Nancy-Université

UFR mathématiques et informatique **École doctorale IAEM Lorraine**

Département de formation doctorale en informatique

Les triggers inter-langues pour la Traduction Automatique Statistique

THÈSE

présentée et soutenue publiquement le 23 juin 2010

pour l'obtention du

Doctorat de l'université Nancy 2

(spécialité informatique)

par

Caroline Lavecchia

Composition du jury

Président :	Michaël Rusinowitch	Directeur de recherche INRIA, LORIA Nancy
Rapporteurs :	Laurent Besacier	Professeur, Université J. Fourier Grenoble
	Holger Schwenk	Professeur, Université du Maine LeMans
Examinateurs :	Frédéric Béchet	Professeur, Université de la Méditerrannée Marseille
	David Langlois	Maître de conférences, Nancy Université – LORIA Nancy
	Kamel Smaïli	Professeur, Nancy Université – LORIA Nancy (Directeur)

Laboratoire Lorrain de Recherche en Informatique et ses Applications — UMR 7503

Loria

Mis en page avec la classe thloria.

A mon fils Loup, ma plus belle réussite

Table des matières

5

Introduction **7**

Partie I Etat de l'art

Chapitre 1 La traduction automatique et son évaluation **13**

1.1 Introduction 13
1.2 Un bref historique 13
1.3 Les différentes approches 14
1.3.1 L'approche experte 14
1.3.2 L'approche empirique 18
1.3.3 Discussion 20
1.4 Évaluation de la qualité des traductions 21
1.4.1 Campagnes d'évaluation 21
1.4.2 Critères de qualité des traductions 22
1.4.3 Évaluation manuelle des traductions 22
1.4.4 Évaluation automatique des traductions 23
1.4.5 Qualité des mesures d'évaluation automatique 27
1.4.6 Quelle mesure automatique choisir ? 27

Chapitre 2 L'approche statistique de la traduction automatique **31**
2.1 Introduction . . . 31
2.2 Généralités . . . 31
2.3 Modélisation du langage . . . 33
2.3.1 Définition d'un modèle de langage . . . 33
2.3.2 Quelques modèles de langage . . . 34
2.3.3 Utilisation des modèles de langage en TA . . . 36
2.4 Modèles de traduction . . . 37
2.4.1 La notion d'alignement . . . 37
2.4.2 les modèles à base de mots . . . 38
2.4.3 les modèles à base de séquences . . . 44
2.5 Décodeur . . . 48
2.5.1 Généralités . . . 48
2.5.2 Cas particuliers : le décodeur PHARAOH . . . 48
2.6 Discussion générale . . . 52

Partie II Contributions à la Traduction Automatique Statistique

Chapitre 3 Les corpus **57**
3.1 Introduction . . . 57
3.2 Le corpus du Parlement Européen : EUROPARL . . . 57
3.3 Le corpus de sous-titres . . . 58
3.3.1 Données brutes : description et problèmes . . . 58
3.3.2 Solution : programmation dynamique . . . 63
3.3.3 Evaluation de l'alignement automatique . . . 66
3.3.4 Discussion . . . 69

Chapitre 4 Le concept de triggers inter-langues **71**
4.1 Introduction . . . 71
4.2 Les triggers inter-langues : définition . . . 71
4.3 Les triggers inter-langues : étude préliminaire . . . 73

4.4 Construction d'un lexique bilingue à partir des triggers inter-langues 76
4.4.1 Influence de la contrainte de symétrie sur les triggers inter-langues . . 76
4.4.2 Evaluation du lexique bilingue 78
4.4.3 Discussion 80

Chapitre 5 La traduction par mot : les triggers 1-1 **83**
5.1 Introduction 83
5.2 Modèle de traduction de mots : les triggers 1-1 83
5.2.1 Le modèle traduction Trig-n 84
5.2.2 Le modèle de traduction Sym-n 84
5.3 Evaluation des modèles de traductions de mots 84
5.3.1 Cadre expérimental 84
5.3.2 Optimisation des modèles de traduction 86
5.3.3 Optimisation du décodeur 91
5.3.4 Validation sur le corpus de test 92
5.4 Discussion 92

Chapitre 6 La traduction par séquence de mots : les triggers m-n **95**
6.1 Introduction 95
6.2 Les triggers inter-langues de séquences 96
6.2.1 Définition 96
6.2.2 Etude préalable 96
6.3 Modèles de traduction de séquences : les triggers m-n 99
6.3.1 Segmentation du corpus source 100
6.3.2 Apprentissage des triggers m-n 102
6.3.3 Optimisation par Recuit Simulé 103
6.4 Evaluation des modèles de traduction de séquences 105
6.4.1 Construction de l'ensemble des triggers m-n 105
6.4.2 Mise en place du Recuit Simulé 106
6.4.3 Influence du Recuit Simulé sur les corpus de développement 107
6.4.4 Validation sur les corpus de test 107
6.4.5 Analyse des premiers résultats 109
6.4.6 Etude comparative des tables de traduction 110
6.5 Discussion 118

Conclusion et perspectives **121**

Annexe **127**

Annexe A Un modèle de langage pour la prise en compte de l'accord en genre et en nombre des mots **127**

A.1 Introduction 127

A.1.1 Le modèle Cache 127

A.1.2 Les modèles Cache-Trait et Cache-Trait-Partiel 128

A.1.3 Expérimentations 129

A.1.4 Conclusion 131

Annexe B Exemple de traductions automatiques produites par PHARAOH sur le corpus de test EUROPARL **133**

Bibliographie **135**

Liste des figures

1.1 Le triangle de Vauquois . . . 14
1.2 TA directe [Jur 06] . . . 15
1.3 TA à base de règles de transfert [Jur 06] . . . 16
1.4 Exemple simple de TA à base de règles de transfert du Français vers l'Anglais de la phrase "*une fin heureuse*" . . . 16

2.1 Modélisation de l'approche statistique suivant le modèle canal bruité [Och 02] . . . 32
2.2 Modélisation de l'approche statistique suivant le modèle log-linéaire [Och 02] . . . 33
2.3 Un exemple d'alignement avec les mots français . . . 38
2.4 Processus de traduction des modèles 1 et 2 d'IBM [Jur 06] . . . 39
2.5 Processus de traduction simplifié des modèles 3 à 5 d'IBM [Jur 06] . . . 42
2.6 Processus de traduction d'un modèle basé sur les séquences de mots [Jur 06] . . . 44
2.7 Illustration de la méthode d'extraction de séquences suivie par Koehn *et al.* Les alignements Français-Anglais et Anglais-Français sont représentés sous forme de matrices dans lesquelles les cellules non-vides représentent les mots liés au sein des ces alignements. Les cellules pleines noires indiquent les mots qui sont liés dans un alignement comme dans l'autre. Dans la combinaison des alignements, les cadres indiquent les paires de séquences qu'il est possible d'extraire d'après l'alignement des mots. . . . 46
2.8 Exemple de treillis d'options de traduction. A chaque mot ou groupe de mots de la phrase source à traduire correspond un ou plusieurs mots dans la langue cible qui constituent des options de traduction. L'ensemble des options de traduction forme un treillis duquel doit-être extraite la meilleure traduction. . . . 49
2.9 Extensions d'hypothèses pour le décodage de la phrase "*please rise then for this minute 's silence*" . . . 50
2.10 Probabilités de distorsion associées à deux extensions possibles de l'hypothèse de traduction "*ensuite*" par la production des séquences "*minute de silence*" et "*pour cette*" . . . 51
2.11 Gestion des hypothèses partielles sous forme de piles . . . 52

3.1 Extrait de fichiers de sous titres de film . . . 59
3.2 Exemple d'introduction de description . . . 60
3.3 Exemple de différence de segmentation . . . 61
3.4 Exemple d'insertion de texte . . . 62
3.5 Décalage temporelle . . . 63
3.6 Alignement dynamique entre les sous-titres anglais e_i et les sous-titres français f_j d'un même film . . . 64
3.7 Exemple de calcul de la fonction match() . . . 65

4.1 Exemples de triggers classiques 72
4.2 Exemples de triggers inter-langues 72
4.3 Associations directes(traits pleins) et indirectes(traits pointillés) révélés par les triggers inter-langues 75
4.4 Influence de la contrainte de symétrie sur le nombre de triggers inter-langues . . . 77

5.1 Evaluation des traductions produites à l'aide des tables de traduction *Trig-n*, *Sym-n* et *Smooth-n* en fonction de n sur le corpus de développement de EUROPARL . 87
5.2 Evaluation des traductions produites à l'aide des tables de traduction *Trig-n*, *Sym-n* et *Smooth-n* en fonction de n sur le corpus de développement de SSTITRES 87

6.1 Évolution du score BLEU sur les corpus de développement EUROPARL 107
6.2 Évolution du score BLEU sur les corpus de développement SSTITRES 108
6.3 Entropie des distributions de probabilités des tables de traduction 116

Résumé

Les recherches menées dans le cadre de mon doctorat concernent le domaine de la Traduction Automatique (TA), ou comment traduire d'une langue source vers une langue cible sans aucune intervention humaine. Mes travaux se sont plus particulièrement concentrés sur l'approche statistique de la TA qui consiste à utiliser différents modèles probabilistes appris sur des quantités importantes de corpus parallèles alignés afin de trouver la traduction la plus vraisemblable d'une phrase source.

Deux problèmes étroitement liés à l'approche statistique de la TA sont abordés dans ce manuscrit : la collecte de corpus parallèles et l'estimation de modèles de traduction à partir de ces corpus. Un système de TA statistique extrait la connaissance dont il a besoin pour produire des traductions à partir de corpus parallèles dans lesquels chaque phrase source est associée à sa traduction dans la langue cible. De nombreux travaux utilisent comme corpus parallèle les actes du Parlement Européen disponibles gratuitement en différentes langues. De tels corpus ne sont pas adéquats pour la traduction de parole spontanée, c'est pourquoi j'ai décidé de construire des corpus parallèles à partir de sous-titres de films afin de construire un système de traduction plus réaliste. Les sous-titres sont des données complexes, ils ne peuvent constituer un corpus parallèle aligné dans leur état brut. Ils nécessitent une phase de pré-traitement et d'alignement. J'ai pour cela proposé une méthode originale basée sur la Programmation Dynamique qui aligne automatiquement les sous-titres. J'ai ainsi constitué une ressource importante et riche pour l'apprentissage des systèmes de TA statistique.

La TA statistique repose sur l'utilisation de plusieurs modèles statistiques comme le modèle d'alignement, la table de traduction ou encore le modèle de distortion. La table de traduction est le modèle le plus indispensable à un système de TA statisque pour calculer la traduction la plus vraisemblable d'une phrase source. En effet, celle-ci donne les probabilités de traduction entre les couples de mots sources et cibles. Il existe différentes méthodes permettant l'estimation de ces tables de traduction. Elles ont habituellement recours à un alignement des mots obtenu automatiquement sur les corpus parallèles. Cette tâche d'alignement est une étape longue et fastidieuse qui fait appel à des algorithmes complexes. Le coeur de mon travail a été de repenser le problème et d'explorer de nouvelles pistes pour estimer les tables de traduction de mots et de séquences de mots, totalement différentes des méthodes état-de-l'art. J'ai proposé une approche originale basée sur le concept de triggers inter-langues qui ne nécessite aucun alignement des mots au sein des corpus parallèles. Les triggers inter-langues permettent de mettre en évidence des unités fortement corrélés en se basant sur l'Information Mutuelle. Dans notre cas les unités sont des séquences de mots sources et cibles. L'idée derrière ce concept est que si une séquence de mots sources est fortement corrélée à une séquence de mots cibles en termes d'IM, alors nous pouvons supposer que la présence de la première dans une phrase source déclenchera la présence de la seconde dans sa traduction et vice versa. J'ai proposé d'utiliser les triggers inter-langues sur les corpus parallèles dans le but de trouver les traductions possibles de séquences de mots et ainsi constituer une table de traduction. L'Information Mutuelle est une mesure de co-occurence qui se calcule simplement en un seul passage sur le corpus parallèle. Pour sélectionner les triggers inter-langues, nous supposons que deux séquences sources et cibles co-occurent si elles apparaissent dans une même paire de phrases du corpus parallèle. De ce fait, ma méthode ne requiert qu'un alignement au niveau des phrases et non au niveau des mots au sein du corpus parallèle.

L'utilisation des triggers inter-langues pour estimer une table de traduction rend mon approche moins complexe mais tout aussi efficace que les approches existantes. Dans un contexte de traduction mot-à-mot, la table de traduction obtenue grâce aux triggers inter-langues conduit à des traductions automatiques de meilleur qualité, en termes de score BLEU, que celles produites avec une table de traduction de mots estimée selon le modèle 3 d'IBM. Dans un contexte de traduction par groupe de mots, la table de traduction basée sur les triggers inter-langues amènent à des traductions automatiques dont le score BLEU est supérieur à 34 et proche de celui des traductions automatiques produites par une table de traduction de séquences estimées à partir de l'alignement des mots suivant les approches état-de-l'art.

Mots-clés: Traduction Automatique Statistique, Triggers Inter-langues, Traduction Automatique à base de séquences

Abstract

During my Ph.D. study, I conducted research in Machine Translation (MT), i.e. finding a possible target translation of a source sentence without any human interference. My works focused on statistical approach of MT which consists in using different probabilistic models trained on large amount of parallel corpora to retrieve the most likelihood translation given a source sentence.

My thesis addresses two issues related to Statistical Machine Translation (SMT) : the collect of aligned parallel corpora and the estimation of translation models given these corpora. An SMT system extracts the knowledge necessary to perform automatic translation from parallel corpora where each source sentence is aligned with its translation in a target language. Most researches dealing with SMT use as parallel corpora the proceedings of the European Parliament available in many languages. Such corpora are not convenient for spontaneous speech translation. That's why I decided to use movie subtitles in order to achieve a more realistic machine translation system. Movie subtitles are considered as difficult data and cannot be used as parallel corpora for SMT without processing. I proposed an original algorithm based on Dynamic Time Wrapping to automatically align movie subtitles. Thus, I obtained parallel corpora that constitute a rich resource to train SMT system.

In SMT, different statistical models are trained on parallel corpora such as alignment model, translation table, or distortion model. The translation table is the major model needed by an SMT system to perform the process. It gives the translation probability between target and source words. Existing methods usually estimate these tables based on word alignment which is obtained through complex and thus time consuming algorithms.

My principal purpose was to rethink the problem and to prospect new options for generating the translation tables, at word and phrase level, which are totally different from state-of-the-art solutions. I proposed an original approach based on inter-lingual triggers, which does not require any alignment at word level. Inter-lingual triggers allow revealing highly correlated source and target word sequences by computing Mutual Information (MI) between them. The idea behind this concept is that if a source sequence is strongly correlated with a target one in terms of MI then we suppose that the occurrence of the first triggers the occurrence of the last and vice versa. I proposed to use inter-lingual triggers on parallel corpora in order to retrieve probable translations of word sequences and thus constitute a translation table. MI is a co-occurrence measure easily computable in one pass on parallel corpora. For selecting inter-lingual triggers, we assume that two sequences co-occur if they appear in at least one pair of sentences of the

parallel corpora. Thus, the method that I proposed does not require alignment at word level but only at sentence level.
The use of inter-lingual triggers makes my approach to estimate translation tables less complex but as efficient as existing approaches. At word level, the translation table obtained with inter-lingual triggers conducted to automatic translations with better quality, in terms of BLEU score, than those produced with a word translation table estimated by the well-know IBM model 3. At phrase level, the translation table based on inter-lingual triggers leads to automatic translations with a BLEU score greater than 34 and very close to those obtained by a phrase translation table estimated with a state-of-the-art method which requires word alignment on the parallel corpora.

Keywords: Statistical Machine Translation, Inter-lingual Triggers, phrase-based Machine Translation

The Babelfish is small, yellow, leechlike, and probably the oddest thing in the Universe. It feeds on brainwave energy, absorbing unconscious frequencies and excreting a matrix of conscious frequencies to the speech centres of the brain, the practical upshot of which is that if you stick one in your ear, you instantly understand anything said to you in any language.

"The Hitchhikers Guide to the Galaxy" - Douglas Adams

Introduction

Selon le mythe de la tour de Babel, l'existence de plusieurs langues sur notre planète résulte d'une punition que Dieu affligea aux hommes pour les punir de leur orgueil. Nemrod, le roi des descendants de Noé, avait en effet commandé à ses sujets de construire une tour assez haute pour atteindre le ciel et rejoindre Dieu. Tous parlaient alors une seule et même langue. Mais pour les empêcher de mener à bien leur projet, Dieu créa de nouvelles langues. Ainsi les hommes ne parvinrent plus à communiquer, la construction de la tour fût abandonnée et les hommes se dispersèrent partout sur la Terre.

Bien que ce ne soit qu'un mythe, l'existence de plusieurs langues est un fait et crée un véritable obstacle entre les êtres humains désirant communiquer. Bien sûr, l'apprentissage de langues étrangères est toujours possible mais il est souvent limité à deux ou trois langues, sauf pour de rares exceptions comme Ionnis Ikonomou, par exemple, fonctionnaire européen à Bruxelles travaillant au département traduction, qui détient le record de l'Homme qui parle le plus de langues au monde. Il est en effet capable de s'exprimer en 42 langues. Cependant, même pour ces rares exceptions, cette connaissance avancée d'autres langues n'est pas suffisante pour ne jamais être confronté à la barrière de la langue. En effet, plus de 6800 langues sont recensées sur notre planète et tout le monde peut un jour se heurter à cette barrière. L'apprentissage des langues étant limité, un autre moyen de franchir cet obstacle est le recours à la traduction.

La barrière de la langue se fait ressentir de plus en plus de par la mondialisation des échanges, mais aussi le tourisme international, ou encore l'accroissement constant des données multilingues sur l'Internet. Les besoins en traduction ont donc considérablement augmentés depuis ces dernières décennies et la nécessité de pouvoir traduire automatiquement et à tout moment toutes sortes d'informations est devenue indéniable.

La traduction est une tâche complexe. Traduire ne consiste pas simplement à prendre chaque mot d'une phrase et à le remplacer par sa traduction dans une autre langue. Il faut faire face à de nombreuses difficultés comme les ambiguïtés que peuvent comporter les composants d'une phrase. Un même mot, par exemple, peut appartenir à plusieurs catégories grammaticales, il est alors question d'ambiguïté grammaticale. Le mot français *souris* en est une parfaite illustration. Il peut à la fois être un nom ou alors un verbe et il est nécessaire de connaître sa catégorie pour le traduire correctement. D'autres mots encore, peuvent avoir plusieurs sens, et l'ambiguïté sera alors dite sémantique. C'est le cas du verbe français *voler* qui peut signifier dérober quelque chose ou voler comme un oiseau. Ces ambiguïtés doivent être levées avant de pouvoir prendre une décision quant à la bonne traduction des mots. Une autre grande difficulté est qu'une traduction ne s'opère pas toujours mot-à-mot. Il faut parfois considérer un ensemble de mots pour en trouver une bonne traduction plutôt que chaque mot pris séparément. L'expression française *appeler un chat un chat* se traduit par exemple en anglais *call a spade a spade* et non pas *call a*

cat a cat, *rat de bibliothèque* se traduit en Anglais par *bookworm* et non par *rat of library*. D'autres problèmes liés à la structure de la langue sont également à prendre en compte. La place du verbe par exemple ne sera pas la même en Français qu'en Allemand où il se place toujours en première, deuxième ou dernière position selon son rôle et la nature de la préposition. Tous ces problèmes rendent la tâche de traduction difficile, aussi bien pour des traducteurs professionnels que pour des machines.

Si par le passé, faire traduire un texte par une machine relevait de l'utopie, les avancées technologiques et les efforts constants de nombreux chercheurs en Traduction Automatique (TA) font que cela paraît aujourd'hui possible. La TA désigne le fait de faire traduire un texte d'une langue source vers une langue cible par une machine sans aucune intervention humaine. C'est dans ce cadre de Traduction Automatique que se situent les travaux présentés dans ce manuscrit qui s'organise en deux parties : une première partie présentant l'état de l'art de la Traduction Automatique et une seconde partie dédiée à nos contributions en TA statistique.

La recherche en Traduction Automatique, qui connaît un important essor depuis les années 90, se scinde en deux grandes écoles : une école au sein de laquelle la machine tire les connaissances nécessaires pour accomplir la tâche de traduction à partir de règles établies par des experts humains, il est alors question d'approche experte et une école pour laquelle la machine extrait les connaissances dont elle a besoin à partir de grande quantité de données textuelles, il s'agit de l'approche empirique. Le chapitre 1 présente succinctement chacune de ces deux écoles ainsi que les principales méthodes qu'elles emploient pour traduire un texte d'une langue source vers une langue cible, avec leurs avantages et leurs inconvénients.

L'approche statistique de la Traduction Automatique appartient à la famille des méthodes empiriques. Il s'agit pour la machine de traduire en utilisant des modèles statistiques appris automatiquement sur de grands corpus monolingues et bilingues. Bien que peu convaincante à ses débuts, l'approche statistique a rapidement rencontré un fort succès de par sa mise en place de façon totalement automatique mais aussi de par la qualité satisfaisante et à moindre coût des traductions automatiques produites. L'équipe de recherche à laquelle j'appartiens possède une solide expérience en apprentissage statistique pour la reconnaissance automatique de la Parole et la modélisation du langage, c'est donc tout naturellement que mes travaux en TA se sont tournés vers l'approche statistique.

Les travaux de cette thèse constituent les premiers pas en TA statistique au sein de notre équipe. Comme cela est présenté plus en détails dans le chapitre 2, l'approche statistique de la traduction repose essentiellement sur deux modèles : un modèle de langage de la langue cible et un modèle de traduction. Le modèle de langage est appris sur des corpus monolingues, il rend compte de la bonne construction des traductions produites dans la langue cible en assignant des probabilités à des suites de mots. Le modèle de traduction est appris sur des corpus parallèles constitués de phrases dans une langue source associées à leur traduction dans une langue cible. Il permet de rendre compte de la vraisemblance des traductions produites compte tenu des phrases sources à traduire. C'est ensuite à un décodeur que revient la lourde tâche de calculer la traduction la plus probable d'une phrase d'entrée source en faisant usage des paramètres de ces deux modèles. Parmi les nombreux paramètres d'un modèle de traduction, le décodeur va essentiellement utiliser la table de traduction qui peut contenir plusieurs millions de paires de séquences, chaque paire étant associée à la probabilité qu'une des deux séquences soit la traduction de l'autre. Le chapitre 2 s'attelle à présenter les principaux modèles de langage et de traduction

utilisés dans la communauté scientifique. Il décrit notamment des méthodes de construction de tables de traduction de séquences. Après avoir exposé l'algorithme de recherche de la meilleure traduction suivi par le décodeur que nous utiliserons pour nos travaux, le chapitre 2 se termine sur une discussion générale où seront notamment soulevées les faiblesses des modèles de traduction existants.

La deuxième partie du manuscrit est consacrée à nos contributions en TA statistique. Celles-ci touchent l'ensemble des étapes essentielles pour la mise en place d'un système de traduction automatique statistique. Notre première contribution concerne la construction automatique de corpus parallèles. En effet, une des principales difficultés rencontrées pour l'apprentissage de modèles de traduction statistiques est de trouver des corpus parallèles bilingues dans lesquels chaque phrase dans une langue source est associée à sa traduction dans une langue cible. De tels corpus ne sont pas facilement disponibles, les plus souvent utilisés sont les actes du Parlement Européen, ils sont présentés dans le chapitre 3. Dans le reste de ce chapitre, nous proposons une méthode originale permettant de mettre en correspondance des sous-titres de films dans des langues différentes et ainsi constituer des corpus parallèles. Les sous-titres de films sont des ressources très abondantes et qui s'apparentent d'avantage à la parole spontanée que le langage soutenu employé durant les commissions du Parlement Européen. Ils constituent donc un meilleur matériel pour l'apprentissage de modèles dédiés à un système de Traduction grand public.

Les chapitres suivants présentent le cœur de nos travaux, à savoir, l'utilisation des triggers inter-langues pour la Traduction Automatique. Les triggers inter-langues s'inspire du concept de triggers largement utilisé en modélisation du langage. Nous les utilisons dans un contexte bilingue de manière tout à fait originale. Nous concentrons, en effet, nos efforts sur la mise en place d'une nouvelle méthode, fondée sur les triggers inter-langues et permettant d'établir des correspondances entre séquences sources et cibles et ainsi générer des tables de traduction nécessaires au bon fonctionnement du décodeur au sein d'un système de Traduction Automatique. Contrairement aux approches les plus couramment utilisées, notre méthode ne nécessite pas d'alignement des mots au sein des paires de traduction du corpus durant la phase d'apprentissage, lequel constitue une tâche longue et difficile. Notre approche se base uniquement sur le concept de triggers inter-langues.

Nous introduisons, au chapitre 4, le concept de triggers inter-langues. Nous étudions notamment son potentiel à mettre en évidence des correspondances lexicales de mots. Pour cela, nous mettons en place un dictionnaire bilingue, obtenu à partir des triggers inter-langues, que nous comparons avec des dictionnaires de référence.

Dans la suite de ce manuscrit, nous exploitons les triggers inter-langues dans le cadre de la Traduction Automatique. Nous proposons en effet de les utiliser pour construire des tables de traduction permettant au décodeur de produire des traductions automatiques dans une approche par mot puis dans une approche par groupe de mots.
Le chapitre 5 s'inscrit dans le cadre de la TA par mot. Nous décrivons différentes méthodes pour l'apprentissage de table de traduction de mots à partir des triggers inter-langues avant de les évaluer dans un processus de TA en terme de qualité des traductions produites par le décodeur. L'évaluation des traductions est un problème délicat que nous abordons brièvement en début de manuscrit mais qui constitue un axe de recherche ouvert. Nous en donnons un bref aperçu en insistant sur l'importance de la mise en place de métriques d'évaluations automatiques tant la tâche d'évaluation des traductions par des experts humains est coûteuse et fastidieuse. Pour nos

travaux, nous avons choisi le score BLEU pour évaluer la qualité des traductions automatiques produites au cours de nos expérimentations.

Dans le chapitre 6, nous nous plaçons cette fois dans un contexte de TA par groupe de mots (ou séquences). Nous exposons une approche originale pour extraire une table de traduction de séquences toujours à partir des triggers inter-langues et sans passer par l'alignement des mots dans le corpus bilingue, comme c'est le cas pour de nombreuses méthodes proposées jusqu'ici dans le domaine. Par la suite, ce chapitre est concentré sur un algorithme de sélection des meilleures paires de traduction de séquence au sein la table de traduction. Le reste du chapitre présente l'évaluation des tables de traduction de séquences (avec et sans sélection) fondées sur les triggers inter-langues. Une discussion sur notre approche et son positionnement par rapport à l'état de l'art vient clore le chapitre.

Nous terminons ce manuscrit par une conclusion sur nos travaux ainsi que des perspectives de recherche à court et long terme.

Première partie

Etat de l'art

Chapitre 1

La traduction automatique et son évaluation

1.1 Introduction

Dans ce chapitre, nous présentons la Traduction Automatique (TA) dans son ensemble en évoquant dans un premier temps les premiers travaux marquants de recherche la concernant. Nous poursuivons ensuite par la description des deux grandes approches la caractérisant : l'approche experte reposant sur des connaissances et l'approche empirique qui tire ses connaissances à partir de grandes quantités de données. Nous présentons pour chacune de ces deux approches une série de méthodes permettant de traduire d'une langue vers une autre. L'approche statistique de la TA n'est que brièvement présentée dans ce chapitre. En effet, nos travaux de thèse portant sur cette approche, le chapitre suivant lui est spécialement dédié. Après le survol des différentes méthodes existantes en TA, nous évoquons ensuite le problème de l'évaluation des traductions automatiques. Evaluer la qualité d'une traduction automatique est primordiale pour la recherche en TA et constitue un défi tout aussi intéressant que celui de la produire.

1.2 Un bref historique

Le Français Georges Artsrouni et le Russe Petr Trojanskij sont les premiers, en 1933, à émettre de réelles idées pour la Traduction Automatique. La première démonstration d'un système de traduction remonte elle, à janvier 1954. Elle fait suite à un projet mené en collaboration par Leon Dostert de l'université de Georgetown et par IBM. Quarante-neuf phrases russes méticuleusement choisies sont alors traduites automatiquement en anglais. Ce tout premier système supporte un vocabulaire de seulement deux cent cinquante mots et utilise en tout et pour tout six règles de grammaire. Bien qu'aujourd'hui un tel système puisse paraître très limité, il fait à l'époque forte impression et met en avant la faisabilité de la Traduction Automatique. Des sommes d'argent considérables sont alors investies et plusieurs projets partout dans le monde émergent. Malheureusement, en 1966 les conclusions du rapport de l'ALPAC (Automatic Language Processing Advisory Committee) mettent un terme à cet enthousiasme et signe la fin des subventions. Elles sont sans équivoque : la Traduction Automatique est lente, moins performante et deux fois plus coûteuse que les traductions réalisées par des experts humains. Bien que la recherche n'ait jamais vraiment été interrompue, il faudra attendre la fin des années soixante-dix et les progrès en Informatique et en traitement des langues pour constater une reprise sérieuse des travaux en Traduction Automatique.

1.3 Les différentes approches

La Traduction Automatique désigne le fait de traduire un texte d'une langue source vers une langue cible sans aucune intervention humaine. Depuis ses premiers pas, deux grandes approches coexistent : une première approche, dite experte, fondée sur les connaissances d'experts humains et une seconde approche, dite empirique, qui extrait les connaissances à partir de quantités importantes de données textuelles.
Le processus naturel de traduction peut être vu comme la succession de trois étapes. La première étape est l'analyse du texte source où l'être humain cherche à définir comment le structurer ou le segmenter pour le traduire. Une fois le texte source structuré, la deuxième étape est le transfert vers la langue cible. Ce dernier consiste à appliquer des règles permettant de passer de la langue source à la langue cible. Enfin, une fois le transfert effectué, la troisième étape est la génération d'une traduction en langue cible du texte source initial. La triangle de Vauquois schématisé dans la figure 1.1 représente ces trois étapes d'analyse, de transfert et de génération. Chacune des

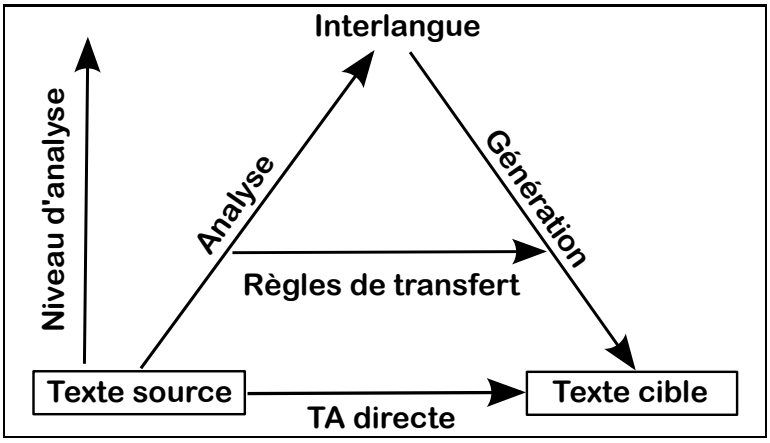

Fig. 1.1 – Le triangle de Vauquois

approches envisagées, quelle soit experte ou empirique, traite ces étapes de façon particulière avec des niveaux d'analyse, de transfert et de génération différents. Nous décrivons dans les sections suivantes les principales méthodes de TA issues de chacune des deux approches.

1.3.1 L'approche experte

L'approche experte repose sur les connaissances de professionnels humains. Ces connaissances sont utilisées durant chaque étape du processus de traduction. Lors de l'analyse d'abord, la phrase source est segmentée en unités liées les unes aux autres. Selon le niveau d'analyse, ces unités peuvent être morpho-syntaxiques, grammaticales ou encore sémantiques. La phase de transfert se charge ensuite de convertir les unités sources en unités cibles en utilisant notamment des règles et des dictionnaires bilingues. Le module de génération, enfin, transforme les unités cibles en texte cible. Les trois principales méthodes dérivées de cette approche sont la Traduction Automatique (TA) directe, la TA à base de règles de transfert et la TA fondée sur l'interlangue. Elles sont détaillées dans les sections suivantes.

1.3.1.1 TA directe

L'approche directe de la TA correspond à la base du triangle de Vauquois. La traduction s'opère ici mot à mot et ne nécessite qu'une analyse partielle de la phrase source. Le processus de traduction repose essentiellement sur l'utilisation d'un grand dictionnaire bilingue qui va spécifier pour chaque mot source un emsemble de règles permettant de le traduire. Les mots de la phrase source, une fois traduits, sont ensuite ré-ordonnancés à l'aide de simples règles afin de générer une traduction. Le procédé est illustré par la figure 1.2.
L'approche directe de la TA est simple à mettre en place en ce sens qu'elle ne nécessite pas de

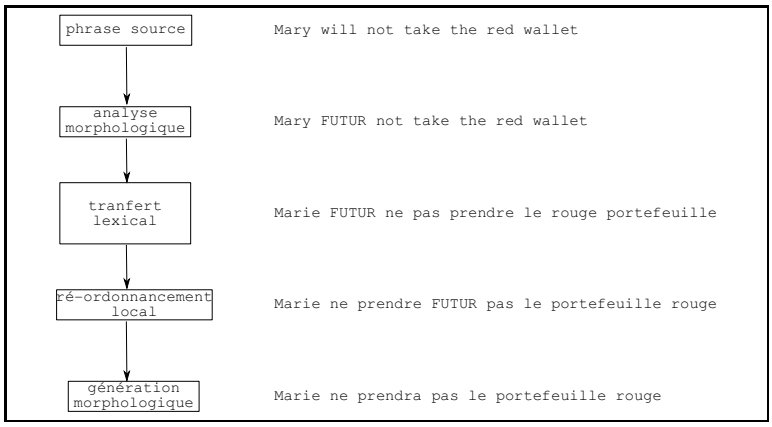

FIG. 1.2 – TA directe [Jur 06]

connaissances appronfondies ni de la langue source, ni de la langue cible. Elle ne fait appel qu'à des régles primaires. L'analyse seulement partielle de la phrase source présente l'avantage de permettre au système de toujours fournir une traduction de la phrase source même si cette dernière n'est pas grammaticalement correcte. En effet, les connaissances utilisées ne sont pas suffisamment importantes pour détecter un phénomène incorrect grammaticalement. Si une structure grammaticale n'est pas reconnue lors de l'analyse de la phrase source, elle est écrite directement en mots cibles sans ré-ordonnancement.
Cependant, le manque d'analyse de la phrase source ainsi que l'absence de lien entre les représentations source et cible du texte à traduire diminuent fortement la fiabilité des résultats obtenus. De plus, la TA directe ne possédant pas de grammaire explicite pour la langue cible, contrairement aux approches dites indirectes (règles de transfert, interlangue), la bonne construction de la phrase, du point de vue de la langue cible n'est pas garantie.
Dans les années 80, des systèmes de traduction tels que SYSTRAN [Wheeler 84] à ces débuts et PAHO [Vasconcellos 85] utilisaient l'approche directe. Aujourd'hui, du fait de sa simplicité et de part la complexité des langues, l'approche directe n'est plus guère utilisée pour la TA grande échelle. Toutefois, elle peut s'avérer efficace et appropriée dans le cadre d'applications spéficifiques pour lesquelles le vocabulaire est restreint et standardisé et où une substitution de mots peut suffir à produire une traduction correcte. Nous pouvons notamment citer le système METEO [Chandioux 96] qui traduit des bulletins météorologiques avec un vocabulaire restreint d'un millier de mots mais aussi le grand fabricant d'appareils médicaux Medtronics [Uwe 06] qui utilise un système de TA directe pour traduire dans plusieurs langues sa base de données de produits.

1.3.1.2 TA à base de règles de transfert

Contrairement à l'approche directe, l'approche à base de règles de transfert requiert un niveau d'analyse de la phrase source plus poussé. En effet, comme l'illustre la figure 1.3, la phrase source est analysée à l'aide d'un analyseur syntaxique et d'une grammaire. Cette phase d'analyse donne lieu à une représentation arborescente qui est ensuite convertie dans la langue cible. Cette étape de transfert, lors du passage de la langue source vers la langue cible fait intervenir des règles bilingues supposées être applicables dans les deux sens source-cible et cible-source. Il s'agit en

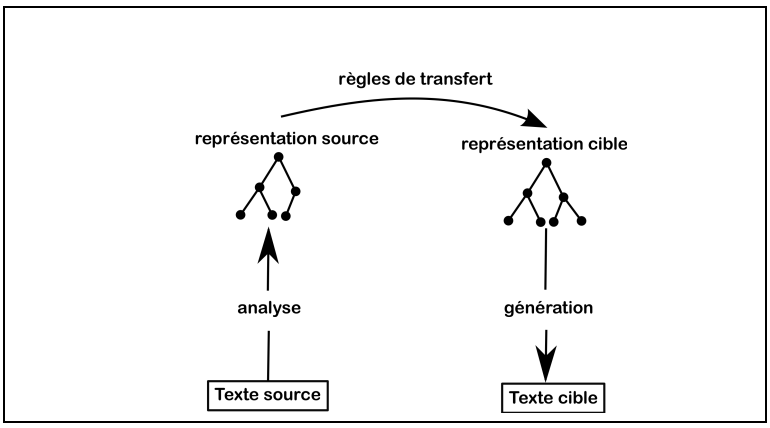

FIG. 1.3 – TA à base de règles de transfert [Jur 06]

fait de règles de correspondance entre les nœuds de l'arbre source et les nœuds de l'arbre cible. Elles peuvent être syntaxiques, lexicales ou encore sémantiques. Elles assurent la transformation entre la représentation de la phrase source et sa représentation dans la langue cible. Enfin, une grammaire de la langue cible va générer la phrase cible à partir de cette représentation. Un cas simple de traduction à base de règles de transfert est illustré dans la figure 1.4. Dans cet exemple,

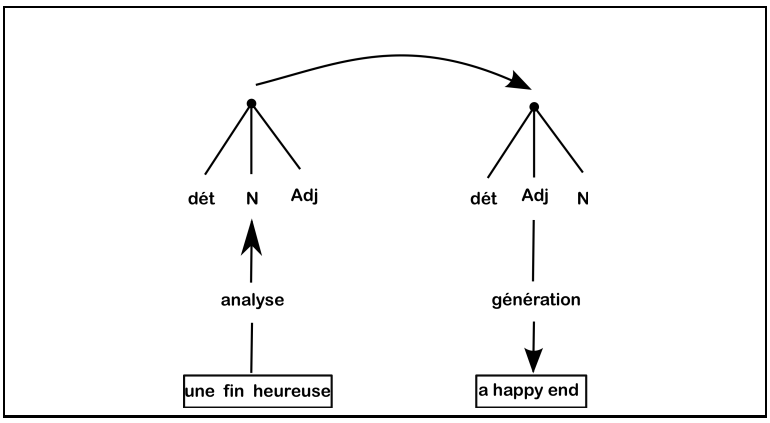

FIG. 1.4 – Exemple simple de TA à base de règles de transfert du Français vers l'Anglais de la phrase "*une fin heureuse*"

l'analyse faite de la phrase source est simplement syntaxique. Cependant, suivant les langues engagées dans la traduction, l'analyse peut être beaucoup plus profonde. Pour une traduction de l'Anglais vers le Chinois par exemple, il est nécessaire de connaître le rôle sémantique des groupes

prépositionnels afin de savoir s'il se place avant ou après le verbe en chinois. En effet, si le groupe prépositionnel exprime une entité qui bénéficie de l'action induite par le verbe, celui-ci sera placé avant le verbe. A l'inverse, s'il exprime une entité passivement impliquée dans les événements dénotés par le verbe, il sera plutôt placé après le verbe.

La TA à base de règles de transfert implique une connaissance avancée de la langue source et de la langue cible ainsi que des liens existant entre ces deux langues. L'inconvénient majeur de cette méthode réside donc en la difficulté de formaliser toute la grammaire réelle d'une langue, ainsi que les grammaires de transfert. Une phrase source trop complexe grammaticalement, par exemple, peut conduire à l'échec du processus de traduction si la grammaire de la langue source n'a pas été suffisamment décrite.
Les systèmes de TA à base de règles sont relativement simples à mettre en place et facilement adaptables pour de nouvelles paires de langues. En effet, les langues sont des modules bien séparés et indépendants. Permettre à un système de traduire vers ou depuis une nouvelle langue, revient donc simplement à ajouter une grammaire de cette nouvelle langue et des grammaires de transfert.

De part le manque de formalisme dans les grammaires, cette approche, au départ, a surtout montré de bons résultats dans des domaines restreints comme par exemple pour la réservation d'hôtels. Mais la disponibilité des grammaires et de lexiques de grande couverture, en fait de nos jours un des paradigmes majeurs en Traduction Automatique. De nombreux systèmes de traduction automatique basée sur cette approche sont disponibles aujourd'hui. C'est notamment le cas de SYSTRAN [Senellart 01], de Babelfish [1] ou encore de PROMT [2] pour n'en citer que quelques uns.

1.3.1.3 TA fondée sur une interlangue

Comme le montre la figure 1.1, l'analyse de la phrase source se fait à des niveaux plus ou moins profonds. Plus l'analyse de la phrase est poussée, plus la représentation en devient abstraite. Le but de la TA fondée sur l'interlangue est de trouver une analyse telle que le transfert des représentations de la langue source vers la langue cible n'aurait plus lieu d'être. A un tel niveau de représentation, la sortie du composant d'analyse du texte source serait directement l'entrée du composant de génération du texte cible sans passer par l'application de règles de transfert. Cette méthode tend vers une représentation du sens de la phrase indépendante de la langue utilisée d'où le terme d'approche interlangue. Et la différence par rapport à la TA à base de règles de transfert tient plus du degré de représentation que d'une réelle distinction.

La principale difficulté de cette méthode est de trouver un vocabulaire de tous les concepts possibles pour l'interlangue. En effet, pour cela il faut trouver les concepts primitifs de la représentation du sens des phrases quelles que soient les langues considérées. Les chercheurs dans ce domaine sont confrontés au problème d'identification de ces concepts primitifs. Il faut définir tous les concepts possibles selon toutes les langues et ceci peut conduire à la génération d'informations inutiles dans certains cas. Il se peut notamment qu'il y ait des différences entre certains concepts de l'interlangue qui n'ont pas lieu d'être. Par exemple, le concept *frère* correspond en Japonais à deux concepts selon que le frère soit une personne plus jeune ou plus vieille. Cette différence doit être spécifiée au sein de l'interlangue même si elle est inutile pour traduire du Français vers

[1] http ://fr.babelfish.yahoo.com
[2] http ://www.online-translator.com/ ?prmtlang=fr

l'Anglais.
L'avantage principal que présente la TA fondée sur l'interlangue, une fois celle-ci bien définie, est la facilité à mettre en place un système de traduction pour de nouvelles paires de langues. Contrairement à la méthode à base de règles de transfert qui nécessite une grammaire pour la langue cible, une pour la source et deux grammaires de transfert, l'ajout d'une nouvelle paire de langues dans le cas d'un système interlangue ne requiert qu'une grammaire supplémentaire pour chacune des deux langues.

Aujourd'hui, le seul système opérationnel basé sur l'interlangue est celui résultant du projet KANT (Knowledge-based Accurate Natural-language Translation) [T. Mitamura 91]. Un autre grand projet, aujourd'hui terminé, suivant cette approche était le projet DLT (Distributed Language Translation) [Witkam 88] qui utilisait notamment l'Esperanto comme interlangue au sein de son système comprenant douze langues européennes.

1.3.2 L'approche empirique

L'approche empirique est née des suites de l'abondance croissante de données textuelles. De nombreux chercheurs ont vu, au travers de cette recrudescence, une porte ouverte à l'élaboration d'un nouveau type d'architectures dans lequel la connaissance nécessaire à la traduction n'émane plus d'experts linguistiques mais provient directement de données réelles. Deux grandes familles se distinguent alors : l'approche par analogie, également appelée approche à base d'exemples, et l'approche statistique. Toutes deux tirent profit des données brutes et contrairement aux méthodes de l'approche experte, ne nécessitent aucune connaissance a priori pour mettre en place un système de traduction.
Dans ce qui suit, nous présentons l'approche par analogie puis nous introduisons brièvement l'approche statistique qui sera détaillée par la suite dans le chapitre 2.

