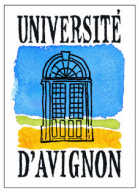


UNIVERSITÉ D'AVIGNON
ET DES PAYS DE VAUCLUSE
MINISTÈRE DE L'ENSEIGNEMENT
SUPÉRIEUR ET DE LA RECHERCHE

ACADÉMIE D'AIX-MARSEILLE
UNIVERSITÉ D'AVIGNON ET DES PAYS DE VAUCLUSE

THÈSE

présentée à l'Université d'Avignon et des Pays de Vaucluse
pour obtenir le diplôme de DOCTORAT

SPÉCIALITÉ : Informatique

École Doctorale 536 « Sciences et Agrosciences »
Laboratoire d'Informatique (EA 4128)

Traduction automatique statistique et adaptation à un domaine spécialisé

par
Raphaël Rubino

Soutenue publiquement le 30 novembre 2011 devant un jury composé de :

M.	Hervé BLANCHON	Maître de Conférences, LIG, Grenoble	Rapporteur
M.	Kamel SMAÏLI	Professeur des Universités, LORIA, Nancy	Rapporteur
M.	Laurent BESACIER	Professeur des Universités, LIG, Grenoble	Examinateur
M.	Jean SENELLART	Directeur scientifique, Systran, Paris	Examinateur
M.	Philippe LANGLAIS	Professeur agrégé, RALI, Montréal	Co-Directeur de thèse
M.	Fabrice LEFEVRE	Professeur des Universités, LIA, Avignon	Co-Directeur de thèse
M.	Georges LINARÈS	Professeur des Universités, LIA, Avignon	Directeur de thèse

---

Laboratoire Informatique d'Avignon

Remerciements

Merci à ...

mes parents, Anita et Franco, pour tout,

mes directeurs de thèse pour leurs conseils, leur confiance et leur sympathie,

Stéphane Huet, Benjamin Lecouteux et Florian Boudin pour leur aide précieuse,

tous mes collègues du LIA et du RALI.

Résumé

La traduction automatique statistique se base sur des ensembles de textes afin de construire un modèle de traduction. Traduire des textes n'ayant pas été observés au préalable s'avère difficile et mène à des performances en terme de qualité de traduction souvent peu satisfaisantes. Cet aspect est amplifié par la diversité des domaines de spécialité, ne permettant pas de disposer d'une quantité de données suffisante pour construire des systèmes de traduction efficaces.

Dans cette thèse, nous proposons d'adapter des systèmes de traduction automatique aux domaines de spécialité. Deux principaux axes sont étudiés : l'édition *a posteriori* d'hypothèses de traduction issues de systèmes automatiques et l'acquisition de lexiques bilingues spécialisés. L'originalité de nos travaux en post-édition réside dans la combinaison d'approches basées sur l'alignement sous-phrastique et ne nécessitant pas l'intervention humaine dans les processus de traduction et de post-édition. Nos approches pour l'extraction de lexiques se différencie des précédentes par la prise en considération de plusieurs niveaux de représentation des termes, que ce soit syntaxique, sémantique ou pragmatique.

Nous évaluons l'efficacité des approches proposées dans un contexte spécialisé : le domaine médical. Nous comparons nos résultats aux travaux précédents concernant cette tâche. Pour la post-édition, nous mettons en avant la pertinence de notre approche malgré le peu de données spécialisées utilisées et le faible coût de sa mise en place. Pour l'acquisition de lexiques, nous montrons l'intérêt du niveau sémantique dans la représentation des termes et la nécessité de combiner différentes vues correspondant à des aspects linguistiques particuliers.

De manière générale, nous avons amélioré la qualité de traductions issues de systèmes automatiques pour le domaine médical. Finalement, la combinaison des méthodes proposées pour chaque tâche est possible et donne lieu à des perspectives intéressantes.

Table des matières

Chapitre 1

Introduction

Le langage naturel est le mode privilégié par les humains pour communiquer entre eux, de manière parlée, signée ou écrite, ayant évolué depuis des dizaines de milliers d'années. Cette évolution a donné lieu à certains groupes de langues existantes aujourd'hui, ou ayant existé auparavant, gardant une certaine proximité provenant de leur origine commune. Toutefois, d'autres langues ont leurs origines encore incertaines, donnant lieu à de nombreuses questions et débats sur l'origine des langues. Selon Ferdinand de Saussure, "C'est une idée très fausse que de croire que le problème de l'origine du langage soit un autre problème que celui de ses transformations" (De Saussure et al., 2008).

De part son besoin de communiquer, l'homme essaye naturellement de comprendre d'autres langues, celles qu'il ne connaît pas. Aujourd'hui, les avancées technologiques et la communication globalisée accentuent ce besoin d'accéder à des données, pouvant être multimédia, dans des langues que nous ne maîtrisons pas. Une solution possible pour prendre en charge la demande grandissante en traduction est l'automatisation de cette tâche au moyen des ordinateurs. Cette idée de traduction "mécanique" (*Mechanical Translation*) remonte aux prémices de l'informatique. Les recherches scientifiques en traitement de l'information étaient alors majoritairement orientées vers la cryptographie, donnant lieu à l'analogie avec la traduction automatique, pouvant être vu comme un processus de décryptage.

L'histoire de la traduction automatique montre des périodes où les investissements financiers et humains varient, mais l'intérêt pour ce domaine depuis plus de 60 ans a engendré des progrès considérables. De nombreux projets d'envergure sont financés par l'armée, ou par des institutions comme l'Union Européenne (UE), dont le besoin en traduction est très important : 23 langues officielles parlées dans l'UE, 1,86 million de pages (comportant 1 500 caractères, sans les espaces) traduites en 2010, des domaines très divers (droit, finance, science, économie, etc.). Le budget consacré à la traduction représente moins d'1% du budget total de l'UE, ce qui correspond environ à 2 euros par habitant et par an.

La recherche académique et industrielle est actuellement très dynamique, et

de nombreux travaux sont publiés dans la littérature scientifique. Des campagnes d'évaluation internationales (WMT [1], IWSLT [2], etc.) permettent de faire état des avancées réalisées en traduction automatique et motivent des laboratoires de recherche à travers le monde. Cependant, ce procédé à tendance à figer les systèmes de traduction sur les solutions les plus performantes selon une tâche ou un critère d'évaluation particulier. Cela pose notamment le problème de l'évaluation de la qualité des traductions émises par les systèmes automatiques. Qu'est-ce qu'une bonne traduction, et comment l'évaluer ?

Certaines approches, actuellement très populaires, reposent sur l'alignement statistique d'exemples de traductions. La traduction automatique se rapproche alors de la linguistique de corpus. Ces méthodes permettent d'atteindre de bonnes performances dans le cas de vocabulaire contrôlé, ou de systèmes construit pour des domaines de spécialités. Mais il parait encore impossible aujourd'hui de traduire parfaitement des textes très longs et génériques, comme des romans par exemple. L'utilisabilité des traductions dépend donc du besoin et de l'usage fait des hypothèses générées automatiquement par les systèmes.

Il existe de nombreux outils de traduction mis à disposition, que ce soit des logiciels (développé par Systran [3] par exemple), ou des interfaces accessibles en ligne (Google Translate [4], Babel Fish [5], Systran [6], etc.). Des industriels s'intéressent donc à la traduction automatique depuis de nombreuses années et proposent des solutions efficaces selon des contextes particuliers. Systran, IBM, ou encore Xerox, sont des acteurs importants dans le développement et la commercialisation d'outils de traduction automatique. De plus, la communauté scientifique travaille depuis plusieurs années sur des boîtes à outil permettant de construire un système de traduction. Une implémentation libre et populaire est *Moses* [7], que nous utilisons dans nos travaux présentés dans cette thèse.

Ce dynamisme dans la recherche et le développement de solutions de traduction automatique passe aussi par la prise en charge de langues dites "peu dotées". Ces langues sont en général utilisées par peu de personnes et pour lesquelles le manque de ressources est un frein à la construction de systèmes de traduction automatique. La communication globalisée à travers les pays, les cultures et les langues doit pouvoir se faire sans condamner à leur perte ces langues peu dotées. La construction de ressources linguistiques et la traduction automatique sont donc des enjeux majeurs pour la diversité des langues.

Les capacités de stockage et de transfert sur les réseaux permettent aujourd'hui de travailler sur de grandes quantités de données. Il apparaît cependant que l'augmentation importante des corpus de texte, dans le cas de la traduction

1. http ://www.statmt.org/wmt11/
2. http ://iwslt2011.org/
3. http ://www.systran.fr/produits-de-traduction
4. http ://translate.google.fr
5. http ://babelfish.yahoo.com
6. http ://www.systranet.fr/translate
7. http ://www.statmt.org/moses/

automatique du langage écrit, n'a pas mené à une qualité parfaite des traductions. Améliorer la qualité des hypothèses générées automatiquement par un système de traduction demande parfois l'intervention humaine, afin d'éditer *a posteriori* les textes traduits.

Que ce soit au niveau syntaxique, sémantique ou pragmatique, des difficultés persistent donc, et font l'objet de débats sur la compréhension du concept même de traduction automatique. La couverture du vocabulaire, par exemple, est un sujet récurrent en traitement automatique du langage. La richesse terminologique d'une langue est liée aux activités de l'homme et est en perpétuelle évolution, avec un dynamisme parfois surprenant. La diversité des domaines de spécialité implique alors un effort considérable de mise à jour des systèmes de traduction automatique. C'est dans ce contexte que s'inscrit cette thèse, concernant l'adaptation aux domaines de spécialité en traduction automatique statistique.

Nos travaux sont effectués dans le cadre du projet AVISON (Archivage Vidéo et Indexation par le SON) [8], financé par l'Agence Nationale de la Recherche (ANR) au titre de l'édition 2008 du programme Contenus et Interactions (CONTINT). Il vise au développement d'une plateforme d'indexation d'une base de documents audiovisuels multilingues destinée à la formation des chirurgiens. Trois acteurs participent au projet : le Laboratoire Informatique de l'Université d'Avignon et des Pays de Vaucluse (LIA-UAPV), l'entreprise Xtensive Technologies et l'Institut de Recherche contre le Cancer de l'Appareil Digestif (IRCAD). Les objectifs du projet se situent dans la valorisation de la base documentaire de l'IRCAD et l'amélioration de l'interaction avec les utilisateurs.

Ce dernier point permet de dégager un axe de recherche portant sur l'accessibilité multilingue à la base documentaire spécialisée. Le domaine médical implique l'utilisation d'une terminologie spécifique et constitue une des difficultés dans la traduction automatique des données. Cet aspect est une des problématiques développées dans nos travaux. Nous proposons d'étudier les possibilités de traduire automatiquement des termes appartenant à un domaine de spécialité, la médecine, sans disposer de ressources constituées d'exemples de traduction dans ce domaine. Cet aspect nous amène à nous poser la question suivante : quels sont les liens existants entre des termes en relation de traduction ? Le premier thème abordé dans cette thèse porte donc sur l'acquisition automatique de vocabulaire spécialisé pour la construction de lexiques bilingues.

Une seconde problématique réside dans l'automatisation de la tâche d'édition *a posteriori* de traductions. Habituellement effectuée par des humains, la post-édition pour l'amélioration de la qualité des traductions automatiques dans un domaine de spécialité fait intervenir des spécialistes du domaine concerné. Nous proposons une approche de post-édition automatique basée sur l'analyse statistique de corpus, permettant d'améliorer des traductions issues de systèmes automatiques. Disposant de peu d'exemples de traduction dans le domaine médical, nous étudions l'impact de l'utilisation de ces données spécialisées dans la qualité des traductions générées. Le second thème abordé dans cette thèse porte sur la post-édition automatique

8. http ://avison.univ-avignon.fr

d'hypothèses de traduction.

Nos contributions portent donc sur deux aspects distincts, permettant de répondre aux deux problématiques présentées. Notre première contribution porte sur l'édition *a posteriori* de traductions issues de systèmes automatiques, permettant d'en améliorer la qualité dans le contexte d'un domaine de spécialité. Nos motivations résident dans la réduction des coûts liés à la construction de systèmes de traduction automatique statistique dans un domaine de spécialité. Nos travaux en post-édition se démarquent des travaux précédents par la mise en place de systèmes basés sur les statistiques et évitant l'intervention humaine dans le processus de traduction.

Notre seconde contribution concerne l'acquisition automatique de lexiques terminologiques bilingues à partir de ressources textuelles disponibles librement et en grande quantité. Ces ressources sont particulièrement utiles dans de nombreuses tâches en traitement automatique du langage et notamment en traduction automatique, car elles permettent d'accroître la couverture du vocabulaire. Collecter automatiquement ce type de ressources reste donc un défi pour la communauté scientifique. L'originalité de nos travaux se situe dans l'étude de niveaux de représentation des termes médicaux : syntaxique, sémantique et pragmatique. Nous proposons aussi un modèle génératif permettant de modéliser les contextes des termes en intégrant la notion de distance lexicale.

Suite à cette introduction, ce manuscrit est organisé de la façon suivante :

- Le chapitre 2 propose un rapide historique de la traduction automatique, suivi d'une présentation détaillée de l'utilisation des statistiques pour cette tâche, et termine par une description des méthodes existantes pour la traduction automatique dans les domaines de spécialité.
- Le chapitre 3 présente nos contributions en matière de post-édition automatique d'hypothèses de traductions produites par différents systèmes.
- Le chapitre 4 est centré sur les méthodes d'acquisition automatique de lexiques bilingues spécialisés.
- Enfin, le chapitre 5 concerne les conclusions et les perspectives de notre thèse.

Chapitre 2

La traduction automatique pour les domaines de spécialité

Sommaire

2.1 Historique

2.1.1 Vers l'automatisation de la traduction

La traduction consiste à porter un texte écrit dans une langue naturelle, la langue source, vers une autre langue, la langue cible. Ce processus est parfois trivial, pouvant se résumer, d'une manière simpliste, en une traduction *mots-à-mots*, comme dans l'exemple présenté par la figure 2.1.

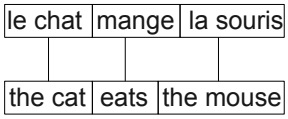


FIG. 2.1 – *Exemple de traduction d'une phrase avec une correspondance source-cible mot-à-mot.*

Il s'agit cependant d'une situation assez marginale qui sous-estime la complexité cognitive du processus de traduction. En effet, dans la plupart des cas, la traduction met en jeu des éléments syntaxiques, pragmatiques, sémantiques, etc. Par exemple, la prise en compte du contexte est nécessaire lorsque des mots à traduire sont polysémiques, comme dans la traduction présentée par la figure 2.2.

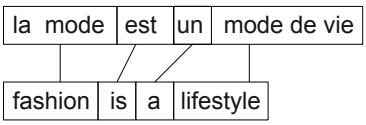


FIG. 2.2 – *Exemple de traduction d'une phrase avec une polysémie dans la langue source.*

Il apparaît aussi que le sens de la phrase source doit être compris, afin de le restituer dans la phrase cible. Garder la signification implique donc une analyse du contenu source avant le transfert vers une autre langue. Ce phénomène est particulièrement important si des expressions idiomatiques sont à traduire, comme le montre l'exemple de la figure 2.3.

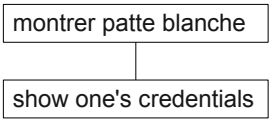


FIG. 2.3 – *Exemple de traduction d'un expression idiomatique.*

Il apparaît donc, selon les aspects présentés par les exemples précédents, que le processus de traduction de textes est complexe. Les notions de sémantique, en plus de la syntaxe, du contexte et du style rédactionnel, sont à considérer pour obtenir des traductions correctes. Un traducteur traduit généralement vers sa langue maternelle,

une parfaite connaissance de la langue source est donc indispensable afin de produire les textes cibles dans le respect des textes sources.

Par ailleurs, il est difficile de répondre à la demande croissante en traduction. Le nombre de langues différentes utilisées à travers le monde, et le développement des techniques de communication, sont des facteurs importants de l'augmentation du besoin en traduction. Rendre le contenu de la masse de données numérisées intelligible pour tous, dans un contexte international, nécessite la mise au point de méthodes de traduction automatique fiables, permettant de limiter le coût lié à l'intervention humaine.

Depuis l'apparition des ordinateurs, automatiser la traduction est un défi, dont l'origine se situe en 1949. Warren Weaver proposa, dans son célèbre *Memorandum*, de répondre au problème de la quantité de traductions à produire en utilisant des ordinateurs. Voici un extrait de sa réflexion :

> *"Thus may it be true that the way to translate from Chinese to Arabic, or from Russian to Portuguese, is not to attempt the direct route [...]. Perhaps the way is to descend, from each language, down to the common base of human communication - the real but as yet undiscovered universal language - and then re-emerge by whatever particular route is convenient."*

La communauté scientifique s'intéresse donc depuis plus de 60 ans à l'automatisation de la traduction de textes. Que ce soit pour gérer le processus de passage de la langue source à la langue cible (traduction automatique, ou *machine translation* en anglais), ou encore pour fournir une aide aux traducteurs humains (traduction assistée par ordinateur, ou *machine aided translation* en anglais). Cette dernière tâche n'est pas décrite dans cette thèse, car nous avons concentré nos travaux sur la traduction automatique.

2.1.2 Les pionniers en traduction automatique

La période de premiers travaux en traduction automatique peut être située entre 1933 et 1956. Des précurseurs tels Andrew Booth ou Warren Weaver ont alors proposé les théories de ce qu'ils appellent *mechanical translation*. Une étude complète retrace l'historique des travaux effectués par ces chercheurs entre les années 1947 et 1954, publiée par Hutchins (1997). Ainsi, dans les années 1950, à l'institut de technologie du Massachusetts (*MIT*), ou encore chez IBM, de nombreuses équipe de scientifiques se penchèrent sur l'automatisation de la traduction.

Certains abordèrent un aspect plus philosophique dans l'automatisation de cette tâche : l'indétermination de la traduction (Quine, 1959), commenté plus tard par Marchaisse (1991). Les principaux problèmes liés à la traduction automatique furent, quant à eux, décrits dans les travaux de Bar-Hillel (1960); Taube (1961). L'effervescence dans ce domaine de recherche est marqué par des travaux de précurseurs comme Y. Bar-Hillel (Bar-Hillel, 1953a,b), ou encore M. Gross (Gross, 1964).

Cette période marqua les débuts des trois approches linguistiques fondamentales à la traduction automatique. La première, appelée *traduction directe*, se repose sur l'utilisation de règles spécifiques au passage d'une langue source vers une langue cible. L'analyse du contenu à traduire et l'étude syntaxique n'était alors pas la priorité. La seconde approche est basée sur une langue intermédiaire, ou *interlangue*, constituée d'un codage neutre et abstrait, indépendant des langues, appelé langue *pivot*. Le processus de traduction est alors décomposé en deux étapes : de la langue source vers l'*interlangue*, puis de l'*interlangue* vers la langue cible. La troisième approche repose elle aussi sur une étape de transfert. Elle permet la transition d'une langue source vers la langue cible au travers d'une représentation désambiguïsée des textes sources et cibles. Trois étapes sont nécessaires à la traduction par transfert : l'analyse du contenu source, le transfert vers la langue cible, et la génération de la traduction (ou synthèse). La figure 2.4 est une représentation populaire de ces fondements sur un même schéma, appelée triangle dit de *Vauquois*, en référence aux travaux de Vauquois et Boitet (1985).

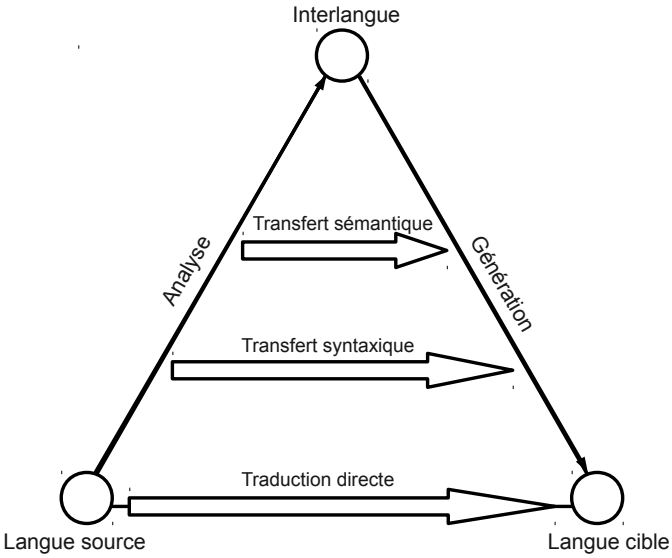


FIG. 2.4 – *Le triangle de Vauquois, illustrant les fondements de la traduction automatique.*

Après cet engouement pour la traduction automatique, apparaissent les premières désillusions, notamment lors de la publication en 1966 du rapport ALPAC (Automatic Language Processing Advisory Committee) (Pierce et Carroll, 1966). Dans ce rapport, la traduction automatique est présentée comme irréalisable dans l'immédiat, et dans un futur proche :

"there is no immediate or predictable prospect of useful machine translation"

Cette déclaration provoqua une baisse drastique des financements dédiés aux recherches en traduction automatique, et impliqua la réduction importante des expérimentations dans ce domaine, et ce pour une dizaine d'années.

Ce n'est que vers les années 1970 que de nouveaux travaux furent entrepris, notamment par des industriels comme SYSTRAN (acronyme de *System*

Translation) (Toma, 1970, 1972, 1977). Malgré les conclusions du rapport ALPAC, la rentabilité de la traduction automatique peut être effective suivant certains conditions, comme garder une intervention humaine dans le processus de traduction, ou se limiter à un domaine de spécialité pour limiter le vocabulaire à traduire.

À l'Université de Montréal, le projet TAUM (Traduction Automatique de l'Université de Montréal) est un très bon exemple de traduction automatique dans un domaine de spécialité. Ce projet donna lieu au système *Météo* (Chandioux, 1976), mis en place en 1976 afin de traduire des bulletins météorologiques entre l'anglais et le français, dont la syntaxe limitée et le vocabulaire restreint permettent d'atteindre de bons résultats.

Toujours dans un contexte de traduction pour les domaines de spécialité, l'Institut textile de France proposa *TITUS* en 1970 (Ducrot, 1973), un système multilingue permettant de traduire des textes dans un langage contrôlé. En 1972, l'Université chinoise de Hong-Kong proposa le système *CULT* (Loh, 1972), développé pour traduire des textes mathématiques du chinois vers l'anglais.

Si certains domaines sont concernés par d'importantes demandes en traductions, depuis les années 1950, jusqu'à la fin des années 1970, le contexte politique de l'époque influença lui aussi l'orientation des systèmes de traduction automatique. Notamment vers la prise en charge de deux langues en particulier : l'anglais et le russe. En 1954, l'Université de Georgetown et IBM proposèrent conjointement un système permettant de traduire une soixantaine de phrases du russe vers l'anglais, en se basant sur 6 règles de grammaire et sur un vocabulaire de 250 mots (Dostert, 1955).

SYSTRAN participe aussi au développement de la traduction automatique entre ces deux langues, dont les premiers systèmes furent présentés par Peter Toma. Portés ensuite à d'autres langues, comme la paire anglais-français en 1976, les systèmes mis en place par SYSTRAN intéressèrent alors l'Union européenne (Pigott, 1988) dans son besoin croissant de traductions vers de nombreuses langues. D'autres organismes, comme l'OTAN ou l'Agence internationale de l'énergie atomique, utilisèrent eux aussi les systèmes développés par Systran. Ces organismes, regroupant plusieurs États, nécessitent depuis lors des traductions dans diverses langues. L'utilisation de la traduction automatique apparaît comme un moyen de faciliter la communication entre les membres tout en réduisant les coûts en temps de travail humain.

De nombreux systèmes, tels que SYSTRAN, Logos (Tschira, 1985) ou METAL (Slocum et al., 1984), furent développés en premier lieu pour traduire des textes *génériques*, c'est à dire pour prendre en charge du vocabulaire n'appartenant pas à un domaine de spécialité particulier. Cependant, les dictionnaires inclus dans ces systèmes ont été adaptés pour certains domaines, afin de répondre à des demandes plus spécifiques et de couvrir le vocabulaire spécialisé. Restreindre ces systèmes à des domaines de spécialité permet d'atteindre des résultats satisfaisant.

C'est dans les années 1970 et 1980 que les systèmes dédiés à la traduction dans un domaine de spécialité connurent un essor important. Nous pouvons remarquer notamment XEROX, ou encore SMART CORPORATION, dont les systèmes se basent sur

un contrôle total du texte à traduire d'un point de vue du vocabulaire et de la syntaxe, afin de limiter au maximum les révisions des traductions émises automatiquement (ces révisions sont connues sour le nom de *post-édition*).

2.1.3 Des approches basées sur les corpus

Jusqu'à la fin des années 1980, les recherches en traduction automatique sont dominées par l'utilisation de règles linguistiques, basées sur l'analyse syntaxique, le transfert lexical, la morphologie, etc. Les systèmes à base de règles sont principalement *Ariane* (Boitet et al., 1982), *METAL*, *SUSY* (Maas, 1977), *Mu* (Nagao et al., 1985) et *Eurotra* (King, 1981). Cependant, si l'approche de la traduction automatique à base de règles a permis de développer cette multitude de systèmes, c'est l'étude empirique des corpus de textes, dont la taille augmente constamment, qui a donné lieu aux approches à la traduction automatique basée sur les corpus.

L'exemple le plus connu de traduction basée sur des corpus remonte à l'époque de Champollion, avec la pierre de Rosette (figure 2.5), contenant trois versions d'un même texte écrit en deux langues : égyptien et grec ancien. L'un des textes est écrit utilisant l'alphabet grec, un autre est écrit en égyptien démotique, et enfin le dernier est écrit en égyptien hiéroglyphique (Champollion, 1828). Ce type de corpus, composés de textes traduits dans plusieurs langues, sont aujourd'hui communément appelés *corpus parallèles*.

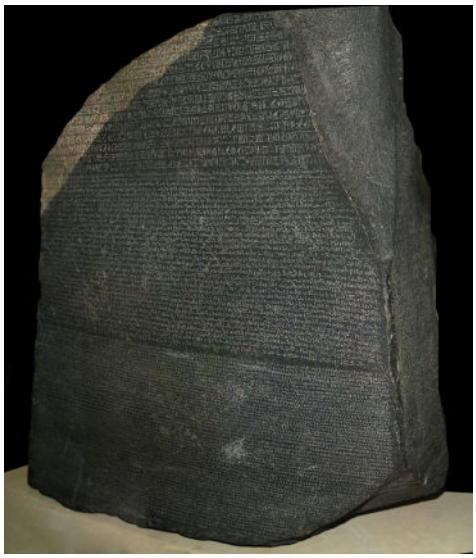


FIG. 2.5 – *La pierre de Rosette, exposée au British National Museum (source Wikipédia).*

Ce bref historique de la traduction automatique est, bien sûr, loin d'être exhaustif. De nombreuses publications couvrent, de manière détaillée, l'ensemble de l'historique pour ce domaine, en présentant les limites et les difficultés rencontrées dans les différentes approches (Hutchins, 1982, 1986; Hutchins et Somers, 1992; Hutchins, 1997, 2000, 2003, 2007). Il apparaît toutefois que depuis une vingtaine d'années, la traduction

automatique basée sur les corpus connaît un développement très important. Ceci est dû à la disponibilité grandissante des corpus parallèles, mais aussi à la puissance croissante des ordinateurs. De ce fait, de nombreuses approches computationnelles, orthogonales aux approches linguistiques, ont été développées afin de répondre au besoin en traduction automatique. Nous avons pu relever trois principales approches computationnelles à la traduction automatique basée sur les corpus :

- la traduction à base de règles (Rule-Based Machine Translation, ou *RBMT*),
- la traduction à base d'exemples (Example-Based Machine Translation, ou *EBMT*),
- la traduction statistique (Statistical Machine Translation, ou *TAS*).

Chaque paradigme se base sur la construction d'une ressource disponible en plusieurs langues et en relation de traduction. La traduction de nouveaux textes est alors possible en s'appuyant sur cette ressource. La méthode de construction et d'utilisation de cette ressource permet de différencier les approches en traduction automatique.

Dans l'approche à base de règles, des instructions permettant la traduction d'une langue source vers une langue cible sont codées et enregistrées dans un programme. Dans une approche idéale, ces instructions sont composées de dictionnaires et de grammaires, écrits dans un langage spécialisé pour la programmation linguistique. Puis, une fois compilées, ces ressources sont utilisées par un moteur de traduction, d'analyse, de transfert et de génération.

L'approche par l'exemple se base sur des textes bilingues dont chaque phrase source a une traduction cible. Grâce à ces ressources, différents alignements peuvent être produits : sur les mots, les groupes de mots, selon des structures arborescentes, etc. Les textes bilingues sont appelés corpus parallèles et permettent de construire automatiquement un ensemble d'exemples de traductions, constituant finalement la base de connaissances du système.

En traduction statistique, des corpus parallèles sont également utilisés, permettant d'estimer les probabilités qu'un texte cible est la traduction d'un texte source. Maximiser ces probabilités permet alors à un *décodeur* de sélectionner des hypothèses de traduction, et ainsi de traduire de nouveaux textes.

Parmi ces trois approches computationnelles, nous pouvons différencier celles pouvant s'affranchir d'une analyse linguistique des corpus : l'approche par l'exemple et la traduction statistique. La traduction basée sur des règles nécessite une analyse linguistique des corpus, afin d'en extraire un ensemble de règles de traductions qui constituent la base du système. Cette différenciation est très importante, car elle permet notamment la combinaison d'approches donnant lieu à des systèmes hybrides. L'avantage de combiner une approche purement computationnelle [1] avec une approche linguistique se situe dans la prise en charges de certains aspects intrinsèques à la traduction. Pour la traduction statistique par exemple, il parait difficile de considérer des contextes de grandes tailles, alors que pour la traduction par l'exemple, ce sont les structures syntaxiques complexes qui ne sont pas gérées.

1. Il faut toutefois préciser que les approches basées sur les corpus de texte reposent sur la branche de la linguistique appelée *linguistique de corpus*

Une approche courante de la traduction automatique hybride repose sur une première étape d'analyse linguistique du texte dans la langue source. Puis, une seconde étape permet de passer dans la langue cible, grâce aux approches non-linguistiques, en traduisant des segments sous-phrastiques. La sélection lexicale dans la langue cible se fait généralement à l'aide d'un modèle de langue. Un système populaire de traduction automatique hybride est Pangloss (Nirenburg et al., 1994).

2.2 Contexte de la thèse

2.2.1 L'approche statistique à la traduction automatique

C'est dans les années 1940 que les premiers théoriciens de l'information et de la communication, dont Claude Shannon et Warren Weaver, imaginèrent que la tâche de traduction de texte pourrait être effectuée par un ordinateur. Si la traduction automatique était considérée comme un problème de cryptographie, c'est d'après le paradigme du canal bruité, proposé en 1948 (Shannon et Weaver, 1948) et illustré par la figure 2.6, que l'approche statistique pour la traduction automatique fut définie.

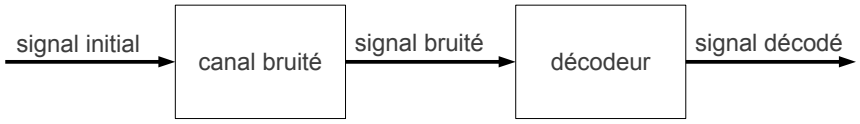


FIG. 2.6 – *Modèle du canal bruité proposé par Shannon et Weaver.*

La traduction automatique statistique (*TAS*) permet d'émettre des hypothèses de traduction dans une langue cible à partir d'une phrase dans une langue source. Cette approche est basée sur l'observation d'exemples de traductions, les corpus parallèles, permettant l'apprentissage automatique de correspondances bilingues. Les corpus parallèles gratuits les plus couramment utilisés proviennent :

- des Nations Unies (Rafalovitch et Dale, 2009) (tableau 2.1),
- du Parlement européen (Koehn, 2005) (tableau 2.2),
- du Parlement canadien (Simard, 1998; Germann, 2001) (tableau 2.3),

De manière générale, ces trois corpus concernent un domaine particulier, celui des débats parlementaires, ou des résolutions adoptées. Or, il est souvent plus difficile d'obtenir ce type de ressources pour d'autres domaines spécialisés, comme celui de la médecine par exemple.

2.2.2 Limites liées au manque de données

La disponibilité des corpus parallèles pose donc des problèmes, que ce soit pour la couverture du vocabulaire ou pour les structures syntaxiques des phrases à traduire. De plus, les textes de domaines spécialisés sont en général plus difficiles à traduire

LANGUE	# MOTS	# CARACTÈRES (M)
anglais	3 067 550	20,7
français	3 442 254	22,8
espagnol	3 581 566	22,9
russe	2 748 898	22,0
chinois	-	5,7
arabe	2 721 463	17,2

TAB. 2.1 – *Taille du corpus parallèle issu des assemblées des Nations Unies.*

automatiquement, de par le manque de corpus et du nombre important de mots hors vocabulaire. Les systèmes de TAS fournissent les meilleures traductions lorsqu'il existe des données d'apprentissage du domaine de taille suffisante (Langlais et al., 2006). Adapter un système à un domaine de spécialité est donc devenu un défi dans la communauté de traitement automatique du langage (TAL) (section 2.4).

2.2.3 Édition a posteriori de traductions

Une des méthodes permettant de palier le manque de données spécialisées est l'édition *a posteriori* de traductions issues de systèmes automatiques. Habituellement effectuée par des humains, cette étape de révision permet de s'assurer de la qualité des traductions et de les adapter, si besoin, aux domaines de spécialités (Krings et Koby, 2001). Cette étape d'édition rajoute un coût au processus de traduction, que ce soit par l'intervention humaine ou la construction d'un système automatisé. De ce fait, la qualité des traductions automatiques agit directement sur l'effort de post-édition. Les chercheurs travaillant en traduction automatique ont donc rapidement envisagé l'automatisation de la post-édition, en se basant sur des approches statistiques similaires à TAS, et permettant de ré-écrire du texte dans un cadre monolingue (section 2.5) (Allen et Hogan, 2000; Allen, 2003).

2.2.4 Couverture du vocabulaire spécialisé

Adapter un système de TAS à un domaine de spécialité implique aussi une couverture du vocabulaire spécifique de ce domaine. Selon Daumé III et Marcu (2006), un corpus d'un domaine particulier contient du vocabulaire spécialisé, mais aussi des mots apparaissant dans des corpus hors domaines (ou *génériques*). Si ces derniers sont pris en charge par les corpus parallèles disponibles, un aspect terminologique reste problématique. L'acquisition de termes et de leurs traductions pour les domaines de spécialités se fait en général manuellement, mais certains travaux montrent des possibilités d'automatisation de cette tâche, et sont détaillés dans la section 2.6 (Fung, 1995; Rapp, 1995).

LANGUE	# PHRASES	# MOTS
bulgare	229 649	-
tchèque	479 636	10 770 230
danois	2 117 839	49 615 228
allemand	1 985 560	48 648 697
grecques	1 344 198	-
anglais	2 032 006	54 720 731
espagnol	1 942 761	55 105 479
estonien	493 198	9 455 337
finlandais	1 929 054	35 799 132
français	2 002 266	57 860 307
hongrois	479 676	10 601 411
italien	1 905 555	52 306 430
lituanien	493 204	9 731 052
letton	473 276	10 024 350
néerlandais	2 147 195	53 459 456
polonais	387 537	8 142 067
portugais	1 942 700	53 799 459
roumain	224 805	5 891 952
slovaque	487 416	10 783 688
slovène	465 985	10 783 688
suédois	2 037 945	45 562 972

TAB. 2.2 – *Taille des données correspondant à la sixième version du corpus parallèle issu des débats du Parlement européen.*

LANGUE	# MOTS (K)	# PHRASES (K)
anglais	22 190	1 428
français	23 771	1 428

TAB. 2.3 – *Taille du corpus parallèle issu des débats du Parlement canadien.*

2.3 Principes de la traduction automatique statistique

Dans cette section, nous abordons les principes fondamentaux de la TAS, avant d'introduire une approche basée sur l'alignement et la traduction de segments sous-phrastiques. Nous expliquons ensuite le fonctionnement d'un décodeur permettant de générer des hypothèses de traductions, suivi par la description d'une implémentation largement utilisée par la communauté en traduction statistique, que nous utilisons aussi dans nos travaux. Puis, nous présentons différentes méthodes d'évaluation automatiques des hypothèses de traduction produites par les systèmes.

2.3.1 Fondements

Brown et al. (1990) proposent un modèle probabiliste selon lequel une phrase P dans la langue source a une traduction possible T dans la langue cible selon la probabilité $p(T|P)$. Cette affirmation peut être interprétée comme la probabilité qu'un traducteur humain produise la phrase cible T, connaissant la phrase source P. Ainsi, ce modèle permet de chercher une phrase $\widehat{T}$, qui soit une traduction possible de P en maximisant la probabilité $p(T|P)$, selon les observations effectuées dans un corpus parallèle. La TAS peut donc se définir selon l'équation :

$$\widehat{T} = \arg\max_{T} p(T|P) = \arg\max_{T} p(P|T).p(T) \tag{2.1}$$

Ce formalisme permet d'isoler deux éléments :
- un modèle de traduction $p(P|T)$, contenant les probabilités inter-langues,
- un modèle de langage $p(T)$, permettant d'évaluer la probabilité d'une séquence de mots.

Cette théorie constitue les fondements de l'approche TAS, dont l'une des premières implémentations a été réalisée dans le cadre du projet *Candide* chez IBM (Berger et al., 1994). Dans les travaux de Brown et al. (1993) sont présentés les modèles statistiques les plus populaires, aujourd'hui connus sous le nom de modèles IBM. Ces modèles permettent l'alignement des mots entre des énoncés source et des énoncés cible, au préalable alignés.

Les modèles IBM sont toujours d'actualité et servent de base à l'alignement au niveau des mots pour la TAS. Cependant, comme montré dans l'exemple 2.2, l'alignement entre deux phrases dans deux langues différentes regroupe parfois des segments sous-phrastiques plus larges qu'un mot. De plus, un mot dans une langue peut se traduire par plusieurs mots dans une autre langue (*la mode* en français, *fashion* en anglais, par exemple). Il est donc nécessaire de prendre en considération le nombre de mots dans la langue cible générés par un mot dans la langue source. C'est ce qu'on appelle généralement la fertilité du mot dans la langue source.

Il aura fallut attendre une dizaine d'années pour que les premiers systèmes se basant sur une méthode d'alignement autre que celle uniquement basée sur les mots permettent d'atteindre des performances significatives. L'alignement et la traduction de segments sous-phrastiques (en anglais, *Phrase-Based Machine Translation* ou PBMT)(Och, 2002; Zens et al., 2002; Koehn et al., 2003; Vogel et al., 2003; Tillmann, 2003) est depuis devenue la référence des approches purement statistiques en traduction automatique. Nous détaillons les étapes pour la construction d'un modèle de traduction de type PBMT dans la section suivante.

2.3.2 Traduction automatique sous-phrastique

L'approche PBMT est basée sur le principe qu'un mot ou un groupe de mots dans une langue peut être traduit par un mot ou un groupe de mots dans une autre langue.

Cette première étape d'alignement nécessite l'utilisation d'un corpus bilingue parallèle, dont l'alignement au niveau des phrases est fourni.

Alignement des mots

En partant de l'équation 2.1, (Brown et al., 1993) introduisent l'alignement au niveau des mots pour l'estimation des probabilités d'un couple de traductions. La probabilité $p(T)$ étant délivrée par le modèle de langage de la langue cible, ils cherchent donc à estimer la probabilité d'observer une phrase P connaissant une phrase T selon l'équation :

$$p(P|T) = \sum_a p(P, a|T) \tag{2.2}$$

$$= \sum_a \prod_{i=1}^{l} (P_i, a_i, l|P_1^{i-1}, a_1^{i-1}, T) \tag{2.3}$$

où l est la longueur de la phrase P, a_i est la position du mot cible aligné avec le mot source P_i. L'alignement des mots est donc obtenu, selon l'algorithme de Viterbi (Viterbi, 1967), en maximisant la probabilité :

$$a' = \arg\max_a p(P, a|T) \tag{2.4}$$

De cette modélisation découlent les modèles IBM, numérotés de 1 à 5, permettant d'associer aux mots sources et cibles des probabilités d'alignement différentes. Ces estimations concernant le ré-ordonnancement des mots peuvent prendre en compte plusieurs paramètres, liés par exemple à la distribution des probabilités dans l'ordre des mots (équiprobables pour le modèle IBM 1), la longueur des phrases source et cible, la position d'un mot cible aligné avec un mot source (information prise en charge dans le modèle IBM 2), la fertilité relative à chaque mot (implémentée dans les modèles IBM 3, 4 et 5), etc. La fertilité d'un mot dans la langue source peut être calculée selon la moyenne des nombres de mots composant sa traduction dans la langue cible, suivant les observations faites dans un corpus parallèle. Les travaux de Och et Ney (2003) tendent à montrer que le modèle IBM4 permet d'obtenir les meilleurs résultats d'alignements.

Alignement de segments

De manière générale, l'alignement de mots entre deux phrases est bi-directionnel, de la langue source vers la langue cible. Ainsi, deux alignements sont obtenus, l'un issu de l'intersection entre les deux langues, l'autre provenant de l'union entre ces dernières. La figure 2.7 permet d'illustrer ce principe, en représentant les mots de deux phrases sur les lignes et les colonnes d'une matrice d'alignement. Les cases noircies présentes les mots (ou ensembles de mots) alignés. Une matrice est construite pour représenter les alignements du français vers l'anglais, et une autre pour l'anglais vers le français. La

mise en commun des alignements présents dans ces deux matrices permet de générer une troisième matrice présentant les alignements issu de l'intersection (cases noircies) ou de l'union (cases grisées) des deux matrices initiales. Un exemple d'alignement de segments sous-phrastiques est présenté par la figure 2.8.

