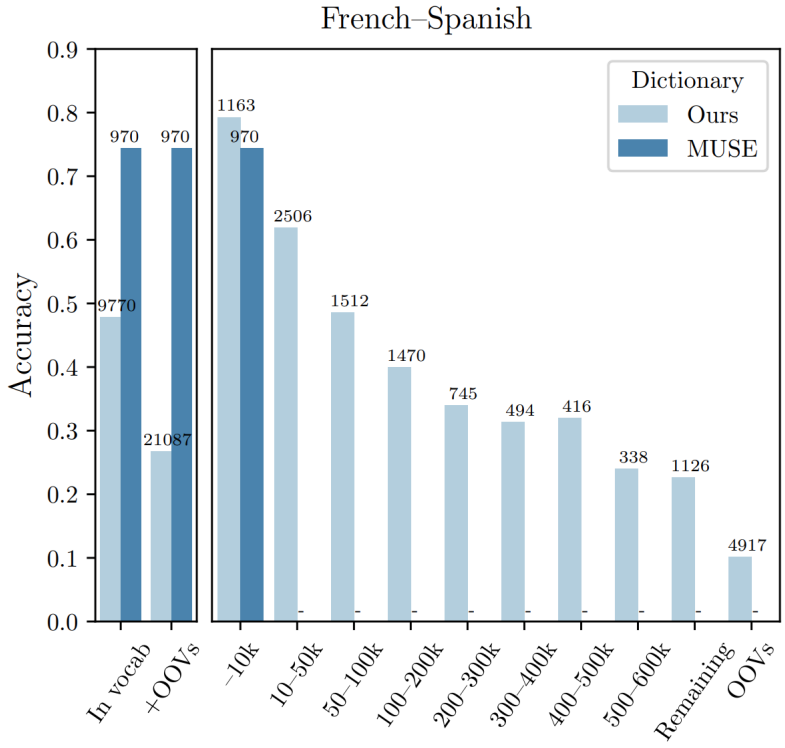

FIGURE 5.1 – Relation entre la Précision (Accuracy) en BLI et la fréquence des mots sources dans le lexique de test. "Ours" correspond au jeu de données MORPH. Le nombre au dessus de chaque barre correspond au nombre de mots sources correctement traduits. Figure issue de l'article de Czarnowska et al. (2019, p. 974).

Cependant, si le jeu de données MORPH corrige certains problèmes de MUSE, il n'est pour autant pas lui-même exempt de tout défaut. Nous donnons dans le Tableau 5.11

l'exemple du lemme source *abanicar* et de ses variations morphologiques ainsi que de leurs traductions comme présenté dans le jeu de données MORPH.

Mot source	Mot cible	Lemme source	Lemme cible	PoSTag
abanique	évente	abanicar	éventer	3 ;PRS ;SBJV ;SG ;V
abanicad	éventez	abanicar	éventer	2 ;IMP ;PL ;POS ;V
42 autres paires		abanicar	éventer	...
abanicaban	éventaient	abanicar	éventer	3 ;IPFV ;PL ;PST ;V
abanicaré	éventerai	abanicar	éventer	1 ;FUT ;SG ;V

TABLEAU 5.11 – Exemple de traductions pour le lemme source *abanicar* en espagnol vers le lemme cible *éventer* en français. Le jeu de données MORPH contient un grand nombre de variations morphologiques (46 ici) et contient les PoSTag correspondants à chacune des paires. Le tableau comprenant l'intégralité des variations du lemme *abanicar* est disponible en Annexe.

Comme montré dans le Tableau 5.11, les listes MORPH ne sont pas utilisables directement : s'il est certes simple d'extraire chaque paire de mots pour obtenir un lexique plus classique (c'est-à-dire composé d'un mot source et d'un mot cible par ligne), on note que pour chacune des paires, on a aussi le lemme et le PoSTag d'indiqué. L'utilisateur du lexique peut donc se questionner sur la pertinence d'utiliser autant de variations morphologiques plutôt que les lemmes. En effet, nous soulignons qu'être capable de traduire toutes les variations morphologiques d'un lemme n'est pas l'objectif premier de la tâche de BLI. C'est pourquoi nous recommandons plutôt l'utilisation des lemmes pour l'évaluation en BLI avec les lexiques MORPH.

De manière similaire, les lexiques MORPH contiennent un grand nombre de propositions de lemmes cibles par mots sources, par exemple, le verbe français *abandonner* possède 21 lemmes candidats différents en italien (*abortire, allentare, arrendere, bandire, cedere, condedere, defezionare, demordere, desistere, disertare, fermare, interrompere, liberare, mollare, piantare, recedere, rinunciare, rinunziare, sfollare, sgomberare, sgombrare*). Si 21 traductions pour un même mot peut déjà paraître élevé, il convient de souligner que pour toutes ces traductions MORPH propose en plus toutes les variations morphologiques de chacun des mots sources et cibles (comme indiqué dans le paragraphe précédent). On peut ainsi atteindre près de 1 000 paires pour un seul lemme source.

Le jeu de données MORPH possède aussi moins de paires de langues que MUSE et, si cela peut être un avantage, n'avoir aucune paire ne comprenant l'anglais reste néanmoins un inconvénient majeur. Cependant, nous recommandons tout de même l'usage du jeu de données MORPH pour la variété qu'il apporte en complément de MUSE, s'il correspond

aux langues étudiées, en ne se concentrant que sur les lemmes et si cela est possible, en réduisant le nombre de traductions différentes par mot source quand elles sont trop nombreuses, même s'il est nécessaire de trouver un moyen de sélectionner les "bonnes" traductions.

5.3 Mesure d'évaluation en BLI

La plupart des travaux de BLI utilisent la précision au rang k (P@k), généralement avec $k \in \{1, 5, 10\}$. Pourtant, Glavaš et al. (2019) conseillent fortement d'utiliser à la place la MAP. Ils indiquent que la MAP est plus informative, car, à l'inverse de la P@k qui pénalise de la même manière une traduction classée au rang $k + 1$ qu'une au rang $k + 100$, la MAP donne une récompense en fonction du rang où la traduction est trouvée.

Nous allons plus loin en pointant que la P@k va seulement chercher à trouver le mot le mieux classé parmi les candidats possibles, alors que la MAP prend en compte chacun des candidats. Le Tableau 5.12 indique le ratio de candidats cibles par mot source dans les lexiques MUSE.

	en-x	x-en	incl. en	no en	avg
MUSE	1,73	1,61	1,67	1,09	1,58

TABLEAU 5.12 – Ratio de mots candidats cibles par mot source dans les listes d'évaluation du jeu de données MUSE.

