Université de Montréal

L'extraction de phrases en relation de traduction dans Wikipédia

par
Lise Rebout

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

Mémoire présenté à la Faculté des arts et des sciences
en vue de l'obtention du grade de Maître ès sciences (M.Sc.)
en informatique

Juin 2012

© Lise Rebout, 2012.

Université de Montréal
Faculté des arts et des sciences

Ce mémoire intitulé:

L'extraction de phrases en relation de traduction dans Wikipédia

présenté par:

Lise Rebout

a été évalué par un jury composé des personnes suivantes:

Guy Lapalme,	président-rapporteur
Philippe Langlais,	directeur de recherche
Pascal Vincent,	membre du jury

Mémoire accepté le 19 juillet 2012

RÉSUMÉ

Afin d'enrichir les données de corpus bilingues parallèles, il peut être judicieux de travailler avec des corpus dits comparables. En effet dans ce type de corpus, même si les documents dans la langue cible ne sont pas l'exacte traduction de ceux dans la langue source, on peut y retrouver des mots ou des phrases en relation de traduction. L'encyclopédie libre Wikipédia constitue un corpus comparable multilingue de plusieurs millions de documents. Notre travail consiste à trouver une méthode générale et endogène permettant d'extraire un maximum de phrases parallèles. Nous travaillons avec le couple de langues français-anglais mais notre méthode, qui n'utilise aucune ressource bilingue extérieure, peut s'appliquer à tout autre couple de langues. Elle se décompose en deux étapes. La première consiste à détecter les paires d'articles qui ont le plus de chance de contenir des traductions. Nous utilisons pour cela un réseau de neurones entraîné sur un petit ensemble de données constitué d'articles alignés au niveau des phrases. La deuxième étape effectue la sélection des paires de phrases grâce à un autre réseau de neurones dont les sorties sont alors réinterprétées par un algorithme d'optimisation combinatoire et une heuristique d'extension. L'ajout des quelques 560 000 paires de phrases extraites de Wikipédia au corpus d'entraînement d'un système de traduction automatique statistique de référence permet d'améliorer la qualité des traductions produites. Nous mettons les données alignées et le corpus extrait à la disposition de la communauté scientifique.

Mots clés : alignement de phrases, réseaux de neurones, corpus comparables, classifieurs, systèmes de traduction statistiques, algorithmes d'optimisation combinatoire.

ABSTRACT

Working with comparable corpora can be useful to enhance bilingual parallel corpora. In fact, in such corpora, even if the documents in the target language are not the exact translation of those in the source language, one can still find translated words or sentences. The free encyclopedia Wikipedia is a multilingual comparable corpus of several millions of documents. Our task is to find a general endogenous method for extracting a maximum of parallel sentences from this source. We are working with the English-French language pair but our method – which uses no external bilingual resources – can be applied to any other language pair. It can best be described in two steps. The first one consists of detecting article pairs that are most likely to contain translations. This is achieved through a neural network trained on a small data set composed of sentence aligned articles. The second step is to perform the selection of sentence pairs through another neural network whose outputs are then re-interpreted by a combinatorial optimization algorithm and an extension heuristic. The addition of the 560 000 pairs of sentences extracted from Wikipedia to the training set of a baseline statistical machine translation system improves the quality of the resulting translations. We make both the aligned data and the extracted corpus available to the scientific community.

Keywords sentence alignment, neural networks, comparable corpora, classifiers, statistical machine translation, combinatorial optimization algorithms.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES SIGLES

BLEU	Bilingual Evaluation Understudy
CLIR	Recherche d'information interlangue (*CrossLingual Information Retrieval*)
CRF	Champs conditionnels aléatoires (*Conditional Random Field*)
RALI	Recherche appliquée en informatique linguistique, laboratoire du département d'informatique de l'Université de Montréal
SMT	Système de traduction statistique (*Statistical Machine Translation*)
SVM	Séparateur à vaste marge
TALN	Traitement automatique des langues naturelles
TF-IDF	Fréquence du terme-Fréquence inverse de document (*term frequency–inverse document frequency*)
WMT 2011	Sixième Atelier de traduction automatique statistique (*Sixth Workshop on Stratistical Machine Translation 2011*)

REMERCIEMENTS

Merci...

Merci tout d'abord à mon directeur de recherche M. Philippe Langlais qui m'a accordé sa confiance et qui m'a soutenue, conseillée et inspirée tout le long de ma maîtrise.

Merci ensuite aux professeurs du DIRO pour leur enseignement et aux membres du laboratoire RALI pour l'ambiance de travail et leurs conseils judicieux.

Merci à mes parents et à mes frères qui m'ont encouragée et qui ont toujours cru en moi ; merci à ma mère en particulier qui a eu le courage de relire mon travail.

Merci à Mathieu pour son soutien indéfectible, sa bonne humeur permanente et son anglais remarquable.

Et merci enfin à Lune et à Camille qui m'ont incitée à entamer cette maîtrise.

INTRODUCTION

Wikipedia , ويكيبيديا, Wikipédia, Viquipèdia[1],... Le nom de la célèbre encyclopédie libre en ligne, malgré des différences de graphie et d'orthographe, est le même dans toutes les langues. Créée en 2001, Wikipédia a connu un succès rapide qui s'est traduit non seulement par la multiplication du nombre d'articles en anglais mais aussi par la création d'un Wikipédia multilingue, reflet de la diversité linguistique de ses rédacteurs, les internautes. Cette masse de données textuelles, disponibles librement, est une mine pour les chercheurs en traitement automatique des langues et en recherche d'information, à condition de savoir l'exploiter efficacement. Prenons par exemple le cas des traductions. L'existence de projets internes à Wikipédia destinés à coordonner les traductions d'articles en d'autres langues[2] nous indique qu'il existe bien des équivalences entre certains articles. Néanmoins, de par le caractère libre et participatif de l'encyclopédie, le contenu des articles fluctue rapidement et deux textes équivalents à une date *d* peuvent subir des transformations très différentes au cours du temps. D'autre part, les rédacteurs ne sont pas dans l'obligation d'indiquer ce qu'ils ont traduit, ni même de rester fidèle au texte source. Nous nous retrouvons donc devant une certaine quantité de phrases en relation de traduction, perdues dans une masse de texte inintéressante quant au bilinguisme. Ces traductions, si elles sont de bonne qualité et en quantité suffisante, peuvent cependant se révéler intéressantes pour servir de carburant à des systèmes de traduction statistique dont les performances dépendent en partie du volume de données sur

[1] "Wikipédia" écrit en anglais, arabe, français et catalan.

[2] Par exemple `http://de.wikipedia.org/wiki/Wikipedia:%C3%9Cbersetzungen` pour les traductions vers l'allemand, `http://fr.wikipedia.org/wiki/Projet:Traduction` pour les traductions vers le français.

lesquelles on les entraîne.

Notre travail est de trouver une méthode pour extraire des phrases en relation de traduction entre la version française et la version anglaise de Wikipédia. Comme nous souhaitons développer une méthode générique, indépendante des langues prises en considération et des ressources bilingues disponibles pour ce couple de langues, notre objectif est de ne pas utiliser de matériel extérieur à Wikipédia, comme un dictionnaire bilingue ou un bitexte déjà aligné. Nous envisageons donc ce problème comme un problème de classification où les instances à classer sont l'ensemble des couples de phrases (fr_i, en_j) où $i \in 1,...,N_{fr}$ et $j \in 1,...,N_{en}$, N_{fr} et N_{en} étant respectivement le nombre de phrases dans la version française et le nombre de phrases dans la version anglaise de Wikipédia. Le volume de texte ne nous permettant pas d'effectuer une classification sur l'ensemble des données[3], nous décidons de restreindre l'espace de recherche aux couples d'articles qui d'une part sont reliés par un lien interlangue dans Wikipédia et qui d'autre part ont un taux de parallélisme important. Nous effectuons cette sélection de deux manières différentes, l'une utilisant un moteur de recherche, ce qui est une méthode classique, l'autre un classifieur. Cette seconde approche est plus originale et nous voulons en tester la validité. Vient ensuite l'étape de la classification des couples de phrases. Nous implémentons là aussi un classifieur. N'ayant pas de matériel bilingue extérieur à notre disposition, nous calculons une série de traits originaux sur les couples de phrases pour nous permettre de discriminer les traductions du reste. L'évaluation du matériel extrait nous indique un gain dans la qualité des traductions produites par un traducteur statistique lorsque nous ajoutons les couples de phrases parallèles issues de Wikipédia à son ensemble d'entraînement. Nous mettons à la disposition de la communauté scientifique l'ensemble des corpus d'entraînement et des corpus de tests constitués de couples d'articles Wikipédia alignés manuellement, ainsi que le meilleur bitexte que nous ayons extrait.

Ce travail s'inscrit dans une série de recherches portant sur les corpus comparables en

[3]Il y a 3.7 millions d'articles en anglais et 1.1 en français.

général. C'est pourquoi nous définissons dans le premier chapitre les différents types de corpus comparables et nous montrons comment Wikipédia s'inscrit dans cette typologie. Nous détaillons ensuite les travaux effectués dans le domaine des corpus comparables bilingues, notamment la construction de dictionnaires et lexiques bilingues et l'extraction de phrases et de segments sous-phrastiques. Le troisième chapitre est consacré à la description de nos travaux à savoir la préparation du corpus, le tri des articles, le tri des phrases et la validation des résultats. Le quatrième et dernier chapitre compile nos résultats et les compare à ceux de travaux similaires au nôtre. Nous concluons en posant un regard critique sur notre méthodologie et en envisageant différents axes de recherche pour améliorer nos performances.

CHAPITRE 1

CORPUS COMPARABLE

1.1 Définition

Il existe plusieurs manières de définir un corpus comparable. Que ces définitions soient basées sur des critères communicationnels, lexicaux ou linguistiques, elles restent peu précises et il n'existe pas de mesure de comparabilité de deux corpus reconnue par l'ensemble de la communauté TALN.

Laffling [1992] nous donne une première définition très large : " [Comparable corpora are] texts which, though composed independently in the respective language communities, have the same communicative function". Les textes composant un corpus comparable expriment donc des idées semblables ou parlent des mêmes faits, mais ils ne sont pas la traduction l'un de l'autre comme le soulignent Bowker et Pearson [2002].

Déjean et Gaussier [2002] expriment la comparabilité d'un corpus en fonction de la couverture lexicale des textes. Ils affirment en effet que "deux corpus de deux langues l1 et l2 sont dits comparables s'il existe une sous-partie non négligeable du vocabulaire du corpus de langue l1, respectivement l2, dont la traduction se trouve dans le corpus de langue l2, respectivement l1." Li et Gaussier [2010] ont ensuite proposé une mesure de comparabilité quantitative définie comme la proportion du vocabulaire pour lequel il existe une traduction dans le corpus. Ils montrent que la qualité des dictionnaires extraits de corpus comparables est directement reliée à ce degré de comparabilité.

Fung et Cheung [2004] ont été les premiers à définir une typologie des corpus comparables. Les différents types de corpus ne sont pas classés selon leur vocabulaire commun mais selon leur degré de parallélisme :

– *Les corpus parallèles* sont alignés phrase à phrase c'est-à-dire qu'ils sont la traduction (idéalement) parfaite l'un de l'autre. Le Hansard canadien, c'est-à-dire la

transcription officielle des débats parlementaires, est un corpus parallèle aligné anglais - français fréquemment utilisé pour des recherches en traduction automatique.

- *Les corpus parallèles bruités* contiennent des phrases qui ne sont pas alignées. Ils peuvent contenir des paragraphes insérés ou supprimés. Les articles issus de service de presse en différentes langues sont ainsi utilisés comme corpus parallèle bruité (par exemple par Gahbiche-Braham et al. [2011]).
- *Les corpus comparables* ne sont pas alignés au niveau des phrases, ils contiennent des documents bilingues qui ne sont pas la traduction l'un de l'autre mais qui portent sur les mêmes thèmes. Des collections de sites Web sur un sujet donné – par exemple la médecine comme dans Chiao et Zweigenbaum [2002a] – ou d'articles de journaux publiés pendant une période donnée constituent de tels corpus comparables.
- *Les corpus quasi-comparables* contiennent des documents non-alignés au niveau des phrases, qui ne sont pas la traduction l'un de l'autre et qui peuvent porter sur le même sujet (in-topic) ou non (off-topic). On peut penser ici au corpus TDT[1] qui est constitué de multiples flux d'informations dans plusieurs langues et formats (agences de presse, radio, télévision, sites Web) (Wu et Fung [2005]).

Cette typologie, même si elle n'est que qualitative, a été utilisée dans plusieurs travaux subséquents et elle est souvent citée comme référence.

Pour finir sur une note géométrique, nous aimerions citer Church (2009) qui représente les corpus comparables et parallèles par des nuages de points : les termes du texte source sont représentés en abscisse, ceux du texte cible en ordonnées et un point est marqué lorsque deux mots correspondent l'un à l'autre ; les diagonales qui apparaissent sur le graphe sont significatives d'alignements ordonnés, les carrés de correspondances non alignées. Alors que pour les corpus parallèles, les diagonales seraient longues, les corpus comparables contiendraient plus de carrés et des diagonales plus courtes et plus

[1] Topic Detection and Tracking : `http://projects.ldc.upenn.edu/TDT3/`

subtiles.

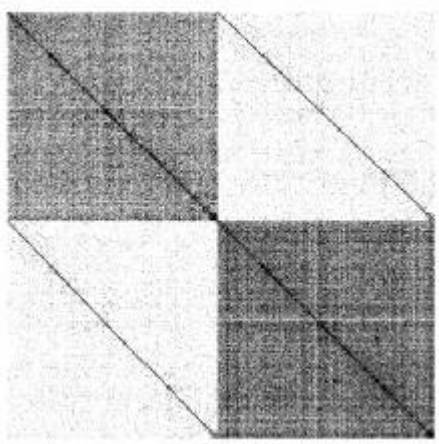

Figure 1.1: Un simple nuage de points.
Nuage de points représentant trois ans du Hansard canadien en anglais et français. Le texte source est concaténé au texte cible. Un point est placé aux coordonnées *i,j* si le mot *i* est identique au mot *j*. Les deux carrés sombres en haut à gauche et en bas à droite montrent les répétitions des mots à l'intérieur d'une langue. les régions claires et les diagonales qui les traversent représentent les correspondances entre les deux langues (Church [1993]).

En somme, les corpus comparables utilisés en traitement automatique des langues peuvent avoir des caractéristiques relativement différentes pour lesquelles il faudra adapter les méthodes utilisées : intuitivement nous pensons que plus le corpus est parallèle, plus les méthodes applicables sur les corpus parallèles ont des chances d'obtenir des résultats satisfaisants.

1.2 Pourquoi les corpus comparables

Traditionnellement les corpus bilingues utilisés dans les recherches en TALN sont des corpus parallèles. Il existe des algorithmes éprouvés pour aligner les phrases de deux textes qui sont la traduction l'un de l'autre (Langlais et al. [1998]). De nombreuses méthodes s'appuient sur la distance de Gale & Church, d'après les travaux des deux chercheurs au début des années 1990 (Church [1993]) qui visent à modéliser le rapport

entre la longueur de la phrase source et celle de la phrase cible. Ce score est ensuite utilisé dans des algorithmes de programmation dynamique qui en général recherchent les phrases parallèles dans un faisceau autour de la diagonale.

Dans les corpus comparables, les phrases, ou les mots, susceptibles d'être la traduction l'un de l'autre ne se trouvent pas forcément le long de la diagonale, ce qui altère les résultats obtenus par les méthodes précédentes. Pourtant il peut être nécessaire d'utiliser ces corpus. En effet, la qualité des résultats produits par les systèmes de traduction automatique, notamment par ceux utilisant des méthodes statistiques (SMT), est fortement corrélée aux volumes des données utilisées pour l'entraînement. Les courbes de performance de Koehn et al. [2003] le montrent bien. Lorsqu'on veut traduire des textes d'un domaine de spécialité, il est également préférable d'entraîner ces traducteurs sur des textes du même domaine (Langlais et al. [2006]). Or pour certaines paires de langues, les corpus parallèles sont très pauvres ou inexistants, que ce soient des corpus parallèles ou de spécialité, car il existe peu de traductions entre ces deux langues. On peut penser par exemple à un couple de langues exotiques comme "albanais - coréen ", mais même dans des couples de langues plus courants, dans certains domaines, il est difficile de trouver de réelles traductions. Le recours aux corpus comparables bilingues, formés de deux corpus monolingues constitués indépendamment peut alors s'avérer utile car, comme on le verra par la suite, ils contiennent du matériel linguistique de qualité.

1.3 Wikipédia comme corpus comparable ?

1.3.1 Description de Wikipédia

Wikipedia est une encyclopédie numérique, multilingue, libre et universelle officiellement née en janvier 2001. Alors que Nupedia, le projet original des deux fondateurs Jimmy Wales et Larry Sanger, était rédigée et corrigée par des experts, n'importe qui peut participer à l'écriture des articles de Wikipédia. L'édition et la publication des textes sont réalisées grâce à un logiciel appelé moteur de wiki dont la création par Ward Cunnin-

gham remonte à 1995. La modification ou l'ajout des textes se fait en ligne directement dans le navigateur Web.

Le succès de Wikipédia est tel que le site de référencement Alexa[2] le place à la 6e place des sites les plus consultés au monde en février 2012 (après Google, Facebook, Youtube, Yahoo ! et Baidu.com).

1.3.2 Le multilinguisme dans Wikipédia

Les premiers articles publiés dans Wikipédia l'ont été en anglais. La croissance de l'encyclopédie a été exponentielle dès les premiers jours : le 22 janvier 2001, on pouvait comptabiliser 24 articles, deux mois plus tard, ce nombre se montait déjà à 2 953 pages.[3]

Après l'anglais, ce sont l'allemand, le catalan et le japonais qui ont été les trois premières langues à avoir elles aussi une version de Wikipédia.[4] La création officielle de la version française de Wikipédia, c'est-à-dire la création d'une page d'accueil en français, a eu lieu quant à elle le 23 mars 2001 mais ce n'est que le 19 mai qu'un article a été publié (Article sur le physicien Paul Héroult).[5]

Depuis, la croissance des deux Wikipédia – français et anglais – a été fulgurante pour atteindre plus de 3 861 000 articles en anglais et plus de 1 208 000 en français. Le nombre de langues présentes dans Wikipédia a lui aussi explosé pour passer de 20 langues fin 2002 à 250 fin 2006. En décembre 2011, les articles de Wikipédia sont rédigés dans 283 langues différentes, dont certaines très rares comme le Hiri Motu parlé en Papouasie-Nouvelle-Guinée ou encore le Min Dong parlé en Chine. Un système de conversion automatique permet également l'affichage dans différents systèmes d'écriture, par exemple en serbe.[6]

Les articles sur un même sujet présents dans deux langues différentes sont en général

[2] http://www.alexa.com/topsites
[3] http://marc.info/?l=wikipedia-l&m=104216623605869&w=2
[4] http://web.archive.org/web/20010331173908/http://www.wikipedia.com/
[5] http://fr.wikipedia.org/wiki/Wikip%C3%A9dia_en_fran%C3%A7ais
[6] http://sr.wikipedia.org/wiki/

liés entre eux par une référence dans la rubrique “Autres langues” (voir figure 1.2). Si on ne considère que l’ensemble de ces articles, nous sommes donc bien en présence d’un corpus comparable selon la définition de Laffling.