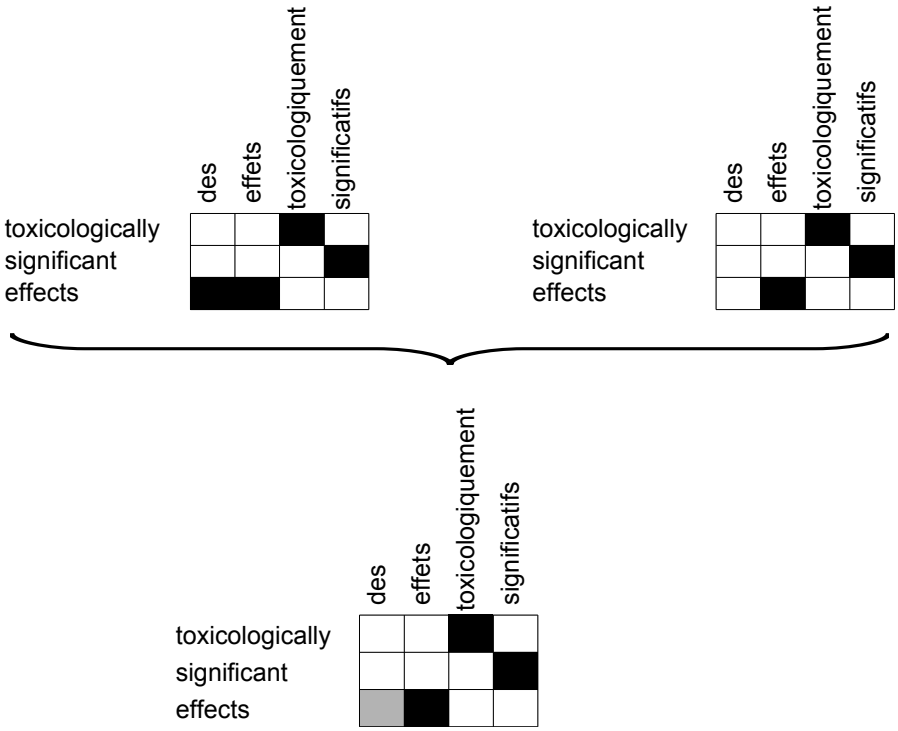


FIG. 2.7 – *Alignement bi-directionnel au niveau des segments entre deux phrases.*

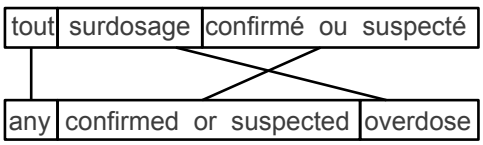


FIG. 2.8 – *Deux phrases en relation de traduction dont les segments sont alignés.*

Les travaux actuels utilisant l'approche PBMT se basent généralement sur cette méthode d'alignement, mais divergent parfois quant aux étapes permettant de délimiter les segments dans une langue alignés avec des segments dans une autre langue (Och et Ney, 2003; Koehn et al., 2003). Ces différences permettent toutefois d'accroître le nombre de paramètres associés à chaque alignement de segments (Och et al., 2004). Une seule exception est à souligner dans les approches d'alignements, les travaux de Marcu et Wong (2002), où les auteurs tentent d'aligner directement les segments sous-phrastiques sans passer par un alignement au niveau des mots effectué par un modèle IBM.

Ainsi, l'estimation des probabilités d'alignement issues de corpus parallèles donne lieu à la construction d'un modèle de traduction, constitué en partie d'une table de

traduction. Cette table regroupe les couples de segments sous-phrastiques sources et cibles alignés, auxquels sont associés leurs scores résultant de l'estimation des probabilités d'observation apprises sur le corpus parallèle. Un extrait d'une table de traduction construite à partir du corpus *Europarl* est présenté dans le tableau 2.4. Les segments sous-phrastiques sont aussi appelés *n*-grammes, correspondant à des sous-séquences de *n* éléments construits à partir d'une séquence donnée (une phrase ou un segment de phrase).

Segment source	Segment cible	Scores
de la sécurité industrielle et de	industrial safety and	(0,5 ; 1)
de la sécurité industrielle et	industrial safety and	(0,5 ; 1)
de la sécurité industrielle	industrial safety	(1 ; 1)
de la sécurité insuffisante	insufficient security	(0,5 ; 1)
de la sécurité insuffisante à nos	insufficient security at our	(1 ; 1)

TAB. 2.4 – *Extrait d'une table de traduction construite sur le corpus parallèle Europarl. Chaque couple de segments est associé aux scores d'alignement source-cible et cible-source.*

2.3.3 Décodeur

La modélisation de l'ensemble des paramètres nécessaires à l'approche PBMT est donc généralement décomposée en plusieurs éléments :

- un modèle de traduction de segments sous-phrastiques, appelé table de traduction,
- un modèle de distorsion (ou de ré-ordonnancement) de segments, permettant de gérer l'alignement des segments,
- un modèle de langage, permettant de s'assurer de la grammaticalité des hypothèses de traduction produites.

L'architecture classique d'un décodeur PBMT est présenté par la figure 2.9

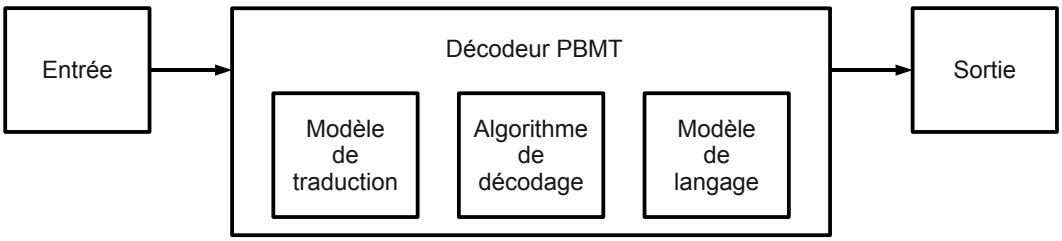


FIG. 2.9 – *L'architecture d'un décodeur PBMT classique.*

Les scores associés à chaque modèle vont permettre au décodeur de sélectionner les segments cibles les plus probables lorsqu'une phrase source est à traduire. Chaque paramètre peut être défini par un ensemble de fonctions f_n ayant comme paramètre les phrases source P et cible T, ainsi que leurs alignements possibles.

En partant du modèle de traduction, le parcours des paires de segments sous-phrastiques (P_i, T_i) va permettre de générer des hypothèses de traduction. Puis, en s'appuyant sur un modèle de langage, les scores de ces hypothèses sont pondérées par la probabilité d'observer ces segments dans l'ordre proposé par le modèle de distorsion. Finalement, le décodeur peut sélectionner une traduction en suivant l'équation :

$$\widehat{T} = \arg\max_{T} p(T|P) \tag{2.5}$$

$$\propto \sum_{n} \lambda_n log(f_n(a, P, T)) \tag{2.6}$$

Les paramètres λ_n sont des poids associés à chaque fonction f_n et permettant de minimiser les erreurs de traduction, selon un corpus parallèle. Cette phase de réglage des poids des paramètres du modèle de traduction est appelée *développement* ou *optimisation*. Un corpus dédié à cette tâche est donc généralement utilisé et est appelé *corpus de développement*. Les poids sont optimisés de manière itérative selon les scores obtenus lors de la traduction du corpus parallèle de développement, en se basant sur une traduction de référence et une métrique automatique.

Cette phase d'optimisation est généralement effectuée selon une méthode proposée par Och (2003) sous la dénomination *MERT* (Minimum Error-Rate Training). Cet algorithme consiste à affecter à chaque fonction f_n le poids λ_n permettant d'obtenir les traductions les plus proches de la traduction de référence du corpus de développement. Concrètement, les λ_n sont sélectionnés chacun leur tour d'après les résultats de traduction du corpus, évalués par la métrique automatique BLEU (voir section 2.3.5). La progression du système est ainsi mesurée après chaque itération permettant de converger vers une solution sous-optimale, ce qui permet d'améliorer le score général mesuré sur l'ensemble du corpus d'optimisation.

2.3.4 Une implémentation : Moses

L'approche PBMT permet actuellement d'obtenir de bonnes performances et les campagnes d'évaluations internationales ont montré une prédominance de ces systèmes. Aussi, nous avons choisi d'adopter ces méthodes dans nos travaux, proposées dans la boîte à outils de traduction automatique *Moses* (Koehn et al., 2007). *Moses* propose l'ensemble des outils nécessaires à la construction d'un modèle de traduction. Un décodeur permet aussi d'utiliser ces outils afin de produire des hypothèses de traduction d'un texte initial.

Plusieurs étapes sont nécessaires à la construction d'un système de traduction avec *Moses* et font appel aux différents outils mis à disposition. Tout d'abord, il est nécessaire d'effectuer un alignement au niveau des mots à partir du corpus parallèle. Puis, l'ensemble des alignements obtenus forment une table servant de base à la construction de la table de traduction. Cette dernière est élaborée selon les scores d'alignement obtenus par les segments sous-phrastiques présents dans le corpus parallèle. Un modèle de ré-ordonnancement est ensuite construit, contant les informations relatives aux positions des segments traduits au sein des phrases.

Afin d'estimer les probabilités d'alignement de mots, *Moses* encapsule l'outil GIZA++, qui implémente les algorithmes des modèles IBM 1-5 (Och, 2000). Nous utilisons aussi une version parallélisable de ce programme, MGIZA (Gao et Vogel, 2008). Les modèles de langage peuvent, quant à eux, être construits à l'aide d'autres outils, comme celui proposé par le SRI (Stolcke, 2002), ou encore par l'IRST (Federico et al., 2008).

Certains systèmes commerciaux permettent aussi de générer des traductions de qualité, mais les briques logicielles mises en œuvre ne sont que rarement dévoilés. Il apparaît toutefois intéressant d'utiliser ce genre de systèmes pour évaluer la qualité des traductions produites. Il est aussi possible de comparer un système état-de-l'art PBMT comme *Moses*, utilisant des ressources parallèles disponibles, avec un système *boîte noire*, simplement utilisé comme décodeur.

2.3.5 Évaluation automatique

Évaluer la qualité des traductions produites par un système automatique peut se faire manuellement, par des humains. On parle alors d'évaluation subjective. Lorsque l'on dispose de références de traduction, l'évaluation automatique (objective) est alors possible. L'évaluation humaine étant généralement plus coûteuse, elle permet cependant d'avoir une meilleur estimation de la qualité des traductions. En effet, elle permet d'intégrer plusieurs critères comme la lisibilité, la justesse grammaticale et syntaxique, etc. (Specia, 2011).

Récemment, les coûts liés à l'évaluation manuelle de traductions ont pu être réduits par l'utilisation de plates-formes en ligne regroupant des personnes pouvant effectuer des tâches parcellisées. Un grand nombre d'évaluateurs sont donc disponibles simultanément et ce qui permet une répartition de l'effort d'évaluation des traductions. Des sous-ensembles de données à évaluer sont formés parmi les hypothèses de traduction produites par un système, puis soumis aux participants à l'évaluation. L'adéquation entre les utilisateurs s'obtient par le croisement des résultats obtenus. Ce protocole d'évaluation est présenté de manière plus détaillé dans les travaux de (Callison-Burch, 2009).

Nous pouvons cependant relever des limites à ces méthodes d'évaluation manuelle, comme le soulignent certains auteurs dans leurs travaux (Sagot et al., 2011). En effet, les plateformes d'évaluation à bas coût risquent d'être vues comme une norme dans le processus de création de ressources linguistiques. De ce fait, faire intervenir des professionnels pour évaluer des traductions automatiques pourrait devenir trop onéreux si les budgets dédiés à cette tâche se calquent sur les coûts réduits des plateformes en ligne.

Mesure de taux d'erreur

D'une manière générale, lorsque l'on dispose d'une traduction de référence pour évaluer une hypothèse de traduction, il est possible de mesurer le nombre de modifications nécessaires à apporter à la sortie d'un système pour atteindre la référence. Une métrique utilisée en évaluation automatique de traduction est le WER (Word Error Rate, ou taux d'erreur mot), permettant d'évaluer la distance par la combinaison des insertions, délétions et substitutions observés entre une hypothèse de traduction et une référence (Nießen et al., 2000). La mesure PER est basée sur WER mais permet de ne pas pénaliser des mots étant à la mauvaise position dans une phrase (Tillmann et al., 1997).

Toujours dans cet esprit de distance d'édition, les métriques TER (Snover et al., 2006) (Translation Error Rate, pour taux d'erreur de traduction) et TERp (Snover et al., 2009) (TER plus) s'orientent vers l'idée de l'effort d'édition à fournir à posteriori pour atteindre la référence à partir d'une hypothèse de traduction. Ces méthodes permettent de mesurer tous les événements pris en charge par WER, et aussi les décalages possibles entre des segments, peu importe leurs tailles. La version *plus* permet d'inclure des données linguistiques, comme des synonymes par exemple, et est optimisé pour être très corrélée avec le jugement humain.

Mesure de précision

L'une des méthodes les plus populaires en évaluation automatique de traductions réside dans la comparaison des segments de 1 à 4 mots en commun entre l'ensemble des hypothèses émises par un système et leurs traductions de référence. Cette approche fut introduite par Papineni et al. (2002) et est appelée BLEU (pour *Bilingual Evaluation Understudy*). Le nombre de n-grammes (généralement avec n entre 1 et 4) apparaissant dans les hypothèses et dans leurs références est normalisé par le nombre total de n-grammes contenu dans les hypothèses. Une pénalité permet de défavoriser les hypothèses trop courtes. Des hypothèses identiques aux références obtiennent le score le plus élevé (score de 1).

Une autre métrique, proposée par Banerjee et Lavie (2005) et connue sous le nom de METEOR (pour *Metric for Evaluation of Translation with Explicit Ordering*), permet de prendre en compte différents paramètres linguistiques comme les étiquettes morpho-syntaxiques, les synonymes, etc. L'ordre des segments est lui aussi pris en compte, et il est possible d'optimiser l'ensemble de ces paramètres afin d'orienter l'évaluation vers le jugement humain.

2.4 L'adaptation des modèles statistiques

Une des limites de l'approche PBMT réside dans l'adéquation entre les données d'apprentissage et de test. Les meilleures traductions sont obtenues lorsque les données

sont du même domaine. De ce fait, un système peut obtenir de très bon résultats avec certaines données, et s'effondrer avec des données à traduire d'un domaine différent. C'est pour cela que l'adaptation aux domaines de spécialités est rapidement devenue populaire auprès des chercheurs en traduction automatique. Nous proposons dans cette section de présenter les différentes approches existantes pour adapter les modèles statistiques des systèmes PBMT.

Un exemple d'expérimentations en adaptation au domaine des actualités (journalisme) est proposé par Koehn et Schroeder (2007). Ils proposent d'étudier l'impact de différents paramètres, comme l'utilisation de corpus hors-domaine de taille importante, de combiner des corpus spécialisés, ou non, pour la construction des modèles statistiques, etc. Le tableau 2.5 regroupe les scores BLEU obtenus selon les paramètres étudiés.

Configuration	Score BLEU (%)
Grand corpus hors-domaine	25,11
Petit corpus spécialisé	25,88
Corpus combinés	26,69
Modèle de langage spécialisé	27,46
Modèles de langage interpolés	27,12
Deux modèles de langage séparés	27,30
Deux modèles de traductions séparés	27,64

TAB. 2.5 – *Résultats obtenus par Koehn et Schroeder (2007) lors de ses expériences pour l'adaptation d'un système PBMT générique au domaine des actualités.*

Les résultats présentés ici correspondent à différentes configurations testées par les auteurs. Dans un premier temps, un modèle de traduction est construit en utilisation un corpus hors-domaine d'une taille importante. Dans un second temps, c'est un corpus plus réduit concernant le domaine de spécialité qui est utilisé. Les auteurs combinent ensuite les corpus afin de construire un modèle de traduction mixte. Puis, ce sont les modèles de langages qui sont testés, en gardant la meilleure configuration du modèle de traduction. Un modèle de langage spécialisé est tout d'abord utilisé, suivi par un modèle de langage mixte basé sur l'interpolation des modèles de langage hors-domaine et spécialisé. Finalement, c'est le système de traduction utilisant deux modèles de langage séparés et deux modèles de traduction séparés qui permet d'obtenir les meilleurs résultats.

2.4.1 Premiers Travaux

C'est dans les travaux de Langlais (2002) que l'adaptation aux domaines de spécialité fut introduite dans une tâche de traduction automatique. L'idée générale est d'utiliser une ressource extérieure : un lexique du domaine concerné, afin d'influencer le décodeur du système de traduction sur ses choix dans la sélection de segments sous-phrastiques issus de son modèle. Ainsi, le *taux d'erreur mot* observé en appliquant cette méthode est fortement diminué. Les auteurs proposent cette approche pour générer

des traductions de meilleure qualité pour une tâche de traduction dans un domaine de spécialité en utilisant un système générique. Ces travaux sont des précurseurs dans l'adaptation aux domaines de spécialité, et ont donné lieu à différentes approches étudiées par plusieurs chercheurs. Les sous-sections suivantes traitent de ces travaux, portant sur l'adaptation des modèles statistiques, que ce soit le modèle de langage ou de traduction.

2.4.2 Modèle de langage

La modélisation du langage est un critère important dans de nombreuses tâches de TAL, que ce soit en reconnaissance de la parole, en recherche d'information, ou en traduction automatique. Pouvoir adapter un modèle à un domaine de spécialité permet donc d'adapter un système à une tâche particulière pour un domaine donné. Suite à ce constat, de nombreux auteurs se sont penchés sur les possibilités liées à l'adaptation des modèles de langage pour la traduction automatique.

Dans les travaux de Eck et al. (2004), c'est le modèle de langage utilisé au décodage qui est adapté au domaine journalistique. Des méthodes de recherche d'information sont appliquées pour récolter de nouvelles données permettant d'augmenter la couverture des données initiales. De manière générale, l'architecture mise en place comprend les étapes suivantes :

- traduire le texte d'un domaine spécialisé avec un modèle de langage générique,
- utiliser la traduction produite pour récolter des documents contenant un vocabulaire et une syntaxe similaire,
- construire un modèle de langage adapté en utilisant les documents récoltés,
- traduire à nouveau le texte initial, avec le modèle de langage adapté.

Cette idée provient d'autres travaux effectués en reconnaissance de la parole, notamment par Mahajan et al. (1999). Un ensemble de possibilités liées à l'adaptation de modèles de langage aux domaines de spécialités sont présentées dans les travaux de Janiszek et al. (2001).

Les travaux de Zhao et al. (2004) proposent une idée légèrement différente, car toujours basée sur la recherche d'information, non plus pour récupérer des documents, mais cette fois pour collecter des phrases. Les auteurs cherchent, à partir d'un premier ensemble d'hypothèses de traductions, des phrases similaires dans un corpus monolingue dont le contenu appartient au domaine de spécialité. Tout comme l'approche présentée dans (Eck et al., 2004), deux étapes de traductions sont nécessaires, entrecoupées d'une phase de recherche d'information et d'une phase d'adaptation du modèle de langage initial.

Interpoler des modèles de langages reste une méthode populaire pour l'adaptation aux domaines de spécialité, comme le prouvent les travaux de Hasan et Ney (2005). Les auteurs proposent une méthode de construction de modèles de langage basée sur des classes spécifiques aux domaines. La classification est effectuée au niveau des phrases afin de capturer leurs différences syntaxiques, en se basant sur des expressions régulières. Cette méthode permet une réduction significative de la perplexité en

utilisant les modèles de langage spécifiques par rapport à un modèle de langage générique. De plus, dans une tâche de traduction automatique entre l'anglais et l'espagnol, ils mesurent une baisse du taux d'erreur mot et un gain en terme de score BLEU.

2.4.3 Modèle de traduction

Lors de la campagne d'évaluation en traduction automatique NIST 2003, Byrne et al. (2003) proposent d'entraîner un modèle de traduction sur des données parallèles collectées grâce à des méthodes de recherche d'information. Ils utilisent la distance cosinus entre les phrases d'un document source à traduire et un corpus parallèle. Un ensemble de phrases bilingues parallèles adaptées au document à traduire est ainsi collecté. Chaque ensemble permet l'apprentissage d'un modèle de traduction adapté, qui est combiné au modèle générique, afin d'augmenter la quantité de segments alignés présents dans le modèle de traduction. Cette approche permet d'augmenter les scores BLEU d'un point environ sur une tâche de traduction entre l'anglais et le chinois.

Cette méthode basée sur la recherche d'information reste très populaire dans les approches d'adaptation aux domaines de spécialité, et les travaux de (Hildebrand et al., 2005) en donnent un très bon aperçu. L'idée générale est d'adapter le modèle de traduction à un domaine de spécialité en enrichissant le corpus d'apprentissage avec des phrases parallèles du domaine provenant de corpus parallèles hors domaine. Cette idée est particulièrement séduisante car elle permet d'utiliser des données génériques en sous-échantillonnant les phrases se rapprochant du domaine, selon des méthodes de recherche d'information mesurant la similarité entre deux phrases. Lorsqu'un ensemble de phrases de test est à traduire, il est tout d'abord utilisé afin de collecter des phrases parallèles similaires dans un corpus hors domaine. Puis, un modèle de traduction et un modèle de langage sont construits sur les phrases collectées, avant d'être utilisés pour traduire les phrases de test. Cette approche étend aux modèles de traductions l'adaptation des modèles de langage, basée sur la recherche d'information spécialisée, et présentées dans la section 2.4.2.

Dans les travaux de (Ueffing, 2006), une approche d'auto-entrainement est proposée, permettant au système de traduction automatique de prendre en considération ses propres hypothèses de traduction. Un algorithme en deux étapes est présenté, en suivant la démarche :

1. Traduction d'un corpus de test spécialisé avec un système générique.
2. Retirer les traductions erronées.
3. Aligner le corpus source avec les bonnes traductions.
4. Apprendre un modèle de traduction adapté.
5. Combiner le modèle adapté avec le modèle générique.
6. Répéter les étapes 1 à 5 en se basant sur le modèle construit précédemment.

La méthode d'auto-apprentissage mène à des gains allant jusqu'à 2 points de BLEU en comparaison avec un système de traduction générique. Le fait d'itérer ne permet pas d'améliorer significativement ces gains.

2.5 La Post-édition de traductions

La qualité des traductions produites par un système automatique n'est souvent pas suffisante pour permettre une automatisation complète du processus de traduction, notamment à cause des erreurs liées à l'ambiguïté des mots, la prise en compte du contexte, la couverture du vocabulaire, etc. L'intervention d'un humain est donc parfois utile pour corriger certaines erreurs que commettent les systèmes automatiques.

Nous proposons dans cette section de définir la post-édition avant d'introduire le concept de post-édition automatique par une approche statistique (SPE, pour *Statistical Post-Edition*). De nombreux travaux ont montré l'intérêt d'utiliser un système PBMT pour éditer *a posteriori* les hypothèses issues d'un système de traduction basé sur des règles (RBMT, pour *Rule-Based Machine Translation*). L'adaptation à un domaine de spécialité peut aussi s'effectuer lors de cette seconde phase dans le traitement des données à traduire. Nous présentons ces travaux dans les sous-sections 2.5.3 et 2.5.3.

2.5.1 Définition

La post-édition de traductions fait référence à des modifications *a posteriori* des hypothèses produites par un système. Une étape d'édition peut impliquer des modifications et des corrections d'erreurs. Ainsi, l'édition d'une traduction automatique T' revient à générer la phrase T'', comportant les modifications grammaticales et syntaxiques nécessaires à l'amélioration de la lisibilité de T', tout en gardant le sens de la phrase source S (voir la figure 2.10). Cette étape implique un coût pouvant dépasser celui induit par la traduction humaine. Savoir si la post-édition de traductions automatiques est moins coûteuse que la traduction humaine dépend de la qualité des hypothèses produites par les systèmes.

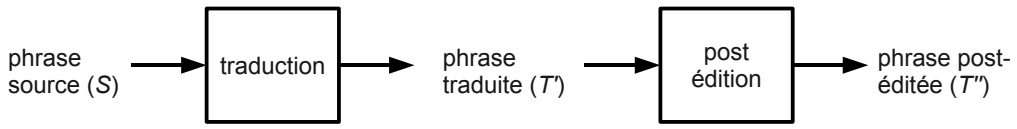


FIG. 2.10 – *Principe général de la post-édition de traductions.*

Si les systèmes de traduction automatique peuvent produire des erreurs récurrentes, la post-édition manuelle devient alors rapidement un processus répétitif. L'idée de capturer ces événements est alors apparu possible, afin de construire un modèle permettant la réécriture dans un cadre monolingue. En se basant sur des sorties de systèmes automatiques et des références de traduction, l'observation des différences au

niveau *n*-grammes permet d'estimer les probabilités de rencontrer ces erreurs. Il reste cependant nécessaire de disposer d'une quantité suffisante de données d'apprentissage pour le modèle de post-édition, afin de couvrir un ensemble d'exemples contenant les erreurs pouvant être produites sur un corpus de test.

2.5.2 Coût de la post-édition

L'étape de post-édition manuelle implique l'intervention d'humains. Le coût de cette intervention peut être mesuré selon des critères comme le nombre de modifications à apporter pour obtenir une traduction grammaticalement correcte, le nombre de mots ou de segments n'ayant pas été traduits, le temps nécessaire à la post-édition d'une phrase, etc. Ces critères se rapprochent de ceux utilisés généralement pour l'estimation de la qualité de traductions issues de systèmes automatiques. Dans les travaux de Specia (2011), les auteurs définissent des niveaux d'effort à fournir pour la post-édition au niveau des phrases.

Un premier niveau correspond à la reformulation intégrale d'une hypothèse de traduction émise par un système automatique. Le coût de la post-édition revient alors à re-générer une traduction complète, et est donc équivalent à une traduction manuelle. C'est le cas le plus extrême concernant le coût de post-édition. Un second niveau correspond à l'édition de certains segments sous-phrastiques, dans le cas où certaines portions d'une phrase ont été mal traduites par le système automatique. Ce type de modifications entraîne un coût inférieur à celui induit par la reformulation complète d'une phrase traduite, et est donc plus rapide qu'une traduction manuelle. Un troisième niveau concerne la modification minime de certains éléments d'une phrase, comme la traduction de mots hors-vocabulaire, ou l'ajout de ponctuation par exemple. La traduction issue d'un système automatique et correspondant à ce cas implique peu d'efforts de post-édition, et implique donc un coût moindre en comparaison à une traduction manuelle. Enfin, un quatrième niveau correspond à aucune post-édition, c'est à dire que la traduction générée par un système automatique correspond aux attentes en terme de grammaticalité et de couverture du vocabulaire.

2.5.3 Approches statistiques

Les premières idées d'apprentissage automatique à partir textes post-édités proviennent des travaux de Knight et Chander (1994); Allen et Hogan (2000), proposant d'utiliser un corpus parallèle composé de trois éléments, à savoir un texte source, sa traduction par un système automatique, et ce texte post-édité manuellement. Les travaux de Simard et al. (2007a) s'en inspirent et introduisent l'approche PBMT appliquées à la post-édition statistique. Ainsi, il est possible d'utiliser les méthodes d'alignement de mots et de segments présentés dans la section 2.3.2. Le processus de post-édition statistique n'est donc pas différent de celui de TAS, dans ce cas, il s'agit de *traduction* monolingue. Un décodeur PBMT pourra appliquer les règles apprises sur un corpus post-édité.

Ces premières expérimentations permettent d'observer les amélioration possibles de traductions issues d'un système RBMT par une étape de post-édition statistique. De nombreux auteurs ont depuis expérimentés la même approche dans des conditions différentes. Les données utilisées pour construire le modèle de post-édition peuvent provenir ou non de traductions éditées manuellement. Cette supervision de l'apprentissage implique toujours un intervenant humain afin d'appliquer des corrections sur les sorties de systèmes de traduction. Des méthodes n'utilisant pas des textes post-édités manuellement, mais des traductions de références, ont aussi été étudiées, et peuvent être considérées comme des approches non supervisées à la post-édition statistique. Ces deux types d'approches, supervisées ou non, sont présentées dans les deux sous-sections suivantes.

Approches supervisées pour la post-édition statistique

Parmi les approches comportant une étape de post-édition manuelle, nous pouvons remarquer les travaux combinant des systèmes RBMT et PBMT pour la traduction et la post-édition respectivement. Selon Simard et al. (2007a), un système RBMT commercial et le système PBMT Portage (Sadat et al., 2005) combinés permettent, sur une tâche de traduction français-anglais, de dépasser les performances de chaque système individuellement. Par rapport au système de traduction par règle utilisé seul, l'étape de post-édition statistique permet d'accroître en absolu le score BLEU jusqu'à 13,7%. Les auteurs combinent aussi deux systèmes PBMT en série, et observent une amélioration des résultats après la post-édition lorsque les corpus d'apprentissage des modèles de traduction et de post-édition sont différents.

Dans les travaux de Dugast et al. (2007), une étude qualitative plus poussée est menée au niveau linguistique sur les hypothèses de traduction avant et après post-édition. Ils combinent un système RBMT (Systran (Toma, 1977)) et un système PBMT (Portage ou Moses) et mesurent dans cette dernière configuration des gains allant jusqu'à 10 points de BLEU sur une tâche de l'allemand vers l'anglais par rapport à un système utilisé individuellement.

Un peu plus récemment, Potet et al. (2011a) proposent, dans une expérience préliminaire, de combiner deux systèmes PBMT en série, un pour la traduction source-cible, l'autre pour la post-édition cible-cible, entre le français et l'anglais. Les auteurs utilisent des données post-éditées provenant du premier système PBMT à différents niveaux dans l'architecture générale de leur approche :

- en enrichissant le corpus d'apprentissage du modèle de traduction,
- en constituant un corpus d'apprentissage du modèle de post-édition,
- en optimisant les poids du modèle linéaire de traduction avec les données post-éditées.

Des gains sont observés dans les trois configurations par rapport à l'utilisation seule d'un système de traduction par segments, mais restent toutefois faibles (moins d'un demi point de BLEU sur le corpus de test). Aussi, cette approche n'a pas été testée en utilisant des données d'apprentissage de plus grande taille et ne peut donc pas être généralisée.

Il apparaît que, dans l'ensemble de ces expérimentations, l'interaction avec des humains reste nécessaire afin de constituer les corpus parallèles contenant les post-éditions manuelles d'hypothèses de traductions. Il est, en général, plus aisé de disposer de corpus traduits (les corpus parallèles classiques), et il est possible de faire de l'adaptation de systèmes de traduction automatique aux domaines de spécialité, en injectant des données parallèles du domaine dans une étape de post-édition.

Approches non supervisées pour la post-édition statistique

Disposer de corpus de post-édition reste finalement assez rare, et certains auteurs ont préféré utiliser les corpus parallèles *classiques*. Cette approche consiste à introduire une petite quantité de données parallèles du domaine de spécialité dans la phase de post-édition, et ainsi de constituer des corpus parallèles contenant trois parties : la source, la traduction automatique, et la traduction de référence. Ce sont les différences entre les hypothèses de traduction et la référence qui sont capturées par le modèle de post-édition, et non plus celles entre les hypothèses avant et après édition manuelle.

Kuhn et al. (2010) expérimentent cette approche dans leurs travaux, et la compare à l'approche basée sur les post-édition manuelle, comme dans les travaux de Simard et al. (2007a). Ils évaluent les traductions produites par le système RBMT Systran, celles provenant du système statistique par segments Portage, et celles issues de la combinaison RBMT-PBMT pour la traduction et la post-édition entre les couples anglais-français et chinois-anglais. Encore une fois, les auteurs montrent des gains lors de la combinaison par rapport aux systèmes utilisés seuls. Pour la traduction du français vers l'anglais, la post-édition améliore de 10,2% en absolu le score BLEU obtenu par le système Systran seul. Cependant, pour cette tâche, le système de traduction statistique par segments obtient des scores similaires à ceux obtenus par la post-édition des hypothèses produites par le système à base de règles. L'un des principaux avantages présentés dans ces travaux réside dans le fait qu'un corpus parallèle *classique* est constitué indépendamment des sorties du système à base de règles, contrairement aux corpus constitués en partie sur les post-édition manuelle de sorties de ce type de systèmes.

Plus récemment, les travaux de Béchara et al. (2011) présentent une approche de post-édition statistique combinée à un système de traduction par segments. Les auteurs proposent de mettre en cascade deux systèmes PBMT et d'évaluer les gains possibles liés à la post-édition. Deux séries d'expériences sont menées : la première concerne l'application *naïve* de la post-édition sur l'ensemble des phrases du corpus de test, la seconde repose sur la modélisation du contexte dans la langue source. La boîte à outils *Moses* est utilisée pour construire les systèmes de traduction et de post-édition. Les langues concernées sont le français et l'anglais. Selon leur meilleure configuration, les auteurs rapportent un gain en terme de BLEU de plus de 2 points, permettant de passer de 61,6% à 63,9% dans le sens du français vers l'anglais.

2.5.4 Adaptation aux domaines

Deux grandes approches peuvent donc être appliquées pour la post-édition statistique : utiliser des post-éditions manuelles ou utiliser des traductions humaines. Ces deux manières de constituer le corpus d'apprentissage destiné à l'apprentissage du modèle de post-édition permettent d'atteindre, selon certaines configurations, des performances supérieures à l'utilisation de systèmes de traduction à base de règles ou de segments utilisés seuls. L'approche utilisant les données de post-édition manuelle obtient en général de meilleurs résultats que celle utilisant les traductions humaines. Cependant, cette dernière ressource, bien que rare pour des domaines de spécialité, est plus largement disponible que des corpus post-édités. L'approche basée sur l'utilisation de petits corpus parallèles de spécialité, introduits lors de l'étape de post-édition, reste une alternative intéressante pour l'adaptation aux domaines de spécialité en traduction automatique.

Isabelle et al. (2007); Simard et al. (2007b) mettent en place un protocole d'évaluation pour cette méthode en s'inspirant des travaux de Simard et al. (2007a) pour la combinaison en série d'un système de traduction par règle et d'un système de traduction statistique par segments. La principale différence entre ces deux travaux réside dans les données d'apprentissage utilisées pour construire le modèle de post-édition. Des gains significatifs (plus de 20 points de BLEU pour la traduction du français vers l'anglais) sont reportés suivant leur méthode et permettent effectivement d'adapter un système de traduction à un domaine de spécialité en introduisant les données du domaine dans une seconde étape.

de Ilarraza et al. (2008) proposent d'appliquer cette approche pour traduire des données spécialisées entre l'espagnol et le basque avec un système de traduction automatique RBMT (Matxin (Alegria et al., 2005)) non spécialisé, et une petite quantité de données introduites dans un système PBMT de post-édition (environ 50 k phrases). Très peu de données sont utilisées pour le domaine, mais les auteurs introduisent des information morphologiques dans les deux systèmes mis en place. Ils mesurent des gains relatifs en terme de score BLEU allant jusqu'à 200% dans leurs expériences comparant la post-édition statistique à un système de traduction par règles seul. Des performances moindres mais toujours significatives sont observées entre la post-édition statistique et un système de traduction statistique.

2.6 L'acquisition de lexiques multilingues

Dans les systèmes de TAS, la couverture du vocabulaire de spécialité peut être accrue en utilisant des ressources extérieures, comme un lexique ou une ontologie bilingue. Ces ressources sont particulièrement utiles en traduction automatique et aussi dans d'autres domaines de recherche, mais restent coûteuses à développer et le dynamisme de certains domaines (comme l'informatique par exemple) nécessite la mise à jour constante du vocabulaire de spécialité. Il parait donc intéressant de construire automatiquement ces ressources.

Si les méthodes les plus efficaces pour construire des lexiques bilingues s'appuient sur des corpus parallèles (Gale et Church, 1991; Koehn et Knight, 2000, 2001), la disponibilité de ces corpus reste un problème. Les ressources monolingues dans les domaines de spécialité sont, par contre, beaucoup plus courantes. Cependant, et contrairement aux corpus parallèles, aucun alignement n'est présent entre deux textes dans deux langues différentes. Extraire des couples de traductions est toutefois possible en se basant sur d'autres paramètres que l'alignement au niveau des phrases. Cette section présente les méthodes les plus populaires. Puis, nous détaillons d'une manière plus approfondie une approche basée sur les contextes lexicaux, pouvant être appliquée dans des domaines de spécialité ou non.

2.6.1 Liens morphologiques

Dans cette section, nous nous intéressons particulièrement aux relations orthographiques entre des mots étant en relation de traduction. Après une brève introduction, nous proposons de détailler les travaux effectués précédemment sur les cognats et les homographes inter-langues. Nous présentons ensuite un ensemble de travaux portant sur les analogies morphologiques entre des traductions.

Généralités

En linguistique, les morphèmes représentent les plus petites sections portant un sens et composant les mots. Il a été démontré que des similitudes morphologiques peuvent exister entre des traductions de langues reliées par leurs origines, comme certaines langues romanes par exemple (une excellente étude des lexiques romans a été menée par Posner (1996)). De ce fait, un grand nombre de travaux ont permis d'extraire automatiquement des couples de mots étant des traductions en comparant leur orthographe, leurs morphologies, et les relations existantes dans la construction des mots entre deux langues.

Dans une langue, l'emprunt d'un mot d'une autre langue est un phénomène linguistique particulier, pouvant permettre de repérer des traductions. Entre le français et l'anglais par exemple, certains domaines de spécialité contiennent du vocabulaire dont l'orthographe est similaire dans une langue ou dans l'autre. Nous pensons notamment au vocabulaire lié à l'informatique et aux nouvelles technologies de l'information et de la communication, où très souvent un mot spécialisé ayant été importé de l'anglais fait son apparition en français. Ainsi, certains domaines de spécialité peuvent faciliter la reconnaissance de traductions de mots, grâce aux emprunts inter-langues, ou translittérations (Koehn et Knight, 2002; Prochasson, 2009).

Analogie

La traduction d'unités lexicales par analogie met en relation quatre entités, notées généralement $[x : y :: z : t]$, et se lisant "x est à y ce que z est à t". La correspondance

entre des représentations linguistiques peut faire l'objet d'un raisonnement par analogie, dont les fondements théoriques sont présentés dans les travaux de Pirrelli et Yvon (1999). Le bien fondée de cette approche, appelée *apprentissage analogique*, est appuyée par les expériences effectuées par Lepage (2003). Des systèmes de traduction basés sur l'analogie formelle (l'analogie des caractéristiques orthographiques d'unités lexicales) ont depuis été développés et proposés lors de campagnes d'évaluation, comme IWSLT'07. Les méthodes d'apprentissage de la morphologie d'une langue en se basant sur l'étude des relations analogique sont parfaitement détaillées dans les travaux de Lavallée (2010).

La traduction de termes médicaux par analogie est proposée par Langlais et al. (2009), suite aux travaux effectués par Langlais et Patry (2007) pour la traduction automatique de mots inconnus dans plusieurs langues européennes. Les auteurs décrivent un algorithme permettant de résoudre de manière efficace une équation analogique (soit $[x : y :: z :?]$). Ils mettent en place un classifieur pour sélectionner les bonnes traductions candidates. Leur approche s'avère très intéressante car indépendante des langues. En effet, sur les 5 langues étudiées, les performances de leur système ne baissent pas, même face à des langues à morphologie riche, comme le finnois par exemple. Selon leur meilleure configuration, jusqu'à 30% des termes à traduire ont été traités et la précision atteinte est de 100%.

Cognats et homographes inter-langues

Selon Pierre Zweigenbaum, des cognats sont des mots « identiques ou de graphie proche d'une langue à l'autre, comme *gouvernement* et *government* » (Zweigenbaum et Habert, 2006). Cette particularité présente entre certains mots dans deux langues permet de reconnaître des traductions par simple comparaison orthographique. Nous pouvons par exemple utiliser la distance d'édition connue sous le nom de distance de Levenshtein (Levenshtein, 1966). Cette mesure consiste à comparer deux mots, selon le nombre d'insertions, délétions et substitutions, à effectuer sur un mot pour le faire correspondre à l'autre. La figure 2.11 illustre cette méthode par un exemple de calcul de distance entre deux mots en français : Avignon et aviron. La distance résultante est égale à 2.

Les cognats ont la particularité de couvrir une dimension sémantique similaire, en plus de leur relation orthographique. Ils ont une signification proche ou identique (Dijkstra et al., 1999). Il faut donc faire attention aux *faux-amis* lors de l'extraction de traductions à partir de corpus. En effet, le mot français *pain* ne porte pas le même sens que son homographe anglais (où *pain* signifie *douleur* en français). Baser un système d'extraction lexical bilingue uniquement sur l'orthographe des mots peut donc s'avérer être une mauvaise piste. Il faut pouvoir repérer des mots qui vont partager des similarités orthographique, sans partager leur(s) sens. Nous pouvons appeler ces mots des homographes inter-langues.

Combiner une approche orthographique avec une approche sémantique peut permettre de repérer des traductions dans un corpus en évitant les *faux-amis*. Koehn

		A	V	I	G	N	O	N
	0	1	2	3	4	5	6	7
A	1	0	1	2	3	4	5	6
V	2	1	0	1	2	3	4	5
I	3	2	1	0	1	2	3	4
R	4	3	2	1	1	2	3	4
O	5	4	3	2	2	2	2	3
N	6	5	4	3	3	2	3	2

FIG. 2.11 – *Exemple de calcul de la distance de Levenshtein entre les mots Avignon et aviron.*

et Knight (2002) mettent en application cette méthode de combinaison orthographique et sémantique. Ils proposent d'utiliser deux paramètres basés sur la comparaison orthographique des mots entre l'allemand et l'anglais. Ils s'intéressent tout d'abord aux mots identiques, puis aux mots partageant des similarités morphologiques sans être totalement identiques. Ces deux paramètres sont ensuite combinés avec des mesures effectuées sur les contextes des mots selon une fenêtre d'observation, sur les fréquences d'apparition dans un corpus, et ce qu'ils appellent la similarité des mots. Ce dernier aspect est intéressant, car il permet d'établir des relations telles que : *chat* est à *chien* en français, ce que *cat* est à *dog* en anglais.

Cet aspect sémantique pouvant permettre de mettre en relation des mots étant des traductions est particulièrement populaire depuis l'apparition des modèles thématiques. L'analyse sémantique latente en est l'exemple le plus connu, permettant notamment de faire de l'indexation et de la recherche documentaire, mais aussi, depuis quelques années, de mettre en exergue des mots dans des langues différentes partageant des similarités thématiques. Nous présentons les bases de l'analyse sémantique latente dans la prochaine section, suivi des applications possibles pour l'alignement de mots multilingues, ou l'acquisition de lexiques bilingues.