1.3.2.1 L'approche par analogie ou basée sur l'exemple

L'approche par analogie, appelée aussi approche à base d'exemples, a été proposée en 1984 par Makoto Nagao [Nagao 84]. Ce dernier part de l'idée que pour traduire automatiquement un texte, il faut modéliser le processus effectué par les êtres humains. En effet, le processus naturel de traduction implique souvent de se référer à un exemple. Un être humain pour traduire une phrase ou une expression d'une langue source vers une langue cible va s'aider d'exemples qu'il a déjà rencontrés. Il va les adapter et s'en servir comme modèles pour trouver la traduction de ce qu'il recherche.

L'approche basée sur l'exemple s'appuie sur ce constat que pour traduire, il est possible de se référer à des traductions déjà connues et de procéder ensuite par analogie. Dans un premier temps, il est nécessaire de constituer une base d'exemples importantes. Cette base est composée d'un ensemble de séquences sources associées à leur traduction dans la langue cible. A partir de cette base, le processus de traduction se déroule en trois étapes.
La première phase, dite de correspondance ou de *matching*, sélectionne dans la base d'exemples, la ou les séquences sources qui se rapprochent le plus de la phrase source à traduire. Plusieurs méthodes de sélection existent et diffèrent sur plusieurs points notamment sur le calcul de similarité entre la phrase source à traduire et les exemples de la base. En effet, le calcul de similarité peut porter sur la phrase entière ou au contraire sur des segments de la phrase. Veale *et al.* [Veale 97], par exemple, proposent de trouver la phrase source du corpus d'exemples qui est la

plus semblable à la phrase source à traduire. Brown [Brown 96] lui, fragmente la phrase source en séquences et va rechercher chacune d'entre elles dans le corpus d'exemples. Dans d'autres travaux encore, le calcul de similarité ne se fait pas au niveau des mots constituant la phrase mais au niveau de l'analyse de celle-ci. Par exemple, dans [Nagao 84, Deniz 08], la phrase source à traduire ainsi que les phrases sources du corpus d'exemples sont sousmises à un analyseur et les arbres résultants sont ensuite comparés. Il est nécessaire dans ce cas de faire appel à des connaissances expertes.
La deuxième phase extrait de la base les séquences cibles associées aux séquences sources retenues, il s'agit ici de la phase d'alignement. Enfin la troisième phase d'adaptation arrange et combine ces séquences cibles de manière à générer une traduction de la phrase source. Dans le système PANGLOSS notamment [Brown 96], Brown utilise un modèle de langage pour arranger les séquences cibles au mieux.

Au départ proposée en renfort pour combler les lacunes des systèmes à base de règles, l'approche par analogie devient rapidement une méthode à part entière, faisant ainsi concurrence aux méthodes expertes. Les avantages de cette approche sont nombreux. Tout d'abord, elle s'appuie sur une collection d'exemples de traduction réelle. Par conséquent, contrairement à un système expert qui construit ses traductions seulement à partir de règles, un système fondé sur l'exemple produit ses traductions d'après de véritables exemples, ce qui devrait garantir une meilleure qualité de traduction.

Ensuite, un moyen simple et efficace d'améliorer les performances d'un système de TA fondé sur cette approche, est d'augmenter la taille de la base d'exemples.

Certains travaux proposent notamment des méthodes de généralisation des exemples de traduction pour augmenter la taille de la base [Veale 97, Brown 00, Gangadharaiah 06].

Bien que souvent référencée comme approche empirique, nous pensons que la méthode basée sur l'exemple relève plus d'une méthode hybride que d'une approche purement empirique. En effet, à chaque étape du processus de traduction peuvent intervenir des règles expertes ou bien encore des scores probabilistes comme dans les méthodes statistiques [Hutchins 05]. De nombreux systèmes à base d'exemples emploient actuellement pour leur développement des méthodes tirées de l'approche experte et de l'approche empirique [Langlais 06]. De ce fait, nous pensons que pour être réellement efficace, l'approche fondée sur l'exemple nécessite, en plus de corpus bilingues des connaissances *a priori* des langues impliquées dans la traduction.

1.3.2.2 L'approche statistique

L'approche statistique n'est que briévement abordée dans cette section. En effet, les travaux de cette thèse portant sur cette approche, le chapitre suivant lui est entièrement dédiée.

L'approche statistique de la TA consiste à trouver la traduction la plus probable d'une phrase source selon deux principaux modèles probabilistes : un modèle de traduction et un modèle de langage [Brown 93].
L'apprentissage du modèle de traduction s'opère à partir de corpus bilingues alignés. De tels corpus sont obtenus en alignant chaque segment d'un corpus source avec sa traduction dans le corpus cible. L'alignement peut se faire à différents niveaux comme la phrase, le paragraphe ou encore le document. Le modèle de langage est quant à lui appris sur un corpus de la langue cible. L'approche statistique consiste alors à dériver, par des méthodes statistiques, le modèle de traduction et de langage à partir des fréquences d'événements dans le corpus aligné et dans le

corpus monolingue.
Le modèle de traduction permet de trouver les traductions les plus probables de groupes de mots. Il peut être vu comme un dictionnaire bilingue dans lequel chaque entrée met en relation de traduction des groupes de mots sources et avec des groupes de mots cibles. Chaque association entre un groupe de mot et sa traduction se voit attribuer une probabilité. Le modèle de langage, quant à lui, assigne des probabilités à des suites de mots dans la langue cible. C'est lui qui va s'assurer de la bonne construction de la traduction produite.

L'approche statistique, contrairement aux approches expertes, extrait ses connaissances à partir de données textuelles et aborde les trois étapes du processus de traduction de façon particulière. En effet, l'analyse de la phrase source consiste à mettre en correspondance les mots de la phrase à traduire avec les séquences sources présentes dans la table de traduction. Durant la phase de conversion, des mots cibles sont choisis relativement aux couples sélectionnés dans la table. La phrase cible est ensuite générée en arrangeant, grâce à un modèle de langage, les mots cibles trouvés.

Les systèmes de traductions statistiques connaissent un véritable succès depuis plusieurs années. Le système Google Translate a notamment remporté le rang de meilleur système de Traduction lors de la campagne d'évaluation NIST 2006 [3] chargée de mettre en compitition différents systèmes de traduction des deux approches experte ou empirique. La société Language Weaver [4], tout comme Microsoft[5] , commercialise également un logiciel de traduction automatique reposant sur l'approche statistique. De plus, la mise en place d'un système de traduction statistique est rendue très accessible notamment grâce à la disponiblité de plusieurs systèmes de traduction en code source libre tels que MOSES [Koehn 07b], ou encore JOSHUA [Li 09].

1.3.3 Discussion

Nous venons de voir les deux grandes écoles existantes en Traduction Automatique : l'approche experte dont les connaissances nécessaires au processus de traduction émanent de règles manuellement apportées par des experts humains et l'approche empirique qui tire automatiquement les connaissances dont elle a besoin à partir de quantités importantes de données textuelles.

Selon l'approche experte de la TA, et plus particulièrement pour l'approche à base de règles, les traductions sont construites à l'aide d'importants dictionnaires et de règles linguistiques sophistiquées. Les systèmes de TA à base de règles fournissent un bon niveau de qualité de traduction dans des domaines non-spécifiques, même si le manque de fluidité dans les traductions leur est souvent reproché. L'adaptation à un domaine spécifique est possible en ajoutant la terminologie adéquate dans les dictionnaires et les règles linguistiques. Toutefois, cette adaptation représente un coût important en terme de temps et de moyens.

Selon l'approche empirique, les traductions sont produites à partir d'événements rencontrés dans des corpus bilingues et à partir desquels le système va puiser ses connaissances. Dans le cadre plus spécifique de la TA statistique, ce sont des modèles de traduction et de langage appris sur les corpus qui vont permettre de construire une traduction sachant une phrase source à traduire. L'apprentissage des modèles est un processus rapide, automatique et peu coûteux

[3] http ://www.itl.nist.gov/iad/mig//tests/mt/2006/doc/mt06eval_official_results.html
[4] http ://www.languageweaver.com/page/home/
[5] http ://research.microsoft.com/enus/projects/mt/

mais qui nécessite une quantité importante de corpus bilingue pour fournir des traductions de bonne qualité. Plus il y a de données d'apprentissage, plus le système fournira des traductions satisfaisantes. Les traductions automatiques par un système de TA statistique sont faciles à lire, toutefois, de part le processus totalement automatique, elles peuvent parfois être imprévisible et manquer de cohérence.

La frontière entre les méthodes de ces deux approches n'est toutefois pas étanche. De plus en plus de systèmes de traduction tirent profit de chacune d'entre elles en les combinant. De tels systèmes sont qualifiés de systèmes hybrides. Nous avons vu par exemple qu'un système de TA fondé sur l'exemple pouvait faire appel à des règles linguistiques pour améliorer la qualité des traductions produites [Langlais 06]. Nous pouvons également citer le système Matrex [Tinsley 08] qui fait appel à la fois à l'approche statistique et à l'approche par analogie. Plus récemment, la compagnie SYSTRAN a mis au point un moteur de traduction hybride combinant la technologie à base de règles et le traitement statistique de données. Ce système [Schwenk 09] a été classé premier en 2009 pour la qualité de ses traductions de l'Anglais vers le Français lors de la campagne d'évaluation "Workshop on Statistical Machine Translation". Il a notamment devancé le système purement statistique de Google.

Notre équipe de recherche possède une solide expérience en apprentissage statistique qui a notamment démontré de bonnes capacités en reconnaissance automatique et en modélisation du langage. C'est pourquoi nous nous sommes naturellement tournés vers l'approche statistique de la traduction automatique dont ce manuscrit fait l'objet.

Quelles soient les méthodes envisagées, chaque système de traduction automatique se doit de confronter les traductions qu'il produit à une évaluation. En effet, il est nécessaire de pouvoir juger d'une traduction automatique si elle est de bonne ou de mauvaise qualité ne serait-ce que pour pouvoir améliorer un système. Nous évoquons ce problème dans la section suivante.

1.4 Évaluation de la qualité des traductions

Si la traduction automatique demeure toujours un défi scientifique à relever, il en va de même pour son évaluation. La nécessité de juger de la qualité des traductions automatiques s'est vite présentée, d'abord pour pouvoir améliorer les sorties d'un système de traduction mais aussi afin de comparer les systèmes entre eux notamment au cours des nombreuses campagnes d'évaluation en Traduction Automatique. Ce travail a d'abord été fait manuellement grâce aux jugements d'experts humains. Mais au vu de la quantité de données de plus en plus importante à évaluer, il a été nécessaire de mettre au point des techniques d'évaluation automatique. Dans ce qui suit, nous présentons quelques une des grandes campagnes d'évaluation en TA. Nous décrivons ensuite les différents critères de qualité d'une traduction avant de poursuivre sur les moyens existants pour les évaluer.

1.4.1 Campagnes d'évaluation

Il existe de nombreuses campagnes d'évaluation en Traduction Automatique. En plus de favoriser les échanges scientifiques et le travail coopératif, elles permettent d'évaluer, par le moyen de métriques automatiques mais aussi manuelles, la qualité des traductions produites par différents systèmes. Elles offrent ainsi la possibilité aux différents groupes internationaux commerciaux et de recherche en TA de se confronter les uns aux autres. Nous présentons ici briévement quelques

unes de ces campagnes d'évalution.

Deux campagnes d'évaluation menées dans le cadre du projet CESTA (Campagne d'évaluation des systèmes de traduction automatique) en 2005 et 2006 ont permis de réaliser des évaluations automatiques de systèmes de TA et de les comparer à des jugements humains [Hamon 08]. Nous pouvons également citer les campagnes d'évaluation NIST Open Machine Translation, qui se déroulent tous les ans depuis 2001, afin de promouvoir et aider la recherche en TA textuel. L'objectif lors de ces campagnes est de fournir des traductions automatiques qui soient en adéquation avec les traductions originales humaines. L'institut NIST propose également une série de workshops intitulée MetricsMATR visant à développer de nouvelles mesures automatiques pour l'évaluation des traductions. Le workshop WMT (Workshop on Statistical Machine Translation) met également en place depuis 2005 des campagnes d'évaluation dédiées aux langues européennes. En 2009, la campagne consistait en trois tâches : traduction automatique pour 10 langues européennes, évaluation automatique des traductions produites et enfin une nouvelle tâche visant à établir un nouveau système de traduction en combinant les productions de chaque système participant à la campagne [Callison-Burch 09].

D'autres campagnes s'intéressent plus particulièrement à la Traduction Parole Parole. C'est le cas des campagnes du projet TC-STAR et du workshop IWSLT (International Workshop on Spoken Language Translation). Le projet européen TC-STAR organise annuellement une campagne d'évaluation portant sur la Reconnaissance Automatique de la Parole, sur la traduction d'une langue parlée, et enfin sur la conversion de texte en parole. Les campagnes d'évaluation annuelles du workshop IWSLT proposent des tâches de traduction aussi bien en contexte textuel que dans un cadre Parole Parole.

1.4.2 Critères de qualité des traductions

Comment juger si une traduction est de bonne ou de mauvaise qualité ? Ce jugement se porte essentiellement sur deux critères. Le premier est le critère d'intelligibilité. Il permet de rendre compte de la bonne construction de la traduction dans la langue cible. Il indique si la traduction est une phrase qui a un sens dans la langue cible, indépendamment de la phrase source qu'elle traduit. Le deuxième critère est le critère d'adéquation ou de fidélité par rapport à la phrase source. Il mesure le degré de similitude entre la phrase source et la phrase cible du point de vue du contenu sémantique. En d'autres termes, si la traduction est lisible et qu'elle exprime bien le même contenu que la phrase source, alors elle sera jugée de bonne qualité.
Un moyen d'évaluer ces deux critères d'intelligibilité et de fidélité est de leur attribuer un score, soit en utilisant le jugement d'experts humains, soit en utilisant des mesures automatiques comme nous allons le voir dans les sections suivantes. Ces scores vont permettre de dire ensuite si oui ou non la traduction est de bonne qualité.

1.4.3 Évaluation manuelle des traductions

Initialement, l'évaluation de la qualité des traductions faisait exclusivement appel aux compétences d'experts humains. En effet, chaque évaluateur se voit confier un ensemble de paires de traduction. Chaque paire de traduction fait l'objet de différents types d'évaluation à l'issue desquelles des scores lui sont attribués. Nous donnons ci-dessous à titre d'exemple les scores produits par les évaluateurs experts lors de la campagne d'évaluation DARPA94 :

– Le A-score (*Adequacy score*)

Il indique si le sens de la phrase source est correctement retranscrit dans sa traduction. Chaque juge donne son avis sur la question et attribue à chaque traduction un score allant de 1 (intégralité du sens conservé) à 5 (rien à voir entre la phrase source et sa traduction). L'attribution du A-score ne doit pas tenir compte de la bonne construction de la traduction dans la langue cible. Toutefois, une mauvaise construction peut influencer sa valeur. Il est plus aisé d'attribuer un bon A-score si la phrase est bien construite plutôt que s'il lui manque un verbe ou si les mots ne sont pas correctement ordonnés.

– F-score (*Fluency score*)

Il s'agit d'une mesure d'intelligibilité qui rend compte du degré de bonne construction de la phrase dans la langue cible. Chaque juge doit évaluer si la traduction est correctement écrite et si elle a un sens sur une échelle de 1 (la traduction est comparable à une phrase écrite par un natif de la langue cible) à 5 (la phrase est incompréhensible).

– I-score (*Informativeness score*)

Tout comme le A-score, le I-score est une mesure de fidélité. Seul le protocole d'évaluation change. Les juges répondent cette fois à une série de questions à choix multiples sur le contenu des traductions afin de juger si les informations apportées sont équivalentes à celles de la phrase source. Ces trois scores sont assez subjectifs. Une même phrase peut être jugée de façon totalement différente par deux juges. De plus, il semble inapproprié de dissocier le critère d'intelligibilité du critère d'adéquation pour évaluer la qualité d'une traduction, l'un des critères pouvant largement influencer l'autre.

L'évaluation manuelle est une démarche lourde en terme de temps et de coûts financiers. Elle nécessite l'intervention d'experts, bilingues ou non selon qu'il s'agisse de mesurer le degré de fidélité d'une traduction par rapport à la phrase source ou plus simplement son degré de bonne construction dans la langue cible. Chaque expert doit évaluer une quantité non négligeable de traductions et une même traduction doit être évaluée par plusieurs juges afin de s'assurer de la fiabilité des résultats. Tous ces travaux demandent des heures de travail. Au vu de la quantité de données de plus en plus importantes à évaluer, il était nécessaire de mettre au point des techniques d'évaluation automatique.

1.4.4 Évaluation automatique des traductions

Depuis les années 2000, de nombreux travaux sont consacrés à la mise en place de mesures automatiques de la qualité des traductions. Dans les sections suivantes, nous détaillons les mesures les plus classiquement employées en Traduction Automatique. Même si l'évaluation humaine reste le moyen le plus fiable de juger de la qualité d'une traduction, les mesures d'évaluation automatiques permettent à moindre coût de traiter un grand nombre de données en un temps limité. Elles affectent un score à une ou plusieurs paires de traduction en comparant une hypothèse de traduction à une traduction dite de référence produite par un expert. Chaque mesure emploie ensuite des techniques de comparaison qui leur sont propres. Ainsi, grâce à ces scores attribués automatiquement aux sorties des systèmes, il est possible de rapidement comparer et classer les performances de différents systèmes de TA lors des campagnes d'évaluation. Ces mesures automatiques constituent également de bonnes fonctions objectives à maximiser lors du paramétrage d'un système dans le but d'améliorer ses performances.

1.4.4.1 WER, PER, TER

Le score WER (Word Error Rate), initialement utilisé en Reconnaissance Automatique de la Parole, compare une phrase hypothèse à une phrase référence en se fondant sur la distance de Levenshtein [Levenshtein 66]. Il est également utilisé en TA pour évaluer la qualité d'une traduction hypothèse par rapport à une traduction de référence. Pour cela, le principe est de calculer le nombre minimum d'opérations (insertion, suppression ou substitution de mot) à effectuer sur une traduction automatique pour la rendre identique à la traduction de référence. Le nombre d'opérations à effectuer, noté $d_L(ref, hyp)$ est ensuite divisé par la taille de la traduction référence, notée N_{ref} comme le montre la formule suivante :

$$WER = \frac{1}{N_{ref}} \times d_L(ref, hyp) \tag{1.1}$$

Moins il y a de modifications à effectuer, meilleur est le score. Lorsqu'une hypothèse est confrontée à plusieurs références, le score WER retenu est le score minimum d'après toutes les références. L'ordre des mots importe pour le calcul du score WER. En effet, un groupe de mots existant à la fois dans la traduction automatique et dans la traduction de référence sera pénalisé s'il ne se trouve pas à la même position dans les deux phrases. Pourtant, certains groupes de mots peuvent être à différents endroits dans une phrase sans pour autant en changer le sens. C'est le cas par exemple des compléments circonstanciels en Français. Le score PER a été proposé pour pallier ce problème.
Le score PER (Position-independent Word Error Rate) compare les mots de la traduction automatique avec ceux de la référence sans tenir compte de leur ordonnancement dans la phrase. Il a été proposé par Tillmann en 1997 [Tillmann 97b].

$$PER = \frac{1}{N_{ref}} \times d_{per}(ref, hyp) \tag{1.2}$$

d_{per} calcule la différence entre les occurrences des mots apparaissant dans la traduction automatique et dans la traduction de référence.
Le score TER (Translation Edit Rate) permet de mesurer le nombre minimum d'opérations qu'une personne doit apporter à une traduction automatique pour que celle-ci soit identique à une des références humaines correspondantes [Snover 06]. Les opérations considérées sont : l'insertion, la suppression et le remplacement d'un mot, tout comme le score WER, mais aussi le déplacement d'une suite de mots. Ce score est défini par :

$$TER = \frac{Nb(op)}{\bar{N}_{ref}} \tag{1.3}$$

où $Nb(op)$ est le nombre minimum d'opérations, calculé par programmation dynamique et $\bar{N}_{ref}$ est la taille moyenne en mots des références. Une fois de plus, cette mesure rend compte d'une certaine distance entre une traduction automatique et une référence humaine. Cependant tout comme les scores WER et PER, la mesure TER ne permet pas vraiment de juger si une traduction est acceptable ou non du fait qu'aucune reformulation de la traduction de référence n'est acceptée. D'autres variantes du score TER ont été proposées, il s'agit des mesures TERp[Snover 09] et HTER[Snover 06].

1.4.4.2 BLEU

Le score BLEU (BiLingual Evaluation Understudy), proposé par Papineni en 2001 [Papineni 01], a été la première mesure automatique acceptée comme référence pour l'évaluation des traductions. Le principe de cette méthode est de calculer le degré de similitude entre une traduction

automatique et une ou plusieurs traductions de référence en se basant notamment sur la précision n-gramme.
Le score BLEU est défini d'après la formule suivante :

$$BLEU = BP \times e^{(\sum_{n=1}^{N} w_n \ log \ p_n)} \tag{1.4}$$

Il s'agit de la moyenne géométrique des précisions n-grammes[6], p_n, obtenue avec des n-grammes d'ordre 1 jusqu'à N et des poids w_n positifs. p_n est le nombre de n-grammes de la traduction automatique présents également dans une ou plusieurs traduction de référence, divisé par le nombre de n-grammes total de la traduction automatique. Le fait de confronter la traduction automatique à plusieurs références laisse plus de liberté au système de traduction quant au choix de la traduction d'un mot, surtout lorsque celui-ci possède plusieurs synonymes. BP est une pénalité calculée pour défavoriser les traductions automatiques courtes par rapport aux références ; BLEU étant une mesure de précision, les phrases courtes seraient trop avantagées sans cette pénalité. Elle est définie par :

$$BP = \begin{cases} 1 & si \quad c > r \\ e^{1-\frac{r}{c}} & si \quad c \leq r \end{cases} \tag{1.5}$$

où c est la taille de la traduction automatique et r la taille la plus proche de c parmis la taille des traductions de référence.
Les précisions n-grammes sont généralement combinées jusqu'au 4-grammes avec des poids w_n uniformes. Une traduction automatique se voit attribuer un score BLEU de 1 lorsqu'elle est identique à une des références. Au contraire, elle aura un score de 0 si aucun de ses n-grammes n'est présent dans une référence.
La méthode BLEU a prouvé ses capacités d'évaluation des traductions automatiques. De nombreux travaux de recherche dans ce domaine l'utilisent, notamment [Koehn 04] et [Yu 04]. Dans [Yu 04], le score BLEU est employé dans un algorithme d'alignement de pages web. BLEU a également été testé lors de l'exercice d'évaluation NIST MT. Il s'est avéré que le classement des systèmes de TA donné par BLEU était le même que le classement fourni par des juges experts.

1.4.4.3 NIST et WNM

Les scores NIST et WNM se basent sur la méthode BLEU et l'étendent pour tenter de combler quelques unes de ses lacunes.
Le score NIST [Doddington 02], tout comme le score BLEU, repose sur la précision n-gramme. Cependant, il considère, non pas la moyenne géométrique des n-grammes communs à la traduction automatique et à la référence comme le fait BLEU, mais la moyenne arithmétique. Une des lacunes du score BLEU réside dans le fait que tous les n-grammes dans le calcul de la précision ont la même importance. Selon Doddington il paraît plus approprié d'accorder davantage de poids aux n-grammes importants, c'est-à-dire ceux qui sont porteurs d'information. Par conséquent, il pondère le compte des n-grammes dans le calcul de la moyenne par leur importance, comme le montre la formule suivante :

$$NIST = \sum_{n=1}^{N} BP \times \frac{\sum_{ngram \in M} Info(ngram)}{L_{hyp}} \tag{1.6}$$

n est l'ordre des n-grammes considérés. BP est, comme dans BLEU, une pénalité destinée aux traductions automatiques courtes. M est l'ensemble des n-grammes communs à la traduction

[6] Un n-gramme est une suite de n mots (cf. section 2.3 pour plus de détails)

hypothèse et à la référence. L_{hyp} est le nombre de n-grammes dans la traduction évaluée. Enfin, $Info(ngram)$ est une fonction dépendante du nombre d'occurrences du n-gramme passé en paramètre dans les traductions références. Le score NIST considère les n-grammes fréquents moins importants que les n-grammes qui apparaissent peu de fois.
Babych et Hartley ont également proposé dans [Babych 04] une extension de la méthode BLEU appelée WNM pour Weighted N-graM en partant du même constat qu'une équivalence entre mots clés qui apportent beaucoup d'information pour le sens de la phrase est beaucoup plus significative qu'une équivalence entre mots outils de la langue qui ne sont nécessaires que pour la bonne structure de la phrase. Ils proposent donc de pondérer les n-grammes à l'aide de deux scores : la mesure standard tf.idf [Salton 68] et la mesure S-Score [Babych 03]. Ces deux mesures permettent d'accorder plus de poids aux mots ou suite de mots porteurs de sens qu'aux mots outils de la langue. Elles rendent compte de la prépondérance d'un mot dans un corpus. Ces deux scores sont calculés pour chaque n-gramme du corpus de référence. Ils sont ensuite intégrés dans le calcul de la précision, du rappel et de la F-mesure en pondérant chaque équivalence de n-grammes par les scores qui lui sont associés. Rappelons que la précision est le nombre d'équivalences de n-grammes divisé par le nombre de n-grammes de la traduction candidate et que le rappel est le nombre d'équivalences de n-grammes divisé par le nombre de n-grammes de la référence. La F-mesure est une moyenne harmonique de la précision et du rappel. Le score WNM prend donc en compte la précision et le rappel des n-grammes, contrairement au score BLEU qui ne considère pas directement le rappel. En effet, pour le score BLEU, c'est seulement par l'intermédiaire de la pénalité BP que le score tient compte d'un problème éventuel de recouvrement en défavorisant les traductions automatiques trop courtes par rapport aux références.

1.4.4.4 METEOR

METEOR (Metric for Evaluation of Translation with Explicit Ordering) est une mesure proposée par Banerjee et Lavie en 2005 [Banerjee 05] et basée sur la correspondance entre les unigrammes d'une traduction automatique et ceux d'une traduction de référence. La mise en correspondance des mots de la traduction automatique avec les mots de la traduction de référence est appelé un alignement de mots. Le calcul de du score METEOR se déroule en deux étapes : trouver le meilleur alignement de mots entre la traduction candidate et la traduction référence et déterminer un score à partir de cet alignement.

L'alignement se fait en plusieurs itérations. Il lie dans un premier temps les unigrammes de même forme orthographique, puis les mots de même racine et enfin les synonymes présents parmi les mots qui ne sont pas encore alignés. Une fois établi, l'alignement permet de calculer le score METEOR comme suit :

$$METEOR = (1 - Pen)\frac{P \times R}{\alpha P + (1-\alpha)R} \tag{1.7}$$

Il s'agit d'une moyenne harmonique pondérée de la précision P, qui est le nombre de correspondances entre unigrammes établies par l'alignement divisé par la taille de la traduction candidate et du rappel R, qui est le nombre de correspondances entre unigrammes divisé par la taille de la traduction référence. α est un facteur de pondération visant à accorder plus ou moins d'influence au rappel ou à la précision dans le calcul de la moyenne. Pen, enfin, est un coefficient de pénalité destiné à réduire le score des traductions n'ayant pas de correspondances d'ordre plus élevé que l'unigramme. Pen est défini par :

$$Pen = \gamma \times (\frac{Nb(chunks)}{Nb(unigram.\ align.)})^{\beta} \tag{1.8}$$

$Nb(unigram.\ align.)$ est le nombre d'unigrammes de la traduction automatique mis en correspondance avec un mot de la traduction de référence. Un *chunk* est un sous-ensemble de mots contigus de la traduction automatique alignés avec un sous-ensemble de mots contigus de la traduction de référence. Dans le cas extrême où seuls des unigrammes sont alignés le facteur de pénalité vaut γ. Dans le cas contraire, plus les n-grammes alignés sont d'ordre important, moins il y a de *chunks* dans la traduction candidate et plus le facteur de pénalité est faible. α, β et γ sont des paramètres du score METEOR qui nécessitent une phase d'optimisation.
Ici encore, le rappel n-gramme est directement pris en compte dans le calcul du score contrairement au score BLEU. De plus, les correspondances sont établies aussi bien sur les formes orthographiques, que sur les synonymes ou les mots de même racine. Rappelons que BLEU se base exclusivement sur la forme orthographique des mots.

Nous venons de présenter une série non exhaustive de scores permettant d'évaluer automatiquement les traductions issues de systèmes de TA (et *a fortiori* les systèmes eux-mêmes) en les comparant à une ou plusieurs traductions de référence. Plus elles sont semblables, plus la qualité de la traduction automatique est qualifiée de correcte. Cette comparaison est basée pour toutes ces mesures sur les équivalences qu'il peut y avoir entre les n-grammes de la traduction candidate et les n-grammes des traductions de référence. Les équivalences sont définies différemment selon chaque méthode et elles ont un poids plus ou moins important dans la traduction. Le problème est maintenant d'évaluer ces mesures pour trouver laquelle est la plus efficace et la plus proche des jugements d'experts humains.

1.4.5 Qualité des mesures d'évaluation automatique

Dans l'idéal, une mesure automatique de la qualité des traductions automatiques devrait pouvoir prédire le jugement humain et donc se substituer à lui. Pour constater ce qu'il en est en réalité, un coefficient de corrélation est calculé entre les résultats assignés par une mesure automatique et les jugements émis par les experts. Plus la corrélation est forte, plus la mesure est jugée apte à prédire l'avis des experts.
Les deux coefficients de corrélation les plus utilisés dans la littérature sont les coefficients de Spearman et de Pearson. Ils permettent de mesurer la dépendance pouvant exister entre les jugements humains et les mesures automatiques. Le coefficient de Spearman sert notamment à comparer l'adéquation entre les classements obtenus manuellement et ceux obtenus automatiquement. Lors du workshop WMT (Workshop on Statistical Machine Translation) qui s'est déroulé en 2008 par exemple, il a été utilisé pour mesurer la corrélation au niveau du classement des systèmes de TA en compétition [Callison-Burch 08]. Le coefficient de Pearson est une autre mesure de corrélation, notamment employé lors des campagnes d'évaluations DARPA/TIDES. Il permet de mesurer le degré de dépendance linéaire entre deux variables. Ainsi, de nombreux travaux de recherche font appel à au coefficient de Pearson pour estimer à quel point les notes d'adéquation (A-Score) et de bonne construction (F-Score) des traductions automatiques (cf. section 1.4.3) évaluées par des experts sont liées aux valeurs des mesures automatiques [Gimenez 08, A. Lavie 04, Banerjee 05]. Cette comparaison porte aussi bien au niveau de la phrase qu'au niveau d'un système.

1.4.6 Quelle mesure automatique choisir ?

Grâce aux coefficients de corrélation, il devrait être possible d'établir quel score automatique corrèle le mieux avec l'avis des experts humains et ainsi choisir celui qui se substituera à l'éva-

luation humaine très coûteuse en temps et en moyens. Cependant, les études menées sur les différentes mesures ont des avis très divergents et par conséquent il reste difficile de privilégier une mesure par rapport à une autre.

Papineni *et al.* [Papineni 01], par exemple, ont montré une forte corrélation entre le score BLEU et les jugements d'experts humains dans une tâche consistant à classer cinq systèmes de TA Chinois-Anglais selon la qualité de leurs traductions. Turian et al [Turain 03] affirment le contraire après avoir mené une étude indiquant que la F-Mesure basée sur les correspondances unigrammes entre une traduction automatique et une référence corrèle plus fortement avec le jugement humain que BLEU ou NIST. Banerjee *et al.* [Banerjee 05] ont mené une série d'expériences visant, entre autre, à comparer leur méthode avec BLEU, mais aussi à comparer l'apport de chacun des constituants du score METEOR à savoir la précision et le rappel. Les résultats ont montré dans un premier temps que METEOR corrèle plus fortement avec le jugement humain que BLEU au niveau de l'évaluation d'un corpus de test. Au niveau de l'évaluation traduction par traduction, les tests ont montré que le rappel a une plus forte corrélation avec le jugement humain que la précision et que la combinaison des deux par la moyenne harmonique améliore encore un peu ce taux de corrélation.

Au vu de toutes les études menées, il semblerait que le choix de la mesure automatique d'évaluation des traductions soit fortement lié au contexte de test (corpus, paire de langues). Turian, en 2003, avait déjà écrit qu'une mesure automatique donnant de bons résultats sur un corpus, ne fonctionnait pas nécessairement aussi bien sur d'autres [Turain 03]. Par exemple, les performances du score METEOR en terme de corrélation avec le jugement humain dépendent fortement de l'optimisation de ses paramètres α, β, γ (cf. formule 1.7) [Agarwal 08]. Gimenez *et al.* [Giménez 06] ont proposé un outil appelé IQMT permettant de combiner plusieurs scores automatiques. De ce fait, il est possible d'optimiser les systèmes de traduction en tenant compte non pas d'un seul score mais d'une combinaison de scores. Ceci permet entre autres de tirer profit de chacune des mesures et ainsi améliorer la qualité des traductions sur les différents aspects linguistiques traités par chacune des mesures.

Cependant, plusieurs travaux en arrivent à la même conclusion que le jugement humain n'est lui-même pas fiable. En effet, pour une même traduction les experts humains ne sont pas toujours d'accord et pire encore, un même juge peut donner deux avis contraires pour une même traduction [Turain 03, Callison-Burch 08]. Il devient alors difficile de mettre en place des mesures automatiques capables de prédire le jugement humain si l'avis des experts humains n'est ni consistant ni fiable dans certains cas.

Les mesures automatiques ne sont pas en mesure de remplacer le jugement humain pour évaluer la qualité d'une traduction. En effet, la langue est tellement riche et complexe que les scores automatiques sont incapables d'en discerner les nombreuses subtilités et en particulier en traduction. Confronter une traduction automatique à une ou plusieurs références limite considérablement le nombre important de façons différentes d'exprimer une idée sans en modifier le sens. Toutefois, ces mesures restent un moyen rapide et efficace d'améliorer les performances d'un système de traduction et de les comparer aux performances d'autres systèmes. Bien qu'une multitude de scores existent aujourd'hui, le plus populaire et le plus utilisé dans les campagnes d'évaluation reste le score BLEU. Il fait état de référence dans la communauté de la Traduction Automatique. Récemment encore, les exercices d'évaluation du Workshop WMT 2008 ont conclu que le score METEOR corrèle le plus avec le jugement humain lorsqu'il s'agit de systèmes tra-

duisant d'une langue source vers l'Anglais, mais que c'est le score BLEU qui corrèle le mieux lorsque la traduction se fait de l'Anglais vers une autre langue [Callison-Burch 08], ce qui s'avère être le cas pour nous. En effet, dans les travaux que nous allons présenter par la suite, nous nous plaçons dans un contexte de traduction de l'Anglais vers le Français. Nous avons donc retenu le score BLEU comme mesure d'évaluation de la qualité des traductions produites par notre système pour sa corrélation avec le jugement humain et de manière à pouvoir se comparer aux autres travaux proposés dans la littérature.

Chapitre 2

L'approche statistique de la traduction automatique

2.1 Introduction

Les contributions pour la Traduction Automatique présentées dans ce manuscrit se concentrent sur l'approche statistique. Dans ce chapitre, nous décrivons cette approche en présentant les grands travaux dans le domaine. Nous verrons notamment les différents modules impliqués dans un système de traduction statistique. Nous commençons par décrire les différents modèles de langage puis les modèles de traduction prenant comme unité de traduction le mot puis les séquences de mots. Nous présentons ensuite le module de décodage qui permet à partir des modèles de produire une phrase dans une langue étant donnée une phrase d'entrée dans une autre langue. Nous terminons enfin ce chapitre en exposant les points forts et les inconvénients des différents modèles évoqués avant d'introduire nos contributions.

2.2 Généralités

Contrairement aux approches expertes qui se concentrent essentiellement sur les différentes étapes impliquées dans le processus de traduction (analyse, transfert, génération), l'approche statistique considère le problème du point de vue du résultat. Ainsi, traduire d'une langue source vers une langue cible consiste à trouver la phrase cible T qui est la traduction la plus probable de la phrase source S. Pour cela, T doit répondre à deux critères qui feront d'elle la meilleure traduction de S. Elle doit être à la fois fidèle au contenu de S et à la fois vraisemblable dans la langue cible.
Ce compromis a, dans un premier temps, été modélisé en se basant sur le principe du canal bruité déjà utilisé en reconnaissance automatique de la Parole. L'utilisation de la formule de Bayes, fait alors intervenir un modèle de langage de la langue cible et donne lieu à l'équation 2.1.

$$\begin{aligned} T^* &= \underset{T}{\operatorname{argmax}}\, P(T|S) \\ &= \underset{T}{\operatorname{argmax}}\, \frac{P(S|T) \times P(T)}{P(S)} \end{aligned} \tag{2.1}$$

Le terme $P(S)$ ne dépend pas de T, il n'a donc aucune influence sur le calcul de la fonction $argmax$. Il peut, par conséquent être écarté pour finalement aboutir à l'équation fondamentale

de la traduction statistique :

$$T^* = \underset{T}{\operatorname{argmax}}\, P(T) \times P(S|T) \tag{2.2}$$

où $P(T)$ est estimé par un modèle de langage dont les paramètres sont appris sur un corpus

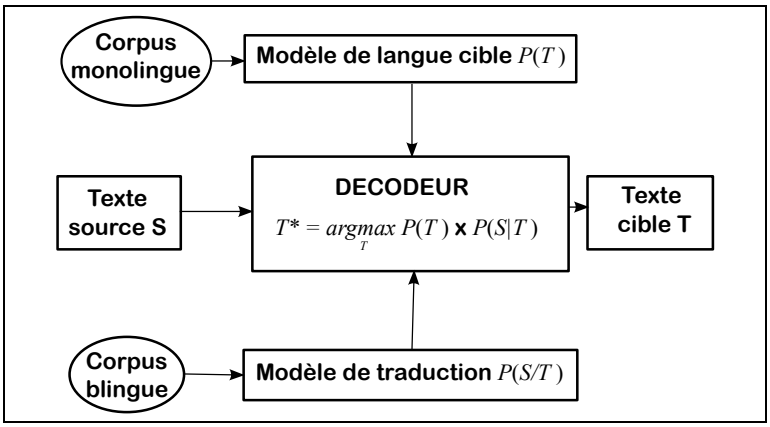

FIG. 2.1 – Modélisation de l'approche statistique suivant le modèle canal bruité [Och 02]

en langue cible et $P(S|T)$ par un modèle de traduction dont les paramètres sont appris sur un corpus d'apprentissage parallèle aligné.

Avec cette première formulation, les deux seules caractéristiques entrant en compte dans la recherche de la meilleure phrase cible T sachant la phrase source S sont sa vraisemblance dans la langue cible et son adéquation avec S. Elles sont simplement multipliées. Cependant, comme dans toutes combinaisons de modèles, il est souvent recommandé de pondérer les modèles afin d'aboutir à de meilleurs résultats. Ainsi, le calcul de la fonction *argmax* s'opère, en réalité, en affectant des poids différents aux modèles comme le suggère l'équation 2.3.

$$T^* = \underset{T}{\operatorname{argmax}}\, P(T)^{\alpha} \times P(S|T)^{\beta} \tag{2.3}$$

De plus, il est également possible de considérer d'autres caractéristiques pouvant être bénéfiques pour produire une traduction correcte de la phrase source S, comme par exemple un modèle de langage de la phrase source. Par conséquent, l'approche statistique a, par la suite, été formalisée de façon plus flexible à l'aide d'une combinaison *log-linéaire* de fonctions caractéristiques :

$$\begin{aligned} T^* &= \underset{T}{\operatorname{argmax}} \prod_i h_i(S,T)^{\lambda_i} \\ &= \underset{T}{\operatorname{argmax}}\, e^{(\sum_i \lambda_i \ log \ h_i(S,T))} \end{aligned} \tag{2.4}$$

où h_i est une fonction caractéristique et λ_i son poids dans la combinaison [Och 02]. Cette formulation permet d'ajouter n'importe quelle fonction caractéristique comme l'illustre la figure 2.2. Toutefois, les études ont montré que les principales fonctions restent le modèle de langage de la langue cible et le modèle de traduction initialement impliqués dans la formulation reprenant

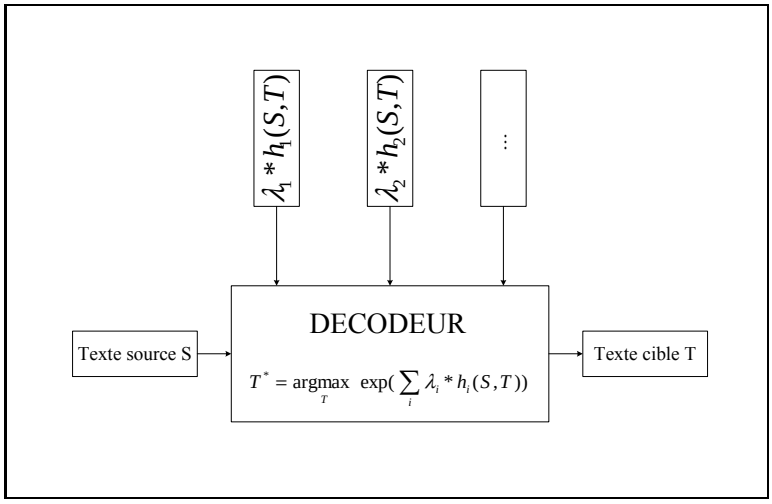

FIG. 2.2 – Modélisation de l'approche statistique suivant le modèle log-linéaire [Och 02]

le modèle du canal bruité (cf. équation 2.3) qui est un cas particulier de l'équation 2.4 avec $h_1 = P(T)$, $\lambda_1 = \alpha$, $h_2 = P(S|T)$ et $\lambda_2 = \beta$.

L'approche statistique nécessite donc trois composants principaux comme le montre la figure 2.1 : un modèle de langage $P(T)$, un modèle de traduction $P(S|T)$ et un décodeur qui va permettre le calcul de la fonction $argmax$. Chacun de ces composants est décrit dans les sections suivantes.

2.3 Modélisation du langage

Comme nous venons de le voir, le processus de traduction du point de vue de l'approche statistique est modélisé par une combinaison log-linéaire de fonctions caractéristiques, une des principales étant le modèle de langage de la langue cible. Cette caractéristique permet de rendre compte de la bonne construction dans la langue cible de la phrase produite. Nous définissons, dans ce qui suit ce qu'est un modèle de langage et présentons les principaux modèles utilisés dans la littérature.

2.3.1 Définition d'un modèle de langage

Un modèle de langage statistique permet d'estimer la vraisemblance d'une suite de mots en lui attribuant une probabilité. Soit $W = w_1 w_2 \dots w_L$ une séquence de L mots dans une langue donnée couvrant un vocabulaire fixé V. Les mots n'appartenant pas au vocabulaire V sont considérés comme mots inconnus et sont tous ramenés au même mot particulier appelé UNK. Alors, la probabilité de W est définie de la façon suivante :

$$P(w_1 w_2 \dots w_L) = \prod_{i=1}^{L} P(w_i | w_1 \dots w_{i-1}) \tag{2.5}$$

Pour chaque mot w_i de la séquence W, il est donc nécessaire de connaître sa probabilité d'apparition sachant son historique $h = w_1 \dots w_{i-1}$. Ces probabilités sont estimées par le modèle

de langage. Elles dépendent de la fréquence d'apparition de la suite hw_i au sein d'un corpus d'apprentissage représentatif de la langue comme l'indique l'équation 2.6.

$$P(w_i|h) = \frac{N(hw_i)}{N(h)} \tag{2.6}$$

La fonction $N(x)$ renvoie le nombre d'occurrences de la suite x dans le corpus d'apprentissage. Dans ce qui suit, nous décrivons plusieurs modèles de langage permettant d'estimer de façons différentes ces probabilités.