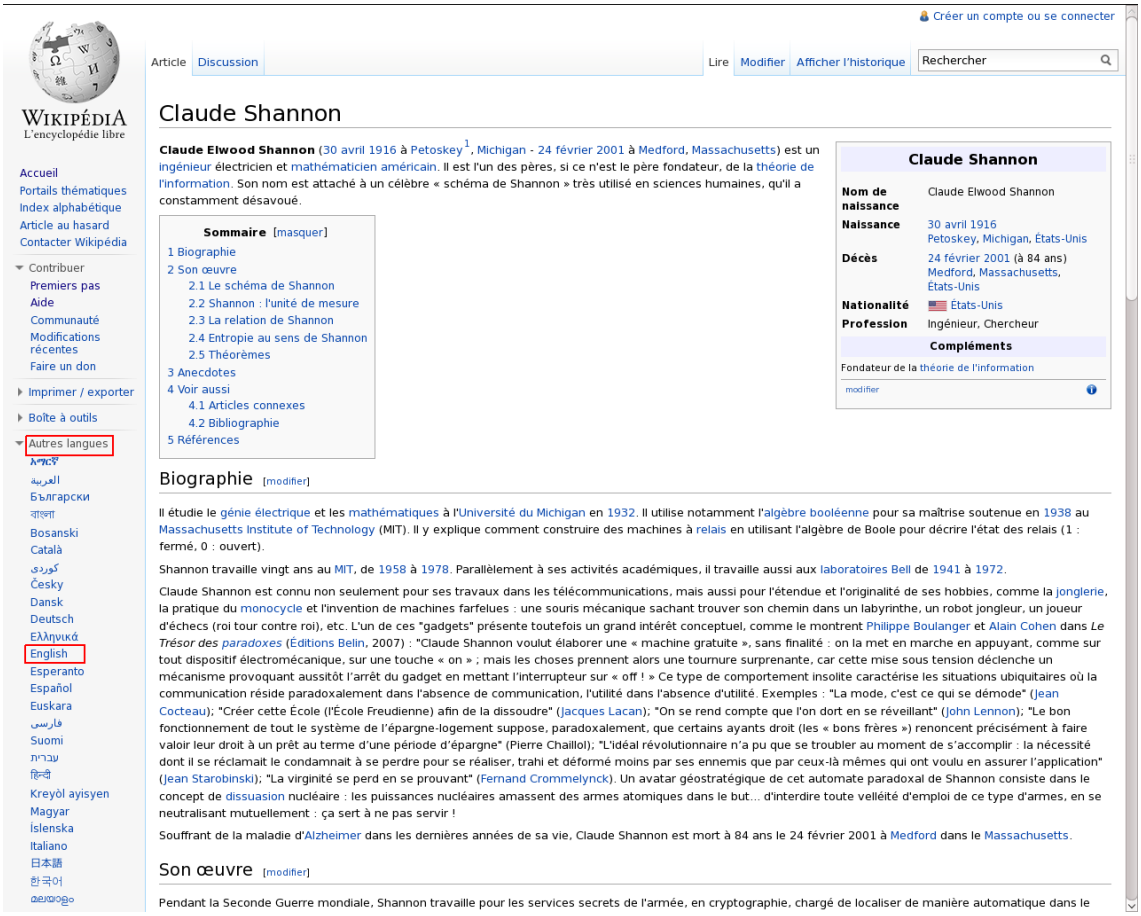

Figure 1.2: Les liens interlangues dans Wikipédia. Exemple de l’article en français sur Claude Shannon. Le premier cadre rouge dans la marge gauche indique le début de la liste des articles sur Claude Shannon dans d’autres langues, le deuxième le lien vers l’article anglais sur le père de la théorie de l’information.

1.3.3 Le contenu de Wikipédia

Rédigée par le citoyen lambda (Foglia [2008]), Wikipédia reflète ses connaissances et ses centres d’intérêt. Les thèmes abordés sont nettement différents de ceux traités dans une encyclopédie traditionnelle et font la part belle à la culture populaire et aux

sujets d'actualité. Chacun peut donc rédiger ce qu'il veut à condition de respecter les principes au coeur de Wikipédia : la neutralité de point de vue, la vérifiabilité des articles et l'absence de travaux inédits. La communauté de Wikipédiens s'autorégule, veille à ce que personne ne déroge à ces principes, corrige les erreurs, améliore ou actualise les articles et tente de contrer les tentatives de vandalisme. Le contenu des articles est donc particulièrement instable. L'article en français consacré à Alan Turing a ainsi subi 40 modifications de mai 2011 (date du téléchargement des données) jusqu'en décembre 2011.

Figure 1.3: Le mode de rédaction de Wikipédia. Source : Foglia [2008]

1.3.4 Statistiques sur Wikipédia anglais et français

Les statistiques compilées dans le tableau 1.I sont issues des statistiques officielles publiées mensuellement par Wikipédia[7].

1.3.5 Syntaxe de Wikipédia

Les articles Wikipédia sont rédigés dans un langage de balises. Il existe une syntaxe particulière pour mettre en forme le texte, pour définir des liens interwiki (balises délimitées par des doubles crochets), pour insérer des références (balises délimitées par

[7] `http://stats.wikimedia.org/pourlemoisdemai2011`

	Wikipedia (total)	Anglais	Français
Nombre de Wikipediens	1 356 748	716 553	68 054
Nombre d'articles	18.8 millions	3.7 millions	1.1 millions
Nombre de nouveaux articles par jour	8037	1025	397
Nombre moyen de révisions par article		78.4	40.5
Taille du dump (décompressé)		31 Go	8 Go

Tableau 1.I: Statistiques sur Wikipédia

des chevrons) et pour utiliser des modèles que le moteur de wiki remplace par du code HTML (balises délimitées par des doubles accolades).

Valérie Chansigaud, administratrice de Wikipédia France, parle de la « complexité croissante de l'écriture comme de l'aspect normatif du projet » (Foglia [2008]). Le tableau 1.II montre un exemple typique de phrases dans Wikipédia avec sa correspondance en HTML et à l'écran.

1.3.6 Wikipédia, un corpus comparable de qualité ?

Utiliser Wikipédia comme corpus comparable présente plusieurs avantages. Tout d'abord, la masse de données est intéressante. Le nombre d'articles anglais et français liés par un lien interlangue se monte à 551 388. Quelques données quantitatives sur ces articles sont fournies dans le tableau 1.III.

Le volume de texte disponible se situe à mi-chemin entre des énormes corpus comparables comme GigaWord (dont la version anglaise contient 1 756 504 kilo-mots[8]) et des corpus comparables plus modestes comme le TDT3 utilisé par Fung et Cheung [2004] qui compte 290 000 phrases anglaises et 110 000 phrases chinoises.

Un autre aspect intéressant est le caractère encyclopédique et actuel des données récoltées. Mis à jour constamment, Wikipédia est le reflet des sujets d'intérêt des internautes. La variété des thèmes abordés permet de constituer un corpus à la fois général et spécialisé. Ainsi, le portail sur les mathématiques compte 5 603 articles en français

[8] `http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2003T05`

'''Alan Mathison Turing''', [[Ordre de l'Empire britannique|OBE]], [[Royal Society|FRS]] ({{Date|23|juin|1912}} − {{Date|7|juin|1954}}) est un [[mathématicien]] [[Royaume−Uni|britannique]], auteur de l'article fondateur de la science [[informatique]]<ref>{{en}} [http://www.thocp.net/biographies/papers/turing_oncomputablenumbers_1936.pdf ''On Computable Numbers with an Application to the Entscheidungsproblem'']</ref> qui allait donner le coup d'envoi à la création des calculateurs universels programmables (ordinateurs).

Alan Mathison Turing, OBE, FRS (23 juin 1912 − <ahref ="/wiki/7_juin" title="7 juin">7 <a href="/wiki/Juin" title="Juin">juin 1954) est un mathématicien britannique, auteur de l'article fondateur de la science informatique^{[1]} qui allait donner le coup d'envoi à la création des calculateurs universels programmables (ordinateurs).

Alan Mathison Turing, OBE, FRS(23 juin 1912 − 7 juin 1954) est un mathématicien britannique, auteur de l'article fondateur de la science informatique[1] qui allait donner le coup d'envoi à la création des calculateurs universels programmables (ordinateurs).

Tableau 1.II: Un exemple de la syntaxe de Wikipédia, avec l'équivalent en HTML et à l'écran.

et 27 287 en anglais. Le style des articles se rapproche également plus du style courant, utilisé au quotidien que celui des corpus comme le Hansard. Grâce à la teneur de son contenu, Wikipédia pourrait donc compléter des corpus parallèles ou des dictionnaires statistiques pour développer des traducteurs automatiques généraux. Comme l'ensemble de l'encyclopédie est sous licence *Creative Commons*, tout le matériel est utilisable librement.[9]

Enfin, il faut noter que certains projets de traduction existent dans Wikipédia. Des bénévoles se proposent de traduire et de relire certains articles intéressants qui n'existent que dans une autre langue. La page du projet en français aide à la coordination des

[9] `http://en.wikipedia.org/wiki/Wikipedia:Text_of_Creative_Commons_Attribution-ShareAlike_3.0_Unported_License`

	Anglais	Francais
Nombre de phrases	33 027 506	24 330 642
Nombre de mots	470 392 619	304 444 932
Nombre moyen de phrases par article	59.90	43.13
Nombre moyen de mots par article	853.11	552.14
Nombre moyen de mots par phrase	14.25	12.51

Tableau 1.III: Statistiques sur les données en français et en anglais dans Wikipédia

travaux en cours. Une balise spéciale peut être incluse à la fin d'un article traduit ou dans sa page de commentaires pour indiquer la langue source.

Néanmoins, la structure et le mode de rédaction de l'encyclopédie peuvent nuire à la qualité du corpus. En effet comme tout un chacun peut rédiger ou modifier des articles, la qualité du texte est parfois médiocre. Par exemple, certaines traductions sont partielles, non relues voire erronées. Il n'est pas rare également que le traducteur prenne de grandes libertés par rapport au texte original. D'autre part, les modifications apportées aux articles sont constantes. Les articles qui sont parallèles à un moment donné peuvent être modifiés à la fois du côté de la langue source et de la langue cible et s'éloigner l'un de l'autre sans que la balise indiquant une traduction ne soit ôtée. Enfin, en terme de comparabilité, Wikipédia constitue un corpus très hétérogène : si l'on reprend l'échelle de comparabilité proposée par Fung et Cheung, on peut trouver parmi les articles liés par un lien interlangue des sous-corpus correspondant à chacune des catégories décrites. Un exemple de chaque type est indiqué dans la liste ci-dessous.

Corpus parallèle Bataille de Trois-Rivières (versions du 25 janvier 2012 [10]).

Corpus parallèle bruité Umbriel (Lune). Le tableau 1.IV montre un extrait du premier paragraphe "Structures de surface" de cet article. L'article français est marqué comme étant la traduction de l'article anglais correspondant. Mais on se rend compte rapidement que certains passages du texte français ne sont pas dans le texte anglais.

[10] `http://fr.wikipedia.org/wiki/Bataille_de_Trois-Rivi%C3%A8res`

Corpus comparable Par exemple l'article sur la lune[11].

Corpus quasi-comparable Certains articles, même s'ils partagent le même titre ont des contenus totalement distincts. L'article sur l'*Humanité*[12], lié par un lien interlangue à l'article anglais *Human nature*[13] est un bon exemple. Alors que le premier dresse la liste de ce qui différencie l'homme des animaux puis revient sur l'histoire de "l'unité de l'humanité" au cours des siècles et sur la protection juridique de l'espèce humaine, le second se penche en détails sur les différentes visions philosophiques, psychologiques et biologiques de la nature humaine.

Wikipédia constitue donc un corpus comparable très hétérogène, un aspect à prendre en compte pour la suite du travail.

[11] `http://fr.wikipedia.org/wiki/Lune` et `http://en.wikipedia.org/wiki/Moon`

[12] `http://fr.wikipedia.org/wiki/Humanit%C3%A9`

[13] `http://en.wikipedia.org/wiki/Human_nature`

Anglais	Français
Umbriel's surface is the darkest of the Uranian moons, and reflects less than half as much light as Ariel, a sister satellite of similar size.	Umbriel est le plus sombre des satellites d'Uranus, qui sont eux-mêmes plus sombres que les satellites des planètes plus proches du Soleil. Par exemple, la surface d'Ariel, un satellite jumeau d'à peu près la même taille, est plus de deux fois plus lumineuse.
Umbriel has a very low Bond albedo of only about 10% as compared to 23% for Ariel.	Umbriel a un albédo de Bond d'environ 10%, à comparer aux 23 % d'Ariel.
The reflectivity of the moon's surface decreases from 26% at a phase angle of 0°(geometric albedo) to 19% at an angle of about °.This phenomenon is called opposition surge.	La réflectivité de la surface du satellite décroît de 26 % à l'angle de phase de 0°(albédo géométrique) à 19 % pour un angle de phase de 1°, c'est l'effet d'opposition.
The surface of Umbriel is slightly blue in color, while fresh bright impact deposits (in Wunda crater, for instance) are even bluer.	Contrairement à ce qui est observé pour Obéron (une autre lune uranienne sombre), la surface observée d'Umbriel est légèrement bleuâtre. En plus, des dépôts récents d'impact apparaissent, très clairs et au bleu encore plus prononcé (notamment dans le cratère Wunda, près de l'équateur).
There may be an asymmetry between the leading and trailing hemispheres ; the former appears to be redder than the latter.	Il se peut qu'il y ait une asymétrie entre les hémisphères avant et arrière. Ce dernier paraîtrait ainsi plus rouge que le premier. Cette asymétrie n'est pas complètement établie car la surface du satellite n'est connue qu'à 40 %.
The reddening of the surfaces probably results from space weathering from bombardment by charged particles and micrometeorites over the age of the Solar System.	Le rougissement de la surface serait la conséquence de l'érosion spatiale due au bombardement par les particules chargées et les micrométéorites depuis le début du système solaire.
However, the color asymmetry of Umbriel is likely caused by accretion of a reddish material coming from outer parts of the Uranian system, possibly, from irregular satellites, which would occur predominately on the leading hemisphere.	Cependant, l'asymétrie de couleur d'Umbriel est probablement aussi causée par l'accumulation de matériaux rougeâtres provenant des parties externes du système d'Uranus, peut-être des satellites irréguliers. Cette accumulation se produirait de préférence sur l'hémisphère avant.
The surface of Umbriel is relatively homogeneous—it does not demonstrate strong variation in either albedo or color.	La surface d'Umbriel est relativement homogène : elle ne présente pas de fortes variations en albédo ou en couleur.

Tableau 1.IV: Extrait du premier paragraphe "Structures de surface" de l'article *Umbriel (Lune)*. Les phrases en rouge n'ont pas de correspondance en anglais.

CHAPITRE 2

ÉTAT DE L'ART : LA RECHERCHE SUR LES CORPUS COMPARABLES

La communauté de recherche en traitement automatique des langues naturelles a commencé à s'intéresser aux corpus comparables dans les années 1990. Plusieurs axes de recherche exploitant les potentialités de ces corpus ont été explorés. Les premiers travaux se sont penchés notamment sur l'identification de traductions de termes en vue de la construction de dictionnaires bilingues et sur la désambiguïsation du sens des mots. Plusieurs études ont également été réalisées dans le domaine de l'alignement de phrases ou de segments sous-phrastiques ainsi que dans le domaine de la recherche d'information multilingue.

2.1 La construction de dictionnaires et la désambiguïsation

Dans les années 1990, c'est à la construction de dictionnaires ou de lexiques et aux tâches de désambiguïsation auxquelles se sont principalement attelés les chercheurs. Plusieurs de leurs travaux s'appuient sur l'hypothèse que "les traductions de deux mots cooccurrant dans une langue source cooccurrent également dans une langue cible". En d'autres termes, si deux mots dans la langue source se retrouvent souvent dans la même phrase, dans le même paragraphe ou dans la même fenêtre de n mots – c'est-à-dire dans le même contexte, on suppose que leurs traductions seront elles-aussi proches l'une de l'autre en langue cible. Ainsi, si on connait la traduction de l'un des deux termes, on peut en déduire la traduction de l'autre terme.

La représentation du contexte de l'ensemble des mots se fait généralement sous la forme d'une matrice de cooccurrences de dimension $n * n$ où n est la taille du vocabulaire ; il existe différents types de matrices de cooccurrences selon l'information qu'elles renferment :

– La matrice de cooccurrences peut stocker pour pour chaque couple de mots (i, j)

la fréquence des cooccurrences. On appelle alors chaque ligne i un vecteur de contexte ou encore un sac de mots.

– Si la matriche stocke la probabilité que deux mots (i, j) se retrouvent dans le même contexte, on parle alors de matrice stochastique.

		1	2	3	4	5	6
blue	1		•			•	
green	2	•		•			
plant	3		•				
school	4						•
sky	5	•					
teacher	6				•		

Figure 2.1: Matrice de cooccurrences selon Rapp [1995]. Les points indiquent les mots qui cooccurrent plus fréquemment que deux mots pris au hasard.

Dagan et Itai [1994] utilisent le Hansard canadien et un ensemble d'articles en anglais pour désambiguïser le sens de mots rencontrés dans des articles allemands ou hébreux. Leur méthode se fonde sur la distribution des cooccurrences dans la langue cible (ici l'anglais). Ainsi lorsqu'ils hésitent entre deux traductions données, ils vont choisir celle dont, statistiquement, le contexte est le plus proche du mot source donné.

Tanaka et Iwasaki [1996] pour leur part représentent les cooccurrences sous la forme d'une matrice stochastique pour trouver la traduction japonaise de termes anglais à partir d'articles de journaux . Grâce à un dictionnaire bilingue existant et à la matrice des traductions associée, ils peuvent comparer les contextes de deux mots, l'un en langue source et l'autre en langue cible. La ligne associée à chaque mot est également appelée vecteur de contexte ou encore sac de mots (*bag of words*) car elle ne contient que des informations reliées à la fréquence des mots dans le contexte.