Avoir plusieurs traductions candidates par mot source est un avantage pour la P@k. En effet, la P@k ne se concentre que sur la traduction candidate la mieux classée. Par exemple, si un mot source possède deux traductions candidates, un système classant une d'entre elle au rang 1 et l'autre au rang 2 {1,2} est considéré de la même manière qu'un système classant les deux traductions {1,100} dans le cas de P@1. En conséquence, avoir plus de traductions candidates donne plus de chances à la P@k de trouver au moins une bonne réponse. Par contre, dans le cas où le système classe les deux traductions {2,3}, la P@1 ne récompense absolument pas le système, au contraire de la MAP qui récompense tout de même le système.

Nous illustrons ces différences entre les deux mesures d'évaluations à l'aide des Tableaux 5.13 et 5.14. Ces tableaux permettent aussi d'illustrer un autre problème de la P@k. Étant donné qu'elle ne se concentre que sur le premier mot candidat correct que le système propose, et que ce mot est souvent le mot le plus fréquent des possibles traduc-

Mot source	Rang	Candidat	P@1	MAP
optional	1	**optionnel**	1	1
	2	chat	0	0
	3	surpris	0	0
	...	...	...	
	100	**facultatif**	0	0,02
Score			1/1	0,51/1

TABLEAU 5.13 – Exemple des récompenses attribuées par la P@1 et la MAP pour un mot source *optional* possédant deux traductions correctes *optionnel* et *facultatif*.

Mot source	Rang	Candidat	P@1	MAP
optional	1	chat	0	0
	2	**optionnel**	0	0,5
	3	**facultatif**	0	0,67
Score			0/1	0,58/1

TABLEAU 5.14 – Exemple des récompenses attribuées par la P@1 et la MAP pour un mot source *optional* possédant deux traductions correctes *optionnel* et *facultatif*.

tions ou alors le plus proche graphiquement, la P@k ignore ces mots et donc aggrave le manque d'évaluation les concernant.

Ainsi, nous appuyons fortement la recommandation de Glavaš et al. (2019) d'utiliser la MAP et démontrons les nouveaux points avancés ici dans la section suivante.

5.4 Expériences

Dans les parties précédentes, nous avons soulevé de nombreux problèmes sur l'évaluation en BLI. La qualité des listes (par exemple, de nombreux noms propres ou de paires graphiquement identiques) ou le protocole d'évaluation (souvent la P@k plutôt que la MAP, mais dans l'ensemble peu de cohérence) sont les deux principaux défauts. Dans cette partie, nous conduisons des expériences pour démontrer l'impact qu'ont ces sujets sur l'évaluation en BLI.

5.4.1 Protocole

Pour appuyer les différents points présentés dans les sections précédentes, nous avons conduit un certain nombre d'expériences sur le corpus Wiki et cinq différentes langues :

anglais, espagnol, français, italien et russe. Les corpus ont été obtenus à l'aide de l'outil WikiExtractor (Attardi, 2015).

Les expériences sont évaluées à l'aide des jeux de données MUSE et MORPH. Pour les lexiques MORPH nous réduisons l'évaluation aux lemmes. Pour le lexique d'entraînement, nous avons utilisé les lexiques d'entraînement pré-construits de MUSE.

Les expériences sont conduites sur deux systèmes :

Alignement (Mapping) nous entraînons des plongements de mots *fastText* et les alignons de manière supervisée à l'aide de l'outil VecMap (Artetxe et al., 2018a).

Entraînement Conjoint (Joint) nous utilisons le *framework joint_align* (Wang et al., 2020) qui entraîne simultanément des plongements de mots *fastText* pour les deux langues puis ajoute une phase d'alignement à partir des plongements entraînés conjointement. Pour l'alignement, la RCSLS est utilisée (Joulin et al., 2018), ce que nous reproduisons.

5.4.2 P@1 vs MAP

Nous indiquons dans le Tableau 5.15 les résultats obtenus sur différentes paires de langues et listes d'évaluation pour deux approches, en comparant les résultats obtenus à l'aide de la P@1 et de la MAP.

		MUSE						MORPH			
		es-fr	fr-it	it-es	en-ru	en-fr	avg	es-fr	fr-it	it-es	avg
Mapping	P@1	84,6	80,5	87,1	44,9	78,8	75,2	57,6	61,9	55,9	58,5
	MAP	87,9	84,4	87,3	51,3	72,8	76,7	45,0	48,7	45,8	46,5
Joint	P@1	65,9	62,6	70,6	34,7	64,5	60,0	43,5	55,3	46,2	48,3
	MAP	71,8	67,5	73,7	39,8	61,1	62,8	37,3	44,8	41,4	41,2
cibles/source		1,02	1,02	1,16	1,63	1,96	1,36	3,37	3,68	2,63	3,22

TABLEAU 5.15 – Résultats détaillés sur différentes paires de langue des approches d'alignement (Mapping) et d'entraînement conjoint (joint) en utilisant deux métriques d'évaluation : P@1 et MAP. Nous indiquons aussi le ratio de mots cibles par mot source (cibles/source) au sein des lexiques d'évaluation.

Nous avons précédemment reporté que la MAP était plus informative, principalement car elle prenait en compte l'intégralité des traductions présentes dans les listes d'évaluation et pas seulement de la mieux classée. Ce tableau confirme cette affirmation et montre que lorsque le ratio de mots cibles par mot source augmente, la P@1 a tendance à dépasser

la MAP, alors que lorsque le ratio se rapproche de 1, la MAP obtient des résultats plus élevés.

On note toutefois une exception, avec la paire de langue *anglais-russe*, où la MAP est au-dessus de la P@1 malgré un ratio de 1,63. En étudiant plus précisément cette paire de langue, nous avons pu nous rendre compte qu'à l'inverse des autres paires, une grande partie des traductions correctes se situait entre le deuxième et cinquième rang, ce qui est récompensé par la MAP. En effet, la P@5 pour cette paire est de 72,0 (+27 points par rapport à la P@1) alors que pour les autres paires de langues, la P@5 est meilleure que la P@1 de moins de 10 points.