2.3.2 Quelques modèles de langage

2.3.2.1 Les modèles n-grammes

a) Description Un des problèmes majeurs dans l'utilisation des modèles probabilistes tient en la longueur de l'historique considéré. Plus la suite à prédire est grande, plus la taille des historiques à considérer est importante. Or, il est généralement impossible de rencontrer dans un corpus d'apprentissage tous les historiques possibles pour tous les mots. Par conséquent, il devient impossible d'estimer toutes les probabilités $P(w_i|h)$. Pour palier ce problème, les modèles de type *n-grammes* ramènent l'historique d'un mot w_i aux $n-1$ mots le précédant. Ainsi, la probabilité d'un mot w_i sachant les mots que le précèdent devient :

$$P(w_i|w_1 w_2 \dots w_{i-1}) \approx P(w_i|w_{i-n+1} \dots w_{i-1}) \tag{2.7}$$

Ainsi, le modèle unigramme revient à la probabilité du mot w_i lui-même, le modèle bigramme considère un historique d'un mot, le modèle trigramme prend en compte un historique de deux mots etc. Bien que réduisant l'historique des mots aux $n-1$ mots précédents, le problème du manque de données reste toujours présent dans les modèles n-grammes. En effet, un n-gramme se verra attribuer une probabilité nulle par le modèle de langage s'il n'apparaît pas dans le corpus d'apprentissage, même si ce n-gramme est une suite de mots possible dans la langue considérée. Pour résoudre ce problème, des méthodes de lissage ont été mises en place pour redistribuer la masse des probabilités des événements ayant une forte probabilité vers ceux ayant une faible probabilité.

b) Les méthodes de lissage Afin d'éviter les sous-estimations d'événements ou les estimations impossibles, il existe des techniques de lissage permettant d'ajuster les probabilités pour réduire l'effet du manque de données. Nous pouvons, par exemple citer la méthode de lissage de Kneser-Ney [Kneser 95], la méthode de *backoff* de Katz [Katz 87] ou bien encore la méthode d'interpolation linéaire de Jelinek et Mercer [Jelinek 80]. Ces méthodes peuvent être généralisées de la façon suivante :

$$P(w_i|w_1 \dots w_{i-k+1} \dots w_{i-1}) = \begin{cases} P(w_i|w_{i-k+1} \dots w_{i-1}) & si\ hw_i\ existe \\ \lambda(w_{i-k+1} \dots w_{i-1}) \times P(w_i|w_{i-k+2} \dots w_{i-1}) & sinon \end{cases} \tag{2.8}$$

$h = w_{i-k+1} \dots w_{i-1}$ est l'historique du mot w_i et $\lambda(w_{i-n+1} \dots w_{i-1})$ est un poids de repli appelé *backoff*. L'idée globale de ces méthodes est de se replier sur le n-gramme d'ordre $k-1$ si le n-gramme d'ordre k n'apparaît pas dans le corpus d'apprentissage.

Les modèles de type n-grammes sont dits *locaux*, en ce sens qu'ils ne considèrent que les mots très proches du mot à prédire. De ce fait, ils ne tiennent compte que des dépendances entre les

mots au sein d'une suite d'au maximum 5 ou 6 mots et ne considèrent aucunement l'influence que pourrait avoir le sujet général du corpus d'apprentissage. Les modèles de type n-classes sont également des modèles dits *locaux*. Ils prédisent l'apparition d'un mot en fonction de sa classe (syntaxique, lemmatique, sémantique ...) et des classes des $n-1$ mots précédents sans tenir compte du contexte plus général du corpus. Pour combler ces lacunes, les modèles n-grammes sont souvent combinés à d'autres modèles, dits *dynamiques*, prenant en compte des dépendances longues distances. Dans ce qui suit, nous décrivons seulement les modèles Cache et Triggers dont nous nous sommes inspirés pour nos travaux en Traduction Automatique.

2.3.2.2 Le modèle Cache

Le modèle Cache a été proposé par Kuhn et Demori en 1990 [Kuhn 90]. Il permet de prendre en compte des dépendances entre mots non contigus. En effet, il est fréquent de remarquer qu'un mot qui est déjà apparu dans un corpus a une probabilité plus importante d'être rencontré de nouveau qu'un mot qui n'a pas encore été cité. Par exemple, dans un document qui traite de politique, si le mot *Président* apparaît, il y a de fortes chances qu'il soit employé encore par la suite. Sur ce constat, le modèle Cache permet d'augmenter la probabilité d'un mot w_i récemment apparu dans le contexte gauche, appelé ici Cache, comme l'indique l'équation 2.9.

$$P_{cache}(w_i|w_{i-K}, w_{i-K+1}, \ldots, w_{i-1}) = \frac{N(w_i)}{K} \tag{2.9}$$

$N(w_i)$ est l'occurrence du mot w_i dans le cache de taille K. Le Cache ne se limite pas aux 5 ou 6 mots du contexte gauche. Il peut être de l'ordre de la centaine voire du millier de mots.

Une des critiques importantes qu'il est possible d'apporter sur le modèle Cache est qu'il ne modélise que les relations entre un mot et lui-même. Pourtant, d'après notre exemple, si le mot *Président* apparaît, d'autres mots qui lui sont liés devraient également voir leur probabilité renforcée comme par exemple *République* ou encore *Sarkozy*. Le modèle Trigger a été proposé pour permettre de prendre en compte cette critique.

2.3.2.3 Le modèle Trigger

Le modèle Trigger a initialement été proposé par Lau *et al.* en 1993 [Lau 93]. Le concept a ensuite été repris dans plusieurs travaux [Rosenfeld 96, Tillmann 97a, Black 98]. L'idée est de déterminer les couples de mots (x,y) tel que la présence de x dans l'historique déclenche l'apparition de y, d'où l'utilisation du terme anglais *to trigger* signifiant *déclencher*. Ce modèle généralise le modèle Cache. La présence d'un mot dans le contexte gauche va renforcer la probabilité d'apparition de ce même mot mais aussi d'un ensemble d'autres mots qui lui sont fortement corrélés. Par la suite, nous appellerons un trigger un couple de mots composé d'un mot déclencheur et d'un mot déclenché. Ces couples permettront ensuite de renforcer les probabilités des modèles locaux de type n-grammes et ainsi prendre les dépendances longue distance en considération.
Les triggers sont déterminés en utilisant l'Information Mutuelle (IM). Celle-ci permet, en effet, de quantifier le lien de dépendance existant entre deux variables aléatoires $\mathcal{X}$ et $\mathcal{Y}$. Elle est définie par :

$$IM(\mathcal{X}, \mathcal{Y}) = \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} P(\mathcal{X} = x, \mathcal{Y} = y) log \frac{P(\mathcal{X} = x, \mathcal{Y} = y)}{P(\mathcal{X} = x)P(\mathcal{Y} = y)} \tag{2.10}$$

Lorsque ces deux variables aléatoires ont pour domaine les mots du vocabulaire d'une langue, l'IM permet d'établir la corrélation existant entre deux mots particuliers x et y :

$$IM(x,y) = P(x,y)log\frac{P(x,y)}{P(x)P(y)} \tag{2.11}$$

où $P(x,y)$ est la probabilité jointe d'apparition des mots x et y définie par :

$$P(x,y) = \frac{N(x,y)}{N} \tag{2.12}$$

où N est la taille du corpus d'apprentissage et $N(x,y)$ le nombre de fois où x et y apparaissent ensemble dans une fenêtre de mots de taille déterminée et $P(x)$ (respectivement $P(y)$ est la probabilité d'apparition de x (respectivement y) définie par :

$$P(x) = \frac{N(x)}{N} \tag{2.13}$$

avec $N(x)$ le nombre d'd'occurrences de x dans le corpus d'apprentissage.
Plus l'IM entre x et y est élevée, plus le lien entre ces mots est fort. Seuls les couples ayant une IM supérieure à un seuil fixé sont sélectionnés comme étant des triggers. Plusieurs travaux [Lau 93, Rosenfeld 96] utilisent une variante de l'IM pour sélectionner les triggers. Ils font, en effet, intervenir les événements $\bar{\mathcal{X}}$ et $\bar{\mathcal{Y}}$ dans le calcul de l'IM et la formule utilisée devient :

$$\begin{aligned} IM(x,y) &= P(x,y)log\frac{P(x,y)}{P(x)P(y)} + P(x,\bar{y})log\frac{P(x,\bar{y})}{P(x)P(\bar{y})} \\ &= P(\bar{x},y)log\frac{P(\bar{x},y)}{P(\bar{x})P(y)} + P(\bar{x},\bar{y})log\frac{P(\bar{x},\bar{y})}{P(\bar{x})P(\bar{y})} \end{aligned} \tag{2.14}$$

Lau *et al.* [Lau 93] ont montré que les triggers sont efficaces en combinaison avec des modèles dits locaux comme les modèles de type n-grammes. L'information supplémentaire qu'ils apportent va permettre de renforcer les probabilités de certains mots suivant leurs contextes qu'ils soient immédiat ou distant. Nous verrons dans le chapitre 4 des exemples de triggers sélectionnés sur des corpus Anglais et Français.

2.3.3 Utilisation des modèles de langage en TA

Comme nous l'avons dit dans la section 2.2, le modèle de langage constitue l'une des principales fonctions caractéristiques impliquées dans le processus de traduction automatique statistique. Il va permettre de déterminer la vraisemblance des hypothèses de traduction produites dans la langue cible. Il intervient alors directement dans le calcul effectué par le décodeur au même titre que le modèle de traduction. Chaque modèle va attribuer un score à chaque hypothèse. Ces scores sont ensuite combinés par le décodeur afin de maximiser une fonction objectif et ainsi trouver la meilleure hypothèse parmi toutes celles possibles. Nous présentons plus en détails un algorithme de recherche de la meilleure hypothèse de traduction dans la section 2.5.

Classiquement, la plupart des systèmes de traduction statistique utilise un modèle de langage de type n-grammes d'ordre pouvant aller jusqu'à 5. Comme tout système statistique, les performances d'un système de traduction peuvent être améliorées en augmentant la taille de ses modèles et particulièrement la taille du modèle de langage. Pour ce faire, il est possible soit d'ajouter encore plus de données dans le corpus d'apprentissage, soit de considérer des n-grammes

d'ordre supérieur. Cependant, plus la taille du modèle de langage s'accroît, plus il est difficile de le stocker en mémoire et de s'en servir de manière efficace dans le processus de traduction. Plusieurs travaux proposent d'intégrer un modèle de langage distribué et grande échelle. Brants *et al.* par exemple [Brants 07] entraînent un modèle de langage sur un corpus de plus de 2×10^3 milliards de mots et l'intègre directement dans le décodeur grâce à une architecture distribuée. Zangh *et al.* [Zhang 06] utilisent également un modèle de langage distribué appris sur une quantité très importante de données. Toutefois à la différence des travaux de Brants *et al.*, le modèle de langage est utilisé en deuxième passe du décodeur pour attribuer un nouveau score à la liste des n meilleures hypothèses proposées par le décodeur.

Les modèles de langage utilisés dans le cadre de la Traduction Automatique Statistique sont des modèles de langage de type n-grammes et donc des modèles dits locaux. Ils ne permettent pas de prendre en compte les dépendances entre les mots très éloignés. Pourtant ces dépendances sont réelles. C'est plus particulièrement le cas pour les accords en genre et en nombre des mots comme le montre l'exemple suivant :

> *la petite pomme verte de l'arbre du jardin que j'ai mang**ée** hier était délicieuse*
> *the small green apple of the garden tree I ate yesterday was delicious*

Imaginons un système de traduction de l'Anglais vers le Français. Le verbe *ate* tout comme l'adjectif *delicious* ne portent pas la marque du genre masculin ou féminin. Dans le processus de traduction, c'est le rôle du modèle de langage d'accorder correctement ces mots en Français. Pourtant, dans cet exemple, l'information nécessaire pour que cet accord soit pris en compte est présent dans le mot *pomme* qui est bien trop éloigné de *ate* ou *delicious* pour pouvoir être pris en compte dans l'historique d'un modèle 5-grammes. Il est fréquent de rencontrer dans les sorties de système de traduction des fautes liées aux accords de ce type. Dans les annexes de ce manuscrit, nous proposons un modèle de langage inspiré du modèle longue distance Cache et proposant une première solution à ce problème.

2.4 Modèles de traduction

Les modèles de traduction décrivent le processus de traduction par le biais notamment d'une table de traduction qui calcule les probabilités de traduction d'une unité source en une unité cible et d'un modèle d'alignement qui, lui, calcule les probabilités de déplacement des unités cibles par rapport aux unités qu'elles traduisent dans la phrase source.
Initialement, les modèles de traduction prenaient comme unité le mot mais aujourd'hui ils se basent plutôt sur les groupes de mots appelés séquences ou encore segments. L'apprentissage de la plupart des modèles de traduction de mots repose sur la succession de 5 modèles probabilistes itératifs connus comme étant les modèles IBM [Brown 93], auxquels s'ajoute également le modèle HMM [Vogel 96]. Avant de présenter ces modèles, nous commençons par définir la notion d'alignement sur laquelle ils reposent. Nous verrons enfin différentes méthodes d'apprentissage de modèles de traduction de séquences à partir d'heuristiques apportées sur les modèles IBM, ou par le biais d'algorithmes génératifs.

2.4.1 La notion d'alignement

La plupart des modèles de traduction statistiques font intervenir une variable cachée A appelée alignement. Au sein d'une paire de traduction, l'alignement indique les correspondances

existantes entre les mots de la phrase source et ceux de la phrase cible. Il définit donc pour chaque mot source les mots cibles qu'il traduit. La figure 2.3 montre un alignement entre une phrase française et sa traduction en Anglais. Les correspondances sont symbolisées par des liens. Un mot

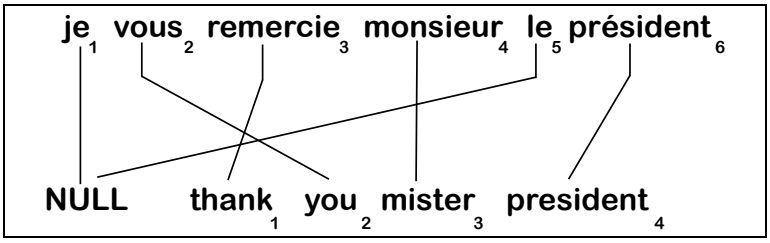

FIG. 2.3 – Un exemple d'alignement avec les mots français

spécial appelé *NULL* est introduit lorsqu'un ou plusieurs mots n'ont pas de correspondance dans la traduction. Toutes les correspondances entre mots sont *a priori* possibles. Cependant, nous verrons par la suite que les modèles de traduction statistiques apportent une forte contrainte sur la distribution des alignements.

2.4.2 les modèles à base de mots

2.4.2.1 Les modèles IBM et le modèle HMM

Dans cette section, par soucis de clarté nous reprenons les notations de Brown et al [Brown 93]. Ainsi, nous cherchons la phrase notée $\mathbf{e} = e_1 \dots e_I$ qui est la traduction la plus probable de la phrase notée $\mathbf{f} = f_1 \dots f_J$. Partant de l'équation fondamentale

$$\mathbf{e}^* = argmax_{\mathbf{e}} P(\mathbf{e}) \times P(\mathbf{f}|\mathbf{e}) \tag{2.15}$$

Brown *et al.* ont proposé une série de 5 méthodes statistiques modélisant le processus de traduction. Chacune de ces méthodes permet de calculer la probabilité $P(\mathbf{f}|\mathbf{e})$ à l'aide de différents paramètres estimés sur un corpus parallèle composé de paires de phrases étant traductions l'une de l'autre. La performance et la complexité des modèles augmentent du premier au cinquième.

Brown *et al.* modélisent l'alignement entre deux phrases comme une variable cachée du processus de traduction et attribuent donc une probabilité à chacun des alignements possibles au sein d'une paire de traduction. Toutefois, ils ne considèrent que les alignements pour lesquels chaque mot d'une phrase $\mathbf{f}$ est connecté à exactement un mot d'une phrase $\mathbf{e}$. De ce fait, étant données une phrase $\mathbf{f}$ contenant J mots et une phrase $\mathbf{e}$ contenant I mots, un alignement entre $\mathbf{f}$ et $\mathbf{e}$ est une suite de J valeurs notée $a_1^J = a_1 a_2 \dots a_J$, chaque valeur étant comprise entre 0 et I. L'alignement indique pour chaque position dans $\mathbf{f}$, la position dans $\mathbf{e}$ qui lui correspond. Ainsi, si le mot à la position j dans $\mathbf{f}$ correspond au mot à la position i dans $\mathbf{e}$ alors $a_j = i$. Suivant cette définition, plusieurs mots f_j peuvent être liés au même mot e_i mais à l'inverse, un même mot f_j ne sera lié qu'à un seul mot e_i. Chaque f_j est donc connecté à 0 ou 1 mot de la phrase $\mathbf{e}$. De ce fait, les modèles d'IBM sont qualifiés d'asymétriques.
Considérant l'ensemble des alignements possibles, la vraisemblance d'une traduction ($\mathbf{f}$,$\mathbf{e}$) est définie par :

$$P(\mathbf{f}|\mathbf{e}) = \sum_a P(\mathbf{f}, a|\mathbf{e}) \tag{2.16}$$

où a est un alignement possible entre $\mathbf{f}$ et $\mathbf{e}$. La somme s'effectue sur tous les alignements possibles.

Les cinq modèles IBM procèdent à leur manière et de façon itérative pour estimer la probabilité $P(\mathbf{f}|\mathbf{e})$. Chaque modèle s'appuie sur les paramètres estimés par le modèle le précédant. Dans la suite de cette section, nous décrivons les modèles IBM de complexité croissante, ainsi que le modèle HMM proposé en amélioration du modèle IBM 2, en indiquant l'apport de chacun par rapport à son prédecesseur.

Pour les modèles IBM 1 et 2, la génération d'une phrase $\mathbf{f}$ sachant une phrase $\mathbf{e}$ s'opère en plusieurs étapes : déterminer la taille J de $\mathbf{f}$ d'après la phrase $\mathbf{e}$ à traduire, ensuite, pour chaque position de $j = 1$ à $j = J$ dans $\mathbf{f}$, sélectionner dans $\mathbf{e}$ le mot e_i qui lui sera connecté, enfin produire le mot cible f_j en fonction de e_i. Ces trois étapes sont illustrées par la figure 2.4.

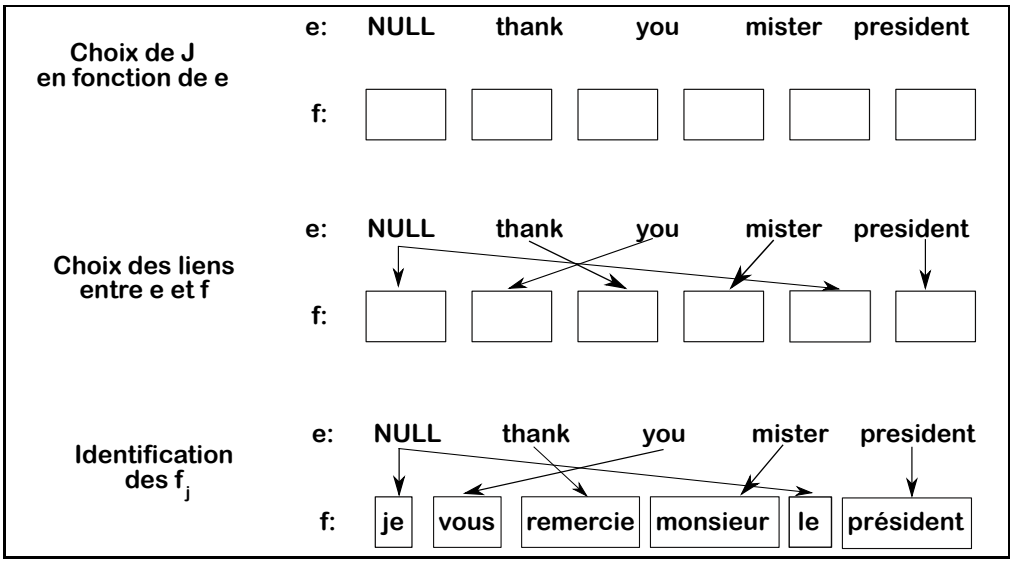

FIG. 2.4 – Processus de traduction des modèles 1 et 2 d'IBM [Jur 06]

Suivant ce processus en trois étapes, les modèles IBM 1 et 2 définissent la probabilité de générer $\mathbf{f}$ avec un alignement a sachant la phrase $\mathbf{e}$ comme un produit de probabilités conditionnelles :

$$P(\mathbf{f}, a|\mathbf{e}) = P(J|\mathbf{e}) \prod_{j=1}^{J} P(a_j|a_1^{j-1}, f_1^{j-1}, J, \mathbf{e}) P(f_j|a_1^j, f_1^{j-1}, J, \mathbf{e}) \tag{2.17}$$

Chaque terme du produit modélise une étape du processus de traduction de $\mathbf{e}$ en $\mathbf{f}$. $P(J|\mathbf{e})$ modélise le choix de la taille J de $\mathbf{f}$ en fonction de $\mathbf{e}$. $P(a_j|a_1^{j-1}, f_1^{j-1}, J, \mathbf{e})$ modélise l'alignement de la position j dans $\mathbf{f}$ avec une position dans $\mathbf{e}$. Cet alignement dépend de $\mathbf{e}$, de J et des mots déjà alignés. Enfin, $P(f_j|a_1^j, f_1^{j-1}, J, \mathbf{e})$ modélise le choix de mot f_j à la position j dans $\mathbf{f}$ en fonction $\mathbf{e}$ et des mots déjà alignés et identifiés de $\mathbf{f}$.
Les modèles IBM 1 et 2 différent sur les simplifications apportées à la définition de $P(\mathbf{f}, a|\mathbf{e})$.

a) Le modèle IBM 1

Brown *et al.* font plusieurs hypothèses simplificatrices. La première est que tous les alignements au sein d'une paire de traduction sont équiprobables. Ainsi, chaque mot f_j peut être connecté à

n'importe quel mot e_i avec la même probabilité. Cette supposition implique que l'ordre des mots dans $\mathbf{f}$ et $\mathbf{e}$ n'affecte en aucun cas la valeur de $P(\mathbf{f}, a|\mathbf{e})$. La deuxième hypothèse est que $P(J|\mathbf{e})$ est une constante. Ces hypothèses conduisent à l'équation suivante :

$$\begin{aligned} P(\mathbf{f}, a|\mathbf{e}) =& \frac{\epsilon}{(I+1)^J} \prod_{j=1}^{J} t(f_j|e_{a_j}) \\ P(\mathbf{f}|\mathbf{e}) =& \frac{\epsilon}{(I+1)^J} \prod_{j=1}^{J} \sum_{i=0}^{I} t(f_j|e_i) \end{aligned} \tag{2.18}$$

où $t(f_j|e_{a_j})$ est la probabilité de traduction de e_{a_j} en f_j. Le modèle 1 repose donc sur une table de traduction de mots dans laquelle sera stockée les probabilités de traduction entre chaque mot source et chaque mot cible.

b) Le modèle IBM 2

Le modèle 2 ignore l'hypothèse que tous les alignements sont équiprobables et par conséquent introduit un modèle d'alignement. Ainsi, chaque alignement possible au sein d'une paire de phrase se voit attribuer une certaine probabilité. Cette probabilité dépend à la fois des mots que l'alignement connecte mais aussi de la taille des phrases. L'ordre des mots a donc cette fois une importance dans le calcul de $P(\mathbf{f}, a|\mathbf{e})$:

$$\begin{aligned} P(\mathbf{f}, a|\mathbf{e}) =& \epsilon \prod_{j=1}^{J} t(f_j|e_{a_j}) a(a_j|j, J, I) \\ P(\mathbf{f}|\mathbf{e}) =& \frac{\epsilon}{(I+1)^J} \prod_{j=1}^{J} \sum_{i=0}^{I} t(f_j|e_i) a(i|j, J, I) \end{aligned} \tag{2.19}$$

où $a(i|j, J, I)$ est la probabilité que le mot e_i à la position $a_j = i$ soit connecté au mot f_j à la position j. Le modèle 2 repose donc sur une table de traduction de mots et sur un modèle d'alignement.

c) Le modèle HMM

Le modèle HMM a été proposé par Vogel *et al.* en 1996 [Vogel 96] en amélioration au modèle IBM 2. En effet, la motivation première a été le constat que pour les langues Indo-européennes, l'alignement des mots a tendance à former des groupes plutôt que d'être distribué de façon arbitraire. Si un mot source est aligné à un certain mot cible, le mot source suivant a tendance, en général à s'aligner avec un mot cible proche du mot cible précédemment aligné. Or, le modèle d'alignement issu du modèle IBM 2 $P(a_j|j, I, J)$ est un modèle d'ordre 0. Ceci signifie que la probabilité que le mot f_j à la position j soit lié au mot e_i à la position a_j ne dépend que de j et de la taille des phrases. Il ne tient pas compte de l'alignement du mot précédent. Au contraire, le modèle HMM introduit un modèle d'alignement du 1^{er} ordre $P(a_j|a_{j-1}, I)$. La probabilité que le mot f_j à la position j soit aligné avec le mot e_i à la position a_j dépend de l'alignement du mot f_{j-1} précédent et de la taille de la phrase e.

Pour les modèles IBM 1 et 2, le processus de génération de f et a sachant e aligne chaque position de f à une position de la phrase e et identifie ensuite chaque mot de f suivant le mot dans e auquel il est lié par l'alignement. Ce processus est différent pour les modèles IBM 3, 4 et 5 que nous décrivons dans les sections suivantes.

d) Les modèles IBM 3, 4 et 5

Pour les modèles IBM 3, 4 et 5, le processus de génération de f et a se déroule en trois étapes. La première étape consiste à choisir, pour chaque mot e_i, le nombre de mots qui lui seront connectés dans f. Ceci introduit la notion de fertilité, c'est-à-dire le nombre de mots dans f générés par un mot e_i. Dans les modèles 1 et 2, la fertilité d'un mot est obtenue de manière implicite en ce sens que l'alignement permet de lier plusieurs f_j à un même e_i. Les modèles 3, 4 et 5 introduisent explicitement un modèle de fertilité noté $n(\phi_i|e_i)$ qui estime pour chaque mot e_i le nombre de mots ϕ_i qui lui seront connectés dans f. Par exemple $n(2|President)$ est la probabilité que le mot *President* génère deux mots dans la traduction. La deuxième étape identifie le ou les mots f_j connectés à chaque mot e_i à l'aide des probabilités $t(f_j|e_i)$ de la table de traduction de mots. Le tableau 2.1 présente des exemples de probabilités de fertilité et de traduction de trois mots anglais. Ces exemples sont intéressants. En effet, pour le mot *farmers*, le modèle de fertilité indique qu'il est plus probable qu'il génère deux mots français qu'un seul. De plus, le modèle de traduction estime que ce mot peut-être traduit aussi bien par *agriculteurs* que par *les*. Ceci nous indique donc que *farmers* se traduit probablement par *les agriculteurs* en Français. De même, le modèle de fertilité indique que *the* peut ne générer aucun mot en Français avec une probabilité de $0,254$.

Enfin la troisième étape attribue aux mots identifiés dans e une position. Cette étape fait inter-

the				*farmers*				*not*			
f	$t(f\|e)$	n	$n(\phi\|e)$	f	$t(f\|e)$	n	$n(\phi\|e)$	f	$t(f\|e)$	n	$n(\phi\|e)$
le	$0,497$	1	$0,746$	agriculteurs	$0,442$	2	$0,731$	ne	$0,497$	2	$0,735$
la	$0,207$	0	$0,254$	les	$0,418$	1	$0,228$	pas	$0,442$	0	$0,154$
les	$0,155$			cultivateurs	$0,046$	0	$0,039$	non	$0,029$	1	$0,107$
l'	$0,086$			producteurs	$0,021$			rien	$0,011$		
ce	$0,018$										
cette	$0,011$										

TAB. 2.1 – Probabilités de fertilité et de traduction extraites de [Brown 93]

venir un modèle de distorsion noté $d(j|a_j, I, J)$ qui estime la probabilité que la position j dans la phrase e soit connectée à la position a_j dans la phrase f. La figure 2.5 illustre le processus de traduction suivi par les modèles 3, 4 et 5.

Dans les modèles 1 et 2, chaque position de la phrase f est connectée à une position dans la phrase e (alignement) et les mots de f sont ensuite identifiés (traduction). Dans les modèles 3, 4 et 5, chaque mot de e produit un ou plusieurs mots dans f (fertilité) qui sont ensuite identifiés (traduction) puis liés à une position dans la phrase e (distorsion). Le cas du mot spécial $NULL$ est considéré à part. En effet, les modèles 3, 4 et 5 ne considèrent pas de probabilités de fertilité pour le mot NULL. Au lieu de cela, à chaque fois qu'un mot réel dans e génère un mot dans f, un nouveau mot "libre" (c-a-d connecté à aucun mot de e) sera produit avec une probabilité $p1 = 1 - p0$. Les probabilités $t(f|NULL)$ permettront d'identifier ces mots "libres". Leur place dans f est choisie parmi les places laissées vides après le positionnement des mots alignés avec de vrais mots dans e.

Les modèles 3, 4 et 5 font donc intervenir plusieurs sous-modèles, à savoir : le modèle de traduction $t(f_j|e_i)$, le modèle de fertilité $n(\phi_i|e_i)$, le modèle de distorsion $d(j|a_j, I, J)$ semblable

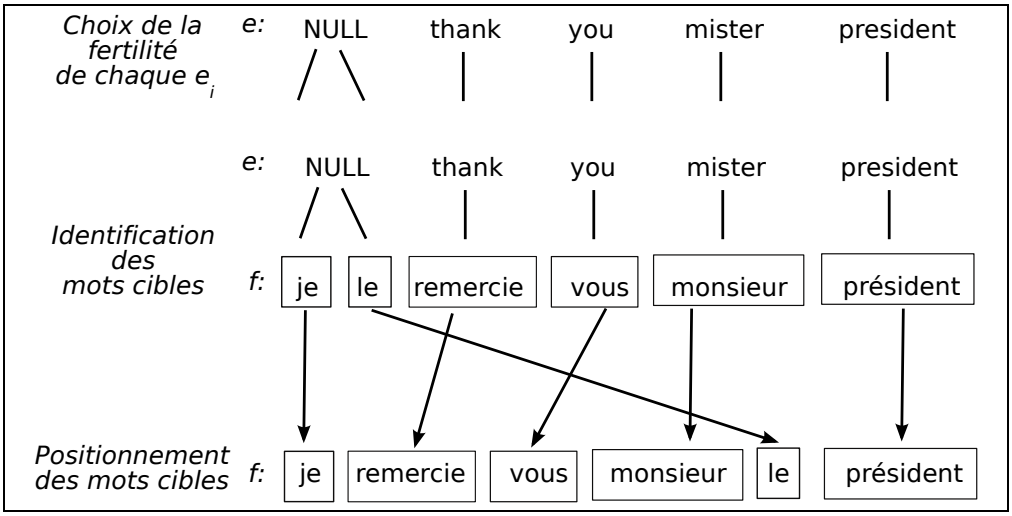

FIG. 2.5 – Processus de traduction simplifié des modèles 3 à 5 d'IBM [Jur 06]

au modèle d'alignement des deux premiers modèles et aux modèles $p0$ et $p1$ qui permettent de générer et placer des mots cibles "libres". Les modèles 3 et 4 différent sur le modèle de distorsion.

Pour le modèle 3, la probabilité d'un lien entre un mot e_i et un mot f_j dépend de la position de ces mots et de la taille des phrases (comme dans le modèle 2) ce qui conduit à l'équation suivante :

$$\begin{aligned} P(\mathbf{f}, a|\mathbf{e}) = \binom{J-\phi_0}{\phi_0} p_0^{J-2\phi_0} p_1^{\phi_0} \times \frac{1}{\phi_0!} \times \prod_{i=0}^{I} \phi_i! \\ \times \prod_{i=1}^{I} n(\phi_i|e_i) \times \prod_{j=1}^{J} t(f_j|e_{a_j}) \times \prod_{j:a_j \neq 0}^{J} d(j|a_j, I, J) \end{aligned} \tag{2.20}$$

Les deux premiers termes de l'équation modélisent la génération et l'insertion des mots "libres" dans f. Le troisième terme est un facteur qui rend compte du nombre d'alignements différents possibles entre un mot de e et les mots dans f auxquels il est lié. Enfin les derniers termes font intervenir les modèles de fertilité, de traduction et de distorsion.

Pour le modèle 4, la probabilité de distorsion dépend, en plus de la position du mot considéré dans e et du mot dans f, des mots eux-mêmes mais aussi de la position des autres mots dans f

également liés au mot considéré dans e.

$$\begin{aligned} P(\mathbf{f}, a|\mathbf{e}) = & \prod_{i=1}^{I} n(\phi_i|e_i) \\ & \times t(f_{\tau_{i,1}}|e_i) d_{=1}(\tau_{i,1} - \overline{\tau_{i-1}}|\mathcal{E}_{i-1}\mathcal{F}(f_{\tau_{i,1}})) \\ & \times \prod_{k=2}^{\phi_i} t(f_{\tau_{i,1}}|e_i) d_{>1}(\tau_{i,k} - \tau_{i,k-1}|\mathcal{F}(f_{\tau_{i,k}})) \\ & \times \binom{J-\phi_0}{\phi_0} p_0^{J-2\phi_0} p_1^{\phi_0} \times \frac{1}{\phi_0!} \\ & \times \prod_{k=2}^{\phi_0} t(f_{\tau_{0,k}}|e_0) \end{aligned} \tag{2.21}$$

$\tau_{i,k} \in [1, J]$ est la position du $k^{\text{ième}}$ mot dans f produit par le mot e_i. Le modèle de distorsion $d_{=1}(\Delta_j|\mathcal{E}\mathcal{F})$ détermine la place du premier mot produit par e_i et le modèle de distorsion $d_{>1}(\Delta_j|\mathcal{F})$ positionne les autres mots également générés par e_i si toutefois il en existe. Le modèle 4 fait également intervenir dans son modèle de distorsion les classes des mots. Ainsi, $\mathcal{F}(f)$ dénote la classe du mot f alors que $\mathcal{E}_{i-1}$ dénote la classe du mot précédant e_i et ayant produit au moins un mot dans la phrase$f$. Enfin, $\overline{\tau_{i-1}}$ indique la moyenne des positions des mots dans f reliés à ce mot précédant e_i. Nous invitons le lecteur à lire [Brown 93] pour de plus amples détails.

Brown *et al.* ont montré que les modèles 3 et 4 sont déficients. Certaines probabilités de ces modèles sont distribuées sur des unités qui ne sont pas des suites de mots en soi. Ils ont donc proposé le modèle 5, très semblable au modèle 4 à la différence que ce dernier donne des résultats beaucoup plus satisfaisants.

2.4.2.2 Apprentissage des modèles IBM et HMM

L'apprentissage des paramètres de ces modèles se fait de manière itérative suivant l'algorithme Expectation-Maximisation [Dempster 77] en utilisant notamment les redondances existantes dans les corpus parallèles utilisées.

L'outil Giza++ [Och 03b] permet l'apprentissage de tous les paramètres de ces modèles uniquement à partir d'un corpus parallèle d'apprentissage aligné au niveau de la phrase. Les modèles de complexité croissante sont entraînés successivement. Une fois déterminé, chaque ensemble de paramètres d'un modèle constitue les paramètres de départ du modèle suivant.

2.4.2.3 Discussion

L'unité de traduction des modèles d'IBM et du modèle HMM est le mot ce qui apportent plusieurs difficultés dans la modélisation du processus traduction. En effet, il faut tenir compte du fait qu'un mot peut se traduire par plusieurs mots , mais aussi que l'ordre des mots d'une phrase n'est pas nécessairement respecté lorsqu'ils sont traduits. Ces différents phénomènes sont respectivement traités par le modèle de fertilité et le modèle de traduction de mots et aussi par le modèle d'alignement ou de distorsion.

Toutefois, plusieurs problèmes persistent. Prenons quelques exemples pour lesquels la traduction se ferait du Français vers l'Anglais. Les modèles précédemment décrits autorisent qu'un

même mot anglais se traduise par plusieurs mots français. Il est donc possible de traduire la suite française *remporte une médaille* par le mot anglais *medal*, en supposant que le modèle de fertilité privilégie une fertilité de 3 et non 1 ou 2 pour *medal*. En revanche, le modèle d'alignement n'autorise pas qu'un même mot français soit traduit par plusieurs mots anglais. Rappelons, en effet, que plusieurs mots cibles peuvent être alignés au même mot source mais que l'inverse est faux suivant les modèles proposés dans [Brown 93]. La traduction du mot français *sortira* posera donc problème puisqu'il se traduit en anglais par *will go out* et que les modèles à base de mots que nous venons de décrire ne permettent pas de rendre compte de ce phénomène.

Très vite, l'idée de prendre comme unité de traduction une séquence de mots, et non plus un mot seul, est apparue en traduction statistique pour modéliser le fait que plusieurs mots peuvent être alignés à plusieurs mots. Plusieurs modèles à base de séquences ont été proposés à partir de 1999. Nous les évoquons dans la section suivante.

2.4.3 les modèles à base de séquences

L'utilisation de séquences de mots comme unités de traduction rend le processus de traduction plus simple que ceux modélisés par Brown et al. ou Vogel *et al.*
Généralement, les modèles à base de séquences découpent le processus de traductions en trois étapes. Dans un premier temps, la phrase source $\mathbf{e}$ est découpée en I séquences notées $\bar{e_1}\bar{e_2}\dots\bar{e_I}$. Chacune des séquences $\bar{e_i}$ est ensuite traduite dans une langue cible par le biais d'une table de traduction de séquences de la forme $t(\bar{e_i}|\bar{f_j})$ qui attribue une probabilité de traduction, entre autres, à tous les couples $(\bar{e_i}, \bar{f_j})$, avec $\bar{e_i}$ et $\bar{f_j}$ respectivement une séquence de mots sources et une séquence de mots cibles. Enfin, les séquences cibles sont éventuellement réordonnés à l'aide d'un modèle de distorsion pour produire la phrase cible finale $\mathbf{f} = \bar{f_1}\bar{f_2}\dots\bar{f_I}$. La figure 2.6 reprend l'exemple de traduction utilisé pour expliquer les modèles à base de mots mais illustre cette fois le processus d'un modèle basé sur les séquences de mots.
Le processus de traduction par séquence de mots repose donc en grande partie sur une table de

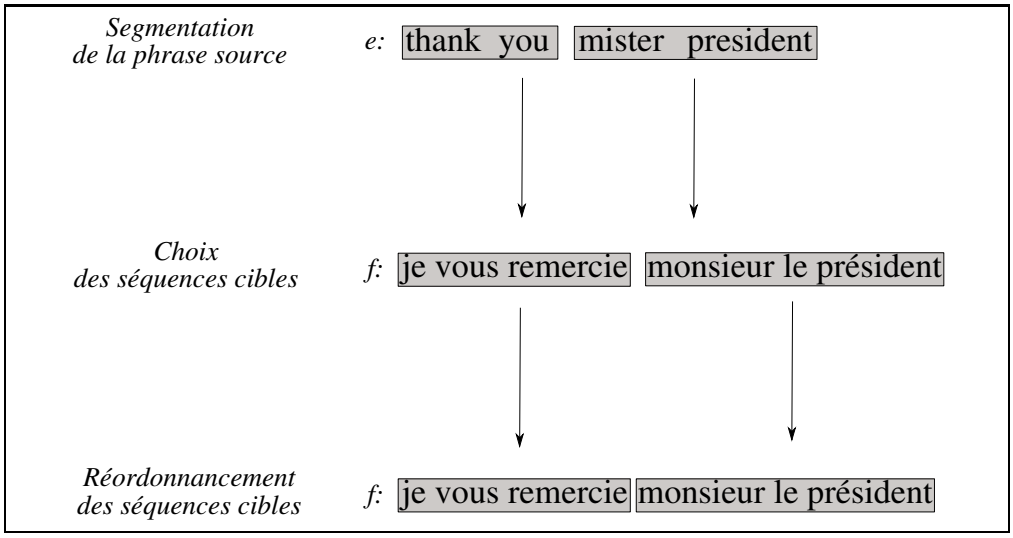

FIG. 2.6 – Processus de traduction d'un modèle basé sur les séquences de mots [Jur 06]

traduction qui va établir les correspondances lexicales entre les séquences de la langue et celles de

la langue cible associées à des probabilités de traduction. Dans la section suivante, nous décrivons plusieurs méthodes d'apprentissage de cette table.

2.4.3.1 Apprentissage des modèles de traduction de séquences de mots

Och *et al.* comptent parmi les premiers chercheurs à avoir proposé un modèle de traduction basée non plus sur les mots mais sur les séquences [Och 99]. Leur approche par patrons d'alignement (*alignment template approach*) permet d'aligner une séquence de classes de mots sources avec une séquence de classes de mots cibles. L'utilisation des classes de mots plutôt que des mots eux-mêmes implique une généralisation des patrons. L'alignement des séquences de classes se fait par l'intermédiaire de l'alignement des mots calculé automatiquement à partir des paramètres d'un modèle d'alignement qui estime les probabilités qu'un mot à la position i soit aligné avec un mot à la position j.

Pour améliorer la qualité de l'alignement au niveau du mot, Och et al proposent de symétriser le processus d'alignement. Pour cela, un modèle HMM [Vogel 96] est appris dans les deux sens de traduction $e \rightarrow f$ et $f \rightarrow e$ et va permettre de déterminer, pour chaque paire de phrases, l'alignement A_1 qui maximise $P(\mathbf{f}, \mathbf{a}|\mathbf{e})$ et l'alignement A_2 qui maximise $P(\mathbf{e}, \mathbf{a}|\mathbf{f})$. A_1 et A_2 sont appelés alignements de Viterbi. Ils sont ensuite combinés pour former l'alignement final A. Dans un premier temps, A est constitué de l'intersection de A_1 et A_2, c'est-à-dire de tous les couples de mots alignés à la fois dans A_1 et dans A_2. Ensuite, lui sont ajoutés itérativement des couples qui ne font pas partie de l'intersection et répondant à certaines heuristiques. Les études ont montré que cette combinaison d'alignements aboutit à une précision plus grande de l'alignement de mots comparé aux alignements de chacune des directions de traduction. Toutes les paires de séquences consistantes avec l'alignement A ainsi déterminé sont ensuites extraites. Une paire de séquences est consistante avec A si les mots de la séquence source sont alignés seulement avec les mots de la séquence cible. La probabilité d'utiliser un patron d'alignement est ensuite estimée par la fréquence relative de l'alignement de sa séquence de classes de mots sources avec sa séquence de classes de mots cibles.

Les travaux de Och *et al.* ont par la suite inspiré plusieurs systèmes de traduction à base de séquences [Zens 02, Tillmann 03, Zhang 03] et notamment celui et Koehn et al [Koehn 03]. Ces derniers proposent en effet un système pour lequel la probabilité $P(\mathbf{f}|\mathbf{e})$ est formulée de la manière suivante :

$$\begin{aligned} P(\mathbf{f}|\mathbf{e}) =& P(\bar{f}_1^I|\bar{e}_1^I) \\ =& \prod_{i=1}^{I} t(\bar{e}_i|\bar{f}_i) d(a_i - b_{i-1}) \end{aligned} \tag{2.22}$$

où $d(a_i - b_{i-1})$ dénote le modèle de distorsion avec a_i la position du premier mot de la séquence cible traduite par la séquence source $\bar{e}_{i-1}$ et b_{i-1} la position du dernier mot de la séquence cible traduite par la $(i-1)^{\text{ème}}$ séquence source $\bar{e}_{i-1}$. Dans leurs travaux, les auteurs utilisent un modèle de distorsion très simple : $d(a_i - b_{i-1}) = \alpha^{\|a_i - b_{i-1} - 1\|}$. Les paramètres $t(\bar{e}_i|\bar{f}_j)$ de la table de traduction sont appris suivant la méthode proposée par Och *et al.*[Och 99]. Nous en donnons ci-dessous un aperçu au travers d'un exemple.