2.6.2 Modélisation thématique

Afin de modéliser les aspects thématiques relatifs aux mots d'un corpus et d'en étudier les applications à la traduction automatique, nous présentons dans cette section l'analyse de la sémantique latente (LSA) (Deerwester et al., 1990). Cette méthode permet de construire des classes thématiques sous la forme de sacs de mots pondérés constitués par le vocabulaire d'un texte. Après différentes extensions, cette modélisation est utilisée pour traiter des données multilingues et en extraire les similitudes thématiques.

Analyse de la sémantique latente

Ce type de modélisation à partir d'un corpus de documents est une approche courante en indexation documentaire et en recherche d'information, plus connue sous le nom d'indexation sémantique latente (ou LSI en anglais). L'objectif de LSA est de représenter les mots selon leur sens, induit par l'analyse de grands corpus de textes, et basé sur les contextes d'apparition des mots. L'hypothèse que des documents peuvent être représentés dans un espace thématique s'appuie donc sur des aspects sémantiques similaires existant entre des mots apparaissant dans des contextes similaires. L'approche peut être résumée par la réciprocité de ses hypothèses :

- deux mots proches sémantiquement apparaissent dans des contextes similaires,
- deux contextes sémantiques similaires contiennent des mots proches.

Dans un espace thématique, chaque dimension est un concept (ou thème). Chaque thème constitue une distribution de probabilités sur les mots contenus dans des documents. Cette méthode s'appuie sur une matrice *mots-documents* contenant les occurrences des mots dans les documents du corpus. Cette matrice, pouvant atteindre de grandes dimensions selon la taille du corpus, est réduite par une décomposition en valeurs singulières (SVD) afin d'en dériver la structure du modèle LSA. Toute matrice rectangulaire M, par exemple la matrice mots-documents $m * d$, peut être décomposée en un produit de trois matrices, selon l'équation $M = U.\Sigma.V'$, où U et V' sont des matrices orthonormales de gauche et droite respectivement, contenant les vecteurs singuliers, et Σ est une matrice diagonale composée des valeurs singulières. Cette méthode est étroitement liée à la décomposition en *eigenvalue-eigenvector* d'une matrice symétrique carrée. Le nombre de k dimensions souhaitées est fixé manuellement. L'approximation de la matrice initiale permet de conserver les k premières valeurs singulières de Σ, les autres valeurs étant ignorées. La matrice de valeurs singulières réduite dans l'espace de dimension k est notée Σ_k, selon $M_k = U_k.\Sigma_k.V'_k$.

De nombreuses limitations ont été observées sur ce modèle, notamment concernant la difficulté d'interprétation des dimensions (des thèmes) résultantes en langage naturel. De plus, la base probabiliste de LSA correspondant à une distribution gaussienne ne correspond pas forcément aux données d'entrées, dont la distribution est plutôt régie par une loi de Poisson. L'analyse de corpus réduit peut aussi induire des erreurs dans l'estimation des classes thématiques, car LSA ne se détache pas assez des documents composant le corpus d'apprentissage dans sa représentation sémantique. Une approche probabiliste à LSA (PLSA) a été introduite par (Hofmann, 1999) afin de palier ces problèmes, en se basant sur le principe du modèle aspect (Hofmann et al., 1999) et un algorithme tempéré d'espérance-maximisation. Cette nouvelle approche se base sur un modèle plus fiable d'un point de vue statistique que LSA et son utilisation en recherche d'information offre une alternative intéressante aux méthodes se basant sur les distances vectorielles entre requêtes et documents.

L'allocation latente de Dirichlet

Une approche plus récente proposée par (Blei et al., 2003) introduit le premier modèle de mélanges de variables latentes basé sur l'allocation de Dirichlet (LDA, pour Latent Dirichlet Allocation). D'une manière générale, LDA peut être vu comme un modèle génératif pour des documents textuels, étant représentés par un mélange de thèmes (ou *topics*) constituant les variables latentes. Un thème est caractérisé par une distribution de probabilités sur le vocabulaire présent dans les documents. Les paramètres du modèle LDA sont :

- un document d constitué de N mots w_n (avec $n \in [1..N]$),
- les paramètres de Dirichlet des K mélanges de thèmes latents : α (avec $\alpha = [\alpha_1, \alpha_2, ..., \alpha_K]$),
- la matrice contenant la distribution des probabilités conditionnelles d'observation d'un mot w sachant un thème k : β, soit $\beta_{k,w} = p(w|k)$,
- un vecteur de mélange de thème issu de la distribution de Dirichlet : θ,
- une séquence de thèmes associés à un document : $\kappa = [k_1, k_2, ..., k_N]$,

Finalement, chaque mot du corpus d'apprentissage peut être associé à un thème selon la probabilité $p(w_n|k_n, \beta)$, et la probabilité d'un document $p(d|\alpha, \beta)$ est obtenue en calculant l'intégrale sur θ de la somme des probabilités d'un mot sachant tous les thèmes possibles. La figure 2.12 permet d'illustrer l'approche LDA et montre que la distribution des thèmes découle directement des paramètres de l'allocation de Dirichlet. Cette approche est utilisée dans de nombreux travaux, notamment pour la classification automatique de documents (Phan et al., 2008), l'adaptation de modèles de langages (Tam et Schultz, 2006), ou encore l'annotation conceptuelle de texte (Camelin et al., 2011). Un aspect multilingue est développé plus récemment par Boyd-Graber et Blei (2009) et est détaillé dans la section suivante.

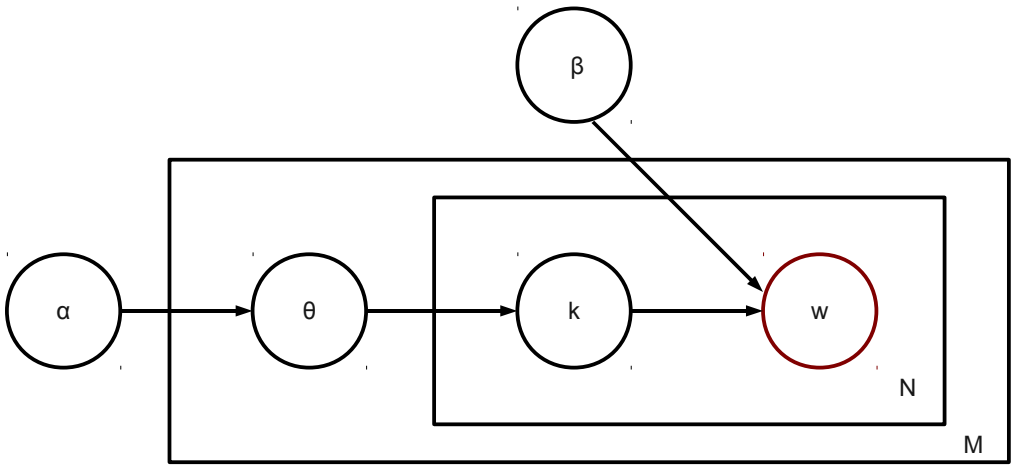


FIG. 2.12 – *Représentation graphique de LDA selon Blei et al. (2003), avec en rouge la variable observable.*

Des modèles multilingues

Suite à l'introduction de l'approche LSA, Bader et Chew (2008) proposent de modéliser des corpus parallèles dans un espace thématique pour permettre la recherche d'information inter-langue. L'idée originale a été présentée par Berry et al. (1995), et se base sur la construction d'une matrice *mots-documents* construite sur des documents anglais et français. Chaque document est la combinaison de deux textes, un dans chaque langue, l'un étant la traduction du second. Ainsi, les dimensions de l'espace thématique résultant contiennent un mélange de mots anglais et français, ayant été observés dans des documents similaires. De ce fait, une requête dans une des deux langues permet de retrouver des documents dans les deux langues grâce aux regroupements thématiques du vocabulaire bilingue. Les auteurs émettent aussi l'hypothèse que des couples de mots étant des traductions peuvent être repérés et extraits des modèles thématiques construits avec cette méthode.

D'autres travaux relatent la modélisation multilingue dans des espaces thématiques, comme l'approche géométrique basée sur PLSA introduite par Gaussier et al. (2004). Dans ces travaux, trois méthodes d'acquisition de vocabulaire bilingue sont présentées, une basée sur des vecteurs contextuels de mots à traduire (méthode présentée dans la section 2.6.3), une autre s'appuyant sur l'analyse de corrélation canonique, et enfin un modèle thématique issu de PLSA. Ces trois méthodes sont évaluées sur une collection de documents comparables provenant de la campagne d'évaluation en recherche d'information inter-langue CLEF2003. C'est finalement la méthode basée sur les contextes qui permet d'obtenir les meilleurs résultats, mais la nouvelle représentation géométrique proposée par les auteurs permet d'imaginer de nouvelles solutions pour la recherche d'information ou l'extraction de lexiques bilingues.

Un autre modèle, le modèle *HM-bitam*, introduit par Zhao et Xing (2008) et étant une extension d'un premier modèle présenté dans Zhao et Xing (2006), s'appuie sur des corpus parallèles avec un alignement au niveau des mots. Les gains possibles en appliquant un modèle bilingue basé sur LSA à la TAS fut montré par Tam et al. (2007), ce qui a par la suite incité d'autres chercheurs à s'intéresser à ce type de modélisation thématique dans un cadre multilingue. C'est suivant cette philosophie de modélisation thématique multilingue que Boyd-Graber et Blei (2009) propose d'étendre leur approche basée sur LDA à des documents comparables non alignés. Ils introduisent leur méthode sous le nom de *MuTo* (pour *Multilingual Topic*). Cette méthode permet d'extraire des relations de traductions entre des mots à partir de documents non parallèles. Afin d'évaluer leur approche, les auteurs utilisent deux corpus : un parallèle (*Europarl*) et un comparable (Wikipédia).

Dans Boyd-Graber et Blei (2009), une première étape consiste à trouver des thèmes similaires entre plusieurs langues, en se basant sur différentes méthodes pour aligner des mots et ainsi définir des points d'ancrage entre les langues. Ces méthodes concernent notamment l'utilisation de ressources parallèles (corpus ou dictionnaire multilingue aligné), la similarité morphologique (orthographique) entre des mots étant des traductions, et l'analyse de corrélation canonique multilingue (MCCA, pour Multilingual Canonical Correlation Analysis, introduit dans les travaux de Haghighi

et al. (2008)). Il est alors possible, disposant d'un ensemble de mots et de leurs traductions, de repérer des thèmes composés des même mots, mais dans plusieurs langues.

Lorsque le modèle thématique multilingue est validé selon la fiabilité des différents thèmes construits automatiquement, une seconde étape consiste à retrouver des documents étant des traductions, en fonctions de documents initiaux dans la langue source. Cette étape d'inférence sert à évaluer les capacités du modèle à établir automatiquement des relations inter-documents dans plusieurs langues. Pour les expériences sur le corpus parallèle, *MuTo* doit retrouver les traductions d'un document initial. Par exemple, une expérience sur Wikipédia permet de retrouver les articles concernant le même thème que le document initial, mais dans d'autres langues. Les auteurs présentent leur approche comme une manière de faciliter la recherche d'information inter-langues, mais aussi comme la possibilité d'aligner des traductions au niveau des mots présents dans les thèmes multilingues. Cet aspect n'a cependant pas été évalué, et est proposé dans les perspectives de ces travaux.

Modéliser des thèmes dans un cadre multilingue a aussi été étudié par Mimno et al. (2009), qui introduisent le *Polylingual Topic Model* (PLMT). Comme dans les travaux de Boyd-Graber et Blei (2009), les auteurs modélisent les corpus Europarl et Wikipédia dans des espace thématiques multilingues, en considérant 11 et 12 langues respectivement pour chaque corpus. Cependant, ces corpus sont initialement alignés au niveau des documents. Ainsi, un document est en fait un n-uplet composé des documents alignés dans les langues étudiées, les auteurs se basant sur l'hypothèse que des documents en relation de traduction - ou étant comparables - partagent des similarités thématiques. La modélisation est effectuée grâce à LDA et les classes (ou thèmes) résultantes de l'apprentissage sont constituées du vocabulaire présent dans toutes les langues présentes dans les documents.

Les auteurs évaluent leur modèle selon sa capacité à assigner des thématiques à des ensembles de documents de test. Cette inférence est effectuée à l'aide de la méthode d'échantillonnage de Gibbs. Une seconde évaluation concerne l'alignement de mots au sein des classes du modèle thématique. Les auteurs utilisent un lexique bilingue construit manuellement pour mesurer la précision dans les alignements de mots effectués. Ils obtiennent des scores avoisinant les 50% de bons alignements entre l'anglais, l'espagnol, l'italien et le français (chaque alignement concerne un couple composé d'un mot anglais avec un mot dans une autre langue). Les auteurs montrent aussi que les 12 langues présentes dans leur corpus extrait de Wikipédia constituent des sous-ensembles de documents uniformes, que ce soit au niveau du contenu, mais aussi du style éditorial. Ces aspects permettent aux auteurs de mettre en place une chaîne de traitement identique pour les 12 langues étudiées.

2.6.3 Comparabilité des contextes

Si des corpus parallèles permettent d'aligner des phrases, puis des mots, afin d'aligner des segments sous-phrastiques associés à leurs probabilités d'apparition, les

corpus non-parallèles ne contiennent pas ce type d'information. Cependant, selon la citation de Firth (1957), "un mot peut se reconnaître en fonction de son entourage lexical" [2]. Des contextes similaires entre des mots de deux langues signifie donc que ces mots peuvent être en relation de traduction. L'utilisation de corpus monolingues est alors possible, avec des repères inter-langues basés sur du vocabulaire dans les contextes des mots à traduire.

Il est toutefois assez évident que des textes abordant des sujets différents auront très peu de vocabulaire en commun, contrairement à des textes concernant le même domaine de spécialité. Il apparaît donc judicieux de constituer des corpus bilingues contenant un vocabulaire comparable dans chaque langue. C'est ce constat qui donne lieu aux études portant sur les corpus comparables, base des travaux sur la comparabilité des contextes entre des mots étant des traductions.

Les corpus comparables

Il est toujours difficile de disposer de corpus parallèles dans des domaines de spécialité et notamment pour le domaine médical. De plus, extraire du vocabulaire de spécialité à partir de textes bilingues parallèles revient à inverser le processus de traduction déjà coûteux en temps et en efforts humains. C'est une des raisons pour laquelle les corpus comparables sont des sources de données très intéressantes pour l'extraction de lexiques bilingues, car disponibles en grandes quantités (sur Internet par exemple) et dans plusieurs langues. C'est dans les travaux de (Laffling, 1992) que la notion de corpus comparable prend forme sous la définition : "des textes qui, composés indépendamment dans leur langue respective, ont le même objectif communicatif" [3].

Une définition des corpus comparables est aussi donnée par Déjean et Gaussier (2002) en se focalisant sur le lexique : "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 $l1$". Cette définition implique une notion de degré de comparabilité des corpus. Des corpus parallèles sont des corpus parfaitement comparables. A l'inverse, des corpus rédigés dans deux langues différentes, n'étant pas composés de traductions entre les langues $l1$ et $l2$, ne contiendront qu'une partie commune négligeable de leur vocabulaire étant des traductions. Si aucun mot dans la partie du corpus de la langue $l1$ n'a de traduction dans la partie du corpus de la langue $l2$, et inversement, ce qui peut paraître étonnant, les corpus sont totalement indépendants.

Les premières approches

C'est dans les années 1990 que les premiers travaux d'extraction lexicographique basée sur le contexte des mots furent publiés. Une première approche fut présentée

2. "You shall know a word by the company it keeps."
3. "which, though composed independently in the respective language communities, have the same communicative function"

par (Rapp, 1995), s'appuyant sur les motifs de co-occurrence entre deux mots dans une langue, et ceux de leurs traductions dans une autre langue. Ils mesurent la similarité inter-langue des contextes des mots, et tentent ainsi d'extraire des couples de mots étant des traductions. Les auteurs présentent un exemple sous la forme de matrices de co-occurrences permettant d'illustrer les similarités existantes entre des mots de l'anglais et de l'allemand (voir les tables 2.6 et 2.7). Ces mots constituent les lignes et les colonnes de ces matrices. L'objectif est de ré-ordonner les mots d'une matrice pour maximiser la ressemblance entre les motifs de co-occurrences.

Les résultats de ces travaux indiquent que les comptes de co-occurrences ne représentent pas réellement les liens existant entre les mots et qu'une mesure d'association doit être utilisée pour les renforcer. Afin de valider leur approche, des expériences sont menées sur 100 candidats à traduire, s'appuyant sur un corpus anglais de 33 millions de mots, et sur un corpus allemand de 46 millions de mots. La partie du corpus en anglais est tiré du *Brown Corpus*, de textes du *Wall Street Journal*, de l'encyclopédie électronique *Grolier* et de résumés scientifiques de différents domaines. Le corpus en allemand est constitué principalement d'articles de journaux tirés du *Frankfurter Rundschau*, *Die Zeit* et *Mannheimer Morgen*. Un tiers des candidats sont correctement alignés avec leur traduction, selon les deux matrices de co-occurrences construites.

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

		1	2	3	4	5	6
blau	1		•	•			
grün	2	•				•	
Himmel	3	•					
Lehrer	4						•
Pflanze	5		•				
Schule	6				•		

TAB. 2.6 – *Exemple de motifs de cooccurrences monolingues de mots anglais (à gauche) et allemands (à droite), selon (Rapp, 1995)*

		1	2	5	6	3	4	
blue	1		•	•				blau
green	2	•				•		grün
sky	5	•						Himmel
teacher	6						•	Lehrer
plant	3		•					Pflanze
school	4				•			Schule

TAB. 2.7 – *Réordonnancement de la matrice de cooccurrences de mots en anglais pour la correspondance des motifs avec les mots en allemands.*

Une autre approche est proposée par (Fung, 1995) et introduit le concept d'hétérogénéité du contexte d'un mot. L'idée consiste à mesurer le nombre de mots différents apparaissant à gauche et droite d'un mot. Cette empreinte permet de mettre

en relation des mots dans des langues différentes ayant une hétérogénéité de contexte similaire. Chaque mot peut être représenté par deux valeurs, l'hétérogénéité à gauche et à droite, notée pour un mot W :

- hétérogénéité à gauche, $x = a/c$
- hétérogénéité à droite, $y = b/c$
- a : nombre de mots précédant immédiatement W
- b : nombre de mots suivant immédiatement W
- c : nombre d'occurrences de W

A partir de corpus anglais et chinois, composés respectivement de 22 147 et 7 942 mots uniques, tirés de transcriptions de débats législatifs de Hong-Kong, les auteurs ont extrait 58 candidats à traduire. Les vecteurs à deux dimensions obtenus pour chaque mot et dans chaque langue sont comparés à l'aide de la distance euclidienne. Cette méthode permet d'atteindre un rappel de 50% dans les 10 premiers candidats retournés par le système.

Si les deux approches décrites permettent d'extraire automatiquement des couples de traductions indépendamment des langues, elles nécessitent cependant de grandes quantités de données, et les coûts liés au temps de calcul peuvent devenir importants. C'est pourquoi d'autres approches furent proposées par la suite, utilisant des ressources linguistiques bilingues afin d'avoir des points d'ancrage dans chaque langue. L'objectif est de modéliser les mots sources et cibles sous la forme de vecteurs, dont les dimensions sont comparables entre les langues. Cette méthode permet de placer des mots dans des langues différentes dans un même espace en s'appuyant sur leurs *vecteurs de contexte*.

Les vecteurs de contexte

C'est dans les travaux de Fung et McKeown (1997) que la modélisation des contextes de mots par des vecteurs est utilisée pour l'extraction de couples de traductions. Les auteurs proposent un algorithme permettant la construction de vecteurs de contexte en partant d'une liste de traductions connues (appelés *seed-words*). À partir de cette liste sw, composée de n mots langue source w_i^s accompagnés de leur traduction en langue cible w_i^t (avec $i \in [1..n]$), tous les candidats à la traduction (c_j^s ou c_k^t pour les candidats dans la langue source et cible respectivement) sont soumis à l'algorithme :

1. mesurer la corrélation du couple (c_j^s, w_i^s), $\forall\ i \in [1..n]$, pour former le vecteur de contexte v_j^s,
2. mesurer la corrélation du couple (c_k^t, w_i^t), $\forall\ i \in [1..n]$, pour former le vecteur de contexte v_k^t,
3. comparer les vecteurs v_j^s et v_k^t selon une mesure de distance vectorielle,
4. les vecteurs v_j^s et v_k^t les plus proches permettent de repérer des traductions potentielles.

Pour leurs expériences, les auteurs s'appuient sur des corpus japonais et anglais constitués d'articles de journaux traitant de la finance. Un dictionnaire bilingue extrait

de ressources disponibles sur Internet est construit et sert de pivot entre les langues. Les observations de co-occurrences entre un mot candidat et les mots du dictionnaire pivot sont limitées à une taille de contexte, c'est à dire à un nombre de mots entourant le candidat. Les auteurs proposent de faire varier le contexte d'observation en fonction de la fréquence d'apparition du mot pivot courant, afin d'éviter le bruit introduit par des mots pivots trop fréquents. Ainsi, le contexte est restreint si le mot pivot est très fréquent, et au contraire, la taille du contexte est étendue si le pivot est peu fréquent. Dans la suite de leurs travaux, les auteurs étudient l'application de la méthode par vecteur de contextes en variant la comparabilité des corpus utilisés, allant de parallèles à comparables, en passant par un corpus parallèle bruité, constitué de phrases sans limites claires et contenant des insertion et des délétions (Fung, 1998).

Rapp (1999) étend son approche précédente basée sur les matrices de co-occurrences de mots en utilisant un lexique bilingue, pouvant être enrichi au fur et à mesure de l'extraction de couples étant des traductions. Cette ressource est un pivot entre les langues (anglais et allemand) et permet de minimiser les ré-ordonnancements au sein de la matrice. La mesure de corrélation entre les mots est renforcée par l'utilisation du rapport de vraisemblance logarithmique (*log-likelihood ratio*). La matrice est composée de vecteurs, représentant les contextes des mots, dont les composantes sont les valeurs obtenues par la mesure de vraisemblance basée sur le compte des co-occurrences observées dans le corpus. Une mesure de distance vectorielle permet alors une comparaison inter-langues les vecteurs extraits des matrices. Les auteurs utilisent la distance de Manhattan (*City-block*), après avoir comparé les résultats avec d'autres distances, comme la distance de Jaccard (Romesburg, 1984), la distance euclidienne, ou encore la distance cosinus (Losee, 1998). Ils obtiennent un rappel de 72% parmi les candidats retournés en première position par leur système.

Les mesures de distance

Dans l'ensemble des travaux sur l'alignement de mots basés sur les vecteurs de contexte, les mesures de distance utilisées sont très importantes. En effet, deux étapes dans l'alignement de mots basé sur leurs contextes nécessitent l'utilisation de métriques, pour :

- mesurer les liens existant entre un mot à traduire et un mot du lexique pivot,
- estimer la distance entre des vecteurs de contexte source et cible.

Ainsi, de nombreux travaux proposent d'évaluer les possibilités en terme d'extraction lexicographique bilingue pour chaque mesure, que ce soit au niveau des co-occurrences de mots ou des distances vectorielles.

La comparaison des vecteurs de contexte se fait généralement par une mesure de la distance entre les deux vecteurs. Par exemple, la mesure classique du cosinus de l'angle formé par les deux vecteurs dans l'espace des contextes permet d'établir la proximité contextuelle. Il apparaît que cette mesure de distance vectorielle apporte de bons résultats dans la plupart des travaux. Nous privilégions donc la présentation des mesures d'associations entre mots dans cette section.

Les travaux de Evert (2004) présentent une étude détaillée des relations pouvant exister entre des mots apparaissant dans les mêmes contextes. Mesurer la valeur représentant les liens existants entre des mots permet de retrouver des couples étant en relation de co-occurrence, ou pouvant être considérés comme une paire de mots étant des collocations. Les mesures d'associations permettent de prendre en compte plusieurs paramètres, comme la fréquence d'occurrence d'un mots individuellement, ou le nombre de mots constituant l'ensemble du corpus.

L'analyse des co-occurrences de mots proposée par les auteurs est directionnelle, c'est-à-dire qu'un mot est choisi comme base, puis ses co-occurrents sont déterminés. Ces derniers sont uniquement pris en compte dans une fenêtre d'observation d'une taille limitée en nombres de mots. Ainsi, dans les fenêtres d'observations définies, quatre événements concernant la co-occurrence de deux mots m_1 et m_2 peuvent être rassemblés dans une table de contingence 2.8, où $\overline{m_1}$ et $\overline{m_2}$ représentent respectivement l'absence des mots m_1 et m_2 dans la fenêtre d'observation.

	m_1	$\overline{m_1}$	
m_2	θ_{11}	θ_{12}	$R_1 = f(m_2)$
$\overline{m_2}$	θ_{21}	θ_{22}	$R_2 = N - f(m_2)$
	$C_1 = f(m_1)$	$C_2 = N - f(m_1)$	$N = \theta_{11} + \theta_{12} + \theta_{21} + \theta_{22}$

TAB. 2.8 – *Table de contingence des co-occurrences entre deux mots selon une fenêtre d'observation.*

A partir de cette table, les auteurs détaillent quatre approches principales à la mesure de l'association entre deux mots :

- l'importance de la relation, mesurée par la probabilité d'observer cette table de contingence,
- le degré d'association, obtenu par l'estimation du maximum de vraisemblance,
- l'information mutuelle partagée par les deux mots,
- la combinaison d'heuristiques.

Si une multitude de mesures d'association peuvent être implémentées, leur impact sur l'extraction de lexiques bilingues est variable. Ainsi, une des étude les plus complète dans la comparaison de ces mesures est présentée par Laroche et Langlais (2010). Les auteurs remarquent des variations dans les résultats obtenus selon les mesures utilisées pour l'association des mots mais aussi pour la distance vectorielle. Il apparaît cependant que la mesure du *odds-ratio* normalisée (voir l'équation 2.7) combinée avec la distance cosinus permet d'obtenir les meilleurs scores en terme de rappel et de précision sur une tâche d'extraction de terminologie médicale. La mesure du rapport des chances (*odds ratio*) se situe dans la troisième approche parmi celles listées précédemment, et correspond donc au calcul de l'information mutuelle présente entre deux mots, tout comme le coefficient de Dice (Dice, 1945), ou encore la distance de Jaccard.

$$\text{odds-ratio}_{disc} = log \frac{(\theta_{11} + \frac{1}{2})(\theta_{22} + \frac{1}{2})}{(\theta_{12} + \frac{1}{2})(\theta_{21} + \frac{1}{2})} \tag{2.7}$$

2.6.4 Vecteurs de contexte et extraction terminologique

La terminologie est un ensemble de vocabulaire spécifique à un domaine de spécialité permettant à des experts de désigner des concepts propres à leur domaine. Ainsi en médecine, lorsque des nouveautés impliquent l'apparition de nouveaux concepts, de nouvelles techniques ou de nouveaux outils, le vocabulaire médical est enrichi par l'apparition de nouveaux termes. L'extraction automatique de ces termes est donc une tâche intéressante dont les enjeux sont importants dans la communauté TAL, notamment en ce qui concerne la traduction automatique de documents médicaux.

Les méthodes d'acquisition de termes médicaux et de leurs traductions à partir de corpus comparables sont généralement identiques à celles décrites dans les sections précédentes. Ainsi, une des méthodes les plus courantes consiste à utiliser la représentation vectorielle des contextes de mots, suivant la description faite dans la section 2.6.3.

Pour cette tâche, nous pouvons notamment remarquer les travaux effectués par Déjean et al. (2002) qui combinent différents modèles, dont celui des vecteurs de contexte, afin de retrouver des traductions parmi une liste de termes médicaux issu du thésaurus MeSH (*Medical Subject Headings*) [4]. Les auteurs construisent un corpus comparable spécialisé à partir de résumé anglais et allemands provenant de la ressource Medline [5]. Un traitement linguistique est effectué sur ces textes, avant d'appliquer la méthode des vecteurs de contexte, concernant principalement l'annotation des mots en classes grammaticales, la délétion des mots outils, et la lemmatisation des mots restants. Les auteurs atteignent une précision dépassant les 50% sur les premières traductions candidates retournées par leur système de combinaison. Une application de ces travaux à la recherche d'information inter-langue est présentée dans Déjean et al. (2005).

Afin de mesurer l'impact des mesures d'association entre un terme et un mot du lexique pivot, mais aussi celui des distances vectorielles, Chiao et Zweigenbaum (2002) étudient et évaluent différentes combinaisons. D'après les auteurs, utiliser une mesure d'association entre un terme et un mot du lexique permet d'obtenir des vecteurs contenant des informations plus caractéristiques sur le contexte d'un terme, par rapport à l'utilisation des fréquences de co-occurrences. Deux mesures de distance vectorielle sont testées, Jaccard et cosinus. Les deux obtiennent des résultats proches, mais le léger écart entre ces résultats favorise l'utilisation de la première mesure. Cette configuration permet d'atteindre un rappel de 33% pour 74% de précision au premier rang. Des expériences plus poussées et plus de détails en général sur ces travaux, ainsi que son application à la recherche d'information translangue, sont présentés dans Chiao (2004).

4. http ://www.ncbi.nlm.nih.gov/mesh
5. http ://www.ncbi.nlm.nih.gov/pubmed/

Il est courant de rencontrer des termes médicaux composés de plusieurs mots, comme *vésicule biliaire* (*gallbladder* en anglais) ou *anémie à cellules falciformes* (*sickle-cell anemia* en anglais). L'acquisition de ces termes particuliers peut poser problème dans un contexte multilingue, car la fertilité des termes médicaux entre le langues n'est pas toujours égale. Daille et Morin (2005) proposent d'identifier dans un premier temps ces termes multi-mots dans la langue source et cible, ainsi que leurs variations. Puis dans un second temps, les termes sources et cibles sont alignés à l'aide de la méthode des vecteurs de contexte, selon les similarités contextuelles observées. Les auteurs utilisent des ressources linguistiques sur les corpus utilisés pour leurs expériences, comme la segmentation des phrases, mais aussi l'annotation des catégories grammaticales et des lemmes pour chaque mot du corpus. Lorsque des termes sources et cibles sont composés de mulit-mots, 88% de bonnes traductions sont trouvées dans les 20 premiers résultats retournés par leur système.

Extraire des termes et leurs traductions dans le domaine médical à partir de corpus comparables, en s'appuyant sur les contextes des termes, nécessite l'utilisation d'un lexique bilingue. Ce lexique permet de disposer de points d'ancrage entre les langues. Si la taille de ce lexique influence directement la taille des vecteurs, le choix des mots présents dans le lexique est un facteur important dans la qualité de l'alignement des termes sources et cibles. Les travaux de Prochasson et al. (2009) mettent en avant cet aspect et proposent d'utiliser les termes scientifiques et les translittérations comme points d'ancrage forts dans la comparaison inter-langue des vecteurs de contexte. Cela leur permet d'atteindre 22,4% de précision au premier rang des résultats retournés par leur système pour le couple de langues français-japonais. L'influence de ces points d'ancrage est étudiée en détails dans Prochasson (2009). Morin (2007) étudie cette influence dans l'utilisation de termes complexes comme points d'ancrage inter-langues.

Les corpus comparables dans des domaines de spécialités peuvent avoir différentes tailles, ce qui implique un déséquilibre des ressources entre les langues (Morin, 2009). Si les corpus de grande taille permettent une couverture large du vocabulaire spécialisé, que ce soit en nombre de termes différents mais aussi en fréquence d'apparition de ces termes, les corpus de petite taille impliquent des fréquences d'apparition faibles pour certains termes. Ces mots *rares*, ainsi que leurs traductions, peuvent cependant être extraits automatiquement, selon les travaux présentés par Prochasson et Fung (2011). Il peut aussi être moins risqué d'utiliser des corpus comparables de grande taille mais moins spécialisés, ou partageant plusieurs domaines, comme Wikipédia par exemple, pour s'assurer d'une couverture suffisante des termes à traduire. Cette ressource encyclopédique a été utilisée pour l'acquisition de lexiques médicaux, comme le montre les travaux de Laroche et Langlais (2010), en obtenant des bons résultats en terme de rappel et de précision.

2.7 Conclusion

Nous avons présenté dans cette section la traduction automatique, ainsi que les méthodes statistiques permettant de générer des hypothèses de traduction basées

sur des corpus bilingues parallèles. L'implémentation logicielle utilisée pour nos expériences basées sur l'approche de traduction statistique par segments a été détaillée, tout comme les méthodes d'évaluation automatique de la qualité des traductions. Nous nous sommes ensuite concentrés sur l'adaptation aux domaines de spécialité en présentant les travaux visant à modifier directement les modèles statistiques des systèmes de traduction par segments.

Il apparaît cependant que l'acquisition de ressources extérieures, à partir de corpus monolingues par exemple, reste une des solutions les plus viables permettant la couverture du vocabulaire de spécialité. Les approches essayant d'utiliser d'autres ressources que les corpus parallèles nous paraissent très intéressantes, car elles permettent de contourner le problème lié à leur faible disponibilité. Il apparaît cependant que très peu de travaux font état de l'adaptation de systèmes de traduction automatique pour le domaine médical.

De plus, la construction de systèmes de traduction dédiés à un domaine en particulier est coûteux du point de vue des ressources à utiliser, car ces dernières demandent l'intervention de traducteurs maîtrisant les aspects terminologiques du domaine concerné. Il nous parait donc intéressant d'étudier les différentes possibilités permettant d'adapter, à moindre coût, un système de traduction automatique à un domaine spécialisé. Nous proposons donc des méthodes automatiques impliquant le moins possible l'intervention de spécialistes et de traducteurs humains.

Dans les travaux décrits dans cette thèse, la tâche principale est d'adapter des systèmes de traduction automatique *état-de-l'art* au domaine médical. Si ce domaine est directement impliqué par le contexte du projet *ANR* dans lequel s'inscrit cette thèse, nous pensons toutefois qu'il est intéressant de proposer des approches *génériques* pouvant être appliquée à tout domaine de spécialité. Nous proposons deux axes majeurs de recherche dans cette thèse :

- l'acquisition automatique de vocabulaire spécialisé pour la construction de lexiques bilingues,
- la post-édition automatique des hypothèses de traduction produites par un système.

Ces deux axes permettent de séparer le problème d'adaptation aux domaines de spécialité en deux sous-parties, l'une présentant des aspects terminologiques, l'autre orientée vers la réécriture de traductions automatiques. Ainsi, nous pensons pouvoir répondre au problème concernant les mots hors vocabulaire, mais aussi à la correction d'erreurs syntaxiques et grammaticales induites par un système de traduction générique, ou hors domaine, lorsque des textes spécialisés sont à traduire.

Chapitre 3

La post-édition automatique de traductions

Sommaire

Dans ce chapitre, nous présentons une nouvelle approche pour l'adaptation aux domaines de spécialité en traduction automatique. Nous proposons une méthode de post-édition automatique basée sur un système de traduction statistique par segments sous-phrastiques (approche présentée dans la section 2.3.2). A partir d'hypothèses émises par des systèmes de traduction automatique génériques, un système de post-édition statistique (SPE) permet de les modifier afin de les adapter à un domaine de spécialité.

L'objectif principal est de mettre en relation des hypothèses de traduction produites par un système de traduction automatique avec leurs traductions de référence. De ce fait, nous voulons capturer les erreurs commises par le système de traduction en comparant les hypothèses et les références. Ainsi, nous voulons nous affranchir d'une étape de post-édition manuelle, en nous basant sur des approches statistiques appliquées en série. En partant du principe qu'une *petite* quantité de données spécifiques à un domaine est disponible, relativement à la quantité disponible hors-domaine, nous voulons mesurer l'impact de notre approche selon plusieurs configurations.

Dans un premier temps, nous appliquons la post-édition statistique sur des traductions de brèves journalistiques, dans le cadre de la campagne d'évaluation en traduction automatique *WMT11*. Cette première série d'expériences est présentée dans la section 3.1, et permet de comparer les résultats entre des systèmes construits sur des ensembles de données différents. Dans un second temps, nous voulons évaluer les gains possibles liés à la post-édition statistique de traductions dans le domaine médical, dont les expériences sont détaillées dans la section 3.2. Cette seconde série d'expériences nous permet de comparer les résultats entre des systèmes figés, construits sur des données génériques ou spécialisées, et des systèmes adaptés à des genres de documents particuliers par l'ajout en cascade d'un système de post-édition statistique.

3.1 Expériences préliminaires

Dans cette section, nous présentons les expériences préliminaires que nous avons menées en post-édition automatique de traductions lors de la campagne d'évaluation *WMT11* (Workshop on Machine Translation 2011). Pour la première fois, le Laboratoire Informatique d'Avignon (LIA) a proposé un système de traduction automatique pour la tâche de traduction du français vers l'anglais, pour un genre de documents particulier : les brèves journalistiques.

Les brèves journalistiques concernent plusieurs domaines, comme la politique, l'économie, ou encore les nouvelles technologies. De ce fait, la terminologie de chaque domaine est incluse dans les documents à traduire. Aussi, la syntaxe utilisée dans les brèves est particulière : nombreuses citations, phrases courtes, style télégraphique, etc. Le vocabulaire contient aussi des entités nommées, comme des noms de personnes ou de marques par exemple.

Le système final, soumis à l'évaluation pendant la campagne *WMT11*, est constitué

de la combinaison des systèmes de traduction du LIA et du Laboratoire Informatique de Grenoble (LIG). Nous ne présentons cependant que la partie relative au système du LIA dans la prochaine sous-section. Puis, dans la sous-section suivante, nous détaillons les expérimentations menées en post-édition de traductions issues du système du LIA.

3.1.1 Le système du LIA pour WMT11

Le système de traduction automatique proposé par le LIA est basé sur la boîte à outils libre *Moses* [1] (détaillée dans la section 2.3.4), permettant de construire un système de traduction automatique statistique par segments. Les corpus mis à disposition des participants pour l'apprentissage des modèles (de traduction et de langage) sont présentés dans le tableau 3.1. La taille des corpus est exprimée en nombre de phrases, avant tout pré-traitement. Ces pré-traitements concernent principalement le retrait des phrases trop longues, généralement celles supérieures à 80 mots, ou encore les phrases composées uniquement, ou majoritairement, de caractères non-alphanumériques.

Parmi ces corpus, la plus grande quantité de données correspond aux débats parlementaires européens et des débats de l'Assemblée Générale des Nations Unies. Ces données sont considérées hors-domaine, car elles ne sont pas composées de brèves journalistiques. Elles sont composées de débats et de décrets. Un plus petit corpus, concernant le domaine des actualités, est mis à disposition pour la tâche de traduction. Ce corpus appelé *News Commentary*, est tiré du site Internet *Project Syndicate* [2], et est composé d'articles rédigés par des experts commentant l'actualité économique, politique et scientifique. Certains contributeurs sont des personnalités reconnues, comme Tony Blair ou Christine Lagarde.

Après avoir effectué des pré-traitements sur les corpus d'apprentissage, un modèle de langage 5-grammes anglais est construit, lissé selon la méthode Kneser-Ney (Kneser et Ney, 1995). Le corpus parallèle de brèves journalistiques (que nous qualifions par extension de corpus dans un domaine spécialisé) est ensuite utilisé afin de collecter des paires de phrases proches de ce genre de documents à partir des corpus parallèles hors-domaine. Cette méthode, appelée sous-échantillonnage, et présentée dans la section 2.4.3, nous permet d'accroître la quantité de données parallèle se rapprochant des brèves journalistiques. Plusieurs expériences sont menées suivant la quantité de données extraites des corpus parallèles hors domaine.

Le tableau 3.2 contient l'ensemble des résultats obtenus par le système de traduction du LIA après les pré-traitements classiques effectués sur les données : les phrases trop courtes ou trop longues sont enlevées, une analyse lexicale (*tokenization*) est faite sur les corpus, etc. Des expériences sont détaillées par Potet et al. (2011b), visant à évaluer d'autres critères, comme l'élagage du modèle de traduction ou l'ajout de balises autour des signes de ponctuation. Ces méthodes ne sont finalement pas intégrées au système soumis pour évaluation, car les gains restent trop marginaux en terme de score BLEU (la métrique automatique d'évaluation de traductions, détaillée dans la section 2.3.5).

1. http ://www.statmt.org/moses
2. http ://www.project-syndicate.org/

CORPUS	LABEL	TAILLE (PHRASES)
Apprentissage bilingue Anglais-Français		
News Commentary v6	*news-c*	116 k
Europarl v6	*euro*	1.8 M
United Nation corpus	*UN*	12 M
10^9 corpus	*giga*	23 M
Apprentissage monolingue Anglais		
News Commentary v6	*mono-news-c*	181 k
Shuffled News Crawl corpus (de 2007 à 2011)	*news-s*	25 M
Développement		
newstest2008	*test08*	2 051
newstest2009	*test09*	2 525
Test		
newstest2010	*test10*	2 489
newstest2011	*test11*	3 003

TAB. *3.1* – *Ressources utilisées par le LIA lors de la campagne d'évaluation en traduction automatique WMT11*

Parmi ces expériences non incluses dans la soumission finale, nous retenons tout de même la post-édition automatique que nous évaluons sur les données de la campagne d'évaluation. Le filtrage de la table de traduction (Johnson et al., 2007) n'a pas permis d'améliorer les scores BLEU, comme l'indique la colonne *Filtrage* du tableau présentant les résultats du LIA.