Les matrices de cooccurrences sont également à la base des travaux de Fung et McKeown [1997], Rapp [1995, 1999], Kaji et Morimoto [2002] et Chiao et Zweigen-

baum [2002b]. Rapp [1995] essaie de trouver la permutation des deux matrices (celle pour la langue source et celle pour la langue cible) qui les rend les plus similaires et en déduit que la traduction du mot source à la ligne *n* se trouve à la même ligne en langue cible. Il est l'un des seuls à ne pas utiliser de dictionnaire bilingue intitial, ce qui permet pourtant de réduire considérablement les temps de calcul car certaines correspondances sont déjà connues. Les méthodes utilisant un dictionnaire bilingue initial, même de petite taille, l'enrichissent au fur et à mesure du processus par bootstrapping. Koehn et Knight [2002] construisent leur dictionnaire initial grâce aux cognates, c'est-à-dire aux les mots qui ne diffèrent que par quelques lettres.

Différentes mesures de similarité entre les vecteurs de contexte ont été testées comme le coefficient de Dice, la similarité cosinus ou bien le coefficient de Jaccard. Au cours des années 2000, les méthodes d'extraction de dictionnaires ou de lexiques bilingues se sont diversifiées ou enrichies. Ainsi, Sadat et al. [2003] utilisent des informations morphologiques pour filtrer les candidats. Gaussier et al. [2004] pour leur part se servent de l'analyse sémantique latente et Shao et Ng [2004] de la translitération des termes pour enrichir les informations fournies par les vecteurs de contexte.

Au milieu des années 2000, l'intérêt des chercheurs s'est tourné vers les expressions multi-mots, la translitération et la découverte d'entités nommées dans les corpus comparables. Ainsi Daille et Morin [2005] calculent, en plus des vecteurs de contexte, des vecteurs de similarité qui contiennent les unités lexicales dont le vecteur de contexte est similaire au vecteur de contexte de l'unité à traduire. Ils utilisent ensuite des méthodes linguistiques et statistiques pour détecter les expressions multi-mots. Klementiev et Roth [2008] étudient la répartition temporelle – en fonction de la date des documents considérés – des entités nommées pour apprendre un modèle de translitération discriminatif ; les nouvelles paires d'entités nommées trouvées servent à réentrainer le classifieur. Sproat et al. [2006] s'appuient également sur l'hypothèse que la variation temporelle de la fréquence des entités nommées en langue cible est corrélée à la variation temporelle de la fréquence des entités nommées en langue source. L'indice de Pearson leur permet

de trouver les paires d'entités nommées similaires et ils "propagent le score" en accordant une confiance plus grande aux paires se trouvant dans des documents contenant déjà des paires avérées de translitérations. Saravanan et Kumaran [2008] utilisent un séparateur à vaste marge (SVM) pour identifier les paires d'entités nommées correspondantes. Ji [2009] représente pour chaque langue les relations entre les entités nommées sous la forme d'un graphe. L'alignement automatique entre les deux graphes lui permet d'identifier de manière itérative la translitération entre les entités.

Ces quatre dernières années ont connu un regain d'intérêt pour la recherche de traductions à l'aide de corpus comparables. Les méthodes proposées tentent d'améliorer celle de Rapp [1995] basée sur les vecteurs de contexte. Yu et Tsujii [2009] et Garera et al. [2009] utilisent les relations de dépendance syntaxique plutôt que des sacs de mots pour définir leur contexte. Laroche et Langlais [2010] étudient différentes manières de calculer et de comparer les vecteurs de contexte. Des modèles hybrides sont également proposés comme celui de Lee et al. [2010] qui calcule des traits lexicaux, contextuels, temporels et grammaticaux sur les couples de noms communs pour alimenter un algorithme Espérance-Maximisation, ou celui d'Andrade et al. [2011] qui apprend les paramètres d'une combinaison linéaire de ces vecteurs de contexte pour en trouver la combinaison optimale. Morin et Prochasson [2011] proposent quant à eux une méthode inspirée des méthodes de méta-recherche en combinant les résultats calculés à partir de mots similaires (synonymes, ...) aux termes à comparer.

L'autre grand axe de recherche avec les corpus comparables tente non pas d'aligner des mots mais d'aligner des phrases ou des segments sous-phrastiques.

2.2 L'extraction de phrases ou de segments sous-phrastiques

Les outils classiques d'alignement de phrases utilisés avec des corpus parallèles, comme celui de Gale et Church [1993] qui compare la longueur des phrases en langue source et en langue cible ne peuvent pas être utilisés ici car ils s'appuient sur l'hypothèse

que l'index des phrases sources et cibles dans le texte sont proches.

Trouver des phrases parallèles dans des documents qui ne le sont pas comporte plusieurs défis :

1. Identifier des paires de documents intéressants ou restreindre l'espace de recherche
2. Trouver les phrases pertinentes
3. Sélectionner les candidats au parallélisme
4. Évaluer la qualité du matériel finalement extrait

2.2.1 Identifier des paires de documents intéressants ou restreindre l'espace de recherche

En général, les corpus comparables sont immenses. Par exemple, l'extrait de Gigaword utilisé par Abdul-Rauf et Schwenk [2009] compte 1.7 millions de lignes en arabe. Le corpus quasi-comparable TDT 3 de Fung et Cheung [2004] totalise 110 000 phrases en chinois et 290 000 phrases en anglais. Supposons que notre corpus comparable compte n lignes dans la langue source et m lignes dans la langue cible avec une phrase par ligne[1]. Essayer tous les couples de phrases possibles pour trouver les couples de phrases parallèles reviendrait donc à conduire $n * m$ expériences, le problème est donc d'ordre quadratique. Afin d'éviter d'avoir à conduire plus de 10 milliards de tests (dans le cas de petits corpus de 100 000 lignes par langue), il est nécessaire de mieux cibler les documents susceptibles de contenir des traductions.

Pour les corpus de nouvelles, la date des documents est généralement considérée comme un critère discriminant. Munteanu et Marcu [2005], Abdul-Rauf et Schwenk [2009] utilisent une fenêtre de +/- 5 jours, Tillmann [2009] de +/- 7 jours. Cette étape étant parfois insuffisante, des outils de recherche d'information interlangue (CLIR : CrossLingual Information Retrieval) sont alors utilisés. Ces outils nécessitent généralement du matériel bilingue existant comme des dictionnaires ou des textes parallèles alignés.

[1] Dans le reste du document, nous utiliserons "ligne" ou "phrase" indifféremment.

Ainsi Utiyama et Isahara [2003] adoptent une méthode classique. Ils traduisent en anglais tous les mots significatifs des textes japonais à l'aide de dictionnaires bilingues, sélectionnent les deux traductions les plus fréquentes pour chaque mot et ne gardent que l'article jugé le plus pertinent par le modèle probalistique BM25. Comme ce modèle ne mesure que la similarité entre des sacs de mots, Utiyama et Isahara définissent une nouvelle mesure qui tient compte de la similarité entre les phrases de même indice – on retrouve ici l'opposition entre les carrés et les diagonales de Church [2009]. En plus de trier les documents par date, Munteanu et Marcu [2005] construisent un dictionnaire statistique à partir d'un corpus parallèle pour traduire vers l'arabe les termes d'indexation de leurs documents anglais. Ces traductions constituent leur requête pour sélectionner les documents arabes. Fung et Cheung [2004] utilisent eux-aussi un dictionnaire bilingue. Mais ils traduisent vers l'anglais l'ensemble des mots des documents chinois, représentent chaque document par un vecteur de mots pondérés (TF-IDF) et sélectionnent les paires de documents en fonction de la similarité cosinus entre leurs vecteurs.

Les autres méthodes n'utilisant pas de matériel bilingue qui ont été explorées se fondent généralement sur la structure des documents. Ainsi Resnik et Smith [2003] comparent la structure des pages Web pour trouver des pages anglaises et françaises parallèles. Ils recherchent par exemple une page Web "parent" qui contient deux liens, l'un "English" et l'autre "Français" ou des pages "jumelles", l'une en anglais qui contient un lien "Français" et l'autre en français qui contient un lien "English". Adafre et de Rijke [2006] et Smith et al. [2010], dans leurs travaux sur Wikipédia, considèrent comme comparables les paires d'articles liés par un lien interlangue. Cette méthode est intéressante. Cependant Adafre et de Rijke n'ont travaillé que sur un échantillon de 30 paires d'articles. Dans le cas du couple de langues français-anglais, si l'on souhaite travailler avec l'ensembe des articles comme l'ont fait Smith et al., cela reviendrait à explorer plus de 550 000 paires. Comme nous l'avons vu dans la section précédente, certaines de ces paires de documents sont clairement "incomparables". Il nous a donc paru judicieux

d'effectuer une sélection préalable basée sur le contenu des articles avant de commencer à chercher des phrases pertinentes.

2.2.2 Trouver les phrases pertinentes

Si l'on considère que chaque phrase du document source a une traduction dans le document cible, il existe un déséquilibre entre le nombre d'exemples "positifs", c'est-à-dire les couples de phrases parallèles, de l'ordre de $O(n)$ et le nombre d'exemple "négatifs", de l'ordre de $O(n^2)$. Pour l'éviter, certains auteurs ont à nouveau réduit l'espace de recherche pour se concentrer sur les phrases pertinentes.

Utiyama et Isahara [2003] par exemple calculent une similarité lexicale basée sur la fréquence des mots et sur le "degré de parallélisme des articles". Munteanu et Marcu [2005] éliminent les couples de phrases dont le rapport de longueur est inférieur à $\frac{1}{2}$. Ils sélectionnent ensuite les seuls couples de phrases dont la phrase cible candidate contient au moins la moitié des traductions des mots de la phrase source. Fung et Cheung [2004] utilisent la même méthode pour sélectionner les couples d'articles et les couples de phrases, à savoir la représentation par des vecteurs de mots. Abdul-Rauf et Schwenk [2009] eux-aussi sélectionnent les couples de phrases de la même manière qu'ils ont sélectionné les documents. Ils réutilisent la méthode de CLIR mise en oeuvre précédemment : ils traduisent les phrases de la langue source avec un SMT et soumettent les traductions obtenues à un engin de recherche d'information. Ils sélectionnent ensuite les cinq phrases ayant le meilleur score puis effectuent un dernier tri basé sur le ratio de la longueur des phrases.

Dans l'une de leurs expériences, Smith et al. [2010] contournent ce problème en utilisant une méthode de ranking. Il ne s'agit plus de classer les couples de phrases en "parallèles" ou "non parallèles" mais de classer les phrases cibles – incluant une phrase "null" – de manière à trouver celle qui a le plus de chance de correspondre à la phrase source. S'il n'y a pas de phrase correspondante, c'est "null" qui devrait arriver en tête du classement.

2.2.3 Sélectionner les candidats au parallélisme

En raison du bruit engendré par les approches sus-mentionnées, les couples de phrases obtenus doivent à nouveau être triés pour extraire les couples parallèles. Alors que certaines recherches s'appuient sur un classifieur, d'autres utilisent simplement un score de similarité calculé pour chaque couple et un seuil au-dessus duquel le couple est considéré comme parallèle.

Les mesures de similarité

Abdul-Rauf et Schwenk [2009] par exemple ont testé différentes mesures de similarité entre les phrases : la distance de Levenshtein, le Translation Edit Rate (TER) et une nouvelle mesure (TERp) prenant en compte entre autres des paraphrases, des synonymes et de la racinisation des mots. La méthode d'Utiyama et Isahara [2003] repose sur l'hypothèse qu'un couple de phrases extrait d'un couple d'articles fortement parallèles a de plus grandes chances d'être parallèle. Un nouveau score de similarité pour les phrases est donc calculé en prenant en compte le score de similarité obtenu au niveau des articles. Fung et Cheung [2004] ont développé une méthode itérative (boot-strapping) qui enrichit le dictionnaire bilingue utilisé à partir des phrases extraites, ce qui permet une resélection des documents et des phrases parallèles. Cette méthode permet non seulement de confirmer le parallélisme des premiers couples extraits mais aussi de découvrir de nouvelles phrases parallèles. Adafre et de Rijke [2006] ont testé deux mesures de similarité pour trouver les phrases parallèles. D'une part, ils traduisent en anglais les phrases néerlandaises grâce à un SMT et calculent le coefficient de Jaccard entre les deux phrases. D'autre part, ils constituent un dictionnaire bilingue grâce aux titres des pages liées par un lien interlangue. Dans le texte néerlandais, ils remplacent chaque lien interwiki – lien indiqué entre doubles crochets – par le titre anglais correspondant si le lien est présent dans le dictionnaire. Comme les liens sont effectués manuellement dans Wikipédia, il est possible que certains liens interwikis n'aient pas été marqués. C'est

pourquoi ils enrichissent le texte néerlandais en recherchant les liens manquants et en les traduisant grâce au dictionnaire des titres. Ils obtiennent ainsi une représentation des phrases néeerlandaises comme un “sac de liens”. Ils font la même chose avec le texte anglais, sans traduire. Ils calculent ensuite le coefficient de Jaccard entre ces deux ensembles.

	Français	Anglais
Liens présents dans la phrase	Ordre de l’Empire britannique Royal Society Mathématicien Royaume-Uni Informatique	Order of the British Empire Royal Society Mathematician Royaume-Uni Computer science
Liens rajoutés	23 juin 7 juin Science	June 23 June 7 Science

Tableau 2.I: Représentation en “sacs de liens” de la phrase du tableau 1.II

Les classifieurs

Munteanu et Marcu [2005] définissent une série de traits généraux (longueur des phrases, pourcentage de mots traduits) et de traits d’alignement calculés sur un modèle de traduction probabiliste IBM1 calculé à partir d’un corpus parallèle. Ils entrainent un classifieur à maximum d’entropie permettant de discriminer les couples. Ce classifieur est souvent utilisé en traitement des langues naturelles car il ne suppose pas d’indépendance statistique entre les traits utilisés. Smith et al. [2010] innovent quant à eux en proposant un classifieur implémentant l’algorithme des champs conditionnels aléatoires (*Conditional Random Field* - CRF), un algorithme qui modèlise le séquençage des textes. Les traits utilisés sont ceux de Munteanu et Marcu [2005] mais les traits d’alignement sont enrichis par les alignements d’un modèle de Markov caché (HMM). Smith utilise aussi des traits propres à Wikipédia comme la structure des liens à l’intérieur des phrases, la présence d’une image ou d’un lien associé à la phrase. Le dictionnaire bi-

lingue utilisé pour calculer ces traits est lui aussi généré par une méthode de ranking à partir de quelques articles Wikipédia alignés.

Pour les deux tâches d'extraction et de filtrage des phrases comparables, en raison de leur complexité, il nous a semblé judicieux d'employer un classifieur sensible aux variations des différents traits.

2.2.4 Évaluer la qualité du matériel finalement extrait

Dans les travaux précédents, on retrouve deux méthodes pour valider la pertinence du matériel extrait.

La méthode intrinsèque consiste à extraire un échantillon (par exemple Fung et Cheung [2004], Utiyama et Isahara [2003]) et à évaluer la qualité des couples de phrases, soit manuellement (Utiyama et Isahara [2003]), soit en calculant un score de similarité lexicale.

La seconde méthode utilisée permet non seulement de mesurer la qualité des traductions extraites mais aussi leur utilité. Elle consiste à entraîner un SMT sur un corpus parallèle enrichi des phrases extraites. On espère alors que ses performances seront meilleures que celles du SMT entraîné uniquement sur le corpus parallèle. La comparaison se fait généralement par les scores BLEU (Bilingual Evaluation Understudy[2]) calculés sur des corpus de test hors-domaine ou in-domaine (Abdul-Rauf et Schwenk [2009], Munteanu et Marcu [2005], Smith et al. [2010], Tillmann [2009]).

En somme, extraire des phrases parallèles à partir de corpus comparables commence par l'extraction de couples de documents pertinents. Cette étape est particulièrement importante pour Wikipédia étant donné le volume de texte disponible. Suivent alors une ou deux phases de sélection de phrases à partir de critères de similarité ou des classifieurs utilisant des traits lexicaux, statistiques ou d'alignement. La phase d'évaluation la plus robuste consiste ensuite à entraîner un SMT avec le matériel extrait et à estimer l'amélioration du score BLEU obtenu sur des corpus de test. Il est difficile de comparer

[2]Papineni et al. [2002]

les travaux entre eux car les corpus sont tous différents qu'il s'agisse de leur langue, de leur taille ou de leur degré de comparabilité. Néanmoins pour référence, nous pouvons citer ceux de Smith et al. [2010] sur Wikipédia.

Smith et al. réalisent leurs expériences sur deux corpus d'entraînement enrichis des couples de phrases extraits de Wikipédia pour le SMT :

- corpus d'entraînement "Medium" : Europarl[3] pour les couples de langue anglais-allemand et anglais-espagnol, JRC/Acquis[4] pour le couple anglais-bulgare ;
- corpus d'entraînement "Large" : Les données "Medium" enrichies de phrases parallèles tirées d'Internet, de données des Nations-Unies, de guides de conversation, de documentation technique, ...

Ils utilisent également deux corpus de tests différents :

- Test A : phrases issues de requêtes Bing pour les couples anglais-allemand et anglais-espagnol ; phrases tirées du corpus JRC/Acquis pour le couple anglais-bulgare ;
- Test Wiki : échantillon des phrases parallèles extraites filtrées manuellement (500 phrases).

Couple de langue	Anglais-Allemand	Anglais-Espagnol	Anglais-Bulgare
Nombre de couples de phrases extraits	1 694 595	1 914 978	146 465
Amélioration du score BLEU obtenue avec les données d'entraînement "Medium"			
Corpus de test A	+3.0	+3.3	+1.6
Corpus de test Wiki	+5.2	+6.1	+10.1
Amélioration du score BLEU obtenue avec les données d'entraînement "Large"			
Corpus de test A	+0.2	-0.1	-0.2
Corpus de test Wiki	+3.1	+2.2	+3.5

Tableau 2.II: Résumé des résultats obtenus par Smith et al. [2010]

[3] http://www.statmt.org/europarl/

[4] http://langtech.jrc.it/JRC-Acquis.html

CHAPITRE 3

DÉFINITION DU PROBLÈME ET MÉTHODOLOGIE

L'ensemble de ce travail peut se résumer en quatre grandes étapes :

– Préparation du corpus

– Tri des articles

– Tri des phrases

– Validation des résultats

3.1 Préparation du corpus

La fondation Wikimédia met à la disposition de la communauté l'ensemble de la base de données de Wikipédia sous la forme d'un fichier xml par langue[1]. Comme ce fichier regroupe l'ensemble des articles, des pages de redirection et des "méta-pages" destinées aux contributeurs Wikipédia, sa taille atteint plus de 31 gigaoctets pour l'anglais et 7 gigaoctets pour le français.[2] Afin de pouvoir travailler avec ce corpus, il faut d'une part supprimer toutes les balises XML et Wikipédia pour obtenir du texte brut et d'autre part conserver l'information pertinente contenue dans ces balises. Cette information est en effet nécessaire pour trouver les articles liés par un lien interlangue et pour récupérer les liens interwiki contenus dans le texte. Comme Smith et al. [2010], nous ne gardons que les articles français qui ont un article anglais correspondant et vice versa. Nous les segmentons en phrases (une phrase par ligne). Seuls les articles de dix phrases ou plus ont finalement été considérés, les articles de moins de dix lignes comprenant essentiellement des séquences de mots typiques de Wikipédia que l'on retrouve dans les autres articles. Pour leur part, Smith et al. gardent tous les articles. On verra par la suite qu'ils extraient plus de phrases parallèles que nous mais toutes ces phrases répétitives,

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

[2] Il existe des dumps plus volumineux encore, contenant l'ensemble de l'historique des pages.

considérées comme parallèles, vont certes faire grossir le volume du matériel extrait mais pas forcément en améliorer la qualité.