La figure 2.7 illustre l'algorithme d'extraction de séquences suivi par Koehn *et al.* et appliqué sur une paire de phrases extraite du corpus d'apprentissage EUROPARL (cf. chapitre 3 pour

plus de détails sur ce corpus). La première étape consiste à calculer l'alignement de Viterbi dans le sens anglais-français puis dans le sens français-anglais. Ces alignements sont obtenus à l'aide de l'outil Giza++. Ils sont respectivement représentés par les matrices qui se situent dans la partie supérieure de la figure. Les cellules non vides indiquent les mots qui sont alignés. Les alignements sont ensuite combinés, tout d'abord en ne conservant que les couples de mots alignés au sein de l'intersection. Ils sont représentés dans les deux matrices par les cellules les plus foncées. L'intersection est ensuite étendue itérativement en y ajoutant des couples de mots faisant partie de l'union mais sous certaines conditions. En effet, un couple est ajouté s'il est voisin d'un couple déjà présent dans l'alignement résultant de la combinaison et si le mot source ou le mot cible constituant le point n'est encore aligné avec aucun autre mot de la matrice. Un couple est considéré comme voisin d'un autre s'il lui est adjacent. Enfin, s'il reste des couples de l'union pour lesquels ni le mot source, ni le mot cible n'est aligné dans la combinaison alors ils sont également ajoutés. L'alignement résultant est représenté dans la matrice de la partie inférieure de la figure. Toutes les paires de séquences consistantes avec cet alignement sont ensuite extraites. Une paire de séquences est consistante si les mots sources au sein de la paire sont alignés uniquement avec les mots cibles au sein de cette même paire et inversement. Dans la figure 2.7, les paires de séquences qu'il est possible d'extraire sont encadrées par un trait épais. Il s'agit d'un sous-ensemble non-exhaustif des paires de séquences possibles avec cette paire de phrases.

La probabilité des paires de séquences ainsi extraites est définie suivant leurs fréquences relatives :

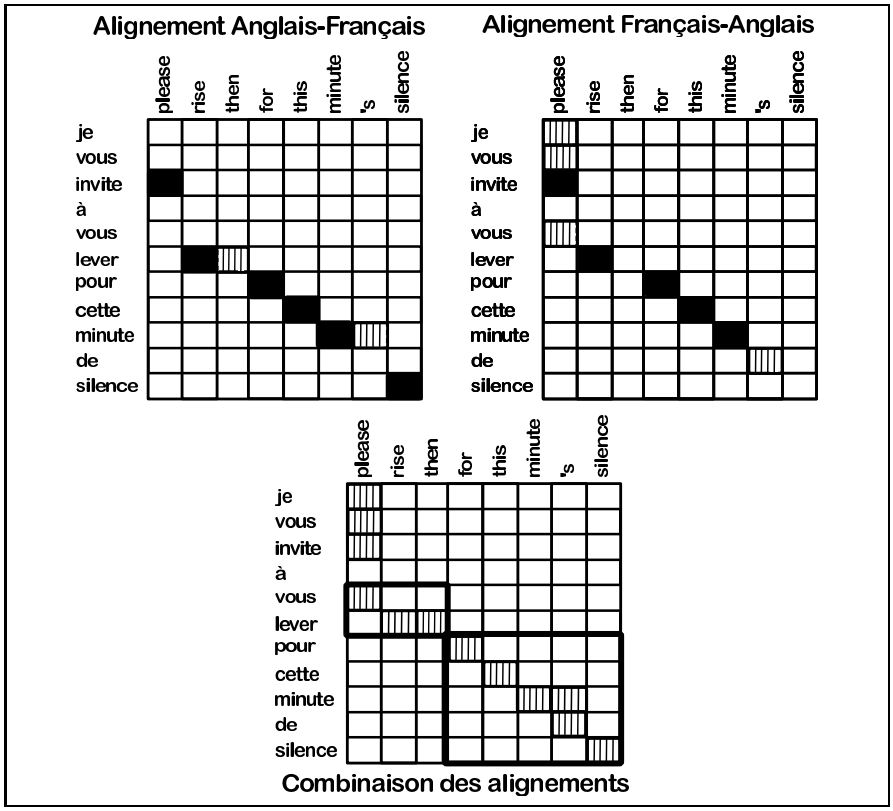

FIG. 2.7 – Illustration de la méthode d'extraction de séquences suivie par Koehn *et al.* Les alignements Français-Anglais et Anglais-Français sont représentés sous forme de matrices dans lesquelles les cellules non-vides représentent les mots liés au sein des ces alignements. Les cellules pleines noires indiquent les mots qui sont liés dans un alignement comme dans l'autre. Dans la combinaison des alignements, les cadres indiquent les paires de séquences qu'il est possible d'extraire d'après l'alignement des mots.

$$t(\bar{f}|\bar{e}) = \frac{N(\bar{f},\bar{e})}{\sum_{\bar{f}} N(\bar{f},\bar{e})} \tag{2.23}$$

où $N(\bar{f},\bar{e})$ est le nombre de fois où $\bar{f}$ et $\bar{e}$ sont alignés dans le corpus d'apprentissage.

Koehn *et al.* comparent cette méthode d'extraction avec une méthode similaire mais qui impose en plus que les constituants d'une paire de séquences soient des sous-arbres dans l'analyse syntaxique des phrases comme proposé dans [Yamada 01]. Les études ont montré que cette contrainte supplémentaire réduit considérablement le nombre de paires de séquences extraites et conduit par conséquent à des qualités de traduction moins performantes.

Nous venons de voir et de citer quelques méthodes d'extraction de paires de séquences reposant sur l'alignement préalable des mots. Plusieurs autres méthodes proposent d'établir les tables de traductions de séquences sans passer par cette étape. C'est le cas notamment des travaux de Marcu *et al.* [Marcu 02] qui présentent un modèle de probabilités jointes pour mettre en évidence les équivalences de mots et de séquences directement à partir d'un corpus bilingue. Alors que la plupart des modèles tentent de capturer la façon dont la phrase source est alignée avec la phrase cible en établissant des correspondances lexicales au niveau du mot, Marcu et al. proposent un modèle qui génère la phrase source et la phrase cible simultanément. Ainsi, dans les approches classiques, le modèle de traduction explique comment les mots sources sont mis en correspondances avec les mots cibles et comment ces mots cibles sont arrangés pour former la phrase cible en estimant la probabilité conditionnelle $P(\mathbf{f}|\mathbf{e})$. Le modèle de probabilité jointe, va lui, décrire la manière dont les phrases source et cible d'une paire de traduction sont générées simultanément en estimant la probabilité jointe $P(\mathbf{e},\mathbf{f})$. Ce modèle donne lieu à deux distributions : $t(\bar{e},\bar{f})$ qui est la probabilité jointe des séquences $\bar{e}$ et $\bar{f}$ et $d(i,j)$ qui est la probabilité de distorsion entre les positions i et j au sein d'une paire de phrases. Ces distributions sont apprises en utilisant l'algorithme EM. Elles sont marginalisées pour devenir des probabilités conditionnelles et peuvent ainsi être employées pour le calcul de la meilleure traduction sachant une phrase source.

2.4.3.2 Discussion

Utiliser comme unité de traduction des séquences de mots plutôt que le mot seul s'est avéré beaucoup plus performant en terme de qualité de traductions. En effet, contrairement à la modélisation au niveau du mot, la modélisation au niveau des séquences de mots permet de prendre en compte le contexte local. Considérons la suite de mots anglaises suivante : *electric ring*. Un modèle de traduction à base de mots n'est capable de lever l'ambiguïté sur le mot *ring* qui peut signifier en français aussi bien *plaque* que *anneau* ou encore *sonner*, que par les fréquences de co-occurrences de ces paires de mots dans le corpus d'apprentissage. En revanche, considérer cette même suite de mots comme une unité à part entière et non plus comme deux mots séparés permet au modèle de traduction de tenir compte du contexte de *ring*. Ainsi, sachant que le mot précédent est *electric*, celui-ci est plus apte à traduire cette suite de mots par *plaque électrique* plutôt que par *anneau électrique*.
Un modèle de traduction à base de séquences de mots permet également d'éviter bon nombre de problèmes de réordonnancement des mots. L'ordre des mots au sein d'une séquence est explicitement pris en compte. Reprenons notre exemple, et supposons que le modèle de traduction à base de mots a su traduire *ring* en *plaque*. Le modèle de distorsion doit agir ensuite pour placer les mots français dans l'ordre adéquat, à savoir *plaque électrique* et non pas *électrique plaque* comme le suggérerait l'ordre des mots dans la suite anglaise. Ce réordonnancement est directement géré

dans un modèle à base de séquences, *electric ring* se traduit en une unité par *plaque électrique*, le modèle de distorsion n'a aucun besoin d'agir dans ce cas là.

Les études menées dans [Koehn 03] ont montré que les meilleures performances en terme de qualité de traduction suivant le score BLEU sont obtenues avec la méthode d'extraction de paires de séquences inspirée des travaux de Och *et al.* Nous avons donc utilisé cette méthode pour apprendre le modèle de traduction de référence qui constituera le modèle "baseline" auquel nous nous comparerons lors de nos différentes expérimentations.

2.5 Décodeur

2.5.1 Généralités

Le décodeur est la partie centrale de tout système de traduction. En effet, étant donné une phrase d'entrée, le décodeur est chargé de produire la ou les n meilleures traductions possibles, à l'aide des paramètres estimés par les modèles impliqués dans le processus de traduction et appris préalablement sur un corpus d'apprentissage bilingue aligné. Le décodeur est le composant dans un système de traduction statistique qui va implanter la fonction *argmax* de l'équation 2.4.

Chercher la meilleure traduction d'une phrase source parmi toutes les phrases cibles possibles est un problème NP-complet en traduction automatique à base de mots [Knight 99]. Les décodeurs emploient donc des algorithmes gloutons ou de programmation dynamique pour rendre le processus de décodage possible et efficace. Nous pouvons notamment citer les travaux de [Wang 97, Tillmann 00, Och 01, Germann 03] qui proposent des algorithmes de décodage dédiés aux modèles à base de mots.

Le décodeur PHARAOH [Koehn 04] est très largement utilisé pour la traduction à base de séquences de mots. Il constituait notamment le système de traduction baseline lors de la campagne d'évaluation WMT06 [7].
Depuis 2007, PHARAOH a laissé la place à un autre décodeur appelé MOSES[8][Koehn 07b]. Ce dernier reprend le même algorithme que son prédécesseur mais offre plusieurs extensions. Il permet notamment le décodage de réseaux de confusion et laisse la possibilité d'intégrer des modèles de traduction factorisés et ainsi peut prendre en compte des connaissances linguistiques [Koehn 07a]. Nos travaux en traduction automatique ont débuté en 2005. Nous avons démarré nos expérimentations avec le décodeur PHARAOH et par soucis de cohérence des comparatifs des résultats, nous avons continué à utiliser ce décodeur. Nous détaillons son algorithme dans la section suivante.

2.5.2 Cas particuliers : le décodeur PHARAOH

Pour traduire d'une langue source vers une langue cible, le problème d'un décodeur consiste à trouver parmi toutes les phrases de la langue cible possibles, celle qui sera la meilleure traduction de la phrase source. Dans le but de réduire l'espace de recherche, Pharaoh ne va considérer que les phrases cibles composées de séquences de mots qui sont des traductions possibles de séquences de la phrase d'entrée selon la table de traduction. Ces traductions possibles sont appelées options

[7] http ://statmt.org/wmt06/
[8] http ://www.statmt.org/moses

de traduction. Une option de traduction est, en fait, une entrée de la table de traduction qu'il est possible de mettre en correspondance avec une portion de la phrase d'entrée. La figure 2.8 représente un exemple de treillis d'options de traduction qu'il serait possible d'appliquer sur la phrase d'entrée "*please rise then for this minute 's silence*". Le problème consiste alors à trouver

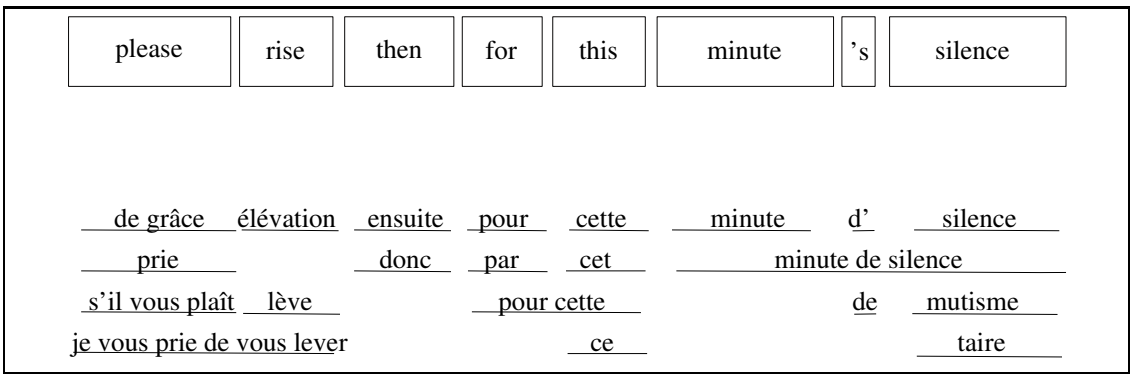

FIG. 2.8 – Exemple de treillis d'options de traduction. A chaque mot ou groupe de mots de la phrase source à traduire correspond un ou plusieurs mots dans la langue cible qui constituent des options de traduction. L'ensemble des options de traduction forme un treillis duquel doit-être extraite la meilleure traduction.

une combinaison de ces options de traduction permettant de produire la meilleure sortie possible. Pour cela, PHARAOH utilise un algorithme de recherche par faisceau (*beam search algorithm*).

2.5.2.1 Construction de l'espace de recherche

L'algorithme de recherche par faisceau est un algorithme de recherche en largeur qui permet de parcourir un graphe, dont les nœuds représentent des états, de manière itérative et niveau par niveau en utilisant une file d'attente. L'objectif est de trouver un chemin optimal dans le graphe entre un nœud initial et un nœud final. Un chemin est étendu par ajout de nœuds. Une évaluation heuristique est utilisée sur chaque nœud afin d'estimer le meilleur chemin y passant. Chaque nœud est ensuite ajouté à une file d'attente donnant la priorité aux nœuds les moins coûteux. A chaque niveau, certains chemins candidats sont élagués et seul un nombre prédéfini d'états servant à construire un chemin complet liant le nœud initial au nœud final est conservé. Ce nombre prédéfini d'états conservés est appelé largeur du faisceau. Le chemin le moins coûteux est considéré comme étant le meilleur.

Dans le cas du décodeur PHARAOH, un état est considéré comme une hypothèse de traduction. Partant de l'hypothèse de traduction initiale pour laquelle aucun mot de la phrase d'entrée n'est encore traduit, le décodeur va progressivement l'étendre en utilisant une option de traduction qui couvre des mots de la phrase source qui ne sont pas encore traduits. Il va ainsi créer de nouvelles hypothèses jusqu'à aboutir à des états finaux pour lesquels tous les mots de la phrase source sont traduits. La figure 2.9 montre quelques étapes de décodage de la phrase anglaise prise en exemple précédemment. E : représente la phrase anglaise que le décodeur doit traduire. Les astérisques indiquent que le mot n'a pas été traduit. La présence de la première lettre du mot indique qu'il a été traduit par un ou plusieurs mots français. F : représente la phrase partielle produite. Elle est générée de gauche à droite. Un état du graphe de recherche est donc une hypothèse partielle composée d'un lien vers l'état précédent, des mots de la phrase d'entrée couverts, des mots produits en sortie et d'un coût.

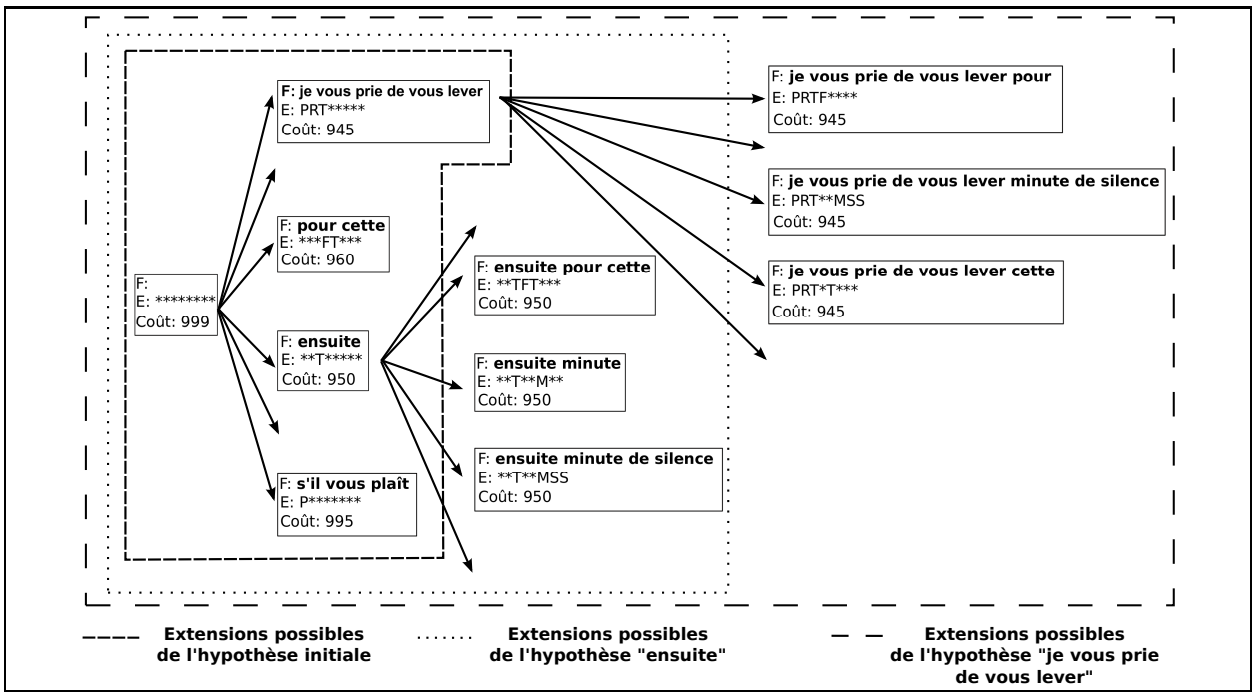

FIG. 2.9 – Extensions d'hypothèses pour le décodage de la phrase "*please rise then for this minute 's silence*"

Le coût d'une hypothèse tient compte du coût de l'hypothèse produite jusqu'ici ainsi que de l'estimation du coût futur. Le coût d'un état du graphe, ou d'une hypothèse partielle, est le coût de l'état précédant multiplié par le coût de production de la nouvelle séquence. Le coût de production d'une nouvelle séquence est un produit de probabilités :

$$COUT_{production}(\bar{f},\bar{e}) = P_{TM}(\bar{e}|\bar{f})^{\lambda_{TM}} * P_D(\bar{e}|\bar{f})^{\lambda_D} * P_{LM}(\bar{f})^{\lambda_{LM}} * \omega^{long(\bar{f})\lambda_W} \quad (2.24)$$

où $\bar{f}$ est la dernière séquence produite de l'hypothèse partielle et $\bar{e}$ les mots de la phrase d'entrée couverts par cette nouvelle séquence. P_{TM}, P_D et P_{LM} dénotent respectivement les probabilités des modèles de traduction, de distorsion et de langage. P_{TM} rend compte de la vraisemblance que $\bar{f}$ soit traduit en $\bar{e}$. P_D rend compte du nombre de mots séparant les deux dernières séquences traduites dans la phrase d'entrée. Plus les deux séquences sont éloignées l'une de l'autre dans la phrase d'entrée, plus le coût de distorsion sera important. Le modèle de distorsion qu'emploie PHARAOH est un modèle simple qui définit la probabilité de distorsion de deux séquences comme suit :

$$P_D(\bar{e_i}|\bar{f_j}) = d(a_j - b_{j-1}) = \alpha^{|a_j - b_{j-1} - 1|} \quad (2.25)$$

a_i est la position dans la phrase d'entrée du premier mot de la séquence traduite par la $i^{\text{ème}}$ séquence de l'hypothèse. b_{i-1} est le dernier mot de la séquence dans la phrase d'entrée traduite par la $i-1^{\text{ème}}$ séquence de l'hypothèse. α est un paramètre dont la valeur est à définir en phase d'optimisation. La figure 2.10 donne les probabilités de distorsion de $P_D(minute's\ silence|minute\ de\ silence)$ et $P_D(for\ this|pour\ cette)$ sachant l'hypothèse courante *ensuite*. Du point de vue du modèle de distorsion, le coût de production de *minute de silence* est plus élevé que le coût de production de *pour cette*. En effet, dans le premier cas, les deux dernières séquences anglaises traduites sont séparées de 2 mots alors qu'elles sont contigües dans le second cas.

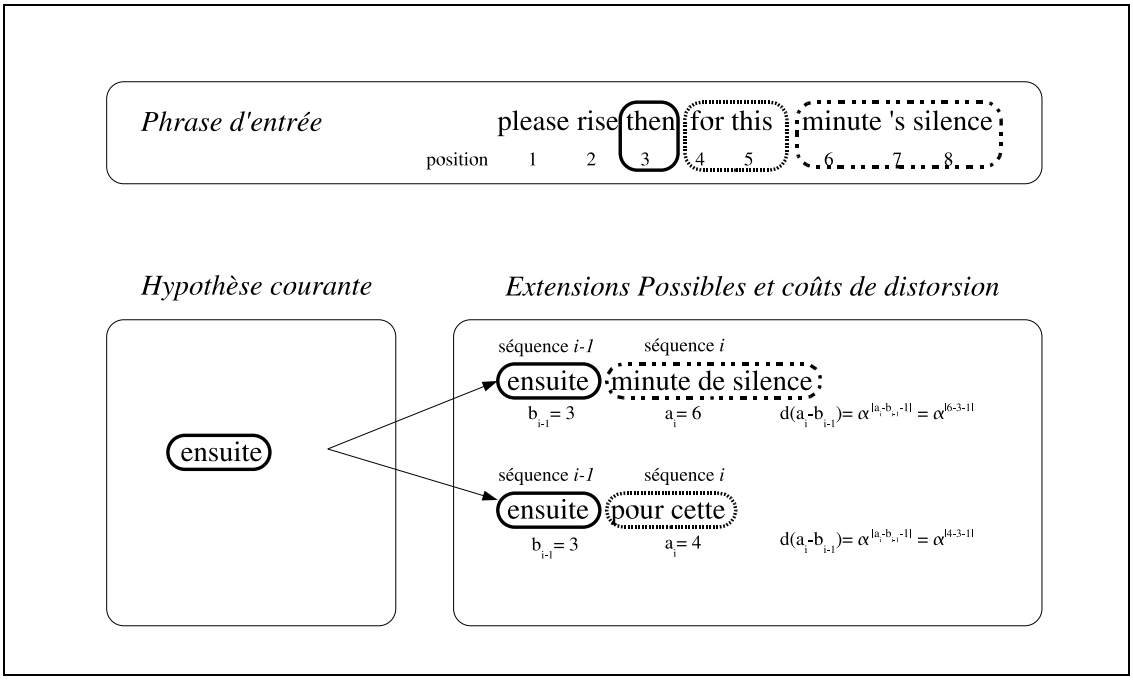

FIG. 2.10 – Probabilités de distorsion associées à deux extensions possibles de l'hypothèse de traduction "*ensuite*" par la production des séquences "*minute de silence*" et "*pour cette*"

P_{LM} estime la vraisemblance de la séquence $\bar{f}$ dans la langue cible. Enfin, ω est un modèle destiné à pénaliser les séquences produites trop longues ou au contraire trop courtes par rapport au nombre de mots couverts dans la phrase d'entrée. Chacune des probabilités est pondérée par un coefficient λ permettant de déterminer l'importance de chaque modèle dans le processus de décodage.
L'estimation du coût futur d'une hypothèse est liée aux mots de la phrase d'entrée qui ne sont pas encore couverts par l'hypothèse. Seuls les coûts de traduction et de modèle de langage de chaque options de traduction à appliquer sont considérés afin de couvrir le restant de la phrase d'entrée. Pour une hypothèse partielle, il existe différentes façons d'appliquer des options de traduction pour traduire le restant de la phrase d'entrée. Le décodeur utilise donc un algorithme de programmation dynamique afin de trouver le chemin final le moins coûteux et ainsi estimer le coût futur de l'hypothèse.

2.5.2.2 Optimisation de l'espace de recherche

Les différentes hypothèses de traduction partielles sont organisées sous forme de piles. Le décodeur gère une pile par nombre de mots traduits de la phrase d'entrée. Si la phrase d'entrée contient n mots, le décodeur gérera n piles et chacune des piles de 1 à n. La pile 1 contiendra les hypothèses de traduction couvrant de 1 mot de la phrase d'entrée, la pile 2 celles qui couvrent 2 mots... Les hypothèses sont placées dans une pile suivant le nombre de mots couverts de la phrase d'entrée. La figure 2.11 illustre ce procédé. Lorsqu'une nouvelle hypothèse est générée en utilisant une option de traduction, elle est déplacée dans une nouvelle pile. La priorité est donnée aux hypothèses les moins coûteuses. A chaque ajout d'une hypothèse dans une pile, celle-ci est élaguée afin d'optimiser l'espace de recherche. Plusieurs moyens d'élagage existent. Le premier est

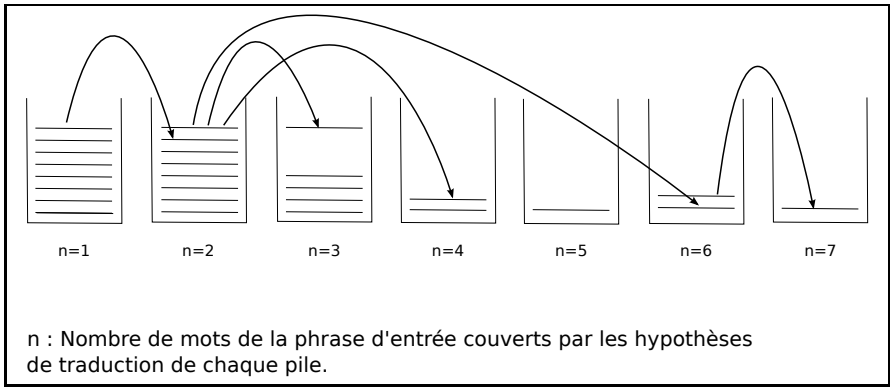

FIG. 2.11 – Gestion des hypothèses partielles sous forme de piles

la recombinaison des hypothèses. Deux hypothèses sont recombinées si elles couvrent les mêmes mots de la phrase d'entrée, si les deux derniers mots produits sont identiques et si la dernière séquence traduite dans la phrase d'entrée a la même position de fin pour les deux hypothèses. Dans ce cas, l'estimation du coût futur sera le même pour ces deux hypothèses. Celle ayant le coût le plus élevé est donc écartée sans risque de perdre le chemin le moins coûteux à la fin du processus. La recombinaison des hypothèses permet de réduire l'espace de recherche mais cette réduction reste insuffisante spécialement lorsqu'il s'agit de traduire de longues phrases. Il existe un autre moyen d'élagage propre à l'algorithme de recherche par faisceau qui consiste à écarter de chaque pile les hypothèses ne faisant pas partie du faisceau. C'est le coût d'une hypothèse qui détermine si oui ou non celle-ci est dans le faisceau. Ne considérer que le coût de production des hypothèses comme moyen d'élagage favoriserait les hypothèses ayant traduit les parties simples de la phrase d'entrée en premier. C'est pour cette raison que le coût d'une hypothèse tient compte de son coût de production mais aussi de l'estimation de son coût futur. Seul un nombre d'hypothèses se trouvant parmi les moins coûteuses est conservé dans le graphe de recherche. Le nombre d'hypothèses à conserver est défini par la taille du faisceau qui peut être déterminée par un seuil. Dans ce cas, les hypothèses ayant un coût plus élevé d'un certain facteur par rapport à la meilleure hypothèse de la pile sont supprimées de cette pile. La taille du faisceau peut également être fixée par avance. Dans ce cas, les n meilleures hypothèses sont conservées dans chaque pile. Ce moyen de réduire l'espace de recherche, contrairement à la recombinaison d'hypothèses présente des risques. En effet, si l'estimation du coût futur est inadéquate, il est possible de supprimer de l'espace de recherche certaines hypothèses qui auraient mené à l'hypothèse ayant le meilleur score à la fin du processus de décodage.

2.6 Discussion générale

La mise en place d'un système de traduction automatique statistique requiert principalement trois composants : une table de traduction qui établit des correspondances lexicales probables entre des mots ou des groupes de mots, un modèle de langage qui estime des probabilités d'apparition de séquences de mots cibles et un décodeur qui exploite les paramètres de ces modèles pour calculer la meilleure traduction automatique d'une phrase source.

L'apprentissage de ces modèles s'effectue sur des quantités importantes de données. Le mo-

dèle de langage est estimé sur un corpus monolingue alors que la table de traduction est estimé sur un corpus bilingue aligné composé de phrases sources associées à leur traduction dans une langue cible. L'acquisition de corpus bilingue aligné reste aujourd'hui une tâche difficile. En TA statistique, il est courant d'utiliser les corpus extraits des actes du Parlement Européen (cf. chapitre 3). Ces corpus retranscrivent de la parole peu naturelle puisqu'il s'agit des discours tenus par les membres du Parlement Européen dans un langage soutenu et spécifique. Nous pensons qu'il serait bénéfique d'estimer les tables de traduction sur des corpus retranscrivant de la parole plus naturelle et se rapprochant davantage de la parole spontanée en vu de mettre en place un système de traduction pour la parole spontanée de tous les jours. Nous proposons donc dans le chapitre 3 d'exploiter les sous-titres de films disponibles en différentes langues afin de construire automatiquement des corpus parallèles alignés.

Une fois les corpus bilingues alignés à disposition, l'estimation des tables de traduction est possible. Que l'unité de traduction soit le mot ou le groupe de mots, le point marquant concernant les méthodes d'apprentissage des tables de traduction est le recours quasi systématique aux modèles IBM. En effet, les tables de traduction de mots, comme nous l'avons évoqué dans le chapitre précédent, sont un des sous-modèles constituant les modèles IBM. Par conséquent, pour estimer une table de traduction de mots, il suffit d'entraîner les modèles IBM avec l'outil Giza++ par exemple et d'en extraire la table de traduction. Pour estimer une table de traduction de séquences, les méthodes les plus populaires s'appuient une fois de plus sur les modèles IBM (cf. chapitre précédent) . Cette fois encore, les modèles IBM sont entraînés mais c'est le modèle d'alignement et non plus la table de traduction de mots qui est utilisé. Celui-ci va en effet permettre de calculer de manière automatique les alignements des mots au sein des paires de traduction et c'est à partir de ces alignements que sera extraite la table de traduction de séquences.
Proposés en 1993, les modèles IBM ont révolutionné la recherche en Traduction Statistique et constituent depuis une véritable réfèrence. Toutefois, ces modèles, et notamment les modèles d'ordre supérieur à 3, présentent certains désavantages : ils sont très complexes et leur apprentissage nécessite une puissance de calcul importante et un temps non négligeable (plus d'une dizaine d'heures pour des corpus d'apprentissage de plus de 500K paires de traduction). La TA statistique est un domaine de recherche encore trop peu exploré relativement à d'autres activités de recherche. Aujourd'hui, beaucoup de travaux consistent en l'optimisation des nombreux paramètres caractérisant les modèles de traduction existant. Pour notre part, nous ne sommes pas satifaits des résultats obtenus à ce jour avec les modèles actuels et plutôt que de partir de modèles existant et de tenter d'en améliorer les performances, nous avons choisi dans nos travaux de proposer une idée tout à fait originale pour l'apprentissage de modèles de traduction de mots et de séquences de façon rapide et simple sans utiliser les modèles IBM et en se basant sur le concept de triggers inter-langues.
Nous commencerons dans le chapitre 4 par donner une définition du concept de trigger inter-langue avant de l'exploiter pour établir des correspondances lexicales au niveau du mot. Nous exploiterons ces correspondaces pour établir un dictionnaire bilingue Français-Anglais. Le chapitre 5 s'inscrit dans le cadre de la TA statistique par mot. Nous y proposerons des méthodes basées sur les triggers inter-langues et destinées à estimer des tables de traductions de mots. Dans le chapitre 6, nous nous placerons cette fois dans un contexte de TA statistique par groupe de mots. Nous verrons qu'il est possible d'étendre le concept de triggers inter-langues afin d'établir des correspondances lexicales au niveau du groupe de mots. Nous exploiterons ces correspondances pour estimer des tables de traduction de séquences.

Deuxième partie

Contributions à la Traduction Automatique Statistique

Chapitre 3

Les corpus

3.1 Introduction

L'apprentissage et la mise au point des modèles de traduction statistique requièrent une quantité importante de corpus parallèles alignés, que ce soit au niveau du paragraphe ou de la phrase ou du mot. L'alignement de ces corpus indique les parties source et cible qui sont des traductions l'une de l'autre.
Dans ce chapitre, nous décrivons les deux corpus que nous avons utilisés pour mettre en place nos modèles de traduction et pour les évaluer. Nous commençons par les corpus extraits des actes du Parlement Européen. Nous avons fait le choix d'utiliser ces corpus dans le but de situer nos travaux en Traduction Automatique Statistique par rapport aux travaux existants. En effet, ceux-ci sont très fréquemment cités dans la littérature et ils constituent aussi les données d'apprentissage et de développement fournies lors des campagnes d'évaluation WMT pour estimer des modèles de traduction. Nous présentons ensuite des corpus originaux que nous avons construits automatiquement à partir de fichiers de sous-titrage de films. Le recours aux sous-titres comme données textuelles pour la mise en place des modèles de traduction constitue une des originalités de nos contributions en Traduction Statistique. Contrairement aux actes du Parlement Européen, les sous-titres se rapprochent davantage de la parole spontanée. Ils constituent également un matériel très intéressant de par la richesse et la grande diversité du vocabulaire composé aussi bien de mots courants que de mots familiers et même vulgaires. En utilisant de telles données, nous espérons tendre vers un meilleur système de traduction automatique de la parole spontanée. En outre, une application directe pourrait être un outil de sous-titrages automatique temps réel de films dans une langue cible d'après la bande-son dans une langue source.

3.2 Le corpus du Parlement Européen : EUROPARL

Les corpus extraits des actes du Parlement Européen (PE) font partie des plus fréquemment utilisés en traduction automatique statistique. Ces corpus sont la transcription, en plusieurs langues, de discours des différents membres du PE [Koehn 05]. Le tableau 3.1 rassemble les données d'apprentissage, de développement et de test employées pour les expériences présentées dans les chapitres suivants. Le langage employé au cours des sessions du Parlement Européen est très formel et ne rappelle guère la parole spontanée. Les discours sont par ailleurs ciblés sur des sujets bien spécifiques tels que l'environnement ou les affaires étrangères et concernent donc un domaine assez limité. Afin de toucher une plus large communauté, nous avons décidé d'utiliser également des corpus qui se rapprocheraient davantage de la parole spontanée. Nous

		Français	**Anglais**
App.	Phrases	$596K$	
	Mots	$17,3M$	$15,8M$
	Singletons	$26,6K$	$22,2K$
	Vocabulaire	$77,5K$	$60,3K$
Dev.	Phrases	1444	
	Mots	$15,0K$	$14,0K$
Test	Phrases	500	
	Mots	$4,9K$	$5,2K$

TAB. 3.1 – Données d'apprentissage (App.), de développement (Dev.) et de test (Test) relatives aux corpus extraits des actes du Parlement Européen

nous sommes pour cela intéressés aux sous-titres anglais et français de films disponibles en grande quantité sur Internet.

3.3 Le corpus de sous-titres

Le sous-titrage est une technique cinématographique permettant l'affichage de texte en bas de l'écran lors de la diffusion d'un film. Ce texte peut être une transcription ou une traduction des paroles des acteurs ou encore une aide aux personnes malentendantes. Plusieurs sites Internet proposent des fichiers destinés à intégrer des sous-titres dans une vidéo au format divx. Ces sites regorgent de sous-titres multilingues de films célèbres ou de séries télévisées. Ces sous-titrages sont proposés par des internautes qui ont retranscrit un film ou qui proposent une traduction de la version originale. Nous avons donc pensé récupérer ces fichiers pour constituer notre corpus parallèle. Toutefois, l'état brut de ces fichiers ne nous permet pas de les utiliser en tant que corpus bilingues pour la Traduction Automatique Statistique. En effet, les corpus bilingues requièrent un alignement au niveau de la phrase, du paragraphe ou même du groupe de mots. Il est nécessaire de savoir quelle partie du corpus en langue source est la traduction de quelle partie du corpus en langue cible.
Nous présentons dans la première section les difficultés que nous avons rencontrées pour aligner ces fichiers avant de proposer une méthode d'alignement automatique. Notre but étant de trouver non pas un alignement parfait qui nécessiterait des interventions manuelles mais le meilleur alignement possible dans de bref délai et le plus simplement possible. Nous terminerons avec l'évaluation des alignements automatiques générés à partir de nos fichiers de sous-titrage.

3.3.1 Données brutes : description et problèmes

Nous avons collecté à partir d'Internet une soixantaine de sous-titrages de films en anglais et en français. Nous présentons dans cette section le format brut des sous-titres et montrons les phénomènes qui nous empêchent de les utiliser tels quels pour l'apprentissage de modèles de traduction.

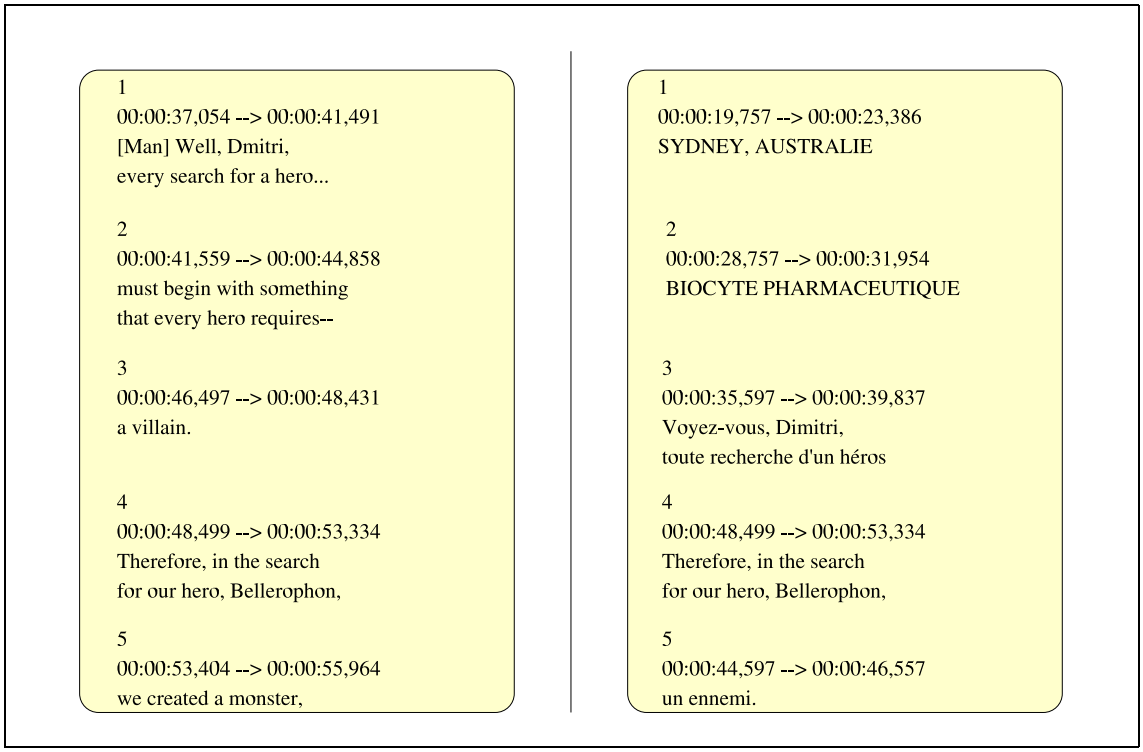

FIG. 3.1 – Extrait de fichiers de sous titres de film

3.3.1.1 Format des fichiers

Un fichier est découpé en un ensemble de sous-titres. Un sous-titre est un texte qui va s'afficher en bas de l'écran lors de la diffusion d'un film. Il décrit une scène, ou retranscrit les paroles des acteurs. La taille d'un sous-titre est limité par la taille de l'écran. Ainsi, un sous-titre peut être une phrase, une partie d'une phrase ou un groupe de phrases. La figure 3.1 présente un extrait des transcriptions anglaise et française du film *Mission Impossible 2*. Un sous-titre est représenté par un identifiant, par une trame de temps indiquant quand le sous-titre apparaît sur l'écran et quand il disparaît et enfin par la transcription des paroles ou la description de la scène. Premier constat qu'il est possible de faire à partir de notre exemple : les sous-titres ne sont pas alignés. Le sous-titre 1 de la version anglaise n'est pas la traduction du sous-titre 1 de la version française mais celle du sous-titre 3. Le deuxième constat direct est que les trames de temps des sous-titres ne correspondent pas non plus. Ces décalages se produisent très fréquemment au sein des fichiers et les causes sont multiples. Nous détaillons dans la section suivante tous les cas de figure causant un décalage dans les alignements.

3.3.1.2 Les problèmes d'alignement

– *Insertion de descriptions de scènes*
Comme nous le montre la figure 3.1, dans certaines versions, des descriptions de scène sont présentes alors que dans d'autres non. Ainsi, les deux premiers sous-titres français situent la scène géographiquement avant de transcrire les dialogues. Cette description est absente dans la version anglaise ce qui provoque un décalage de deux sous-titres. Ceci s'explique par le fait que deux scripts d'un même film ne sont pas forcément déposés par la même

personne. Une personne peut décider d'introduire des descriptions, alors qu'une autre peut décider de les supprimer. Nous avons observé par ailleurs que le texte qui n'est pas la transcription de ce que disent les acteurs est noté soit totalement en majuscules, soit entre crochets ou encore entre #. Dans la figure 3.2, les sous-titres anglais numérotés 11 à 15 sont de la description que nous ne retrouvons pas dans le script français. Les sous-titres qui étaient déjà décalés au niveau de l'identifiant d'une version à l'autre le sont alors encore plus.
Une solution simple à ce problème a été dans un premier temps de débarrasser les scripts de ces descriptions facilement identifiables par les crochets, les dièses ou les majuscules. Nous allons voir que cela n'a pas suffit à recaler les sous-titres. En effet, en plus de l'introduction de description dans le sous-titrage, nous sommes également confrontés à une différence de segmentation au niveau des phrases mais aussi à l'ajout de portions de dialogue.