Les résultats présentés ici démontrent un lien entre les résultats de la P@k par rapport à la MAP et le ratio de mots cibles par mots sources dans les listes d'évaluation. Plus ce ratio est élevé (c'est-à-dire, plus il y a de mots cibles par mots sources), plus la P@k obtient de meilleurs résultats par rapport à la MAP. Si cela peut déjà être un problème en soit, ce phénomène est aggravé par le fait que les mots sur lesquels la P@k va se concentrer sont les mots "faciles" à traduire (les mots les plus fréquents et proches graphiquement). En effet, sur la liste MUSE *anglais-français*, parmi les 747 mots sources possédant au moins deux candidats traductions cibles, la première proposition du système VecMap (Mapping) est dans 69 % des cas le mot le plus fréquent, 74 % des cas le mot le plus proche graphiquement et correspond à au moins un de ces deux critères dans 92 % des mots sources possédant plus de deux traductions.

Dans le Tableau 5.16, nous illustrons ce phénomène avec trois mots sources et leurs candidats. Nous indiquons aussi le nombre d'occurrences de chacun des mots cibles (#occ).

Mot source	Mot cible	Rang	#occ.
customs	coutumes	1	7 221
	douanes	2	4 165
arch	arche	1	7 407
	voûte	3	541
reveal	révéler	1	7 577
	dévoiler	5	1 858

TABLEAU 5.16 – Trois mots sources de la liste d'évaluation *anglais-français* et les traductions cibles correspondantes ainsi que le rang où le système VecMap (Mapping) les a classé ainsi que le nombre d'occurrences des mots cibles dans leur corpus.

Toutes ces données supportent le fait qu'avoir de multiples traductions possibles par mot source facilite l'augmentation des résultats mesurés en P@1, par rapport à la MAP,

alors qu'intuitivement, avoir plusieurs traductions par mots devrait complexifier le processus en rendant plus difficile de retrouver toutes les traductions (qui peuvent être des synonymes ou dû à des mots sources polysémiques). C'est pourquoi nous reporterons les prochains résultats seulement en MAP plutôt qu'en P@1.

5.4.3 Paires de mots graphiquement proches

Dans le Tableau 5.17, nous reportons les résultats sur différentes paires de langues, pour lesquelles nous avons découpé les listes d'évaluation en plusieurs sous-listes. Dans la première sous-liste (*not id.*), nous supprimons toutes les paires de mots graphiquement identiques. Puis, en repartant de cette sous-liste, nous séparons les paires restantes en deux sous-listes en nous basant sur la distance de Levenshtein : *Far* qui contient tous les mots avec une distance de Levenshtein supérieure à 3 et *Close* qui réunit les paires de mots les proches graphiquement (distance inférieure à 4).

		MUSE						MORPH			
		es-fr	fr-it	it-es	en-ru	en-fr	avg	es-fr	fr-it	it-es	avg
Mapping	*not id.*	88,5	84,0	84,2	50,9	63,8	74,3	41,4	47,5	36,0	41,6
	Far	78,9	71,3	63,3	51,4	46,8	62,3	16,2	19,7	11,5	15,8
	Close	91,2	88,7	87,6	36,4	68,3	74,4	62,9	71,4	58,9	64,4
Joint	*not id.*	68,6	64,0	67,1	39,3	48,9	57,6	33,3	43,4	30,5	35,7
	Far	62,4	55,1	52,8	40,1	35,4	49,2	13,9	19,7	10,6	14,7
	Close	70,4	67,3	69,4	33,7	53,1	58,8	49,0	63,5	45,9	52,8

TABLEAU 5.17 – Résultats en MAP quand les lexiques d'évaluations sont découpés en se basant sur la proximité graphique des différentes paires de mots.

Ce tableau montre clairement une facilité à traduire les mots graphiquement proches et ce, pour les deux méthodes étudiées. En laissant de côté la paire *anglais-russe* (qui est composée de deux langues avec un alphabet différent et pour lesquelles la distance de Levenshtein est surtout une évaluation de la différence de longueur des deux mots), la différence entre *Far* et *Close* est d'au minimum 8 points (sur la paire de langue *espagnol-français* et le lexique MUSE) et monte jusqu'à un maximum de plus de 50 points (*fr-it*, MORPH).

Étant donné que les travaux les plus récents en BLI se sont majoritairement évalués à l'aide des jeux de données MUSE (Hakimi Parizi and Cook, 2020; Wang et al., 2020; Ormazabal et al., 2021; Severini et al., 2022), les performances démontrées peuvent donc

être considérées comme très optimistes, étant donné la grande proportion de paires identiques dans ces lexiques. C'est pourquoi nous recommandons d'évaluer plus précisément les approches en allant étudier les paires de mots éloignées graphiquement.

Pour continuer sur les proximités graphiques entre les langues, nous illustrons en Figure 5.2 la relation entre la distance de Levenshtein moyenne entre les paires de mots d'un lexique et le score de MAP obtenu pour les approches d'alignement et d'entraînement conjoint. Ce graphique montre clairement que les meilleurs résultats sont obtenus sur les listes d'évaluations les plus proches graphiquement et non pas sur les paires de langues les plus proches. En effet, les résultats sont bien meilleurs pour les listes MUSE que pour les listes MORPH pour des paires de langues identiques (ces résultats sont aussi dus aux mots moins fréquents). On note aussi que la paire *anglais-russe* (qui sont deux langues très éloignées) présente des résultats similaires à la paire *français-italien* dans le cas du lexique MORPH qui contient énormément de mots éloignés graphiquement.

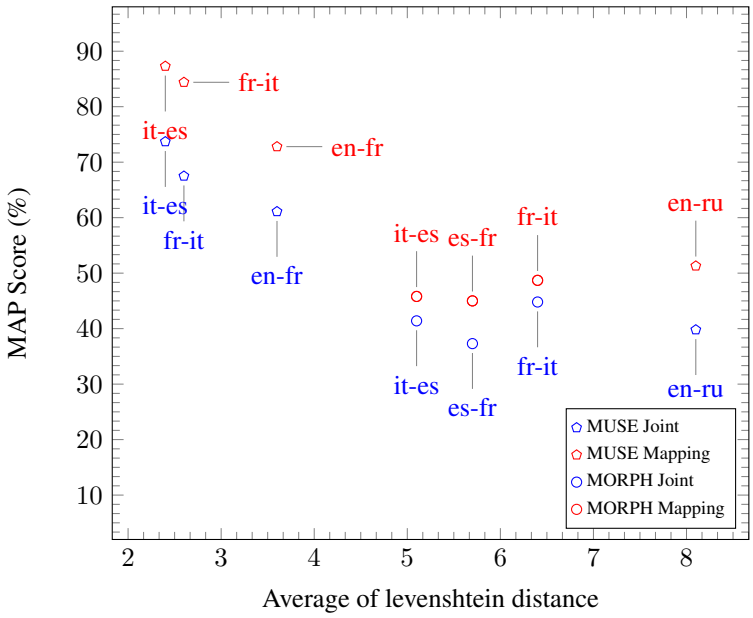

FIGURE 5.2 – MAP en relation avec la distance de Levenshtein moyenne des listes d'évaluation pour différentes paires de langues étudiées.