Cette étape a été réalisée en analysant les fichiers XML à l'aide de l'interface de programmation SAX (Simple API for XML)[3] pour Java. SAX permet de traiter les fichiers XML comme un flux, de manière événementielle, ce qui évite le chargement entier du fichier en mémoire.

3.2 Tri des articles

Une fois les articles nettoyés et segmentés, nous procédons à leur tri pour ne garder que les couples d'articles qui ont le plus de chance de contenir des phrases qui seront la traduction l'une de l'autre. En effet, en raison du nombre considérable de documents et de l'hétérogénéité du corpus, nous avons choisi de nous rapprocher d'un corpus quasi-parallèle. Nous espérons ainsi augmenter la précision de nos résultats finaux.

Nous avons expérimenté deux méthodes pour obtenir des articles au degré de parallélisme suffisant.

3.2.1 Utilisation d'un classifieur

Définition du problème de classification

La première méthode est celle dont nous voulons éprouver la pertinence. Elle consiste à entraîner un classifieur capable de détecter les articles qui sont en grande partie la traduction l'un de l'autre. On définit le *degré de parallélisme* d'un couple d'articles français-anglais par la formule (3.1).

$$tx_{para} = \frac{2 * n_{para}}{n_{fr} + n_{en}} \tag{3.1}$$

[3] `http://www.saxproject.org/`

où n_{para} est le nombre de couples de phrases parallèles et n_{fr} et n_{en} sont respectivement le nombre de phrases du texte français et du texte anglais.

Nous définissons deux classes :

– La classe 1 comprenant les articles dont $tx_{para} \geq \frac{2}{3}$;

– La classe 0 comprenant les articles dont $tx_{para} < \frac{2}{3}$.

Ce sont les articles de la classe 1 que nous souhaitons conserver pour la suite de nos travaux.

Corpus d'entraînement

Pour entraîner ce classifieur, nous avons annoté manuellement quatre-vingts couples d'articles de plus de dix lignes. Pour chaque phrase du texte français, nous indiquons s'il existe une phrase correspondante dans le texte anglais et son indice le cas échéant. Suivant Smith et al. [2010], nous considérons comme parallèles les phrases qui sont la traduction littérale l'une de l'autre et les phrases qui sont quasiment la traduction l'une de l'autre avec quelques mots manquants.

Sur l'ensemble des couples d'articles traités, dix-sept sont considérés comme quasi-parallèles (classe 1), c'est-à-dire 21.25 %. La liste des couples d'articles du corpus d'entraînement est disponible dans l'annexe A. Elle indique également quels sont les documents de la classe 1.

Définition des traits

Les traits utilisés pour représenter les couples d'articles doivent permettre de distinguer les deux classes. Nous souhaitons travailler sans matériel bilingue extérieur donc la seule ressource bilingue à notre disposition est le dictionnaire constitué par les titres des articles liés par un lien interlangue. Tout comme Adafre et de Rijke [2006], nous remplaçons tout d'abord chaque lien français par l'équivalent en anglais s'il se trouve dans notre dictionnaire. Chaque article peut dont être représenté par une séquence de

liens anglais. Nous calculons la distance de Levenshtein entre ces deux séquences. Ensuite, toujours suivant Adafre et de Rijke, nous enrichissons cette séquence en cherchant les liens manquants dans les deux textes. Cette fois, plutôt que de considérer la séquence des liens, nous les envisageons comme un ensemble non ordonné. Cela devrait nous permettre de mieux détecter les permutations de phrases dans le texte.

Nous utilisons donc les traits suivants :

- Nombre de phrases dans le texte français ;
- Nombre de phrases dans le texte anglais ;
- Similarité Levenshtein entre les séquences de liens ;
- Taille de l'ensemble enrichi de liens provenant du texte français ;
- Taille de l'ensemble enrichi de liens provenant du texte anglais ;
- Taille de l'intersection des deux ensembles.

L'annexe D présente un exemple du calcul de ces traits.

Nous avons également testé d'autres traits mais ils dégradaient les résultats des classifieurs :

- Nombre de caractères dans le texte français ;
- Nombre de caractères dans le texte anglais ;
- Nombre de liens dans le texte français ;
- Nombre de liens dans le texte anglais ;
- Similarité Cosinus entre les vecteurs de liens ;
- Taille de l'union des deux ensembles enrichis ;
- Similarité cosinus entre les deux ensembles enrichis.

Classifieur utilisé

Grâce au logiciel Weka et à son API Java[4], nous avons testé deux grandes familles de classifieurs :

- les arbres de décision ;

[4] `http://www.cs.waikato.ac.nz/ml/weka/`

– les réseaux de neurones.

Les arbres de décision sont des classifieurs qui partitionnent les données au fur et à mesure : à chaque noeud, les données sont séparées en deux en fonction d'une variable de coupure et d'un seuil. L'apprentissage permet de construire l'arbre, c'est-à-dire de déterminer, parmi un ensemble de variables donné, les variables de coupure de chaque noeud et leurs seuils. Il existe différents algorithmes qui varient, entre autres selon la méthode utilisée pour ce choix et selon la méthode utilisée afin d'éviter le surapprentissage. Un exemple d'arbre de décision obtenu à partir de notre corpus d'entraînement et de l'algorithme RandomTree est reproduit dans l'annexe B.

Les réseaux de neurones utilisés ici sont des perceptrons multicouches. Ils sont constitués d'une combinaison non linéaire (le réseau) de classifieurs linéaires (les neurones ou perceptrons). Ils permettent donc de construire des classifieurs pour des problèmes non séparables linéairement. Les paramètres à apprendre lors de l'entraînement sont le vecteur de poids et le biais de chaque perceptron. L'algorithme classique d'apprentissage du réseau de neurones est l'algorithme dit de rétropropagation[5] qui permet de minimiser une fonction de coût grâce à une descente de gradient. Cet algorithme est itératif. Plusieurs hyperparamètres font varier les performances du réseau :

– Le nombre de neurones et de couches cachées.
– Le taux d'apprentissage (et le momentum) : c'est le coefficient par lequel est multiplié le gradient lorsqu'on actualise les poids. Le momentum permet d'accélerer la descente de gradient lorsque la fonction de coût décroît rapidement.
– Le nombre d'itérations maximum : il permet d'éviter le surapprentissage.

Étant donné le faible volume de données pour l'apprentissage, notamment de données positives, la recherche de la valeur optimale des hyperparamètres a été réalisé par validation croisée à trois plis.

[5] On "propage" le gradient en commençant par le calculer pour la dernière couche de neurones puis successivement pour les couches précédentes jusqu'à la première

3.2.2 Méthode de référence : utilisation d'un moteur de recherche

La seconde méthode, qui nous sert de référence, est semblable à celle d'Utiyama et Isahara [2003] et consiste à considérer la tâche de trier les articles comme un problème de recherche d'information multilingue. Contrairement à l'approche précédente, cette méthode requiert un dictionnaire bilingue pour obtenir la traduction des mots des articles français en anglais. En revanche, elle ne nécessite pas d'apprentissage. En effet les documents français servent de requêtes qui sont soumises à un moteur de recherche renvoyant les documents anglais les plus proches. Pour notre travail, nous avons utilisé la librairie Java de Lucene[6] pour définir notre propre moteur d'indexation et notre moteur de recherche.

Chaque article anglais est donc indexé à l'aide de Lucene : le texte est mis en minuscules, chaque mot est réduit à sa racine (*racinisation* ou *stemming* en anglais) et les mots vides sont supprimés. Les documents sont représentés sous la forme d'un vecteur de mots pondéré par les valeurs TF-IDF qui mesurent l'importance d'un mot pour un document par rapport à un corpus de référence. Pour les articles français, nous avons tout d'abord réduit chaque mot à sa forme canonique (*lemmatisation*) grâce à l'analyseur syntaxique TreeTagger[7]. Ensuite, nous recherchons chaque mot porteur de sens (substantifs, adjectifs, noms propres, verbes et nombres) dans un dictionnaire bilingue pour obtenir leur équivalent anglais. L'ensemble des traductions trouvées constituent la requête soumise à Lucene.

Lucene analyse la requête de la même manière que le texte anglais (mise en minuscules, racinisation,...), calcule un score de similarité dérivant de la similarité cosinus entre le vecteur de mots constituant la requête et les vecteurs de mots représentant les textes indexés et renvoie les documents les plus pertinents.[8]

Dans notre travail, les documents avec un score supérieur à 0.5 et ayant un lien

[6] `http://lucene.apache.org/core/`

[7] `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/`

[8] Les formules utilisées par Lucene sont décrites sur son site à l'adresse `http://lucene.apache.org/core/3_6_0/api/core/org/apache/lucene/search/Similarity.html`

interlangue avec le document de la requête ont été conservés. Ce choix a été fait de manière à avoir environ le même nombre de documents que par la première méthode.

3.3 Tri des phrases

3.3.1 Définition du problème de classification

Une fois les documents pertinents pour la suite de notre travail sélectionnés, il s'agit de trouver quels sont les couples de phrases parallèles parmi tous les couples possibles. Cette tâche peut là-encore être considérée comme un problème de classification. Les deux classes sont :

- Classe 1 : phrases qui sont la traduction l'une de l'autre ;
- Classe 0 : les autres.

Comme nous l'avons mentionné dans la section 2.2.2, le problème est fortement déséquilibré car le nombre de couples de phrases parallèles est de l'ordre de $O(n)$ et le nombre d'exemple "négatifs", de l'ordre de $O(n^2)$, n étant le nombre de phrases de l'article en français.

3.3.2 Corpus d'entraînement

Comme corpus d'entraînement, nous reprenons les dix-sept couples d'articles quasi-parallèles du corpus d'entraînement du classifieur précédent. Ils contiennent 415 exemples de la classe 1 et 21 534 exemples de la classe 0. Les couples de phrases parallèles peuvent être téléchargés à partir du site Internet du RALI[9] dans l'onglet "Ressources".

3.3.3 Définition des traits

Nous avons calculé un certain nombre de traits sur les phrases prises individuellement, puis sur les couples de phrases. Un exemple de calcul de traits est donné dans l'annexe E.

[9] `http://rali.iro.umontreal.ca/rali/`

Traits calculés sur les phrases individuellement

- Le nombre de mots de la phrase (française ou anglaise) ;
- Le nombre de caractères de la phrase ;
- Une série de traits permettant de positionner la phrase par rapport à l'ensemble du texte :
  - Le score du classifieur précédent, interprété comme le degré de parallélisme du texte ;
  - Trois traits permettant de comparer la fréquence des mots de la phrase par rapport à celle de l'ensemble des mots du texte.

    Nous calculons la fréquence de chaque mot du texte dans l'ensemble du document. Pour chaque phrase, nous regardons quelle est la plus haute et la plus basse fréquence et nous calculons la fréquence moyenne des mots. Cela nous permet de détecter notamment la présence de mots hapax, c'est-à-dire de mots qui ne sont présents qu'une seule fois dans l'ensemble de l'article. Enright et Kondrak [2007] ont montré que la détection des hapax pouvait contribuer à l'alignement de documents.

Traits calculés sur les couples de phrases

Traits de comparaison des indices des phrases dans leur document respectif L'observation des couples de phrases alignées permet de détecter une certaine régularité dans leur positionnement respectif. La représentation graphique des indices source et cible des phrases parallèles de notre corpus d'entraînement montre que les points sont situés autour d'une droite proche de la diagonale (voir figure 3.1). Plus les indices sont élevés, plus les points ont tendance à s'éloigner de cette droite. Comme le corpus d'entraînement ne compte pas un nombre important de documents, indiquer l'indice des phrases aurait pu mener à des interprétations erronées de la part du classifieur si cet indice se retrouve en dehors de la plage de données utilisées – notamment pour les documents très longs.

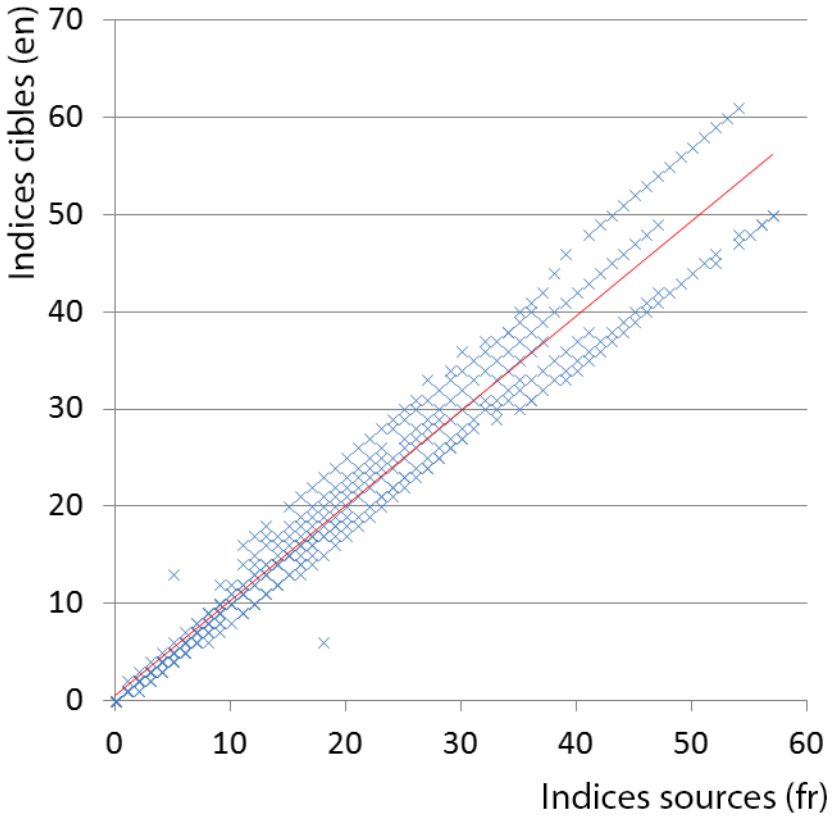

Figure 3.1: Répartition des indices des couples de phrases parallèles dans les données d'entraînement

Nous avons donc pensé calculer deux traits qui pourraient modéliser ce comportement :

– Mesure de la distance du point par rapport à la diagonale pondérée par l'indice de la phrase française

$$compIndice = \frac{|i_{fr} - i_{en}|}{i_{fr} + 1} \tag{3.2}$$

où i_{fr} et i_{fr} sont respectivement l'indice de la phrase française et l'indice de la phrase anglaise.

– Estimation de la densité de probabilité selon une mesure un peu plus fine : nous calculons l'équation de la droite moyenne par régression linéaire et nous calculons la valeur de l'écart-type sur l'ensemble des données positives de l'ensemble d'en-

traînement (c'est-à-dire les phrases parallèles). Nous représentons la distribution de nos points par une distribution gaussienne centrée sur la droite moyenne et dont la variance a été calculée précédemment. (voir figure 3.2)

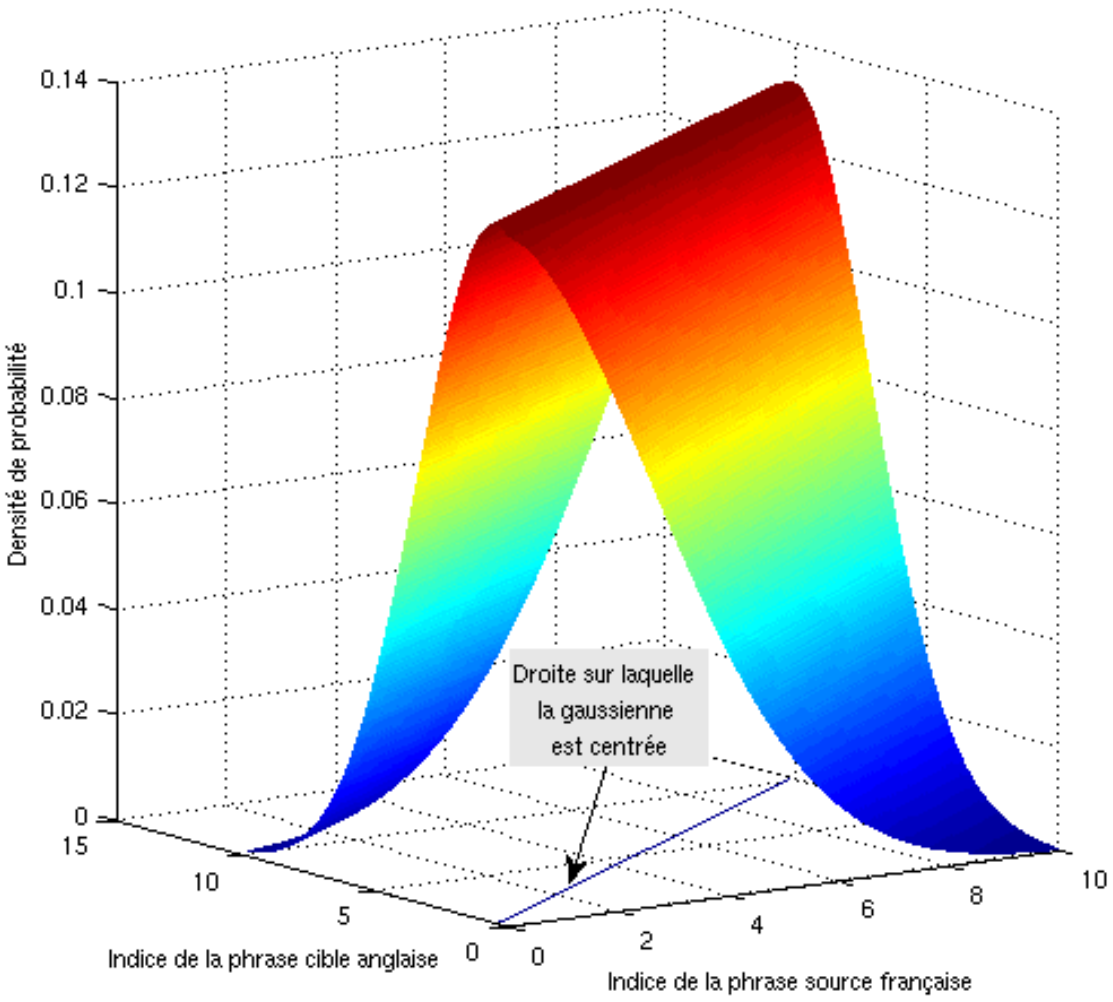

Figure 3.2: Modélisation par une gaussienne de la répartition des indices des couples de phrases parallèles

$$compIndice2 = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(i_{en}-(ai_{fr}+b))^2}{2\sigma^2}} \tag{3.3}$$

où σ est l'écart-type, a et b sont les paramètres de l'équation de la droite moyenne et i_{fr} et i_{en} sont respectivement l'indice de la phrase française et l'indice de la phrase anglaise.