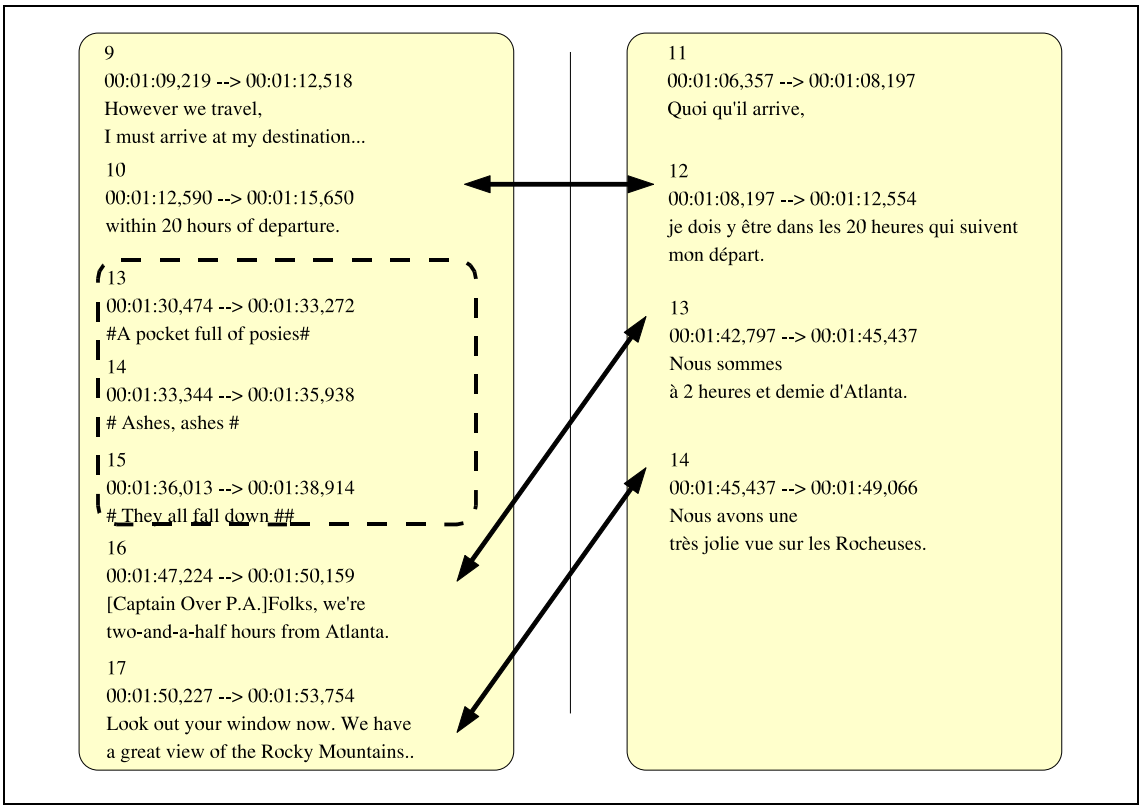

FIG. 3.2 – Exemple d'introduction de description

– *Différence de segmentation*
Ce que nous appelons ici segmentation d'une phrase est sa répartition sur un ou plusieurs sous-titres. Par exemple dans la figure 3.3, la phrase *However we travel, I must arrive at my destination within 20 hours of departure* est segmenté en deux sous-titres. Son équivalence en français *Quoi qu'il arrive, je dois y être dans les 20 heures qui suivent mon départ* est également répartie sur deux sous-titres mais la segmentation ne se fait pas au même endroit. Une meilleure solution ici serait d'aligner le sous-titre anglais 9 avec les sous-titres français 11 et 12. En effet, *However we travel, I must arrive at my destination* se traduit par *Quoiqu'il arrive, je dois y être* et *within 20 hours of departure* est la traduction de

dans les 20 heures qui suivent mon départ. L'idéal ici au niveau de l'alignement serait de concaténer les segments deux à deux pour pouvoir caler la phrase entière anglaise avec son équivalent français. Nous verrons par la suite comment nous avons géré ce problème. Le dernier problème auquel nous avons été confronté est l'insertion ou l'omission de parole.

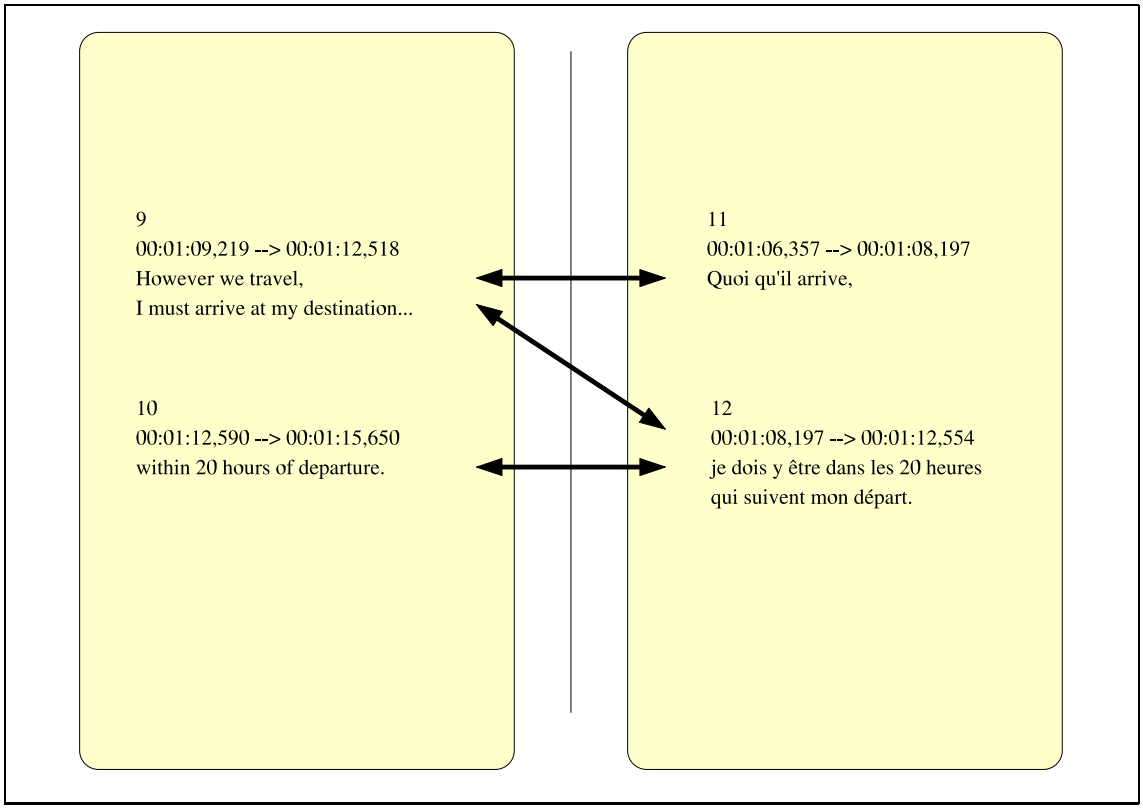

FIG. 3.3 – Exemple de différence de segmentation

– *Insertion/Omission de sous-titres*
En plus d'une segmentation différente des phrases dans les scripts s'ajoute le problème des sous-titres de transcription de parole qui sont présents dans une version et absents dans l'autre. Autant il est simple d'identifier automatiquement une description et de la supprimer, autant il est compliqué de juger automatiquement si une phrase émise par un acteur a été transcrite en anglais et en français. Dans l'extrait présenté dans la figure 3.4, nous pouvons distinguer deux types d'insertion. Dans le sous-titre anglais 17 aligné avec le sous-titre français 14 , le segment *Look at your window* a été ajouté. Il ne constitue cependant pas la totalité du sous-titre. Dans ce cas, pour un alignement correct, la solution serait, soit de supprimer les sous-titrages équivalents 17 et 14 avec la perte de données que cela représente, soit d'aligner cette paire en y laissant ce bruit. Supprimer automatiquement la partie erronée paraît difficile dans un tel cas de figure. Le sous-titre anglais 19 est lui entièrement rajouté par rapport à la version française. Il n'est absolument pas retranscrit dans le script français. Il est nécessaire de le supprimer entièrement puisqu'il n'est en équivalence avec aucun autre sous-titre. La question qui se pose alors est comment juger automatiquement qu'un sous-titre a ou non sa traduction dans l'autre version. Dans la section suivante, nous présentons comment nous avons résolu ce problème.

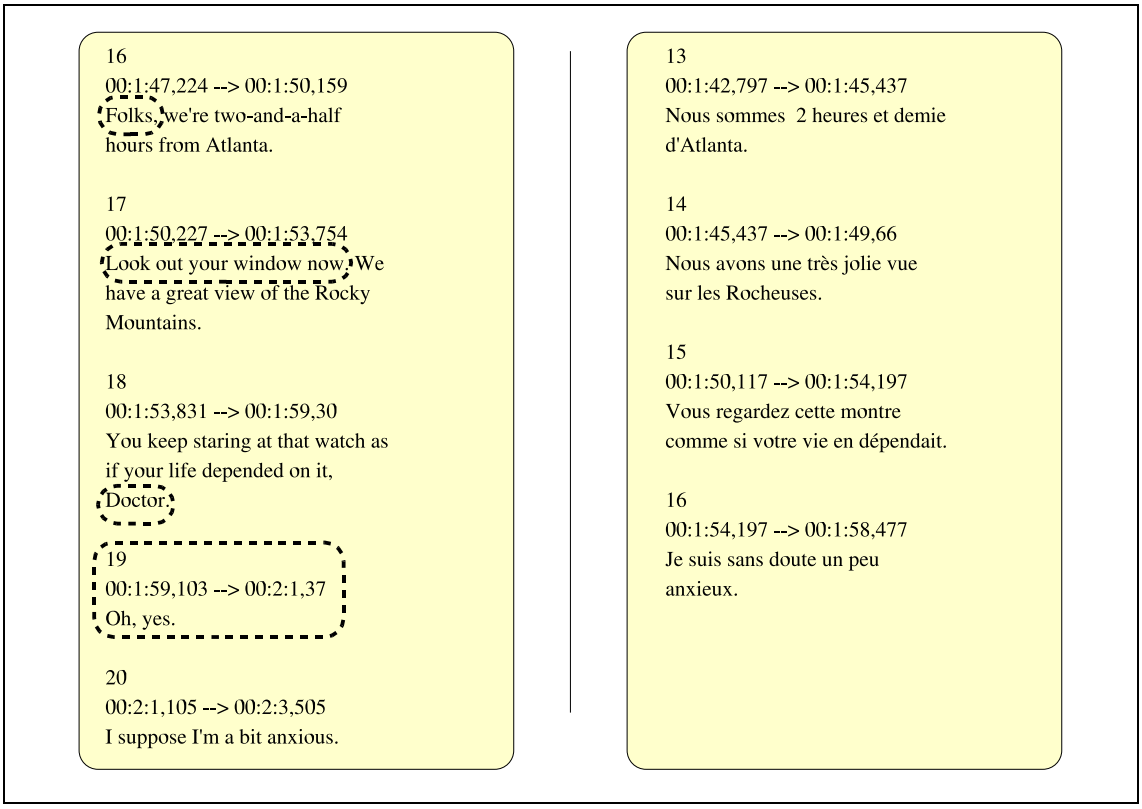

FIG. 3.4 – Exemple d'insertion de texte

A défaut de pouvoir se baser sur l'identifiant des sous-titres, nous avons pensé pouvoir les aligner en se référant à leurs temps de début et temps de fin. Malheureusement, il est apparu impossible de se fier à ces deux attributs pour aligner nos sous-titres. En effet, comme le montre la figure 3.1, de la même façon qu'il se produit un décalage dans le nombre de sous-titres, il se produit également un décalage temporel qui s'accroît de plus en plus à mesure que le film arrive à son terme. La figure 3.5 est un extrait de la fin du film *Mission Impossible 2*, au niveau de la trame de temps, il se produit un décalage de plus d'une minute trente entre le sous-titre anglais 626 et son correspondant français.
Ce décalage temporel est variable au cours du temps, il s'accroît, diminue, s'accroît de nouveau. Il est difficile d'en déduire une quelconque règle pour aligner correctement et automatiquement les sous-titres, même si dans certains travaux, les auteurs utilisent justement les trames de temps pour aligner les sous-titres [Mangeot 05]. L'identifiant et la trame de temps des sous-titres ne nous permettent pas de les aligner de façon automatique sans intervenir manuellement. Le seul point de référence auquel nous pouvons nous fier est le texte du sous-titre.

Aligner manuellement toutes les paires de fichiers anglais et français est possible mais inenvisageable du point de vue de la quantité de données à traiter et de la perte de temps considérable que cela engendre. Nous proposons donc dans la section suivante une méthode automatique d'alignement des fichiers de sous-titres basée sur l'algorithme de Viterbi.

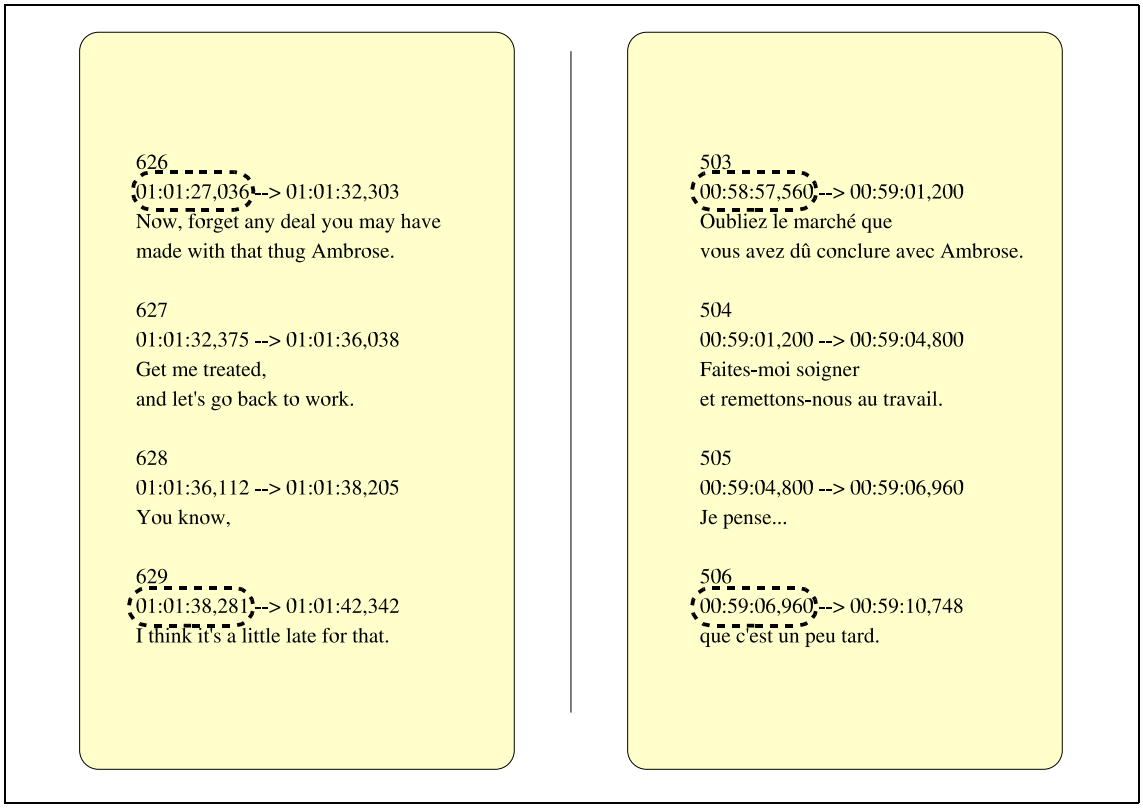

FIG. 3.5 – Décalage temporelle

3.3.2 Solution : programmation dynamique

Plusieurs solutions ont déjà été envisagées pour résoudre le problème d'alignement de corpus parallèles [Moore 02, Brown 91, Melamed 96, Vandeghinste 04, Gale 91b] (nous invitons le lecteur à se référer à l'article [Singh 05] pour une étude comparative de ces méthodes). La plupart des travaux sont basés sur la programmation dynamique. Ils utilisent pour cela une fonction objectif qui permet de déterminer la distance existante entre les différents segments des corpus parallèles, un segment pouvant être un paragraphe, une phrase ou une suite de mots.

3.3.2.1 Optimisation de l'alignement

Le problème consistant à aligner deux fichiers de sous-titrage d'un même film est un problème classique de programmation dynamique. Etant donné tous les sous-titres anglais et français d'un film, l'objectif est de trouver le meilleur alignement possible entre eux.

Nous allons pour cela considérer tous les alignements possibles et trouver quel est le meilleur selon un critère d'optimisation qui sera dans notre cas la Fmesure. Nous allons utiliser cette mesure pour trouver le meilleur chemin allant du premier sous-titre de chacune des deux version de sous-titrage jusqu'au dernier.

La figure 3.6 modélise le problème. Chaque nœud (e_i, f_j) représente un alignement potentiel entre le sous-titre anglais e_i et le sous-titre français f_j. Un chemin valide commence au nœud (e_0, f_0) et se termine au nœud (e_E, f_F), E et F étant le nombre respectif de sous-titres dans la version anglaise et française du film considéré. Les arcs représentent les transitions entre alignements. A partir d'un nœud, trois transitions sont possibles :

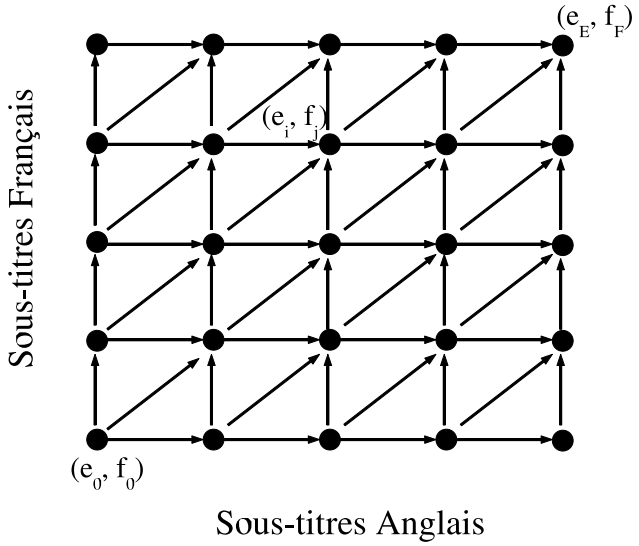

FIG. 3.6 – Alignement dynamique entre les sous-titres anglais e_i et les sous-titres français f_j d'un même film

- transition verticale de (e_i, f_j) à (e_i, f_{j+1}). Le sous-titre e_i est aligné avec deux sous-titres consécutifs f_j et f_{j+1}. La progression se fait seulement dans le fichier français. (cf. figure 3.3 où le sous-titre anglais 9 est aligné avec les sous-titres français 11 et 12)
- transition diagonale de (e_i, f_j) à (e_{i+1}, f_{j+1}). La progression se fait dans les deux fichiers. (cf. figure 3.2 où les sous-titres anglais 16 et 17 sont respectivement alignés avec les sous-titres français 13 et 14))
- transition horizontale de (e_i, f_j) à (e_{i+1}, f_j). Les sous-titres consécutifs e_i et e_{i+1} sont alignés avec le même sous-titre f_j. La progression se fait uniquement dans le fichier anglais.

Les transitions verticale et horizontale nous permettent de gérer les concaténations de sous-titres et les suppressions de sous-titres. Une transition verticale signifie soit qu'un même sous-titre anglais est aligné avec deux sous-titres français, soit qu'il y a un sous-titre français de plus qui n'est pas rapporté en anglais. De la même façon, une transition horizontale indique soit que deux sous-titres anglais sont alignés avec le même sous-titre français, soit qu'un des deux sous-titres anglais n'est pas présent dans la version française du film. Nous verrons comment décider si de telles transitions sont le cas de concaténation de sous-titres ou au contraire le cas d'insertion ou d'omission.

Pour chaque nœud (e_i, f_j), nous définissons un score $S(e_i, f_j)$ de correspondance basé sur la Fmesure (F_M) et sur le score du nœud précédent et calculé comme suit :

$$S(e_i, f_j) = max \begin{cases} S(e_i, f_{j-1}) + \beta_{F_m}(F_M(e_i, f_j) + \epsilon) & (\uparrow) \\ S(e_{i-1}, f_{j-1}) + \alpha_{F_m}(F_M(e_i, f_j) + \epsilon) & (\nearrow) \\ S(e_{i-1}, f_j) + \lambda_{F_m}(F_M(e_i, f_j) + \epsilon) & (\rightarrow) \end{cases}$$

α_{F_m}, β_{F_m} et λ_{F_m} sont des paramètres optimisés dans le but d'obtenir les meilleures performances d'alignement sur un corpus de test. Ils accordent un poids différent à chacune des transitions

possibles. La valeur de ces coefficients dépend de la valeur de la Fmesure F_M. Notons également que pour prévenir des problèmes liés à des scores nuls, nous utilisons une constante de lissage ϵ. La Fmesure est donc utilisée comme mesure de distance entre deux sous-titres. Il s'agit d'une moyenne harmonique entre la précision (P) et le rappel (R) :

$$F_M(e_i, f_j) = 2 \times \frac{R(e_i, f_j) \times P(e_i, f_j)}{R(e_i, f_j) + P(e_i, f_j)} \tag{3.1}$$

Si la Fmesure entre deux sous-titres est supérieure ou égale à $0,5$ alors nous pouvons considérer que plus de la moitié du sous-titre anglais est couvert par le sous-titre français et que par conséquent l'un a de fortes chances de correspondre à l'autre.
Le rappel et la précision sont calculés en comparant les mots du sous-titre anglais aux traductions des mots du sous-titre français.

$$R(e_i, f_j) = \frac{match(e_i, tr(f_j))}{N(e_i)} \qquad P(e_i, f_j) = \frac{match(e_i, tr(f_j))}{N(f_j)} \tag{3.2}$$

$$match(e_i, tr(f_j)) = \sum_{k=1}^{N(e_i)} \delta(e_i^k, tr(f_j^l)) \forall l \in [1 - N(f_j)] \tag{3.3}$$

$tr(f_j)$ est la traduction mot-à-mot en anglais du sous-titre français f_j. Cette traduction est obtenue à l'aide d'un dictionnaire Français-Anglais. $N(x)$ est le nombre de mots contenus dans le sous-titre x. $match(e_i, tr(f_j))$ est le nombre de mots identiques entre le sous-titre e_i et la traduction mot-à-mot du sous-titre f_j. Enfin $\delta(x, y)$ est la fonction de Kronecker valant 1 si x et y sont identiques et 0 sinon. Un exemple de calcul de la fonction $match()$ est donné dans la figure 3.7. Pour chaque mot du sous-titre français, un dictionnaire bilingue fournit les traductions

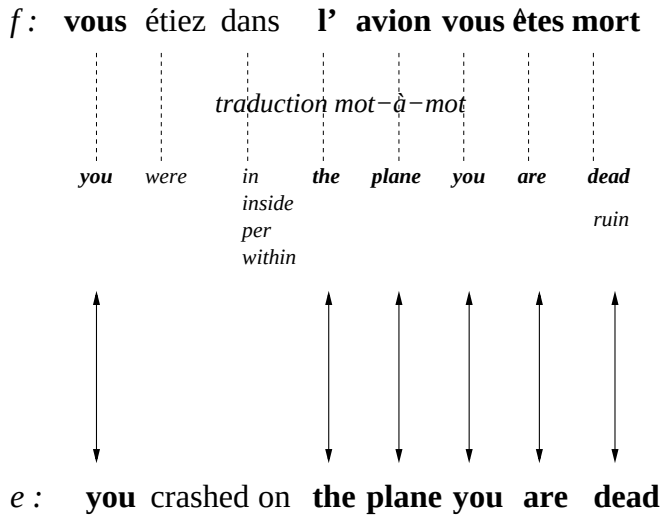

FIG. 3.7 – Exemple de calcul de la fonction match()

possibles. Si une des traductions est présente dans le sous-titre anglais, la fonction $match()$ est incrémentée. Dans notre exemple, la fonction renvoie la valeur 6 puisque 6 mots du sous-titre anglais se retrouvent dans les traductions possibles des mots du sous-titre français. Afin que la mise en correspondance des sous-titres soit plus précise, la fonction $match()$ est également

incrémentée lorsqu'une même forme orthographique apparaît dans les deux sous-titres français et anglais. Cela permet ainsi de faire correspondre les noms propres sans qu'ils ne soient présents dans le dictionnaire bilingue utilisé.
Les scores $S(e_i, f_j)$ sont calculés ainsi de la gauche vers la droite. Une fois tous les nœuds évalués, et partant du meilleur score associé au dernier sous-titre anglais, le meilleur alignement est obtenu par retour arrière jusqu'au premier sous-titre. Nous présentons, dans la section suivante, les expériences menées avec cette méthode d'alignement automatique.

3.3.3 Evaluation de l'alignement automatique

3.3.3.1 Les données de test

Pour tester notre méthode d'alignement, nous l'avons dans un premier temps appliquée sur une quarantaine de films, soient 44318 sous-titres français et 46003 sous-titres anglais. L'alignement automatique obtenu est ensuite comparé à un alignement de référence construit manuellement. La comparaison ne s'est faite que sur un corpus extrait des 40 films. Etablir un alignement manuel de tous les films dans leur intégralité aurait demandé beaucoup trop de temps. Un film représentant en moyenne un millier de sous-titres, nous en avons extrait aléatoirement 35 dans la version anglaise ainsi que leur correspondant dans la version française pour chacun des 40 films. Ceci nous amène à un total de 1353 sous-titres anglais (T_E) et 1334 sous-titres français (T_F) que nous avons aligné manuellement pour obtenir finalement un corpus de référence de 1364 (#A) paires de sous-titres.

3.3.3.2 Protocole d'évaluation

Pour l'évaluation de notre méthode d'alignement automatique, nous avons suivi la procédure suivante :

1. Supprimer des ensembles de sous-titres T_E et T_F les sous-titres de description de scènes.
2. Aligner automatiquement T_E et T_F. Pour le calcul de la fonction $match()$ impliquée dans le processus d'alignement, nous avons utilisé un dictionnaire Français-Anglais extrait du projet XDXF[9] contenant 41398 entrées[10].
3. Supprimer de l'alignement obtenu les paires dont la Fmesure est nulle. La Fmesure est nulle si aucun mot du sous-titres anglais n'a de correspondant dans le sous-titres français ($match() = 0$). Nous supposons alors que si aucune correspondance lexicale n'a été trouvée au sein de la paire de sous-titres alors ils ne sont pas des traductions l'un de l'autre.
4. Comparer l'alignement automatique à l'alignement de référence en terme de précision et de rappel.

3.3.3.3 Choix des paramètres

Un premier test a été mené afin d'étudier l'effet du paramètre α_{F_m} qui est le poids accordé à la transition diagonale (progression dans les deux fichiers de sous-titres) dans la recherche du meilleur alignement. Nous estimons que si la Fmesure n'est pas nulle, nous devons alors privilégier la transition diagonale. En effet, une Fmesure positive entre deux sous-titres signifie qu'il existe au moins un lien entre eux et nous supposons, dans ce cas, qu'ils doivent être alignés ensembles et avec aucun autre sous-titre. Si au contraire la Fmesure est nulle entre deux sous-titres, nous

[9] http ://xdxf.revdanica.com
[10] Nom de fichier :comn_sdict05_French-English.tar.bz2

ne privilégions aucune transition. Pour cette première expérience, nous faisons donc varier α_{F_m} de 1 (la transition diagonale n'est pas favorisée) à 100 lorsque la Fmesure entre deux sous-titres est positive. La valeur des paramètres β_{F_m} et λ_{F_m} est fixée à 1. Le tableau 3.2 présente les taux de rappel (Rap.), précision (Prec.) et Fmesure (Fm.) de l'alignement automatique. $\#Tot.$ est le

α_{F_m}	$\#C$	$\#I$	$\#Tot.$	Rap.	Prec.	Fm.
1	1063	842	1905	0,779	0,558	0,650
2	1124	213	1337	0,824	0,841	0,832
3	1124	114	1238	0,824	0,908	0,864
4	1121	99	1220	0,822	0,919	0,868
5	1121	98	1219	0,822	0,920	0,868
6	1120	97	1217	0,821	0,920	0,868
7	1119	97	1216	0,820	0,920	0,867
8	1118	96	1214	0,820	0,921	0,867
9	1119	94	1213	0,820	0,923	0,868
10	1118	94	1212	0,820	0,922	0,868
20	1116	93	1209	0,818	0,923	0,867
100	1114	92	1206	0,817	0,923	0,867

TAB. 3.2 – Performance en fonction du paramètre α_{F_m} lorsque $F_m > 0$

nombre de paires alignées automatiquement. $\#C$ indique le nombre de paires correctement alignées (c'est-à-dire présentes dans le corpus de référence) de l'alignement automatique. $\#I$ indique à l'inverse le nombre de paires incorrectes (les paires non présentes dans le corpus de référence). La précision (Prec.) représente le taux de paires correctes dans l'alignement automatique.

$$Precision = \frac{\#C}{\#Tot.} \tag{3.4}$$

Le rappel représente la proportion de paires de sous-titres du corpus de référence retrouvées automatiquement, $\#A$ étant le nombre de paires de sous-titres dans le corpus de référence.

$$Rappel = \frac{\#C}{\#A} \tag{3.5}$$

La Fmesure (Fm.) est une moyenne harmonique du rappel et de la précision.

$$Fmesure = (2 \times Rappel \times Precision)/(Rappel + Precision) \tag{3.6}$$

Les résultats montrent que le paramètre α_{F_m} a un impact important sur les performances. La Fmesure augmente en effet avec la valeur de α_{F_m} et ce jusqu'à une valeur de 7. Au delà, elle devient instable. Maximiser la Fmesure revient à maximiser à la fois le rappel et la précision. Or, rappelons que notre objectif est de retrouver le plus d'alignements corrects possibles sans ajouter pour autant trop de bruit. Le tableau 3.2 indique que cet objectif est atteint lorsque la précision est optimale et non pas la Fmesure. En fait, lorsque la précision augmente, le nombre de faux-positifs[11] diminue. Considérant ceci, nous fixons la valeur de α_{F_m} à 9 dans les expériences suivantes. Cette valeur conduit à seulement 94 paires erronées mais à un rappel de 82% et à une

[11] paire de sous-titres de l'alignement automatique incorrecte (absente du corpus de référence)

précision maximale de $92,3\%$.

En observant les paires de sous-titres incorrectes de l'alignement automatique, nous avons constaté que la plupart des sous-titres qui ne devraient pas être alignés ont une Fmesure trop élevée compte tenu du fait qu'ils ne sont pas la traduction l'un de l'autre. Une des raisons est que le poids accordé aux mots outils dans la fonction *match*() est le même que le poids accordé aux mots porteurs de sens de la phrase. Ce phénomène peut s'avérer très problématique, notamment lorsqu'il s'agit d'établir les correspondances entre sous-titres courts, comme l'illustre le tableau 3.3.

	E1 : Wallis **hold** on to this F1 : Wallace **tiens** moi cela	E1 : Wallis hold on **to** this F2 : Ulrich pense **à**
N(e)	5	5
N(f)	4	3
match	1	1
Prec.	1/4	1/3
Rec.	1/5	1/5
Fm.	0,22	0,23

TAB. 3.3 – Exemple d'une mauvaise correspondance due aux mots outils

Deux paires de sous-titres peuvent avoir la même valeur de Fmesure si leurs constituants ont la même longueur et le nombre de liens entre les mots des sous-titres au sein d'une paire est le même pour les deux paires. L'alignement $(E1, F2)$ est erroné mais sa Fmesure est supérieure à celle de l'alignement $(E1, F1)$ qui lui est correct. Pour chacun des alignements, un seul lien est trouvé (tiens, hold) pour le premier et (to, à) pour le second. La différence de Fmesure s'explique par la différence de taille des sous-titres français. Le fait d'accorder le même poids à tous les liens introduit des erreurs. Il paraît évident dans cet exemple que la correspondance entre *tiens* et *hold* est plus importante que celle entre *to* et *à*. Entre outre, nous pouvons remarquer que le nom propre *Wallace (Wallis)* n'appartient pas au dictionnaire. Une meilleure couverture de la part du dictionnaire incluant ce nom propre aurait permis d'obtenir une Fmesure de $0,44$ pour la paire $E1, F1$ et donc de privilégier cet alignement plutôt que le deuxième.

Pour réduire l'impact des mots outils, nous avons modifié la formule 3.3 en introduisant un facteur de pondération γ :

$$match(e_i, tr(f_j)) = \sum_{k=1}^{N(e_i)} \gamma \times \delta(e_k^i, tr(f_l^j)) \forall l \in [1 - N(f_j)] \tag{3.7}$$

Nous attribuons au facteur γ une valeur inférieure à 1 lorsque les mots e_k^i ou f_l^j sont des mots outils et une valeur de 1 sinon. Tous les déterminants, pronoms, prépositions, verbes courts et courants comme les auxiliaires ont été considérés comme mots outils dans nos expériences.
Le fait d'accorder moins d'importance aux mots outils dans le calcul de la fonction *match*() n'a malheureusement pas d'impact positif sur les performances comme l'indique le tableau 3.4. En effet, le fait de faire varier le poids des mots outils n'améliore en rien l'alignement automatique du point de vue de la Fmesure. La perte en précision et en rappel est significative au fur et à mesure que la valeur du facteur γ diminue. Nous expliquons ce phénomène par la faible taille

γ	$\#C$	$\#I$	$\#Tot$	Rec.	Prec.	Fm.
1,0	1119	94	1213	0,820	0,923	0,868
0,9	1097	134	1231	0,804	0,891	0,845
0,8	1097	134	1231	0,804	0,891	0,845
0,7	1097	134	1231	0,804	0,891	0,845
0,6	1097	133	1230	0,804	0,892	0,846
0,5	1097	133	1230	0,804	0,892	0,846
0,4	1056	171	1227	0,774	0,861	0,815
0,3	1044	189	1233	0,765	0,847	0,804
0,2	1040	192	1232	0,762	0,844	0,801
0,1	1039	194	1233	0,762	0,843	0,800
0,0	869	55	951	0,657	0,942	0,774

TAB. 3.4 – Impact du poids accordé aux mots outils dans le calcul de la fonction *match*()

des sous-titres. En moyenne, ceux-ci sont composés de 7 à 10 mots dont plusieurs sont des mots outils. Ne pas les prendre en compte rend la mise en correspondance des sous-titres plus difficile. Une étude des paires de sous-titres proposées dans l'alignement automatique ayant la plus forte précision (avec $\alpha_{F_m} = 9$ et $\gamma = 1$, soit la première ligne du tableau 3.4), a révélé que sur les 1119 paires comptabilisées comme correctes, 182 reposaient uniquement sur la mise en correspondance de mots outils. Par conséquent, en minimisant leur importance dans la mise en correspondance des sous-titres, nous perdons une quantité non négligeable de bonnes paires de sous-titres. Ceci pourrait également expliquer la dernière ligne du tableau. Lorsque nous ignorons les mots outils ($\gamma = 0$), nous constatons en effet que le nombre de paires proposées par notre algorithme d'alignement diminue considérablement, seulement 951 paires contre plus de 1200 quand ils sont pris en compte. Rappelons que dans notre procédure, nous supprimons les paires de sous-titres (e_i, f_j) pour lesquelles la Fmesure est nulle. C'est pourquoi toutes les paires de sous-titres reposant uniquement sur des correspondances de mots outils sont supprimées de l'alignement lorsque ceux-ci ont un poids nul dans le calcul de la fonction *match*(). C'est le cas ici de 289 paires.

3.3.4 Discussion

Travailler sur des corpus parallèles de sous-titres semble être selon nous un bon moyen de tendre vers des applications de traduction automatique de la parole spontanée. En effet, les sous-titres de films sont constitués d'une multitude d'expressions courantes, d'hésitations, d'expressions familières qui se rapprochent davantage de la parole spontanée que les textes du Parlement Européen par exemple. Nous avons proposé une méthode d'alignement automatique de sous-titres et les résultats obtenus nous permettent de l'utiliser pour constituer des corpus parallèles alignés nécessaires à l'apprentissage de modèle de traduction. En effet, au vu des évaluations faites, notre méthode permet d'aligner automatiquement les sous-titres avec une précision de $92,3\%$. Ces résultats font de notre approche une méthode concurrentielle aux méthodes état-de-l'art d'alignement de corpus bruités [Singh 05]. Les tests ont été réalisés sur une quarantaine de films seulement. Beaucoup d'autres sont à disposition sur Internet. Il est donc possible de constituer un large corpus d'apprentissage pour un modèle de traduction.

Pour constituer des corpus nécessaires à la mise en place de notre système de traduction, nous avons retenu les sous-titrages des 36 films pour lesquels la précision de l' alignement était

la meilleure. Ainsi le tableau 3.5 rassemble les données d'apprentissage, de développement et de test employées pour les expériences présentées dans les chapitres suivants. Nous pouvons noter

		Français	**Anglais**
App.	Phrases	$27,5K$	
	Mots	$0,19M$	$0,20M$
	Singletons	$7,0K$	$5,4K$
	Vocabulaire	$14,6K$	$11,7K$
Dev.	Phrases	1959	
	Mots	$13,6K$	$14,7K$
Test	Phrases	250	
	Mots	$2,02K$	$2,05K$

Tab. 3.5 – Données d'apprentissage (App.), de développement (Dev.) et de test (Test) relatives aux corpus de sous-titres (SSTITRES)

que 45% des mots du corpus d'apprentissage (de la partie française ou de la partie anglaise) apparaissent seulement une fois. Par ailleurs, $13,8\%$ des mots anglais du corpus de test sont des mots hors vocabulaire. Tous ces éléments montrent la difficulté de ce corpus et expliquent les faibles performances que nous allons décrire en comparaison de celles obtenues sur le corpus des actes du Parlement Européen.

Les corpus présentés dans ce chapitre seront référencés dans la suite de ce manuscrit en tant que EUROPARL pour le corpus extrait des actes du Parlement Européen et SSTITRES pour le corpus construit automatiquement à partir des sous-titres de films.

Nous venons de décrire les données que nous avons à disposition pour la mise en place d'un système de traduction automatique statistique. Ces données vont nous permettre l'apprentissage et l'optimisation de modèles de traduction ainsi que l'évaluation de notre système. Dans les chapitres suivants, nous proposons et testons une nouvelle approche pour l'apprentissage des tables de traduction basée sur le concept de triggers inter-langues.

Chapitre 4

Le concept de triggers inter-langues

4.1 Introduction

Dans ce chapitre, nous présentons le concept de triggers inter-langues. Nous commençons par en donner une définition avant de poursuivre sur une étude préalable de leur potentiel dans un cadre de Traduction Automatique. Nous verrons notamment comment tirer profit de l'information apportée par les triggers inter-langues pour construire un lexique bilingue Français-Anglais.

4.2 Les triggers inter-langues : définition

Comme nous l'avons évoqué dans la section 2.3, le concept de triggers est très souvent cité en modélisation statistique du langage et plus particulièrement en reconnaissance de la parole. Un trigger est un ensemble composé d'un mot *déclencheur* et d'une liste de mots qu'il déclenche appelés *déclenchés*. Les triggers permettent entre autre d'améliorer et de généraliser le modèle Cache [Kuhn 90]. En effet, alors que le modèle Cache favorise la probabilité d'un mot w_i récemment apparu dans le contexte gauche, un modèle de triggers va plus loin et accorde une probabilité plus importante à une liste de mots corrélés au mot w_i [Lau 93, Tillmann 96].

Les triggers sont sélectionnés selon la valeur de l'Information Mutuelle (IM) qui permet d'établir la corrélation existante entre deux mots x et y. L'IM entre x et y est définie par :

$$IM(x,y) = P(x,y) log \frac{P(x,y)}{P(x)P(y)} \tag{4.1}$$

où $P(x,y)$ est la probabilité jointe d'apparition des mots x et y et $P(x)$ (respectivement $P(y)$) est la probabilité d'apparition de x (respectivement y). Seuls les couples (x,y) ayant une IM supérieure à un seuil fixé sont sélectionnés comme étant des triggers. Ainsi dans un modèle de language de type Trigger, chaque mot appartenant à un vocabulaire défini est associé à une liste de n mots qu'il déclenche et cette association constitue un trigger. Par abus de langage, nous appellerons par la suite *trigger* un couple de mots (x,y) tel que x déclenche y. La figure 4.1 illustre un exemple de triggers anglais. La présence du mot *Kasparov* dans le contexte gauche va renforcer l'apparition des mots *chess* qui signifie *échecs* en Français et *champion*.

Nous proposons dans ce qui suit d'utiliser ce concept sur des corpus bilingues alignés d'où le terme employé de triggers inter-langues. Les triggers sont sélectionnés, non plus au sein d'une même langue mais entre deux langues.

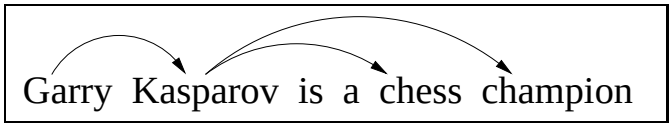

FIG. 4.1 – Exemples de triggers classiques

Un trigger inter-langue est donc défini comme étant un mot déclencheur, dans une langue, associé à une liste de mots déclenchés dans une autre langue. Ainsi, chaque mot x appartenant au vocabulaire V d'une langue est associé à n mots, ou triggers inter-langues, appartenant à une autre langue et qui lui sont le plus fortement corrélés. Plus formellement,

$$\forall x \in V, Trig_n(x) \text{ est l'ensemble des } n \text{ triggers inter-langues de } x$$

De la même manière que les triggers utilisés en modélisation du langage, les triggers inter-langues sont déterminés suivant la valeur de l'Information Mutuelle entre deux mots x et y. Mais cette fois, les mots x et y n'appartiennent pas à la même langue. L'IM entre x et y est calculée sur un corpus aligné au niveau de la phrase constitué de paires (P, P') où P est la traduction de P', elle est définie par :

$$IM(x,y) = P(x,y) log \frac{P(x,y)}{P(x) \times P(y)} \tag{4.2}$$

$$P(x) = \frac{N(x)}{N} \qquad P(x,y) = \frac{N(x,y)}{N} \tag{4.3}$$

où $N(x)$ (respectivement $N(y)$) est le nombre de phrases du corpus aligné dans lesquelles le mot x (respectivement y) apparaît, $N(x,y)$ est le nombre de paires de phrases du corpus aligné dans lesquelles les mots x et y co-occurrent et N est le nombre de paires de phrases constituant le corpus aligné. Seuls les couples (x,y) ayant une IM supérieure à un seuil fixé sont sélectionnés comme étant des triggers inter-langues.

La figure 4.2 illustre un exemple de triggers inter-langues de l'Anglais vers le Français. Cette

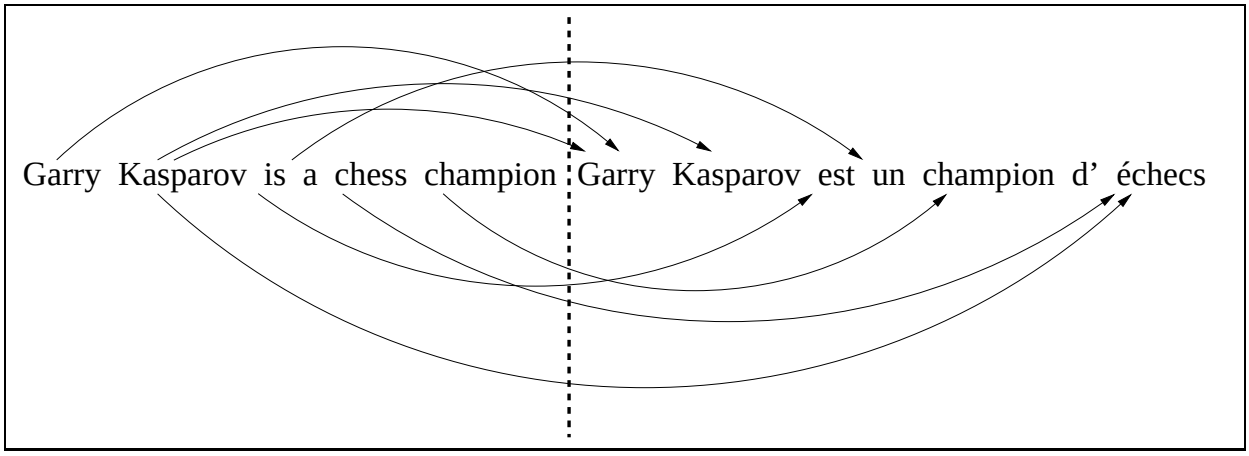

FIG. 4.2 – Exemples de triggers inter-langues

fois, la présence du mot *Kasparov* dans la phrase anglaise va déclencher l'apparition du mot *échecs* dans sa traduction en Français.

Notons que ce concept de triggers inter-langues a été utilisé, non pas sur des corpus alignés dans lesquels chaque phrase source est associée à sa traduction, mais sur des documents bilingues comparables traitant d'un même sujet, afin adapter les modèles de langage des langues faiblement dotées à partir d'autres langues très riches en termes de corpus [Kim 04].