5.5 Synthèse

Nous avons étudié dans ce chapitre le processus d'évaluation en BLI en nous concentrant d'abord sur différents types de méthodes puis sur la qualité des lexiques utilisés

pour l'évaluation. Nous avons évalué le BLI selon différents critères : la fréquence des mots, l'éloignement graphique des paires étudiées ou encore la présence des mots dans un dictionnaire monolingue. Tous les résultats que nous avons présentés dans ce manuscrit démontrent que l'évaluation en BLI a lieu sur des paires de mots relativement simples à traduire voire même peu utiles. Nous avons donc cherché dans ce chapitre à formuler un certain nombre de recommandations qui, nous l'espérons, permettront de mieux s'intéresser aux réelles difficultés du BLI. En effet, même si certains systèmes proposent des résultats très satisfaisants, il est aujourd'hui très difficile d'obtenir des résultats intéressants sur des catégories de mots particuliers, comme par exemple les mots peu fréquents ou éloignés graphiquement, comme nous avons pu le démontrer à l'aide des différentes expériences conduites au sein de ce chapitre.

Premièrement, dans un cadre d'évaluation utilisant MUSE nous recommandons la suppression des paires de mots graphiquement identiques. Comme nous l'avons montré au cours de ce chapitre, les lexiques MUSE sont composés d'énormément de paires de mots identiques qui, en plus d'être d'un intérêt minime dans un cadre de traduction, sont pour beaucoup des paires incorrectes. Ensuite, l'utilisation de plusieurs jeux de données est aussi une piste intéressante pour renforcer les résultats des systèmes, ainsi, nous conseillons, lorsque c'est possible, d'évaluer les systèmes de BLI sur les jeux de données MUSE et MORPH (ou tout autre jeu de données correspondant aux langues étudiées). Finalement, nous recommandons d'étudier les lexiques sous différents angles, par exemple en les séparant en différents groupes basés sur la distance de Levenshtein ou la fréquence des mots. Les résultats que nous avons présentés précédemment démontrent clairement que les systèmes performent bien mieux sur les paires de mots graphiquement proches ou sur les mots les plus fréquents.

Nous militons aussi à nouveau pour l'utilisation de la MAP en lieu et place de la P@k, particulièrement si les lexiques d'évaluation comportent plusieurs mots cibles par mot source. La mesure d'évaluation de MAP prenant en compte tous les mots recherchés, la mesure est plus à même d'étudier la polysémie qui, comme on a pu le voir en Chapitre 4 d'une autre manière, peut être un problème en BLI.

Finalement, comme simplement regarder les résultats en MAP ne donne pas réellement les informations sur les forces et faiblesses d'un système, nous recommandons aussi une évaluation plus précise en sélectionnant certains mots sources pour étudier les candidats proposés par les différents systèmes.

Conclusion

Conclusion et contributions

Nous nous sommes intéressés dans cette thèse à l'extraction de lexiques bilingues à partir de corpus comparables pour les domaines spécialisé et général. Le premier concerne les corpus spécialisés qui se concentrent sur un domaine précis (comme la médecine ou l'énergie) et ont par conséquent un vocabulaire spécifique. Les domaines spécialisés sont complexes à étudier du fait de la difficulté à réunir des corpus de qualité sur un seul et unique domaine. Le second se rapporte simplement au domaine général, pour lequel les corpus vont, à l'inverse du domaine spécialisé être composés d'un vocabulaire très large et aborder de nombreux sujets différents.

Le premier chapitre de ce manuscrit illustre les différences entre ces deux types de corpus, ainsi que la nuance entre les corpus comparables (que nous utilisons) et les corpus parallèles (qui sont des textes en relation de traduction) et présente les différents types de ressources nécessaires qu'utilisent les systèmes de BLI. Cette présentation des types de données est suivie par une introduction des différents jeux de données que nous utilisons au cours de cette thèse.

Les approches de BLI utilisent des représentations sémantiques des mots, de manière à pouvoir les comparer. Les approches historiques considéraient qu'un mot ne possédait qu'une représentation basée sur l'ensemble des contextes de ses occurrences au sein d'un texte, mais des approches plus récentes ont commencés à produire des représentations pour chacune des occurrences d'un mot, en fonction de son contexte actuel. Nous présentons dans le Chapitre 2 différents types de représentations sémantiques des mots.

Ces représentations sont originellement créées pour des tâches monolingues et dans le câde du BLI, il devient nécessaire de les comparer entre les langues. Le Chapitre 3 introduit donc différentes méthodes, que l'on peut séparer en deux familles distinctes. D'abord les méthodes d'alignement, qui utilisent des représentations pré-entraînées et les projettent dans un espace commun pour permettre la comparaison. Mais aussi les méthodes d'entraînement conjoint, qui se basent sur les méthodes d'apprentissage des représentations en les adaptant pour introduire dans l'objectif d'apprentissage une information bilingue.

Les chapitres 4 et 5 regroupent nos contributions.

Dans le Chapitre 4, nous nous intéressons à la problématique de BLI dans le cadre du domaine spécialisé, avec les contraintes que nous avons déjà présentées : des données plus rares et des systèmes peu adaptés à si peu de données et un vocabulaire plus complexe et restreint. Nous proposons d'appliquer des techniques de sélection de données aux corpus généraux pour éviter d'ajouter des données qui seraient trop éloignés du domaine originel ce qui a pour conséquence d'obtenir des représentations sémantiques plus fidèles au domaine spécialisé étudié. Ce système permet d'améliorer les résultats en BLI en domaine spécialisé de 6 points, tout en réduisant les temps de calcul d'un facteur 10 par rapport à l'approche d'augmentation de données proposée par Hazem and Morin (2016, 2018).