Traits de comparaison des longueurs des phrases en caractères et en nombre de mots L'observation des variations dans le nombre de mots ou de caractères dans les données positives de l'entraînement montre également une répartition autour d'une droite proche de la diagonale (voir les figures 3.3 et 3.4). Nous calculons donc quatre nouveaux traits en reprenant les formules (3.2) et (3.3) et en y remplaçant les indices par le nombre de mots puis par le nombre de caractères.

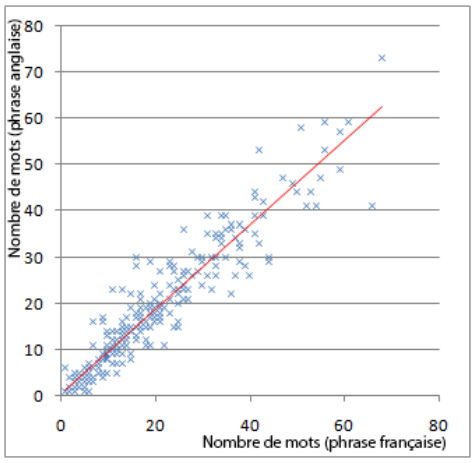

Figure 3.3: Répartition des nombres de mots pour les couples de phrases parallèles dans les données d'entraînement

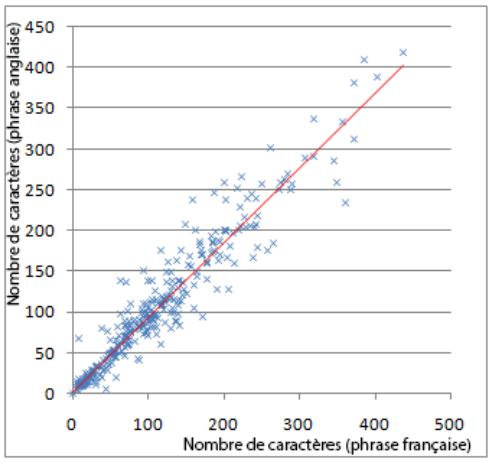

Figure 3.4: Répartition des nombres de caractères pour les couples de phrases parallèles dans les données d'entraînement

Traits de comparaison des liens contenus dans la phrase Comme pour le classifieur de documents, nous travaillons d'une part avec les liens originaux et d'autre part avec la liste enrichie de liens. Nous calculons pour ces deux ensembles :

– le nombre de liens qui se retrouvent à la fois dans la phrase française et anglaise, pondéré par le nombre de mots de la phrase source ;

- le nombre de liens qui ne se retrouvent que dans l'une des phrases, pondéré par le nombre de mots de la phrase source ;
- l'indice de Jaccard entre l'ensemble des liens de la phrase française et celui de la phrase anglaise[10].

Traits de comparaison des chiffres présents dans les phrases Suivant Patry et Langlais [2011] qui détectent des documents parallèles dans un corpus comparable en se basant sur les nombres qu'ils contiennent, nous calculons trois traits de comparaison des nombres :

- le nombre de nombres qui se retrouvent à la fois dans la phrase française et anglaise, pondéré par le nombre de mots de la phrase source ;
- le nombre de nombres qui ne se retrouvent que dans l'une des phrases, pondéré par le nombre de mots de la phrase source ;
- l'indice de Jaccard entre l'ensemble des nombres de la phrase française et celui de la phrase anglaise.

Traits de comparaison concernant le contenu non-textuel Tout comme Smith et al. [2010], nous observons si les phrases contiennent un lien vers une adresse Web, une illustration, si elles font partie d'une liste ou si elles sont incluses dans des balises titres. Chacun de ces quatre traits sont des variables ternaires pouvant prendre l'une des trois valeurs suivantes :

- 1 si les deux phrases contiennent le même lien, la même illustration, si elles font toutes les deux parties d'une liste ou d'un titre ;
- 0 si les deux phrases ne contiennent pas de lien, pas d'illustration, si elles ne font pas partie d'une liste ou d'un titre ;
- -1 si l'une des deux phrases contient un lien ou une illustration et pas l'autre (ou si les liens ou les illustrations sont différentes) ou si l'une fait partie d'un titre ou

[10] Jaccard, P. (1908). Nouvelles recherches sur la distribution florale. Bull Soc Vaudoise Sci Nat, 44 : 223–270.

d'une liste et pas l'autre.

3.3.4 Classifieur utilisé

Comme dans la section précédente, nous avons entrainé un réseau de neurones et un arbre de décision. Comme nous le verrons dans le chapitre suivant, les résultats obtenus avec le réseau de neurones sont nettement meilleurs que ceux obtenus avec l'arbre. Dans la suite de nos expériences, nous ne garderons donc que les scores obtenus avec le réseau de neurones. La validation des résultats s'est faite par validation croisée. Comme nous avions dix-sept articles à notre disposition, nous avons réalisé une validation croisée à dix-sept plis, l'ensemble d'entraînement étant constitué à chaque pli de seize articles, le bitexte restant servant d'ensemble de test.

3.3.5 Interprétation et raffinement des résultats

Le résultat retourné par le réseau de neurones pour chaque couple de phrases correspond à un score que l'on peut interpréter comme la probabilité que les phrases soient la traduction l'une de l'autre. En général, seuls les exemples ayant un score supérieur à 0.5 sont considérés comme positifs. Déplacer ce seuil permet de changer le rappel et la précision du classifieur, ce qui peut au final améliorer les résultats.

D'autre part, il est possible que la phrase française d'indice i soit ainsi couplée à plusieurs phrases anglaises d'indice j et k. Comme nous ne recherchons que des correspondances 1-1, il faut à nouveau trier ces résultats.

Deux méthodes ont été utilisées pour cela, la méthode des plus hauts scores et celle de l'algorithme hongrois. Nous avons ensuite enrichi les résultats en nous basant sur une heuristique d'extension.

Méthode des plus hauts scores

Cette méthode a été utilisée par Adafre et de Rijke [2006] pour sélectionner les meilleurs couples de phrases. Alors qu'Adafre et de Rijke triaient les couples en fonction d'une mesure de similarité entre les phrases, nous utilisons le score retourné par le réseau de neurones. La méthode consiste à trier de manière décroissante les couples de phrases en fonction de ce score. Nous sélectionnons le couple de plus haut score et nous éliminons tous les autres couples impliquant la même phrase française ou anglaise. Nous recommençons en prenant le second plus haut score et ainsi de suite jusqu'à ce que nous n'ayons plus d'exemples.

Algorithme hongrois

L'algorithme hongrois (ou algorithme de Kuhn-Munkres) est un algorithme d'optimisation combinatoire pour des problèmes d'affectation dans des graphes bipartites. Il permet de trouver l'assignement optimal entre deux ensembles disjoints, c'est-à-dire celui qui maximise (ou minimise, selon le problème) la somme des poids affectés aux arêtes qui relient les éléments des deux ensembles. Chaque élément ne peut être sélectionné qu'une seule fois. Nous considérons donc ici le problème comme un problème d'affectation entre les phrases françaises et les phrases anglaises de manière à maximiser la somme des scores retournés par le classifieur.

Supposons que le classifieur nous renvoie les résultats suivants :

$$Sc(fr_1, en_1) = 0.6$$
$$Sc(fr_1, en_2) = 0.7$$
$$Sc(fr_2, en_1) = 0.0$$
$$Sc(fr_2, en_2) = 0.6$$

Notre premier algorithme sélectionnerait le couple (fr_1, en_2) seulement[11]. L'algo-

[11] Le couple (fr_2, en_1) n'est pas retourné car son score est nul.

rithme hongrois quant à lui sélectionne les couples (fr_1, en_1) et (fr_2, en_2).

Une description du fonctionnement de l'algorithme est donnée dans l'annexe C.

Heuristique d'extension

Nous voulons tester les effets de l'heuristique suivante qui traduit le caractère séquentiel de la traduction d'un texte :

Si la phrase cible d'indice *j* est considérée comme la traduction de la phrase source d'indice *i* et que celle d'indice *j+2* est considérée comme la traduction de la phrase d'indice *i+2* ; si d'autre part la phrase source d'indice *i+1* et la phrase cible d'indice *j+1* ne sont pas impliquées dans une autre relation de traduction ; si enfin le score associé au couple *(i+1,j+1* n'est pas nul, nous considérons la phrase cible *j+1* comme la traduction de la phrase source *i+1*.

3.4 Mesure de la qualité du corpus extrait : protocole expérimental

Pour mesurer la qualité du bitexte extrait, nous entraînons et testons un SMT sur différents corpus et nous comparons les scores BLEU obtenus (Bilingual Evaluation Understudy, Papineni et al. [2002]). Nous utilisons la plateforme Moses[12], un système de traduction statistique à base de segments. Moses permet de préparer les données d'entraînement, de construire les modèles de traduction et de langue, de les paramètrer et de les tester sur des corpus de test.

Les étapes à réaliser sont :

Entrainement d'un "truecaser" Comme il est préférable de travailler sur des corpus où tous les mots – sauf les noms propres – sont en minuscules, il faut entraîner un outil sur le corpus d'entraînement pour supprimer les majuscules dans le texte source et remettre les majuscules dans la traduction obtenue.

[12] `http://www.statmt.org/moses/`

Préparation des corpus Il s'agit de tokenizer les corpus d'entraînement, de les nettoyer et de les mettre en minuscules.

Construction des modèles de langues Cette étape consiste à répertorier l'ensemble des n-grammes[13] et leur probabilité dans le corpus d'entraînement en langue cible. Nous nous limitons ici aux n-grammes d'ordre inférieur ou égal à 5, une pratique courante.

Construction des modèles de traduction Moses utilise Giza++[14] pour obtenir les modèles de traduction IBM4 à partir des bitextes et pour construire les tables de traduction.

Paramétrage des modèles La probabilité d'une certaine traduction est calculée comme le produit pondéré des probabilités issues du modèle de langue, du modèle de traduction, du modèle de distorsion et de la pénalité de longueurs. Le modèle de distorsion pénalise le déplacement de mots dans la traduction et les pénalités de longueur les différences entre les tailles des phrases source et cible. Pour optimiser les poids associés à chacun des modèles, nous nous servons d'un corpus de développement.

Décodage Un texte en langue source est soumis au décodeur qui calcule la traduction la plus probable à partir des modèles et des poids calculés précédemment.

3.4.1 Corpus d'entraînement

Nous entraînons un SMT sur plusieurs corpus d'entraînement que nous combinons également entre eux pour mesurer les gains apportés par chacun d'entre eux. Nos corpus de base sont les suivants :

Europarl+Newscommentary Corpus constitué d'extraits des comptes-rendus du Parlement européen[15] et de commentaires de nouvelles. Il est tiré des données fournies

[13] Sous-séquences de n mots
[14] `http://www.statmt.org/moses/giza/GIZA++.html`
[15] `http://www.statmt.org/europarl/`

pour le sixième atelier de traduction automatique statistique 2011 (*WMT 2011 : sixth Workshop on Stratistical Machine Translation*[16]) de la *Conference on Empirical Methods on Natural Language Processing*. Nous nommerons ce corpus Europarl dans la suite.

Wiki Corpus constitué par les phrases extraites par le classifieur précédent. Le tableau 3.I fournit quelques statistiques sur les phrases que nous avons extraites à l'aide du meilleur classifieur. Les résultats des différentes expériences menées sont détaillés dans le chapitre suivant.

Titres Corpus constitué par les titres des articles liés par un lien interlangue.

WikiLucene Corpus constitué par les couples de phrases extraits des articles sélectionnés à l'aide de Lucene.

Le tableau 3.I indique le nombre de phrases, le nombre de mots et le nombre de mots distincts que contient chaque corpus ainsi que les différentes combinaisons de corpus. La différence entre le nombre de mots distincts du corpus Europarl et celui des corpus combinés nous permet de remarquer l'apport de chaque corpus au niveau du vocabulaire.

	Europarl		Wiki		Titres		WikiLucene	
	Français	Anglais	Français	Anglais	Français	Anglais	Français	Anglais
Nombre de phrases	1 940 639		560 694		580 288		1 453 565	
Nombre de mots	60 255 293	54 055 841	11 164 680	10 705 099	1 524 825	1 469 122	29 430 579	27 342 121
Nombre de mots distincts	140 320	110 401	398 779	382 248	346 045	336 655	665 324	611 868

	Europarl + Wiki		Europarl +Titres		Europarl + Wiki + Titres		Europarl + WikiLucene + Titres	
	Français	Anglais	Français	Anglais	Français	Anglais	Français	Anglais
Nombre de phrases	2 501 333		2 520 927		3 081 621		3 954 898	
Nombre de mots	71 419 973	64 760 940	61 780 118	55 524 963	72 944 798	66 230 062	91 210 697	82 867 084
Nombre de mots distincts	463 164	426 424	445 636	408 470	664 759	622 375	876 678	808 175

Tableau 3.I: Statistiques sur les corpus d'entraînement

[16] `http://www.statmt.org/wmt11/translation-task.html`

3.4.2 Corpus de développement

Nous avons constitué un corpus composé d'une part d'un corpus de nouvelles (issu de WMT 2011) et d'autre part d'articles Wikipédia que nous avons nettoyés et alignés manuellement. Le corpus de développement Wikipédia compte 813 lignes, celui des nouvelles 2 489.

3.4.3 Corpus de test

Nous testons nos différentes configurations sur trois corpus de test. L'un est un corpus de nouvelles, celui utilisé dans WMT 2011. Il contient 3 003 lignes.

Le deuxième est composé de 2 000 paires de phrases tirées des comptes-rendus du Parlement Européen. Il provient de l'atelier WMT 2008[17].

Le dernier corpus est composé d'articles Wikipédia nettoyés et alignés manuellement. Comme pour le corpus de développement, nous nous sommes assurés que les articles choisis ne faisaient pas partie des articles sélectionnés par le premier classifieur ou Lucene[18]. Le corpus de test Wikipédia compte 800 lignes. Les corpus de test et de développement issus de Wikipédia sont mis à la disposition de la communauté sur le site du RALI[19].

Nous appelons les trois corpus de test respectivement NewsTest, WikiTest et EuroTest. Ils vont permettre d'évaluer les performances des différents corpus d'entraînement.

Le tableau 3.II indique le taux de mots inconnus de chaque corpus de test par rapport au vocabulaire contenu dans le corpus d'entraînement. On remarque les termes du corpus EuroTest sont largement couverts par le corpus d'entraînement Europarl. Au contraire, il y a de nombreux mots inconnus dans WikiTest par rapport à Europarl. Rajouter le corpus Wiki ou WikiLucene à l'entraînement permet de faire fortement baisser ce taux :

[17] `http://www.statmt.org/wmt08/shared-task.html`

[18] Pour trouver des articles adéquats, nous avons entre autres consulté la page du *Projet :Traduction* de Wikipédia qui recense par mois les traductions en cours, en relecture ou terminées. Nous avons choisi des articles traduits après le dump de la base de données.

[19] `http://rali.iro.umontreal.ca/rali/?q=fr/node/1293`

on passe de plus de 17 % à 5 et 6 %. Le corpus EuroTest est donc dit *in-domain* car il est proche du corpus d'entraînement. Le corpus WikiTest, tout comme le corpus NewsTest, sont eux *out-domain*.

	NewsTest		WikiTest		EuroTest	
Nombre de mots distincts	12 172		4 309		7 561	
Wiki	11.0 %	(8.9)	8.9 %	(384)	6.5 %	(492)
Europarl	11.0 %	(1 341)	17.3 %	(744)	0.8 %	(60)
Europarl + Wiki	5.8 %	(706)	7.4 %	(322)	0.5 %	(41)
Europarl +Titres	6.5 %	(798)	10.0 %	(431)	0.6 %	(44)
Europarl + Wiki +Titres	4.9 %	(599)	6.6 %	(284)	0.5 %	(36)
Europarl + WikiLucene + Titres	4.3 %	(524)	5.3 %	(231)	0.4 %	(32)

Tableau 3.II: Taux de mots inconnus dans les corpus de test en fonction des corpus d'entraînement. (Le nombre de mots inconnus est indiqué entre parenthèses.)

	NewsTest		WikiTest		EuroTest	
Nombre de bigrammes distincts	49 031		12 271		32 988	
Wiki	33.9 %	(16 637)	31.2 %	(3 824)	31.3 %	(10 325)
Europarl	24.3 %	(11 929)	37.6 %	(4 614)	5.9 %	(1 945)
Europarl + Wiki	19.9 %	(9 737)	25.5 %	(3 130)	5.5 %	(1 804)
Europarl +Titres	23.4 %	(11 474)	35.1 %	(4 312)	5.9 %	(1 932)
Europarl + Wiki +Titres	19.6 %	(9 586)	25.0 %	(3 068)	5.5 %	(1 800)
Europarl + WikiLucene + Titres	17.1 %	(8 381)	21.1%	(2 591)	5.1 %	(1 666)

Tableau 3.III: Taux de bigrammes inconnus dans les corpus de test en fonction des corpus d'entraînement. (Le nombre de bigrammes inconnus est indiqué entre parenthèses.)

CHAPITRE 4

RÉSULTATS ET ANALYSE

4.1 Tri des articles

4.1.1 Résultats obtenus par les classifieurs

Résultats obtenus par les réseaux de neurones

Les meilleurs résultats du réseau de neurones classifiant les paires d'articles ont été obtenus avec les hyperparamètres suivants :

- Learning rate : 0.9
- Momentum : 0.3
- Nombre maximal d'itérations : 500 époques
- Nombre de neurones cachés : 0

Comme le nombre optimal de neurones cachés est estimé à 0 (l'entrée du réseau est directement connectée à la sortie), c'est finalement un classifieur à régression logistique qui fournit les meilleurs résultats.

Le nombre d'instances correctement classées se monte à 75 soit un taux de réussite de 93.75 %. La matrice de confusion de ce classifieur est donnée dans le tableau 4.I. Elle a été calculée par validation croisée à trois plis.