L'utilisation de l'Information Mutuelle permet de quantifier le lien de corrélation pouvant exister entre deux unités. Plus sa valeur est importante, plus les deux unités sont corrélées. Les travaux menés en modélisation du langage ont montré que les triggers améliorent les performances des modèles n-grammes en ce sens que la présence d'un mot dans une phrase favorise l'apparition d'un mot présent dans la liste de ses triggers. Dans un contexte bilingue où des corpus bilingue aligné sont disponibles, l'utilisation des triggers permet d'établir des liens de corrélation entre des mots d'une langue source et des mots d'une langue cible. Nous espérons pouvoir établir des correspondances lexicales et donc des traductions en nous servant des triggers inter-langues.

4.3 Les triggers inter-langues : étude préliminaire

La première étape de nos travaux a consisté à selectionner des couples de triggers inter-langues sur les corpus bilingues alignés à notre disposition et à étudier leur potentiel éventuel pour la traduction. Les tableaux 4.1 et 4.2[12] présentent des exemples de triggers inter-langues sélectionnés sur le corpus d'apprentissage d'EUROPARL. Nous pouvons constater plusieurs phénomènes liés à l'utilisation de l'Information Mutuelle comme mesure de sélection des triggers. Le premier

f	$Trig_{n=10}(f)$	IM	e	$Trig_{n=10}(e)$	IM	e	$Trig_{n=10}(e)$	IM
abandonner	abandon	0,794	leave	laisser	1,604	abandon	abandonner	0,794
	leave	0,198		quitter	1,027		renoncer	0,436
	to	0,198		congé	0,705		ne	0,104
	abandoning	0,189		laisse	0,485		à	0,099
	we	0,159		ne	0,457		abandonne	0,098
	up	0,147		à	0,452		abandon	0,090
	give	0,123		soin	0,398		de	0,090
	abandoned	0,107		congés	0,339		et	0,089
	not	0,084		parental	0,332		d'	0,079
	and	0,082		laissons	0,329		pas	0,074
sonnerie	bell	0,271	I	je	232,694	bell	sonnerie	0,271
	ringing	0,040		j	75,283		glas	0,070
	ring	0,037		que	53,315		sonne	0,066
	votes	0,029		ai	39,097		alarme	0,065
	rung	0,028		voudrais	37,751		sonnette	0,061
	vote	0,027		suis	27,295		vote	0,031
	i	0,027		me	24,368		pas	0,030
	bells	0,025		président	23,298		retenti	0,030
	that	0,022		pense	20,895		députés	0,028
	announcing	0,022		monsieur	20,364		n'	0,026

TAB. 4.1 – Exemples de triggers inter-langues français-anglais et anglais-français révélant des synonymes

[12]Les valeurs d'IM indiquées dans les tableaux sont multipliées par 10^3

point remarquable au vu de ces exemples, est que les triggers retenus en première position, c'est-à-dire ceux dont la valeur de l'IM associée est la plus forte, sont souvent une traduction correcte du mot déclencheur. En effet, le meilleur trigger du mot français *abandonner* est le mot anglais *abandon*, de même que le trigger ayant la plus forte IM pour le mot anglais *president* est le mot français *président*. Autre fait, non moins intéressant, les triggers semblent avoir la capacité de détecter en plus d'une traduction correcte, des synonymes, ainsi que des formes fléchies de cette traduction. Dans nos exemples, les deux meilleurs triggers du mot anglais *leave* sont *laisser* et *quitter*. La forme conjuguée de *laisser*, à savoir *laisse*, est également présente dans la liste des meilleurs triggers. Le mot anglais *abandon* est associé prioritairement à *abandonner* et à *renoncer* qui sont deux traductions tout à fait acceptables de ce mot.

L'utilisation de l'IM ne présente toutefois pas que des avantages quant à la mise en correspon-

f	$Trig_{n=10}(f)$	IM	e	$Trig_{n=10}(e)$	IM	e	$Trig_{n=10}(e)$	IM
monsieur	mr	128,12	mr	monsieur	128,12	president	président	167,68
	president	117,44		président	107,82		monsieur	117,44
	commissioner	21,95		m,	82,67		le	48,31
	I	20,36		le	38,42		madame	37,25
	ladies	18,57		je	26,59		je	19,71
	gentlemen	18,50		rapport	16,59		chers	15,36
	you	13,70		collègues	11,37		collègues	14,94
	like	10,75		chers	10,52		présidente	11,08
	thank	10,28		voudrais	10,26		voudrais	10,73
	report	7,38		commissaire	9,72		commissaire	10,52
feu	fire	0,66	green	vert	9,00	the	la	74,49
	light	0,42		livre	6,62		l	56,57
	green	0,36		verts	5,14		le	54,30
	go-ahead	0,28		groupe	2,95		de	52,71
	flames	0,15		green	2,47		les	43,91
	give	0,11		verte	2,21		et	41,76
	fires	0,09		nom	1,20		des	39,29
	gave	0,08		sur	1,20		du	39,19
	the	0,07		unitaire	1,20		à	33,77
	to	0,06		mme	1,17		que	28,70

TAB. 4.2 – Exemples de triggers inter-langues français-anglais et anglais-français révélant des associations dites indirectes

dance de mots sources et de mots cibles. Bien qu'elle permette, à première vue, de détecter les traductions potentielles d'un mot, l'IM révèle également des couples de mots qui relèvent plus d'une association indirecte que d'une traduction potentielle [Melamed 00]. En effet, dans le tableau 4.2 par exemple, le mot français *monsieur* est associé aux deux meilleurs triggers *mr* et *president* ayant une valeur d'IM similaire. Comme l'illustre la figure 4.3, ce phénomène est dû au fait que *monsieur* et *mr* sont souvent rencontrés dans les mêmes paires de traductions, mais qu'en plus *mr* est très souvent utilisé avec *president* dans la partie anglaise du corpus. De ce fait, les associations directes entre *monsieur* et *mr* et entre *mr* et *president* créent un lien indirect entre *monsieur* et *president* symbolisé sur la figure par les traits en pointillés. Au vu de cet exemple, l'utilisation de séquences de mots paraît judicieuse pour palier ce problème lié aux associations

indirectes. En effet, si *mr* et *president* apparaissent très fréquemment ensembles, c'est qu'il a y de fortes chances pour qu'il forme une séquence de mots indissociable tout comme *monsieur* et *président*. L'emploi de séquences de mots au lieu de mots simples pour le calcul de l'IM permettrait alors d'établir une correspondance entre *monsieur le Président* et *mr President*. Nous en reviendrons dans le chapitre 6. Les associations indirectes expliquent également la présence de

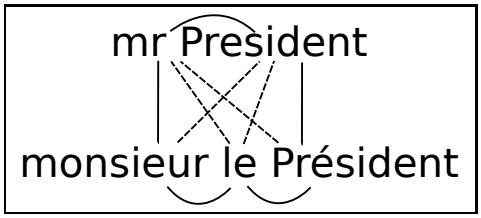

FIG. 4.3 – Associations directes(traits pleins) et indirectes(traits pointillés) révélés par les triggers inter-langues

mots outils, tels que *le, à, des, je* en français ou *I, you, like, and, to, a* en anglais dans les listes de triggers inter-langues. Il s'agit, en effet, de mots très fréquents et par conséquent présents dans la majorité des phrases des corpus bilingues alignés. Le tableau 4.3 présente les 4 meilleurs triggers inter-langues du mot français *voiture*. Si les deux premiers triggers correspondent bien à une traduction potentielle de *voiture*, ou à une forme fléchie d'une traduction, le troisième meilleur trigger *a* est un mot outil anglais qui ne s'apparente absolument pas à une traduction. Les occurrences indiquent que le mot *voiture* apparait 147 fois sur 255 dans la même paire de phrases que *a* alors qu'il n'est rencontré avec *vehicle* que 16 fois sur 255. Même si l'occurrence

x	$y \in Trig_{n=4}(x)$	$N(x)$	$N(y)$	$N(x,y)$	$IM(x,y)$
voiture	car	255	653	189	0,0020
	cars		687	49	0,0004
	a		200801	147	0,0001
	vehicle		409	16	0,0001
a	un	200801	130468	86129	0,097
	une		141461	89245	0,094
	d		197384	89707	0,045
	de		401732	153854	0,033

TAB. 4.3 – Exemples des triggers inter-langues pour le mot français porteur de sens *voiture* et le mot outil anglais *a*

du mot *a* est beaucoup plus importante que celle de *vehicle*, la formule de l'IM ne permet pas d'écarter le mot outil anglais des triggers inter-langues de *voiture*. En revanche, si nous nous intéressons maintenant aux 4 meilleurs triggers inter-langues du mot anglais *a*, nous remarquons que *voiture* n'en fait pas partie. En effet, ce dernier apparaît dans 255 phrases du corpus aligné et seulement 16 fois avec *a*. Le mot outil *a* couvre 200801 phrases du corpus alignés et co-occurre beaucoup plus que 16 fois avec d'autres mots outils anglais dont ses traductions.
De façon générale, lorsqu'un mot apparait relativement peu souvent dans le corpus aligné, il est fréquemment associé à un ou plusieurs mots outils qui vont par conséquent bruiter la liste de ses triggers inter-langues. A l'inverse si un mot a une occurrence très importante, l'impact des

mots outils sera moindre. Nous verrons dans la section suivante le moyen que nous utilisons pour minimiser la présence des mots outils dans la liste des triggers langues.

4.4 Construction d'un lexique bilingue à partir des triggers inter-langues

Dans un premier temps, nous avons utilisé les triggers inter-langues pour la construction d'un lexique bilingue Français-Anglais établissant des correspondances lexicales entre les mots d'un vocabulaire français V_F et les mots d'un vocabulaire anglais V_E.
L'algorithme de construction du lexique est le suivant :

1. Calculer l'Information Mutuelle pour chaque couple de mots (f_i,e_j) appartenant respectivement aux vocabulaires français V_F et anglais V_E
2. Déterminer, pour chaque mot f_i, l'ensemble de ses triggers $Trig_n(f_i)$ en sélectionnant les n meilleurs couples (f_i,e_j) suivant la valeur de l'IM décroissante
3. Déterminer de la même façon, pour chaque mot e_j, l'ensemble de ses triggers $Trig_n(e_j)$
4. Ajouter le couple (f_i,e_j) au lexique bilingue s'il répond à la contrainte de symétrie suivante :

$$f_i \in Trig_n(e_j) \quad et \quad e_j \in Trig_n(f_i) \tag{4.4}$$

Par le biais de cette contrainte de symétrie, nous espérons éliminer un maximum de bruit de la liste des triggers inter-langues (cf. section 4.3).

4.4.1 Influence de la contrainte de symétrie sur les triggers inter-langues

Avant d'ajouter un couple de triggers inter-langues (f_i,e_j) à notre lexique bilingue, nous lui imposons donc de remplir la condition de symétrie 4.4, et ce, dans le but de supprimer le bruit relatif aux mots outils. Parmi les triggers inter-langues de *voiture* (cf. tableau 4.3), la contrainte de symétrie nous permet d'écarter du dictionnaire bilingue le couple de mots (*voiture*,*a*). En effet, *a* fait partie des triggers inter-langues de *voiture* mais *voiture* n'est pas présent dans la liste des triggers inter-langues de *a*. Le tableau 4.4 présente les couples de triggers inter-langues retenus comme traduction après avoir appliqué la contrainte de symétrie sur les triggers inter-langues des tableaux 4.1 et 4.2. Pour les mots tels que *abandonner*, *sonnerie* et *feu* dont la fréquence n'excède pas 350 dans le corpus bilingue aligné, la contrainte de symétrie semble permettre d'écarter les mots outils de la liste des triggers inter-langues. En revanche, pour les mots très fréquents comme *Président* qui apparaît dans plus de 44000 phrases du corpus français, les mots outils bruitent toujours la liste des triggers inter-langues.
Nous avons étudié l'impact de la contrainte de symétrie sur le nombre de triggers inter-langues retenus pour chaque mot français f_j du corpus d'apprentissage d'EUROPARL. Pour cela, nous avons calculé l'Infomation Mutuelle pour chaque couple de mots (f_j, e_j) appartenant aux vocabulaires français et anglais du corpus d'apprentissage d'EUROPARL. Nous avons ensuite déterminé, pour chaque mot f_j, l'ensemble de ses triggers $Trig_n(f_j)$ en sélectionnant les n meilleurs couples (f_j, e_j) suivant la valeur de l'Information Mutuelle, et en faisant varier n de 10 à 200. La figure 4.4 donne l'évolution du nombre de triggers inter-langues $Trig_n(f_j)$ (barres "sans contrainte de symétrie") en fonction de n. Nous avons ensuite appliqué la contrainte de symétrie sur chacun des ensembles $Trig_n(f_j)$. Nous avons pour cela déterminé pour chaque mot e_j, l'ensemble de ses triggers $Trig_n(e_j)$ de la même façon que pour les ensembles $Trig_n(f_j)$ en faisant varier n de 10 à 200. Enfin, dans chaque ensemble $Trig_n(f_j)$, nous n'avons gardé que les triggers $e_i \in Trig_n(f_j)$

f	$e \in Trig_{n=10}(f)$	IM	e	$f \in Trig_n(e)$	IM
monsieur	mr	128,12	feu	fire	0,66
	president	117,44		go-ahead	0,28
	commissioner	21,95		flames	0,15
	i	20,36			
	ladies	18,58			
	gentlemen	18,51			
	you	13,70			
	like	10,76			
	thank	10,28			
abandonner	abandon	0,79	sonnerie	bell	0,27
	abandoning	0,19		ringing	0,04
	abandoned	0,11		ring	0,04
				rung	0,03
				bells	0,03

TAB. 4.4 – Couples de triggers inter-langues retenus après symétrisation

tel que $f_j \in Trig_n(e_i)$. L'évolution du nombre de triggers inter-langues retenus est donné en fonction de n par la série "avec contrainte de symétrie" de la figure 4.4. Quelle que soit la valeur de n,

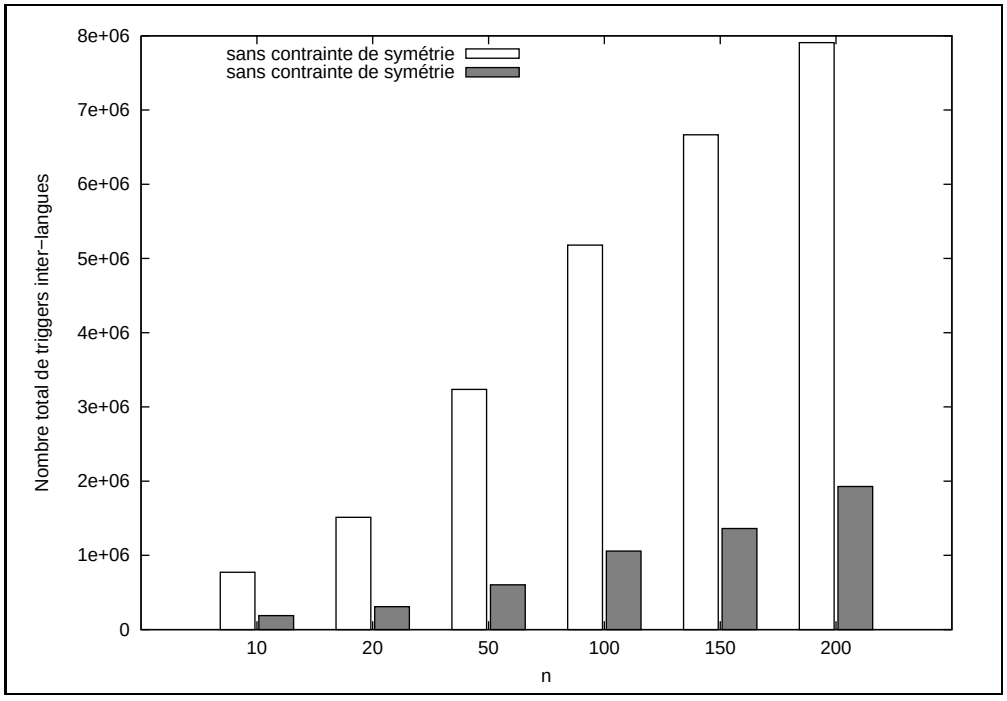

FIG. 4.4 – Influence de la contrainte de symétrie sur le nombre de triggers inter-langues

la contrainte de symétrie réduit considérablement le nombre total de triggers inter-langues. En effet, lorque nous sélectionnons pour chaque mot français les 10 meilleurs triggers inter-langues, leur nombre total s'élève alors à $773K$ dont seulement 24% sont finalement conservés dans le

lexique bilingue.

4.4.2 Evaluation du lexique bilingue

Pour construire notre lexique bilingue Français-Anglais, nous avons créé les vocabulaires à partir des mots apparaissant au moins 7 fois dans le corpus d'apprentissage d'EUROPARL. Le vocabulaire français V_F est ainsi composé de 26811 mots et le vocabulaire anglais V_E de 19588 mots. Nous nous sommes limités aux mots les plus fréquents de manière à pouvoir plus facilement repérer si leurs traductions potentielles proposées par les triggers inter-langues étaient correctes ou non. En effet, un mot apparaissant moins de 7 fois dans un corpus a toutes les chances d'être un mot très peu employé dans la langue, ce qui rend l'évaluation de ses traductions moins évidente. Nous avons ensuite déterminé pour chaque mot f_j (respectivement e_i) du vocabulaire français (respectivement anglais), l'ensemble des ses triggers inter-langues $Trig_{n=5}(f_j)$ (respectivement $Trig_{n=5}(e_i)$). La valeur de n a été fixée arbitrairement à 5. Nous pensons en fait qu'il s'agit d'un nombre de traductions généralement suffisant pour traduire un même mot. Enfin, nous n'avons retenu que les couples de mots (f_i,e_j) répondant à la condition de symétrie 4.4. Le tableau 4.5 présente un extrait du lexique bilingue ainsi construit. Pour évaluer la pertinence

Entrée	**Traductions possibles**		
humide	wetland	wet	rainforest
humble	humble	opinion	modest
mécaniquement	automatically	systematically	necessarily
méconnaissance	ignorance	lack	knowledge
sonnette	alarm	sound	bells
urgence	urgent	urgency	emergency

TAB. 4.5 – Un extrait du lexique bilingue Français-Anglais construit à partir des triggers inter-langues

du lexique bilingue ainsi construit, nous l'avons comparé à deux dictionnaires bilingues de référence : un dictionnaire distribué par ELRA [13] contenant 70832 entrées et un dictionnaire libre de droits disponible sur Internet [14] et contenant 41398 entrées. La comparaison porte uniquement du Français vers l'Anglais et pour que celle-ci soit justifiée, seuls les mots français présents à la fois dans un dictionnaire de référence et dans notre lexique bilingue sont pris en considération. Ici, notre lexique bilingue construit à partir des triggers inter-langues, que nous appelerons TrigDic, partage 10405 entrées françaises avec le dictionnaire fourni par ELRA et 11265 avec le dictionnaire distribué gratuitement sur Internet. L'évaluation est effectuée en terme de rappel, c'est-à-dire le nombre d'entrées du dictionnaire de référence correctement retrouvées dans notre lexique bilingue TrigDic. Les taux de rappel sont précisés dans le tableau 4.6. La colonne "Rang 1" indique le taux de rappel en ne prenant en compte que la traduction possible proposée en première position dans la liste des triggers inter-langues. La colonne "Rang 5" indique le même taux, mais cette fois, toutes les traductions présentes dans TrigDic, à savoir 5 par entrée française, sont prises en considération.

Les résultats montrent que si nous ne considérons que les traductions proposées en première

[13] M0033-3 SCI-FRAN-EURADIC
[14] http ://xdxf.revdanica.com/down/index.php

	Rang 1	**Rang 5**
ELRA	53%	65%
Internet	41%	52%

TAB. 4.6 – Evaluation du lexique bilingue TrigDic en terme de rappel

position (c'est-à-dire, pour un mot français, le trigger inter-langue dont l'IM est la plus élevée et qui répond à la contrainte de symétrie), le taux de rappel est alors de 53% par rapport au dictionnaire ELRA et de 41% par rapport à celui d'Internet. Ce taux de rappel augmente si davantage de triggers inter-langues sont pris en compte dans le calcul du taux de rappel. En effet, en considérant les cinq traductions proposées dans notre lexique bilingue, le taux de rappel atteint 65% pour le dictionnaire ELRA et 52% pour le dictionnaire tiré d'Internet.

En établissant le dictionnaire fourni par ELRA comme référence pour l'évaluation de notre dictionnaire TrigDic, nous pouvons dire que notre méthode de construction d'un lexique bilingue à partir de triggers inter-langues fournit une traduction pertinente pour 65% des mots français. Pour les 35% de mots français restant, les traductions proposées dans notre dictionnaire TrigDic ne se retrouvent pas dans le dictionnaire de Référence. Toutefois, une étude appronfondie a révélé qu'il ne s'agissait pas pour autant de traductions erronées mais que le dictionnaire auquel nous nous comparions n'était pas adéquat. En effet, une étude des résultats a permis de soulever plusieurs points importants, et notamment :

- Le dictionnaire ELRA propose en moyenne 2 traductions pour chaque mot français. Ce nombre est insuffisant compte tenu des nombreux sens ou des synonymes que peut avoir un mot français. Pour le mot *occidentale*, par exemple, le dictionnaire ELRA propose les traductions : *westerner* et *occidental*. Dans notre dictionnaire TrigDic, les traductions retenues sont *western*, *west* et *weu* (pour Western European Union). Le couple (*occidentale*,*western*) est une traduction correcte dans notre dictionnaire, pourtant le mot *occidentale* fait partie des 35% de mots français qui ne seront pas comptabilisés comme correct dans le taux de rappel puisque ce couple de traduction, aussi juste soit-il, ne figure pas dans le dictionnaire ELRA.
- Les traductions proposées par le dictionnaire ELRA pour des mots fréquemment employés tels que *voiture* ne sont pas toujours les traductions les plus intuitives. Il propose en effet comme traductions pour *voiture* : *auto*, *coach*, *conveyance* et *motorcar*, alors qu'intuitivement la traduction attendue serait plutôt *car*. Dans TrigDic, les traductions retenues sont moins surprenantes : *car*,*cars* et *vehicle*. Une fois de plus, le mot *voiture* ne sera pas comptabilisé comme correct dans le calcul du taux de rappel puisqu'aucune des traductions proposées dans TrigDic n'est présente dans le dictionnaire ELRA. Le tableau 4.7 donnent d'autres exemples de mots pour lesquels le dictionnaire ELRA proposent des traductions peu fréquemment employées alors que le dictionnaire TrigDic en fournit d'autres correctes, parmi les triggers inter-langues retenus, plus couramment utilisées.

Entrée	**ELRA**	**TrigDic**
chevaux	horseflesh	horses, animals, horse
chimère	bubble	illusion, fantasy, dream, fancy
déléguée	deputy	delegated, united, delegate, legislative

TAB. 4.7 – Comparaison des entrées du dictionnaire de ELRA et de notre lexique bilingue TrigDic

4.4.3 Discussion

Au vu de cette première étude, nous avons montré que l'utilisation des triggers inter-langues se révèle un bon moyen pour la construction d'un lexique bilingue Français-Anglais. Par rapport au dictionnaire ELRA, que nous avons choisi comme référence, les triggers inter-langues permettent, dans 65% des cas, de retrouver, au moins une même traduction d'un mot français. De plus, après une étude appronfondie, nous pouvons dire que parmi les 35% de mots français pour lesquels nous n'avons trouvé aucune traduction commune avec le dictionnaire ELRA, certains étaient toutefois associés à des traductions correctes par le biais des triggers inter-langues. En effet, le dictionnaire ELRA ne propose en moyenne que deux traductions par mot français. Il écarte ainsi des synonymes ou encore des double-sens que capturent fréquemment les triggers. De plus, le dictionnaire ELRA fournit des traductions parfois peu communes et qu'il n'est pas courant d'employer. De ce fait, de nombreuses traductions proposées dans TrigDic, en dépit de leur justesse, ne figurent pas dans le dictionnaire ELRA. Par conséquent, les 35% de mots français pour lesquels nous n'avons trouvé aucune traduction commune avec le dictionnaire ELRA relèvent davantage d'un choix de dictionnaire de référence *a posteriori* inadéquat que de mauvaises traductions établies par les triggers inter-langues.

Le recours à des fonctions de similarité, telle que l'Information Mutuelle, pour établir des correspondances lexicales entre deux langues n'est pas une méthode nouvelle dans la communauté de la traduction [Gale 91a, Fung 95, Melamed 00]. Dans ces travaux, l'algorithme de sélection des paires de mots équivalents est généralement le suivant :

1. Choix de la fonction de similarité qui servira de score pour établir ou non une équivalence entre les mots d'un vocabulaire source V_S et ceux d'un vocabulaire cible V_T.
2. Calcul du score S pour tous les couples de mots s_i, t_j appartenant respectivement à V_S et V_T.
3. Ordonnancement des couples (s_i, t_j) suivant les valeurs décroissantes des scores de similarité
4. Suppression des paires de mots dont le score ne dépasse pas un certain seuil

Pour calculer le degré de similitude entre deux mots x et y, Gale *et al.* [Gale 91a] utilisent la fonction $\phi^2(x, y)$ qui est une mesure semblable à celle du χ^2. Ils montrent que 98% des paires de mots sélectionnées dans un contexte d'expérimentation particuliers, se révèlent être des traductions. Dans ses travaux, Dan Melamed [Melamed 00], quant à lui fait appel à la fonction $G^2(x, y)$ qui selon ses études obtient de meilleurs résultats que la fonction ϕ^2. Ces deux fonctions de similarité ont la particularité de prendre en compte, à la fois la présence de x et/ou y dans les paires de phrases d'un corpus bilingue aligné, mais aussi leur non-apparition $\bar{x}$, $\bar{y}$. Fung [Fung 95], quant à lui, utilise l'Information Mutuelle sur des vecteurs, pour trouver les paires de mots semblables et dont les fréquences sont faibles. Alors qu'une variante de l'IM existe de manière à prendre en compte la non-apparition d'évènements, Fung a utilisé la formule de l'IM classique exploitant uniquement la présence et non l'absence d'évènements pour définir des liens de corrélation.
Quelle que soit la fonction utilisée, les différents travaux montrent leur fort potentiel à établir les correspondances lexicales entre les mots d'une langue cible et ceux d'une langue source. Tous utilisent un procédé itératif en estimant un premier lexique bilingue et en ré-estimant ses paramètres en fonction des paires de mots ayant les meilleurs scores de similarité. Dan Melamed notamment [Melamed 00] utilise l'algorithme de liens concurrents (*competitive linking*) suivant :

1. Ordonner les paires de mots du lexique bilingue de façon décroissante suivant leur score de similarité associé
2. Pour chaque paire de mots de la liste ainsi ordonnée

- Lier dans le corpus d'apprentissage toutes les occurrences de cette paire de mots
- Ecarter ces liens du corpus d'apprentissage

3. Ré-estimer un nouveau lexique bilingue

Le lexique bilingue est ainsi affiné au fur et à mesure des itérations en privilégiant d'abord les couples de mots ayant les scores les plus forts.
Dans les travaux cités ci-dessus, chaque lexique bilingue obtenu est évalué au sein d'une tâche d'alignement de mots. Un algorithme de programmation dynamique est généralement utilisé pour aligner les phrases d'un corpus de test à l'aide des paramètres du lexique bilingue. Les liens ainsi construits sont ensuite comptabilisés en lien correct ou lien incorrect. Gale *et al.* rapportent que sur 800 paires de phrases, leur procédure d'alignement met en correspondance 61% des mots anglais avec des mots français et que 95% des liens établis sont corrects. Dan Melamed établit que son modèle obtient de meilleures performances que le modèle IBM1 lorsqu'il s'agit de trouver les mots français correspondants aux mots anglais d'un extrait de la Bible. Fung annonce quant à lui une précision moyenne de $73,1\%$ sur les paires de mots établies dans le cadre de ses expérimentations.

Ces travaux nous ont conforté dans l'idée que l'utilisation des triggers inter-langues peut se révéler prometteuse en Traduction Automatique. En effet, les triggers inter-langues sont sélectionnés grâce à l'Information Mutuelle et les travaux menés dans le passé ont montré l'aptitude de telles fonctions de similarité à établir de bonnes correspondances lexicales. Dans nos travaux, nous proposons de calculer des modèles de traduction, établissant les correspondances lexicales entre une langue cible et une langue source, en nous servant du potentiel des triggers inter-langues. Contrairement aux travaux déjà menés, nous proposons d'établir les correspondances lexicales aussi bien au niveau du mot qu'au niveau des séquences de mots et ce, en une seule itération (nous ne faisons qu'une seule passe sur le corpus d'apprentissage) et sur des quantités de données beaucoup plus importantes. En outre, nous évaluons nos modèles de traduction, non pas dans un exercice d'alignement de phrases au niveau du mot, mais dans le cadre d'un processus complet de traduction d'un texte, d'une langue source vers une langue cible. En effet, nous proposons d'intégrer nos modèles de traduction basés sur les triggers inter-langues au sein d'un décodeur et d'évaluer la qualité des traductions ainsi produites. Dans le chapitre suivant, nous présentons les différents modèles de traduction de mots sur lesquels nous avons travaillé ainsi que les expériences que nous avons conduites, avant de passer au niveau des séquences de mots.

Chapitre 5

La traduction par mot : les triggers 1-1

5.1 Introduction

Nous avons montré dans le chapitre précédent le potentiel des triggers inter-langues pour la construction d'un lexique bilingue Français-Anglais. L'étude préalable que nous avons menée nous a poussés à exploiter le concept de triggers inter-langues dans le cadre de la Traduction Automatique.
Pour ce faire, nous proposons dans la première section de ce chapitre d'utiliser les triggers inter-langues afin d'établir des distributions de probabilité, également appelées tables ou modèles de traduction, du type $p(e|f)$, e et f étant respectivement des mots anglais et français appartenant à un vocabulaire prédéfini. Ainsi, chaque couple de mots se voit attribuer une probabilité indiquant la chance que celui-ci forme une paire de traduction. Dans la suite de nos travaux, nous traduisons automatiquement des données textuelles de l'Anglais vers le Français, c'est pourquoi nous nous bornons à estimer les probabilités $p(e|f)$. Toutefois l'estimation des probabilités $p(f|e)$ peut s'obtenir de la même façon en inversant simplement l'Anglais et le Français. Rappelons que dans le domaine de la modélisation du langage, Kim [Kim 04] propose d'enrichir des langues faiblement dotées à partir d'autres langues très riches en termes de corpus par le biais de distributions de probabilité reposant également sur des triggers inter-langues. Il se sert en effet des triggers inter-langues pour établir des correspondances entre des documents traitant de même sujet mais dans des langues différentes.
Dans nos travaux, nos tables de traduction sont exploitées dans un processus complet de traduction, à savoir la traduction d'un texte d'une langue source vers une langue cible grâce à l'information apportée par les triggers inter-langues. L'évaluation de ces tables est faite en termes de qualité des traductions ainsi produites. La deuxième section de ce chapitre est consacrée à la description du cadre expérimental ainsi qu'à l'évaluation des tables de traduction.

5.2 Modèle de traduction de mots : les triggers 1-1

Nous avons retenu trois modèles de traduction de mots, tous utilisant les couples de mots sélectionnés comme étant des triggers inter-langues. Nous présentons, dans un premier temps deux modèles ainsi que leur évaluation avant de présenter le troisième modèle proposé en amélioration aux deux premiers. Les modèles proposés différent sur les règles, plus ou moins restrictives, appliquées à chaque trigger inter-langue afin de décider si oui ou non il doit être intégré à la table de traduction. Chaque couple de mots retenu dans un modèle est ensuite associé à une probabilité calculée en fonction de son Information Mutuelle préalablement calculée.

5.2.1 Le modèle traduction Trig-n

Au sein du modèle de traduction que nous appelons Trig-n (où n est le nombre de triggers retenus pour chaque mot français), nous estimons que tous les triggers inter-langues peuvent être assimilés à des traductions à la seule condition que leur Information Mutuelle fasse partie des meilleures. Par conséquent, un mot anglais e_i est une traduction probable du mot français f_j s'il fait partie de ses n meilleurs triggers inter-langues suivant les valeurs de l'IM décroissantes. La probabilité associée à la traduction (f_j, e_i) est ensuite calculée en normalisant la valeur de l'Information Mutuelle du couple :

$$\forall f_j \in V_F, \forall e_i \in Trig_n(f_j) \quad P_{Trig-n}(e_i|f_j) = \frac{IM(f_j, e_i)}{\sum_{e_k \in Trig_n(f_j)} IM(f_j, e_k)} \tag{5.1}$$

5.2.2 Le modèle de traduction Sym-n

Dans le chapitre précédent, l'étude des triggers inter-langues a montré la présence de bruit dans les traductions potentielles. Il s'agissait notamment de mots très fréquents comme les mots outils de la langue par exemple. L'application de la contrainte de symétrie 4.4 permettait alors de réduire ce bruit. Nous avons voulu analyser si l'impact de cette contrainte de symétrie s'avère également positif dans un cadre de Traduction Automatique. C'est pourquoi, au sein de la table de traduction que nous appelons Sym-n (où n est le nombre de triggers retenus pour chaque mot français), sont considérés comme traductions possibles, l'ensemble Sym_n des couples (f_j, e_i) qui respectent la contrainte de symétrie. Ainsi, un mot anglais e_i appartient aux traductions possibles du mot français f_j, si e_i fait partie des n meilleurs triggers inter-langues de f_j et inversement si f_j est un des n meilleurs triggers inter-langues de e_i. Plus formellement,

$$(f_j, e_i) \in \text{Sym}_n \text{ si } e_i \in Trig_n(f_j) \text{ et } f_j \in Trig_n(e_i)$$

La probabilité associée à ce couple est calculée de la même manière que pour la table de traduction Trig-n, en normalisant la valeur de l'IM du couple :

$$\forall f_j \in V_F, \forall e_i \in Sym_n(f_j) \quad P_{Sym-n}(e_i|f_j) = \frac{IM(f_j, e_i)}{\sum_{e_k \in Sym_n(f_j)} IM(f_j, e_k)} \tag{5.2}$$

La contrainte de symétrie nous permet d'affiner la liste des triggers inter-langues de f_j pour ne retenir comme traductions probables que les plus pertinents. Nous supposons en effet que si e_i est un des n mots les plus corrélés avec f_j et que f_j est également dans les n mots les plus déclenchés par e_i, alors il y a de fortes chances que e_i soit une traduction de f_j. Comme nous l'avons montré dans le chapitre précédent, la contrainte de symétrie réduit considérablement le nombre de couples retenus parmi les triggers inter-langues.

5.3 Evaluation des modèles de traductions de mots

5.3.1 Cadre expérimental

Dans nos expérimentations, nous nous plaçons dans un contexte de traduction textuelle dans le sens Anglais vers Français. Nous avons donc besoin de déterminer des tables de traduction du type $P(e_i|f_j)$ où e_i et f_j sont des mots appartenant respectivement aux vocabulaires Anglais V_E et Français V_F prédéfinis. Les expériences décrites par la suite sont menées indépendamment à la fois sur le corpus EUROPARL et sur le corpus SSTITRES (cf. chapitre 3 pour la description de

ces corpus). Notre objectif est de comparer en terme de score BLEU la qualité des traductions produites par le décodeur PHARAOH avec nos tables de traduction fondées sur les triggers inter-langues avec la qualité des traductions produites avec une table de traduction issue des modèles IBM.

5.3.1.1 Apprentissage des tables de traduction

Pour construire nos tables de traduction, nous calculons tout d'abord en une seule passe sur le corpus d'apprentissage, l'Information Mutuelle pour chaque couple possible (f_j, e_i) tel que f_j et e_i appartiennent respectivement à V_F et V_E. Les vocabulaires V_F et V_E sont respectivement constitués de tous les mots appartenant aux données françaises et anglaises du corpus d'apprentissage bilingue aligné. Aucun filtre sur le nombre d'occurrences des mots dans le corpus n'est appliqué pour construire les vocabulaires. Nous déterminons ensuite pour chaque mot f_j un ensemble de triggers inter-langues $Trig_n(f_j)$ constitués des n mots anglais ayant la plus forte IM avec f_j. Nous définissons de la même façon pour chaque mot anglais e_i un ensemble de triggers inter-langues $Trig_n(e_i)$. Rappelons, en effet, que pour les tables de traduction Trig-n, seuls les triggers de chaque mot français f_j nous intéressent mais que nous avons besoin des triggers de chaque mot anglais e_i pour construire les tables de traductions Sym-n. Enfin, pour chaque ensemble $Trig_n(f_j)$ de triggers inter-langues défini, nous appliquons les contraintes précédemment citées pour décider quels triggers intégrer ou non aux différentes tables de traduction et avec quelle probabilité.
Dans la suite de nos travaux, nous comparons nos tables de traduction fondées sur les triggers inter-langues avec la table de traduction de mots issue du modèle IBM 3. Comme nous l'avons décrit dans la section 2.4.2.1, le modèle IBM 3 se décompose en plusieurs sous-modèles dont une table de traduction de mots, un modèle de distorsion qui définit les probabilités d'alignement entre les mots de la phrase source et ceux de la phrase cible et un modèle de fertilité qui estime pour chaque mot source le nombre de mots cibles qui lui seront connectés. L'apprentissage des paramètres de ces sous-modèles s'effectue itérativement suivant l'algorithme Expectation-Maximisation sur le corpus d'apprentissage à partir des paramètres du modèle IBM 2. Chaque sous-modèle influe sur l'estimation des paramètres des autres sous-modèles. Autrement dit, les paramètres de la table de traduction dépendent des paramètres du modèle d'alignement et du modèle de fertilité. Nous avons entraîné le modèle IBM 3 à l'aide de l'outil Giza++ [Och 03b] sur les mêmes corpus d'apprentissage que pour nos modèles fondés sur les triggers inter-langues, et nous nous sommes intéressés seulement à la table de traduction $t(e_i|fj)$.

5.3.1.2 Optimisation et évaluation des tables de traduction

L'optimisation et l'évaluation des tables de traduction sont effectuées sur une tâche de traduction automatique de données textuelles dans le sens Anglais-Francais en utilisant le décodeur Pharaoh. Pour ce faire, ce décodeur requiert en paramètres d'entrée le corpus Anglais à traduire, une table de traduction $p(e_i|f_j)$ qui définit les probabilités de traduction pour chaque couple de mots (e_i, f_j) de la table, et un modèle de langage du Français. Les tables de traduction que nous utilisons, sont soient celles fondées sur les triggers inter-langues, soit la table de traduction du modèle IBM 3. Le modèle de langage Français est un modèle trigramme appris sur le même corpus d'apprentissage que les tables de traduction à l'aide de l'outil CMU toolkit [Rosenfeld 95]. Pharaoh calcule ensuite à l'aide de ces paramètres la traduction française la plus probable du corpus d'entrée. Le score BLEU attribué aux différentes traductions ainsi produites est ensuite utilisé comme fonction objectif à maximiser.

L'optimisation consiste à trouver l'ensemble des paramètres qui va garantir une qualité de traduction optimale en terme de score BLEU sur un corpus de développement. Les paramètres à estimer sont nombreux. Concernant le décodeur PHARAOH, il faut déterminer les poids des modèles impliqués dans le calcul de la meilleure traduction. Comme nous l'avons décrit dans la section 2.5, PHARAOH estime la probabilité de traduction $p(e|f)$ à partir de quatre modèles de la façon suivante :

$$p(e|f) = p_{TM}(e|f)^{\lambda_{TM}} * p_D(e|f)^{\lambda_D} * p_{LM}(f)^{\lambda_{LM}} * \omega^{long(f)\lambda_W} \tag{5.3}$$

où p_{TM} est le modèle de traduction, p_D le modèle de distorsion, p_{LM} le modèle de langage et ω un modèle destiné à pénaliser les phrases produites trop longues ou trop courtes par rapport à la phrase d'entrée. La qualité des traductions produites dépend fortement de la table de traduction fournie à PHARAOH. Toutefois, l'estimation des poids λ_{TM}, λ_D, λ_{LM} et λ_W demeure également très importante pour garantir une meilleure qualité de traduction.
Concernant nos tables de traduction Trig-n et Sym-n, il est également nécessaire de déterminer sur le corpus de développement le nombre de triggers optimal n à selectionner pour chaque mot du vocabulaire français.

Dans nos expériences, nous déterminons un jeu complet de paramètres à chaque utilisation de PHARAOH avec une table de traduction différente. Une fois les paramètres optimaux déterminés sur le corpus de développement, ils sont appliqués sur le corpus de test. Ce dernier est traduit à l'aide du décodeur Pharaoh pour chacune des différentes tables de traduction testées et donne alors lieu à différentes traductions automatiques. Les sorties ainsi produites sont comparées en terme de score BLEU.