Le chapitre 5 se concentre majoritairement sur le BLI en domaine général. Dans un premier temps, nous avons étudié différents scénarios d'évaluation, en faisant varier certains paramètres au sein des systèmes ou en ne s'évaluant que sur certaines sous-parties des listes d'évaluation. Ces expériences nous ont permis de réaliser l'incohérence et les problèmes des données utilisées pour l'évaluation, ce qui nous a amené à chercher à quel point il était possible d'abuser de ces problèmes pour en démontrer l'absurdité. Lors de la campagne d'évaluation BUCC 2020, nous avons proposé d'utiliser en combinaison aux approches classiques utilisant des plongements de mots, un système binaire sélectionnant une traduction pour un mot source de la liste d'évaluation dans le cas où un mot graphiquement identique apparaîtrait dans le corpus cible. Ce système très simple obtient de meilleurs résultats sur les mots peu fréquents que les meilleures approches utilisant des réseaux de neurones. Ces résultats nous ont menés à conduire une étude plus précise des données d'évaluation utilisées en domaine général et plus globalement, du processus d'évaluation complet. Les systèmes actuels ne s'évaluent que trop rarement sur des mots peu fréquents ou éloignés graphiquement, ce qui est dommageable car il y a encore beaucoup à faire pour ces cas de figure. Nous avons alors cherché à proposer des solutions et des conseils pour obtenir une évaluation en BLI plus uniforme et précise. Pour conclure, l'évaluation en lexique bilingue doit être accompagnée d'une étude précise des résultats et pas seulement de la présentation d'un score. Il est nécessaire de chercher à comprendre plus précisément ce qui se passe pour pouvoir améliorer les systèmes.

Perspectives

Nos contributions se sont concentrées dans un premier temps sur les domaines de spécialité, avec la sélection de données, nous avons démontré qu'il était possible d'améliorer

les résultats en ne sélectionnant qu'un pourcentage des données générales, vues comme les plus proches du domaine étudié. La quantité de données générales à ajouter au corpus de spécialité change d'un mot à l'autre. Développer une méthode de sélection sensible à cela est donc une piste à étudier.

Nous avons aussi présenté une approche utilisant les plongements de mots contextualisés pour la tâche de BLI (Schuster et al., 2019). Un des avantages des plongements de mots contextualisés est la possibilité d'isoler les occurrences des mots en fonction de leur contexte et donc de leur sens. Il devient donc envisageable de sélectionner les occurrences d'un mot qui sont porteuses du sens qui nous intéresse au sein du domaine spécialisé et donc d'éventuellement supprimer ce problème de polysémie.

Dans le domaine général, nous avons étudié manuellement certaines listes du jeu de données MUSE, probablement trop peu. Malheureusement, faire le nécessaire sur l'intégralité des listes nécessiterait des connaissances dans un grand nombre de langues ou bien l'intervention d'annotateurs externes, ce qui pourrait être intéressant pour vérifier que les sources d'erreurs sont les mêmes pour toutes les paires de langues ou si l'on retrouve des spécificités dues à certaines langues.

Plus globalement, il serait nécessaire de restituer des listes d'évaluation plus équilibrées entre MUSE et MORPH. Même si ces jeux de données ont malheureusement peu de paires de langues en commun, une combinaison pourrait apporter un équilibre qui, accompagné d'un système d'évaluation adapté, pourrait permettre de s'assurer de la comparabilité des résultats entre les différentes approches.

Ce système d'évaluation pourrait directement permettre d'étudier les résultats en fonction des catégories de mots que nous avons définies comme étant plus complexes à évaluer (par exemple les mots peu fréquents ou les paires éloignées graphiquement). Il serait alors possible de conduire une étude complète et dans un cadre commun des différents systèmes qui ont été proposés ces dernières années et permettrait de mieux comprendre quelles sont les forces des différents systèmes, car ils ont souvent été évalués dans des scénarios différents.

Liste des publications

Word Representations, Seed Lexicons, Mapping Procedures, and Reference Lists : What Matters in Bilingual Lexicon Induction from Comparable Corpora ? (Laville et al., 2020c)

Martin Laville, Mérième Bouhandi, Emmanuel Morin et Philippe Langlais

Abstract : Methods for bilingual lexicon induction are often based on word embeddings (WE) similarity. These methods must be able to project the WE to the same space. Uncontextualized WE proved to be useful for this task. We compare them to contextualized WE and Bag of Words, using specialized and general datasets. We also evaluate the impact of seed lexicons and check the existing reference lists validity, claiming that extracting the translation of some words in those lists is not useful and confirming the need to have more fine-grained reference lists.

TALN/LS2N Participation at the BUCC Shared Task : Bilingual Dictionary Induction from Comparable Corpora (Laville et al., 2020b)

Martin Laville, Amir Hazem et Emmanuel Morin

Abstract : This paper describes the TALN/LS2N system participation at the Building and Using Comparable Corpora (BUCC) shared task. We first introduce three strategies : (i) a word embedding approach based on fastText embeddings ; (ii) a concatenation approach using both character Skip-gram and character CBOW models, and finally (iii) a cognates matching approach based on an exact match string similarity. Then, we present the applied strategy for the shared task which consists in the combination of the embeddings concatenation and the cognates matching approaches. The covered languages are French, English, German, Russian and Spanish. Overall, our system mixing embeddings concatenation and perfect cognates matching obtained the best results while compared to individual strategies, except for English-Russian and Russian-English language pairs for which the concatenation approach was preferred.

Data Selection for Bilingual Lexicon Induction from Specialized Comparable Corpora (Laville et al., 2020a)

Martin Laville, Amir Hazem, Emmanuel Morin et Philippe Langlais

Abstract : Narrow specialized comparable corpora are often small in size. This particularity makes it difficult to build efficient models to acquire translation equivalents, especially for less frequent and rare words. One way to overcome this issue is to enrich the specialized corpora with out-of-domain resources. Although some recent studies have shown improvements using data augmentation, the enrichment method was roughly conducted by adding out-of-domain data with no particular attention given to how to enrich words and how to do it optimally. In this paper, we contrast several data selection techniques to improve bilingual lexicon induction from specialized comparable corpora. We first apply two well-established data selection techniques often used in machine translation that is : Tf-Idf and cross entropy. Then, we propose to exploit BERT for data selection. Overall, all the proposed techniques improve the quality of the extracted bilingual lexicons by a large margin. The best performing model is the cross entropy, obtaining a gain of about 4 points in MAP while decreasing computation time by a factor of 10.