Classe réelle \ Classe prédite	a	b
a=0	61	2
b=1	3	14

Tableau 4.I: Réseau de neurones : matrice de confusion

Cette matrice nous permet de calculer la précision, le rappel et la f-mesure sur la classe 1, c'est-à-dire la classe des articles quasi-parallèles.

$$\text{rappel} = \frac{\text{vrais positifs}}{\text{vrais positifs} + \text{faux négatifs}} = \frac{14}{17} = 82.35\%$$

$$\text{précision} = \frac{\text{vrais positifs}}{\text{vrais positifs} + \text{faux positifs}} = \frac{14}{16} = 87.5\%$$

$$F = \frac{2 \times (\text{précision} \times \text{rappel})}{\text{précision} + \text{rappel}} = 0.8235$$

Le temps d'entraînement du classifieur est très court puisque 0.34 seconde suffisent pour construire le modèle.

Résultats obtenus par les arbres de décision

Le meilleur arbre de décision a été obtenu par l'algorithme J48[1].

Nombre d'instances correctement classées : 71 soit un taux de réussite de 88.75 %

Classe réelle \ Classe prédite	a	b
a=0	59	4
b=1	5	12

Tableau 4.II: Arbre de décision : matrice de confusion

La matrice de confusion 4.II nous permet de calculer précision, rappel et f-mesure sur la classe 1.

$$\text{rappel} = \frac{12}{16} = 70.6\ \%$$
$$\text{précision} = \frac{12}{17} = 75\ \%$$
$$\text{F} = 0.727$$

Temps nécessaire pour construire le modèle : 0.02 seconde.

[1]Ross Quinlan (1993). C4.5 : Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo, CA.

Conclusion

Les résultats obtenus par régression logistique sont meilleurs que ceux obtenus à partir d'un arbre de décision. Cela est dû certainement au faible volume de l'ensemble d'entraînement et au déséquilibre entre les classes ce qui ne permet pas la construction d'un arbre assez précis.

La précision obtenue sur la classe 1 est de 87.5 %, c'est-à-dire que sur 100 paires de documents retournées comme parallèles, 87,5 % le sont véritablement. Un rappel de 82.35 % signifie que 82.35 % des paires de documents parallèles sont identifiées comme telles par le classifieur. On perd donc 17 % des articles parallèles.

Classification de l'ensemble des articles

Lorsqu'on applique le classifieur à régression logistique sur l'ensemble des 367 797 couples d'articles de plus de dix phrases, on obtient 38 829 quasi-parallèles soit un taux de 10.55 %. Ces résultats sont comparables à ceux de Patry et Langlais [2011] qui recensent 44 447 paires d'articles Wikipédia parallèles pour la paire de langues français-anglais. La tâche la plus coûteuse en temps est celle du calcul des traits. Elle nécessite en effet environ 9.6 secondes par couple d'articles soit environ 60 jours pour traiter l'ensemble des couples d'articles liés français-anglais, si l'on ne travaille que sur un processeur. C'est la recherche des n-grammes dans le dictionnaire constitué par les titres des articles qui est particulièrement gourmande. Une structure en arbre Radix permettrait d'améliorer fortement les temps de calcul. La classification est quant à elle relativement rapide car elle prend moins de 2 minutes.

Éliminer près de 90 % des articles revient sûrement à éliminer aussi des couples de phrases parallèles mais va nous permettre de réduire le temps de calcul subséquent et nous espérons ainsi obtenir une meilleure précision. Nous verrons par la suite qu'il est préférable de privilégier la qualité à la quantité. Il est également intéressant de constater que près de 40 000 couples d'articles sont identifiés comme étant en relation de traduc-

tion même si les versions anglaises et françaises de Wikipédia sont rédigées de manière indépendante. Ce volume de données est loin d'être inintéressant.

4.1.2 Résultats obtenus par le moteur de recherche

L'utilisation du moteur de recherche pour trier les articles est plus rapide que le réseau de neurones. Le temps nécessaire pour construire l'index des articles anglais se monte à quinze minutes. Construire les requêtes à partir des articles français nécessite environ quatre heures. C'est le temps de recherche dans l'index qui est particulièrement long car il prend environ 7 jours. Lucene trouve des documents pertinents pour toutes les requêtes soumises. Cependant ce n'est pas toujours l'article anglais correspondant à l'article français qui est en tête de la liste. Si l'on ne considère que le premier document retourné par Lucene, le score obtenu varie énormément (entre 0.0063 et 130 avec une moyenne de 0.43). Le graphique 4.1 représente la répartition des scores des premiers documents obtenus. Elle semble suivre une loi log-normale d'espérance -1.32 et de variance 0.758.

Nous ne sélectionnons que les paires documents dont le score de classification est supérieur à 0.5 et dont la longueur est suffisante (comme précédemment). Nous obtenons un ensemble de 43 564 documents. Le nombre de documents est supérieur à celui obtenu par la méthode précédente. Nous constatons également que ce sont en général des documents plus longs. Cependant les résultats des expériences suivantes nous laissent penser que ces documents contiennent beaucoup de bruit. Il n'y a que 12 490 articles en commun avec l'ensemble des articles classés comme parallèles par le réseau de neurones. Cette faible concordance entre les deux ensembles semble confirmer que l'un de deux est très bruité.

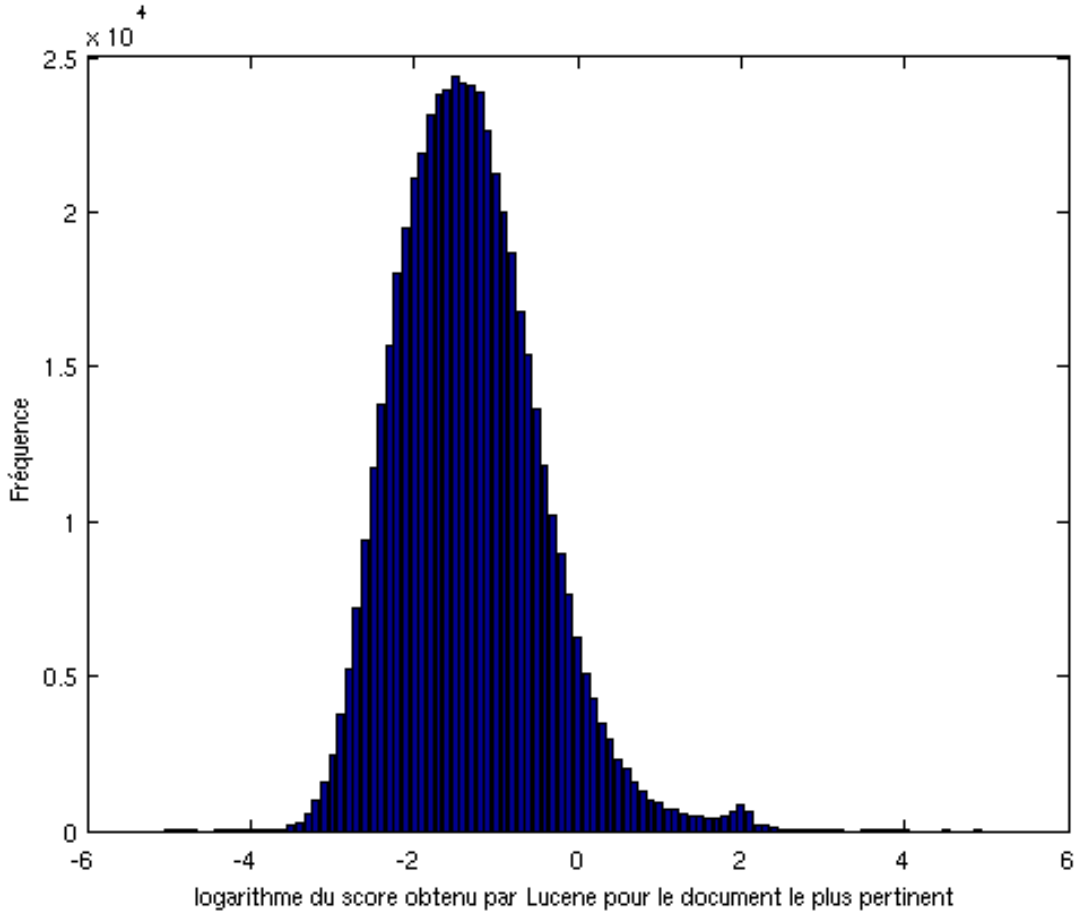

Figure 4.1: Répartition des scores des documents les plus pertinents renvoyés par Lucene (échelle logarithmique)

4.2 Tri des phrases

4.2.1 Résultats du classifieur et du raffinement

L'évaluation des performances des différentes combinaisons classifieur-raffinement ont été réalisées simultanément.

Réseau de neurones

Nous ne reproduisons ici que les meilleurs scores obtenus pour chaque classifieur et chaque méthode de raffinement.

Pour le réseau de neurones, les meilleurs hyperparamètres sont :

– Learning rate : 0.2

– Momentum : 0.3

– Nombre maximal d'itérations : 1 000 époques

– Nombre de neurones cachés : 16

La sortie brute du réseau de neurones obtient les performances suivantes :

Classe prédite / Classe réelle	a	b
a=0	21341	193
b=1	177	238

Tableau 4.III: Réseau de neurones : matrice de confusion

$$\text{rappel} = \frac{238}{238+177} = 57.3\ \%$$

$$\text{précision} = \frac{238}{238+193} = 55.2\ \%$$

F = 0.563

Ces résultats ne sont pas très bons mais on ne peut pas encore en déduire quels sont les couples de phrases parallèles car certaines phrases sont impliquées dans plusieurs couples "positifs". Les meilleurs résultats obtenus avec les différentes méthodes de raffinement ("hauts scores" ou "hongrois") sont compilés dans les tableaux ci-dessous.

La F-mesure la plus haute est obtenue lorsque qu'on fixe le seuil à partir duquel un exemple est considéré comme positif à 0.1, qu'on raffine les résultats en utilisant l'algorithme hongrois et qu'on les enrichit. Avec une précision de plus de 73 %, plus des trois quarts des couples de phrases considérés comme parallèles par le classifieur le sont vraiment. En revanche, près de deux couples de phrases sur cinq sont perdus (rappel de 62.7 %).

Classe réelle \ Classe prédite	a	b
a=0	21332	202
b=1	119	296

rappel = 71.3 %
précision = 59.4%
F = 0.648

Tableau 4.IV: Réseau de neurones, seuil 0, raffinement “hauts scores”

Classe réelle \ Classe prédite	a	b
a=0	21439	95
b=1	164	251

rappel = 60.5 %
précision = 72.5 %
F = 0.66

Tableau 4.V: Réseau de neurones, seuil 0.1, raffinement “hauts scores” + Extension

Classe réelle \ Classe prédite	a	b
a=0	21445	89
b=1	173	242

rappel = 58.3 %
précision =73.1%
F = 0.649

Tableau 4.VI: Réseau de neurones, seuil 0.1, raffinement hongrois

Classe réelle \ Classe prédite	a	b
a=0	21440	94
b=1	155	260

rappel = 62.7 %
précision = 73.4 %
F = 0.676

Tableau 4.VII: Réseau de neurones, seuil 0.1, raffinement hongrois + Extension

Lorsque l’on compare les résultats “bruts” et les résultats raffinés, on remarque que le raffinement permet d’augmenter la précision. Les deux méthodes consistent en effet à supprimer des exemples “superflus”. Avant le raffinement, les deux méthodes obtiennent des F-mesures similaires à partir de seuils différents (0 pour la méthode des plus hauts scores, 0.1 pour la méthode hongroise) Dans le premier cas, le rappel est haut mais la précision est basse. Dans le second cas, c’est l’inverse.

L’enrichissement permet de faire augmenter le rappel sans perdre de précision (+ 3 points gagnés lorsque l’algorithme hongrois est utilisé à la phase précédente). Il permet donc de retrouver des couples de phrases parallèles que le classifieur n’a pas su détecter. Ces résultats confirment notre hypothèse que les traductions se retrouvent sous forme de grappes et que la séquence des phrases est respectée.

Arbres de décision

Le meilleur arbre de décision est l'arbre J48. Comme avec le réseau de neurones, le raffinement avec l'algorithme hongrois et l'enrichissement des résultats permettent d'obtenir les performances les plus élevées. Néanmoins, les arbres de décision se révèlent être là-encore moins précis et moins couvrants. La complexité de la tâche est certainement la cause de cette différence de résultats.

Classe réelle \ Classe prédite	a	b	
a=0	21436	98	rappel = 53.9 %
b=1	191	224	

précision = 69.6 %
F = 0.608

Tableau 4.VIII: Arbre J48, seuil 0.1, raffinement hongrois + Extension

4.2.2 Conclusion

Nous utilisons le meilleur modèle obtenu pour extraire les couples de phrases parallèles des articles sélectionnés à l'étape précédente, soit par le classifieur, soit par le moteur de recherche. Nous extrayons 560 694 couples de phrases à partir du corpus d'articles issus du réseau de neurones et 1 453 565 à partir des articles sélectionnés par Lucene. La différence entre ces deux nombres est étonnante étant donné le faible écart entre le nombre d'articles sélectionnés. La section 4.3.2 fournit une explication possible à ce phénomène.

Le nombre de couples de phrases extraites par Smith et al. [2010] se monte à environ 1.7 million pour le couple de langues anglais-allemand et 1.9 million pour le couple anglais-espagnol. Ces chiffres sont du même ordre que ceux que nous obtenons avec Lucene. La méthode utilisant le réseau de neurones pour sélectionner les articles semble donc être très – peut-être trop – sélective et dans de prochains travaux, il serait intéressant de faire varier le "seuil de parallélisme" à partir duquel on garde les articles, celui que

nous avions fixé à 66 %.

L'étape suivante va nous permettre de vérifier la qualité du matériel extrait.

4.3 Mesure de la qualité du corpus extrait

4.3.1 Scores BLEU des différents SMT

Le tableau 4.IX résume les scores BLEU obtenus sur les corpus de test en entraînant Moses sur les différents corpus d'entraînement.

Le score *WER* est le "Word Error Rate" (Taux d'erreur au niveau des mots), le score *SER*, le "Sentence Error Rate" (Taux d'erreur au niveau des phrases). Plus ces scores sont bas, meilleur est le traducteur.

Le score BLEU est basé sur la précision en n-grammes entre la phrase traduite à tester et la phrase de référence. Pour la traduction du français vers l'anglais les meilleurs systèmes obtiennent des scores autour de 30.[2] Les résultats obtenus sur le corpus de test WikiTest sont donc particulièrement bas et les traductions produites sont essentiellement du bruit. On note ici l'importance du corpus d'entraînement pour obtenir un SMT de qualité. Ici le taux de mots inconnus dans le corpus WikiTest est élevé (17.3 % lorsqu'on entraîne avec Europarl) ce qui explique la faiblesse du score. L'ajout des phrases extraites de Wikipédia améliore certes la couverture lexicale du corpus d'entraînement mais on remarque que le nombre de bigrammes inconnus reste élevé (on passe de 37.6 % avec Europarl à 25.5 % avec Europarl+Wiki).

Sur le corpus de test Newstest, l'ajout des titres et des couples de phrases extraites de Wikipédia par le classifieur permet de gagner 1.55 points BLEU. Ceci ne s'explique pas seulement par la meilleure couverture lexicale du corpus d'entraînement mais aussi par la qualité du modèle de traduction que l'on a pu en extraire. En effet, le gain lexical obtenu en ajoutant les titres est presque aussi bon que celui obtenu en ajoutant les lignes Wikipédia (on passe de 11 % de mots inconnus à 6.5 % avec les titres et à 5.8 % avec

[2] http ://www.statmt.org/matrix/ et Callison-Burch et al. [2011]

Corpus de test : WikiTest

Corpus d'entraînement	WER(%)	SER(%)	BLEU(%)
Wiki	74.02	94.93	11.78
Europarl	77.69	96.83	9.44
Europarl + Titres	75.28	94.93	10.37
Europarl + Wiki	73.84	94.80	11.89
Europarl + Wiki + Titres	73.26	94.04	**11.98**
Europarl + WikiLucene + Titres	74.88	94.93	11.60

Corpus de test : NewsTest

Corpus d'entraînement	WER(%)	SER(%)	BLEU(%)
Wiki	61.90	99.60	18.48
Europarl	59.10	99.43	20.73
Europarl + Titres	58.81	99.33	21.06
Europarl + Wiki	57.87	99.30	22.16
Europarl + Wiki + Titres	57.74	99.30	**22.28**
Europarl + WikiLucene + Titres	58.29	99.17	21.70

Corpus de test : EuroTest

Corpus d'entraînement	WER(%)	SER(%)	BLEU(%)
Wiki	65.82	99.45	15.26
Europarl	53.56	97.40	28.02
Europarl + Titres	53.48	97.35	27.96
Europarl + Wiki	53.69	97.60	27.93
Europarl + Wiki + Titres	53.53	97.55	27.99
Europarl + WikiLucene + Titres	53.72	97.70	**28.32**

Tableau 4.IX: Comparaison des scores BLEU obtenus en fonction du corpus d'entraînement

le corpus Wiki). En revanche, les titres améliorent peu le taux de bigrammes inconnus contrairement au corpus Wiki (on passe de 24.3 % à 23.4 % avec les titres et à 19.9 %

avec le corpus Wiki). Le corpus WikiLucene, quant à lui, même s'il permet d'améliorer les taux de mots inconnus et de bigrammes inconnus ne permet pas des gains aussi importants sur le corpus de test Newstest. Comme nous le verrons dans la section 4.3.2, cela est probablement dû à la qualité médiocre des couples de phrases extraits.

Sur le corpus de test EuroTest, l'ajout de nouvelles traductions au corpus d'entraînement Europarl ne semble pas améliorer les résultats sauf avec l'ajout de WikiLucene où l'on observe un gain de 0.3 % du score BLEU. En revanche, même dans cette configuration, le WER et le SER sont moins bons qu'avec la configuration de base. L'amélioration du score BLEU n'est peut-être pas très significative. La relative stabilité des scores BLEU s'explique par la similarité entre le corpus de base Europarl et ce corpus de test. Comme ces deux corpus sont dans le même domaine, ajouter des données supplémentaires hors-domaine a tendance à apporter du bruit.

On remarque également qu'entraîner un SMT sur les seules données tirées de Wikipedia ne suffit pas pour obtenir des résultats satisfaisants, sûrement à cause du volume insuffisant de données. En revanche utiliser Wikipédia comme données complémentaires est utile pour la traduction de nouvelles car son contenu est souvent d'actualité et il s'inspire du style journalistique.