5.3.2 Optimisation des modèles de traduction

5.3.2.1 Etude de l'impact du nombre de triggers n

Notre premier objectif est de déterminer le nombre optimal n de triggers inter-langues à retenir pour chaque mot français f_j appartenant au vocabulaire. Pour cela, nous utilisons le décodeur Pharaoh sur le corpus de développement avec chacune des tables de traduction Trig-n et Sym-n séparément en faisant varier n de 10 à 200. Les poids λ_{TM}, λ_D, λ_{LM} et λ_W de chaque modèle impliqué dans le calcul de la meilleure traduction sont fixés aux valeurs par défaut[15]. Les traductions produites par PHARAOH sont ensuite évaluées en terme de score BLEU. Les expériences ont été conduites parallèlement sur les corpus EUROPARL et SSTITRES et sont reportées respectivement dans les figures 5.1 et 5.2. La qualité en terme de score BLEU des traductions produites par le décodeur avec les tables Trig-n et Sym-n sont respectivement représentées par les séries étiquetées *Trig-n* et *Sym-n*. Les séries *Smooth-n* correspondent au modèle Smooth-n utilisant également les triggers et que nous détaillerons par la suite dans la section 5.3.2.3. A titre de comparaison, la courbe linéaire *IBM 3* indique le score BLEU associé aux traductions produites par PHARAOH avec la table de traduction issue du modèle IBM 3.
Pour EUROPARL (figure 5.1), que ce soit pour les traductions produites avec les tables Trig-n ou avec les tables Sym-n, le score BLEU progresse de façon remarquable (+2 points) lorsque le nombre de triggers inter-langues retenus pour chaque mot français passe de $n = 10$ à $n = 20$. Il se stabilise ensuite quelle que soit la valeur de n. En effet, pour les tables Trig-n, la différence

[15] un poids d'une valeur de 1 pour les modèles de traduction, de langage et de distorsion et une pénalité de 0 pour les traductions trop courtes

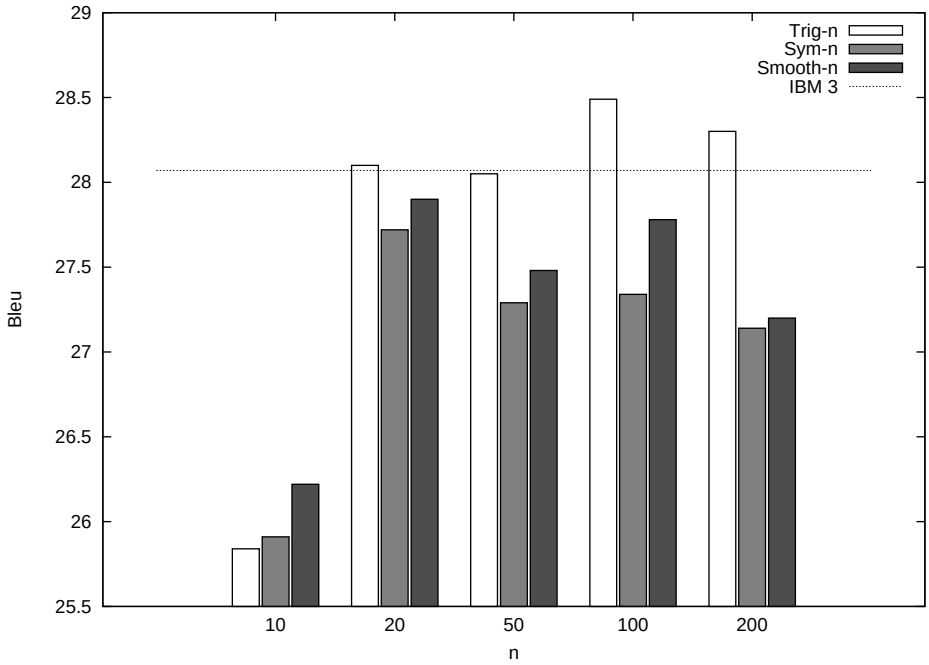

FIG. 5.1 – Evaluation des traductions produites à l'aide des tables de traduction *Trig-n*, *Sym-n* et *Smooth-n* en fonction de n sur le corpus de développement de EUROPARL

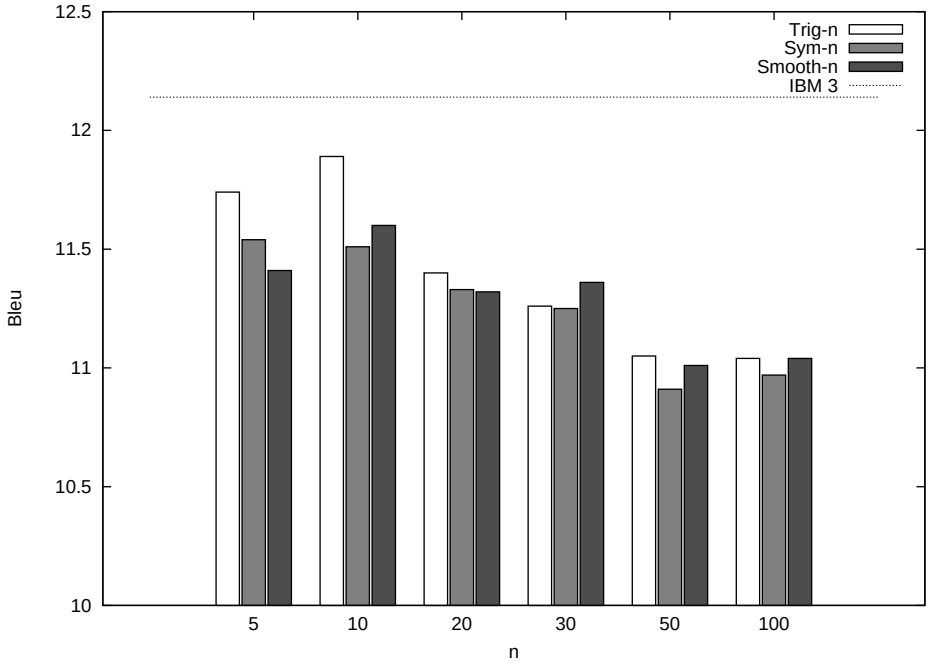

FIG. 5.2 – Evaluation des traductions produites à l'aide des tables de traduction *Trig-n*, *Sym-n* et *Smooth-n* en fonction de n sur le corpus de développement de SSTITRES

de score BLEU pour n variant de 20 à 200 n'excède pas $0,3$ points et il en va de même pour les table Sym-n. Pour les tables Trig-n, la valeur optimale de BLEU est de $28,49$, elle est atteinte lorsque n vaut 100 (c'est-à-dire que 100 triggers inter-langues sont conservés comme traduction pour chaque mot f_j). Pour les tables Sym-n, le score maximum est de $27,72$ pour une valeur de n de 20.
Ces premiers résultats indiquent que le décodeur nécessite une grande largesse de choix dans les traductions possibles d'un mot pour atteindre de bonnes performances. Un minimum de 20 traductions potentielles par mot français est nécessaire pour atteindre les meilleures résultats avec les tables Trig-n et Sym-n. Ce nombre est conséquent par rapport au nombre de traductions proposées pour un même mot dans un dictionnaire bilingue. Concernant la table de traduction issue du modèle IBM 3, la conclusion que le décodeur requiert un large éventail de choix dans les traductions reste vraie. En effet, elle compte en moyenne 23 traductions par mot français, et le score BLEU associé aux traductions produites par PHARAOH avec cette table est de $28,07$. Notons qu'en ne retenant que 20 triggers inter-langues par mot français dans notre table de traduction Trig-n ($n = 20$), les traductions produites par PHARAOH obtiennent un score BLEU très légèrement supérieur. Il vaut alors en effet $28,10$.

Pour SSTITRES (figure 5.2), l'impact du nombre de triggers retenus n semble nettement moins évident. Le score BLEU est moins sensible à la valeur de n aussi bien concernant les tables Trig-n que Sym-n. Il ne varie en effet pas de plus de $0,9$ points pour le décodage avec les tables Trig-n, et de plus de $0,7$ pour le décodage avec les tables Sym-n et ce, quelle que soit la valeur de n. Et à l'inverse des résultats sur EUROPARL, les meilleures traductions en terme de score BLEU sont obtenues pour une valeur de n inférieure ou égale à 10. Pour les tables Trig-n, la valeur optimale de BLEU de $11,89$ est atteinte lorsque n vaut 10 et pour les tables Sym-n, le score maximum de $11,54$ est atteint pour $n = 5$. Concernant la table de traduction du modèle IBM 3, elle compte en moyenne 6 traductions par mot français, ce qui est comparable aux tables Trig-n et Sym-n avec lesquelles les meilleures performances sont atteintes. Notons cependant que le score BLEU des traductions produites avec la table du modèle IBM 3 obtiennent un score BLEU de $12,14$. La faiblesse des résultats, ainsi que le nombre optimal n de traductions proposées dans les tables de traduction pour chaque mot français, s'expliquent par le fait que SSTITRES est un corpus beaucoup moins fourni que EUROPARL et dans lequel les redondances sont moins nombreuses et rendent donc le traitement statistique moins efficace. Par conséquent, inclure plus de 10 traductions potentielles pour chaque mot français induit de bruiter trop fortement la table de traduction et n'améliore aucunement la qualité des traductions en terme de score BLEU.

5.3.2.2 Etude de l'impact de la contrainte de symétrie

Nous nous sommes ensuite intéressés à l'influence de la contrainte de symétrie sur la qualité des traductions produites par le décodeur Pharaoh.
Pour EUROPARL (figure 5.1), l'impact de la contrainte de symétrie est positif uniquement lorsque chaque mot français est associé à 10 triggers inter-langues dans la table de traduction. En effet, pour une valeur de $n = 10$, le score BLEU est moins élevé pour les traductions automatiques produites avec la table Trig-n que pour celles produites avec la table Sym-n. Dans ce cas, nous pouvons dire que la contrainte de symétrie permet effectivement de sélectionner de meilleurs couples de traduction que la simple valeur de l'Information Mutuelle. Malheureusement, ce constat ne s'applique pas lorsque plus de 10 triggers sont retenus. En effet, le score attribué aux traductions automatiques produites avec les tables Trig-n demeure toujours plus élevé par rapport à celles produites avec les tables Sym-n.

La contrainte de symétrie, bien qu'apportant une meilleure qualité au niveau des couples de mots retenus dans la table de traduction en éliminant certains mots outils et certaines associations indirectes, engendre donc une diminution du score BLEU. Il existe plusieurs raisons à cela. Tout d'abord, nous avons vu que le score BLEU s'appuyait sur la correspondance entre les mots d'une traduction automatique et les mots d'une traduction de référence. Les mots outils apparaissent souvent dans les traductions de référence. Ainsi, un mot outil présent dans la liste des traductions possibles d'un mot français, même s'il ne constitue pas une traduction correcte, peut être comptabilisé dans le calcul du score BLEU s'il est tout de même choisi comme traduction par le décodeur. La contrainte de symétrie en éliminant de nombreux mots outils de la table de traduction supprime par la même des correspondances entre les traductions automatiques et les traductions de référence et fait donc diminuer le score BLEU.
Nous avions également pu constater dans la section 4.4.1 que l'application de cette contrainte de symétrie sur les triggers inter-langues en diminuait considérablement le nombre. Elle doit par conséquent en plus de supprimer du bruit, écarter des couples de mots qui seraient de bonnes traductions. Prenons l'exemple du mot français " maison " présent dans le vocabulaire français du corpus EUROPARL. Dans la liste de ses 20 meilleurs triggers inter-langues se trouve le mot anglais " house " qui est une traduction correcte. Cependant " house " dans le corpus EUROPARL se traduit plus souvent par " assemblée " que par " maison " qui ne se trouve donc pas dans les 20 meilleurs triggers de " house ". En conséquence, le couple ("maison", "house") est écarté de la table de traduction.
Pour SSTITRES (figure 5.2), quel que soit le nombre de triggers retenus, la contrainte de symétrie dégrade les performances du décodeur. Ce corpus, qui plus est difficile, dispose de peu de données. Filtrer encore par la suite la liste de triggers n'apporte donc aucun bénéfice.

La première série d'expériences faites avec les tables de traduction Trig-n a montré que plus le décodeur avait un large choix dans les traductions possibles d'un mot, meilleur était le score BLEU associé aux traductions qu'il produisait. La contrainte de symétrie restreint trop le nombre de traductions possibles pour un mot français. Pour palier ce problème de suppression de couples de traductions qui pourraient s'avérer corrects et/ou utiles pour le décodeur, nous avons proposé un troisième modèle de traduction qui utilise la contrainte de symétrie sans toutefois écarter définitivement les triggers non-symétriques de la table de traduction.

5.3.2.3 Le modèle de traduction Smooth-n

Afin de ne pas affecter une probabilité nulle aux couples (f_j, e_i) qui ne satisfont pas la contrainte de symétrie *" e_i déclenche f_j et f_j déclenche e_i "*, nous proposons d'utiliser une technique de lissage pour estimer une troisième table de traduction que nous appelons par conséquent Smooth-n. En modélisation du langage, ces techniques dites de *smoothing* permettent de lisser les probabilités de manière à ce que chaque évènement, même impossible, se voit affecter une probabilité [Ney 94]. Nous proposons d'employer le même type de technique. Pour ce faire, nous réduisons la probabilité des triggers symétriques des tables Sym-n. La masse ainsi récupérée est répartie uniformément sur les triggers non symétriques. Cette nouvelle estimation est calculée de la manière suivante :

$$\forall e_i \in Trig_n(f_j) \quad P_{smooth-n}(e_i|f_j) = \begin{cases} P_{Sym-n}(e_i|f_j) - \epsilon & \text{si } e_i \in Sym_n(f_j) \\ \gamma & \text{sinon} \end{cases} \tag{5.4}$$

Pour un mot français, nous retirons une quantité ϵ à chaque probabilité de traduction assignée à ses triggers inter-langues symétriques et nous redistribuons la masse récoltée uniformément sur

ses autres triggers inter-langues non symétriques. γ est la somme de probabilités soustraites aux triggers symétriques divisée par le nombre de triggers non symétriques. Ainsi, chaque trigger non symétrique recevra la même probabilité non nulle dans la table de traduction Smooth-n.

La dernière série des histogrammes des figures 5.1 et 5.2, notée Smooth-n, indique la qualité des traductions automatiques produites par le décodeur PHARAOH avec les tables de traductions Smooth-n avec n, le nombre de triggers inter-langues retenus pour chaque mot du vocabulaire, variant de 10 à 200.
Sur EUROPARL, les traductions produites avec les tables Smooth-n obtiennent un meilleur score BLEU que celui des traductions produites avec les tables Sym-n et ce, quelle que soit la valeur de n. Ceci confirme le fait que la contrainte de symétrie supprime trop de couples de mots parmi lesquels certains aident le décodeur à produire de meilleures traductions du point de vue du score BLEU. Malheureusement, n'accorder aux couples de mots non symétriques qu'une probabilité très faible par rapport aux triggers symétriques fait que les tables Smooth-n ne parviennent pas à ce que le décodeur atteigne les mêmes performances qu'avec les tables Trig-n.
Pour SSTITRES, concernant la qualité des traductions automatiques générées par le décodeur, nous pouvons constater qu'elle est également meilleure en terme de score BLEU avec les tables Smooth-n qu'avec les tables Sym-n, sauf pour $n = 10$. Par conséquent, nous pouvons en déduire que la contrainte de symétrie est en effet trop restrictive, et que le fait de conserver les triggers non-symétriques permet bien d'améliorer les performances. En revanche, malgré ces efforts de lissage, notre système reste le plus performant lorsque chaque trigger inter-langue est considéré comme une traduction potentielle qu'il respecte la contrainte de symétrie ou non (série *Trig-n*).

En conclusion à ces premières séries de tests, nous constatons que le décodeur PHARAOH obtient de meilleures performances lorsqu'il a un plus large choix de traductions possibles pour un même mot en dépit du fait qu'il peut y avoir de nombreux couples de mots ne représentant pas une réelle traduction. Resteindre le nombre de traductions potentielles d'un mot présente le risque trop important d'écarter des couples utiles sans lesquels le décodeur PHARAOH ne peut obtenir des performances optimales en terme de score BLEU.

Dans la suite, nous avons donc décidé de poursuivre nos expérimentations avec les tables de traduction avec lesquelles la qualité des traductions a été la meilleure en terme de score BLEU. Il s'agit de la table de traduction Trig-n avec $n = 100$ qui a permis d'obtenir un score BLEU optimal de $28,49$ pour le corpus EUROPARL et de la table Trig-n avec $n = 10$, pour le corpus SSTITRES, avec laquelle le score BLEU optimal est de $11,89$. Elles seront notées par la suite indépendemment Triggers 1-1 pour souligner qu'il s'agit des tables de traduction de mots basées sur les triggers du type "1 mot déclenche 1 mot".
Rappelons qu'à titre de comparaison, dans les mêmes conditions, nous avons évalué la qualité des traductions produites sur les mêmes corpus de développement par le décodeur PHARAOH avec la table de traduction du modèle IBM 3. Les scores BLEU résultant sont de $28,07$ pour EUROPARL et $12,14$ pour SSTITRES (courbe linéaire *IBM 3* dans les figures 5.1 et 5.2). Les tables de traduction basées sur les triggers obtiennent donc des résultats comparables à la table issue du modèle IBM 3. Mais alors que l'apprentissage du modèle IBM 3 requiert l'apprentissage au préalable des modèles IBM 1 et 2 et que chacun d'eux nécessitent plusieurs itérations, la sélection des triggers ne demande qu'un passage sur le corpus d'apprentissage suivi d'une normalisation. Notre approche est dont plus simple et plus rapide à mettre en œuvre.

5.3.3 Optimisation du décodeur

Une fois notre table de traduction basée sur les triggers inter-langues optimisée sur le corpus de développement, nous avons par la suite établi les paramètres optimaux du décodeur PHARAOH.

Le processus de traduction automatique effectué par le décodeur PHARAOH implique quatre modèles : une table de traduction qui permet de trouver les traductions de chaque composant de la phrase à traduire, un modèle de langage qui contrôle la vraisemblance de la traduction produite, un modèle de distorsion qui permet d'ordonner différemment les mots de la phrase à traduire et enfin une pénalité qui rend compte de la différence de taille entre la phrase à traduire et la traduction proposée. Chacun de ces modèles a un poids dans le calcul des hypothèses de traduction comme le rappelle l'équation 5.3. La qualité des traductions automatiques produites dépend fortement de la valeur de ces poids. La manière la plus simple d'établir la valeur de ces paramètres est de procéder à la main en testant une multitude de combinaisons de poids et en gardant le jeu de paramètres avec lequel les traductions automatiques ont un score BLEU optimal. Une autre façon d'estimer la valeur optimale de chaque poids est d'utiliser l'algorithme Minimum Error Rate Training (MERT) [Och 03a] sur un corpus de développement avec comme objectif la maximisation du score BLEU. Dans un premier temps, le décodeur est lancé avec un jeu de paramètres sur les phrases sources du corpus de développement. Il produit alors pour chacune d'entre elles une liste des n meilleures hypothèses de traduction. Les valeurs des paramètres sont ensuite optimisées sur ces listes. Le processus est itéré jusqu'à convergence des valeurs. MERT bien que fournissant des résultats satisfaisants n'est pas un algorithme idéal. Il est notamment confronté au problème du maximum local. L'optimisation des poids des modèles demeure donc un domaine à améliorer en recherche tant il joue un rôle important sur les performances d'un système de traduction statistique. Nous avons fait le choix d'établir les poids des modèles de

conf, initiale	tm	lm	d	w	BLEU
Triggers $1-1$	$0,9$	$0,8$	$0,4$	-3	$\mathbf{31,02}$
Modèle IBM 3	$0,6$	$0,7$	$0,4$	-1	$29,23$

Tab. 5.1 – Score BLEU obtenu sur le corpus de développement de EUROPARL

façon empirique en testant un grand nombre de combinaisons de valeurs. Pour cela, nous avons fait varier les poids des modèles de traduction, de langage et de distorsion de 0 à 1 avec un pas de $0,1$ et la pénalité de longueur de -3 à 3 comme conseillé dans [Koehn 04]. Pour information,

Conf. Initiale	tm	lm	d	w	BLEU
Triggers $1-1$	$0,6$	$0,3$	$0,3$	0	$\mathbf{12,49}$
Modèle IBM 3	$0,8$	$0,6$	$0,6$	0	$12,39$

Tab. 5.2 – Score BLEU obtenu sur le corpus de développement de SSTITRES

nous donnons ces paramètres optimaux pour chacune des tables de traduction dans les tableaux 5.1 et 5.2 respectivement pour le corpus EUROPARL et le corpus SSTITRES, ainsi que les performances obtenues sur le corpus de développement pour les tables de traduction basées sur les triggers inter-langues et pour la table de traduction extraite à partir du modèle 3 d'IBM. Dans ces tableaux, les paramètres de PHARAOH sont notés : *tm* pour le poids de la table de

traduction, *lm* pour le poids du modèle de langage, *d* pour le poids du modèle de distorsion, et *w* pour la pénalité sur la longueur de la phrase cible générée.

5.3.4 Validation sur le corpus de test

Après l'étape d'optimisation des différents paramètres des tables de traduction basées sur les triggers inter-langues, mais aussi ceux du décodeur PHARAOH, une dernière étape de validation sur un corpus de test est nécessaire pour tirer les premières conclusions concernant notre approche.
Le tableau 5.3 présente les performances du décodeur PHARAOH sur le jeu de 500 phrases que nous avons choisi pour constituer le corpus de test d'EUROPARL. Les traductions automatiques générées à l'aide de notre table de traduction basée sur les triggers inter-langues obtiennent un score BLEU supérieur aux traductions produites avec la table du modèle 3 d'IBM. Toutefois, cette différence n'est pas statistiquement significative selon le test t ($t(48) = 0,664, p = 0,510$). PHARAOH produit donc des traductions de qualité similaire , en terme de score BLEU, quelle que soit la table utilisée dans le processus de décodage.
Nous aboutissons à la même conclusion concernant les expériences menées sur le corpus SS-

	Triggers 1-1	Modèle IBM 3
Test(500)	30,97	29,57

Tab. 5.3 – Scores BLEU obtenu sur le corpus de test de EUROPARL

TITRES. Le tableau 5.3 indique, de la même façon, le score BLEU associé aux traductions automatiques fournies par PHARAOH sur un corpus de test de 250 phrases avec notre table de traduction et avec la table de traduction du modèle 3 d'IBM. Dans ce contexte, les traductions produites avec la table du modèle 3 d'IBM obtiennent un meilleur score BLEU que celles produites avec notre table de traduction basée sur les triggers inter-langues. Toutefois, cette

	Triggers 1-1	Modèle IBM 3
Test(250)	12,04	12,34

Tab. 5.4 – Scores BLEU obtenu sur le corpus de test de SSTITRES

différence n'est pas statistiquement significative. Ainsi, tout comme lors des expériences menées sur EUROPARL, les deux tables de traductions permettent chacunes au décodeur PHARAOH de produire des traductions de qualité équivalente en terme de score BLEU.

5.4 Discussion

Nous avons vu dans le chapitre précédent que les triggers inter-langues permettent d'établir de bonnes correspondances lexicales entre les mots du Français et de l'Anglais. Pour le montrer, nous les avons utilisés pour construire un dictionnaire bilingue de mots qui s'est avéré de bonne qualité d'après les expériences menées.
Dans ce chapitre, nous avons souhaité tester la qualité des correspondances lexicales mises en évidence par les triggers inter-langues dans un contexte de traduction automatique statistique par mot. Est-ce qu'avec une table de traduction de mots construite à partir des triggers, un

décodeur est capable de traduire un texte source au moins aussi bien qu'il le ferait avec une table de traduction de mots apprise avec les modèles état-de-l'art proposés par Brown [Brown 93] ? Nous avons montré que c'est effectivement le cas sur deux corpus différents : le corpus EUROPARL classiquement utilisé par la communauté de recherche en Traduction Automatique et le corpus SSTITRES, plus difficile et moins courant, construit à partir des sous-titres de films. En effet, les traductions automatiques produites par le décodeur PHARAOH avec la table de traduction apprise selon le modèle 3 d'IBM obtiennent un score BLEU similaire à celui des traductions produites par le même décodeur avec une table de traduction mise en place à partir d'une liste de triggers inter-langues automatiquement sélectionnés. A score BLEU équivalent, notre méthode présente l'avantage d'être simple et de permettre la construction très rapide d'une table de traduction sans prendre en compte l'alignement des mots au sein du corpus d'apprentissage. A contrario, les modèles d'IBM sont entraînés suite à un processus particulier : il faut d'abord entraîner le modèle 1, qui sert d'initialisation à l'entraînement du modèle 2, et ainsi de suite. Par ailleurs, chaque modèle demande plusieurs itérations sur le corpus d'apprentissage et l'estimation de la table de traduction dépend de l'estimation du modèle d'alignement qui définit des probabilités d'alignement entre les mots d'une phrase source et les mots d'une phrase cible. Les triggers inter-langues, quant à eux, ne demandent qu'un passage sur le corpus d'apprentissage suivi d'une normalisation pour générer une table de traduction. Notre approche est donc plus simple à mettre en œuvre et elle permet au décodeur PHARAOH d'obtenir des performances légèrement meilleures, même si cette différence n'est pas significative, sur le corpus EUROPARL que celles obtenues avec la table de traduction issue du modèle IBM 3 plus complexe et plus longue à mettre en place.

Ces premiers résultats dans un contexte de Traduction Automatique démontrent, une fois de plus, le potentiel des triggers inter-langues. Nous nous sommes jusqu'à présent cantonnés à n'établir des correspondances lexicales qu'au niveau du mot. Dans le chapitre suivant, nous exploitons le formalisme des triggers inter-langues pour mettre en évidence les correspondances lexicales entre des suites de mots et ainsi nous placer dans le contexte de traduction automatique statistique le plus performant : la traduction par séquence de mots.

Chapitre 6

La traduction par séquence de mots : les triggers m-n

6.1 Introduction

La grande majorité des systèmes de TA statistique actuels utilisent comme unité de traduction, non plus un mot, mais un groupe de mots appelé communément une séquence ou encore un segment. Ainsi, les correspondances lexicales établies au sein de la table de traduction ne sont plus considérées seulement entre des mots mais aussi entre les groupes de mots.
Ce recours aux séquences permet notamment de tenir compte du contexte local des mots et ainsi désambiguïser les mots ayant plusieurs sens, mais cela évite également de nombreux problèmes liés à l'ordre des mots. Comme nous l'avons évoqué dans la section 2.4.3.1, il existe plusieurs méthodes permettant l'extraction d'une table de traduction de séquences à partir de corpus bilingue aligné. La plupart d'entre elles applique de simples heuristiques à partir de l'alignement des mots du corpus.
Dans ce chapitre, nous proposons une nouvelle méthode d'extraction d'une table de traduction de séquences s'appuyant sur le concept de triggers inter-langues et ne nécessitant pas d'alignement préalable des mots au sein du corpus d'apprentissage.
Dans ce qui suit, nous généralisons le concept de triggers inter-langues et nous l'étendons aux séquences de mots. Ainsi, nous ne nous bornons plus à la relation "1 mot déclenche 1 mot" mais "m mots déclenchent n mots" d'où l'appellation triggers m-n. Après avoir défini le concept de triggers inter-langues de séquences, nous commençons dans un premier temps par étudier les correspondances lexicales de séquences de mots qu'ils permettent de détecter. Nous proposons, ensuite, un nouvel algorithme tirant parti des triggers inter-langues afin de construire une table de traduction de séquences. La dernière section de ce chapitre est dédiée à l'évaluation des tables ainsi construites ainsi qu'à une étude comparative avec une table de traduction état-de-l'art établie à partir de l'alignement des mots du corpus d'apprentissage.

6.2 Les triggers inter-langues de séquences

6.2.1 Définition

Comme nous l'avons présenté dans le chapitre 4, les triggers inter-langues sont sélectionnés selon la valeur de l'Information Mutuelle (IM) entre deux variables aléatoires $\mathcal{X}$, $\mathcal{Y}$:

$$IM(\mathcal{X}=x,\mathcal{Y}=y) = P((\mathcal{X}=x,\mathcal{Y}=y)log\frac{P((\mathcal{X}=x,\mathcal{Y}=y)}{P((\mathcal{X}=x)P(\mathcal{Y}=y)} \tag{6.1}$$

Dans le cas des triggers de mots (ou triggers 1-1 comme nous les avons appelés), x est un mot de la langue source et y un mot de la langue cible. Nous avons montré dans les chapitres précédents que les triggers 1-1 permettaient d'établir des correspondances lexicales de mots de bonne qualité.
Au vu de son potentiel à établir de bonnes correspondances lexicales de mots, nous proposons de généraliser le concept de triggers inter-langues et de l'étendre aux séquences de mots. Pour ce faire, nous ne nous bornons plus à ce que x et y soient des mots appartenant aux vocabulaires, mais au contraire ils peuvent tout aussi bien être une séquence de mots apparaissant dans le corpus d'apprentissage. Ainsi, pour chaque séquence source d'au moins un mot et d'au plus m mots, nous calculons son Information Mutuelle avec chaque séquence cible d'au moins un mot et d'au plus n. Nous retenons alors comme triggers inter-langues, les k séquences cibles ayant la meilleure IM avec la séquence source. L'ensemble des k triggers m-n d'une séquence de mots f est noté Trig-m-n$_k(f)$.

Dans la section suivante, nous menons une première étude sur les couples de séquences ainsi révélés par les triggers de séquences.

6.2.2 Etude préalable

L'Information Mutuelle est une mesure de corrélation entre les mots. Dans le cas des triggers 1-1, nous avons vu qu'elle permettait de sélectionner des correspondances lexicales équivalentes à des traductions. Mais cette mesure présente également le désavantage de mettre en avant des associations dites indirectes entre les mots qui ne sont pas des traductions. C'est le cas, comme nous l'avons vu, pour le mot français " monsieur " dont le premier trigger inter-langue est " mr " et le deuxième " president ". Si le premier constitue bien une traduction du mot français, en revanche le deuxième est qualifié d'association indirecte dû au fait de rencontrer très souvent " mr " suivi de " president " dans le corpus EUROPARL (cf. figure 4.3). Calculer l'information mutuelle entre des groupes de mots pour sélectionner des triggers de séquences permet de limiter ce problème liée aux associations indirectes comme le montre le tableau 6.1. Il indique les 10 meilleurs triggers inter-langues de la séquence française *monsieur le président* parmi les séquences anglaises d'au moins un mot et d'au plus trois. Nous remarquons que *monsieur le président* est plus fortement corrélé avec le groupe de mots *mr president* en tant qu'unité qu'avec chacun de ses constituants pris séparément. Dans ce cas présent, l'utilisation des séquences règle également le problème lié au fait que le mot *le* de la séquence n'est pas traduit en Anglais.
Comme il a été dit, un des avantages que présente les séquences par rapport aux mots pour la traduction automatique est qu'elles capturent le contexte. Par conséquent, lorsqu'un mot possède plusieurs traductions, le problème de choisir celle qui convient ne se pose pas, puisque le mot est considéré avec son contexte et non plus seul. Le tableau 6.2 présente les correspondances lexicales proposées par les triggers inter-langues pour le mot *pièces* considéré seul puis au sein de

f	e ∈ Trig-m-n$_{k=10}(f)$	IM
monsieur le président	mr president	0,078
	president	0,068
	mr	0,061
	mr president I	0,025
	president I	0,023
	I	0,010
	mr president the	0,009
	president the	0,008
	like to	0,006
	like	0,006

TAB. 6.1 – Les 10 meilleurs triggers inter-langues de la séquence *monsieur le président*

f	e ∈ Trig-m-n$_k(f)$	IM
pièces	coins	37,52
	notes	9,51
	spare	9,39
	parts	8,78
pièces justificatives	receipts	0,29
pièces détachées	spare parts	0,58
	parts	0,40
	spare	0,39
pièces de monnaies	of coins	0,46
	number of coins	0,33
	national currency	0,28

TAB. 6.2 – Triggers de séquences sélectionnés sur le corpus EUROPARL

séquences. Seul il correspond le plus fortement avec le mot anglais *coins* et avec le mot *notes* qui signifie *billets* en français et qui relève ici d'une association dite indirecte. Pourtant, en dehors de son contexte, *pièces* peut aussi bien indiquer une pièce de monnaie, une pièce d'une maison, une pièce automobile ou encore une pièce d'identité par exemple. A chaque contexte y correspond une traduction différente. En procédant mot-à-mot, un système de traduction ferait difficilement la différence entre toutes ces possibilités. Si maintenant, le système dispose de correspondances lexicales au niveau de groupes de mots le problème est simplifié. D'après les triggers de séquences du tableau 6.2, établir les correspondances par groupe de mot enlève toute ambiguïté en cas de sens multiples. Ainsi *pièces justificatives* est le plus fortement associé à *receipts*, *pièces détachées* à *spare parts* et *pièces de monnaies* à *of coins*. Même si cette dernière correspondance lexicale ne constitue pas vraiment une traduction à cause du bruit généré par le mot outil *of*, elle permet toutefois de retrouver la bonne signification en anglais et de restituer le bon contexte de *pièces*. D'après l'étude que nous avons menée sur les corpus EUROPARL et SSTITRES, les triggers inter-langues de séquences semblent capables, tout comme les triggers inter-langues de mots, d'établir des correspondances lexicales de bonne qualité. Les tableaux 6.3 et 6.4 montrent des exemples de triggers inter-langues de séquences caractéristiques, extraits des corpus EURO-

PARL et SSTITRES. La première colonne montre les suites de mots déclenchentes en français. La deuxième colonne montre les meilleures séquences anglaises déclenchées correspondantes en utilisant les triggers inter-langues. Nous avons comparé les couples d'équivalences proposés par les triggers inter-langues, avec ceux extraits à partir de l'alignement des mots, comme proposé par [Koehn 03]. La troisième colonne liste les séquences les plus probables obtenues par cette méthode de l'état de l'art appelée la méthode " Référence ". Pour chaque case du tableau, les traductions sont données de la plus pertinente à la moins pertinente selon le critère utilisé (l'Information Mutuelle pour les triggers de séquences, les probabilités pour la méthode de référence).

f	e $\in$ Trig-m-n$_{k=3}(f)$	Référence
allons y	let's go 's go let's	let's go let's let
allumer	light to turn on turn on	turn
bonjour	hi hello good morning	hi good good morning
bonne nuit	good night good night	good night good good night buddy
ça dépend	it depends depends to her	schedule I will check schedule I will it depends
calme toi	calm down calm down	I promise
d'accord	ok o. k. all right	okay I o. k. stand over o.k. stand
passé	happened what happened past	get get pass get pass my
petit déjeuner	breakfast to breakfast say breakfast	breakfast sunday
porter plainte	press charges charges press	but you

TAB. 6.3 – Comparaison des traductions obtenues à partir du corpus SSTITRES avec les triggers de séquences et avec la méthode référence proposée par Koehn

Une analyse qualitative montre que les séquences déclenchées que nous obtenons sont souvent pertinentes en terme de traduction. Par ailleurs, les triggers permettent de retrouver les sens dif-

férents d'un même mot, comme c'est le cas pour le mot " allumer " (tableau 6.3). Ils restituent également correctement les synonymes (voir par exemple " attention particulière" ou " bonnes expériences " dans le tableau 6.4). Enfin, lorsqu'une suite de mots est traduite en plusieurs mots, les triggers inter-langues ont tendance à donner un meilleur score à la séquence de traduction entière plutôt qu'à ses sous-parties (par exemple, voir "porter plainte" dans le tableau 6.4).
La méthode que nous utilisons comme point de comparaison (méthode référence), n'extrait les correspondances de séquences que si elles sont consistentes avec l'alignement des mots. L'alignement des mots est calculé sur le corpus d'apprentissage par l'outil Giza++. Il s'agit d'un alignement automatique, par conséquent il peut être erroné ou incomplet. Dans ce cas, il est possible de manquer certaines correspondances. Le corpus EUROPARL dispose de données importantes et l'alignement des mots calculé sur le corpus d'apprentissage conduit à des couples de correspondance de séquences proches voire même quelquefois semblables aux couples révélés par les triggers inter-langues (cf. tableau 6.4). Ce n'est pas le cas sur le corpus SSTITRES. La tâche d'alignement des mots y est très difficile étant donné le peu de données ainsi que leur complexité. Si les mots ne sont pas correctement alignés, cela peut engendrer des couples de séquences surprenants comme par exemple (porter plainte, but you) ou encore (calme toi, I promise) dans le tableau 6.3. Les triggers inter-langues de séquences ne se basent pas sur l'alignement des mots et sont donc moins sensibles au manque de données. Ainsi, ils sont capables de trouver des couples que l'alignement des mots n'a pas été en mesure de révéler, notamment (porter plainte, press charges) ou encore (calme toi, calm down).

Tout comme avec les triggers de mots, nous avons testé le potentiel des triggers de séquences dans un cadre de traduction automatique d'un texte dans le sens Anglais Français. Dans les sections suivantes, nous expliquons comment nous en avons tiré profit pour établir une table de traduction de séquences nécessaire au décodeur PHARAOH pour produire automatiquement la traduction de corpus Anglais. Nous verrons également comment nous avons optimisé cette table en nous aidant de l'algorithme du Recuit Simulé. Nous comparerons notre méthode, en terme de qualité de traductions produites, avec celle proposée par Koehn [Koehn 03] et qui fait état de référence dans les campagnes d'évaluation WMT.

6.3 Modèles de traduction de séquences : les triggers m-n

Nous proposons dans cette section une nouvelle approche pour constuire une table de traduction de séquences à partir d'un corpus bilingue aligné sans passer par l'alignement des mots mais en utilisant seulement les triggers inter-langues.
L'algorithme que nous proposons est le suivant :

1. Segmentation du corpus source en séquences pertinentes. De cette façon, nous sélectionnons au préalable les séquences sources qui seront présentes dans notre table de traduction et ce dans le but de constituer une table de traduction propre et de taille raisonnable.
2. Sélection des k meilleurs triggers-langues m-n pour chaque séquence source. Nous appelerons l'ensemble des couples ainsi constitué les triggers m-n, en opposition aux triggers 1-1 qui n'établissent que des correspondances lexicales au niveau du mot.
3. Sélection, parmi l'ensemble des triggers m-n ainsi constitué, des traductions potentielles à intégrer dans la table de traduction de séquences. Nous utiliserons pour cela l'algorithme du Recuit Simulé.

Nous détaillons ci-dessous chaque étape de notre algorithme d'extraction d'une table de traduction de séquences. Le reste du chapitre est dédié à la phase d'expérimentation. Nous verrons

f	e ∈ Trig-m-n$_{k=3}(f)$	Référence
abaisser	lower to lower lowering	lower reduce reducing
attention particulière	attention particular attention special attention	attention emphasis account
atterrir	land to land land at	deciding go land
bonnes expériences	good experience positive experiences experience	good experience good record good reputation
croix rouge	red cross cross red	red cross red cross is red cross movement
danger	danger risk the danger	danger risk dangerous
date butoir	deadline date the deadline	deadline target date end-date
excusez moi	sorry am sorry I am sorry	excuse me I am sorry forgive me
merci	thank you thank you	thank you thank you very much thank
traverser	cross to cross through	cross through now

Tab. 6.4 – Comparaison des traductions obtenues à partir du corpus EUROPARL avec les triggers de séquence et avec la méthode référence proposée par Koehn

notamment les performances de notre table de traduction de séquences en terme de qualité des traductions automatiques en score BLEU sur un corpus de test. Elles seront enfin comparées aux performances d'une table de traduction extraite à partir d'heuristiques sur l'alignement des mots.

6.3.1 Segmentation du corpus source

La plupart des méthodes utilisant les séquences en TA nécessitent que les corpus d'apprentissage soient alignés au niveau du mot. Elles font donc appel aux modèles IBM dont l'apprentissage requiert un temps de calcul et une capacité de mémoire considérables. Koehn *et al.* [Koehn 03], par exemple, extraient toutes les paires de séquences qui sont en adéquation avec la combination des deux alignements de mots Source-Cible et Cible-Source établis grâce aux modèles IBM. Cette

méthode repose sur de simples heuristiques appliquées à l'alignement des mots et de ce fait, toute suite de mots sources contigus peut être la traduction de toutes suites de mots cibles contigus. Les paires de traduction ainsi obtenues ne sont pas toujours naturelles et peuvent conduire à des données bruitées. De plus, les séquences sources comme cibles ne sont pas toujours linguistiquement justifiées comme le montre le tableau 6.5 qui présente un extrait de la table de traduction obtenue selon la méthode proposée par Koehn [Koehn 03] sur le corpus d'apprentissage EUROPARL. Dans un processus naturel de traduction d'une phrase française, il nous paraît peu vraisemblable

!	|||	mesdames et messieurs les députés
) according to which	|||	d après lequel
. for	|||	. je voudrais pour
. for	|||	. pendant des
: what	|||	. quels
: what	|||	: " et
: what	|||	: " qu en
a council	|||	"

TAB. 6.5 – Extrait d'une table de traduction de séquences

de choisir le groupe de mots *. je voudrais pour* comme unité à traduire en un seul bloc. Un être humain aurait en effet plutôt tendance à segmenter la phrase en groupe syntagmatique.
L'extraction de paires de séquences à partir de l'alignement des mots engendrent de nombreuses paires qu'il ne semble pas très naturel d'utiliser pour produire une traduction. Les travaux de Zettlemoyer [Zettlemoyer 07] ont d'ailleurs révélé qu'il était possible de réduire de moitié le nombre de couples de séquences de la table de traduction ainsi extraite à partir de l'alignement des mots sans pour autant engendrer une baisse de qualité dans les traductions automatiques produites par un décodeur.

Afin de ne pas polluer notre table de traduction avec des séquences peu naturelles, nous proposons de segmenter au préalable la partie source du corpus d'apprentissage indépendamment de la partie cible. De ce fait, nous sélectionnons les séquences sources qui seront présentes dans notre table de traduction totalement indépendamment de la partie cible. Les triggers inter-langues nous permettront par la suite d'identifier les traductions des séquences sources sélectionnées. En sélectionnant les séquences sources indépendamment du corpus cible et par conséquent indépendamment de l'alignement des mots, nous pourrons ainsi éviter de trouver des paires de traductions bruitées comme celles présentes dans le tableau 6.5.

Pour sélectionner les séquences sources sur le corpus d'apprentissage, nous utilisons une méthode développée il y a quelques années dans notre équipe [Zitouni 03]. Il s'agit d'un processus itératif qui sélectionne les séquences en groupant les unités dont l'Information Mutuelle est forte. A la première itération, une unité est un mot, aux itérations suivantes une unité est soit un mot soit un groupe de mots. Seules les groupes de mots qui améliorent la perplexité sur le corpus de développement sont fusionnés en unité pour les itérations suivantes. Le procédé est répété jusqu'à convergence de la perplexité. Il aboutit à une liste de séquences de mots et à un corpus réécrit en termes de séquences.
Une fois la partie source du corpus d'apprentissage segmentée, nous proposons ensuite d'identifier les traductions potentielles de chaque séquence source en utilisant les triggers inter-langues.

6.3.2 Apprentissage des triggers m-n

La partie source de notre corpus parallèle est maintenant écrite en termes de séquences. La question qui se pose à nous est comment trouver la traduction de ces séquences. Nous proposons pour cela d'utiliser les triggers inter-langues.
Nous avons vu dans les sections précédentes que les triggers permettent grâce à l'Information Mutuelle de mettre en corrélation des unités sources et des unités cibles. Pour les triggers 1-1, les unités sont simplement des mots, mais rien ne nous empêche de considérer une séquence de mots comme une unité. Ainsi nous pouvons associer chaque séquence source extraite à une suite de un ou plusieurs mots cibles contigus. Nous parlons alors de triggers m-n puisque m mots sources sont associés à n mots cibles.
Nous supposons par la suite que le nombre de mots de la traduction d'une séquence source dépend de la taille de cette séquence. De ce fait, nous assumons que chaque séquence source de m mots peut être traduite par une séquence de j mots cibles contigus avec $j \in [m - \Delta m, m + \Delta m]$. Nous permettons par conséquent à une séquence source d'être traduite par des séquences cibles de tailles différentes dépendantes de la taille de la séquence source. Le tableau 6.6 présente un exemple des triggers inter-langues retenues pour la séquence française *porter plainte*. Dans nos

	triggers m-n		
séquence	2-1	2-2	2-3
porter plainte	press	press charges	can press charges
	charges	can press	not press charges
	easy	not press	you can press

Tab. 6.6 – Triggers inter-langues de la séquence *porter plainte*

travaux, nous admettons que pour les séquences courtes, la valeur de Δm est de 1. Ainsi, dans l'exemple cité, la séquence *porter plainte* est associée à l'aide des triggers inter-langues, à des suites d'au moins 1 mot et d'au plus 3. Pour une taille donnée, nous avons sélectionné les 3 meilleurs triggers inter-langues, ce qui fait un total de 9 traductions potentielles. Cependant, une seule est correcte : *press charges*.
Dans un cas plus général, chaque séquence source est associée à k traductions potentielles, k augmentant avec la taille de la séquence. C'est pourquoi, nous proposons de sélectionner les traductions pertinentes et d'écarter celles qui introduisent des erreurs. Dans ce qui suit, nous appelons triggers m-n l'ensemble des séquences sources associées à leurs traductions potentielles de tailles variables. La question maintenant est comment sélectionner parmi l'ensemble des triggers m-n, les meilleures paires de traductions, c'est-à-dire celles qui conduiront aux meilleures traductions automatiques.
Le problème auquel nous sommes confrontés est clairement un problème d'optimisation. Il nous faut trouver un sous-ensemble des triggers m-n qui conduira à des traductions automatiques de score BLEU optimal. En d'autres termes, partant d'une table de traduction initiale et *a fortiori* d'un score BLEU initial, notre objectif est de modifier cette table de traduction afin d'atteindre un score BLEU optimal. Nous utilisons comme table de traduction initiale, notre table de traduction de mots Triggers 1-1. Le score BLEU associé aux traductions automatiques produites avec cette table constitue le score initial à optimiser. Nous proposons de modifier cette table de traduction de mots initiale en y injectant aléatoirement des paires de séquences extraites de l'ensemble des triggers m-n jusqu'à parvenir à une table de traduction composée de paires de mots et de paires de séquences et conduisant à un score BLEU optimal. Pour mener à bien ce

processus d'otimisation, nous utilisons l'algorithme du Recuit Simulé que nous adaptons à notre problématique comme cela est expliqué dans la section suivante.