About Evaluating Bilingual Lexicon Induction (Laville et al., 2022)

Martin Laville, Emmanuel Morin et Philippe Langlais

Abstract : With numerous new methods proposed recently, the evaluation of Bilingual Lexicon Induction have been quite hazardous and inconsistent across works. Some studies proposed some guidance to sanitize this ; yet, they are not necessarily followed by practitioners. In this study, we try to gather these different recommendations and add our owns, with the aim to propose an unified evaluation protocol. We further show that the easiness of a benchmark while being correlated to the proximity of the language pairs being considered, is even more conditioned on the graphical similarities within the test word pairs.

Annexe

Mot source	Mot cible	Lemme source	Lemme cible	PoSTag
abanique	évente	abanicar	éventer	3 ;PRS ;SBJV ;SG ;V
abanicad	éventez	abanicar	éventer	2 ;IMP ;PL ;POS ;V
abaniques	éventes	abanicar	éventer	2 ;PRS ;SBJV ;SG ;V
abanicaremos	éventerons	abanicar	éventer	1 ;FUT ;PL ;V
abaniqué	éventai	abanicar	éventer	1 ;PFV ;PST ;SG ;V
abanicabais	éventiez	abanicar	éventer	2 ;IPFV ;PL ;PST ;V
abanicamos	éventâmes	abanicar	éventer	1 ;PFV ;PL ;PST ;V
abaniquemos	éventions	abanicar	éventer	1 ;PL ;PRS ;SBJV ;V
abanicabas	éventais	abanicar	éventer	2 ;IPFV ;PST ;SG ;V
abaniquen	éventent	abanicar	éventer	3 ;PL ;PRS ;SBJV ;V
abaniquéis	éventiez	abanicar	éventer	2 ;PL ;PRS ;SBJV ;V
abanica	évente	abanicar	éventer	3 ;PRS ;SG ;V
abanicó	éventa	abanicar	éventer	3 ;PFV ;PST ;SG ;V
abanicaseis	éventassiez	abanicar	éventer	2 ;PL ;PST ;SBJV ;V
abanicáis	éventez	abanicar	éventer	2 ;PL ;PRS ;V
abanicasteis	éventâtes	abanicar	éventer	2 ;PFV ;PL ;PST ;V
abanicarías	éventerais	abanicar	éventer	2 ;COND ;SG ;V
abanicaríais	éventeriez	abanicar	éventer	2 ;COND ;PL ;V
abanicase	éventât	abanicar	éventer	3 ;PST ;SBJV ;SG ;V
abanique	évente	abanicar	éventer	1 ;PRS ;SBJV ;SG ;V
abanicaréis	éventerez	abanicar	éventer	2 ;FUT ;PL ;V
abanico	évente	abanicar	éventer	1 ;PRS ;SG ;V
abanicábamos	éventions	abanicar	éventer	1 ;IPFV ;PL ;PST ;V
abanica	évente	abanicar	éventer	2 ;IMP ;POS ;SG ;V
abanicaría	éventerait	abanicar	éventer	3 ;COND ;SG ;V
abanicarás	éventeras	abanicar	éventer	2 ;FUT ;SG ;V
abanicará	éventera	abanicar	éventer	3 ;FUT ;SG ;V
abanicaba	éventais	abanicar	éventer	1 ;IPFV ;PST ;SG ;V
abanicarán	éventeront	abanicar	éventer	3 ;FUT ;PL ;V
abanicaste	éventas	abanicar	éventer	2 ;PFV ;PST ;SG ;V
abanicaba	éventait	abanicar	éventer	3 ;IPFV ;PST ;SG ;V
abanicases	éventasses	abanicar	éventer	2 ;PST ;SBJV ;SG ;V
abanican	éventent	abanicar	éventer	3 ;PL ;PRS ;V
abanicaría	éventerais	abanicar	éventer	1 ;COND ;SG ;V
abanicamos	éventons	abanicar	éventer	1 ;PL ;PRS ;V
abanicásemos	éventassions	abanicar	éventer	1 ;PL ;PST ;SBJV ;V
abanicarían	éventeraient	abanicar	éventer	3 ;COND ;PL ;V
abanicaron	éventèrent	abanicar	éventer	3 ;PFV ;PL ;PST ;V
abaniquemos	éventons	abanicar	éventer	1 ;IMP ;PL ;POS ;V
abanicar	éventer	abanicar	éventer	NFIN ;V
abanicase	éventasse	abanicar	éventer	1 ;PST ;SBJV ;SG ;V
abanicas	éventes	abanicar	éventer	2 ;PRS ;SG ;V
abanicaríamos	éventerions	abanicar	éventer	1 ;COND ;PL ;V
abanicasen	éventassent	abanicar	éventer	3 ;PL ;PST ;SBJV ;V
abanicaban	éventaient	abanicar	éventer	3 ;IPFV ;PL ;PST ;V
abanicaré	éventerai	abanicar	éventer	1 ;FUT ;SG ;V

Tableau 5.18

BIBLIOGRAPHIE

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

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2016. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2289–2294, Austin, Texas. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 451–462.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018a. Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018b. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL'18)*, pages 789–798, Melbourne, Australia.

Mikel Artetxe, Sebastian Ruder, Dani Yogatama, Gorka Labaka, and Eneko Agirre. 2020. A call for more rigor in unsupervised cross-lingual learning. *arXiv preprint arXiv :2004.14958*.

Giusepppe Attardi. 2015. Wikiextractor. `https://github.com/attardi/wikiextractor`.

Jean-Julien Aucouturier and Francois Pachet. 2008. A scale-free distribution of false positives for a large class of audio similarity measures. *Pattern Recogn.*, 41(1) :272–284.

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain adaptation via pseudo in-domain data selection. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP'11)*, pages 355–362, Edinburgh, Scotland, UK.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The wacky wide web : a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, 43(3) :209–226.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5 :135–146.

Dhouha Bouamor, Nasredine Semmar, and Pierre Zweigenbaum. 2013. Context Vector Disambiguation for Bilingual Lexicon Extraction from Comparable Corpora. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL'13)*, pages 759–764, Sofia, Bulgaria.

Yun-Chuang Chiao and Pierre Zweigenbaum. 2002. Looking for candidate translational equivalents in specialized, comparable corpora. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING'02)*, pages 1208–1212, Tapei, Taiwan.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv :1710.04087*.