CORPUS PARALLÈLE	FILTRAGE	
	AVEC	SANS
news-c + *euro* (1.77 M)	28,1	28,0
news-c + 1.77 M de *UN*	27,2	-
news-c + 1.77 M de *giga*	27,1	-
news-c + 1.77 M avec RI	28,2	-
news-c + 3 M avec RI	29,1	29,0
news-c + 5 M avec RI	28,8	-
***news-c* + 10 M avec RI**	**29,3**	29,2
Toutes les données	28,9	29,0

TAB. *3.2* – *Scores BLEU (%) sur test*10 *obtenus avec le système du LIA en utilisant différents corpus bilingues pour l'apprentissage. Les meilleurs résultats sont mis en gras.*

Les résultats obtenus par le système du LIA indiquent que la recherche d'information (notée RI) dans les corpus parallèles hors-domaine permet d'atteindre les meilleurs résultats. Les moins bons scores BLEU sont obtenus par la combinaison du corpus de brèves (*news-c*) et du corpus hors-domaine noté *giga*. La quantité de phrases extraites de ce corpus a été fixée à 1,77 M afin de correspondre à la taille du corpus *euro* servant de repère pour évaluer les gains obtenus. En collectant des phrases à partir des corpus hors-domaine (tous les corpus parallèles bilingues sauf *news-c*) selon leur

proximité avec les phrases du corpus *news-c*, le score BLEU est similaire à celui obtenu avec la combinaison *news-c* et *euro* avec 1,77 M de phrases. Lorsque 3 M de phrases sont collectées par cette méthode, le score BLEU est amélioré d'un point. C'est avec 10 M de phrases collectées depuis les corpus hors-domaines que les meilleurs scores sont obtenus.

3.1.2 La post-édition statistique

Lors de la campagne d'évaluation WMT11, nous avons mené une série d'expérimentations visant à évaluer les possibilités de corriger les erreurs contenues dans les hypothèses de traduction produites par le système du LIA. Le tableau 3.3 contient des sorties de ce système et les traductions de référence, permettant d'illustrer les erreurs rencontrées dans les hypothèses produites par le système. Ces erreurs peuvent être récurrentes ou isolées, et nous supposons qu'il est possible de les capturer et de les corriger *a posteriori*.

SORTIES LIA	RÉFÉRENCES
barack obama will be the fourth u.s. president to receive the nobel peace prize	barack obama becomes the fourth american president to receive the nobel peace prize
your next smartphone operating systems will master two	your next smartphone will run two operating systems
nobody wants to spend the winter in the mountain without snow .	nobody wants to spend the end of the year in the mountains when there 's no snow .
in previous cases , the engines failed on the first floor .	previously , the engines have always malfunctioned during the first phase .

TAB. 3.3 – *Exemples de sorties issues du système PBMT du LIA, comparées à leurs traductions de référence.*

Pour cela, nous proposons une seconde étape, la post-édition, en série de la première étape de traduction automatique, en mettant en relation les hypothèses de traduction produites par le système de traduction statistique par segments avec leur référence de traduction. Nous constituons ainsi un corpus parallèle dans la langue cible, aligné au niveau des phrases, et permettant l'apprentissage d'un modèle de traduction monolingue, pouvant être considéré comme un système de réécriture. De cette manière, nous voulons essayer de rapprocher les hypothèses de traduction, produites par un système automatique, avec leurs traductions de référence.

Nous disposons d'un ensemble de données restreint dans le domaine de spécialité, un ensemble d'apprentissage et un de développement (*tuning*). Pour respecter les contraintes de la campagne d'évaluation concernant les données utilisées, nous décidons de construire le modèle de post-édition dans la langue cible à partir du corpus *news-c* traduit par le système du LIA. Le corpus obtenu, noté *news-c_liatrad*, est aligné avec la référence dans la langue cible du corpus *news-c* avec l'outil GIZA++ (présenté dans la section 2.3.4).

Le modèle de post-édition est construit en utilisant les mêmes paramètres que le modèle de traduction issu de l'apprentissage décrit dans la section 3.1.1. Ce modèle contient un ensemble de couples de segments sous-phrastiques, basées sur les observations des différences entre une sortie du système de traduction du LIA et la traduction de référence du corpus initial. Le principe général est de capturer les erreurs commises, et de leur associer des probabilités calculées selon leurs fréquences d'apparition dans le corpus *news-c_liatrad*. Les erreurs les plus fréquentes auront des probabilités fortes, contrairement aux erreurs les plus rares. Les poids associés à chaque paramètre du modèle de post-édition (voir section 2.3.3) sont optimisés selon l'algorithme MERT, en prenant comme corpus d'entrée la dernière itération du développement du système présenté en section 3.1.1.

Les corpus de test *test10* et *test11* sont tout d'abord traduits par le système français-anglais du LIA, avant d'être fournis au système de post-édition anglais. Appliquer l'ensemble du modèle de post-édition à l'ensemble des hypothèses émises par le système du LIA nous parait risqué, mais nous voulons en mesurer les effets et nous en servir comme base afin d'évaluer l'impact de la post-édition statistique. Nous sommes conscients du fait qu'appliquer la post-édition statistique peut corriger des erreurs issues de la traduction, mais aussi introduire de nouvelles erreurs. Nous estimons donc que seulement certaines phrases ont besoin d'être post-éditée.

Afin d'évaluer les gains apportés par la post-édition statistique sur chaque phrase du corpus de test, nous utilisons la métrique automatique basée sur le taux d'erreur de traduction (*TER*, détaillée dans la section 2.3.5). Ainsi, chaque traduction est évaluée avant et après l'étape de post-édition. De ce fait, nous pouvons calculer un score *oracle* sur l'ensemble du corpus de test, en post-éditant uniquement les phrases permettant un gain en TER, et en gardant les phrases produites par le système du LIA le cas échéant.

Pour faire cette évaluation, nous utilisons la métrique automatique BLEU (voir description dans la section 2.3.5) avant et après l'étape de post-édition. L'outil utilisé implémentant BLEU est *multi-bleu*. Les corpus utilisés pour l'apprentissage du modèle de traduction du système du LIA peuvent être les même que ceux utilisés pour la construction du modèle de post-édition. Cependant, pour mesurer les effets de notre approche en terme de score BLEU, nous considérons deux configurations : une ne contenant pas de données spécialisées (notée *euro + UN*) et une autre contenant en plus le corpus *news-c*, ce dernier est alors présent dans les deux phases d'apprentissage (traduction et post-édition). Les résultats obtenus en post-édition sont comparés avec ceux mesurés en sortie du système du LIA, et présentés dans le tableau 3.4.

	test10			*test11*		
	LIA	*LIA + SPE*	*LIA + SPE$_{oracle}$*	*LIA*	*LIA + SPE*	*LIA + SPE$_{oracle}$*
euro + UN	28,4	28,2	28,9	27,9	27,8	28,5
news-c + euro + UN	28,3	28,3	28,5	28,1	28,1	28,4

TAB. 3.4 – *Scores BLEU (%) obtenus avec le système du LIA et en post-édition automatique pour l'adaptation au domaine journalistique.*

Ces résultats sont moins bons que ceux obtenus par le LIA, présentés dans le tableau 3.2, en raison de la quantité de données du domaine utilisée. En effet, nous privilégions dans cette approche la correction d'erreurs *a posteriori*, contrairement aux méthodes de sous-échantillonnage permettant l'acquisition de corpus parallèles se rapprochant du domaine. Ainsi, nous réduisons la quantité de données utilisée, donc le coût impliquée par la collecte de corpus liée à la recherche d'information, et le temps de calcul nécessaire à la construction du modèle de traduction.

Nous observons cependant plusieurs phénomènes dans les résultats de post-édition :

- de manière générale, le score BLEU n'est pas amélioré, et le plus souvent dégradé, par la post-édition statistique,
- il existe un gain potentiel lorsque l'on sélectionne les phrases à post-éditer, comme l'indique les scores *oracle*,
- les gains en oracles ne sont pas significatif lorsque le corpus de spécialité est utilisé pour l'apprentissage, selon l'intervalle de confiance accordé à la métrique BLEU (Koehn, 2004).

Si ces premiers résultats ne sont pas vraiment significatifs en terme d'amélioration de la qualité des hypothèses de traductions, nous pensons que traduire des brèves journalistiques de manière efficace requiert une masse de données très importante. En effet, une multitude de sujets sont abordés (économie, droit, politique, etc.) et cet aspect nous incite à penser que la quantité limitée de données spécialisées peut apporter de meilleurs résultats dans d'autres domaines, plus restreints. Nous voulons donc poursuivre les expérimentations pour le domaine médical dans la prochaine section, en augmentant légèrement la quantité de données parallèles, et en utilisant plusieurs configurations pour construire les systèmes de traduction par segments.

3.2 La post-édition pour l'adaptation au domaine médical

Dans cette section, nous appliquons la chaîne de traitement automatique constituée d'un système de traduction et d'un système de post-édition, les deux suivant l'approche PBMT, pour l'adaptation au domaine spécialisé de la médecine. Nous voulons vérifier si reformuler des hypothèses *a posteriori* permet d'améliorer la qualité générale des traductions dans le domaine médical, si le système initial contient peu ou pas de données du domaine. Nous voulons aussi évaluer les gains possibles de l'approche statistique basée sur l'alignement de segments sous-phrastiques dans le cadre de la post-édition statistique.

Nous considérons deux systèmes de traduction PBMT, un basé sur *Moses* dont les configurations pour l'apprentissage et le décodage sont identiques à la description faite dans la section 3.1.1, et un système propriétaire, dans sa version gratuite accessible en ligne, développé par Google [3]. Ce système de traduction statistique est utilisé comme *boîte noire* car nous n'intervenons pas dans les phases d'apprentissage et de décodage.

3. http ://translate.google.fr

Nous notons $PBMT$ le système de traduction initial, *com* le système commercial, et SPE le système de post-édition statistique, placé en aval d'un système de traduction.

3.2.1 Cadre expérimental

Afin de mesurer l'impact de notre approche, nous constituons trois sous-ensembles d'expériences. Le premier est constitué d'expériences identiques aux travaux préliminaires en post-édition statistique. Le second est centré sur les méthodes de sélection des phrases à post-éditer. Le troisième permet d'élaguer la table de post-édition. L'évaluation effectuée pour chaque série d'expériences s'appuie sur un corpus parallèle médical utilisé comme test, et la métrique BLEU permettant de comparer les hypothèses issues des systèmes avec leur traduction de référence.

Traduction du corpus de spécialité

Cette série d'expériences est préliminaire à la post-édition et concerne la traduction d'un corpus de test du Français vers l'Anglais. Les modèles statistiques peuvent être entraînés sur les données spécialisées et sur les données génériques. Différentes configurations sont possibles suivant les combinaisons de données effectuées. En utilisant une métrique d'évaluation automatique, les résultats en terme de score BLEU forment notre système de base (*baseline*), nous permettant ainsi de mesurer les gains possibles résultant des séries d'expériences en post-édition.

Afin d'évaluer notre approche de la manière la plus complète et de réduire les coûts de traitements nécessaires à nos expériences, nous décidons d'utiliser un second système *baseline*, disponible en ligne gratuitement, en plus du système PBMT état-de-l'art. C'est le système noté *com*, permettant de traduire automatiquement le corpus de test de la langue source vers la langue cible, sans pouvoir intervenir sur les paramètres du modèle de traduction ou du décodeur.

Post-édition naïve et oracle

La première phase d'évaluation de notre approche en post-édition s'effectue sur la mise *en série* de la post-édition statistique après l'étape préliminaire de traduction. Le corpus du domaine médical utilisé pour l'entraînement est parfaitement adapté à la construction d'un modèle de post-édition, comme décrit dans la section 3.1.2. L'apprentissage du modèle de post-édition se fait suivant les même paramètres intrinsèques au système PBMT utilisé pour la phase de traduction. L'évaluation se fait grâce à un corpus parallèle de test, qui est traduit de la langue source vers la langue cible par le premier système, puis post-édité dans la langue cible par le second système. Les hypothèses issues du second système sont comparées à leurs traductions de référence dans le corpus de test.

Les scores ainsi obtenus résultent de l'application *naïve* de la post-édition sur un ensemble de phrases composant un corpus de test. De ce fait, nous ne savons pas si cela va dégrader ou améliorer la qualité des hypothèses de traduction. Nous séparons donc les phrases post-éditées en deux ensembles : celles étant améliorées par la post-édition et celles étant dégradées. Nous calculons alors un score oracle d'une manière identique à la méthode utilisée dans la section 3.1.2, en gardant uniquement les traductions améliorées par la post-édition, les autres provenant du système de traduction.

Afin d'évaluer les gains apportés par la post-édition statistique au niveau du corpus et au niveau des phrases individuellement, nous utilisons la métrique BLEU. Chaque phrase de test a donc deux scores BLEU, le premier obtenu après traduction et le second après post-édition. Les scores BLEU sont donc utilisés au niveau local sur chaque phrase, mais aussi au niveau global sur le corpus de test.

Selon les résultats préliminaires obtenus lors des expériences en post-édition statistique sur les brèves journalistiques, nous remarquons que les effets de la post-édition sont très variables. Par rapport à la phase de traduction de la langue source vers la langue cible, certaines hypothèses de traduction sont dégradées par la post-édition statistique, alors que d'autres sont améliorées. Il nous semble donc important de pouvoir sélectionner les hypothèses de traduction à post-éditer, afin de se rapprocher au maximum du score oracle mesuré. Nous décrivons ce processus de sélection de traductions dans la sous-section suivante.

Sélection des hypothèses de traduction à post-éditer

Dans une seconde phase d'évaluation, nous nous concentrons sur la sélection automatique des hypothèses de traduction à post-éditer. Nous voulons appliquer la post-édition uniquement sur les traductions dont la qualité peut être améliorée. De manière générale, nous pensons qu'en agissant au niveau de chaque phrase pour corriger des erreurs, nous pouvons améliorer le résultat de notre approche en terme de score BLEU sur l'ensemble du corpus de test. Pour cela, nous devons construire un modèle permettant de classer des phrases de test selon les gains possibles après post-édition.

Le corpus utilisé pour l'apprentissage du modèle de post-édition sert de base pour la construction du modèle de classification. Chaque phrase de ce corpus est étiquetée selon que son score BLEU se dégrade ou s'améliore suite à la post-édition. Nous définissons un score $\Delta BLEU$ selon : $\Delta BLEU = BLEU_{SPE} - BLEU_{PBMT}$. L'architecture permettant d'obtenir ce score est détaillée par la figure 3.1.

Un classifieur de type Séparateur à Vaste Marge (*SVM*) (Boser et al., 1992) nous permet de prendre en charge des vecteurs de séquences n-gramme issus des phrases d'apprentissage. Nous fixons la taille maximum des séquences à 3 mots, et choisissons le noyau linéaire pour le classifieur. Nous voulons ainsi définir deux ensembles linéairement séparables de phrases, suivant les classes auxquelles elles appartiennent. Les deux classes que nous utilisons peuvent être décrites de la sorte :

– $\Delta BLEU > 0$: la phrase est améliorée par SPE,

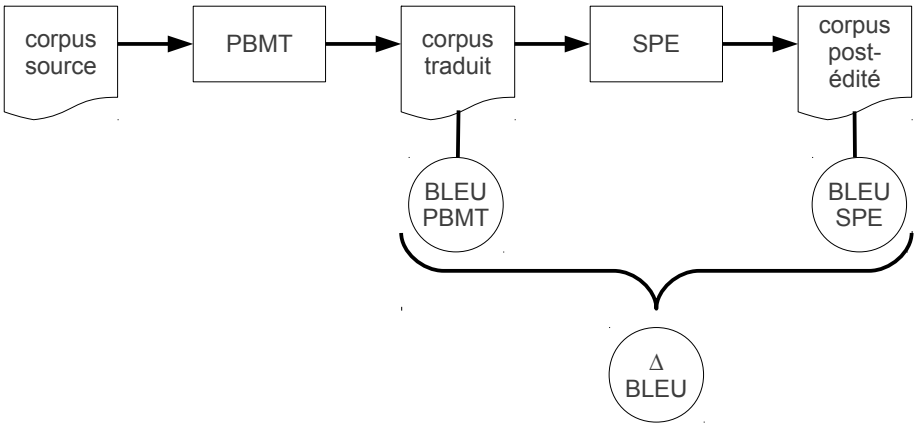


FIG. 3.1 – *Architecture de notre système permettant de mesurer les scores $\Delta BLEU$ pour chaque phrase d'un corpus dont nous disposons de la traduction de référence.*

– $\Delta BLEU \leq 0$: la phrase n'est pas améliorée par SPE.

Ainsi, pour une phrase de test, le classifieur peut inférer la classe dont elle fait parti en s'appuyant sur son modèle, et décider de post-éditer ou non cette phrase.

Élagage de la table de post-édition

Lors de la construction du modèle de post-édition, nous nous trouvons dans un contexte monolingue. Ainsi, un ensemble de segments différents ou identiques sont alignés pour former la table de post-édition. Si les segments sont identiques, le système de traduction mis en place dans la phase préliminaire a produit des hypothèses identiques aux traductions de référence. Si les segments sont différents cela signifie que des erreurs ont été commises par le système de traduction, et c'est dans ce cas que nous voulons intervenir avec la post-édition. Cependant, il peut se produire un événement entraînant la dégradation des résultats de post-édition. Un segment issu de la traduction peut être traduit correctement un nombre réduit de fois, et être mal traduit un plus grand nombre de fois.

L'élagage de la table de post-édition concerne donc une troisième phase d'évaluation. Nous voulons mesurer l'impact de chaque paire de segments de la table du système de post-édition en terme de gain BLEU. Les paires n'apportant aucune amélioration du score BLEU sont retirées de la table. Afin de mesurer l'impact de chaque paire de segments de la table de post-édition statistique, nous utilisons des corpus dont nous disposons de la traduction de référence, c'est à dire le corpus d'apprentissage et celui dédié au développement, les deux se concernant le domaine de spécialité. L'évaluation est finalement effectuée grâce au corpus de test, en utilisant la métrique automatique BLEU.

3.2.2 Les ressources

Le corpus parallèle générique

Afin de construire les modèles de traduction pour les systèmes PBMT et de fournir des hypothèses à post-éditer, nous avons besoin de corpus bilingues parallèles. En ce qui concerne les données hors domaine de spécialité, nous considérons deux corpus : celui issu des sessions du Parlement Européen (*Europarl*) et celui provenant des Nations Unies (*UN*). Les détails de ces deux corpus sont donnés dans le tableau 3.1. Pour les données du domaine médical, nous utilisons le corpus extrait de documents PDF publiés par l'Agence Européenne de Médecines (*EMEA* (Tiedemann, 2009)).

Le corpus parallèle médical

Nous considérons le corpus médical *EMEA* comme une ressource particulièrement intéressante, car elle contient du texte issu de documents pouvant aborder divers sujet dans le domaine de la médecine. De plus, ces documents sont construits de plusieurs manières. Par exemple, un document traitant de l'insuline peut être rédigé comme une prescription médicale (posologie), alors qu'un autre peut concerner la description de symptômes, et donc être rédigé avec un style plus littéraire. Le tableau 3.5 contient des phrases extraites du corpus parallèle médical.

PHRASES SOURCES	PHRASES CIBLES
Le principe actif d'Abilify est l'aripiprazole, un médicament de la famille des neuroleptiques.	The active substance in Abilify, aripiprazole, is an antipsychotic medicine.
Un repas riche en graisses n'a pas d'effet sur la pharmacocinétique de l'aripiprazole.	There is no effect of a high fat meal on the pharmacokinetics of aripiprazole.
La posologie initiale recommandée pour Abilify est de 10 ou 15 mg/jour.	The recommended starting dose for Abilify is 10 or 15 mg/day.
Le potentiel cancérogène du paclitaxel n'a pas été étudié.	The carcinogenic potential of paclitaxel has not been studied.

TAB. 3.5 – *Extrait du corpus médical utilisé lors de nos expériences sur la post-édition de traductions.*

Nous avons observé des redondances importantes au niveau des phrases dans ce corpus spécialisé. Il nous paraît intéressant de constituer différents corpus afin de mener des expériences selon des situations particulières. Dans un premier temps, nous nettoyons le corpus afin d'en enlever les phrases trop courtes (inférieures à 5 mots) et trop longues (supérieures à 80 mots). Nous retirons ensuite les phrases dont le contenu est uniquement numérique et celle contenant des caractères non-alphanumériques.

Nous constituons alors trois configurations à partir de ce corpus. Le tableau 3.6 contient les détails des ressources nécessaires à nos expérimentations sur la post-édition. Le corpus EMEA disponible en ligne [4] est constitué initialement de 1,1 M de paires de phrases français-anglais. Après les pré-traitements effectués, le corpus contient 390 k paires de phrases.

LABEL	TAILLE (PHRASES)
Apprentissage	
$c1Emea_{train}$	390 k
$c2Emea_{train}$	156 k
$c3Emea_{train}$	156 k
Développement/optimisation	
$c1Emea_{dev}$	
$c2Emea_{dev}$	2 k
$c3Emea_{dev}$	
Test	
$c1Emea_{test}$	
$c2Emea_{test}$	2 k
$c3Emea_{test}$	

TAB. 3.6 – *Détails des ressources utilisées et leurs découpages pour les expériences en post-édition statistique.*

La première configuration ($config1$, noté $c1$ pour préfixer le label du corpus utilisé) consiste à utiliser l'ensemble des données spécialisées, sans effectuer des traitements spécifiques. La seconde ($config2$ ou $c2$ comme préfixe) est basée sur un corpus exempt de toute redondance de phrases, et les corpus de développement et de test sont extraits aléatoirement parmi l'ensemble du corpus médical. La troisième configuration ($config3$ ou $c3$ comme préfixe) est aussi basée sur les données spécialisées exemptes de redondance de phrases, et les corpus de développement et de test sont extraits selon la chronologie d'apparition des documents permettant de construire le corpus. Cette troisième configuration nous parait la plus adaptée, car elle permet d'évaluer l'approche de post-édition statistique de manière plus générale, sans prendre en compte les caractéristiques intrinsèques du corpus médical utilisé. Les trois configurations nous permettent donc de juger de la pertinence de la post-édition statistique dans des cas plus ou moins favorables.

Pour les configurations $config2$ et $config3$, le corpus d'entraînement du modèle de traduction ne contient aucune phrase des corpus de développement et de tests. La configuration $config1$ contient des phrases redondantes parmi les ensembles d'apprentissage, de développement et de test. Cependant, pour les trois configurations, il peut arriver que du vocabulaire se trouvant dans le corpus d'apprentissage soit aussi dans le corpus de test, suivant la chronologie d'apparition des documents constituant le corpus dans sa globalité. La dernière configuration tend à éviter ce genre de phénomène. Afin de disposer d'un corpus de développement de taille raisonnable,

4. http ://opus.lingfil.uu.se/EMEA.php

nous décidons de garder 2 k paires pour optimiser les systèmes de traduction et de post-édition. Ainsi, notre corpus de développement correspond environ à ceux utilisés pendant les campagnes d'évaluation WMT en nombre de paires de phrases.

Le modèle de traduction

Notre méthode pour séparer les données du corpus spécialisé en trois sous-ensembles, l'apprentissage, le développement et le test, donne lieu aux différents systèmes, dont nous rassemblons les détails dans le tableau 3.7. Nous désignons les systèmes basés sur *Moses* avec le label *PBMT*, avec plusieurs variantes selon les corpus d'apprentissage utilisé. Le label *LM* est utilisé pour désigner le type de modèle de langage associé au modèle de traduction, pouvant être construit sur les données d'apprentissage génériques ou médicales.

CORPUS PARALLÈLE	LABEL
Système commercial	
-	*com*
Systèmes PBMT domaine unique	
euro + *UN*	$PBMT_{gen}LM_{gen}$
EMEA	$PBMT_{med}LM_{med}$
Systèmes PBMT domaines combinés	
euro + *UN*	$PBMT_{gen}LM_{med}$
euro + *UN*	$PBMT_{gen}LM_{gen+med}$
euro + *UN* + EMEA	$PBMT_{gen+med}LM_{med}$
euro + *UN* + EMEA	$PBMT_{gen+med}LM_{gen+med}$

TAB. 3.7 – *Détails des systèmes utilisés pour les expériences en post-édition statistique. La taille des données médicales varie selon les configurations présentées dans le tableau 3.6*

La combinaison de modèles de traduction se fait grâce à l'option présente dans *Moses* permettant d'interpoler plusieurs modèles. La méthode utilisée dans nos travaux consiste à utiliser en priorité le modèle de traduction spécialisé, et se replier sur le modèle générique si aucune paire de segments n'est disponible dans le premier modèle. En ce qui concerne les modèles de langages, nous avons la possibilité d'interpoler un modèle spécialisé avec un modèle générique grâce à l'outil *SRILM*. Cette interpolation s'effectue grâce à un facteur λ, estimé selon la perplexité des modèles sur des données d'apprentissage et de développement.

Le modèle de post-édition

Pour chaque système présenté dans le tableau 3.7, nous appliquons une étape de post-édition basée sur le corpus $Emea_{train}$. Pour chaque configuration, le corpus d'apprentissage est traduit du Français vers l'Anglais, puis est utilisé pour la construction du modèle de post-édition. De plus, chacun de ces modèles peut être

combiné à un des modèles de langage, soit médical, générique, ou la combinaison des deux, selon la liste donnée dans le tableau 3.7.

L'optimisation des poids du modèle de traduction est effectuée par l'algorithme MERT, présenté dans la section 2.3.3. Le corpus d'entrée est celui issu de la dernière itération du développement du système de traduction, aligné avec sa traduction de référence. L'étape de développement du système de post-édition est donc toujours effectuée sur les données spécialisés, comme présenté dans le tableau 3.6.

Le modèle de langage

Pour mener à bien nos expériences, nous utilisons deux types de modèles de langage : un générique et un spécialisé. Pour le modèle de langage générique, nous utilisons les données monolingues en langue cible mises à disposition pour WMT11 détaillées dans le tableau 3.1. Pour le modèle de langage spécialisé, nous utilisons les corpus d'entraînements et de développements relatifs à chaque configuration décrite dans le paragraphe précédent. Trois modèles de langage sont donc construit pour le domaine de spécialité, et trois autres pour l'interpolation des modèles (générique et spécialisé). Nous utilisons l'outil du SRILM pour construire ces modèles ou pour les interpoler. La construction des modèles de langage est effectuée selon les mêmes paramètres que ceux utilisés pour lors de la campagne d'évaluation présentée dans la section 3.1.1.

Lorsque les données génériques sont utilisées, nous appliquons des seuils afin de ne prendre en considération que les séquences (*n-gram*) avec une fréquence d'apparition dans le corpus supérieure à 2. Pour la construction du modèle médical, aucun seuil n'est fixé afin de garder l'ensemble du vocabulaire spécialisé, même les séquences les moins fréquentes. La méthode de seuillage est principalement utilisée afin de réduire la quantité de données à modéliser, mais elle permet aussi parfois de réduire la perplexité sur des données de test.

Pour la combinaison des modèles, nous étudions deux possibilités : la combinaison linéaire des modèles ou la construction d'un modèle sur les données génériques et médicales concaténées. La combinaison linéaire des modèles nécessite l'estimation d'un coefficient α de combinaison. Ce coefficient est estimé selon les scores de perplexité obtenus par les modèles de langage, générique ou spécialisé, sur les données de développement dans le domaine de spécialité, selon $config3$. La concaténation des données permet de construire un modèle de langage contenant les données génériques et spécialisées. Nous construisons deux modèles à partir des données concaténées : un contenant tout le vocabulaire présent dans les corpus (noté $LM_{gen+med}$), un autre constitué d'un vocabulaire générique limité à 1 million de mots (noté $LM_{gen+med}L$). Ce vocabulaire est constitué des mots les plus fréquents parmi les corpus hors domaine médical. Le vocabulaire issu du corpus médical est, quant à lui, inclus dans son intégralité.

Afin de limiter les coûts, en terme de temps de calcul, induits par nos expériences, nous décidons d'évaluer nos modèles selon leur perplexité sur les données de

développement de $config3$. Ainsi, nous sélectionnons le modèle intégrant les données génériques et spécialisées obtenant la perplexité la plus basse. Les expériences en traduction et en post-édition utilisant un modèle de langage mixte seront donc basées sur le modèle sélectionné. Les scores de perplexité des différents modèles utilisés pour la traduction et la post-édition sont détaillés dans le tableau 3.8.

		INTERPOLATION	CONCATÉNATION	
LM_{gen}	LM_{med}	LM_{gen} + LM_{med}	$LM_{gen+med}$	$LM_{gen+med}L$
648,7	9,1	10,0	227,7	140,4

TAB. 3.8 – *Scores de perplexité des modèles de langage utilisés pour nos expérimentations, selon le corpus de développement du domaine médical.*

Ces scores de perplexité montrent une forte adéquation des données de développement avec le modèle de langage spécialisé (LM_{med}). Le modèle de langage générique ne couvre que très partiellement les séquences n-gramme présentes dans le corpus spécialisé. La différence importante (un facteur de 10, jusqu'à 20) entre les perplexités de LM_{med} et des modèles de langage mixtes provient de la taille des corpus. En effet, le corpus générique a tendance à *noyer* les données médicales. Ce phénomène est très marqué pour le modèle $LM_{gen+med}$, où tout le vocabulaire générique est inclus dans le modèle. Pour $LM_{gen+med}L$, la différence avec LM_{med} est plus réduite, mais reste toutefois importante, et appuie notre hypothèse selon laquelle la quantité de données spécialisées est trop faible face aux données génériques pour pouvoir construire un modèle mixte se reposant sur la concaténation des données.

3.2.3 Construction de systèmes de traduction plus ou moins spécialisés

Dans cette section, nous présentons les résultats pour les systèmes présentés dans le tableau 3.7. Nous commençons par évaluer les systèmes PBMT et le système commercial sur le corpus de test du domaine médical $Emea_{test}$, selon les trois configurations liées au découpage des données médicales. Les résultats de ces expériences sont détaillés dans le tableau 3.9. Si le système PBMT construit uniquement sur des données génériques obtient les moins bons scores de BLEU, ce sont les combinaisons de données qui permettent d'atteindre les meilleurs résultats. Le système commercial est, quant à lui, placé entre les modèles de traduction génériques et ceux incluant des données médicales. Nous pensons donc que des données spécialisées sont certainement utilisées par ce système.

Les résultats obtenus pour les trois configurations permettent de vérifier la cohérence du découpage du corpus médical. En effet, la différence entre $config1$, avec une redondance des données entre les ensembles d'apprentissage, de développement et de test, et $config3$, atteint 30 points de BLEU pour les meilleures systèmes PBMT. Les résultats liés à la sélection aléatoire des trois sous-ensembles de données, pour $config2$, se situent entre les deux autres configurations. Des hypothèses de traduction produites par différents systèmes selon $config3$ sont présentés dans le tableau 3.10.

SYSTÈME DE TRADUCTION	SCORES BLEU			P-VALUE
	$config1$	$config2$	$config3$	
$PBMT_{gen}LM_{gen}$	27,0	27,9	29,9	0.002
$PBMT_{gen}LM_{gen} + LM_{med}$	49,9	42,2	38,2	0.002
$PBMT_{gen}LM_{med}$	50,4	42,5	39,2	0.002
com	48,3	46,5	44,9	0.007
$PBMT_{med}LM_{med}$	76,5	61,6	46,4	0.001
$PBMT_{gen+med}LM_{gen} + LM_{med}$	77,3	61,8	47,3	0.75
$PBMT_{gen+med}LM_{med}$	77,5	61,3	47,2	

TAB. 3.9 – *Résultats de traduction du corpus médical suivant plusieurs configurations.*

source	la mise en place de l'immunité a été démontrée
$PBMT_{gen}LM_{gen}$	the establishment of the immunity was demonstrated
com	the development of immunity has been demonstrated
$PBMT_{gen+med}LM_{med}$	onset of immunity has been demonstrated
référence	onsets of immunity have been demonstrated
source	rcp se présente sous la forme de lyophilisat
$PBMT_{gen}LM_{med}$	rcp consists of lyophilisat
com	rcp is in the form of a lyophilisate
$PBMT_{med}LM_{med}$	spc is supplied as a lyophilised powder
référence	rcp is a lyophilisate
source	en cas d'apparition d'effets indésirables
com	in case of occurrence of side effects
$PBMT_{gen+med}LM_{gen} + LM_{med}$	if side effects occur
$PBMT_{gen+med}LM_{med}$	if any of the side effects
référence	if side effects occur

TAB. 3.10 – *Hypothèses de traduction produites par différents systèmes PBMT selon $config3$.*

Les moins bons scores sont obtenus par $config3$, associés à la sélection chronologique des documents composant les trois sous-ensembles de données. C'est donc pour cette configuration qu'il y a le moins de redondance dans données médicales utilisées pour l'apprentissage, le développement et le test. Nous analyserons donc en priorité les résultats liés à cette configuration. Le système commercial suit la même tendance de baisse suivant les configurations, et ceci est dû principalement, selon nous, à la difficulté de traduction des données de $config3$. Le vocabulaire utilisé dans cet ensemble de test est plus complexe, tout comme les structures syntaxiques.

Nous observons un gain en absolu de plus de 8% de BLEU lors de l'insertion des données médicales dans le modèle de langage, et environ 1 point de gain BLEU est lié à l'utilisation du modèle de langage médical au lieu du modèle construit sur les données génériques et médicales concaténées. En ce qui concerne le système de traduction construit sur les données combinées, utiliser les données médicales pour le modèle de langage permet d'atteindre les résultats les plus élevés. La différence avec le modèle de langage interpolé n'est pas significative. Il paraît donc plus avantageux

d'utiliser le modèle de langage médical. En effet, dû à la disproportion dans la quantité de données de chaque type, beaucoup plus de données génériques sont disponibles, les probabilités associées aux n-grammes du domaine de spécialité sont affectées.

3.3 Évaluation de la post-édition

Afin d'évaluer la post-édition statistique selon les différents systèmes de traduction automatique, nous procédons à des expérimentations de post-édition. Chaque système de traduction présenté dans la section précédente correspond à un scénario possible, concernant la disponibilité et la quantité de données monolingues et bilingues spécialisées, la capacité à construire un système adapté à un domaine particulier, la possibilité d'utiliser un système accessible en ligne, etc. Nous voulons ainsi montrer la pertinence de notre approche en post-édition statistique pour l'adaptation à un domaine spécialisé.

Tout d'abord, le corpus d'apprentissage du domaine médical est traduit de la langue source vers la langue cible par chaque système de traduction automatique individuellement. Puis, utilisé avec sa traduction de référence, chaque corpus traduit permet de construire un modèle de post-édition. Nous proposons d'étudier dans cette section les différents systèmes de traduction automatique, évalués dans la section précédente, afin de mesurer l'impact de SPE pour l'adaptation au domaine médical.

3.3.1 Post-édition à partir d'un système de traduction commercial

Pour cette série d'expériences impliquant le système de traduction commercial, nous ne contrôlons aucun paramètre lors de la traduction, mais disposons de l'implémentation *Moses* pour construire le modèle de post-édition. Nous gardons la même configuration du système dans toutes les expériences menées en traduction PBMT et en post-édition. L'étape d'optimisation des paramètres du modèle de post-édition est basée sur la traduction du corpus de développement par le système commercial, aligné avec sa référence. Puis, les itérations de l'algorithme *MERT* permettent de converger vers les poids optimaux en terme de score BLEU.

Ensuite, le corpus de test traduit par le système commercial est décodé d'après le modèle de post-édition et le modèle de langage. Comme dans les expériences préliminaires en post-édition, nous mesurons le score BLEU sur l'ensemble du test, mais aussi un score BLEU *oracle*, permettant d'évaluer un BLEU maximum si les phrases dégradées par la post-édition sont maintenues à leur état post-traduction. Nous présentons les résultats obtenus dans le tableau 3.11, selon les trois configurations de données correspondant à des cas d'utilisation du corpus médical. La comparaison entre les systèmes de post-édition est statistiquement significative (p-value$=$ 0.001 pour les scores BLEU et p-value$=$ 0.05 pour les scores *oracle*).

Nous obtenons de très bons résultats sur ces expériences de post-édition. Pour $config3$, les scores obtenus en traduction avec le système $PBMT_{gen+med}LM_{med}$, soit

SYSTÈME	SCORES BLEU (*oracle*)		
	$config1$	$config2$	$config3$
com	48,3	46,5	44,9
+ $SPE_{med}LM_{gen}$	76,3 (76,9)	59,6 (61,6)	48,9 (51,4)
+ $SPE_{med}LM_{gen} + LM_{med}$	79,6 (80,2)	61,9 (64,3)	47,9 (53,5)
+ $SPE_{med}LM_{med}$	79,8 (80,3)	61,8 (64,3)	46,8 (53,3)

TAB. *3.11* – *Résultats de SPE sur des sorties issues d'un système commercial.*

47,2%, sont améliorés par l'application *naïve* de la post-édition statistique, avec $SPE_{med}LM_{gen}$ et $SPE_{med}LM_{gen} + LM_{med}$. Avec 1,6% de gain BLEU pour le modèle de langage construit sur les données hors domaine médical, et 0,6% avec le modèle de langage combinant général et médical, nous estimons que les gains sont significatifs et permettent d'envisager une étude plus poussée sur la post-édition statistique pour l'adaptation au domaine médical.

Les résultats de post-édition sur des sorties de traduction issues du système *com* montrent une amélioration plus importante des scores BLEU, par rapport à l'utilisation d'un système de traduction statistique par segments seule. Cet aspect est très important pour appuyer notre approche, car la même quantité de données spécialisées est utilisée dans l'ensemble de nos expérimentations. Ainsi, l'utilisation de ces données spécialisées dans un système SPE permet d'atteindre de meilleurs scores BLEU, par rapport à la construction d'un système PBMT spécialisé sur les mêmes données.

Les scores *oracle* atteignent jusqu'à 53,5% de BLEU pour $config3$, et montrent qu'un gain de 5 points environ peut être atteint. Pour $config1$, la post-édition améliore fortement (jusqu'à 28% en absolu) les résultats obtenus par le système *com* seul. Les résultats sont plus mitigés pour $config2$, où les gains sont présents, mais non significatifs. Pour ces deux configurations, les scores *oracle* permettent encore des améliorations du BLEU, de manière plus réduite mais non négligeable.

Selon le découpage du corpus médical, nous voulons étudier la quantité de données nécessaire à la post-édition pour égaler et dépasser le score de *baseline* obtenu par le système *com* sur $config3$. Nous proposons la figure 3.2 afin d'illustrer la progression des scores BLEU et *oracle* selon la quantité de données d'apprentissage utilisée pour construire le modèle de post-édition.

Nous remarquons que le score *oracle* est au dessus de la baseline de 4 points BLEU avec peu de données du domaine de spécialité (10 mille paires de phrases). Cet aspect est très important car le manque de données spécialisées est le principal problème que nous abordons dans nos travaux, or ces résultats montrent que nous pouvons appliquer la post-édition avec peu de données du domaine médical, en sachant sur quelles phrases appliquer la post-édition. La courbe représentant les scores BLEU se détache du système PBMT servant de ligne de base dès que le corpus d'apprentissage dépasse les 100 mille paires de phrases.

Ces résultats nous permettent de penser que la post-édition statistique, basée sur l'alignement sous-phrastique entre des sorties de systèmes de traduction et leurs

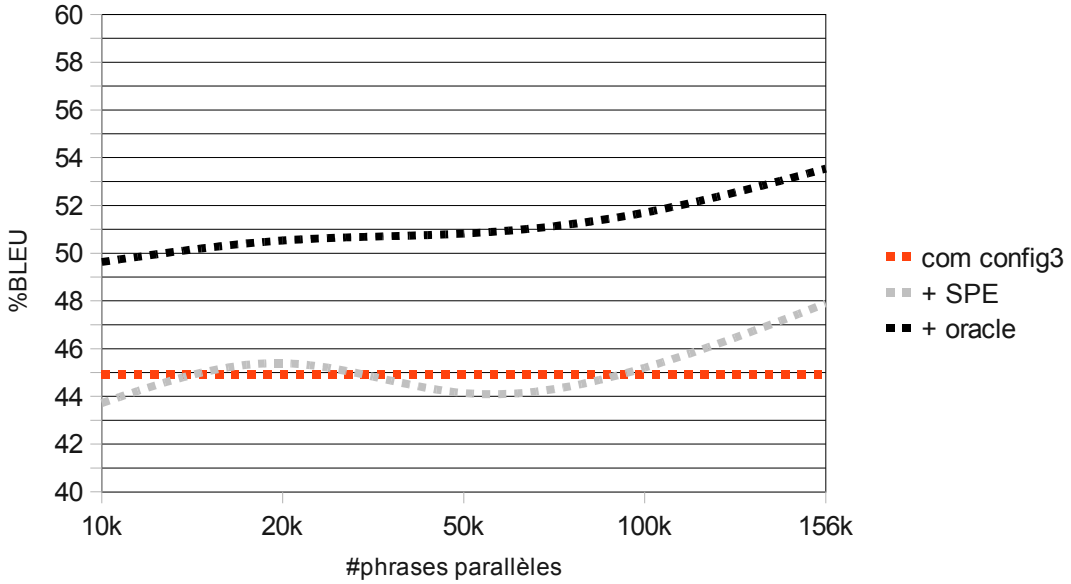


FIG. 3.2 – *Progression des scores BLEU et oracle en fonction de la quantité de données dédiées à la construction du modèle de post-édition. Le système de traduction est com, selon conf ig3, avec le système de post-édition* $SPE_{med}LM_{gen} + LM_{med}$.

références, peut améliorer des hypothèses de traductions issues d'un système PBMT. Cependant, nous voulons pouvoir tester plus de configurations, au niveau des données utilisées pour le système de traduction. Nous proposons dans les sections suivantes d'étudier l'impact de la post-édition statistique sur des systèmes de traduction par segments au niveau de l'état-de-l'art.

3.3.2 Post-édition à partir d'un modèle de traduction générique

Pour l'approche de traduction PBMT, nous disposons de plusieurs modèles de traductions, selon les corpus d'apprentissage utilisés, mais aussi plusieurs modèles de langages. Nous proposons dans un premier temps d'évaluer l'impact de notre approche sur le système PBMT basé uniquement sur des données génériques ($PBMT_{gen}LM_{gen}$).

Avec un modèle de langage générique

Nous voulons vérifier la possibilité d'adapter un système de traduction construit sur des données génériques (hors domaine médical) au domaine de la médecine. Nous utilisons les deux modèles de langages, individuellement ou combinés. Les résultats de la post-édition de sorties du système générique sont présentés dans le tableau 3.12.