Nos conditions expérimentales sont similaires aux conditions "Medium" de Smith et al. [2010] dont le corpus d'entraînement de référence est constitué des données Europarl et des titres des articles Wikipédia. Leurs corpus de test sont constitués d'une part de 5 000 requêtes issues de Bing[3] (Test A) et d'autre part de 500 couples de phrases Wikipédia (WikiTest). Pour le couple espagnol-anglais, les gains BLEU obtenus sont de 3.3 points sur le corpus Test A et 6.1 points sur le corpus WikiTest. Pour le couple allemand-anglais ces gains sont de 3 points sur Test A et 5.2 sur WikiTest. Nos performances sont certes moins élevées, mais il faut souligner que contrairement à Smith et al., nous n'avons eu recours à aucune ressource bilingue extérieure pour extraire les couples de phrases parallèles.

[3]http ://www.microsofttranslator.com/

4.3.2 Analyse qualitative

Analyse des couples de phrases extraits

Afin de nous donner une idée de la qualité des traductions obtenues par notre processus, nous avons sélectionné au hasard un échantillon de 200 couples de phrases du corpus Wiki. Nous en avons également tiré 200 autres du corpus WikiLucene. Ces couples ont été manuellement classés comme parallèles, quasi-parallèles ou non-parallèles. Le tableau 4.X montre que la précision sur le corpus Wiki (70 % si on ne compte que les phrases strictement parallèles, 76 % si on inclut les phrases semi-parallèles) est sensiblement la même que celle calculée par validation croisée lors de l'entraînement (73.4 %). En revanche, sur WikiLucene, les résultats ne sont pas bons du tout : moins d'un quart des couples de phrases extraits sont effectivement parallèles ou semi-parallèles. Il semble donc que Lucene ne parvienne pas à sélectionner des articles parallèles, ce qui perturbe notre classifieur.

	Wiki		WikiLucene	
Phrases parallèles	70 %	(140)	18.5 %	(37)
Phrases semi-parallèles	6 %	(12)	4 %	(8)
Phrases non-parallèles	24 %	(48)	77.5 %	(155)

Tableau 4.X: Comparaison qualitative des couples de phrases extraits

Comparaison des traductions produites

Nous comparons les traductions du corpus Newstest obtenues par le SMT entraîné sur Europarl seulement et celles obtenues par le SMT entraîné sur le corpus Europarl+Wiki+Titres. Sur les 3 003 phrases, seules 486 sont communes aux deux traductions. Parmi elles, douze se retrouvent telles quelles dans la traduction donnée en référence.

Parmi les 2 517 phrases traduites différemment par les deux systèmes, voici quelques exemples d'amélioration constatée :

Exemple 1 : amélioration de la traduction de la la forme verbale “Ç’a été”.	
Phrase source	”**Ç’a été** une super expérience.”
Traduction Europarl	”**Has been** a super experience.”
Traduction Europarl+Wiki+Titres	”**It was** a super experience.”
Référence	”It was a super experience.”

Exemple 2 : certains termes sont mieux traduits et le temps des verbes est respecté.	
Phrase source	Il **était** frappant de constater **combien** la question de la **partie continentale** elle-même était absente.
Traduction Europarl	It **is** striking **that** the issue of the **continent** itself was absent.
Traduction Europarl+Wiki+Titres	It **was** striking **how** the issue of the **mainland** itself was absent.
Référence	It was striking how the issue of the mainland itself was absent.

Exemple 3 : traduction plus fidèle de l’expression “temps de guerre”	
Phrase source	L’amour en **temps de guerre**
Traduction Europarl	Love in **wartime**
Traduction Europarl+Wiki+Titres	Love in **times of war**
Référence	Love in times of war

Exemple 4 : amélioration de la traduction de l’expression “Earthworks”	
Phrase source	Mystérieux **travaux de terrassement**
Traduction Europarl	Mysterious **workings of terrassement**
Traduction Europarl+Wiki+Titres	Mysterious **earthworks**
Référence	Mysterious **earthworks**

Exemple 6 : meilleur respect de l'accord du déterminant	
Phrase source	"Ils ont été restitués à **leurs** propriétaires légitimes."
Traduction Europarl	"They have been returned to **its** rightful owners."
Traduction Europarl+Wiki+Titres	"They have been returned to **their** rightful owners."
Référence	"They have been returned to their rightful owners."

Exemple 8 : correction de l'oubli de mots	
Phrase source	Zapatero et la **ligne rouge** allemande
Traduction Europarl	Zapatero and the German **Red**
Traduction Europarl+Wiki+Titres	Zapatero and the German **red line**
Référence	Zapatero and the German red line

Exemple 9 : meilleure traduction des propositions relatives et des entités nommées	
Phrase source	”**Environ** 750 **plants de cannabis**, quatre **kilos de haschich**, de l’argent et des équipements **qui servaient** à la production ont été saisis”, a indiqué Louis-Philippe Ruel, porte-parole de la **Sûreté du Québec**.
Traduction Europarl	”**Around** 750 **cannabis plants**, four **kilograms of cannabis**, money and equipment **that** production were seized,” said Louis-Philippe Ruel, spokesman for the **security of Quebec**.
Traduction Europarl+Wiki+Titres	”**About** 750 **plants of cannabis**, four **kilos of hashish**, money and equipment **which served to** production were seized”, has indicated Louis-Philippe Ruel, spokesman for the **Sûreté du Québec**.
Référence	”Approximately 750 cannabis plants, 4 kilogrammes of hashish, money and equipment which served for production were seized” stated KLouis-Philippe Ruel, spokesman of the Quebec Sûreté.

Exemple 10 : correction d’un faux-sens	
Phrase source	C’est le serveur de la BBC **qui a transmis** cette information.
Traduction Europarl	This is the BBC server **which has received** this information.
Traduction Europarl+Wiki+Titres	This is the BBC server **which reported** this information.
Référence	The case was reported by BBC.

Exemple 11 : amélioration du vocabulaire	
Phrase source	Elle a écrit le **scénario** elle-même : c'est une **histoire d'amour** entre une femme originaire de la Bosnie et un homme d'origine **serbe**.
Traduction Europarl	She wrote the **scenario** itself : It is a **history of love** between a woman from Bosnia and a man of **Serb** origin.
Traduction Europarl+Wiki+Titres	She wrote the **screenplay** itself : It is a **love story** between a woman from Bosnia and a man of **Serbian** origin.
Référence	She wrote the script herself - a love story between a woman from Bosnia and a Serbian man.

CONCLUSION

À partir des téléchargements de la totalité des articles Wikipédia en anglais et en français, nous avons extrait un bitexte parallèle de plus d'un demi-million de couples de phrases. Celui-ci permet d'améliorer les performances d'un SMT car il enrichit à la fois le modèle de traduction et les modèles de langue. Les améliorations sont notables sur les corpus de test de nouvelles, le style et le contenu des articles Wikipédia se rapprochant de ceux d'articles de journaux. Notre méthode se démarque des travaux précédents sur plusieurs points.

Tout d'abord nous n'utilisons aucune ressource bilingue extérieure. Le seul dictionnaire que nous utilisons est constitué par l'ensemble des titres des articles Wikipédia marqués comme correspondants. La plupart des recherches précédentes ont recours à un bitexte aligné pour calculer un modèle de traduction duquel sont inférés des traits liés à la probabilité que la phrase cible soit la traduction de la phrase source (Munteanu et Marcu [2005], Smith et al. [2010] entre autres). Le dictionnaire des titres procure certes des informations sur le vocabulaire mais il n'est pas assez riche en phrases pour calculer un modèle de traduction. Néanmoins, nous réussissons à calculer des traits sur les articles et les phrases qui nous permettent de discriminer les phrases parallèles.

Notre seconde originalité consiste en l'utilisation d'un classifieur pour sélectionner les couples d'articles plutôt que d'avoir recours à un moteur de recherche. Grâce à des traits sur les articles soigneusement choisis, le corpus d'articles obtenus est de meilleure qualité que celui obtenu avec Lucene. Le bitexte que nous extrayons ensuite est en effet moins bruité. Lucene est donc utile pour trouver des textes au contenu similaire mais il ne parvient pas à trouver des textes dont la forme est également similaire.

Enfin, au niveau de l'extraction des phrases, nous sommes les seuls à utiliser un réseau de neurones et surtout à raffiner la sortie du classifieur grâce à un algorithme d'optimisation généralement utilisé pour les problèmes d'affectation dans les graphes bipartites. Cette méthode s'avère améliorer à la fois le rappel et la précision par rapport

à la méthode des plus hauts scores utilisée dans d'autres travaux.

Nos résultats se comparent à ceux des meilleurs travaux effectués jusqu'à présent sur l'extraction de phrases parallèle dans Wikipédia. Nous approchons les gains BLEU obtenus par Smith et al. [2010] dont les travaux font référence pour cette tâche. En n'effectuant pas de tri initial des articles, Smith et al. sont confrontés dans la sélection des couples de phrases à un plus grand volume de données. Ils y font face en envisageant le problème comme un problème de ranking et non de classification. Le nombre de couples de phrases parallèles qu'ils extraient est certes plus important mais la faible différence dans les gains BLEU laisse envisager que leur bitexte est bruité ou répétitif.

Les données alignées utilisées pour l'entraînement, le développement et les test ainsi que le bitexte extrait de Wikipédia sont mis à la disposition de la communauté sur le site Internet du RALI[4]. A partir de ce bitexte, il serait intéressant de calculer un modèle de traduction à partir duquel nous pourrions calculer de nouveaux traits sur les couples de phrases. Dans une étape de bootstrapping inspirée de Fung et Cheung [2004], nous pourrions ainsi d'une part améliorer la précision de nos classifieurs et d'autre part augmenter le volume de phrases extraites. Tout le matériel issu de Wikipédia pourra servir entre autres à améliorer des SMT. Il serait intéressant également de ne se concentrer que sur une matière donnée – la médecine ou l'astronomie par exemple – afin de n'extraire que du texte spécialisé pour enrichir des traducteurs spécialisés eux-aussi. Wikipédia par son caractère encyclopédique, multilingue et universel est une source adéquate notamment lorsqu'aucun autre matériel bilingue n'est disponible.

[4] http://rali.iro.umontreal.ca/rali/?q=fr/node/1293

ANNEXE A

LISTE DES ARTICLES UTILISÉS COMME CORPUS D'ENTRAÎNEMENT AU CLASSIFIEUR D'ARTICLES

Seuls les titres français sont indiqués.

Les couples d'articles considérés comme quasi-parallèles sont en rouge.

1196
14e cérémonie des Critics Choice Awards
1927 en football
70 Virginis
Adam-Charles-Gustave Desmazures
Akshaye Khanna
Albin Egger-Lienz
Aleksandr Shustov
Alexander Schleicher ASW 28
Alexandria (Ontario)
Alina Cojocaru
Anatomie des oiseaux
Aron Ralston
Assises nationales des États généraux du Canada français de 1967
Azulejo
Bandon
Battleford—Kindersley
Business Bart
Cecilia (chanteuse)
Championnat du monde junior de hockey sur glace 2006
Cheval de Nangchen
Chiers
Comité des droits de l'enfant
Coupe d'Afrique des nations des moins de 17 ans 2009
Creech
Darkness Dynamite
Déclaration des droits de l'homme et du citoyen de 1793
Diagramme de Ramachandran
Diepkloof
Dina Merrill
Dragon Wawel
Euscorpius
Figwit
Gaston Berger
Google Lively
Guerre Chu-Han
Guerre de Pérouse
Harry's Law
Helmholtz-Gemeinschaft
Hermine Demoriane
James Frederick Ferrier
Jan Długosz
João Rodrigues Cabrilho
Keeley Hawes
Kuwait Airways
Liste des épisodes de Digimon Tamers
Littérature française du XVIIe siècle
LOEWE
Lumines II
Maculelê
Nobuko Otowa
No Limit (album)
Olha Kharlan
Opération Outward
Pat Cox
Pax Americana
Pelni
Peterborough-Est
Peter Cheyney
Rapanui (langue)
Richard E. Grant
Robin Hood
Rodong-1
Rus' de Kiev
Saint-Georges-de-l'Oyapock
Saison NBA 2003-2004
Shaka Ponk
Stone Ocean
Sulidae
Susan Pevensie
Susumu Hirasawa
Système de gestion de contenu
Toastmasters
Torii
Tristan Palmer
Tunnel de base du Ceneri
UL Bohemian RFC
Villars-Épeney
VirusTotal
Vrms

ANNEXE B

ARBRE DE DÉCISION OBTENU LORS DE LA CLASSIFICATION DES ARTICLES

```
RandomTree
==========

Similarité_Levenshtein_entre_les_séquences_de_liens < 0.57
|   Nombre_de_phrases_anglaises < 66.5
|   |   Taille_de_l_intersection_des_deux_sets < 49
|   |   |   Similarité_Levenshtein_entre_les_séquences_de_liens < 0.38
|   |   |   |   Nombre_de_phrases_anglaises < 24.5
|   |   |   |   |   Taille_de_l_intersection_des_deux_sets < 19.5 : 0 (6/0)
|   |   |   |   |   Taille_de_l_intersection_des_deux_sets >= 19.5
|   |   |   |   |   |   Similarité_Levenshtein_entre_les_séquences_de_liens < 0.3 : 1 (2/0)
|   |   |   |   |   |   Similarité_Levenshtein_entre_les_séquences_de_liens >= 0.3 : 0 (2/0)
|   |   |   |   Nombre_de_phrases_anglaises >= 24.5 : 0 (23/0)
|   |   |   Similarité_Levenshtein_entre_les_séquences_de_liens >= 0.38
|   |   |   |   Taille_du_set_de_liens_enrichis_FR < 52
|   |   |   |   |   Taille_de_l_intersection_des_deux_sets < 20.5
|   |   |   |   |   |   Similarité_Levenshtein_entre_les_séquences_de_liens < 0.4 : 1 (1/0)
|   |   |   |   |   |   Similarité_Levenshtein_entre_les_séquences_de_liens >= 0.4
|   |   |   |   |   |   |   Nombre_de_phrases_francaises < 34 : 0 (5/0)
|   |   |   |   |   |   |   Nombre_de_phrases_francaises >= 34 : 1 (1/0)
|   |   |   |   |   Taille_de_l_intersection_des_deux_sets >= 20.5 : 1 (3/0)
|   |   |   |   Taille_du_set_de_liens_enrichis_FR >= 52 : 0 (4/0)
|   |   Taille_de_l_intersection_des_deux_sets >= 49 : 1 (2/0)
|   Nombre_de_phrases_anglaises >= 66.5 : 0 (23/0)
Similarité_Levenshtein_entre_les_séquences_de_liens >= 0.57 : 1 (8/0)
```

Exemple de lecture de l'arbre :

Si la similarité Levenshtein entre les séquences de liens est inférieure à 0.57, on considère le nombre de phrases anglaises. S'il est supérieur ou égal à 66.5 alors la classe est 0. Le 23/0 indiqué entre parenthèses à la fin de la ligne indique que si on applique cet arbre sur le corpus d'entraînement, alors la feuille en question contient 23 exemples de la classe prédite (ici 0), et 0 de l'autre.

ANNEXE C

PSEUDOCODE DE L'ALGORITHME HONGROIS

C.1 Pseudocode

ENTRÉES : Une matrice de coûts

SORTIES : Une matrice de permutation

Étape 1 : Réduction des lignes : Trouver l'élément minimum dans chaque ligne de la matrice. Construire une nouvelle matrice en soustrayant de chaque coût le minimum dans sa ligne

Étape 2 : Réduction des colonnes : Trouver l'élément minimum dans chaque colonne de la matrice. Construire une nouvelle matrice en soustrayant de chaque coût le minimum dans sa colonne

Étape 3 : Tracer le nombre minimum de traits (horizontaux ou verticaux) pour couvrir tous les zéros dans cette nouvelle matrice (appelée la matrice des coûts réduits). Si ce nombre est égal au nombre de lignes (ou colonnes), la matrice est réduite ; aller à l'étape 5. Si ce nombre est inférieur au nombre de lignes (ou colonnes), aller à l'étape 4.

Étape 4 : Trouver l'élément de valeur minimum non-couvert par un trait à l'étape 2. Soustraire cette valeur de tous les éléments non-couverts. Ajouter cette valeur aux éléments situés à l'intersection de deux traits. Retourner à l'étape 3.

Étape 5 : Déterminer la solution optimale. Générer la matrice binaire de permutation qui définit l'affectation optimale.[1]

[1] Ce pseudo-code est inspiré de Jouili, S. "Indexation De Masses De Documents Graphiques : Approches Structurelles" (2011).

C.2 Exemple de fonctionnement

Supposons que la sortie du classifieur soit la matrice suivante :

$$ScoreMat = \begin{pmatrix} 0.8 & 0.9 & 0.3 \\ 0.0 & 0.7 & 0.2 \\ 0.6 & 0.5 & 0.6 \end{pmatrix}$$

Pour retrouver un problème d'optimisation, nous remplaçons chaque élément $s_{i,j}$ de la matrice *ScoreMat* par $\max_{k,l} s_{k,l} - s_{i,j}$.

$$CoutMat = \begin{pmatrix} 0.1 & 0.0 & 0.6 \\ 0.9 & 0.2 & 0.7 \\ 0.3 & 0.4 & 0.3 \end{pmatrix}$$

Étape 1 : réduction des lignes

$$CoutMat' = \begin{pmatrix} 0.1 & 0.0 & 0.6 \\ 0.7 & 0.0 & 0.5 \\ 0.0 & 0.1 & 0.0 \end{pmatrix}$$

Étape 2 : réduction des colonnes

$$CoutMat'' = \begin{pmatrix} 0.1 & 0.0 & 0.6 \\ 0.7 & 0.0 & 0.5 \\ 0.0 & 0.1 & 0.0 \end{pmatrix}$$

Étape 3 : Traçage des traits qui couvrent les zéros

$$CoutMat'' = \begin{pmatrix} \boxed{0.1} & 0.0 & 0.6 \\ 0.7 & 0.0 & 0.5 \\ 0.0 & 0.1 & 0.0 \end{pmatrix}$$

Il y a seulement 2 traits, il faut passer à l'étape 4.

Étape 4 L'élément minimum non-couvert par un trait est 0.1 (encadré dans la matrice de l'étape précédente).

$$CoutMat''' = \begin{pmatrix} 0.0 & 0.0 & 0.5 \\ 0.6 & 0.0 & 0.4 \\ 0.0 & 0.2 & 0.0 \end{pmatrix}$$

Étape 3 - $2^{ème}$ **itération**

$$CoutMat''' = \begin{pmatrix} 0.0 & 0.0 & 0.5 \\ 0.6 & 0.0 & 0.4 \\ 0.0 & 0.2 & 0.0 \end{pmatrix}$$

Il y a trois traits, on passe à l'étape 5.