Paula Czarnowska, Sebastian Ruder, Edouard Grave, Ryan Cotterell, and Ann Copestake. 2019. Don't forget the long tail! a comprehensive analysis of morphological generalization in bilingual lexicon induction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 974–983, Hong Kong, China. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Georgiana Dinu, Angeliki Lazaridou, and Marco Baroni. 2014. Improving zero-shot learning by mitigating the hubness problem.

Ted E Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational linguistics*, 19(1) :61–74.

Long Duong, Hiroshi Kanayama, Tengfei Ma, Steven Bird, and Trevor Cohn. 2016. Learning crosslingual word embeddings without bilingual corpora. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1285–1295, Austin, Texas. Association for Computational Linguistics.

Robert M Fano. 1961. Transmission of information : A statistical theory of communications. *American Journal of Physics*, 29(11) :793–794.

Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 462–471, Gothenburg, Sweden. Association for Computational Linguistics.

John R Firth. 1957. A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*.

Ana Frankenberg-Garcia. 2009. Are translations longer than source texts. *A corpus-based study of explicitation In : Beeby, A., Rodríguez P., & Sánchez-Gijón, P.(eds.) Corpus use and learning to translate (CULT) : An Introduction. Amsterdam & Philadelphia : John Benjamins*, pages 47–58.

Pascale Fung. 1998. A Statistical View on Bilingual Lexicon Extraction : From Parallel Corpora to Non-parallel Corpora. In *Proceedings of the 3rd Conference of the Association for Machine Translation in the Americas on Machine Translation and the Information Soup (AMTA'98)*, pages 1–17, Langhorne, PA, USA.

Goran Glavaš, Robert Litschko, Sebastian Ruder, and Ivan Vulić. 2019. How to (properly) evaluate cross-lingual word embeddings: On strong baselines, comparative analyses, and some misconceptions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 710–721, Florence, Italy. Association for Computational Linguistics.

Stephan Gouws and Anders Søgaard. 2015. Simple task-specific bilingual word embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, pages 1386–1390, Denver, Colorado. Association for Computational Linguistics.

Ali Hakimi Parizi and Paul Cook. 2020. Joint training for learning cross-lingual embeddings with sub-word information without parallel corpora. In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, pages 39–49, Barcelona, Spain (Online). Association for Computational Linguistics.

Zellig S Harris. 1954. Distributional structure. *Word*, 10(2-3) :146–162.

Amir Hazem and Emmanuel Morin. 2016. Efficient data selection for bilingual terminology extraction from comparable corpora. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics : Technical Papers*, pages 3401–3411, Osaka, Japan. The COLING 2016 Organizing Committee.

Amir Hazem and Emmanuel Morin. 2018. Leveraging meta-embeddings for bilingual lexicon extraction from specialized comparable corpora. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 937–949, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

R. Hicklin, Craig Watson, and Brad Ulery. 2005. The myth of the goats: How many people have fingerprints that are hard to match?

Laurent Jakubina and Philippe Langlais. 2016. A comparison of methods for identifying the translation of words in a comparable corpus : recipes and limits. *Computación y Sistemas*, 20(3) :449–458.

Laurent Jakubina and Phillippe Langlais. 2017. Reranking translation candidates produced by several bilingual word similarity sources. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL'17)*, pages 605–611, Valencia, Spain.

Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave. 2018. Loss in translation: Learning bilingual word mapping with a retrieval criterion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2984, Brussels, Belgium. Association for Computational Linguistics.

David Kamholz, Jonathan Pool, and Susan Colowick. 2014. PanLex: Building a resource for panlingual lexical translation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3145–3150, Reykjavik, Iceland. European Language Resources Association (ELRA).

Martin Kay, Martin Röscheisen, et al. 1994. Text-translation alignment. *Computational linguistics*, 19(1) :121–142.

Yova Kementchedjhieva, Mareike Hartmann, and Anders Søgaard. 2019. Lost in evaluation: Misleading benchmarks for bilingual dictionary induction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3336–3341, Hong Kong, China. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X : Papers*, pages 79–86, Phuket, Thailand.

Moshe Koppel and Noam Ordan. 2011. Translationese and its dialects. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies*, pages 1318–1326, Portland, Oregon, USA. Association for Computational Linguistics.

Audrey Laroche and Philippe Langlais. 2010. Revisiting context-based projection methods for term-translation spotting in comparable corpora. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 617–625, Beijing, China. Coling 2010 Organizing Committee.

Martin Laville, Mérième Bouhandi, Emmanuel Morin, and Philippe Langlais. 2020a. Word representations, seed lexicons, mapping procedures, and reference lists : What matters in bilingual lexicon induction from comparable corpora ? In *Canadian Conference on Artificial Intelligence*, pages 349–355. Springer.

Martin Laville, Amir Hazem, and Emmanuel Morin. 2020b. TALN/LS2N participation at the BUCC shared task: Bilingual dictionary induction from comparable corpora. In *Proceedings of the 13th Workshop on Building and Using Comparable Corpora, BUCC@LREC 2020, Marseille, France, May, 2020*, pages 56–60. European Language Resources Association.

Martin Laville, Amir Hazem, Emmanuel Morin, and Phillippe Langlais. 2020c. Data selection for bilingual lexicon induction from specialized comparable corpora. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6002–6012, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Martin Laville, Emmanuel Morin, and Philippe Langlais. 2022. About evaluating bilingual lexicon induction. In *LREC 2022 Workshop Language Resources and Evaluation Conference 25 June 2022*, page 8.

Angeliki Lazaridou, Georgiana Dinu, and Marco Baroni. 2015. Hubness and pollution: Delving into cross-space mapping for zero-shot learning. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, pages 270–280, Beijing, China. Association for Computational Linguistics.

Bo Li and Éric Gaussier. 2010. Improving corpus comparability for bilingual lexicon extraction from comparable corpora. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING'10)*, pages 644–652, Beijing, China.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta : A robustly optimized bert pretraining approach. *arXiv preprint arXiv :1907.11692*.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

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

Tomás Mikolov, Quoc V. Le, and Ilya Sutskever. 2013b. Exploiting similarities among languages for machine translation. *CoRR*, abs/1309.4168.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013c. Distributed representations of words and phrases and their compositionality.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013d. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, pages 746–751, Atlanta, Georgia. Association for Computational Linguistics.

Robert C. Moore and William D. Lewis. 2010. Intelligent selection of language model training data. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL'10)*, pages 220–224, Uppsala, Sweden.