Les gains de la post-édition varient selon la configuration et le modèle de langage

SYSTÈME SPE	SCORES BLEU (*oracle*)		
	$config1$	$config2$	$config3$
$PBMT_{gen}LM_{gen}$	27,0	27,9	29,9
+ $SPE_{med}LM_{gen}$	73,3 (73,3)	56,4 (56,7)	43,6 (43,9)
+ $SPE_{med}LM_{gen} + LM_{med}$	78,1 (78,2)	60,6 (61,1)	45,6 (47,0)
+ $SPE_{med}LM_{med}$	77,2 (77,4)	60,5 (60,9)	46,1 (47,3)

TAB. 3.12 – *Résultats de SPE sur des sorties issues d'un système PBMT utilisant un modèle de traduction et un modèle de langage générique (p-value= 0.001).*

utilisé. Comme dans les résultats de traduction pour $config3$, l'utilisation du modèle de langage médical permet d'atteindre les scores BLEU les plus élevés. Pour $config2$, l'interpolation des modèles de langage ou le modèle de langage médical permettent d'atteindre des résultats similaires, car nous considérons qu'une différence de $0,1\%$ de BLEU n'est pas significative. Pour $config1$, c'est l'interpolation des modèles de langage qui mène aux meilleurs résultats.

Pour le modèle de langage générique, un gain absolu supérieur à 13% de BLEU est observé entre la première phase de traduction et l'application *naïve* de la post-édition pour $config3$. Le score *oracle* indique que la sélection de phrases à post-éditer ne permet pas d'améliorer de manière significative les résultats, car la différence de scores BLEU est égale à $0,3\%$. Ceci peut provenir de trois phénomènes :

- peu d'erreurs sont introduites par la post-édition,
- peu d'erreurs sont corrigées par la post-édition,
- de manière générale, la post-édition introduit des erreurs, mais en corrige un plus grand nombre.

L'utilisation des modèles de langage contenant des données médicales permet une amélioration des résultats de l'ordre de 2 points BLEU environ pour le système $SPE_{med}LM_{gen} + LM_{med}$, et de 2,5 points pour $SPE_{med}LM_{med}$, par rapport au système $SPE_{med}LM_{gen}$. Ces gains, liés à la post-édition, permettent d'atteindre les scores issus de la traduction avec un système PBMT construit sur les données médicales. Nous pouvons comparer les résultats en traduction obtenus par le système $PBMT_{med}LM_{med}$ pour $config3$ (tableau 3.9), avec la combinaison des systèmes $PBMT_{gen}LM_{gen}$ et $SPE_{med}LM_{med}$. Les scores BLEU sont de $46,4\%$ et $46,1\%$ respectivement. La différence non-significative entre ces deux scores indique la possibilité d'adapter un système PBMT générique en utilisant SPE. C'est dans cette seconde phase que les données spécialisées sont introduites.

Cet aspect est très intéressant dans le cas où un système de traduction générique est déjà disponible et que des données spécialisées sont à traduire. Il est moins coûteux de combiner $PBMT_{gen}LM_{gen}$ et $SPE_{med}LM_{med}$ pour atteindre des résultats équivalents à ceux atteints par $PBMT_{med}LM_{med}$, plutôt que de construire un système de traduction utilisant l'ensemble des données (génériques et spécialisées). Cependant, ces systèmes de traduction combinant les types de données produisent des phrases cibles atteignant des résultats supérieurs.

Toutefois, les scores *oracles* indiquent que les gains peuvent être encore supérieurs

si la post-édition est appliquée uniquement sur certaines phrases. Selon le modèle de langage utilisé, uniquement médical ou combiné, les scores BLEU peuvent être améliorés, jusqu'à 1, 5%, par rapport à l'application *naïve* de la post-édition. La sélection des phrases à post-éditer permet donc d'atteindre des scores équivalents à ceux issus du système de traduction $PBMT_{gen+med}LM_{med}$, et supérieurs à ceux obtenus par le système $PBMT_{med}LM_{med}$.

Avec un modèle de langage médical

Lorsque des données spécialisées dans la langue cibles sont disponibles, il est alors possible de construire un modèle de langage adapté au domaine. Nous effectuons des expériences de post-édition statistique sur les sorties du système PBMT possédant un modèle de traduction générique et un modèle de langage médical ($PBMT_{gen}LM_{med}$). Le modèle de langage utilisé lors de la phase de traduction est identique à celui utilisé lors de la phase de post-édition. Les résultats obtenus sont présentés dans le tableau 3.13. Les scores BLEU montrent des gains par rapport à la traduction par le système $PBMT_{gen}LM_{med}$, compris entre 1,8% et 3,5%. Ces gains sont largement inférieurs à ceux obtenus lors de la post-édition des sorties de $PBMT_{gen}LM_{gen}$. Notre approche permet donc d'améliorer de manière plus significative un système PBMT générique, par rapport à un système de traduction utilisant un modèle de langage médical.

SYSTÈME SPE	SCORES BLEU (*oracle*)		
	$config1$	$config2$	$config3$
$PBMT_{gen}LM_{med}$	50,4	42,5	39,2
+ $SPE_{med}LM_{gen}$	70,6 (70,9)	51,4 (52,1)	41,0 (42,1)
+ $SPE_{med}LM_{gen} + LM_{med}$	75,2 (75,9)	55,5 (56,3)	42,5 (44,4)
+ $SPE_{med}LM_{med}$	75,2 (75,5)	55,9 (56,6)	42,7 (44,2)

TAB. *3.13* – *Résultats de SPE sur des sorties issues d'un système PBMT utilisant un modèle de traduction générique et un modèle de langage médical.*

Pour les trois configurations, les scores BLEU n'atteignent pas ceux présentés dans le tableau 3.12. De plus, pour $config3$ par exemple, les scores *oracle* sont à 3 points en dessous des résultats obtenus par traduction avec le système $PBMT_{gen+med}LM_{med}$, et 2 points sous les résultats de $PBMT_{med}LM_{med}$. Utiliser des modèles de langage différents, dans un ordre particulier, entre la phase de traduction et celle de post-édition paraît donc plus appropriée pour notre approche. Si le système PBMT se base sur un modèle de langage générique, suivi d'une post-édition utilisant un modèle de langage médical, les gains sont supérieurs à la configuration inverse.

Avec deux modèles de langage interpolés

Nous pensons qu'il est intéressant de continuer ces expériences, en post-éditant les sorties du système PBMT générique utilisant des modèles de langage interpolés

($PBMT_{gen}LM_{gen} + LM_{med}$). Selon les résultats obtenus précédemment par le système $SPE_{med}LM_{gen}$, étant les plus bas des trois modèles de langage utilisés, nous décidons d'étudier la post-édition avec les modèles de langage interpolés, ou le modèle médical uniquement. Les résultats de ces expériences sont détaillés dans le tableau 3.14. Cependant, nous observons une dégradation des résultats pour $config3$, en post-édition *naïve* et en évaluation *oracle*.

SYSTÈME SPE	SCORES BLEU (*oracle*)		
	$config1$	$config2$	$config3$
$PBMT_{gen}LM_{gen} + LM_{med}$	49,9	42,2	38,2
$+ SPE_{med}LM_{gen} + LM_{med}$	76,1 (76,5)	56,1 (57,0)	42,3 (44,0)
$+ SPE_{med}LM_{med}$	76,1 (76,4)	56,2 (57,0)	42,4 (44,0)

TAB. 3.14 – *Résultats de SPE sur des sorties issues d'un système PBMT utilisant un modèle de traduction générique et des modèles de langage interpolés.*

Ces résultats sont cohérents par rapport au système $PBMT_{gen}LM_{gen} + LM_{med}$, et à l'utilisation des modèles de langage interpolés. En effet, dans toutes les expériences menées jusqu'à présent, le modèle de langage médical permet d'atteindre les meilleurs scores BLEU. Nous pouvons tout de même noter le gain pour $config1$ et $config2$ par rapport aux résultats présentés dans le tableau 3.13. Ces derniers résultats nous incitent à penser que les sorties du système $PBMT_{med}$ peuvent difficilement être améliorées, étant donné les bons scores BLEU obtenus initialement.

3.3.3 Post-édition à partir d'un modèle de traduction médical

Nous décidons tout de même d'expérimenter la post-édition sur des systèmes de traduction statistique par segments basés sur un modèle de traduction spécialisé, et se différenciant par les modèles de langage utilisés. Le tableau 3.15 contient les résultats obtenus par la post-édition statistique sur le système $PBMT_{med}LM_{med}$. Nous pouvons remarquer l'équivalence entre les scores BLEU obtenus avec la post-édition statistique ($SPE_{med}LM_{gen} + LM_{med}$ et $SPE_{med}LM_{med}$), peu importe le modèle de langage utilisé, pour toutes les configurations de séparation des données. Les scores *oracle* sont, eux aussi, quasiment identiques entre $SPE_{med}LM_{gen} + LM_{med}$ et $SPE_{med}LM_{med}$.

SYSTÈME SPE	SCORES BLEU (*oracle*)		
	$config1$	$config2$	$config3$
$PBMT_{med}LM_{med}$	76,5	61,6	46,4
$+ SPE_{med}LM_{gen} + LM_{med}$	81,1 (81,6)	62,7 (63,9)	44,6 (47,4)
$+ SPE_{med}LM_{med}$	81,1 (81,6)	62,7 (63,8)	44,8 (47,5)

TAB. 3.15 – *Résultats de SPE sur des sorties issues d'un système PBMT utilisant un modèle de traduction et un modèle de langage médical.*

Pour $config1$, les scores d'application *naïve* de post-édition sont un demi point de BLEU sous les scores *oracle*. Les résultats de $config2$ montrent une différence d'un point

entre les deux scores. Pour $config3$ en revanche, un gain de 3 points est possible si les phrases à post-éditer sont sélectionnées. Nous pourrions de cette manière améliorer d'un point de BLEU les scores obtenus par $PBMT_{med}LM_{med}$ qui, pour $config3$, se situe à 46,4% de BLEU. Toutefois, les meilleurs scores issus de la traduction, avec les modèles de traduction combinés, ne sont pas significativement améliorés. En effet, le système de traduction $PBMT_{gen+med}LM_{med}$ permet d'obtenir un score de 47,2% de BLEU, contre 47,5% obtenu avec $PBMT_{med}LM_{med}$ combiné avec $SPE_{med}LM_{med}$.

Nous remarquons par cette série d'expériences que deux systèmes, soit $PBMT_{med}LM_{med}$ et $SPE_{med}LM_{med}$, utilisant les même données spécialisées et mis en place en série, permettent d'améliorer les traductions émises. Cette affirmation est uniquement vraie si les phrases à post-éditer peuvent être sélectionnées en amont afin de ne pas introduire trop d'erreurs lors de la phase de SPE. Il est donc évident que certaines hypothèses de traduction émises par le système PBMT contiennent des erreurs pouvant être apprises par un système de SPE, puis corrigées par la post-édition automatique.

Dans le cas où une redondance existe dans les données d'apprentissage, de développement et de test, les gains de la post-édition sont plus important. Pour $config1$, plus de 4% d'écart séparent les scores de traduction et de post-édition *naïve*, en utilisant les mêmes données. Pour $config2$, l'amélioration est plus légère, de l'ordre de 1%, par rapport aux traductions initiales. Nous pensons donc que la post-édition *naïve* peut être appliquée, lorsque les corpus du domaine de spécialité utilisés pour la traduction contiennent de la redondance au niveau des phrases ($config1$), ou au niveau des segments sous-phrastiques ($config2$).

3.3.4 Post-édition à partir de modèles de traduction combinés

Nous proposons à présent d'évaluer la post-édition statistique sur des hypothèses de traduction produites par un système de traduction par segment constitué de deux modèles de traduction combinés. De ce fait, nous voulons savoir s'il est possible d'améliorer les scores BLEU lorsque les données bilingues sont disponibles pour effectuer la traduction de la langue source vers la langue cible. Nous évaluons la post-édition suivant deux systèmes de traduction, l'un avec un modèle de langage mixte (les deux modèles de langage interpolés) et l'autre avec un modèle de langage spécialisé. Pour ces deux configurations, l'ensemble des données spécialisées sont intégrées dès la phase de traduction.

Tout d'abord, nous étudions la post-édition mise en série avec le système $PBMT_{gen+med}LM_{gen} + LM_{med}$, dont les résultats sont présentés dans le tableau 3.16. Nous observons une légère dégradation des résultats pour $config3$, que nous pouvons considérer comme non significative. Pour les deux autres configurations, les gains sont de 4,5 et 1% environ pour $config1$ et $config2$ respectivement. Les scores *oracle* permettent cependant d'améliorer les résultats d'une manière générale, jusqu'à 1 point pour $config3$ par rapport à $PBMT_{gen+med}LM_{gen} + LM_{med}$, et de 0,5 point comparé à $PBMT_{gen+med}LM_{med}$.

SYSTÈME SPE	SCORES BLEU (*oracle*)		
	$config1$	$config2$	$config3$
$PBMT_{gen+med} LM_{gen} + LM_{med}$	77,3	61,8	46,6
$+ SPE_{med} LM_{gen} + LM_{med}$	81,8 (82,3)	62,9 (63,6)	46,6 (47,7)
$+ SPE_{med} LM_{med}$	81,7 (82,2)	62,7 (63,8)	46,4 (47,7)

TAB. 3.16 – *Résultats de SPE sur des sorties issues d'un système PBMT utilisant des modèles de traduction combinés et des modèles de langage interpolés.*

Afin d'évaluer la post-édition statistique sur un modèle de traduction combiné et un modèle de langage spécialisé, nous effectuons des expériences de post-édition des traductions du système $PBMT_{gen+med} LM_{med}$. Ce système a obtenu les meilleurs résultats pour la traduction de données médicales. Les résultats sont présentés dans le tableau 3.17. Nous nous sommes concentrés sur $config3$ afin de limiter le temps de calcul. Nous observons une dégradation non significative des scores BLEU de $config3$. Les scores *oracle* permettent cependant d'améliorer les résultats de $PBMT_{gen+med} LM_{med}$ d'environ 0,7%. Seuls les scores *oracle* dépassent ceux issus de la traduction PBMT, lorsque deux modèles de traduction sont combinés, associé à un modèle de langage médical.

SYSTÈME SPE	SCORES BLEU (*oracle*)
	$config3$
$PBMT_{gen+med} LM_{med}$	47,2
$SPE_{med} LM_{gen} + LM_{med}$	46,3 (48,1)
$SPE_{med} LM_{med}$	47,1 (48,0)

TAB. 3.17 – *Résultats de SPE sur des sorties issues d'un système PBMT utilisant des modèles de traduction combinés et un modèle de langage médical.*

Dans toutes les expériences que nous avons menées en post-édition, les scores *oracle* indiquent des gains possibles, par rapport à l'application *naïve*. Nous proposons deux approches dans les sections suivantes afin de se rapprocher des scores *oracle*. La première est basée sur la sélection des phases issues de la traduction permettant d'obtenir des gains avec la post-édition. La seconde consiste à étudier plus en détail la composition du modèle de traduction contenant les couples de segments sous-phrastiques alignés, afin d'en retirer les éléments induisant des pertes en terme de score BLEU.

3.3.5 Choix des phrases à post-éditer

Motivations

Lors de la post-édition d'un text traduit, les scores BLEU ou TER de certaines phrases sont dégradés, pour d'autres en revanche ils sont améliorés. Nous proposons dans cette section de classifier les phrases à la sortie du système de traduction, avant

l'étape de post-édition, afin de prédire si cette dernière étape va apporter un gain ou non. Pour cela, nous voulons utiliser une méthode d'apprentissage supervisée de type séparateur à vaste marge. Nous regroupons sous un ensemble d'apprentissage des phrases classifiées selon leurs gains définis par la mesure $\Delta BLEU$ décrite dans la section 3.2.1. Une phrase avec un $\Delta BLEU$ positif sera affectée à une classe, tandis qu'une phrase avec un $\Delta BLEU$ négatif ou nul sera affectée à une autre classe.

Modéliser cet ensemble de phrases comme un problème de discrimination à deux classes nous permet de projeter dans cet espace des phrases de test, et de vérifier s'il est possible de prédire un gain de $\Delta BLEU$. Ainsi, seules les phrases dont le $\Delta BLEU$ est positif sont post-éditées. Les autres sont gardées intactes, à la sortie du système PBMT. Nous mettons donc en place un protocole d'expérimentation simple et ciblé sur la classification des phrases issues des systèmes PBMT. La figure 3.3 illustre l'architecture mise en place pour construire le modèle de classification à l'aide d'un SVM, basé sur les phrases issues du système PBMT associées à leur score $\Delta BLEU$.

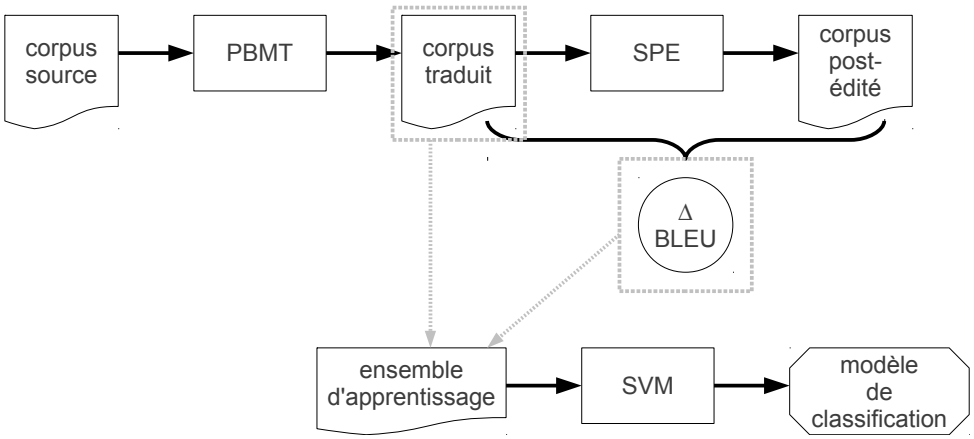


FIG. 3.3 – *Mise en place d'un classifieur de type SVM construisant son modèle à partir des phrases issues d'un corpus traduit et de leur classe respective en fonction de $\Delta BLEU$.*

Protocole expérimental

Afin de former un ensemble d'apprentissage permettant au classifieur de construire son modèle, nous disposons du corpus médical présenté dans la section 3.2.2. Deux ensembles de phrases étiquetées par classes peuvent donc être formés : avec le corpus d'apprentissage et avec le corpus de développement. Afin d'aléger les coûts de calcul et de nous focaliser sur la configuration obtenant les résultats les plus bas, nous considérons uniquement la troisième configuration des données ($config3$) pour les expérimentations en classification.

Pour chacun de ces corpus, nous avons besoins de la traduction source-cible, mais aussi de la post-édition cible-cible. Pour chaque phrase, avant et après post-édition, nous calculons le score $\Delta BLEU$ associé, permettant d'établir la classe résultante. Pour le corpus d'apprentissage traduit avec le système com, 16,8% des phrases sont étiquetées

comme n'apportant pas de gain, et 83,2% étiquetées avec gain. Cette disproportion est lié au fait que ces données sont les même que celles utilisées pour la construction du modèle de post-édition. Nous envisageons donc d'utiliser le corpus de développement afin de construire le modèle de classification.

En plus de la classe liée au score $\Delta BLEU$, nous utilisons la phrase issue du système de traduction PBMT comme paramètre d'apprentissage. Nous pouvons aussi disposer des scores issus du décodeur, associé à chaque segment et chaque phrase produite pour les hypothèses de traductions. Nous étudions la corrélation entre le gain BLEU lié à la post-édition et le score du décodeur associé à chaque hypothèse lors de la phase de traduction. Nous nous basons sur ce score de confiance pour représenter l'ensemble des phrases regroupées par scores issus du décodeur.

La figure 3.4 montre ces regroupements, selon les scores issus du décodeur en post-édition des sorties du système de traduction $PBMT_{gen+med}LM_{med}$. L'ensemble des phrases du corpus de développement sont utilisées et groupées par gain $\Delta BLEU$. Nous avons effectué la même opération sur le corpus d'apprentissage, et nous avons observé la même tendance. Nous pouvons remarquer qu'aucune corrélation particulière existe entre les scores du décodeur PBMT et les gains $\Delta BLEU$. Il semble donc que cette méthode ne peut fonctionner pour isoler un type de phrase à post-éditer, car le paramètre issu du décodeur correspond, pour chaque tranche de score, à des phrases associées aux deux classes.

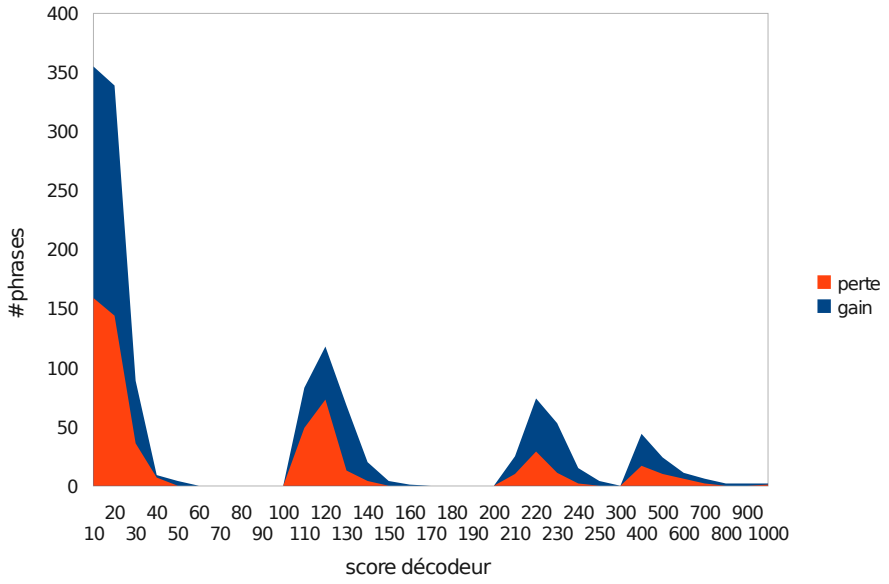


FIG. 3.4 – *Corrélation entre les scores issus du décodeur PBMT et les gains apportés par SPE, pour le corpus de développement.*

Si le score du décodeur ne permet pas d'indiquer le gain possible avec la post-édition, nous voulons confronter les scores BLEU avant et après la post-édition. En ne gardant que les phrases avec un $\Delta BLEU$ positif, nous voulons les regrouper par

tranches de scores BLEU après traduction. Nous effectuons ces mesures directement sur le corpus de test, afin d'estimer le potentiel maximum de la méthode (mesure *oracle*). La figure 3.5a présente la répartition des phrases avec gain lié à la post-édition statistique selon leur score obtenu après la traduction.

Nous pouvons appliquer la méthode de classification basée sur un modèle construit à partir des phrases traduites associées à leur score $\Delta BLEU$ à tous les systèmes de traduction utilisés dans la section 3.2.2. Cependant nous nous concentrons donc sur le système commercial (*com*), car c'est celui qui permet d'obtenir les scores BLEU les plus élevés selon les *oracles* mesurés. Afin d'illustrer le nombre de phrases permettant des gains avec la post-édition, regroupées par $\Delta BLEU$, nous proposons la figure 3.5b présentant ces détails sur le corpus de test.

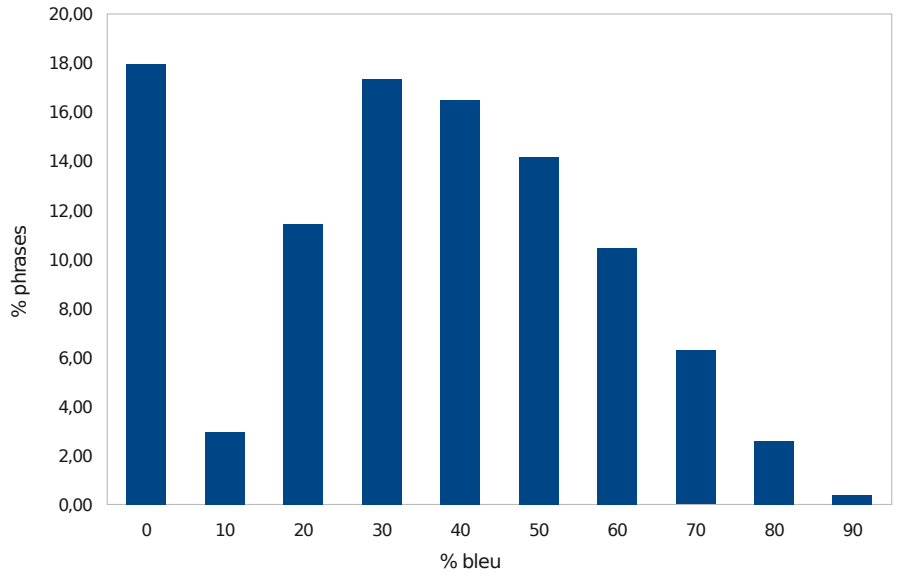
Expériences et résultats

Dans cette section, nous évaluons l'approche par classification basée sur SVM, permettant de sélectionner les phrases à post-éditer pour améliorer les résultats de traduction. Nous nous concentrons sur les résultats obtenus avec le *com*. Notre objectif est de nous rapprocher du score oracle obtenu avec ce sytème et la post-édition statistique, pour $config3$, selon les résultats obtenus avec le système $SPE_{med}LM_{gen} + LM_{med}$ (le score oracle est de 53, 3% de BLEU dans cette configuration). Dans un premier temps, nous présentons les résultats obtenus avec le corpus médical d'apprentissage utilisé ($c3Emea_{train}$) pour construire le modèle de classification. Le corpus de test $c3Emea_{test}$ est ensuite étiqueté automatiquement par le SVM. A chaque phrase est associé une classe et un score. Ce dernier indique l'erreur de prédiction émise par le classifieur lors de l'étiquetage de données non observées (Lin et al., 2007).

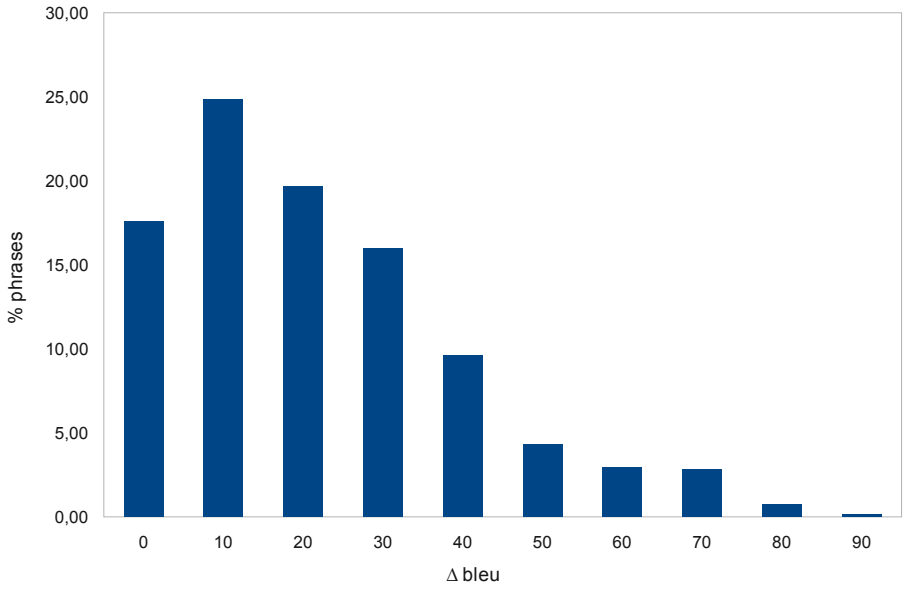
Le tableau 3.18 présente les résultats obtenus en utilisant le corpus d'apprentissage comme base pour la construction du modèle. Nous pouvons appliquer l'ensemble des étiquettes issues du classifieur, ou nous limiter à celles dont les erreurs de prédictions sont basses. Nous proposons de mesurer l'impact de la sélection des phrase à post-éditer selon les tranches d'erreur de prédiction, individuellement ou cumulées. Le système $SPE_{med}LM_{gen} + LM_{med}$ est utilisé dans la phase de post-édition car il correspond aux scores oracles les plus élevés, donc les résultats avec le plus de gain possible si la sélection des phrases à post-éditer est optimale (l'application naïve de la post-édition statistique permet d'atteindre 47,9% de BLEU, l'oracle se situe à 53,5%).

SCORE DE PRÉDICTION	0,5	0,6	0,7	0,8	0,9
# PHRASES	127	133	264	655	821
TRANCHES INDIVIDUELLES	44,8	44,7	44,5	45,1	48,1
TRANCHES CUMULÉES	-	47,7	47,8	48,0	48,0

TAB. 3.18 – *Scores BLEU en SPE, sur des sorties d'un système commercial, après classification des phrases à post-éditer. Les phrases sont regroupées selon leur score de prédiction issu du SVM, avec un modèle construit sur le corpus médical utilisé pour l'apprentissage.*



(a) Distribution des phrases avec gains suite à SPE, selon les scores BLEU après traduction.



(b) Distribution des phrases avec gains suite à SPE, par tranche de $\Delta BLEU$.

FIG. 3.5 – *Regroupement des phrases par score BLEU après traduction, et par $\Delta BLEU$ après SPE.*

Les résultats obtenus en classification permettent d'améliorer de 0,25% le score BLEU par rapport à l'application naïve de SPE. Ces résultats ne sont pas significatifs, mais permettent d'observer un phénomène intéressant, lié à la sélection des phrases à post-éditer selon le score de prédiction du classifieur. Un gain BLEU est observable uniquement lorsque le score de prédiction du classifieur est au dessus de 0,9. Le cumul des phrases selon leurs tranches de score de prédiction permet d'obtenir un gain BLEU, dès que ce score de prédiction est supérieur à 0,7.

Nous pensons toutefois que le corpus $c3Emea_{train}$, utilisé pour entraîner le SVM, n'est pas adapté à la construction du modèle de classification des phrases avec ou sans gain lié à la post-édition statistique. En effet, c'est ce même corpus qui est utilisé pour construire le système de post-édition. Nous proposons alors d'utiliser le corpus de développement $c3Emea_{dev}$ pour construire le modèle de classification. Le processus est identique à celui mis en place lors de l'utilisation du corpus $c3Emea_{train}$. Les résultats obtenu après la phase de classification et de post-édition sont présentés dans le tableau 3.19.

SCORE DE PRÉDICTION	0,5	0,6	0,7	0,8	0,9
# PHRASES	436	489	557	458	60
TRANCHES INDIVIDUELLES	44,8	44,6	45,0	48,5	45,3
TRANCHES CUMULÉES	-	48,4	48,5	48,8	48,9

TAB. 3.19 – *Scores BLEU en SPE, sur des sorties d'un système commercial, après classification des phrases à post-éditer. Les phrases sont regroupées selon leurs scores de prédictions issus du SVM, selon un modèle construit sur le corpus médical utilisé pour le développement.*

L'utilisation du corpus de développement pour la construction du modèle de classification permet d'atteindre de meilleurs résultats en terme de scores BLEU, en comparaison à l'utilisation du corpus d'apprentissage. Lorsque la post-édition est limitée aux phrases du corpus de test regroupées par tranches de score de prédiction issu du SVM, un score BLEU de 48,51% est atteint par l'ensemble de phrases dans la tranche comprises entre 0,8 et 0,9. Lorsque nous cumulons les tranches, et cumulons donc les phrases dans ces tranches, les scores BLEU augmentent progressivement, de 48,4% jusqu'à 48,9%, soit 0,5% et 1% de gain respectivement, par rapport à l'application naïve de SPE.

Le score oracle pour la même configuration, présenté dans le tableau 3.11, n'est pas atteint, et le score BLEU peut encore être amélioré de 3,6% en théorie. Cependant, un certain nombre d'erreurs sont encore présentes dans les sorties du système SPE. Nous en présentons quelques exemples dans le tableau 3.20.

3.3.6 Élagage de la table de post-édition

Dans la section précédente, nous avons proposé une méthode de sélection des phrases sorties d'un système de traduction afin d'effectuer une post-édition automatique basée sur une approche statistique. Cette approche est identique à la

com	presentation of 10 vials of 1 dose
+ SPE	pack size of 10 bottles of 1 dose
référence	pack containing 10 bottles of 1 dose
com	viruses are contained in a bag itself
+ SPE	viruses are supplied in a sachet itself
référence	the viruses are contained in a sachet within
com	the adverse effects most frequently reported
+ SPE	the most frequent adverse reactions reported
référence	the most frequently reported side effects

TAB. 3.20 – *Extraits de sorties du système commercial après SPE, comparées à leurs références de traduction.*

description des systèmes de traduction PBMT faite dans la section 2.3.2, s'appuyant sur un modèle composé d'un ensemble de couples de segments sous-phrastiques alignés. Dans le cas de PBMT, le modèle est composé de segments dans la langue source alignés avec des segments dans la langue cible. Pour la post-édition, les segments alignés sont uniquement dans la langue cible.

Les segments sous-phrastiques du modèle de post-édition forment donc un ensemble de règles de réécriture permettant, en théorie, de capturer les erreurs issues du système de traduction, et de générer des phrases dans la langue cible de meilleure qualité. Or, il est possible que des règles de réécriture dégradent certaines phrases. Ce ne sont donc plus les phrases à post-éditer qui doivent être sélectionnées, mais les couples de segments présents dans le modèle de post-édition. Ceci afin d'appliquer uniquement les règles menant à un gain en terme de gain BLEU.

Dans les prochaines sections, nous présentons en détails l'idée générale permettant de sélectionner des segments sous-phrastiques à partir du modèle de post-édition. Le modèle résultant peut être vu comme une version *élaguée* de la table de réécriture initiale. Nous évaluons notre approche selon les différentes configurations PBMT et SPE présentées dans les sections 3.2.2 et 3.2.2. Nous détaillons les expériences menées pour l'élagage du modèle de post-édition, puis présentons les résultats obtenus en terme de scores BLEU.

Idée générale

L'application naïve de la post-édition présentée dans les sous-sections de 3.3.1 à 3.3.4 ne permet pas d'obtenir des gains en score BLEU dans toutes les configurations. Notamment, lorsque le corpus spécialisé est utilisé dans la première phase de traduction, une dégradation des résultats est observée. Cependant, les scores oracles indiquent que des gains sont possibles, et ceci pour toutes les configurations testées. Nous pensons donc que post-éditer les phrases issues de la traduction automatique peut introduire des erreurs. Cela signifie que certaines règles de réécriture ne doivent pas être appliquées, car elles induisent une dégradation des scores BLEU mesurés après

post-édition.

De cette hypothèse résulte notre proposition de sélection des couples de segments présents dans la table de post-édition. Nous voulons garder uniquement les couples de segments menant à un gain BLEU sur l'ensemble de phrases de test. Cet élagage de la table de post-édition peut être effectué grâce à un corpus parallèle composé des sorties de traduction automatique accompagnées de leurs traductions de référence. Pour chaque phrase du corpus d'entraînement, si un couple de segments issu de la table SPE, noté (seg_{src}, seg_{trg}), est appliqué, nous calculons un score $\Delta BLEU_{seg}$. Ce score est issu de la différence entre les scores BLEU d'une phrase avec et sans l'application de ce couple de segments SPE. Ainsi, pour l'ensemble des phrases du corpus, chaque couple de segments de la table de post-édition est associé à une liste de scores $\Delta BLEU_{seg}$.

De ce fait, nous pouvons retirer de la table les couples induisant des dégradations du score BLEU sur les phrases du corpus. Plusieurs méthodes peuvent alors être mises en place. Une moyenne des $\Delta BLEU_{seg}$ peut être calculée, et les couples ayant une moyenne négative sont retirés. Nous pouvons aussi être plus strict sur la sélection des règles de réécriture, et retirer celles auxquelles est associé à au moins un score $\Delta BLEU_{seg}$ négatif. De ce fait, nous voulons nous assurer de garder uniquement les couples de segments permettant d'obtenir des gains en terme de score BLEU. Notre approche est détaillée par l'algorithme 1.

Algorithm 1 Calcul des $\Delta BLEU$ pour chaque phrase du corpus de développement, et affectation de ce score aux couples de segments SPE concernés. Les couples remplissant la condition prédéfinie (moyenne des $\Delta BLEU_{seg} > 0$, ou $\forall \Delta BLEU_{seg} > 0$) sont gardée dans la table SPE.

$c3Emea_{dev}$: corpus de développement
$tableSpe$: table de post-édition
$tableSpeElag$: table de post-édition élaguée
$segSpe_j$: une paire de segments de $tableSpe$
$deltableu_j$: un score $\Delta BLEU_{seg}$ concernant $segSpe_j$
$listeDeltaBleu_j$: liste des $deltableu_j$ concernant $segSpe_j$
$condition$: condition de sauvegarde d'une paire de segment de $tableSpe$
for $phrase_i \in c3Emea_{dev}$ **do**
  for $segSpe_j \in tableSpe$ **do**
    if $segSpe_j$ appliquée à $phrase_i$ **then**
      Calculer $deltableu_j$
      $deltableu_j$ ajouté à $listeDeltaBleu_j$
    end if
  end for
end for
for $segSpe_j \in tableSpe$ **do**
  if $condition$ **then**
    $segSpe_j$ ajouté à $tableSpeElag$
  end if
end for

Expériences et résultats

Afin d'évaluer notre approche d'élagage de la table de post-édition, nous utilisons le corpus $c3Emea_{dev}$ afin d'évaluer chaque couple de segments provenant de cette table. Un exemple de phrase issue du corpus utilisé, associée à ses règles de réécriture, est présenté dans le tableau 3.21. Nous mesurons le score BLEU du corpus $c3Emea_{test}$, suite à la traduction de la langue source vers la langue cible par le système de traduction *com*, et la post-édition en utilisant la table élaguée. Les résultats sont présentés dans le tableau 3.22. La ligne de $\Delta BLEU$ positif correspond à la table de post-édition élaguée en gardant que les $\Delta BLEU_{seg} > 0$. La ligne de $\Delta BLEU$ moyen correpond aux paires de segments avec $\Delta BLEU_{seg} > 0$.

PHRASE À POST-ÉDITER	RÈGLES DE POST-ÉDITION
for a complete description of side effects reported with purevax rcpch , see instructions .	(a complete description of ; the full list of all) (instructions . ; the package leaflet .)
antibodies to help protect against diseases	(antibodies ; the antibodies) (to help ; will help to)

TAB. 3.21 – *Exemple d'une phrase accompagnée des couples de segments la concernant et impliquant une post-édition, issus de la table SPE.*

Lors de cette première série d'expériences, nous évaluons les gains BLEU acquis grâce à la post-édition sur le corpus $c3Emea_{test}$. Lors de la phase de décodage avec la table de post-édition élaguée, les poids λ_n associés aux fonctions f_n (voir la description dans la section 2.3.3) sont issus de la phase de développement suivant l'apprentissage du modèle de post-édition décrit dans la section 3.1.2.

Ces poids peuvent être réévalués lors d'une nouvelle phase d'optimisation (avec *MERT*), effectuée après élagage de la table de post-édition. Nous proposons cependant d'utiliser trois ensembles de poids différents :

- les poids définis par défaut par *Moses* (noté λ_{init}),
- les poids issus de l'optimisation *MERT* suivant l'apprentissage du modèle SPE (noté λ_{spe}),
- les poids résultants de l'optimisation *MERT* suivant l'élagage de la table SPE (noté λ_{el}).

Nous proposons dans le même temps de combiner notre approche d'élagage de la table de post-édition avec la méthode de sélection des phrases à post-éditer présentée dans la section 3.3.5. Suite à l'élagage de la table de post-édition, nous estimons un score *oracle* basé sur la sélection des phrases dont la post-édition apporte un gain en terme de score BLEU. Cet oracle indique les gains possibles lors de la combinaison des deux approches dans le cas où la classification des phrases à post-éditer permet de sélectionner rigoureusement les phrases dont la post-édition apporte un gain.

D'après ces premiers résultats concernant l'élagage de la table de traduction, les deux méthodes de sélection des couples de segments de la table de post-édition permettent d'obtenir des résultats équivalents en terme de BLEU. Nous remarquons

$\Delta BLEU$	λ_{init}	λ_{spe}	λ_{el}	*oracle*	# COUPLES SEGMENTS
positif	48,5	48,0	48,6	53,6	5 637 194
moyen	48,5	48,0	47,7	53,7	5 637 202

TAB. 3.22 – *Scores BLEU en SPE après élagage de la table de post-édition selon les $\Delta BLEU_{seg}$ moyens évalués sur le corpus $c3Emea_{dev}$. Les différents poids pouvant être associés au modèle SPE sont testés.*

tout de même que la table de post-édition élaguée en utilisant la méthode des $\Delta BLEU_{seg}$ strictement positifs contient 8 couples de segments en moins par rapport à la table élaguée avec l'autre méthode. L'utilisation des poids λ_{init} permet d'obtenir des scores BLEU plus élevés que lors de l'utilisation des poids issus du développement de SPE, λ_{spe}.

Nous avons aussi effectué une étape d'optimisation *MERT* à partir de la table SPE élaguée, en utilisant le corpus $c3Emea_{dev}$ pour optimiser les poids du modèle SPE basé sur cette nouvelle table. Nous remarquons que les résultats obtenus ne permettent pas d'améliorer les gains en terme de BLEU avec la méthode d'élagage basée sur la moyenne des $\Delta BLEU_{seg}$. Pour la méthode basée sur des $\Delta BLEU_{seg}$ strictement positifs, le gain BLEU est non significatif (0,1%) par rapport à l'utilisation des poids λ_{init}.

De ce fait, nous pensons que la manière dont nous avons effectué le développement post-élagage n'est pas optimale, notamment à cause de l'utilisation de la procédure par défaut qui consiste à prendre les poids λ_{init} comme base de l'étape d'optimisation. Il serait peut être judicieux d'utiliser les poids λ_{spe} afin de converger plus rapidement vers une solution sous-optimale obtenue par l'algorithme MERT. Finalement, les scores BLEU obtenus après la phase d'optimisation suivant l'élagage de la table de post-édition sont équivalents à ceux obtenus avec les poids λ_{init}.

3.4 Discussion

Dans ce chapitre, nous avons présenté une approche permettant la post-édition automatique de traductions. En nous basant sur les méthodes issues de PBMT (voir la description dans la section 2.3.2), un système de post-édition est utilisé en série d'un système de traduction automatique. Ainsi, aucune intervention humaine n'est nécessaire pendant les phases de traduction et de post-édition. Tout le processus de traitement est basé sur l'analyse statistique des corpus, sans ajout de données annotées manuellement, ce qui n'avait jamais été envisagé à notre connaissance.