Étape 5

$$PermMat = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

ANNEXE D

CALCUL DES TRAITS SUR LA PAGE “LIBERTÉ D’ÉDUCATION”

Texte français : 10 lignes

1. La liberté d’éducation est le droit pour toute personne de créer une école et le droit pour les parents, leurs enfants ou les étudiants d’être éduqués dans l’école de leur choix ou directement par les parents (Instruction à la maison [[Instruction à la maison]]).
2. Dans certains pays l’inscription dans un système public, ou dirigé par le gouvernement, est obligatoire et les individus ne sont pas autorisés à fonder des écoles sans autorisation.
3. En principe la liberté d’éducation entraîne la suppression de tout monopole.
4. <h2>Protection légale de la liberté d’éducation </h2>
5. La liberté d’éducation a été incluse dans plusieurs constitutions (Article 2 du premier Protocole additionnel), la Constitution belge [[Constitution belge]] et la Constitution hollandaise [[Loi fondamentale du Royaume des Pays-Bas]] et dans la Convention européenne des droits de l’homme art 2 du premier protocole.
6. Aux États-Unis la liberté d’éducation n’est pas explicitement garantie par la Constitution [[Constitution des États-Unis d’Amérique]] mais elle a été réglementée comme faisant partie des ”libertés des citoyens des États-Unis”.
7. <h2>Voir aussi </h2>
8. <h3>Articles connexes</h3>
9. *Liberté d’enseignement [[Liberté d’enseignement]]
10. *Instruction à la maison [[Instruction à la maison]]

Texte anglais : 8 lignes

1. Freedom of education is a constitutional (legal) concept that has been included in the European Convention on Human Rights, Protocol 1, Article 2 [[European Convention on Human Rights Protocol 1.2C Article 2 - education]] and several national constitutions, e.g. the , the Belgian [[Belgium]] constitution (former article 17, now article 24) and the Dutch constitution [[Constitution of the Netherlands.C2.A723 : Freedom of education]](article 23).
2. This is the right for parents to have their children educated in accordance with their religious and other views.
3. Brown v. Board of Education [[Brown v. Board of Education]] was a landmark United States Supreme Court [[United States Supreme Court]] case that overturned segregation [[Racial segregation]] in US schools based on one's race.
4. In Holland, a political battle raged throughout the nineteenth century over the issue of the state monopoly on tuition-free education.
5. It was opposed under the banner of "Freedom of Education" and the Separation of Church and State [[Separation of Church and State]].
6. The Dutch called it "De Schoolstrijd" [[Schoolstrijd]] (The Battle of the Schools).
7. The Dutch solution was the Separation of School and State by funding all schools equally, both public and private.
8. <h2>References</h2>

Le tableau D.I indique (en gras) la valeur des différents traits calculés sur le couple d'articles.

	Français		Anglais	
Nombre de phrases	**10**		**8**	
Séquence de liens présents dans le texte	Liens du texte français	Équivalent dans le dictionnaire des titres	Liens du texte anglais	Liens dans le dictionnaire des titres
	Instruction à la maison	Homeschooling	European Convention on Human Rights Protocol 1.2C Article 2 - education	–
	Constitution belge	–	Belgium	Belgium
	Loi fondamentale du Royaume des Pays-Bas	Constitution of the Netherlands	Constitution of the Netherlands.C2.A723 : Freedom of education	–
	Constitution des États-Unis d'Amérique	–	Brown v. Board of Education	Brown v. Board of Education
	Liberté d'enseignement	–	United States Supreme Court	–
	Instruction à la maison	Homeschooling	Racial segregation	Racial segregation
			Separation of Church and State	Separation of Church and State
			Schoolstrijd	–
Similarité Levenshtein	**0.01449**			
Ensemble de liens présents dans le texte	Liens du texte français	Fréquence	Liens du texte anglais	Fréquence
	convention européenne des droits de l'homme – european convention on human rights	1	european convention on human rights	1
	instruction à la maison – homeschooling	2	separation of church and state	1
	de la liberté – on liberty	1	brown v. board of education	1
	liberté d'éducation – freedom of education	4	freedom of education	2
	le droit – le droit	2	united states	1
	constitution – constitution	3	constitution	2
	enseignement – teaching	1	segregation	1
	gouvernement – government	1	separation	1
	suppression – deletion	1	education	1
	protection – protection	1	monopoly	1
	états Unis – United states	2	national	1
	protocole – protocol	2	protocol	1
	éducation – education	1	article	4
	monopole – monopoly	1	century	1
	personne – person	1	concept	1
	article – article	1	funding	1
	liberté – liberty	1	holland	1
	système – system	1	private	1
	public – public	1	supreme	1
	choix – choice	1	battle	2
	école – school	2	school	1
	elle – elle	1	court	1
	pays – country	1	dutch	3
	sans – sans	1	issue	1
	être – being	1	other	1
			thout	1
			free	1
			mark	1
			over	1
			rees	1
			this	1
Taille de l'ensemble des liens	**35**		**39**	
Intersection des deux ensembles	Liens			Fréquence
	european convention on human rights			1
	freedom of education			2
	united states			1
	constitution			2
	education			1
	monopoly			1
Taille de l'ensemble des liens	**8**			

Tableau D.I: Calcul des traits du couple d'articles "Liberté d'éducation"

ANNEXE E

EXEMPLE D'UN CALCUL DE TRAITS SUR DEUX COUPLES DE PHRASES

Les couples de phrases sont tirés de l'article sur l'artiste Hans Hollein.[1]

Phrase française : <TITRE>Principales réalisations	
Phrase anglaise : <TITRE>Main works/galleries	
Indice de la phrase source	4
Indice de la phrase cible	10
Comparaison des indices 1 (voir l'équation (3.2))	1.2
Comparaison des indices 2 (voir l'équation (3.3))	0.023105973209605623
Nombre de mots dans la phrase source	2
Nombre de mots dans la phrase cible	2
Nombre de caractères dans la phrase source	24
Nombre de caractères dans la phrase cible	20
Fréquence moyenne des mots de la phrase source dans l'ensemble du document source	0.004484304932735426
Fréquence moyenne des mots de la phrase cible dans l'ensemble du document cible	0.0029069767441860465
Fréquence maximale des mots de la phrase source dans l'ensemble du document source	0.004484304932735426
Fréquence maximale des mots de la phrase cible dans l'ensemble du document cible	0.0029069767441860465
Fréquence minimale des mots de la phrase source dans l'ensemble du document source	0.004484304932735426
Fréquence minimale des mots de la phrase cible dans l'ensemble du document cible	0.0029069767441860465
Comparaison 1 de la longueur de la phrase en mots (voir l'équation (3.2))	0.6801301580459603
Comparaison 2 de la longueur de la phrase en mots (voir l'équation (3.3))	0.10891314044254659
Comparaison 1 de la longueur de la phrase en caractères (voir l'équation (3.2))	0.954159921972187
Comparaison 2 de la longueur de la phrase en caractères (voir l'équation (3.3))	0.01890609485545677
Nombre pondéré de liens similaires	0.0
Nombre pondéré de liens différents	0.0
Indice de Jaccard entre le nombre de liens	1.0
Nombre pondéré de liens enrichis similaires	0.0
Nombre pondéré de liens enrichis différents	0.0
Indice de Jaccard entre le nombre de liens enrichis	1.0
Nombre pondéré de nombres similaires	0.0
Nombre pondéré de nombres différents	0.0
Indice de Jaccard entre le nombre de nombres	1.0
Présence d'une balise "liste"	0 (car aucune des deux phrases ne contient une balise "liste")
Présence d'une balise "titre"	1 (car les deux phrases contiennent une balise "titre")
Présence d'une image	0 (car aucune des deux phrases ne contient une référence à une image)
Présence d'une adresse Web	0 (car aucune des deux phrases ne contient une référence à un site Web)
Score du classifieur précédent	0.9857334204278129

[1] en français : `http://fr.wikipedia.org/wiki/Hans_Hollein`, en anglais : `http://en.wikipedia.org/wiki/Hans_Hollein`

Phrase française : <TITRE>Principales réalisations	
Phrase anglaise : <LISTE>1964-1965 : Retti candle shop, Vienna [http ://www.bluffton.edu/ sullivanm/hollein/retti.jpg Photo]	
Indice de la phrase source	4
Indice de la phrase cible	11
Comparaison des indices 1 (voir l'équation (3.2))	1.4
Comparaison de indices 2 (voir l'équation (3.3))	0.011756521450696533
Nombre de mots dans la phrase source	2
Nombre de mots dans la phrase cible	8
Nombre de caractères dans la phrase source	24
Nombre de caractères dans la phrase cible	98
Fréquence moyenne des mots de la phrase source dans l'ensemble du document source	0.004484304932735426
Fréquence moyenne des mots de la phrase cible dans l'ensemble du document cible	0.009084302325581396
Fréquence maximale des mots de la phrase source dans l'ensemble du document source	0.004484304932735426
Fréquence maximale des mots de la phrase cible dans l'ensemble du document cible	0.02906976744186046
Fréquence minimale des mots de la phrase source dans l'ensemble du document source	0.004484304932735426
Fréquence minimale des mots de la phrase cible dans l'ensemble du document cible	0.0029069767441860465
Comparaison 1 de la longueur de la phrase en mots (voir l'équation (3.2))	0.156388825334476
Comparaison 2 de la longueur de la phrase en mots (voir l'équation (3.3))	0.028876101603742543
Comparaison 1 de la longueur de la phrase en caractères (voir l'équation (3.2))	0.1055550714831239
Comparaison 2 de la longueur de la phrase en caractères (voir l'équation (3.3))	3.34107396146759E-5
Nombre pondéré de liens similaires	0.0
Nombre pondéré de liens différents	0.0
Indice de Jaccard entre le nombre de liens	1.0
Nombre pondéré de liens enrichis similaires	0.0
Nombre pondéré de liens enrichis différents	0.5 (On a trouvé le mot "candle" dans notre dictionnaire)
Indice de Jaccard entre le nombre de liens enrichis	0.0
Nombre pondéré de nombres similaires	0.0
Nombre pondéré de nombres différents	1 (il y a deux nombres dans la phrase anglaise)
Indice de Jaccard entre le nombre de nombres	0.0
Présence d'une balise "liste"	-1 (car la phrase cible contient une balise "liste" mais pas la phrase source)
Présence d'une balise "titre"	-1 (car la phrase source contient une balise "liste" mais pas la phrase cible)
Présence d'une image	-1 (car la phrase cible contient une référence vers une image mais pas la phrase source)
Présence d'une adresse Web	0 (car aucune des deux phrases ne contient une référence à un site Web)
Score du classifieur précédent	0.9857334204278129

BIBLIOGRAPHIE

S. Abdul-Rauf et H. Schwenk. On the use of comparable corpora to improve SMT performance. Dans *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, page 16–23, 2009.

S.F. Adafre et M. de Rijke. Finding similar sentences across multiple languages in wikipedia. Dans *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, page 62–69, 2006.

D. Andrade, T. Matsuzaki et J. Tsujii. Learning the optimal use of dependency-parsing information for finding translations with comparable corpora. *ACL HLT 2011*, page 10, 2011.

Lynne Bowker et Jennifer Pearson. *Working with specialized language : a practical guide to using corpora*. Routledge, London ; New York, 2002. ISBN 0415236983 9780415236980 0415236991 9780415236997.

C. Callison-Burch, P. Koehn, C. Monz et O. F. Zaidan. Findings of the 2011 workshop on statistical machine translation. Dans *Proceedings of the Sixth Workshop on Statistical Machine Translation*, page 22–64, 2011.

Y.C. Chiao et P. Zweigenbaum. Looking for candidate translational equivalents in specialized, comparable corpora. Dans *Proceedings of the 19th international conference on Computational linguistics-Volume 2*, page 1–5, 2002a.

Y.C. Chiao et P. Zweigenbaum. Looking for candidate translational equivalents in specialized, comparable corpora. Dans *Proceedings of the 19th international conference on Computational linguistics-Volume 2*, page 1–5, 2002b.

K. Church. Repetition and language models and comparable corpora. Dans *Proceedings*

of the 2nd Workshop on Building and Using Comparable Corpora : from Parallel to Non-parallel Corpora, page 1–1, 2009.

K.W. Church. Char_align : a program for aligning parallel texts at the character level. Dans *Proceedings of the 31st annual meeting on Association for Computational Linguistics*, page 1–8, 1993.

I. Dagan et A. Itai. Word sense disambiguation using a second language monolingual corpus. *Computational Linguistics*, 20(4):563–596, 1994.

B. Daille et E. Morin. French-english terminology extraction from comparable corpora. *Natural Language Processing–IJCNLP 2005*, page 707–718, 2005.

H. Déjean et E. Gaussier. Une nouvelle approche à l'extraction de lexiques bilingues à partir de corpus comparables. *Lexicometrica, Alignement lexical dans les corpus multilingues*, page 1–22, 2002.

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

Marc Foglia. *Wikipedia, média de la connaissance démocratique ? : quand le citoyen lambda devient encyclopédiste*. FYP, Limoges (France), 2008. ISBN 9782916571065 291657106X.

P. Fung et P. Cheung. Multi-level bootstrapping for extracting parallel sentences from a quasi-comparable corpus. Dans *Proceedings of the 20th international conference on Computational Linguistics*, page 1051, 2004.

P. Fung et K. McKeown. Finding terminology translations from non-parallel corpora. Dans *Proceedings of the 5th Annual Workshop on Very Large Corpora*, page 192–202, 1997.

S. Gahbiche-Braham, H. Bonneau-Maynard et F. Yvon. Two ways to use a noisy parallel news corpus for improving statistical machine translation. *ACL HLT 2011*, page 44, 2011.

W.A. Gale et K.W. Church. A program for aligning sentences in bilingual corpora. *Computational linguistics*, 19(1):75–102, 1993.

N. Garera, C. Callison-Burch et D. Yarowsky. Improving translation lexicon induction from monolingual corpora via dependency contexts and part-of-speech equivalences. Dans *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, page 129–137, 2009.

E. Gaussier, J.M. Renders, I. Matveeva, C. Goutte et H. Dejean. A geometric view on bilingual lexicon extraction from comparable corpora. Dans *Proceedings of ACL*, volume 4, 2004.

H. Ji. Mining name translations from comparable corpora by creating bilingual information networks. Dans *Proceedings of the 2nd Workshop on Building and Using Comparable Corpora : from Parallel to Non-parallel Corpora*, page 34–37, 2009.

H. Kaji et Y. Morimoto. Unsupervised word sense disambiguation using bilingual comparable corpora. Dans *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, page 1–7, 2002.

A. Klementiev et D. Roth. Named entity transliteration and discovery in multilingual corpora. *Learning Machine Translation*, 2008.

P. Koehn et K. Knight. Learning a translation lexicon from monolingual corpora. Dans *Proceedings of the ACL-02 workshop on Unsupervised lexical acquisition-Volume 9*, page 9–16, 2002.

P. Koehn, F. J. Och et D. Marcu. Statistical phrase-based translation. Dans *Proceedings of the 2003 Conference of the North American Chapter of the Association for*

Computational Linguistics on Human Language Technology-Volume 1, page 48–54, 2003.

John Laffling. On constructing a transfer dictionary for man and machine. *Target*, 4(1):17–31, janvier 1992. ISSN 09241884. URL `http://openurl.ingenta.com/content/xref?genre=article&issn=0924-1884&volume=4&issue=1&spage=17`.

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

P. Langlais, M. Simard et J. Véronis. Methods and practical issues in evaluating alignment techniques. Dans *Proceedings of the 17th international conference on Computational linguistics-Volume 1*, page 711–717, 1998.

A. Laroche et P. Langlais. Revisiting context-based projection methods for term-translation spotting in comparable corpora. Dans *Proceedings of the 23rd International Conference on Computational Linguistics*, page 617–625, 2010.

L. Lee, A. Aw, M. Zhang et H. Li. EM-based hybrid model for bilingual terminology extraction from comparable corpora. Dans *Proceedings of the 23rd International Conference on Computational Linguistics : Posters*, page 639–646, 2010.

B. Li et E. Gaussier. Improving corpus comparability for bilingual lexicon extraction from comparable corpora. Dans *Proceedings of the 23rd International Conference on Computational Linguistics*, page 644–652, 2010.

E. Morin et E. Prochasson. Bilingual lexicon extraction from comparable corpora enhanced with parallel corpora. Dans *Proceedings of the 4th Workshop on Building and Using Comparable Corpora : Comparable Corpora and the Web*, page 27–34, 2011.

D.S. Munteanu et D. Marcu. Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, 31(4):477–504, 2005.

K. Papineni, S. Roukos, T. Ward et W. J. Zhu. BLEU : a method for automatic evaluation of machine translation. Dans *Proceedings of the 40th annual meeting on association for computational linguistics*, page 311–318, 2002.

A. Patry et P. Langlais. Identifying parallel documents from a large bilingual collection of texts : application to parallel article extraction in wikipedia. Dans *Proceedings of the 4th Workshop on Building and Using Comparable Corpora : Comparable Corpora and the Web*, page 87–95, 2011.

R. Rapp. Identifying word translations in non-parallel texts. Dans *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, page 320–322, 1995.

R. Rapp. Automatic identification of word translations from unrelated english and german corpora. Dans *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, page 519–526, 1999.

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

F. Sadat, M. Yoshikawa et S. Uemura. Learning bilingual translations from comparable corpora to cross-language information retrieval : hybrid statistics-based and linguistics-based approach. Dans *Proceedings of the sixth international workshop on Information retrieval with Asian languages-Volume 11*, page 57–64, 2003.

K. Saravanan et A. Kumaran. Some experiments in mining named entity transliteration pairs from comparable corpora. *CLIA 2008*, page 26, 2008.

L. Shao et H.T. Ng. Mining new word translations from comparable corpora. Dans

Proceedings of the 20th international conference on Computational Linguistics, page 618, 2004.

J.R. Smith, C. Quirk et K. Toutanova. Extracting parallel sentences from comparable corpora using document level alignment. Dans *Human Language Technologies : The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, page 403–411, 2010.

R. Sproat, T. Tao et C.X. Zhai. Named entity transliteration with comparable corpora. Dans *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, page 73–80, 2006.

K. Tanaka et H. Iwasaki. Extraction of lexical translations from non-aligned corpora. Dans *Proceedings of the 16th conference on Computational linguistics-Volume 2*, page 580–585, 1996.

C. Tillmann. A Beam-Search extraction algorithm for comparable data. Dans *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, page 225–228, 2009.

M. Utiyama et H. Isahara. Reliable measures for aligning Japanese-English news articles and sentences. Dans *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, page 72–79, 2003.

D. Wu et P. Fung. Inversion transduction grammar constraints for mining parallel sentences from quasi-comparable corpora. *Natural Language Processing–IJCNLP 2005*, page 257–268, 2005.

K. Yu et J. Tsujii. Extracting bilingual dictionary from comparable corpora with dependency heterogeneity. Dans *Proceedings of Human Language Technologies : The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume : Short Papers*, page 121–124, 2009.