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

Dragos Stefan Munteanu and Daniel Marcu. 2006. Extracting parallel sub-sentential fragments from non-parallel corpora. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 81–88, Sydney, Australia. Association for Computational Linguistics.

Aitor Ormazabal, Mikel Artetxe, Gorka Labaka, Aitor Soroa, and Eneko Agirre. 2019. Analyzing the limitations of cross-lingual word embedding mappings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4990–4995, Florence, Italy. Association for Computational Linguistics.

Aitor Ormazabal, Mikel Artetxe, Aitor Soroa, Gorka Labaka, and Eneko Agirre. 2021. Beyond offline mapping: Learning cross-lingual word embeddings through context anchoring. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, pages 6479–6489, Online. Association for Computational Linguistics.

Barun Patra, Joel Ruben Antony Moniz, Sarthak Garg, Matthew R. Gormley, and Graham Neubig. 2019. Bilingual lexicon induction with semi-supervision in non-isometric

embedding spaces. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 184–193, Florence, Italy. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Patrizia Pierini. 2008. Opening a pandora's box: Proper names in english phraseology. *Linguistik Online*, 36.

Milos Radovanovic, Alexandros Nanopoulos, and Mirjana Ivanovic. 2010. Hubs in space : Popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research*, 11(sept) :2487–2531.

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

Reinhard Rapp, Pierre Zweigenbaum, and Serge Sharoff. 2020. Overview of the fourth bucc shared task : Bilingual dictionary induction from comparable corpora. In *13th Workshop on Building and Using Comparable Corpora (BUCC)*, pages 6–13.

Nils Reimers and Iryna Gurevych. 2019. Alternative weighting schemes for elmo embeddings. *arXiv preprint arXiv :1904.02954*.

Anthony Rousseau. 2013. Xenc : An open-source tool for data selection in natural language processing. *The Prague Bulletin of Mathematical Linguistics*, pages 73–82.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2019. A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, 65 :569–631.

Tal Schuster, Ori Ram, Regina Barzilay, and Amir Globerson. 2019. Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing. In *Proceedings of the 2019 Conference of the North American Chapter of the*

Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers), pages 1599–1613, Minneapolis, Minnesota. Association for Computational Linguistics.

Silvia Severini, Viktor Hangya, Masoud Jalili Sabet, Alexander Fraser, and Hinrich Schütze. 2022. Don't forget cheap training signals before building unsupervised bilingual word embeddings. *arXiv preprint arXiv :2205.15713*.

Serge Sharoff, Reinhard Rapp, and Pierre Zweigenbaum. 2013. Overviewing important aspects of the last twenty years of research in comparable corpora. *Building and using comparable corpora*, pages 1–17.

Samuel L. Smith, David H. P. Turban, Steven Hamblin, and Nils Y. Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax.

Anders Søgaard, Sebastian Ruder, and Ivan Vulić. 2018. On the limitations of unsupervised bilingual dictionary induction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 778–788, Melbourne, Australia. Association for Computational Linguistics.

Jörg Tiedemann. 2012. Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Zirui Wang, Jiateng Xie, Ruochen Xu, Yiming Yang, Graham Neubig, and Jaime G Carbonell. 2020. Cross-lingual alignment vs joint training : A comparative study and a simple unified framework. In *International Conference on Learning Representations*.

Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. 2015. Normalized word embedding and orthogonal transform for bilingual word translation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, pages 1006–1011, Denver, Colorado. Association for Computational Linguistics.

Zheng Zhang, Ruiqing Yin, Jun Zhu, and Pierre Zweigenbaum. 2019. Cross-lingual contextual word embeddings mapping with multi-sense words in mind. *arXiv :1909.08681*.

COLLEGE ED MATHEMATIQUES
DOCTORAL STIC
PAYS DE LA LOIRE

Nantes Université

Titre : Évaluation en extraction de lexiques bilingues à partir de corpus comparables

Mot clés : plongements de mots bilingues, évaluation, extraction de lexiques bilingues

Résumé : L'extraction de lexique bilingue (BLI) a pour objectif la création, de manière automatique à partir de corpus bilingues, de lexiques entre deux langues. Le BLI est utilisé le plus souvent en domaine général, où les lexiques extraits peuvent par exemple servir en traduction automatique ou en recherche d'information. Les systèmes de BLI fonctionnent alors sur de grandes quantités de données et les résultats semblent hautement satisfaisants. Cependant, les données d'évaluation contiennent de nombreuses erreurs, ce qui pourrait conduire à une remise en question des systèmes. Un second contexte d'utilisation plus marginal du BLI est celui des domaines de spécialité, où l'objectif est l'obtention de traductions absentes des dictionnaires classiques. Les corpus spécialisés (qui ne concernent qu'un seul sujet) sont peu fournis en données et il est compliqué pour les systèmes de BLI d'obtenir d'aussi bons résultats qu'en domaine général. Il faut donc chercher à adapter les approches pour prendre en compte cette particularité. Dans cette thèse, nous améliorons les résultats obtenus en BLI en domaine de spécialité en proposant l'utilisation de techniques de sélection de données. Puis, nous nous intéressons au processus d'évaluation en domaine général et plus particulièrement à certains biais présents dans les données d'évaluation comme la surprésence de paires de mots très fréquents ou graphiquement identiques et proposons un processus d'évaluation plus précis et unifié qui prend en compte ces faiblesses dans les données.

Title: Evaluating bilingual lexicon induction using comparable corpora

Keywords: bilingual word embeddings, evaluation, bilingual lexicon induction

Abstract: Bilingual lexicon extraction (BLI) has as its objective the creation, in an automatic manner from bilingual corpora, of lexicons between two languages. It is most often used in the general domain, where the extracted lexicons can be used in machine translation or information retrieval. BLI systems work on large amounts of data and the results seem to be highly satisfactory. However, the evaluation data contains many errors, which could lead to a re-evaluation of the systems. A second and more marginal context of use of BLI systems is in specialized domains, where the objective is to obtain translations that are not available in classical dictionaries. Specialized corpora (about only one subject) are poorly supplied with data and it is complicated for BLI systems to obtain as good results as in the general domain. It is therefore necessary to adapt the approaches to take into account this particularity. In this thesis, we improve the results obtained in specialized domains by proposing the use of data selection techniques. Then, we focus on the evaluation process in general domain and more particularly on some biases present in evaluation data such as the overpresence of very frequent or graphically identical word pairs and we propose a more accurate and unified evaluation process that takes into account these weaknesses.