Nous proposons dans cette section une synthèse des résultats obtenus lors des différentes expériences menées en post-édition, accompagnée d'une comparaison avec les travaux précédents issus de la littérature. Puis, une analyse des résultats est proposée, suivie par des propositions de travaux futurs, avant de conclure ce chapitre concernant l'édition *a posteriori* de traductions.

3.4.1 Synthèse des résultats

Dans les expériences menées en post-édition statistique présentées dans ce chapitre, nous avons remarqué une différence importante dans les gains BLEU selon les genres de documents concernés. Tout d'abord, les expériences préliminaires effectuées dans le cadre de la campagne d'évaluation *WMT11* n'ont pas permis d'améliorer les scores de traductions issues du système PBMT. Les gains en oracle sont minimes, probablement liés à la mixité des domaines couverts par les brèves journalistiques. Le domaine de la médecine est, quant à lui, beaucoup plus spécialisé et donc restreint.

Les résultats obtenus en post-édition pour le domaine médical montrent une amélioration plus importante des scores BLEU, par rapport à la traduction par un système PBMT générique. Cependant, l'ajout de données médicales dans la phase de traduction mène à des améliorations moindres, mais tout de même significatives. Il parait donc judicieux de construire un système de post-édition statistique spécialisé en cascade d'un système de traduction par segments générique. En effet, si peu de données spécialisées sont disponibles, il parait moins coûteux et plus efficace d'intégrer ces données à un système de post-édition statistique, plutôt que de re-développer entièrement un système de traduction. De plus, ce scénario est pertinent car il correspond à la réalité : peu de données spécialisées et des systèmes de traduction automatique générique disponibles.

L'utilisation d'un système de traduction automatique commercial, ou grand publique, combiné à un système de post-édition statistique spécialisé, permet d'atteindre les performances les plus élevées parmi toutes les configurations testées. Utiliser un système *boîte noir* générique suivi d'un système de post-édition statistique médical reste la solution la moins coûteuse et la plus performante. Cette configuration correspond à un cas d'utilisation réel d'un outil de traduction en ligne afin d'obtenir des hypothèses de traduction pouvant être post-éditées. L'amélioration des scores BLEU est assez importante pour valider notre approche de post-édition statistique basée sur des segments sous-phrastiques.

Cette conclusion est vraie dans l'application *naïve* de la post-édition et est appuyée par les deux méthodes proposées afin de se rapprocher des scores oracles mesurés. La classification des phrases à post-éditer en utilisant un classifieur de type SVM permet d'améliorer les scores BLEU par rapport à la post-édition statistique *naïve*, sans pour autant atteindre les oracles. L'élagage du modèle de post-édition permet lui aussi d'atteindre de meilleurs résultats par rapport à la première approche. Nous avons toutefois utilisé uniquement la métrique automatique BLEU pour l'évaluation de la qualité des hypothèses de traduction produites par nos systèmes. Il nous semble intéressant de procéder à une évaluation manuelle avant et après la post-édition afin d'analyser" de manière qualitative les améliorations et les dégradations induites par nos approches.

Il reste cependant des gains possibles, jusqu'à 4% de BLEU dans certaines configurations. Nous estimons que l'approche par classification des phrases à post-éditer combinée avec celle basée sur l'élagage de la table de post-édition pourrait

permettre d'améliorer encore les résultats. De plus, cette dernière approche peut être appliquée de manière itérative. Il nous parait envisageable de réduire la table de post-édition aux couples de segments permettant un gain BLEU, puis de réévaluer les scores $\Delta BLEU$ à partir de cette table réduite, et de recommencer ce processus jusqu'à obtenir une stabilisation des scores BLEU sur le corpus de développement.

Si la post-édition statistique basée sur l'alignement de segments sous-phrastiques et incluant des données annotées manuellement est une approche déjà étudiée dans la littérature, nous proposons deux nouvelles approches permettant des gains BLEU *a posteriori* de la traduction automatique. Ces deux approches peuvent se résumer en :

- la classification des phrases à post-éditer,
- la réduction de la table de post-édition.

Chacune de ces approches permet une amélioration des scores BLEU par rapport à l'application *naïve* de la post-édition statistique. Nous présentons dans la prochaine section les travaux issus de la littérature concernant la post-édition de traduction, en nous situant parmi les différentes approches proposées.

3.4.2 Travaux précédents

À notre connaissance, aucune étude n'à été effectuée sur l'utilisation en série d'un système de traduction PBMT et d'un système de post-édition statistique pour l'adaptation à un domaine de spécialité, sans ajout de données nécessitant l'intervention d'un humain. Cependant, de nombreux d'autres travaux concernent l'adaptation aux domaines de spécialités en traduction automatique. Utiliser l'édition *a posteriori* de traductions produites par un système automatique afin d'en améliorer la qualité et de les adapter à un domaine particulier est toutefois peu étudié. Les travaux de Isabelle et al. (2007) se concentrent sur une approche de post-édition statistique basée sur l'alignement de segments sous-phrastiques, identique à celle présentée dans la section 2.5.4.

Dans leurs études, les auteurs combinent un système de traduction à base de règles avec la post-édition statistique et obtiennent des améliorations en terme de TER et BLEU sur une tâche concernant la traduction d'offres d'emplois. Leur meilleur configuration pour le système RBMT seul, effectuant la traduction de la langue source vers la langue cible, permet d'atteindre 31,2% de BLEU pour la traduction du français vers l'anglais. Leur système de traduction par segments effectuant la même tâche de traduction source-cible atteint quant à lui 41,0% de BLEU. Combiner le système de traduction à base de règles et la post-édition statistique permet d'améliorer le score BLEU d'environ 4%, pour atteindre 44,9%.

Il nous parait cependant difficile de comparer notre approche directement avec des travaux issus de la littérature. Nous pouvons cependant remarquer les travaux de Koehn et Schroeder (2007) sur l'adaptation d'un système de traduction construit sur les données du Parlement Européen aux brèves journalistiques, lors de la campagne d'évaluation *WMT07*. Les auteurs ont évalué leur système lors de l'utilisation de données génériques et spécifiques pour la construction de modèles de langage et de

traduction. La meilleure configuration est composée des modèles de langage interpolés, associés avec deux modèles de traduction combinés de la même manière que celle utilisée dans nos travaux. Les auteurs obtiennent une amélioration de 2,5% de BLEU en comparaison avec un système de traduction par segments générique.

La combinaison en série d'un système de traduction à base de règles avec la post-édition statistique basée sur les segments sous-phrastiques permet donc d'améliorer la qualité des traductions produites automatiquement. L'adaptation à un domaine spécialisé est aussi possible, même si la quantité de données parallèles du domaine est largement inférieure à celle hors-domaine (ou générique). Une étude très complète de l'impact de la post-édition statistique sur des traductions issues du système RBMT Systran est présentée dans les travaux de Dugast et al. (2007). Les modifications lexicales induites par la post-édition statistique permettent de corriger jusqu'à 22% des erreurs produites par le système de traduction pour le couple de langues anglais-français. Les auteurs donnent une quantité importante de détails sur l'ensemble des améliorations et dégradations induites par leur approche. Cependant, la tâche diffère trop de celle sur laquelle nous nous sommes penchés dans cette thèse pour pouvoir en comparer les résultats.

La traduction par segment combinée à la post-édition statistique, sans ajout de données étiquetées ou post-éditées manuellement, est une approche récente ayant été assez peu étudiée jusquà présent. Cette approche apparaît comme valable, à la vue des résultats obtenus, et du faible coût nécessaire à la construction des système de traduction et de post-édition. De plus, il apparaît que plus le domaine est spécialisé, plus les gains avec la post-édition statistique sont élevés, lorsque nous comparons les résultats obtenus pour les brèves journalistiques et ceux obtenus pour le domaine médical. Il est cependant nécessaire d'évaluer de manière subjective les résultats obtenus car certains gains BLEU sont très faibles et ne permettent pas de rendre compte des améliorations dans la qualité des hypothèses de traduction.

Chapitre 4

L'acquisition de lexiques bilingues médicaux

Sommaire

Dans ce chapitre, nous présentons nos contributions pour la construction de lexiques bilingues dans le domaine médical, pouvant être inclus dans un système de traduction automatique ou dans un système de post-édition statistique. Les lexiques bilingues sont des ressources particulièrement utiles dans de nombreuses tâches relatives au traitement automatique du langage. Si l'utilisation de corpus parallèles permet de construire des lexiques bilingues de manière robuste et efficace, ces corpus ont tendance à faire défaut pour des domaines de spécialité. De nombreux travaux relatent de l'utilisation d'une autre ressource multilingue pour l'acquisition de lexiques bilingues : les corpus comparables. Plus de détails sur ces ressources sont donnés dans la section 2.6.3.

Ne disposant pas d'alignement au niveau des phrases ou des mots, contrairement aux corpus parallèles, les corpus comparables impliquent l'utilisation d'autres repères permettant de mettre en relation des unités lexicales étant des traductions. Nos travaux s'inscrivent dans ce contexte : comment prendre en considération les informations graphiques, sémantiques et lexicales, afin de repérer dans une langue cible les traductions d'un mot dans une langue source. Notre objectif est de modéliser les termes médicaux que nous voulons traduire selon ces différents aspects. Notre objectif étant d'aligner ces termes entre le français et l'anglais, selon des approches génériques, permettant de garder une indépendance par rapport aux langues.

Dans un premier temps, nous voulons étudier les aspects relatifs aux ressources bilingues nécessaires à ce type de tâches, sous-entendu le corpus comparable et le lexique pivot. Nous effectuons des expériences sous la forme de recherche documentaire afin de retrouver des traductions. Dans un second temps, nous proposons d'étudier les aspects permettant de repérer des traductions de termes sans disposer d'alignement interlangue au préalable. Trois approches (ou *vues*) sont proposées :

- contextuelle : similarité des vecteurs de contextes
- thématique : classification thématique des termes
- orthographique : détection de cognats

Puis, dans un troisième temps, nous présentons la combinaison de ces trois approches afin d'en étudier la complémentarité. Enfin, dans un dernier temps, nous étendons l'approche contextuelle *classique* en modélisant l'ensemble des contextes d'un terme sous la forme d'une matrice. Nous proposons alors un modèle génératif basé sur des modèles de mélanges gaussiens (*GMM* pour *Gaussian Mixture Models*) qui permet d'évaluer la vraisemblance contextuelle entre des termes de deux langues.

Afin de mener à bien nos expériences en acquisition de traductions de vocabulaire spécialisé, nous utilisons un ensemble de termes médicaux comme candidats à la traduction, chacun étant accompagné d'une traduction de référence dans la langue cible. Les deux langues concernées sont le Français et l'Anglais, comme langues source et cible respectivement. Nous présentons les détails des ressources utilisées pour nos expériences sur l'acquisition de vocabulaire spécialisé dans le tableau 4.1.

Corpus	Langue	Documents	Mots	Label
Corpus Comparable				
Wikipédia	Français	872 111	3 994 040	$wiki_{FR}$
Wikipédia	Anglais	3 223 790	14 059 292	$wiki_{EN}$
Lexiques Bilingues expériences préliminaires				
Général	Français	-	3 200	$gLex1_{FR}$
Général	Anglais	-	3 200	$gLex1_{EN}$
Spécialisé	Français	-	1 800	$sLex1_{FR}$
Spécialisé	Anglais	-	1 800	$sLex1_{EN}$
extraction de lexique bilingue				
Spécialisé	Français	-	9 000	$sLex2_{FR}$
Spécialisé	Anglais	-	9 000	$sLex2_{EN}$
Candidats expériences préliminaires				
MeSH	Français	-	10 000	$cand1_{FR}$
MeSH	Anglais	-	10 000	$cand1_{EN}$
extraction de lexique bilingue				
MeSH	Français	-	3 000	$cand2_{FR}$
MeSH	Anglais	-	3 000	$cand2_{EN}$

TAB. *4.1* – *Détails des ressources utilisées pour les expériences d'extraction de vocabulaire médical.*

4.1 Les ressources bilingues

Selon la description des ressources utiles à la construction de vecteurs de contexte faite dans la section 2.6.3, nous nous intéressons à la possibilité d'utiliser l'encyclopédie en ligne Wikipédia [1] comme corpus comparable. Son contenu est en évolution constante (plus de détails dans les travaux de Almeida et al. (2007)) et est librement disponible au téléchargement [2]. Sa structure est pratique, car elle permet de mettre en relation des articles encyclopédiques (ou documents) en différentes langues décrivant des éléments similaires. Ainsi, un alignement multilingue au niveau des documents est possible automatiquement et permet de constituer un corpus comparable. Cependant, nous voulons nous concentrer sur l'extraction de termes médicaux, en se basant sur leurs contextes, pouvant être composés de mots présents dans le lexique pivot.

Nous considérons donc l'ensemble des documents dans une langue comme une partie du corpus comparable. Notre corpus est composé des documents extraits de Wikipédia en Français et en Anglais. Afin de faciliter l'accès à cet ensemble de documents, nous utilisons l'outil d'indexation et de recherche d'information disponible en ligne NLGbase [3], composé d'un moteur de recherche interrogeant la base encyclopédique Wikipédia. Les résultats d'une requête sont retournés classés par la

1. www.wikipedia.org
2. http ://dumps.wikimedia.org/
3. http ://www.nlgbase.org/

valeur du cosinus mesurée entre les mots clés saisis et les documents (pour plus de détails sur le fonctionnement du moteur de recherche, voir (Charton et Torres-Moreno, 2010). Il nous permet aussi d'extraire automatiquement la liste des mots d'un document ordonnées par leurs scores *tf.idf*.

En ce qui concerne le lexique pivot, nous faisons varier sa composition selon le domaine de spécialité. Certains éléments du lexique peuvent être très fréquent, contrairement au vocabulaire spécialisé qui sera plus rare. C'est pourquoi il nous paraît intéressant d'utiliser deux lexiques pivots : un composé de vocabulaire général, l'autre composé de vocabulaire spécialisé. Nous voulons aussi mesurer l'impact de la taille du lexique, en faisant varier le nombre de mots qui le compose.

4.1.1 De la recherche d'information...

L'objectif des expériences menées sur les ressources bilingues est de valider l'utilisation de Wikipédia comme corpus comparable et d'un lexique comme pivot entre les deux langues étudiées. Pour pouvoir retrouver des traductions de termes médicaux dans les documents composant Wikipédia, nous utilisons l'algorithme 2. Le résultat est un ensemble de documents dans la langue cible, pouvant contenir la traduction du terme dans la langue source ayant été utilisé comme requête initiale. Nous mesurons un score oracle sur la présence de la traduction dans les documents cibles. De cette manière, nous estimons la possibilité d'extraire automatiquement un contexte source, qui une fois traduit, se trouve à proximité du terme langue cible cherché.

4.1.2 ... à l'extraction terminologique

Pour notre première série d'expériences permettant d'évaluer la présence des traductions cherchées, nous utilisons les deux lexiques pivots (générique et spécialisé) concaténés en un seul lexique. De ce fait, nous espérons pouvoir traduire le maximum de mots collectés des documents en langue source, afin de former des requêtes pour l'extraction de documents en langue cible. Nous faisons varier le nombre de documents langue source ainsi que le nombre de mots extraits de ces documents. Ces paramètres influent directement sur la taille du contexte traduit. Si aucun mot du contexte source ne peut être traduit, aucun document cible n'est retourné par le système, donc la traduction du terme initial n'est pas trouvée. Ainsi, nous mesurons deux *oracles* :

- Un *oracle* (noté *oracle*) calculé sur l'ensemble des termes à traduire.
- Un *oracle* restreint (noté $oracle_R$) aux termes dont au moins un mot de leurs contextes peut être traduit.

Chaque oracle est incrémenté si la traduction est trouvée dans les documents retournés par le système, puis normalisé afin de donner un résultat compris entre 0 et 1. La taille du contexte traduit est variable selon les termes source et les mots extraits des documents sources. Nous présentons donc cette taille sous la forme d'une moyenne calculée sur l'ensemble des termes source pour une configuration donnée. Un extrait des résultats obtenus lors de ces expériences sont présentés dans le tableau 4.2.

Algorithm 2 Recherche de documents contenant la traduction d'un terme médical.

doc_src : documents langue source
doc_trg : documents langue cible
nb_w : nombre de mots de contexte
$cand_{src_i}$: terme à traduire provenant de $cand1_{FR}$ (langue source)
$cand_{trg_i}$: référence du terme à traduire ($cand1_{EN}$)
$oracle$: score oracle sur la présence de la traduction
$oracle = 0$
$cand_{src_i}$ est la requête permettant d'extraire doc_src
for $doc\_src_j \in doc\_src$ **do**
  C_{src_i} est le contexte de $cand_{src_i}$ composé des nb_w premiers mots du document (selon $tf.idf$)
end for
$C_{src} \leftarrow \forall C_{src_i}$
C_{trg} est la traduction de C_{src} selon le lexique pivot
for $W_{trg_k} \in C_{trg}$ **do**
  W_{trg_k} est la requête permettant d'extraire doc_trg_k
end for
$doc\_trg \leftarrow \forall doc\_trg_k$
if $cand_{trg_i} \in doc\_trg$ **then**
  $oracle++$
end if

En suivant le même protocole, nous avons mené une série d'expériences en utilisant les deux lexiques pivots indépendamment. Nous présentons un extrait des résultats obtenus grâce au lexique contenant un vocabulaire général dans le tableau 4.3, et grâce au lexique spécialisé dans le tableau 4.4. Ces résultats montrent que le lexique spécialisé permet de retrouver un nombre de traductions plus élevé que le lexique général car les oracles mesurés sont meilleurs en utilisant le lexique médical. Le vocabulaire général permet d'accroître le nombre de mots traduits dans le contexte d'un terme, donc d'augmenter le nombre de points d'ancrage entre les langues, mais plus de documents collectés dans la langue cible ne contiennent pas la traduction recherchée. L'utilisation d'un lexique spécialisé afin d'établir des points d'ancrage entre les langues paraît donc plus approprié dans ce type d'expériences. Finalement, les meilleurs résultats sont obtenus avec un lexique pivot mixte.

L'ensemble des résultats de recherche de traduction de termes médicaux parmi les documents de Wikipédia donnent un bon aperçu des possibilités liées à l'utilisation de ce corpus comme ressource comparable pour l'extraction de lexique bilingue. En premier lieu, le nombre de candidats couverts par l'encyclopédie est très élevé, même en français (Wikipédia est plus développée en anglais). C'est un aspect très important si l'on veut disposer d'une ressource permettant de travailler sur des termes de spécialité. En second lieu, jusqu'à 80% de bonnes traductions sont trouvées dans les documents retournés par le système mis en place lors de ces expériences, et ceci sans ordonner les documents composant le corpus. Nous estimons donc que Wikipédia est une bonne

#documents	#mots	taille contexte	*oracle*	$oracle_R$
1	10	2,06	0,29	0,58
	100	11,07	0,49	0,67
	200	19,95	0,53	0,72
10	1	1,74	0,21	0,54
	10	7,99	0,48	0,69
	20	14,78	0,53	0,75
20	1	2,28	0,25	0,56
	5	6,81	0,44	0,67
	10	12,09	0,51	0,72
50	1	3,31	0,30	0,59
	2	5,45	0,37	0,63
	5	11,06	0,47	0,70

TAB. 4.2 – *Extrait des résultats des expériences en recherche de traduction parmi les documents de Wikipédia, en utilisant un lexique pivot mixte.*

ressource comparable pour l'extraction de termes médicaux dans plusieurs langues.

documents	mots	contexte	*oracle*	$oracle_R$
1	10	1,36	0,09	0,42
10	10	3,28	0,26	0,43
20	5	3,02	0,24	0,44
50	1	1,76	0,14	0,43

TAB. 4.3 – *Extrait des résultats des expériences en recherche de traduction parmi les documents de Wikipédia, en utilisant un lexique pivot générique.*

documents	mots	contexte	*oracle*	$oracle_R$
1	10	1,99	0,19	0,55
10	10	6,42	0,39	0,63
20	5	5,52	0,35	0,61
50	1	2,92	0,22	0,55

TAB. 4.4 – *Extrait des résultats des expériences en recherche de traduction parmi les documents de Wikipédia, en utilisant un lexique pivot spécialisé.*

4.2 Vers une approche multivue

Dans cette section, nous étudions les différents aspects relatifs à la graphie, au contexte et aux thèmes permettant de repérer des traductions de termes médicaux. Après avoir mené une série d'expériences sur les ressources utilisées pour l'extraction de lexique bilingue, nous voulons à présent étudier la possibilité de construire ces lexiques. Pour cela, nous implémentons tout d'abord l'approche classique basée sur des vecteurs de contexte présentée dans la section 2.6.3.

Nos études sur ces vecteurs portent sur la taille de la fenêtre permettant de limiter, en nombre de mots, l'environnement d'un terme. Nous voulons capturer l'information se trouvant dans le contexte local d'un terme, mais aussi dans un contexte plus large. Nous pensons que certains termes sont caractérisés par leur environnement proche, alors que d'autres termes le sont plus par un environnement distant.

Nous faisons ainsi référence aux caractéristiques syntaxiques, sémantiques et pragmatiques concernant les termes à traduire. D'autres travaux ont fait l'objet d'études dans la prise en charge des contraintes locales et globales permettant de modéliser le langage, notamment dans la construction de modèles de langage pour la reconnaissance de la parole (Bellegarda, 1998).

Nous émettons ensuite l'hypothèse qu'un terme et sa traduction partagent des similarités d'un point de vue thématique. Nous voulons construire une représentation des termes à traduire dans un espace composé de thèmes. Pour cela, nous modélisons le corpus comparable dans un espace sémantique à l'aide de l'Allocation Latente de Dirichlet (LDA) (Blei et al., 2003). Cette approche est particulièrement adaptée à nos besoins : une représentation sémantique par sacs de mots et des dimensions (des thèmes) indépendantes.

Finalement, pour notre étude sur les paramètres graphiques de termes, nous présentons une série d'expériences d'extraction terminologique par minimisation de la distance de Levenshtein (Levenshtein, 1966) entre des termes de la langue source et de la langue cible.

4.2.1 Les vecteurs de contexte

Afin de représenter d'une manière robuste les contextes de termes, nous voulons évaluer la portée d'une telle représentation, selon les fenêtre d'observations possibles. Pour un terme, nous pouvons nous situer dans l'entourage lexical immédiat, et capturer ainsi des informations limitées à un contexte restreint. Ces informations peuvent être qualifiées d'informations syntaxiques. Élargir la fenêtre d'observation permet ensuite de prendre en considération des informations à différents niveaux, pouvant se rapprocher des collecter de informations sémantiques. Par exemple, couvrir un document entier contenant un terme à traduire permet d'avoir une vue plus large du contexte du terme. Chaque taille de contexte permet alors de représenter une vue pour un terme, dont les performances en terme de modélisation contextuelle et d'extraction de traductions sont évaluées dans les expériences effectuées dans cette section.

Afin de construire un vecteur représentant le contexte d'un terme, les cooccurrences entre ce terme et les mots du lexique pivot dans le corpus sont comptées. Cette tâche pouvant rapidement devenir coûteuse en temps de calcul, notamment à cause de l'accès aux données, nous décidons d'utiliser un outil d'indexation et de recherche d'information : *Indri* issu du projet *Lemur*[4]. A partir d'un ensemble de documents Wikipédia disponibles en ligne, un formatage particulier est nécessaire pour mener

4. http ://www.lemurproject.org/

à bien l'indexation par l'outil utilisé. Nous utilisons le format *Trectext*, basé sur XML (Xtensible Markup Language, qui signifie langage de balisage extensible). Ainsi, tout formatage propre à Wikipédia est retiré du corpus, y compris les liens entre les documents d'une même langue et les liens inter-langues (les liens reliant deux articles entre deux langues). Voici un exemple de document formaté en *Trectext* :

```
<DOC>
<DOCNO> 366451 </DOCNO>
<TITLE> H2G2 : le Guide du voyageur galactique </TITLE>
<TEXT>
  ' h2g2  :  le guide du voyageur galactique '  ou  ...
  ... vers de nouvelles aventures .
</TEXT>
</DOC>
```

Après l'indexation du corpus comparable avec *Lemur*, nous pouvons récupérer des comptes de co-occurrences de manière *quasi* instantanée (de l'ordre de 5 millisecondes pour le corpus en français). Nous avons aussi la possibilité de faire varier la taille de la fenêtre d'observation, *i.e.* le nombre de mots autour d'un terme à traduire. La taille de la fenêtre glissante permet de limiter le contexte d'observation autour d'un terme, et ainsi de capturer les informations de co-occurrences pouvant être locales ou plus globales. En effet, nous estimons qu'un segment sous-phrastique, une phrase, un paragraphe, ou encore le document complet peuvent apporter des informations différentes.

Ces informations basées sur le décompte de co-occurrences sont normalisées par le rapport des chances (mesure *odds ratio*) présenté dans la section 2.6.3. Une fois les vecteurs de contextes construits, nous mesurons le *cosinus* de l'angle formé par deux vecteurs de contexte afin d'en évaluer la similarité. D'autres mesures d'association *terme-mot pivot* et d'autres métriques de similarité vectorielle sont envisageables, cependant nous avons observé de meilleurs résultats avec la combinaison de ces deux méthodes.

En nous basant sur les résultats présentés à la section 4.1.2, et afin de limiter le temps de calcul nécessaire à nos expériences, nous décidons d'utiliser uniquement le lexique spécialisé comme lexique pivot (soit $sLex2_{FR}$ et $sLex2_{EN}$ présentés dans le tableau 4.1). Les candidats à traduire proviennent à nouveau de MeSH (soit $cand2_{FR}$ et leurs références $cand2_{EN}$). Le paramètre que nous faisons varier lors de la construction des vecteurs de contexte est la taille de la fenêtre d'observation autour d'un terme candidat.

Les résultats d'extraction de traductions sont présentés dans le tableau 4.5. Il apparaît très clairement aux vues de ces expériences qu'une taille de contexte de 30 mots autour des termes candidats permet d'obtenir le meilleur rappel, et ceci à tous les rangs d'observations, du premier candidat dans la langue cible, jusqu'aux 100 premiers. Un contexte de la taille du document contenant le terme candidat introduit généralement trop de points d'ancrage incertains entre les langues, et ne permet pas d'obtenir de bons résultats.

	10 mots	20 mots	30 mots	40 mots	document
rang 1	31,1	32,9	33,7	32,4	15,6
rang 10	57,6	59,6	60,6	58,6	37,7
rang 50	69,3	71,8	72,6	71,8	54,6
rang 100	73,4	76,0	76,9	76,6	61,5

TAB. 4.5 – *Scores de rappel pour l'extraction de traductions basée sur des vecteurs de contexte.*

Cependant, nous estimons que certains candidats sont caractérisés par un contexte de petite taille (contexte direct), alors que d'autres le sont par un contexte plus éloigné (contexte indirect). Nous mesurons le nombre de traductions candidates retournées au premier rang par une seule et unique taille de contexte. C'est-à-dire, est-ce qu'il existe des termes et leurs traductions étant uniquement liés par un contexte d'une certaine taille ? Ces résultats sont présentés dans le tableau 4.6.

	10 mots	20 mots	30 mots	40 mots	document
rang 1	4,5	1,7	1,5	1,7	3,5

TAB. 4.6 – *Scores de rappel pour des termes et leurs traductions caractérisés par une seule taille de contexte.*

Ces résultats nous permettent de confirmer l'hypothèse proposée précédemment et nous amènent à penser que ces différents contextes d'observation peuvent être complémentaires. En effet, si certains contextes permettent de retrouver des traductions, le nombre de bonnes traductions ne s'élèvent pas au dessus des 33,7% des termes candidats initiaux. De plus, accroître le nombre de traductions candidates retournées par le système (c'est-à-dire augmenter le rang pour calculer le rappel) fait automatique chuter la précision.

En effet, au premier rang, le système ne retourne qu'un candidat, alors qu'au rang 100, le système a retourné 100 termes dans la langue cible, et un seul peut être la bonne traduction du terme source (selon la référence présente dans la liste de candidats). Nous considérons donc chaque taille de contexte comme un juge, permettant de voter pour un candidat dans la langue cible selon un terme à traduire dans la langue source. Seuls les traductions candidates récoltant les votes à la majorité des juges sont retenues. Si, pour un terme source, aucun terme cible n'a atteint la majorité, le système ne retourne aucun résultat.

Nous effectuons plusieurs combinaisons de taille de contexte, et nous présentons les résultats obtenus dans le tableau 4.7. Ces résultats sont exprimés en pourcentages, d'après les scores de rappel, précision et f-mesure. Ce dernier score est détaillé par l'équation 4.1.

$$\text{f-mesure} = \frac{2 * (\text{précision} * \text{rappel})}{(\text{précision} + \text{rappel})} \tag{4.1}$$

Nous remarquons une baisse conséquente des scores de rappel en fonction du

	30 mots	20+30 mots	10+30+40 mots	tous
rappel	33,7	27,8	19,2	6,2
précision	33,7	50,5	75,9	83,7
f-mesure	33,7	35,8	30,6	11,5

TAB. 4.7 – *Scores pour la combinaison des contextes d'observation.*

nombre de tailles de contextes combinés. Le nombre de bonnes traductions candidates retournées par le système, lorsque nous combinons les tailles de contexte, est donc moins important qu'avec une utilisation individuelle des tailles de contexte. Cependant, si le système retourne un terme dans la langue cible, il est validé par la majorité des juges. Cette méthode nous assure donc une meilleure précision sur les candidats atteignant la majorité des votes, et nous permet d'obtenir 83,7% de bonnes traductions placées au premier rang. Il apparaît toutefois que seulement 6,2% des candidats initiaux sont validés par la majorité des juges lorsque toutes les tailles de contexte sont combinées (de 10 mots jusqu'au document complet). Des extraits de bonnes traductions trouvées au premier rang des résultats retournés par notre système sont présentés dans le tableau 4.8.

TERME SOURCE	TERMES CIBLES
solvants	**solvents**, polymers, oxides, ...
trisomie	**trisomy**, genetic research, monosomy, ...
réticulocytes	**reticulocytes**, fetal blood, transferrin, ...
vaccin coquelucheux	**pertussis vaccine**, bcg vaccine, ...
peptidoglycane	**peptidoglycan**, glycolipids, operon, ...
hypoalbuminémie	**hypoalbuminemia**, proteinuria, ...
paroi thoracique	**thoracic wall**, esophagus, subclavian vein, ...

TAB. 4.8 – *Extrait de correspondances entre les termes sources et cibles selon leurs distances contextuelles, avec un contexte limité à 20 mots, lorsque la bonne traduction est retournée au premier rang par notre système.*

Les tailles de contextes les plus caractéristiques pour des termes peuvent varier au sein d'une langue, mais aussi entre des termes et leurs traductions. Un paramètre permettant de caractériser au mieux un terme par son contexte est donc nécessaire. Une possibilité serait d'agrandir la taille du lexique pivot et donc des vecteurs de contexte.

Nous pensons cependant que le temps de calcul pourrait devenir trop important car augmenter la taille du lexique pivot implique de compter plus de co-occurrences sur l'ensemble du corpus. De ce fait, il paraît lourd en temps de calcul de compter *à la volée* les co-occurrences. Nous estimons qu'une méthode permettant de filtrer les *mauvaises* traductions candidates peut avoir un effet positif sur la précision sans faire chuter le rappel de manière aussi importante. Un modèle permettant de positionner les termes dans un espace thématique pourrait résoudre ce problème.

4.2.2 Le modèle thématique

L'idée générale du modèle thématique est de construire un espace multidimensionnel permettant d'associer des termes dans la langue cible avec des thèmes. Les termes des langues sources et cibles sont modélisés dans cet espace afin d'obtenir leurs positions dans les différentes dimensions (les thèmes). La distribution thématique est indépendante des langues étudiées. Si une traduction candidate ne partage pas de similarité thématique avec un terme dans la langue source, nous pensons que ce n'est pas une traduction. Le schéma général de cette approche est présenté par la figure 4.1. Nous utilisons le corpus comparable détaillé dans le tableau 4.1 afin de construire un modèle thématique.

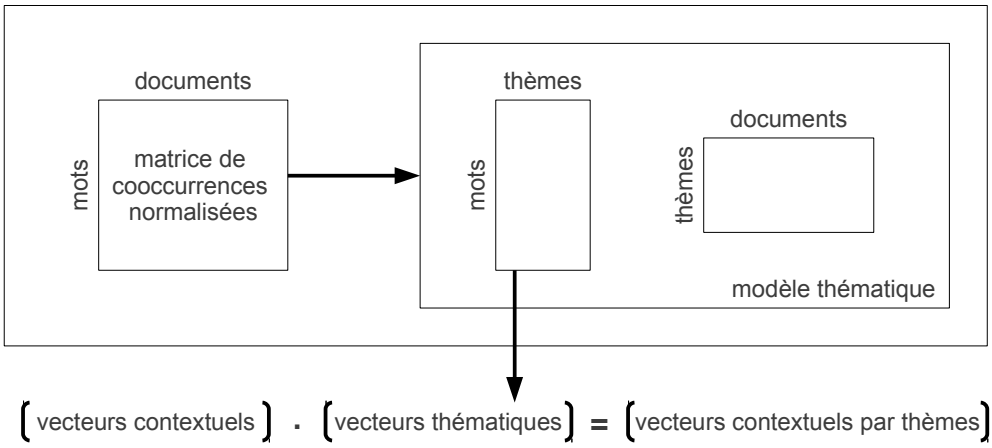


FIG. 4.1 – *Intégration des informations issues du modèle thématique dans les vecteurs de contexte, pour donner une représentation du contexte dans chaque dimension de l'espace thématique.*

D'après la description faite dans la section 2.6.2 des méthodes d'analyse de la sémantique latente, nous utilisons l'approche d'Allocation Latente de Dirichlet (*LDA* pour *Latent Dirichlet Allocation*) dans un contexte monolingue. Pour la partie du corpus comparable dans la langue source, nous réduisons le vocabulaire à celui contenu dans le lexique pivot. De ce fait, nous construisons un espace sémantique pouvant être projeté de la langue source à la langue cible. Cette méthode nous permet d'avoir des dimensions (ou thèmes) alignés entre les langues.

Ainsi, pour chaque terme source ou cible, nous estimons sa distance à chaque

thème composant notre modèle. Le résultat est une liste ordonnée de thèmes associée à chaque terme source et cible. Ces listes pouvant être comparées entre langues, nous voulons mesurer les similarités thématiques existant entre un terme et sa traduction. La comparaison inter-langue des associations entre les termes et les thèmes est mise en œuvre avec l'algorithme 3.

Algorithm 3 Association des termes à traduire avec les dimensions du modèle thématique et comparaison thématique interlangue.

$cand_{src_i}$: terme à traduire provenant de $cand2_{FR}$ (langue source)
$cand_{trg_j}$: référence du terme à traduire ($cand2_{EN}$)
$theme_{src_k}$: thème source issu du modèle thématique $theme_{src}$
$theme_{trg_l}$: thème cible issu du modèle thématique $theme_{trg}$
$dist(cand_{src_i}, theme_{src_k})$: distance entre $cand_{src_i}$ et $theme_{src_k}$
$dist(cand_{trg_j}, theme_{trg_l})$: distance entre $cand_{trg_j}$ et $theme_{trg_l}$
$list\_theme_{src_i}$: liste ordonnée des distances aux thèmes sources pour $cand_{src_i}$
$list\_theme_{trg_j}$: liste ordonnée des distances aux thèmes cibles pour $cand_{trg_j}$
nb_theme_{src} : nombre de thèmes sources considérés
$list\_trans_{src_i}$: liste des traductions candidates pour $cand_{src_i}$
for $cand_{src_i} \in cand2_{FR}$ **do**
  for $theme_{src_k} \in theme_{src}$ **do**
    Calculer $dist(cand_{src_i}, theme_{src_k})$
    Ajouter $dist(cand_{src_i}, theme_{src_k})$ à $list\_theme_{src_i}$
  end for
end for
for $cand_{trg_j} \in cand2_{EN}$ **do**
  for $theme_{trg_l} \in theme_{trg}$ **do**
    Calculer $dist(cand_{trg_j}, theme_{trg_l})$
    Ajouter $dist(cand_{trg_j}, theme_{trg_l})$ à $list\_theme_{trg_j}$
  end for
end for
for $cand_{src_i} \in cand2_{FR}$ **do**
  for $cand_{trg_j} \in cand2_{EN}$ **do**
    if $list\_theme_{trg_j}[1] \cap list\_theme_{src_i}[nb\_theme_{src}]$ **then**
      Ajouter $cand_{trg_j}$ à $list\_trans_{src_i}$
    end if
  end for
end for

Afin de calculer la distance entre un terme et un thème, nous utilisons la mesure d'association entre deux unités lexicales utilisée dans la section 4.2.1. Les deux unités sont, pour la langue source par exemple, le terme candidat (soit $cand_{src_i}$ issu de $cand2_{FR}$) et un mot w_j (issu de $sLex2_{FR}$) présent dans le thème courant (soit z_k). La mesure d'association est pondérée par la probabilité $p(w_j|z_k)$, et l'équation 4.2 en donne les détails.

Cette mesure est calculée pour chaque mot du thème, puis nous faisons varier un

seuil sur le nombre de mots à sélectionner par thème. En effet, si nous ordonnons les mots par leurs probabilités d'appartenance à un thème, certains ont un *poids* très fort et caractérisent donc cette dimension, mais d'autres sont moins représentatifs. De plus, avec la méthode utilisée, l'ensemble du vocabulaire présent dans lexique pivot se retrouve dans chaque thème. Nous estimons donc qu'un nombre limité de mots est nécessaire afin de calculer la distance d'un terme à un thème.

Ainsi, un premier paramètre que nous voulons étudier est le nombre de dimensions du modèle thématique. La variation de la *taille* de notre espace permet d'influer directement sur les probabilités associées à chaque mot, et ceci pour chaque thème. Ne sachant pas au préalable quel est le nombre idéal de thèmes abordés par Wikipédia, nous estimons qu'un seuil minimal de 20 et maximal de 200 apporte différentes visions du corpus selon la granularité choisie. Des extraits de mots présents dans les thèmes sont présentés dans le tableau 4.9.

Thème 1		Thème 2		Thème 3		Thème 4	
art	0.34	air	0.26	moteur	0.10	russie	0.27
peintre	0.11	force	0.14	arrière	0.05	roumanie	0.11
peinture	0.09	aéroport	0.13	puissance	0.05	finlande	0.11
salon	0.04	direction	0.02	vitesse	0.04	moscou	0.10

TAB. 4.9 – *Extraits de thèmes issus du modèle construit avec LDA, où chaque mot est associé à la probabilité conditionnelle de le rencontrer dans le thème courant.*

Un second paramètre est le nombre de thèmes à conserver pour chaque terme dans la langue source et cible. En effet, si nous déterminons le thème le plus proche pour un terme à traduire, sa traduction n'a pas forcément l'équivalent de ce thème en première position dans sa liste. Nous décidons donc de faire varier le nombre de thèmes dans la langue cible entre 1 et 3. Il est cependant évidant qu'en observant les trois premiers thèmes des termes dans la langue cible, beaucoup de ces derniers auront des similarités thématiques avec un terme dans la langue source. Pour limiter l'impact sur la précision du système, nous décidons de garder un seul thème par terme dans la langue source.

$$d(cand_{src_i}, z_k) = \sum_j odds(cand_{src_i}, w_j) p(w_j|z_k) \tag{4.2}$$

Nous présentons les résultats de la comparaison thématique inter-langue dans le tableau 4.10, en étudiant deux paramètres : le nombre de dimensions de l'espace et le nombre de mots retenus par thème pour calculer la distance entre un terme et un thème. Les précisions mesurées sont naturellement très basses, ceci étant dû au grand nombre de termes langue cible à la même position dans l'espace thématique qu'un terme langue source. Cet aspect est lié au domaine de spécialité, car l'ensemble du corpus est utilisé pour construire le modèle initial.

Or, si les termes à traduire sont dans le domaine médical, seuls quelques thèmes du modèle seront concernés. De ce fait, un modèle propre au domaine doit être construit afin d'apporter plus de nuance entre les thèmes, tout en restant dans le

DIMENSIONS :		20	50	100	200
1 mot	rappel	23,6	33,4	33,0	10,4
	précision	0,12	0,06	0,07	0,18
	f-mesure	0,23	0,12	0,14	0,35
2 mots	rappel	35,9	42,2	38,3	18,8
	précision	0,1	0,06	0,07	0,15
	f-mesure	0,19	0,12	0,14	0,31
3 mots	rappel	44,1	45,9	41,2	24,1
	précision	0,08	0,06	0,07	0,14
	f-mesure	0,16	0,12	0,13	0,28

TAB. *4.10* – *Scores pour l'extraction de traductions basée sur les thèmes.*

domaine de spécialité. Afin d'illustrer cet aspect, nous détaillons dans le tableau 4.11 quelques exemples de termes dans les langues source et cible partageant des similarités thématiques selon notre approche.

TERME SOURCE	TERMES CIBLES
plasminogène	**plasmin**, glycolipids, pneumoperitoneum, ...
ulcère cutané	**skin ulcer**, hypoalbuminemia, fluphenazine, ...
cyprotérone	**cyproterone**, thrombasthenia, ageusia, ...
granulocytes éosinophiles	nipah virus, shallots, patellar ligament, ...

TAB. *4.11* – *Extrait des termes sources et cibles partageant des similarités thématiques. Les bonnes traductions sont indiquées en gras.*

4.2.3 Les cognats

Comme nous l'avons présenté dans la section 2.6.1, il est parfois possible d'extraire des traductions en se basant sur l'orthographe des termes. Nous estimons que les termes français et anglais dans le domaine médical sont particulièrement adaptés à ce genre d'approches, car ces deux langues partagent le même alphabet. Il est donc fort probable que des termes de domaines spécialisés aient une racine commune, les préfixes et les suffixes pouvant être différents afin de respecter les contraintes de chaque langue. Cette particularité peut être vue comme une relation très forte existant entre des termes ayant une origine commune, mais ayant pu évoluer indépendamment dans chaque langue.

De plus, il apparaît parfois que des termes français soient *empruntés* par l'anglais, et inversement. Nous sommes alors en présence de *translittérations*. Une définition de la relation de translittération est donnée dans Prochasson (2009) : "Deux mots $w1$ et $w2$ de langue $l1$ et $l2$ sont en relation de translittération s'ils sont traductions l'un de l'autre, s'ils sont phonétiquement proches et dans des systèmes d'écriture différents." Ce phénomène est d'autant plus observable dans des domaines où les néologismes sont fréquents, comme en informatique par exemple (le terme *Internet* illustre cet aspect). Les noms propres sont souvent des translittérations entre différentes langues.

Nous pensons donc qu'il est possible d'extraire des traductions de termes médicaux en s'appuyant sur leurs orthographes. Afin de comparer les termes selon cet aspect, nous utilisons la distance d'édition populaire appelée *distance de Levenshtein*, présentée dans la section 2.6.1. Chaque couple de termes obtient un score concernant leur éloignement d'un point de vue orthographique, et nous permet d'ordonner les candidats à la traduction. Plus le score est bas, plus l'orthographe des termes est similaire, plus les termes du couple sont des traductions possibles.

Nous mesurons les pourcentages de rappel, de précision et de f-mesure à plusieurs rangs (de 1 à 10) dans la liste ordonnée de traductions candidates. Nous effectuons le calcul de la distance d'édition entre les termes complets mais aussi entre les 4 premières lettres des termes, comme présenté dans les travaux de Simard et al. (1993). Les résultats de ces expériences sont détaillés dans le tableau 4.12.

		rang 1	rang 2	rang 3	rang 4	rang 5	rang 10
4 lettres	rappel	34,0	39,0	45,9	65,4	100	100
	précision	15,9	3,9	0,5	0,1	0,03	0,03
	f-mesure	21,7	7,2	1,0	0,2	0,07	0,07
termes	rappel	50,7	54,8	59,6	67,4	77,3	99,3
	précision	83,5	29,6	5,6	1,4	0,5	0,2
	f-mesure	63,1	38,5	10,3	2,8	1,0	0,3

TAB. 4.12 – *Résultats de l'approche basée sur les cognats pour l'extraction de traductions.*

Pour chaque rang, le score de rappel est incrémenté si la bonne traduction est trouvée. La précision est mesurée selon le nombre de bonnes traductions retournées par le système, divisé par le nombre total de résultats retournés. Ce score chute très rapidement lorsque le rang d'observation augmente, car beaucoup de termes dans la langue cible sont retournés par le système. Le tableau 4.13 contient plusieurs exemples de termes anglais et français alignés selon leur distance de Levenshtein.

4.2.4 La combinaison de vues

Dans cette section, nous voulons combiner les trois approches présentées précédemment, c'est-à-dire les vecteurs de contexte, l'approche thématique et les cognats. De ce fait, nous combinons plusieurs représentations des termes de chaque langue selon des vues différentes. Nous estimons que ces approches sont

TERME SOURCE	TERMES CIBLES
globines	**globins**
solvants	**solvents**
quadriplégie	**quadriplegia**, quadruplets
caryotype	cardiomegaly, carotid stenosis, ...
flux génétique	flurazepam, fluorine compounds, flumazenil, ...
lécithines	**lecithins**, leeching, ...
trisomie	**trisomy**

TAB. *4.13* – *Extrait de correspondances entre les termes sources et cibles selon leurs distances de Levenshtein. Les bonnes traductions sont indiquées en gras.*

complémentaires, et trois combinaisons sont étudiées. La première, déjà présentée dans la littérature, consiste à combiner les vecteurs de contexte avec la comparaison orthographique des termes. La seconde est la combinaison du contexte et des thèmes, permettant de filtrer les traductions candidates n'ayant pas de similarités thématiques avec un terme à traduire. La dernière combinaison étudiée est l'association des trois approches, donnant lieu à notre approche multivue.

Chaque combinaison est effectuée en utilisant un vote entre les approches. Si deux approches sont combinées, le résultat final est issu d'un vote à l'unanimité. Si les trois vues sont combinées, deux résultats sont possibles : la majorité absolue et l'unanimité. Comme présenté dans la section 4.2.1 pour les vecteurs de contexte, plusieurs tailles de contexte peuvent être combinées. Nous avons sélectionné la meilleure configuration (c'est à dire un contexte de 30 mots), mais aussi la combinaison de toutes les tailles de contexte, où chaque taille correspond à un juge. Pour chaque combinaison de vues, si aucune des conditions de vote n'est remplie, le système ne retourne aucune traduction candidate. Les résultats de l'approche multivue sont présentés dans le tableau 4.14, selon le rappel, la précision et la f-mesure (exprimés en pourcentages).

Nous remarquons que le vote à l'unanimité pour la combinaison multivue permet d'atteindre des scores de précision de 100% mais dégrade fortement les scores de rappel. Ainsi, pour utiliser les lexiques bilingues construit automatiquement, cette configuration permet d'avoir une bonne confiance dans les résultats retournés par notre système. De manière générale, lorsque l'approche par similarité contextuelle est utilisée selon un contexte limité à 30 mots, les scores de précision sont élevés et les scores

		contexte+thème	contexte+cognats	multivue
	rappel	13,9	19,0	24,2 | 7,6
30 mots	précision	100	99,1	99,3 | 100
e	f-mesure	24,4	31,9	39,0 | 14,1
	rappel	20,8 | 2,6	21,1 | 4,0	26,9 | 1,7
tous contextes	précision	76,2 | 100	76,6 | 97,6	80,4 | 100
	f-mesure	32,7 | 5,0	33,1 | 7,7	40,3 | 3,4

TAB. 4.14 – *Résultats de l'approche multivue selon différentes combinaisons. La première sous-colone indique les résultats des votes à la majorité. La seconde sous-colonne indique les résultats des votes à l'unanimité.*

de rappel sont bas. Peu de traductions sont donc retrouvées mais lorsque le système propose un résultat, celui-ci s'avère être bon à plus de 99%.

Lorsque différentes tailles de contextes sont utilisées comme des juges indépendant dans le vote de l'approche multivue, les résultats en terme de f-mesure sont plus élevés que ceux obtenus avec une seule taille de contexte, selon les votes à la majorité absolue. En comparant les résultats entre une seule et plusieurs tailles de contexte, les scores de précision sont inférieurs et les scores de rappels sont supérieurs. Les meilleurs résultats sont obtenus avec la combinaison multivue incluant toutes les tailles de contexte.

4.3 Un modèle génératif à portées continues

Nous avons proposé précédemment d'étendre l'approche basée sur les vecteurs de contexte en l'associant avec deux autres approches. Il apparaît toutefois que cette représentation reste limitée par la taille du contexte d'observation, et que ce paramètre ne semble pas identique pour tous les termes, selon les résultats présentés dans la table 4.6. De plus, si intuitivement un contexte large devrait permettre d'accroître la quantité d'information contenue dans un vecteur, nous remarquons une baisse sur l'ensemble des résultats issus des expériences utilisant un document entier. Inclure l'ensemble des contextes d'observation est possible si les distances entre un terme et un mot du lexique pivot sont aussi incluses. L'idée générale est de garder le contexte le plus large pour chaque occurrence d'un terme au sein d'un corpus, tout en intégrant les distances à chaque élément du lexique.

Ainsi, trois éléments principaux motivent notre approche de modèle génératif à portées continues :

- modéliser individuellement les contextes d'un terme à traduire en les considérant comme des évènements disjoints,
- éviter la représentation en sac de mots des contextes, s'orienter vers des contextes à portée continue,
- considérer les informations locales et globales, d'un point de vue syntaxique, sémantique et pragmatique.

Nous voulons de ce fait modéliser la probabilité d'apparition d'un terme T dans un contexte C, permettant par la suite d'inférer un terme inconnu à partir d'un contexte observé selon la théorie de Bayes : $p(C|T) = \frac{p(T|C).p(C)}{p(T)}$. Cette inférence doit pouvoir s'effectuer dans entre les langues afin de mettre en relation des termes étant des traductions. Nous proposons donc dans cette section un nouveau modèle qui repose sur des matrices composées des vecteurs de distances. Contrairement aux vecteurs de contexte, les valeurs d'association entre un terme et le lexique bilingue servant de pivot entre les langues ne sont pas *moyennées* en étant soumise à une métrique basée sur les co-occurrences. Chaque matrice est associée à un terme dans la langue source ou cible. La taille des matrices est fixe en nombre de colonnes : la taille du lexique pivot. Le nombre de lignes est cependant variable selon le nombre d'occurrences du terme dans le corpus.

Nous détaillons dans un premier temps l'inclusion des distances dans un vecteur de contexte (section 4.3.1), puis dans un second temps, la construction des matrices de distances (section 4.3.2). Nous présentons ensuite la méthode de comparaison inter-langue des modèles, afin de repérer les couples qui sont en relation de traduction (section 4.3.3). Le protocole expérimental mis en place afin de mener à bien nos expérimentations est alors détaillé (section 4.3.4), suivi par les résultats obtenus selon les différentes configurations testées (section 4.3.5).

4.3.1 Vecteurs de distances

Afin de considérer toutes les occurrences d'un terme dans le corpus, nous devons considérer et modéliser tous les contextes possibles, c'est-à-dire un contexte pour une observation du terme initial. La modélisation d'un contexte pour un terme donné peut se faire grâce à un vecteur. Les mots du lexique pivot permettent de définir les dimensions de ce vecteur, comme dans la méthode des vecteurs de contexte présentée dans la section 2.6.3. Chaque composante du vecteur de distances est constituée de la distance entre le terme à traduire et un mot du lexique pivot. Cette distance est le nombre de mots *inconnus*, c'est à dire que l'on ne sait pas traduire, entre le terme et le mot du lexique. Puis, nous appliquons une règle de normalisation afin de lisser les valeurs de distances obtenues.

La fenêtre d'observation des cooccurrences est fixée au document entier. Lorsqu'un mot du lexique n'est pas dans le document, la composante correspondante du vecteur contient une valeur négative permettant de représenter l'absence d'un mot dans le contexte d'un terme. L'équation 4.3 détaille la construction d'un vecteur de distances, avec L représentant le lexique bilingue servant de pivot entre les langues, $V[k]$ est la dimension k du vecteur de distances V, $dist$ est la distance en nombre de mots et C_t est le contexte du terme t à traduire, ayant f occurrences dans le corpus. L'équation 4.4 formalise la méthode de normalisation des distances.

$$\forall k \in L\ , V[k] = \begin{cases} dist(t, L[k]) & \text{if } L[k] \in C_t \\ -1 & \text{else} \end{cases} \tag{4.3}$$

$$d(t, L[k]) = \frac{1 + log(10)}{1 + log(dist(t, L[k]))} \tag{4.4}$$

Pour chaque occurrence d'un terme dans le corpus, un vecteur de distances est construit. La méthode de construction de ce type de vecteurs est illustrée par un exemple simplifié sur la figure 4.2. Un terme est donc modélisé par un ensemble de vecteurs contenant les distances aux mots du lexique pour chaque occurrence dans le corpus. Cette représentation nous permet d'élaborer un modèle basé sur les distances, regroupées dans une matrice.

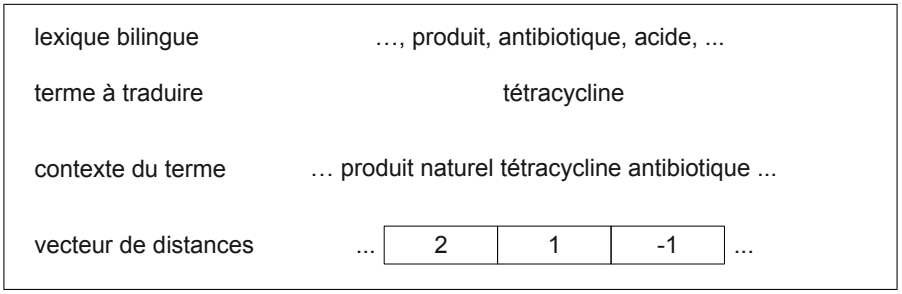


FIG. 4.2 – *Exemple simplifié de la construction d'un vecteur contenant les distances entre un terme à traduire et les mots du lexique pivot.*

4.3.2 Matrices de distances

L'ensemble de vecteurs de distances, correspondant à chaque occurrence d'un terme à traduire, nous permet de construire un modèle pour chaque terme sous la forme d'une matrice de distances. Cette matrice est constituée des vecteurs associés à un terme. Ainsi, un terme dont le nombre d'occurrence dans le corpus est élevé, est associé à un nombre important de vecteurs. Au contraire, un terme peu fréquent a un nombre réduit de vecteurs modélisant ses contextes. La construction des vecteurs, décrite par l'équation 4.3, permet de capturer les informations relatives aux distances *termes-mots du lexique* si les mots du lexique se trouvent dans les documents où apparaissent les termes à traduire.

Si peu de mots du lexique sont dans les contextes des termes, les matrices de distances résultantes sont creuses. De ce fait, trop peu d'information s'y trouve pour permettre de caractériser les termes. Afin de palier ce problème, nous estimons qu'une réduction du nombre de dimensions de la matrice de distances initiale peut permettre de prendre en compte uniquement certaines valeurs caractéristique pour un terme donné.

Selon la méthode de comparaison inter-langues, nous pensons que la réduction du nombre de dimensions des matrices de distance peut aider à minimiser les temps de calcul, permettant de retrouver la traduction d'un terme dans la langue source parmi une liste de candidats dans la langue cible. Nous proposons donc de ne garder des matrices initiales que les valeurs les plus caractéristiques, en faisant varier un paramètre : le nombre de dimensions. Nous effectuons une décomposition en valeurs singulières des matrices initiales, dans le but d'enlever les valeurs nulles et de réduire la taille des matrices.

La matrice diagonale issue de la factorisation d'une matrice de distances et comportant les valeurs singulières pourrait permettre d'effectuer des comparaisons matricielles inter-langues. Cependant, il ne nous paraît pas judicieux d'effectuer des comparaisons entre les modèles des termes sources et cibles à partir de ces matrices projetées dans un espace réduit. Nous proposons plutôt de construire un modèle génératif, en restant dans la philosophie des travaux en modélisation thématique multilingue. Ainsi, des contextes non observés de termes pourraient être associés à des modèles, grâce à une étape d'inférence. Ceci nous permet d'effectuer des comparaisons inter-langues entre un modèles et des contextes brutes.

Une modélisation correspondant à notre idée est possible avec les modèles de mélanges gaussiens, pouvant être estimés dans l'espace réduit contenant les valeurs singulières des matrices de distances. Un modèle de mélanges gaussiens est un modèle probabiliste construit sur la somme des fonctions gaussiennes de densités pondérées. L'équation 4.5 présente le calcul d'un modèle GMM, où w_i sont les poids associés à chaque composant gaussien, et $g(x|\mu_i, \Sigma_i)$ représente les densités de probabilités sous la forme d'une fonction gaussienne.

$$\forall x \in \mathbb{R}\,;\, p(x) = \sum_{i=1}^{M} w_i\, g(x|\mu_i, \Sigma_i) \tag{4.5}$$

4.3.3 Comparaison inter-langues

Grâce au modèle GMM construit pour chaque terme médical dans la langue source ou cible, nous voulons déterminer à quel terme se rapproche le plus un ou plusieurs contextes n'ayant pas été observés auparavant. Dans notre cas, nous connaissons un terme à traduire, mais pas sa traduction. Le modèle est donc construit sur les données observées, c'est-à-dire les contextes d'un terme dans la langue source dont nous cherchons la traduction, et l'inférence est faite sur les données non observées, c'est-à-dire des contextes de termes dans la langue cible. C'est cet aspect permettant l'inférence de nouvelles données qui fait de notre modèle basé sur les matrices de distances un modèle génératif.

Afin d'associer un terme source aux contextes non observés dans la langue cible, nous avons besoin d'une mesure permettant d'évaluer la vraisemblance entre un modèle et des données brutes. Nous nous inspirons des travaux en reconnaissance de

la parole effectués par (Reynolds et al., 1998), où les auteurs effectuent des mesures de vraisemblance entre des données de parole brutes issus d'un locuteur et un modèle GMM construit au préalable et caractérisant un locuteur. Ils proposent d'utiliser la mesure symétrique de vraisemblance croisée, en prenant en compte différents paramètres permettant de mesurer l'adéquation entre :

- un modèle source et des données cibles,
- un modèle cible et des données sources,
- un modèle source et des données sources,
- un modèle cible et des données cibles.

L'équation 4.6 présente le calcul de la vraisemblance croisée (où l représente la fonction de vraisemblance, m et c étant les modèles GMM et les contextes brutes respectivement) utilisée dans nos travaux afin de comparer les modèles dans un contexte inter-langue.

$$VC(m_s, m_t) = log\left(\frac{l(c_s|m_s)}{l(c_s|m_t)}\right) + log\left(\frac{l(c_t|m_t)}{l(c_t|m_s)}\right) \tag{4.6}$$

Ainsi, chaque couple de termes source et cible est associé à une mesure de vraisemblance croisée. D'une manière générale, nous estimons la proximité d'un contexte dans une langue avec un GMM dans une autre langue, tout en prenant en compte la proximité de ce GMM avec les contextes issus de la même langue. L'alignement des termes sources et cibles (soit T^s et T^t respectivement) s'effectue donc d'après les scores de vraisemblances obtenus entre les contextes sources et cibles (soit c_s et c_t) et les modèles GMM sources et cibles (m_s et m_t).

4.3.4 Protocole expérimental

Pour évaluer l'approche par matrices de distances, nous voulons aligner un ensemble de termes dans la langue source avec leurs traductions dans la langue cible. Un échantillon de 200 termes tests sont extraits du thésaurus MeSH, accompagnés d'une traduction de référence. Pour chaque terme source T^s et cible T^t, un ensemble de contextes c_{T^s} et c_{T^t} sont collectés séparément à partir de Wikipédia. Un contexte correspond à un document dans lequel apparaît un terme. Nous construisons un modèle pour chaque terme, correspondant aux descriptions faites dans la section 4.3.2. Les modèles GMM construits sur les matrices de distances réduites sont notés m_{T^s} et m_{T^t}.

Pour chaque matrice de distances, nous calculons la vraisemblance croisée entre la source et la cible selon la description faite dans la section 4.3.3. Dans un premier temps, nous considérons ces scores individuellement, c'est à dire sans comparer les vraisemblances obtenues entre un modèle source m_{T^s} et l'ensemble des modèles cibles m_{T^t}. Dans un second temps, nous considérons l'ensemble des scores de vraisemblances obtenus pour un modèle source m_{T^s}, et nous décidons d'ignorer les résultats retournés par le système dans deux cas :

- les vraisemblances pour un modèle source m_{T^s} et tous les modèles cibles sont trop proches (selon un seuil), ce qui indique que le modèle source ne contient pas d'information caractéristiques du terme source correspondant,
- la vraisemblance entre un modèle cible m_{T^t} et tous modèles sources m_{T^s} sont particulièrement élevées (selon un seuil), ce qui indique que le modèle m_{T^t} est trop *générique*, correspondant à un effet de sur-apprentissage.

Nous faisons varier deux paramètres dans la construction des modèles basés sur les matrices de distances : le nombre de dimensions lors de la décomposition en valeurs singulières, et le nombre de gaussiennes dans le modèle GMM construit sur ces valeurs singulières.

D'une manière générale, et pour les deux types de modèles, pour un terme source T^s, l'ensemble des termes cibles T^t sont évalués, et une liste ordonnée de candidats à la traduction est retournée par notre système selon les scores obtenus (similarité vectorielle pour les vecteurs de distances, vraisemblance pour les matrices de distances). Nous mesurons ainsi à différents rangs de cette liste les scores de précision, rappel et f-mesure.

Nous intégrons un aspect relatif à l'utilisabilité de notre approche pour l'extraction de terminologie bilingue dans les scores de rappel mesurés. En effet, nous mesurons le rappel en divisant le nombre de candidats pris en charge par notre approche par le nombre de candidats initiaux, c'est à dire 200. Ainsi, nous voulons mesurer la capacité de notre système à gérer la quantité de données variable associées les candidats testés. Ces données sont relatives aux fréquences d'apparition des termes dans le corpus, mais aussi aux mots du lexique pivot se trouvant dans les contextes des termes.

Le score de précision indique la quantité de bonnes traductions retournées par notre approche selon le nombre de candidats pris en charge. Nous voulons, de cette manière, évaluer uniquement les capacités du système à extraire des bonnes traductions, sans prendre en compte les problèmes liés à la quantité de données associées à chaque terme à traduire. La f-mesure est calculée selon l'équation 4.1, déjà utilisée dans la section 4.2.1.

4.3.5 Expériences et résultats

Afin de mesurer l'impact du nombre de dimensions des matrices réduites et du nombre de gaussiennes dans les modèles GMM, nous faisons varier ces deux paramètres. Le nombre de dimensions testé varie entre 4 et 64. Le nombre de gaussiennes varie lui aussi entre 4 et 64, et ceci pour chaque nombre de dimensions. Les tables 4.15, 4.16 et 4.17 contiennent l'ensemble des résultats obtenus, en terme de rappel, précision et f-mesure. Trois rangs d'observations sont considérés pour effectuer ces mesures : $Top1$, correspondant au premier résultat retourné par le système pour chaque terme testé, $Top10$, où nous prenons en compte les 10 premiers résultats retournés, et $Top20$, pour les 20 premiers résultats.

Espace réduit à 4 et 8 dimensions

Dimensions		4					8				
Gaussiennes		4	8	16	32	64	4	8	16	32	64
Top1	rappel	22,5	17,5	13,0	10,5	9,5	17,5	17,5	15,5	11,5	9,0
	précision	25,6	26,1	29,2	33,3	42,2	21,1	28,4	33,3	37,7	43,9
	f-mesure	**23,9**	21,0	18,0	16,0	15,5	19,1	**21,7**	21,2	17,6	14,9
Top10	rappel	25,5	21,5	17,0	13,0	11,5	24,0	22,5	17,0	14,0	11,5
	précision	29,0	32,1	38,2	41,3	51,1	28,9	36,6	36,6	45,9	56,1
	f-mesure	**27,1**	25,7	23,5	19,8	18,8	26,2	**27,9**	23,2	21,5	19,1
Top20	rappel	28,5	24,0	18,0	13,5	12,0	26,0	22,5	19,0	14,0	11,5
	précision	32,4	35,8	40,4	42,9	53,3	31,3	36,6	40,9	45,9	56,1
	f-mesure	**30,3**	28,7	24,9	20,5	19,6	**28,4**	27,9	25,9	21,5	19,1

TAB. *4.15* – *Résultats du modèle génératif sur les matrices de distances réduites aux dimensions de 4 et 8.*

Pour les matrices réduites à 4 et 8 dimensions (voir table 4.15), nous obtenons les meilleurs scores de rappel pour un nombre de gaussiennes réduit. C'est avec une matrice réduite de 4 dimensions et un modèle GMM composé de 4 gaussiennes qui permet d'atteindre le meilleur score de rappel, et ceci sur les trois rangs d'observation. Augmenter le nombre de dimensions (de 4 à 8) fait baisser ce score, car le système est incapable de proposer des traductions candidates.

Cette dégradation des scores de rappel est liée aux termes qui n'occurrent pas fréquemment dans le corpus initial. En effet, peu de contextes modélisés dans la matrice des distances impliquent un manque d'informations dans la matrice des valeurs singulières, donc un GMM ne modélisant pas les caractéristiques du terme initial. Augmenter le nombre de dimensions ne permet donc pas d'accroître la quantité d'informations caractéristiques d'un terme, mais va empêcher la modélisation par GMM si le nombre de gaussiennes est trop élevé. Notre système est donc incapable de proposer des candidats à la traduction.

Ce nombre réduit de candidats dans la langue cible proposé par le système permet cependant d'augmenter la précision générale de l'approche. Nous pouvons remarquer dans ces premiers résultats que les scores de précision augmentent si le nombre de dimensions des matrices et le nombre de gaussiennes augmentent. Si le système est incapable de fournir des candidats pour l'ensemble des termes testés, le nombre de bonnes traductions au niveau des rangs observés sur l'ensemble des candidats retournés augmente, jusqu'à atteindre la précision de 56,1% pour les rangs 10 et 20, avec 8 dimensions et 64 gaussiennes. Cette croissance de la précision ne permet cependant pas de contrebalancer la chute importante des scores de rappel, ce qui provoque une baisse des résultats en terme de f-mesure.

D'une manière générale, le rappel est dégradé et la précision est améliorée lorsque, pour un nombre de dimensions fixes, le nombre de gaussiennes du modèle GMM augmente. Nous observons une dégradation des scores de la f-mesure suivant les

scores de rappel. Aux rang 1 et 20 (Top1 et Top20), la meilleure f-mesure est obtenue avec 4 dimensions et 4 gaussiennes pour le GMM. Pour le rang 10 (Top10), c'est la configuration basée sur 8 dimensions et 8 gaussiennes qui permet d'obtenir les meilleurs scores de f-mesure.

Espace réduit à 16 et 32 dimensions

Nous décidons d'augmenter le nombre de dimensions de l'espace réduit, et de tester deux nouvelles configurations, à 16 et 32 dimensions. Le nombre de gaussiennes reste identique aux expériences précédentes, compris entre 4 et 64. Les résultats sont contenus dans la table 4.16. Nous observons à nouveau une baisse du rappel lorsque le nombre de gaussiennes augmente pour un nombre de dimensions donné, tandis que les scores de précision augmentent pour la même configuration.

Dimensions		16					32				
Gaussiennes		4	8	16	32	64	4	8	16	32	64
Top1	rappel	17,0	15,0	15,5	10,5	9,5	12,0	13,0	12,5	9,5	7,0
	précision	18,9	28,0	34,4	36,8	48,7	14,0	21,3	33,3	35,2	43,7
	f-mesure	17,9	19,5	**21,4**	16,3	15,9	12,9	16,1	**18,2**	15,0	12,1
Top10	rappel	20,5	17,0	16,5	13,5	11,0	12,0	13,0	12,5	9,5	7,0
	précision	22,8	31,2	36,7	47,4	56,4	17,0	25,4	36,0	40,7	53,1
	f-mesure	21,6	22,1	**22,8**	21,0	18,4	15,6	19,2	**19,6**	17,3	14,6
Top20	rappel	21,5	17,5	17,0	13,5	11,0	16,0	16,0	14,0	11,5	9,0
	précision	23,9	32,7	37,8	47,4	56,4	18,7	26,2	37,3	42,6	56,2
	f-mesure	22,6	22,8	**23,4**	21,0	18,4	17,2	19,9	**20,4**	18,1	15,5

TAB. *4.16* – *Résultats du modèle génératif sur les matrices de distances réduites aux dimensions 16 et 32.*

Cependant, contrairement aux résultats présentés précédemment, les scores de f-mesure sont les plus élevés avec 16 gaussiennes composant les GMM, selon les trois rangs étudiés. Nous remarquons cependant que les résultats en terme de f-mesure sont plus bas que ceux obtenus dans les configurations à 4 et 8 dimensions, et ceci malgré une légère amélioration de la précision (56,4% au rang 10). Nous pensons que cette baisse est directement liée aux scores de rappel ayant été dégradés par l'augmentation du nombre de dimensions dans la transformation de la matrice de distances.

Augmenter la taille de l'espace réduit permettant de modéliser la matrice de distances et augmenter le nombre de gaussiennes des GMM permet d'atteindre les résultats les plus élevés, d'après les scores de f-mesure, et ceci pour tous les rangs observés. Comme les scores de rappel chutent en suivant l'augmentation des dimensions, globalement moins de termes sont modélisés suivant leurs contextes par l'approche basée sur les GMM. Il reste cependant un nombre de termes pris en charge lors de l'augmentation des dimensions, et ils permettent une modélisation plus fine des distances intra-contextuelles, ce qui donne lieu à des scores de précision plus élevés que ceux obtenus avec un nombre plus petit de dimensions.

Si nous comparons les résultats en terme de rappel selon le nombre de dimensions compris entre 4 et 32, nous remarquons que, d'une manière générale, moins de dimensions permet de construire des modèles GMM pour plus de termes candidats, donc de couvrir plus de termes médicaux. Nous observons au premier rang (Top1), pour les meilleurs configurations dans le nombre de gaussiennes, une baisse de 9,5% de rappel entre des matrices de distances réduites à 4 et 32 dimensions.

Espace réduit à 64 dimensions

La table 4.17 contient la série de résultats obtenus suivant la dernière configuration testée, c'est à dire avec 64 dimensions pour la décomposition des matrices de distances. Le nombre de gaussiennes par GMM reste encore une fois identique aux différentes configuration déjà mises en place (entre 4 et 64). Si l'aspect relatif à l'augmentation des scores de f-mesure selon le nombre de dimensions et de gaussiennes est ici vérifié, car la combinaison de 64 dimensions avec 32 gaussiennes permet d'atteindre une f-mesure de 13,5% au premier rang, nous observons une dégradation générale des scores de rappel et de précision.

Dimensions		64				
Gaussiennes		4	8	16	32	64
Top1	rappel	5,5	6,5	7,0	8,5	8,0
	précision	6,5	9,9	19,2	32,7	39,0
	f-mesure	5,9	7,8	10,2	**13,5**	13,3
Top10	rappel	6,5	9,0	9,0	9,5	8,5
	précision	7,6	13,7	24,7	36,5	41,6
	f-mesure	7,0	10,9	13,2	**15,1**	14,1
Top20	rappel	8,0	10,5	10,5	10,55	10,0
	précision	9,4	16,0	28,8	40,4	48,8
	f-mesure	8,6	12,7	15,4	**16,7**	16,6

TAB. *4.17* – *Résultats du modèle génératif sur les matrices de distances réduites à 64 dimensions.*

Très peu de termes sont modélisés par GMM lorsque leurs matrices de distances sont réduites à 64 dimensions. De ce fait, les scores de rappel sont les plus faibles obtenus, en comparaison avec toutes les configurations testées. Cependant, la dégradation observée suit à l'augmentation du nombre de gaussiennes, lors des expériences entre 4 et 32, ne se reproduit pas ici. Les scores de rappel atteignent leurs maximums locaux avec 32 gaussiennes selon les trois rangs d'observation retenus. L'augmentation du nombre de gaussiennes permet toujours d'augmenter la précision du système, mais les résultats sont inférieurs aux précisions obtenues précédemment.

Analyse des scores de vraisemblance croisée

Comme spécifié dans la section 4.3.4, le manque de données pour certains termes peut induire un sur-apprentissage. Ainsi, les GMM obtenus modélisent uniquement

les informations présentes dans les contextes observés. De ce fait, les scores de vraisemblance croisée calculés entre ce type de modèle et des données brutes ne représentent pas réellement l'adéquation des contextes non observés et les valeurs singulières modélisées par GMM. Nous pensons qu'il est judicieux de mettre en place des seuils, permettant d'accepter uniquement les termes placés au premier rang si leur score de vraisemblance avec un modèle est suffisamment éloigné de celui obtenu par le terme au second rang.

De plus, des scores de vraisemblance trop faibles pour l'ensemble des candidats indiquent clairement que le modèle ne contient pas des informations singulières caractérisant un terme. Ainsi, tous les candidats obtiennent des scores très proches, et généralement bas. Nous décidons d'utiliser deux seuils :

- le premier permet d'éliminer des candidats dont les scores de vraisemblance avec un modèle sont trop bas (noté λ),
- le second permet d'éliminer des candidats dont les scores de vraisemblance sont trop proches de ceux obtenus par les autres candidats (noté $diff$).

Nous présentons les résultats obtenus avec les seuils déterminés de manière empirique sur l'ensemble des candidats de test. La table 4.18 contient les résultats selon deux configurations : avec un espace réduit à 4 et à 16 dimensions (noté d). Pour chacune de ces configurations, nous avons modélisé deux modèle GMM, avec 4 et 64 composants gaussiens (noté g). Les résultats présentés sont ceux obtenus au rang 1 dans la liste ordonnées des traductions retournées

d	g	λ	$diff$	précision	rappel	f-mesure
4	4	> 0	$> 1{,}5$	**100**	15,4	26,7
	64	< 3	-	76,6	35,8	**48,8**
16	4	> 3	> 3	79,1	16,9	27,9
	64	< 3	-	75,3	33,3	46,2

TAB. 4.18 – *Résultats du modèle génératif en utilisant des seuils sur les scores de vraisemblance croisée.*

4.4 Discussion

Nous avons présenté dans ce chapitre nos contributions à l'acquisition de terminologie médicale. Nous avons validé l'utilisation de Wikipédia pour la construction de vecteurs de contexte en menant des expériences de recherche d'information. Cette étape a été nécessaire pour mesurer la pertinence d'utiliser des articles Wikipédia comme documents pouvant constituer des ensembles comparables entre les langues, sans procéder à un alignement au préalable.

4.4.1 Synthèse des résultats

Parmi les différentes approches étudiées pour l'acquisition de vocabulaire spécialisé, nous avons présenté la possibilité d'utiliser les vecteurs de contexte, la modélisation thématique et les cognats, d'une manière individuelle, mais aussi combinée. Nous avons ensuite proposé une nouvelle approche basée sur une modélisation GMM des contextes de mot intégrant la notion de distance lexicale.

Approches individuelles

En construisant des vecteurs de contexte pour des termes médicaux et en les comparant entre les langues, nous avons pu former plus de 30% de couples étant en relation de traduction sur les candidats testés. Il s'est avéré que certains termes sont mieux caractérisés par un contexte directe, ou proche, et d'autres le sont par un contexte plus éloigné. La combinaison de plusieurs tailles de contextes provoque une hausse de la précision, mais moins de candidats sont couverts par l'approche, et le rappel est donc dégradé.

Suite à l'approche basée sur les vecteurs de contextes, nous avons proposé d'intégrer des informations issues d'un modèle thématique. Ce modèle, composé de classes, permet de position dans un espace thématique les termes à traduire, tout en restant indépendant des langues concernées. Les modèles utilisés sont construit selon une méthode non supervisée d'analyse de la sémantique latente présente dans des documents. Si cette méthode nous a permis de retrouver jusqu'à un tiers de bonnes traductions pour les candidats testés, les résultats en terme de précision nous empêchent d'utiliser l'approche thématique seule.

L'hypothèse selon laquelle des termes médicaux français et anglais peuvent partager des similarités orthographiques, nous a permis de mener une série d'expériences sur l'alignement de couples de traductions basée sur la distance d'édition. Si la comparaison des 4 premières lettres des termes a permis de retrouver 34% de bonnes traductions, prendre en compte les termes entiers mène à un de gains allant jusqu'à 63% pour les scores de f-mesure. Cette méthode de comparaison est dépendante des langues étudiées, contrairement aux deux approches présentées précédemment.

Approche multivue

Les différentes combinaisons des approches contextuelle, thématique et graphique permettent d'atteindre des scores de précision de 100% selon la configuration choisie, mais fournissent des scores de rappel très bas (7,6% pour la combinaison des trois vues). Sans utiliser les cognats, le modèle thématique nous permet de valider des hypothèses émises par la comparaison des vecteurs de contexte, et ainsi d'accroître la précision du système jusqu'à 100% en couvrant 13% des candidats testés, ou

encore 76,2% de précision avec 20,8% de rappel si toutes les tailles de contextes sont combinées.

L'approche orthographique appliquée seule permettant d'atteindre de bons résultats, les combinaisons l'impliquant montrent une hausse des résultats par rapport à l'approche par contexte seule. De 33% de f-mesure environ dans la combinaison contexte-cognats, et jusqu'à 40,3% de f-mesure dans la combinaison multivue. Cette dernière configuration permet d'atteindre 99,3% de précision et 24,2% de rappel avec une fenêtre contextuelle de 30 mots.

Les scores de précision élevés sont particulièrement intéressants, car cela permet d'accorder une forte confiance dans les résultats retournés par le système. Ainsi, il est possible d'enrichir progressivement le lexique pivot, grâce aux traductions proposées, pour construire de nouveaux vecteurs de contextes et d'augmenter le nombre de dimensions de l'espace thématique. De ce fait, augmenter le rappel paraît possible avec un algorithme permettant d'itérer à chaque enrichissement du lexique pivot.

Au niveau de l'utilisabilité de notre système, permettant par exemple de fournir des lexiques spécialisés à des experts du domaine médical, atteindre une précision élevée est aussi avantageux. En effet, lorsque le système propose une traduction, cette dernière est correcte. Ainsi, pour plus de 20% des termes testés, la traduction retournée est la bonne. Dans les 80% des cas où le système ne couvre pas le candidat à la traduction, aucun résultat n'est retourné.

Modèle génératif

L'approche basée sur les GMM modélisant les distances lexicales présente un certain nombre d'intérêts théoriques mais s'est révélée difficile à mettre en œuvre. Notre motivation était double :

- développer un modèle qui dépasse les limites d'une approche sac-de-mots estimé dans une fenêtre de taille fixe par l'utilisation d'un modèle paramétrique intégrant les distances lexicales,
- modéliser finement l'ensemble des contextes dans lequel un mot est observé en modélisant l'ensemble des contextes par des distributions multinomiale.

Les résultats des expériences préliminaires que nous avons menées montrent qu'une des principales difficultés de cette approche est liée à la quantité de données nécessaire à l'estimation des modèles. En effet, la méthode que nous proposons s'appuie sur une modélisation statistique de l'ensemble des contextes d'un terme. Elle dépend donc de la quantité de données utilisée et de la fréquence d'apparition des termes en contexte. Plus les observations sont élevées, moins la modélisation des contextes se fige sur des exemples marginaux. Augmenter la quantité de données pourrait donc permettre de résoudre le problème de sur-apprentissage, par exemple en constituant des corpus à partir du Web. Cependant, dans le cas de la traduction de termes peu fréquents, notre modèle est confronté aux limites inhérentes à la disponibilité des corpus permettant d'observer des contextes.

Les résultats obtenus tiennent compte de cette difficulté d'estimation, par les mesures de rappel et de f-mesure. Cette approche nécessite tout de même un approfondissement, car les résultats obtenus ne sont guère satisfaisants. Dans les meilleures configurations, les traductions retournées au rang 1 permettent à peine d'atteindre 20% de f-mesure, et jusqu'à 30% lors de l'observation au rang 20. Encore une fois, les scores de rappel sont assez bas, contrairement à la précision pouvant atteindre 40%.

Nous avons cependant remarqué que l'analyse des scores de vraisemblance croisée permet de repérer les cas de sur-apprentissage des GMM, lorsque trop peu de données contextuelles sont présentes pour un terme. La mise en place de seuils à permis de faire remonter les scores, jusqu'à 100% de précision dans certaines configurations. Les scores de rappel sont, quant à eux, extrêmement bas. Seuls une trentaine de candidats sont pris en charge par la méthode, sur les 200 termes médicaux à traduire composant l'ensemble de test (soit 15% des candidats).

Les résultats de ce modèle génératif ne permettent pas d'atteindre ceux obtenus avec l'approche multivue. Cependant, ce modèle a la capacité d'estimer des traductions selon les contextes lexicaux prenant en considération la notion de distance. Cette approche est intéressante selon plusieurs aspects :

- la modélisation mathématique robuste,
- la flexibilité de l'espace de représentation et du nombre de composants gaussiens,
- l'indépendance par rapport aux langues étudiées,
- la possibilité d'utiliser des corpus de tailles inégales.

4.4.2 Travaux précédents

Dans la littérature, un certain nombre de travaux concernent l'acquisition de termes médicaux à partir de corpus comparables, notamment Wikipédia. Nos résultats peuvent être comparés avec ceux obtenus dans les travaux effectués par Laroche et Langlais (2010), car les termes à traduire et le lexique pivot sont identiques. Ils utilisent aussi Wikipédia pour extraire des corpus comparables. Cependant, ils n'utilisent pas l'ensemble Wikipédia, mais effectuent une extraction de documents contenant les termes à traduire.

Lorsque les auteurs effectuent des expériences sur une liste de 3 000 candidats à traduire, en se basant sur l'approche des vecteurs de contexte, ils atteignent une f-mesure de 13,3% mesurée au premier rang des résultats retournés par leur système. Pour la même approche, nous atteignons une f-mesure de plus de 33%, allant jusqu'à 35% lors de la combinaison des tailles de fenêtres d'observation des co-occurrences. Notre approche multivue permet d'améliorer encore ces résultats, jusqu'à 40% de f-mesure, et 99,3% de précision.

D'une manière plus générale, l'utilisation de Wikipédia pour l'extraction terminologique bilingue s'est popularisée ces dernières années, notamment en raison de la quantité grandissante d'articles encyclopédiques, et l'augmentation du nombre de langues présentes dans cette ressource. De ce fait, il paraît intéressant, comme nous

l'avons montré dans la section 4.1, d'utiliser Wikipédia comme corpus comparable, dont la structure permet en plus d'effectuer un alignement au niveau des documents.

Chapitre 5

Conclusion

Nous avons présenté dans cette thèse nos contributions à l'adaptation aux domaines de spécialité en traduction automatique statistique. Deux aspects ont été étudiés : l'édition *a posteriori* de traductions issues de systèmes automatiques afin d'en améliorer la qualité dans un contexte de spécialité, et l'acquisition de lexiques terminologiques bilingues permettant d'accroître la couverture du vocabulaire spécialisé.

La post-édition de traductions

Si la post-édition est généralement effectuée par des humains, les approches automatiques permettant de corriger des erreurs produites par les systèmes de traduction ont récemment acquis une popularité dans la communauté de traitement automatique du langage. Nos travaux en post-édition s'inscrivent dans ce mouvement tout en se démarquant des travaux précédents. Nous avons proposé de combiner des modèles statistiques pour la traduction et la post-édition et ainsi de disposer d'une chaîne de traitement entièrement automatique. Nous évitons donc toute intervention humaine et proposons d'adapter aux domaines de spécialité des systèmes aux caractéristiques distinctes.

Ainsi, nous avons montré les diverses possibilités dans l'utilisation des données spécialisées et leur intégration à différents niveaux dans le processus de traduction. Il est possible de réduire les coûts liés la construction de systèmes de traduction spécialisés en adoptant une démarche *a posteriori* d'édition de traductions. Ces dernières sont améliorées par la post-édition, qu'elles proviennent d'un système de traduction état-de-l'art ou d'un système grand publique accessible en ligne.

Notre première série d'expériences a permis de montrer les possibilités d'amélioration de la qualité des traductions de brèves journalistiques, correspondant à des textes multi-domaines. Les résultats obtenus ont motivé la poursuite de nos expérimentations dans le domaine médical, où les gains en qualité de traduction suite à la post-édition se sont avérés plus importants. Nous avons montré qu'il est possible d'adapter un système de traduction par segments générique *a posteriori*, sans nécessiter

la construction d'un nouveau système de traduction de la langue source vers la langue cible.

L'utilisation d'une faible quantité de données parallèles du domaine peut se faire dans cette étape de post-édition, et mène à des gains en terme de scores BLEU. Nous avons aussi montré que l'utilisation d'un service de traduction accessible en ligne permet d'effectuer la première étape de traduction à moindre coût, et la post-édition des sorties de ce système permet d'atteindre des gains importants en qualité de traduction.

Suite à ces expérimentations, nous avons proposé deux approches permettant d'appliquer le concept de post-édition en sélectionnant :
- soit les traductions à éditer,
- soit les paires de segments issus de la table de post-édition.

Ces deux approches sont originales et n'ont, à notre connaissance, jamais été proposées dans une étape de post-édition automatique basée sur une approche d'alignement sous-phrastique. Les gains apportés par ces deux approches permettent de valider nos propositions, mais les scores oracles mesurés indiquent des améliorations possibles qui ne sont pas encore atteintes.

Il reste toutefois le problème de la couverture du vocabulaire spécialisé. Les données parallèles utilisées pour construire les modèles de traduction et de post-édition ne contiennent pas forcément tous les termes devant être traduits. Ainsi, la prise en charge des mots hors vocabulaire nécessite le recours à des ressources extérieures, comme des lexiques bilingues spécialisés.

L'acquisition de lexiques bilingues

La prise en charge des mots hors vocabulaire dans les systèmes de traitement automatique du langage reste une problématique d'actualité. Nos travaux en acquisition de lexiques bilingues spécialisés permettent de palier ce manque de données, en étudiant plusieurs niveaux de représentation des termes selon des tailles de contextes différentes. Nous avons proposé, dans un premier temps, d'analyser les ressources bilingues utilisées pour nos expériences en extraction terminologique. Les résultats obtenus appuient notre hypothèse sur l'utilisabilité de Wikipédia comme corpus comparable. Nous avons aussi montré que la taille et le contenu des lexiques pivots agissent directement sur les performances de notre système en terme de recherche d'information.

Si ces expériences préliminaires ont permis de valider les ressources bilingues utilisées, c'est dans un second temps que nous nous sommes concentrés sur la traduction de termes médicaux. Nous avons étudié différents aspects permettant de modéliser les termes : le contexte, les thèmes et la graphie. Si les deux premières approches sont indépendantes des langues et des domaines de spécialité, ce n'est toutefois pas le cas avec l'approche par graphie. En effet, les langues française et anglaise utilisant un alphabet commun, et la terminologie médicale tirant ses racines du grecque ou du latin, il paraît aisé de comparer l'orthographe des termes entre ces deux langues.

Nous avons alors proposé de combiner les différentes vues, donnant lieu à une approche multivue originale. Les résultats obtenus sont bons en comparaison à l'état de l'art. D'une manière globale, les approches en extraction terminologique bilingue présentés dans la littérature mènent à des perspectives d'utilisabilité réduites, principalement à cause de scores relativement bas. Nous proposons, par la combinaison des vues, une méthode d'auto-évaluation des résultats, permettant d'augmenter la confiance générale dans le système, et de produire des lexiques bilingues de qualité.

Les combinaisons possibles dans les différents niveaux de représentation d'un terme permettent justement d'accroître la pertinence des résultats retournés. Nous pensons cependant que la méthode de combinaison des vues, effectuée en l'occurrence par un vote où chaque vue a le même poids dans la décision finale, peut être améliorée dans les perspectives relatives à la tâche d'extraction de lexiques bilingues.

Nous avons aussi présenté une autre approche, permettant de représenter un terme par une matrice de contextes intégrant les distances. Une des motivations réside dans le fait qu'un représentation vectorielle des contextes est limitée par l'agrégation de l'ensemble des observations sous la forme d'un seul vecteur. De ce fait, les spécificités de chaque contexte en particulier ne sont pas intégralement restituées dans la modélisation vectorielle. Dans l'optique de conserver toutes les caractéristiques contextuelles d'un terme, nous proposons une nouvelle approche basée sur un modèle génératif représentant la porté continue des contextes.

Nous avons étendu l'approche contextuelle *classique* en modélisant l'ensemble des contextes d'un terme sous la forme d'une matrice composée de vecteurs intégrant la notion de distance lexicale. Cette modélisation nous permet de contourner la limitation imposée par une taille de fenêtre d'observation figée. De ce fait, nous allons plus loin que la simple représentation par sac de mots du vecteur de contexte et nous intégrons les informations contextuelles locales et globales. Malgré le peu de données disponibles pour évaluer notre approche, les termes pris en charge ont pu être modélisé et confronté à la mesure de vraisemblance croisée permettant de proposer des traductions candidates.

Perspectives

Nous envisageons d'intégrer les lexiques bilingues construits automatiquement à un système de traduction automatique par segments ou à un système de post-édition statistique. En effet, nous n'avons pas évalué l'apport de ce genre de ressources dans un contexte de traduction statistique pour des domaines de spécialité. Puis, nous voulons évaluer notre approche de post-édition statistique dans les cas où un lexique bilingue spécialisé est disponible. Cela nous permettrait d'évaluer les capacités de la post-édition statistique de corrections d'erreurs produites par le système de traduction automatique, sans prendre en compte les aspects relatifs à la couverture terminologique. Nous souhaitons étudier l'impact de la post-édition statistique sur l'amélioration des structures syntaxiques des hypothèses issues d'un système de traduction automatique par segments en procédant à une évaluation subjective des

résultats.

Si nos contributions furent évaluées dans le domaine médical avec le couple de langues français-anglais, il semble intéressant de pouvoir prendre en charge d'autres domaines et d'autres langues. Ainsi, nous pourrons valider l'indépendance de nos approches par rapport aux langues étudiées et aux domaines de spécialité concernés. Nous pensons que seule l'approche permettant l'acquisition de lexiques bilingues basée sur les cognats est dépendante des langues étudiées dans cette thèse et du domaine médical.

Enfin, nous souhaitons poursuivre nos expériences sur le modèle génératif à portées continues en constituant des corpus permettant d'extraire une plus grande quantité de contextes pour les termes. Nous pensons en particulier au Web, où la construction de corpus multilingues comparables est possible par l'intermédiaire de moteurs de recherche (Kilgarriff et Grefenstette, 2003; Do et al., 2009). Ainsi, nous pensons pouvoir prendre en charge plus de candidats à la traduction dans des domaines de spécialité où la fréquence d'apparition des termes à traduire est basse.

Le modèle génératif permet de représenter des contextes de mots selon une portée plus longue que les représentations classiques du langage. Ainsi, nous envisageons les possibilités d'estimer des relations longues entre des mots pour palier les limites de la représentation n-gramme des modèles de langage. Il paraît alors possible d'utiliser la représentation contextuelle de mots sous la forme de GMM dans un cadre monolingue.

Notre approche en post-édition statistique a été étudiée en détails lors de l'application du modèle post-édition sur un ensemble de phrases de test. La classification des phrases à post-éditer a été effectuée en prenant en compte les phrases de test avant post-édition, associées avec leur classe relative au gain lié à la post-édition statistique. D'autres informations peuvent être intégrées à l'ensemble d'apprentissage permettant de construire le modèle de classification, comme les phrases dans la langue source, ou d'autres paramètres utilisés notamment dans la prédiction de la qualité des traductions générées automatiquement (Specia et al., 2009).

Nous envisageons alors plusieurs stratégies pour la classification de phrases à post-éditer. La variation des paramètres constituant l'ensemble d'apprentissage peut donner lieu à différents types d'approches : la classification précoce et tardive. La classification précoce consiste à agréger des paramètres pour entraîner un classifieur, tandis que la classification tardive se base sur l'apprentissage de plusieurs classifieurs, un par groupe de paramètres, dont les sorties sont combinées. Nous pourrions alors faire intervenir différents types de classifieurs.

Finalement, l'élagage de la table de post-édition est une approche que nous voulons continuer d'explorer. Notre algorithme proposé afin d'évaluer les couples de segments présents dans la table SPE peut être appliqué de manière itérative. Lorsqu'un premier élagage est effectué, nous pouvons réévaluer les couples restant dans la table et ainsi procéder à un élagage plus important, jusqu'à la stabilisation des scores associés à chaque couple, et l'obtention de la table de post-édition idéale.

D'une manière plus générale, l'adaptation d'un système générique à différents

domaines de spécialité peut être effectuée à différents niveaux. Dans cette thèse, nous avons proposé d'explorer deux tâches distinctes et nous avons rapporté les résultats des évaluations effectuées. Pouvoir s'adapter à tout domaine spécialisé de manière dynamique, efficace et peu coûteuse est un des objectifs à long terme en traduction automatique.

Liste des illustrations

Liste des tableaux

Bibliographie

(Alegria et al., 2005) I. Alegria, A. de Ilarraza, G. Labaka, M. Lersundi, A. Mayor, K. Sarasola, M. Forcada, S. Ortiz-Rojas, et L. Padró, 2005. An open architecture for transfer-based machine translation between spanish and basque. Dans les actes de *MT Summit X, Workshop on Open-Source Machine Translation*, 7–14.

(Allen, 2003) J. Allen, 2003. Post-editing. *Computers and Translation : a Translator's Guide*, 297–317.

(Allen et Hogan, 2000) J. Allen et C. Hogan, 2000. Toward the development of a post editing module for raw machine translation output : A controlled language perspective. Dans les actes de *CLAW-00*, 62–71.

(Almeida et al., 2007) R. Almeida, B. Mozafari, et J. Cho, 2007. On the evolution of wikipedia. Dans les actes de *ICWSM*.

(Bader et Chew, 2008) B. Bader et P. Chew, 2008. Enhancing multilingual latent semantic analysis with term alignment information. Dans les actes de *COLING*, Volume 1, 49–56.

(Banerjee et Lavie, 2005) S. Banerjee et A. Lavie, 2005. Meteor : An automatic metric for mt evaluation with improved correlation with human judgments. 65–72.

(Bar-Hillel, 1953a) Y. Bar-Hillel, 1953a. A quasi-arithmetic notation for syntactic description. *Language 29*, 47–58.

(Bar-Hillel, 1953b) Y. Bar-Hillel, 1953b. Some linguistic problems connected with machine translation. *Philosophy of science 20*(3), 217–225.

(Bar-Hillel, 1960) Y. Bar-Hillel, 1960. The present status of automatic translation of languages. *Advances in computers 1*, 91–163.

(Béchara et al., 2011) H. Béchara, Y. Ma, et J. van Genabith, 2011. Statistical post-editing for a statistical mt system. Dans les actes de *MT Summit XIII*, 308–315.

(Bellegarda, 1998) J. Bellegarda, 1998. A multispan language modeling framework for large vocabulary speech recognition. *IEEE Transactions on Speech and Audio Processing 6*(5), 456–467.

(Berger et al., 1994) A. Berger, P. Brown, S. Della Pietra, V. Della Pietra, J. Gillett, J. Lafferty, R. Mercer, H. Printz, et L. Ureš, 1994. The candide system for machine translation. Dans les actes de *Workshop on HLT*, 157–162.

(Berry et al., 1995) M. Berry, S. Dumais, et G. O'Brien, 1995. Using linear algebra for intelligent information retrieval. *SIAM review*, 573–595.

(Blei et al., 2003) D. Blei, A. Ng, et M. Jordan, 2003. Latent Dirichlet Allocation. *The Journal of Machine Learning Research 3*, 993–1022.

(Boitet et al., 1982) C. Boitet, P. Guillaume, et M. Quezel-Ambrunaz, 1982. Ariane-78, an integrated environment for automated translation and human revision. Dans les actes de *COLING*, 19–27.

(Boser et al., 1992) B. Boser, I. Guyon, et V. Vapnik, 1992. A training algorithm for optimal margin classifiers. Dans les actes de *Workshop on Computational learning theory*, 144–152.

(Boyd-Graber et Blei, 2009) J. Boyd-Graber et D. M. Blei, 2009. Multilingual topic models for unaligned text. Dans les actes de *Conference on Uncertainty in Artificial Intelligence*, 75–82.

(Brown et al., 1990) P. Brown, S. Della Pietra, V. Della Pietra, F. Jelinek, J. Lafferty, R. Mercer, et P. Roossin, 1990. A Statistical Approach to Machine Translation. *Computational Linguistics 16*(2), 79–85.

(Brown et al., 1993) P. Brown, S. Pietra, V. Pietra, et R. Mercer, 1993. The mathematic of statistical machine translation : Parameter estimation. *Computational linguistics 19*(2), 263–311.

(Byrne et al., 2003) W. Byrne, S. Khudanpur, W. Kim, S. Kumar, P. Pecina, P. Virga, P. Xu, et D. Yarowsky, 2003. The johns hopkins university 2003 chinese-english machine translation system. Dans les actes de *MT Summit IX*, 447–450.

(Callison-Burch, 2009) C. Callison-Burch, 2009. Fast, cheap, and creative : Evaluating translation quality using amazon's mechanical turk. Dans les actes de *EMNLP*, Volume 1, 286–295.

(Camelin et al., 2011) N. Camelin, B. Detienne, S. Huet, D. Quadri, et F. Lefevre, 2011. Unsupervised concept annotation using latent dirichlet allocation and segmental methods. Dans les actes de *EMNLP 1st UNSUP Workshop*, 72–81.

(Champollion, 1828) J. Champollion, 1828. *Précis du système hiéroglyphique des anciens Égyptiens, ou Recherches sur les éléments premiers de cette écriture sacrée, sur leurs diverses combinaisons, et sur les rapports de ce système avec les autres méthodes graphiques égyptiennes*. Imprimerie royale. 499 pages.

(Chandioux, 1976) J. Chandioux, 1976. Météo : un système opérationnel pour la traduction automatique des bulletins météorologiques destinés au grand public. *Meta 21*, 127–133.

(Charton et Torres-Moreno, 2010) E. Charton et J. Torres-Moreno, 2010. Nlgbase : a free linguistic resource for natural language processing systems. Dans les actes de *LREC*, Volume 1, 2621–2625.

(Chiao, 2004) Y. Chiao, 2004. *Extraction lexicale bilingue à partir de textes médicaux comparables : application à la recherche d'information translangue*. Thèse de Doctorat, Université Paris 6. 190 pages.

(Chiao et Zweigenbaum, 2002) Y. Chiao et P. Zweigenbaum, 2002. Looking for Candidate Translational Equivalents in Specialized, Comparable Corpora. Dans les actes de *COLING*, Volume 2, 1–5.

(Daille et Morin, 2005) B. Daille et E. Morin, 2005. French-English terminology extraction from comparable corpora. 707–718.

(Daumé III et Marcu, 2006) H. Daumé III et D. Marcu, 2006. Domain adaptation for statistical classifiers. *Journal of Artificial Intelligence Research 26*(1), 101–126.

(de Ilarraza et al., 2008) A. de Ilarraza, G. Labaka, et K. Sarasola, 2008. Statistical postediting : A valuable method in domain adaptation of rbmt systems for less-resourced languages. Dans les actes de *Workshop Mixing Approaches to Machine Translation 2008*, 35–40.

(De Saussure et al., 2008) F. De Saussure, S. Bouquet, R. Engler, et A. Weil, 2008. *Écrits de linguistique générale*. Gallimard. 353 pages.

(Deerwester et al., 1990) S. Deerwester, S. Dumais, G. Furnas, T. Landauer, et R. Harshman, 1990. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science 41*(6), 391–407.

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

(Déjean et al., 2005) H. Déjean, E. Gaussier, J. Renders, et F. Sadat, 2005. Automatic Processing of Multilingual Medical Terminology : Applications to Thesaurus Enrichment and Cross-language Information Retrieval. *Arificial Intelligence in Medicine 33*(2), 111–124.

(Déjean et al., 2002) H. Déjean, É. Gaussier, et F. Sadat, 2002. An approach based on multilingual thesauri and model combination for bilingual lexicon extraction. Dans les actes de *COLING*, Volume 1, 1–7.

(Dice, 1945) L. Dice, 1945. Measures of the amount of ecologic association between species. *Ecology 26*(3), 297–302.

(Dijkstra et al., 1999) T. Dijkstra, J. Grainger, et W. Van Heuven, 1999. Recognition of cognates and interlingual homographs : The neglected role of phonology* 1,* 2. *Journal of Memory and Language 41*(4), 496–518.

(Do et al., 2009) T. Do, V. Le, B. Bigi, L. Besacier, et E. Castelli, 2009. Mining a comparable text corpus for a vietnamese-french statistical machine translation system. Dans les actes de *Proceedings of the Fourth Workshop on Statistical Machine Translation*, 165–172. Association for Computational Linguistics.

(Dostert, 1955) L. Dostert, 1955. The georgetown-ibm experiment. *Machine translation of languages*, 124–135.

(Ducrot, 1973) J. Ducrot, 1973. Mise en application et en exploitation opérationnelle d'une méthode de traduction automatique de textes documentaires en vue d'accroître leur utilisation dans le monde. *Délégation Générale à la Recherche Scientifique et Technique*.

(Dugast et al., 2007) L. Dugast, J. Senellart, et P. Koehn, 2007. Statistical post-editing on systran's rule-based translation system. Dans les actes de *WMT*, 220–223.

(Eck et al., 2004) M. Eck, S. Vogel, et A. Waibel, 2004. Language model adaptation for statistical machine translation based on information retrieval. Dans les actes de *LREC*, 327–330.

(Evert, 2004) S. Evert, 2004. *The Statistics of Word Cooccurrences : Word Pairs and Collocations*. Thèse de Doctorat, Universität Stuttgart. 353 pages.

(Federico et al., 2008) M. Federico, N. Bertoldi, et M. Cettolo, 2008. Irstlm : an open source toolkit for handling large scale language models. Dans les actes de *Ninth Annual Conference of the International Speech Communication Association*.

(Firth, 1957) J. Firth, 1957. A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*, 1–32.

(Fung, 1995) P. Fung, 1995. Compiling Bilingual Lexicon Entries from a Non-parallel English-Chinese Corpus. Dans les actes de *Workshop on Very Large Corpora*, 173–183.

(Fung, 1998) P. Fung, 1998. A Statistical View on Bilingual Lexicon Extraction : from Parallel Corpora to Non-parallel Corpora. *Lecture Notes in Computer Science 1529*, 1–17.

(Fung et McKeown, 1997) P. Fung et K. McKeown, 1997. Finding Terminology Translations from Non-parallel Corpora. Dans les actes de *Workshop on Very Large Corpora*, 192–202.

(Gale et Church, 1991) W. Gale et K. Church, 1991. Identifying word correspondences in parallel texts. Dans les actes de *Proceedings of the workshop on Speech and Natural Language*, 152–157.

(Gao et Vogel, 2008) Q. Gao et S. Vogel, 2008. Parallel implementations of word alignment tool. 49–57.

(Gaussier et al., 2004) E. Gaussier, J. Renders, I. Matveeva, C. Goutte, et H. Dejean, 2004. A Geometric View on Bilingual Lexicon Extraction from Comparable Corpora. Dans les actes de *ACL*, 526–533.

(Germann, 2001) U. Germann, 2001. Aligned hansards of the 36th parliament of canada, release 2001-1a. *http ://www.isi.edu/natural-language/download/hansard*.

(Gross, 1964) M. Gross, 1964. On the equivalence of models of language used in the fields of mechanical translation and information retrieval. *Information storage and retrieval 2*(1), 43–57.

(Haghighi et al., 2008) A. Haghighi, P. Liang, T. Berg-Kirkpatrick, et D. Klein, 2008. Learning bilingual lexicons from monolingual corpora. Dans les actes de *ACL*, 771–779.

(Hasan et Ney, 2005) S. Hasan et H. Ney, 2005. Clustered language models based on regular expressions for smt. Dans les actes de *EAMT*, 119–125.

(Hildebrand et al., 2005) A. Hildebrand, M. Eck, S. Vogel, et A. Waibel, 2005. Adaptation of the translation model for statistical machine translation based on information retrieval. Dans les actes de *EAMT*, 133–142.

(Hofmann, 1999) T. Hofmann, 1999. Probabilistic latent semantic indexing. Dans les actes de *SIGIR*, 50–57.

(Hofmann et al., 1999) T. Hofmann, J. Puzicha, et M. Jordan, 1999. Unsupervised learning from dyadic data. *Advances in Neural Information Processing Systems 11*, 466–472.

(Hutchins, 2007) J. Hutchins, 2007. *Machine translation : A concise history*, 1–21. Chinese University of Hong-Kong. http ://www.hutchinsweb.me.uk/history.htm.

(Hutchins, 1982) W. Hutchins, 1982. The evolution of machine translation systems. *Practical Experience of Machine Translation. North-Holland, Amsterdam*, 21–37.

(Hutchins, 1986) W. Hutchins, 1986. *Machine translation : past, present, future*. Ellis Horwood Series in Computers and their Applications. 382 pages.

(Hutchins, 1997) W. Hutchins, 1997. From first conception to first demonstration : the nascent years of machine translation, 1947–1954. a chronology. *Machine Translation 12*(3), 195–252.

(Hutchins, 2000) W. Hutchins, 2000. Early years in machine translation : memoirs and biographies of pioneers. *Studies in the history of the language sciences 97*, 400.

(Hutchins, 2003) W. Hutchins, 2003. Has machine translation improved ? some historical comparisons. Dans les actes de *MT Summit IX*, 181–188.

(Hutchins et Somers, 1992) W. Hutchins et H. Somers, 1992. *An introduction to machine translation*. Academic Press New York. 362 pages.

(Isabelle et al., 2007) P. Isabelle, C. Goutte, et M. Simard, 2007. Domain adaptation of mt systems through automatic post-editing. Dans les actes de *MT Summit XI*, 255–261.

(Janiszek et al., 2001) D. Janiszek, R. De Mori, et E. Bechet, 2001. Data augmentation and language model adaptation. Dans les actes de *ICASSP*, Volume 1, 549–552.

(Johnson et al., 2007) J. Johnson, J. Martin, G. Foster, R. Kuhn, et al., 2007. Improving translation quality by discarding most of the phrasetable. Dans les actes de *EMNLP*, 967–975.

(Kilgarriff et Grefenstette, 2003) A. Kilgarriff et G. Grefenstette, 2003. Introduction to the special issue on the web as corpus. *Computational linguistics 29*(3), 333–347.

(King, 1981) M. King, 1981. Eurotra–a european system for machine translation. *Lebende Sprachen 26*(1), 12–14.

(Kneser et Ney, 1995) R. Kneser et H. Ney, 1995. Improved backing-off for n-gram language modeling. Dans les actes de *ICASSP-95*, Volume 1, 181–184.

(Knight et Chander, 1994) K. Knight et I. Chander, 1994. Automated postediting of documents. Dans les actes de *NCAI*, 779–779.

(Koehn, 2004) P. Koehn, 2004. Statistical significance tests for machine translation evaluation. Dans les actes de *EMNLP*, Volume 4, 388–395.

(Koehn, 2005) P. Koehn, 2005. Europarl : A Parallel Corpus for Statistical Machine Translation. Dans les actes de *MT Summit X*, 79–86.

(Koehn et al., 2007) P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, et al., 2007. Moses : Open source toolkit for statistical machine translation. Dans les actes de *ACL*, 177–180.

(Koehn et Knight, 2000) P. Koehn et K. Knight, 2000. Estimating word translation probabilities from unrelated monolingual corpora using the em algorithm. Dans les actes de *NCAI*, 711–715.

(Koehn et Knight, 2001) P. Koehn et K. Knight, 2001. Knowledge sources for word-level translation models. Dans les actes de *EMNLP*, 27–35.

(Koehn et Knight, 2002) P. Koehn et K. Knight, 2002. Learning a Translation Lexicon from Monolingual Corpora. Dans les actes de *ACL workshop on unsupervised lexical acquisition*, Volume 9, 9–16.

(Koehn et al., 2003) P. Koehn, F. Och, et D. Marcu, 2003. Statistical phrase-based translation. Dans les actes de *NAACL*, Volume 1, 48–54.

(Koehn et Schroeder, 2007) P. Koehn et J. Schroeder, 2007. Experiments in domain adaptation for statistical machine translation. Dans les actes de *WMT*, 224–227.

(Krings et Koby, 2001) H. Krings et G. Koby, 2001. *Repairing texts : empirical investigations of machine translation post-editing processes*. The Kent State University Press. 580 pages.

(Kuhn et al., 2010) R. Kuhn, P. Isabelle, C. Goutte, J. Senellart, M. Simard, N. Ueffing, et al., 2010. Recent advances in automatic post-editing. *Multilingual computing and technology*, 43–46.

(Laffling, 1992) J. Laffling, 1992. On Constructing a Transfer Dictionary for Man and Machine. *Target 4*(1), 17–31.

(Langlais, 2002) P. Langlais, 2002. Improving a general-purpose statistical translation engine by terminological lexicons. Dans les actes de *workshop on computational terminology, COLING*, 1–7.

(Langlais et al., 2006) P. Langlais, F. Gotti, et A. Patry, 2006. De la chambre des communes à la chambre d'isolement : adaptabilité d'un système de traduction basé sur les segments. Dans les actes de *TALN*, 217–226.

(Langlais et Patry, 2007) P. Langlais et A. Patry, 2007. Translating unknown words by analogical learning. Dans les actes de *EMNLP*, 877–886.

(Langlais et al., 2009) P. Langlais, F. Yvon, et P. Zweigenbaum, 2009. Improvements in analogical learning : application to translating multi-terms of the medical domain. Dans les actes de *EACL*, 487–495.

(Laroche et Langlais, 2010) A. Laroche et P. Langlais, 2010. Revisiting Context-based Projection Methods for Term-translation Spotting in Comparable Corpora. Dans les actes de *COLING*, 617–625.

(Lavallée, 2010) J. Lavallée, 2010. Moranapho : apprentissage non supervisé de la morphologie d'une langue par généralisation de relations analogiques. Mémoire de Master, Université de Montréal, Québec, Canada. 80 pages.

(Lepage, 2003) Y. Lepage, 2003. De l'analogie rendant compte de la commutation en linguistique. Habilitation à Diriger les Recherches, Université de Grenoble.

(Levenshtein, 1966) V. Levenshtein, 1966. Binary codes capable of correcting deletions, insertions, and reversals. Dans les actes de *Soviet Physics Doklady*, Volume 10, 707–710.

(Lin et al., 2007) H. Lin, C. Lin, et R. Weng, 2007. A note on platt's probabilistic outputs for support vector machines. *Machine Learning 68*(3), 267–276.

(Loh, 1972) S. Loh, 1972. Machine translation at the chinese university of hong kong. Dans les actes de *CETA Workshop on Chinese Language and Chinese Research Materials*.

(Losee, 1998) R. Losee, 1998. *Text retrieval and filtering : analytic models of performance*. Kluwer Academic Publishers. 256 pages.

(Maas, 1977) H. Maas, 1977. The saarbrücken automatic translation system (susy). Dans les actes de *European Congress on Information Systems and Networks, Overcoming the language barrier*, 585–592.

(Mahajan et al., 1999) M. Mahajan, D. Beeferman, et X. Huang, 1999. Improved topic-dependent language modeling using information retrieval techniques. Dans les actes de *ICASSP*, 541–544.

(Marchaisse, 1991) T. Marchaisse, 1991. L'acte du traducteur et le principe d'indétermination. *Le Gré des Langues 2*, 144–157.

(Marcu et Wong, 2002) D. Marcu et W. Wong, 2002. A phrase-based, joint probability model for statistical machine translation. Dans les actes de *EMNLP*, 133–139.

(Mimno et al., 2009) D. Mimno, H. Wallach, J. Naradowsky, D. Smith, et A. McCallum, 2009. Polylingual topic models. Dans les actes de *EMNLP*, 880–889.

(Morin, 2007) E. Morin, 2007. Apport des termes complexes à l'acquisition lexicale multilingue à partir de corpus comparables spécialisés : entre intuition et réalité. Dans les actes de *Terminologies et Intelligence Artificielle*, 11–20.

(Morin, 2009) E. Morin, 2009. Apport d'un corpus comparable déséquilibré à l'extraction de lexiques bilingues. Dans les actes de *TALN*.

(Nagao et al., 1985) M. Nagao, J. Tsujii, et J. Nakamura, 1985. The japanese government project for machine translation. *Computational Linguistics 11*(2-3), 91–110.

(Nießen et al., 2000) S. Nießen, F. Och, G. Leusch, et H. Ney, 2000. An evaluation tool for machine translation : Fast evaluation for mt research. Dans les actes de *LREC*, 39–45.

(Nirenburg et al., 1994) S. Nirenburg, R. Frederking, D. Farwell, et Y. Wilks, 1994. Two types of adaptive mt environments. Dans les actes de *COLING*, 125–128.

(Och, 2002) F. Och, 2002. *Statistical Machine Translation : From Single-Word Models to Alignment Templates*. Thèse de Doctorat, Bibliothek der RWTH Aachen. 163 pages.

(Och, 2003) F. Och, 2003. Minimum error rate training in statistical machine translation. Dans les actes de *ACL*, 160–167.

(Och et al., 2004) F. Och, D. Gildea, S. Khudanpur, A. Sarkar, K. Yamada, A. Fraser, S. Kumar, L. Shen, D. Smith, K. Eng, et al., 2004. A smorgasbord of features for statistical machine translation. Dans les actes de *NAACL*, 161–168.

(Och et Ney, 2003) F. Och et H. Ney, 2003. A systematic comparison of various statistical alignment models. *Computational linguistics 29*(1), 19–51.

(Och, 2000) H. Och, F.J. Ney, 2000. Improved statistical alignment models. Dans les actes de *ACL*, 440–447.

(Papineni et al., 2002) K. Papineni, S. Roukos, T. Ward, et W. Zhu, 2002. Bleu : a method for automatic evaluation of machine translation. Dans les actes de *ACL*, 311–318.

(Phan et al., 2008) X. Phan, L. Nguyen, et S. Horiguchi, 2008. Learning to classify short and sparse text & web with hidden topics from large-scale data collections. Dans les actes de *international conference on World Wide Web*, 91–100.

(Pierce et Carroll, 1966) J. Pierce et J. Carroll, 1966. Language and machines : Computers in translation and linguistics. *Rapport du National Academy of Sciences/National Research Council*. 124 pages.

(Pigott, 1988) I. Pigott, 1988. Mt in large organizations : Systran at the commission of the european communities. *Technology as translation strategy 2*, 159.

(Pirrelli et Yvon, 1999) V. Pirrelli et F. Yvon, 1999. The hidden dimension : a paradigmatic view of data-driven nlp. *Journal of Experimental and Theoretical Artificial Intelligence 11*(3), 391–408.

(Posner, 1996) R. Posner, 1996. *The romance languages*. Cambridge Univ Pr.

(Potet et al., 2011a) M. Potet, E. Esperança-Rodier, H. Blanchon, et L. Besacier, 2011a. Preliminary experiments on using users' post-editions to enhance a smt system. Dans les actes de *EAMT*, 161–168.

(Potet et al., 2011b) M. Potet, R. Rubino, B. Lecouteux, S. Huet, H. Blanchon, L. Besacier, et F. Lefevre, 2011b. The liga (lig/lia) machine translation system for wmt 2011. Dans les actes de *WMT*, 440–446.

(Prochasson, 2009) E. Prochasson, 2009. *Alignement multilingue en corpus comparables spécialisés*. Thèse de Doctorat, Université de Nantes. 136 pages.

(Prochasson et Fung, 2011) E. Prochasson et P. Fung, 2011. Rare word translation extraction from aligned comparable documents. Dans les actes de *ACL*, 1327–1335.

(Prochasson et al., 2009) E. Prochasson, E. Morin, et K. Kageura, 2009. Anchor points for bilingual lexicon extraction from small comparable corpora. Dans les actes de *MT Summit XII*, 284–291.

(Quine, 1959) W. Quine, 1959. Meaning and translation.

(Rafalovitch et Dale, 2009) A. Rafalovitch et R. Dale, 2009. United nations general assembly resolutions : A six-language parallel corpus. *MT Summit XII*, 292–299.

(Rapp, 1995) R. Rapp, 1995. Identifying Word Translations in Non-parallel Texts. Dans les actes de *Proceedings of the 33rd ACL Conference*, 320–322. ACL.

(Rapp, 1999) R. Rapp, 1999. Automatic Identification of Word Translations from Unrelated English and German Corpora. Dans les actes de *Proceedings of the 37th ACL conference*, 519–526. ACL.

(Reynolds et al., 1998) D. Reynolds, E. Singer, B. Carlson, G. O'Leary, J. McLaughlin, et M. Zissman, 1998. Blind clustering of speech utterances based on speaker and language characteristics. Dans les actes de *ICSLP*.

(Romesburg, 1984) H. Romesburg, 1984. Cluster analysis for researchers. *Belmont : California, Lifetime Learning Pubications xiii, 334p.-. En General (KR, 198406344)*.

(Sadat et al., 2005) F. Sadat, J. Johnson, A. Agbago, G. Foster, R. Kuhn, J. Martin, et A. Tikuisis, 2005. Portage : A phrase-based machine translation system. Dans les actes de *The Association for Computational Linguistics (ACL) 2005 Workshop on Building and Using Parallel Texts : Data-Driven Machine Translation and Beyond*.

(Sagot et al., 2011) B. Sagot, K. Fort, G. Adda, J. Mariani, B. Lang, et al., 2011. Un turc mécanique pour les ressources linguistiques : critique de la myriadisation du travail parcellise.

(Shannon et Weaver, 1948) C. Shannon et W. Weaver, 1948. The mathematical theory of communication. *Bell Systems Technical Journal 27*(1948), 379–423.

(Simard, 1998) M. Simard, 1998. The baf : a corpus of english-french bitext. Dans les actes de *LREC*, Volume 1, 489–494. Citeseer.

(Simard et al., 1993) M. Simard, G. Foster, et P. Isabelle, 1993. Using cognates to align sentences in bilingual corpora. Dans les actes de *Proceedings of the 1993 conference of the Centre for Advanced Studies on Collaborative research : distributed computing-Volume 2*, 1071–1082. IBM Press.

(Simard et al., 2007a) M. Simard, C. Goutte, et P. Isabelle, 2007a. Statistical phrase-based post-editing. Dans les actes de *NAACL-HLT*, 508–515.

(Simard et al., 2007b) M. Simard, N. Ueffing, P. Isabelle, et R. Kuhn, 2007b. Rule-based translation with statistical phrase-based post-editing. Dans les actes de *Proceedings of the Second Workshop on Statistical Machine Translation*, 203–206.

(Slocum et al., 1984) J. Slocum, W. Bennet, J. Bear, M. Morgan, et R. Root, 1984. Metal : the lrc machine translation system. Dans les actes de *Machine Translation today : the state of the art (Proc. third Lugano Tutorial, 2–7 April 1984)*.

(Snover et al., 2006) M. Snover, B. Dorr, R. Schwartz, L. Micciulla, et J. Makhoul, 2006. A study of translation edit rate with targeted human annotation. Dans les actes de *Proceedings of Association for Machine Translation in the Americas*, 223–231.

(Snover et al., 2009) M. Snover, N. Madnani, B. Dorr, et R. Schwartz, 2009. Fluency, adequacy, or hter ? exploring different human judgments with a tunable mt metric. Dans les actes de *Proceedings of the Fourth Workshop on Statistical Machine Translation*, Volume 30, 259–268.

(Specia, 2011) L. Specia, 2011. Exploiting objective annotations for measuring translation post-editing effort. Dans les actes de *15th Annual Conference of the European Association for Machine Translation, EAMT*, Volume 11.

(Specia et al., 2009) L. Specia, N. Cancedda, M. Dymetman, M. Turchi, et N. Cristianini, 2009. Estimating the sentence-level quality of machine translation systems. *European Association for Machine Translation*, 28.

(Stolcke, 2002) A. Stolcke, 2002. Srilm-an extensible language modeling toolkit. Dans les actes de *Proceedings of the international conference on spoken language processing*, Volume 2, 901–904. Citeseer.

(Tam et al., 2007) Y. Tam, I. Lane, et T. Schultz, 2007. Bilingual lsa-based adaptation for statistical machine translation. *Machine translation 21*(4), 187–207.

(Tam et Schultz, 2006) Y. Tam et T. Schultz, 2006. Unsupervised language model adaptation using latent semantic marginals. Dans les actes de *Proc. of INTERSPEECH*, 2206–2209. Citeseer.

(Taube, 1961) M. Taube, 1961. *Computers and common sense : the myth of thinking machines*. Columbia university press.

(Tiedemann, 2009) J. Tiedemann, 2009. News from opus—a collection of multilingual parallel corpora with tools and interfaces. *Recent advances in natural language processing V : selected papers from RANLP 2007 309*, 237.

(Tillmann, 2003) C. Tillmann, 2003. A projection extension algorithm for statistical machine translation. Dans les actes de *Proceedings of the 2003 conference on Empirical methods in natural language processing*, 1–8. Association for Computational Linguistics.

(Tillmann et al., 1997) C. Tillmann, S. Vogel, H. Ney, A. Zubiaga, et H. Sawaf, 1997. Accelerated dp based search for statistical translation. Dans les actes de *Fifth European Conference on Speech Communication and Technology*.

(Toma, 1970) P. Toma, 1970. Systran machine translation system. Rapport technique, DTIC Document.

(Toma, 1972) P. Toma, 1972. Optimization of systran system. Rapport technique, DTIC Document.

(Toma, 1977) P. Toma, 1977. Systran as a multilingual machine translation system. Dans les actes de *Proceedings of the Third European Congress on Information Systems and Networks, Overcoming the language barrier*, 569–581.

(Tschira, 1985) K. Tschira, 1985. 'looking back at a year of german-english mt with logos. *Tools for the Trade, Translating and the Computer 5*.

(Ueffing, 2006) N. Ueffing, 2006. Self-training for machine translation. Dans les actes de *NIPS workshop on Machine Learning for Multilingual Information Access*.

(Vauquois et Boitet, 1985) B. Vauquois et C. Boitet, 1985. Automated translation at grenoble university. *Computational Linguistics 11*(1), 28–36.

(Viterbi, 1967) A. Viterbi, 1967. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *Information Theory, IEEE Transactions on 13*(2), 260–269.

(Vogel et al., 2003) S. Vogel, Y. Zhang, F. Huang, A. Tribble, A. Venugopal, B. Zhao, et A. Waibel, 2003. The cmu statistical machine translation system. Dans les actes de *Proceedings of MT Summit*, Volume 9, 54. Citeseer.

(Zens et al., 2002) R. Zens, F. Och, et H. Ney, 2002. Phrase-based statistical machine translation. *KI 2002 : Advances in Artificial Intelligence*, 35–56.

(Zhao et al., 2004) B. Zhao, M. Eck, et S. Vogel, 2004. Language model adaptation for statistical machine translation with structured query models. Dans les actes de *Proceedings of the 20th international conference on Computational Linguistics*, 411–es. Association for Computational Linguistics.

(Zhao et Xing, 2006) B. Zhao et E. Xing, 2006. Bitam : Bilingual topic admixture models for word alignment. Dans les actes de *Proceedings of the COLING/ACL on Main conference poster sessions*, 969–976. Association for Computational Linguistics.

(Zhao et Xing, 2008) B. Zhao et E. Xing, 2008. Hm-bitam : Bilingual topic exploration, word alignment, and translation. *Advances in Neural Information Processing Systems 20*, 1689–1696.

(Zweigenbaum et Habert, 2006) P. Zweigenbaum et B. Habert, 2006. «faire se rencontrer les parallèles : regards croisés sur l'acquisition lexicale monolingue et multilingue.». *Revue de sociolinguistique en ligne GLOTTOPOL 8*, 22–44.

Titre : Traduction automatique statistique et adaptation à un domaine spécialisé

Résumé

Nous avons observé depuis plusieurs années l'émergence des approches statistiques pour la traduction automatique. Cependant, l'efficacité des modèles construits est soumise aux variabilités inhérentes au langage naturel. Des études ont montré la présence de vocabulaires spécifique et général composant les corpus de textes de domaines spécialisés. Cette particularité peut être prise en charge par des ressources terminologiques comme les lexiques bilingues. Toutefois, nous pensons que si le vocabulaire est différent entre des textes spécialisés ou génériques, le contenu sémantique et la structure syntaxique peuvent aussi varier. Dans nos travaux, nous considérons la tâche d'adaptation aux domaines spécialisés pour la traduction automatique statistique selon deux axes majeurs : l'acquisition de lexiques bilingues et l'édition a posteriori de traductions issues de systèmes automatiques. Nous évaluons l'efficacité des approches proposées dans un contexte spécialisé : le domaine médical. Nos résultats sont comparés aux travaux précédents concernant cette tâche. De manière générale, la qualité des traductions issues de systèmes automatiques pour le domaine médical est améliorée par nos propositions. Des évaluations en oracle tendent à montrer qu'il existe une marge de progression importante.

Mots-clés : Traduction automatique statistique, domaine spécialisé, post-édition, lexique bilingue, terminologie

---

Title : Domain Adaptation for Statistical Machine Translation

Abstract

These last years have seen the development of statistical approaches for machine translation. Nevertheless, the intrinsic variations of the natural language act upon the quality of statistical models. Studies have shown that in-domain corpora contain words that can occur in out-of-domain corpora (common words), but also contain domain specific words. This particularity can be handled by terminological resources like bilingual lexicons. However, if the vocabulary differs between out and in-domain data, the syntactic and semantic content may also vary. In our work, we consider the task of domain adaptation for statistical machine translation through two major axes : bilingual lexicon acquisition and post-edition of machine translation outputs. We evaluate our approaches on the medical domain. The quality of automatic translations in the medical domain are improved and the results are compared to other works in this field. Oracle evaluations tend to show that further gains are still possible.

Keywords : Statistical machine translation, specific domain, post-edition, bilingual lexicon, terminology