6.3.3 Optimisation par Recuit Simulé

Le Recuit Simulé (RS) est une technique inspirée d'un processus physique utilisé en métallurgie alternant des cycles de refroidissement lent et de réchauffage (recuit) qui tendent à minimiser l'énergie d'un matériau. L'algorithme du RS s'inspire donc de la thermodynamique. Il est employé pour trouver une solution optimale à un problème difficilement résolvable en utilisant des méthodes combinatoires. L'avantage de cette approche est qu'elle permet d'éviter la contrainte d'explosion combinatoire tout en réglant le problème d'optimum local rencontré par d'autres algorithmes d'optimisation. En effet, sous certaines conditions l'algorithme du Recuit Simulé converge vers un optimum global [Aarts 85].
Notre problème peut être résolu grâce à cet algorithme. En effet, nous avons un ensemble de triggers m-n constituant des correspondances lexicales de séquences et nous souhaitons en intégrer un sous-ensemble à notre table de traduction de mots, pour améliorer la qualité de traduction en terme de score BLEU. Naturellement, compte tenu du temps nécessaire à un décodeur pour traduire un corpus source, il est tout à fait déraisonnable de tester une par une chaque combinaison possible parmi tous les triggers m-n. C'est pour cette raison que nous proposons d'utiliser l'algorithme du RS qui va nous permettre de sélectionner une combinaison pertinente.
Pour cela, nous commençons avec un système de traduction dont la table de traduction ne contient que des traductions de mots (triggers 1-1). Nous évaluons ses performances en termes de score BLEU sur un corpus de développement. Ensuite, nous ajoutons de manière aléatoire à la table de traductions de mots un sous-ensemble des triggers m-n. Le processus est répété jusqu'à obtenir un score BLEU optimal. L'algorithme que nous suivons est le suivant :

Algorithme 1 : Algorithme du Recuit Simulé

```
Commencer avec une température T élevée
tant que La température finale n'est pas atteinte ou que l'équilibre du système n'est pas
atteint faire
    tant que L'équilibre du système n'est pas atteint faire
        Perturber le système d'énergie E_i de l'état i à l'état j
        Calculer la nouvelle énergie du système E_j
        si E_j − E_i >= 0 alors
            Accepter l'état j
        sinon
            L'état j est accepté avec une probabilité random(P) < e^((E_i−E_j)/T) with
            P ∈ [0 − 1]
        finsi
    fintq
    Diminuer la température
fintq
```

Il est nécessaire de définir chaque paramètre de l'algorithme de façon à pouvoir l'adapter à notre problème de traduction.

La température

Il s'agit d'un paramètre très important dans l'algorithme du Recuit Simulé qu'il est nécessaire de calibrer au mieux. La température contrôle tout le processus d'optimisation. A température élevée, tous les changements d'état ont approximativement la même probabilité d'être acceptés. Au contraire, à température basse les changements d'état qui détériorent la fonction à optimiser ont une faible probabilité d'être acceptés. Enfin, lorsque la température est nulle, aucune dégradation de la fonction à optimiser n'est permise. La température initiale se doit donc d'être élevée afin de prospecter un maximum de configuration contenant potentiellement la meilleure. Elle diminue ensuite progressivement par palier au cours des itérations de l'algorithme.

La configuration initiale

Dans notre adaptation du Recuit Simulé, les états correspondent à des systèmes de traduction Automatique. Tous ces sytèmes utilisent le décodeur PHARAOH avec en paramètres d'entrée le même corpus à traduire et le même modèle de langage trigramme. Ils différent sur la table de traduction utilisée. Le système initiale repose sur une table de traduction de mots. Nous utilisons pour cela la table de traduction Triggers 1-1 optimale déterminée dans le chapitre précédent.

Perturbation du système

Perturber le système consiste à passer d'un état i à un état j. Dans notre cas, changer d'état se ramène à modifier la table de traduction de notre système. Pour cela, nous nous proposons d'ajouter aléatoirement à la table de traduction de notre système à l'état i un sous-ensemble des triggers m-n préalablement déterminés pour aboutir à la table de traduction de notre système pour l'état j. A chaque pertubation de notre système, un processus complet de décodage est lancé sur le corpus de développement afin d'évaluer les performances du nouveau système.

Energie du système

L'énergie du système est le critère à optimiser comme convenu dans l'algorithme originel du Recuit Simulé. En ce qui nous concerne, le critère que nous souhaitons optimiser est le score BLEU associé aux traductions produites par le décodeur PHARAOH sur le corpus de développement. Dans notre cas, il nous faut donc maximiser l'énergie pour atteindre au final un score BLEU optimal.

Etat d'équilibre

L'état d'équilibre est atteint lorsque l'énergie du système franchit un palier ou n'évolue plus. Pour déterminer la taille d'un palier, il est possible de fixer au préalable le nombre de changement d'états à effectuer au cours d'un palier. Dans notre cas, nous avons choisi de franchir un palier lorsque le score BLEU associé aux traductions automatiques n'évolue plus entre deux états. A chaque état d'équilibre, la température est baissée afin de resteindre de plus en plus les configurations possibles et de converger vers une solution optimale.

Le critère d'arrêt

Il est inutile de poursuivre le processus de perturbation du système lorsque la température est trop basse, ce qui signifie que le taux d'acceptation de nouveaux états devient très faible et lorsque l'énergie n'évolue plus. Nous concernant, le processus d'ajout aléatoire des triggers m-n

stoppe lorsque le score BLEU converge ou lorsque la température descend sous un certain seuil qu'il sera nécessaire de déterminer.

L'algorithme du Recuit Simulé, comme nous l'avons dit, permet dans certaines conditions d'atteindre un extrema global. Le fait de transiter d'un état à un autre un nombre important de fois permet en effet au système de converger vers un état stable. Selon notre problèmatique, la fonction à maximiser est le score BLEU des traductions automatiques produites à chaque changement d'état par le décodeur PHARAOH. Rappelons qu'un changement d'état dans notre cas consiste à modifier la table de traduction en injectant aléatoirement des correspondances lexicales tirées de l'ensemble des triggers m-n. L'algorithme du Recuit Simulé nous garantit d'aboutir à un score BLEU au minimum équivalent à celui de l'état initial, à savoir celui associé aux traductions produites avec la table de traduction de mots Triggers 1-1. Au terme du processus, seuls les triggers m-n qui ont conduit à améliorer le score BLEU sur le corpus de développement sont conservés dans la table de traduction finale.

La section suivante présente les expérimentations qui ont été menées afin de valider notre approche.

6.4 Evaluation des modèles de traduction de séquences

Afin d'évaluer la table de traduction de séquences mise en place selon l'approche que nous avons proposée à la section précédente, nous nous plaçons comme précédemment dans un contexte de traduction Anglais vers Français. Les tests ont été réalisés parallèlement sur le corpus EUROPARL et sur le corpus SSTITRES. Nous décrivons dans ce qui suit les données expérimentales que nous avons utilisées à chaque étape de notre approche.

6.4.1 Construction de l'ensemble des triggers m-n

Détection des séquences françaises La première étape pour la mise en place de notre table de traduction de séquences est la sélection au préalable des N séquences sources (françaises dans notre cas) de notre table de traduction. Pour cela nous segmentons le corpus source en utilisant la méthode décrite section 6.3.1. Nous avons ainsi extrait du corpus EUROPARL un ensemble de $N = 11212$ séquences françaises de taille $m \in 2, 3$. De même, nous avons extrait du corpus SSTITRES $N = 15860$ séquences de taille $m \in 2, 3$.

Sélection des correspondances lexicales Pour chaque séquence française f_j de longueur $m \in 1, 2, 3$ (ce qui inclut les séquences françaises retenues après segmentation du corpus français mais aussi les mots simples du vocabulaire), nous sélectionnons un ensemble $Trig_{m-n_k}$ de k triggers inter-langues. Dans nos expérimentations, nous définissons la taille n des séquences anglaises comme appartenant à l'intervalle $[\max(0, m - \Delta m), \min(3, m + \Delta m)]$. Pour chaque taille de séquences anglaises possible, nous sélectionnons les $l = 10$ meilleurs triggers inter-langues de f_j. Ainsi, pour une séquence française de longueur $m = 2$, et $\Delta m = 1$, nous obtenons un total de $k = 30$ séquences anglaises comme étant des triggers inter-langues : 10 de longueur 1, 10 de longueur 2 et enfin 10 de longueur 3. Nous constituons ainsi l'ensemble des triggers m-n. Parmi ces triggers candidats au statut de traduction, nous devons maintenant en sélectionner un sous-ensemble qui sera intégré à la table de traduction de séquence. Cette sélection est opérée par l'algorithme du Recuit Simulé.

6.4.2 Mise en place du Recuit Simulé

configuration initiale Nous prenons comme état initial pour le Recuit Simulé, la configuration du système de traduction par mots, utilisant le décodeur PHARAOH ainsi que la table de traduction fondée sur les triggers 1-1 qui nous a conduit à la meilleure qualité de traduction en terme de score BLEU dans les expériences menées jusqu'ici. Sur le corpus EUROPARL, il s'agissait de la table de traduction dans laquelle chaque mot français est associé à ses $k = 100$ meilleures triggers inter-langues de mots anglais et sur le corpus SSTITRES, seuls les $k = 20$ meilleurs ont été retenus dans la table. Pour les paramètres du décodeur, nous reprenons les paramètres optimaux définis sur le corpus de développement et reportés dans le chapitre précédant. Pour information, nous redonnons ces paramètres dans les tableaux 6.7 et 6.8 respectivement pour le corpus EUROPARL et le corpus SSTITRES ainsi que les performances obtenues sur le corpus de développement. Dans ces tableaux, les paramètres de PHARAOH sont notés : *tm* pour le poids de la table de traduction, *lm* pour le poids du modèle de langage, *d* pour le poids du modèle de distorsion, et *w* pour la pénalité sur la longueur de la phrase cible générée.

conf, initiale	tm	lm	d	w	BLEU
Triggers $1-1$	$0,9$	$0,8$	$0,4$	-3	$\mathbf{31,02}$
Modèle IBM 3	$0,6$	$0,7$	$0,4$	-1	$29,23$

Tab. 6.7 – Score BLEU initial obtenu sur le corpus de développement de EUROPARL

Conf. Initiale	tm	lm	d	w	BLEU
Triggers $1-1$	$0,6$	$0,3$	$0,3$	0	$\mathbf{12,49}$
Modèle IBM 3	$0,8$	$0,6$	$0,6$	0	$12,39$

Tab. 6.8 – Score BLEU initial obtenu sur le corpus de développement de SSTITRES

La configuration initiale de l'algorithme du Recuit Simulé requiert également comme nous l'avons dit une valeur de température adéquate afin d'accepter au commencement un grand nombre de changements d'état. Nous avons réalisé plusieurs séries de tests pour paramétrer la valeur de la température initiale. Ces tests nous ont conduit à une valeur de $T_0 = 10^{-4}$.

Perturbation du système Plusieurs valeurs de paramètres ont été testées. Nous ne présentons dans cette section que ceux qui ont conduits aux meilleures performances. Pour passer d'un état du système à un autre, nous modifions la table de traduction utilisée par PHARAOH en y injectant des couples de traduction de séquences appartenant aux triggers m-n. Pour cela, 10 séquences françaises associées chacune à 10 de leurs triggers inter-langues de séquences sont sélectionnées au hasard parmi l'ensemble des triggers m-n. Cet ajout donne lieu à une nouvelle table de traduction et par conséquent à un nouvel état. Ce nouvel état engendre la production de nouvelles traductions automatiques par PHARAOH et associées à un nouveau score BLEU.

Equilibre thermique L'équilibre thermique pour une température donnée est atteint quand le score BLEU n'évolue plus. Nous assumons que le score BLEU n'évolue plus lorsqu'il ne varie pas de plus de 10^{-3} entre deux états. Lorsque l'équilibre thermique est atteint, la température doit être diminuée de façon très douce afin de relancer le processus en acceptant de moins en

moins de changements d'état ne causant pas d'amélioration du score BLEU. Nous diminuons la température de 4% chauqe fois que l'équilibre thermique et atteint. Le processus est ainsi relancé jusqu'à convergence complète du score BLEU ou lorsque la température descend sous le seuil de 10^{-8}.
L'évolution du SCORE BLEU sur les corpus de développement de EUROPARL et SSTITRES est présenté et étudié dans la section suivante.

6.4.3 Influence du Recuit Simulé sur les corpus de développement

Les figures 6.1 et 6.2 indiquent la progression du score BLEU associé aux traductions automatiques produites par PHARAOH tout au long du processus du Recuit Simulé à partir des corpus de développement de EUROPARL et SSTITRES. Ceci confirme le comportement attendu de

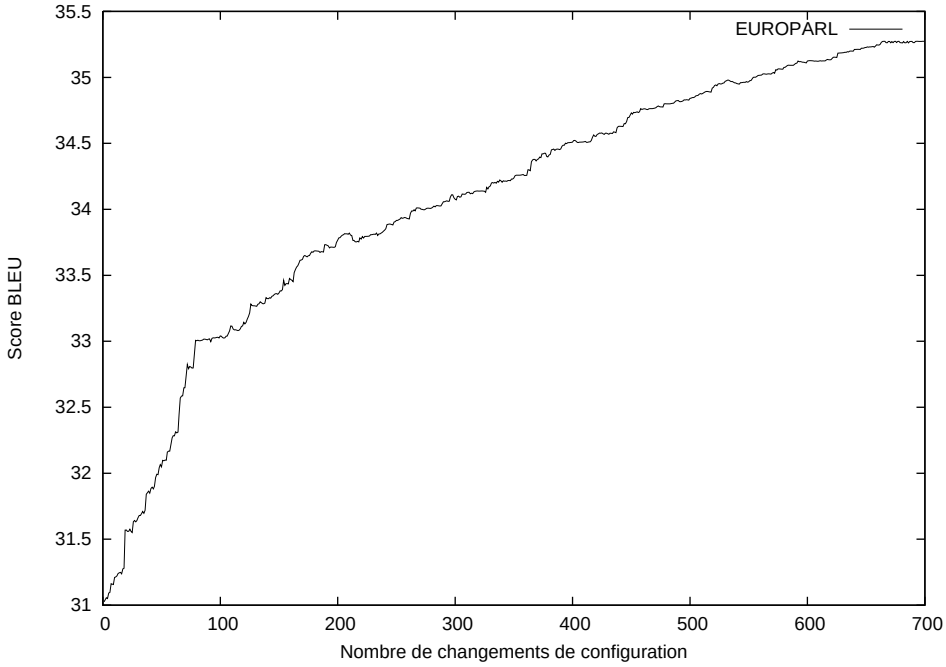

FIG. 6.1 – Évolution du score BLEU sur les corpus de développement EUROPARL

l'algorithme du recuit simulé : la fonction objectif augmente peu à peu en autorisant des baisses locales nécessaires pour ne pas tomber dans un maximum local.

Suite à ce processus, le système intégrant les traductions de séquences atteint un score BLEU de 35, 27 sur EUROPARL et 14, 14 sur SSTITRES. Sur les deux corpus, nos expériences montrent la validité de notre méthode de traduction à base de séquences qui améliorent les résultats de la méthode de base.

6.4.4 Validation sur les corpus de test

Après optimisation avec l'algorithme du Recuit Simulé, nous évaluons la table de traduction obtenue à la fin du Recuit Simulé sur un corpus de test. Nous comparons pour cela les performances de traduction de PHARAOH obtenues avec la table de traduction basée sur les triggers

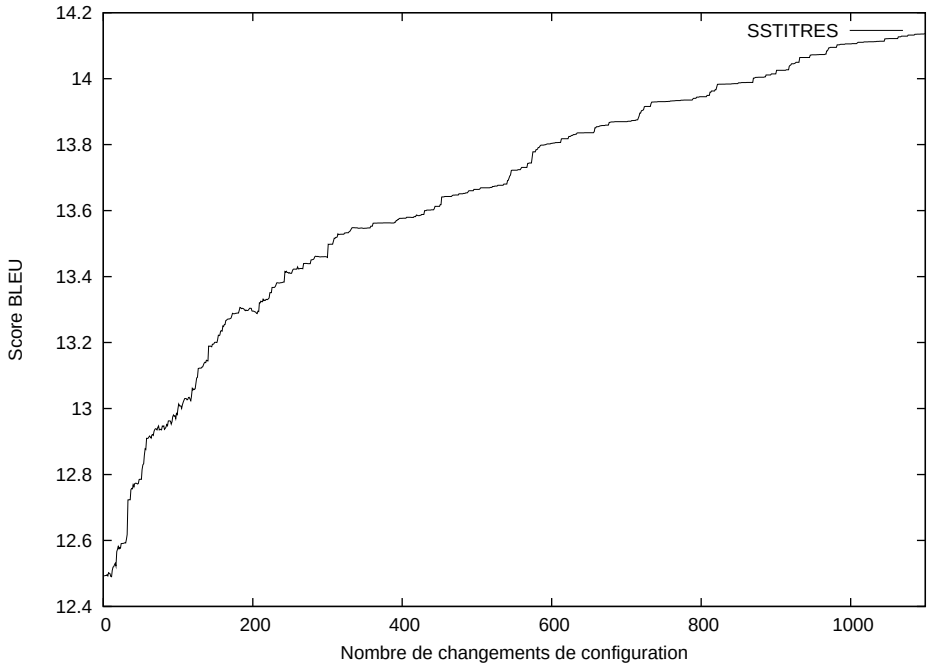

FIG. 6.2 – Évolution du score BLEU sur les corpus de développement SSTITRES

inter-langues avec celles obtenues avec une table de traduction construite à partir de l'alignement des mots. Les résultats sont présentés dans les tableaux 6.9 et 6.10. Dans chaque tableau,

	Triggers inter-langues		Etat de l'art	
	1-1	m-n	IBM3	Référence
Dev	31,02	35,27	29,23	35,07
Test	30,97	32,75	29,57	37,15

TAB. 6.9 – Évaluations des systèmes de traduction en terme de score BLEU sur les corpus de développement (Dev) et de test (Test) extraits de EUROPARL

la colonne "1-1" rappelle les performances obtenues avec la table de traductions de mots basée sur les triggers 1-1. La colonne "m-n"indique la qualité des traductions produites avec la table de traduction basée sur les triggers de séquences et optimisée avec l'algorithme du Recuit Simulé. En comparaison à notre approche, la colonne "IBM3" indique les performances obtenues avec la

	Triggers inter-langues		Etat de l'art	
	1-1	m-n	IBM3	Référence
Dev	12,49	14,14	12,39	7,65
Test	12,04	9,62	12,34	7,73

TAB. 6.10 – Évaluations des systèmes de traduction en terme de score BLEU sur les corpus de développement (Dev) et de test (Test) extraits de SSTITRES

table de traduction calculée à partir du modèle 3 d'IBM et la colonne "Référence" celles de la table de traduction générée à partir de l'alignement bi-directionnel des mots calculé sur le corpus d'apprentissage comme décrit dans [Koehn 03].

6.4.5 Analyse des premiers résultats

Sur le corpus EUROPARL, nous constatons que notre approche donne des résultats légèrement meilleurs que les méthodes IBM3 et Référence sur le corpus de développement. Ceci est vrai aussi bien pour une traduction mot-à-mot que pour une traduction par séquence. Par ailleurs, sur le corpus de développement comme sur le corpus de test, l'ajout de couples séquences dans la table de traduction conduit à un score BLEU plus élevé que celui obtenu avec une table de traduction de mots. Notre table de traduction de mots basée sur les triggers 1-1 conduit à des traductions automatiques de score BLEU légèrement supérieur au score de celles produites avec la table extraite du modèle IBM3. En revanche, la table de traduction de séquences "Référence" conduit à des traductions automatiques avec un score BLEU plus élevé comparé au score de celles produites avec la table de traduction basée sur les triggers m-n. Pour expliquer ce dernier résultat, rappelons que notre stratégie de sélection de séquences est beaucoup plus stricte que celle de référence, et nous en obtenons finalement beaucoup moins. En conséquence, nos séquences se retrouvent plus rarement dans le corpus de test, et leur impact positif est donc plus faible.
Sur le corpus SSTITRES, comme attendu, les performances sont plus faibles. Sur le corpus de développement, nous obtenons les mêmes conclusions que pour le corpus EUROPARL, sauf pour la table de traduction "Référence". En effet, l'utilisation de traductions de séquences construites suivant l'approche de Koehn, détériore la qualité des traductions automatiques produites par rapport à celles produites avec les tables de traduction de mots. En revanche, sur le corpus de test, les deux tables de traduction de séquences dégradent les performances même si notre méthode est plus robuste sur ce point. Notons que toutefois, à ce niveau de score BLEU, les traductions automatiques produites restent dans les deux cas incompréhensibles. Pour ce corpus, beaucoup plus diversifié que EUROPARL, l'effet de sur-apprentissage sur le corpus de développement se fait donc plus critique concernant la sélection par Recuit Simulé des meilleurs couples de séquences sur le corpus de développement. Notons tout de même que notre méthode aboutit à une dégradation moins forte par rapport à la table de traduction de Référence beaucoup plus sensible à la complexité du corpus SSTITRES. En effet, rappelons que l'extraction des couples de séquences dans la table de traduction Référence se fait à partir de l'alignement automatique des mots sur le corpus d'apprentissage. La tâche d'alignement des mots sur le corpus SSTITRES est rendue très difficile de par le peu de données ainsi que leur diversité. Par conséquent, la qualité de l'alignement des mots se fait ressentir au niveau de la table de traduction de séquences. Les séquences ainsi extraites, ne permettent pas au décodeur PHARAOH de produire des traductions de meilleure qualité que celles produites avec une table de traduction de mots.

Quel que soit le corpus envisagé, la différence de performances de PHARAOH est significative lorsqu'il s'agit d'utiliser une table de traduction de séquences construites avec les triggers m-n ou à partir de l'alignement des mots. Sur le corpus SSTITRES, cette différence s'explique par le manque de données qui pénalise davantage l'extraction des séquences par alignement des mots mais aussi par l'hétérogénéité des corpus. Toutefois, il est difficile de tirer des conclusions sur un corpus aussi peu fourni, si ce n'est que notre méthode de sélection de séquences par les triggers inter-langues semble plus robuste au manque de données du fait de ne pas utiliser d'alignement préalable des mots. Dans la section suivante, nous menons une étude comparative entre notre table de traduction construite à l'aide des triggers inter-langues et la table de traduction "Réfé-

rence" extraite à partir de l'alignement des mots afin d'expliquer les différences de score BLEU observées.

6.4.6 Etude comparative des tables de traduction

Dans les tests menés précédemment, nous avons observé une différence significative de $4,4$ points BLEU entre les traductions automatiques produites avec la table de traduction fondée sur les triggers inter-langues et celles obtenues avec une table de traduction extraite à partir d'heuristiques sur l'alignement des mots. Afin d'analyser et de comprendre cette différence et ainsi améliorer notre méthode, nous présentons dans cette section une étude comparative des deux tables de traduction relative aux nombres de couples contenus dans les tables mais aussi à la distribution des probabilités associées à ces couples.

6.4.6.1 Taille des tables de traductions

Le tableau 6.11 indique le nombre d'entrées dans les différentes tables de traductions utilisées jusqu'à maintenant dans nos expériences menées sur le corpus EUROPARL. Pour les tables de

Table de traduction	**Taille**
Triggers 1-1	5,18M
IBM3	1,81M
Triggers m-n RS	5,20M
Référence	21,14M

Tab. 6.11 – Taille des tables de traductions en nombre d'entrées extraites à partir du corpus d'apprentissage EUROPARL

traduction de mots, la table extraite à partir du modèle 3 d'IBM contient cinq fois moins d'entrées que notre table de traduction basée sur les triggers 1-1. Rappelons que pour chaque mot français du vocabulaire, nous gardons les 100 meilleurs triggers inter-langues comme traduction potentielle. En moyenne, dans la table notée IBM3, un mot français est associé à 23 traductions potentielles. Malgré cette différence notable dans la taille des tables, elles aboutissent toutes deux à des résultats similaires en terme de score BLEU sur la qualité des traductions produites par PHARAOH avec chacune d'entre elles.
Au niveau des tables de traduction de séquences cette fois, la table de traduction "Référence" extraite à partir de l'alignement des mots est beaucoup plus conséquente que notre table de traduction construite à partir des triggers m-n et de l'algorithme du Recuit Simulé (elle est notée "Triggers m-n RS"). Rappelons que selon notre méthode, nous démarrons le Recuit Simulé avec notre table de traduction de mots qui contient $5,18$ millions de couples de mots. Par conséquent, seulement $26K$ couples de séquences ont été introduits au cours du Recuit Simulé. Notre table de traduction couvre donc beaucoup moins de séquences que la table de traduction Référence. La colonne "Référence" du tableau 6.12 donne le détail des séquences françaises couvertes par la table de traduction construite à partir de l'alignement des mots. Elle compte au total plus de 13 millions de séquences françaises d'au moins deux mots. Dans notre cas, comme nous l'avons dit dans la section 6.4.1, nous n'en avons extrait que 11212 sur le corpus EUROPARL et une sous-partie seulement a été intégrée à la table de traduction avec l'algorithme du Recuit Simulé. Bien que l'ajout de ces séquences ait permis d'optimiser le score BLEU sur le corpus de développement, leur impact s'est avéré moindre sur le corpus de test du fait non seulement de leur

	Table de traduction		
Vocabulaire Fr.	Référence	Triggers m-n	Référence 2
unigramme	46K	56K	30K
bigramme	0,65M	0,20M	0,29M
trigramme	1,84M	0,38M	0,54M
>trigramme	10,53M	1,16M	1,54M

TAB. 6.12 – Nombre des séquences françaises couvertes par les tables de traductions de séquences

nombre limité mais aussi d'un effet de sur-apprentissage.
Afin d'améliorer l'effet de l'emploi des triggers m-n par rapport aux triggers 1-1, il convient donc dans un premier temps de sélectionner un nombre plus important de séquences françaises sur le corpus d'apprentissage. Pour cela, nous utilisons de nouveau l'algorithme de détection de séquences décrit à la section 6.3.1 afin de constituer une liste de séquences pertinentes. Nous ajoutons ensuite à cette liste les n-grammes les plus fréquents dans le corpus d'apprentissage. Nous segmentons enfin notre corpus source à l'aide de toutes ces séquences. La colonne "Triggers m-n " du tableau 6.12 indique le nombre de séquences désormais couvertes sur le corpus d'apprentissage. Nous couvrons un total de $1,81$ millions de séquences, soit seulement 14% du nombre de séquences couvertes par la table de traduction de Référence. Dans notre cas, la liste de séquences françaises est déterminée à partir de la seule partie française du corpus bilingue, indépendemment de la partie anglaise. Rappelons que chaque séquence est choisie à l'aide de l'Information Mutuelle entre les mots la constituant et que le corpus segmenté a une perplexité plus faible que le corpus non segmenté. Par conséquent, nos séquences sont donc plus pertinentes et constituent davantage des groupes de mots qu'un être humain choisirait comme bloc à traduire, c'est pourquoi elles sont beaucoup moins nombreuses que celles de la table "Référence". Cette dernière regroupe en effet toutes les paires de séquences consistantes avec l'alignement des mots, par conséquent les séquences françaises comme anglaises ne constituent pas toujours des groupes de mots qu'un humain choisirait de traduire en un seul et même bloc. Il peut même s'agir d'une suite de mots n'apparaissant que très rarement dans le corpus et ne constituant pas un groupe syntagmatique dans une phrase (cf. tableau 6.5 de la section 6.3.1). Souvent ces couples de séquences non pertinentes sont associées à une probabilité de 1 dans la table de traduction. La colonne "Référence 2" du tableau 6.12 indique la couverture des séquences françaises de la table de traduction de Référence lorsque tous les couples de traduction ayant une probabilité de 1 sont écartés. Le nombre de séquences couvertes diminue alors radicalement. En fait, ces couples représentent environ $50,34\%$ de la table de traduction. Avec la table de référence dans sa totalité, les traductions automatiques fournies par PHARAOH obtiennent un score BLEU de $37,15$ sur le corpus de test. Si les couples de traductions avec une probabilité de 1 sont écartés de la table, PHARAOH produit des traductions automatiques qui aboutissent cette fois à un score BLEU similaire de $37,16$ sur le même corpus de test. En d'autres termes, PHARAOH dans son processus de décodage n'utilise pas ces paires de séquences. Nous pouvons donc les qualifier de bruit. Sans ce bruit, notre table de traduction basée sur les triggers inter-langues couvre un nombre de séquences françaises non négligeable comparé à la table de traduction de référence.
Pour construire notre table de traduction de séquences, nous avons pour chaque séquence française constitué un ensemble de 90 triggers inter-langues d'au moins 1 mot anglais et d'au plus 3 à raison de 30 triggers inter-langues par longueur. Parmi ces 90 triggers inter-langues, les $n = 20$ meilleurs suivant la valeur de l'Information Mutuelle et indépendamment de leur taille

sont conservés comme traductions potentielles. Afin d'éviter tout effet de sur-apprentissage sur le corpus de développement, ces triggers inter-langues sont intégrés automatiquement à la table de traduction sans passer par l'étape de sélection par l'algorithme du Recuit Simulé. L'Information Mutuelle a été normalisée comme suit pour associer une probabilité à chaque couple de séquences :

$$\forall f_j \in V_F, \forall e_i, e_k \in Trig_n(f_j) \;\; p_{NORM1}(e_i|f_j) = \frac{IM(f_j, e_i)}{\sum_{k=1}^{n} IM(f_j, e_k)} \tag{6.2}$$

où f_j est une séquence appartenant à la liste des séquences retenues sur le corpus d'apprentissage et notée ici V_F, et e_i, e_k sont des séquences anglaises appartenant à l'ensemble des n meilleurs triggers inter-langues de séquence de f_j noté $Trig_n(f_j)$. Cette nouvelle table de traduction compte un total de 36 millions d'entrées. Le score BLEU associé aux traductions produites par PHARAOH sur le corpus de test d'EUROPARL avec cette nouvelle table de traduction est indiqué dans le tableau 6.13 à la ligne intitulée Triggers m-n. Les quatre premières lignes rappellent

Table de traduction	BLEU
Triggers 1-1	30,97
IBM 3	29,57
Référence	37,15
Triggers m-n RS	32,75
Triggers m-n	30,83

Tab. 6.13 – Evaluations des tables de traduction en terme de score BLEU sur le corpus de test de EUROPARL

les performances obtenues sur le même corpus de test avec les tables de traduction de mots et de séquences exposées précédemment. Comme nous pouvons le constater, l'ajout des triggers inter-langues de séquences n'améliore pas la qualité des traductions par rapport à l'utilisation des triggers 1-1 de mots. Dans la section précédente, nous avons pourtant montré que l'ajout de triggers de séquences pertinemment choisis avec l'algorithme du Recuit Simulé, permettait à PHARAOH de produire des traductions de meilleure qualité en BLEU que celles obtenues avec une table de traduction de mots. Afin de mieux cerner le problème et comprendre pourquoi l'ajout des triggers m-n n'améliore pas la qualité de traduction de PHARAOH, nous nous sommes penchés sur les distributions de probabilités de notre table de traduction de séquences. L'étude est décrite dans la section suivante.

6.4.6.2 Etude des distributions de probabilités

Nous avons vu jusqu'à maintenant que les triggers inter-langues permettent d'établir des correspondances lexicales de bonne qualité que ce soit au niveau de mots ou de groupes de mots. Appliqués en Traduction Automatique, les triggers de mots obtiennent des performances similaires à celles de la table de traduction de mots du modèle 3 d'IBM. Il n'en va pas de même pour les triggers de séquences qui n'apportent aucun avantage par rapport aux triggers de mots et qui obtiennent des résultats bien moins performants que pour une table de traduction de séquences construite à partir de l'alignement des mots. Le bénéfice de l'utilisation de séquences en Traduction Automatique Statistique n'est pourtant plus à prouver [Och 99, Marcu 02, Koehn 03]. Dans le but de comprendre pourquoi notre table de traduction de séquences Triggers m-n échouait

dans cette voie malgré le fait d'établir de bonnes correspondances lexicales à première vue, nous nous sommes intéressés à la distribution des probabilités de notre table de traduction. Rappelons que pour associer une probabilité aux triggers inter-langues, nous normalisons la valeur de leur Information Mutuelle comme indiqué par l'équation 6.2.

Séquence anglaise	**IM(e,f)**$\times 10^{-}4$	$p_{NORM1}(e\|f) \times 10^{-1}$
confidence in	2,34	2,39264
confidence	2,05	2,09611
confidence in the	1,17	1,19632
in the	0,47	0,48057
in	0,36	0,36810
on confidence in	0,33	0,33742
currency represents	0,30	0,30675
on confidence	0,30	0,30675
our confidence in	0,24	0,24540
in electronic	0,22	0,22495
our confidence	0,22	0,22495
build	0,21	0,21472
electronic	0,21	0,21472
trust in the	0,21	0,21472
euro	0,20	0,20450
monetary	0,20	0,20450
public confidence	0,20	0,20450
blow	0,19	0,19427
; to	0,18	0,18405
trust in	0,18	0,18405
Entropie		3, 59

TAB. 6.14 – Traductions de la séquence française f ="la confiance dans" dans la table Triggers m-n

Les tableaux 6.14 et 6.16 donnent respectivement les entrées de la table de traduction Triggers m-n pour les séquences françaises "la confiance dans" et "de la libre concurrence". La colonne $IM(e, f)$ rappelle l'Information Mutuelle associée au trigger inter-langue et la colonne $p_{NORM1}(e|f)$ indique la probabilité calculée à partir de l'IM comme le suggère l'équation 6.2. Comme nous pouvons le constater, les distributions de probabilités pour ces deux séquences françaises ne sont pas très discriminantes. En effet, aucune valeur ne nous permet de trancher vraiment pour une traduction particulière. A titre de comparaison, les tableaux 6.15 et 6.17 présentent les traductions de ces mêmes séquences françaises extraites de la table de traduction Référence construite à partir de l'alignement des mots. A contrario, les distributions de probabilités ici creuse un écart important entre, d'une part la meilleure traduction potentielle et d'autre part les autres traductions possibles de chacune des deux séquences anglaises.

Pour déterminer dans quelle mesure la distribution de probabilité de chaque séquence permet ou non de prendre une décision quant à la traduction, nous nous sommes intéressés à la valeur de leur Entropie [DeNero 06]. L'Entropie H est une fonction mathématique qui permet dans notre cas de quantifier la notion de distribution discriminante ou non, elle se calcule de la manière

Séquence anglaise	$p_{REF}(e\|f) \times 10^{-1}$
confidence in	6,61292
trust in	0,80645
trustworthiness of	0,16129
trusting in	0,16129
this confidence	0,16129
the trust in	0,16129
then confidence in	0,16129
's faith in	0,16129
's confidence in	0,16129
perceived trustworthiness of	0,16129
people's faith in	0,16129
my view confidence in	0,16129
in my view confidence in	0,16129
have confidence in	0,16129
faith in	0,16129
consumer confidence in	0,16129
confidence and trust in	0,16129
belief in	0,16129
Entropie	$2,22$

TAB. 6.15 – Traductions de la séquence française f ="la confiance dans" dans la table Référence

Séquence anglaise	**IM(e,f)**$\times 10^{-}4$	$p_{NORM1}(e\|f) \times 10^{-1}$
free competition	1,60	1,90250
of free competition	1,02	1,21284
free	0,94	1,11772
competition	0,83	0,98692
of free	0,75	0,89180
free competition and	0,73	0,86801
competition and	0,50	0,59453
free competition which	0,34	0,40428
having to ensure	0,33	0,39239
dogmas	0,28	0,33294
ensure the	0,26	0,30916
market	0,23	0,27348
competition which	0,22	0,26159
equality of	0,19	0,22592
free market	0,19	0,22592
Entropie		$3,58$

TAB. 6.16 – Traductions de la séquence française "de la libre concurrence" dans la table Triggers m-n

Séquence anglaise	$p_{REF}(e\|f) \times 10^{-1}$
of free competition	3,42857
of the free market	0,571429
of competition	0,571429
free competition	0,571429
fair competition	0,571429
the idea of free competition	0,285714
the distortion of competition	0,285714
of the competitive	0,285714
of full and fair competition	0,285714
of free competition we	0,285714
of free competition and	0,285714
of control and supervision	0,285714
is a competitive	0,285714
in free competition	0,285714
free competition and	0,285714
flouted the idea of free competition	0,285714
direction of free competition	0,285714
competition	0,285714
a competitive	0,285714
about free competition	0,285714
Entropie	$3,67$

TAB. 6.17 – Traductions de la séquence française f ="de la libre concurrence" dans la table Référence

suivante :

$$H(d(f)) = -\sum_{e} p(e|f) log_2 p(e|f) \tag{6.3}$$

où $d(f)$ est la distribution de probabilité associée à la séquence française f. Plus la valeur de l'Entropie est grande, plus il y a d'incertitude quant à la bonne traduction de la séquence française f. Plus la valeur de l'Entropie est proche de 0, plus la distribution de probabilités est discriminante. La distribution de probabilité de la séquence française "la confiance dans" a une entropie de $3,59$ dans la table de traduction Triggers m-n et une entropie de $2,22$ dans la table de traduction Référence. La distribution de cette séquence est donc plus discriminante dans la table de traduction Référence. A l'inverse, la distribution de probabilité de la séquence "de la libre concurrence" est plus faible et donc plus discriminante dans la table de traduction Triggers m-n. Mais qu'en est-il sur l'ensemble des séquences françaises des tables de traductions ? Nous avons calculé l'Entropie sur les distributions de chaque séquence française des tables de traduction Triggers m-n et Référence. La figure 6.3 représente la répartition des distributions de probabilités des différentes tables de traduction étudiées sur différents intervalles de valeurs pour l'Entropie. La série "NORM1" correspond à la table de traduction Triggers m-n comme construite jusqu'à maintenant. La série "REF" reprend la table de traduction Référence. Plus de 90% des séquences françaises de la table de traduction Triggers m-n ont une Entropie de distribution supérieure à 4, ce qui signifie que pour $1,6M$ de séquences françaises, notre table de traduction ne permet pas de prendre une décision pour trouver la bonne traduction. En revanche, dans la

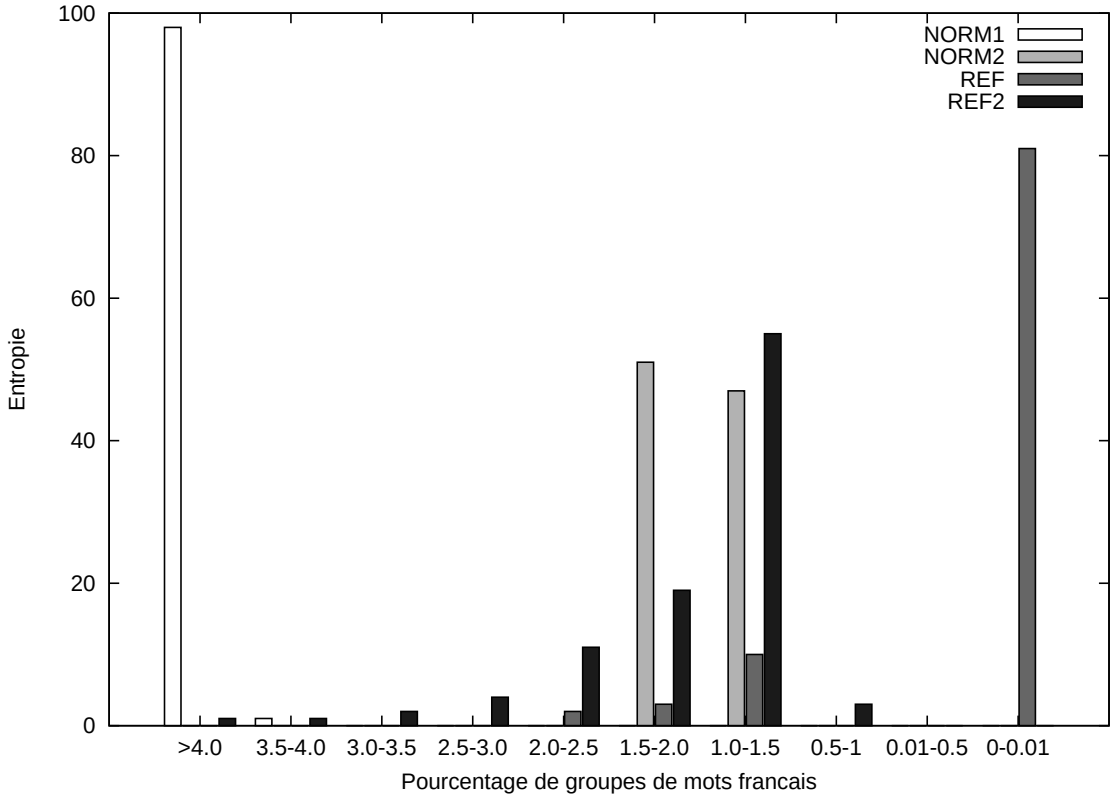

FIG. 6.3 – Entropie des distributions de probabilités des tables de traduction

table de traduction Référence, plus de 80% des séquences françaises, soit plus de $10M$, ont une Entropie de distribution comprise entre 0 et $0,01$, ce qui veut dire que la table de traduction Référence pour ces séquences permet de faire un choix parmi toutes les traductions possibles. Cependant, ce dernier constat est biaisé par le nombre de traductions dans la table Référence pour lesquelles la probabilité associée est de 1. Nous avons vu que ces couples représentaient plus de 90% de la table et que de plus ils n'intervenaient pas dans le calcul des meilleures hypothèses par le décodeur PHARAOH. La série "REF2" reprend la table de traduction Référence amputée de tous ces couples. La plupart des séquences françaises restantes ont une entropie de distribution comprise entre 1 et $2,5$, ce qui reste nettement inférieur à celle des séquences françaises de notre table de traduction Triggers m-n. Nos probabilités calculées à partir de l'Information Mutuelle ne sont pas assez discriminantes. Ceci peut expliquer le fait que l'introduction des triggers inter-langues de séquences n'améliore par les performances de PHARAOH par rapport à la table de traduction de mots Triggers 1-1. Les données que nous ajoutons introduisent une certaine forme d'incertitude et ne permettent pas à PHARAOH de les utiliser à bon escient. Nous avons donc calculé de nouvelles probabilités pour notre table de traduction Triggers m-n, toujours à partir de l'Information Mutuelle associée aux triggers inter-langues, mais cette fois, afin de rendre ces probabilités plus discriminantes, nous avons également tenu compte de leur ordre dans la liste triées sur la valeur de l'Information Mutuelle croissante. L'équation est la suivante :

$$\forall f_j \in V_F, \forall e_i, e_k \in Trig_n(f_j) \quad P_{NORM2}(e_i|f_j) = \frac{IM(f_j,e_i) \times e^{(-i+1)}}{\sum_{k=1}^{n} IM(f_j,e_k) \times e^{(-k+1)}} \tag{6.4}$$

Nous avons ajouté le facteur $e^{(-k+1)}$ qui permet d'accentuer la probabilité des triggers inter-langues en début de liste. Les colonnes intitulées "$p_{NORM2}(e|f)$" dans les tableaux 6.18 et 6.19 indiquent les nouvelles probabilités des traductions possibles des séquences françaises "la confiance dans" et "de la libre concurrence".

Séquence anglaise	**IM(e,f)**$\times 10^{-4}$	$p_{NORM1}(e\|f) \times 10^{-1}$	$p_{NORM2}(e\|f) \times 10^{-1}$
confidence in	2,34	2,39264	7,12144
confidence	2,05	2,09611	2,29515
confidence in the	1,17	1,19632	0,48189
in the	0,47	0,48057	0,07121
in	0,36	0,36810	0,02007
on confidence in	0,33	0,33742	0,00677
currency represents	0,30	0,30675	0,00226
on confidence	0,30	0,30675	0,00083
our confidence in	0,24	0,24540	0,00025
in electronic	0,22	0,22495	0,00008
our confidence	0,22	0,22495	0,00003
build	0,21	0,21472	0,00001
electronic	0,21	0,21472	0,00000
trust in the	0,21	0,21472	0,00000
euro	0,20	0,20450	0,00000
monetary	0,20	0,20450	0,00000
public confidence	0,20	0,20450	0,00000
blow	0,19	0,19427	0,00000
; to	0,18	0,18405	0,00000
trust in	0,18	0,18405	0,00000
Entropie		$3,59$	$1,13$

TAB. 6.18 – Traductions de la séquence française f ="la confiance dans" dans la table Triggers m-n

L'Entropie de la première séquence passe de $3,59$ à $1,13$ et de $3,58$ à $1,19$ pour la deuxième. Cette deuxième normalisation permet donc de rendre les distributions de probabilités plus discriminantes. L'entropie des distributions de ces nouvelles probabilités sur l'ensemble des séquences françaises est représentée par la série "NORM2" de la figure 6.3. Cette fois, pour la majorité des séquences, l'Entropie se situe entre 1 et 2, ce qui veut dire que pour cette majorité, il est possible de prendre une décision parmi toutes les traductions possibles proposées dans la table. Le tableau 6.20 rappelle les performances obtenues par le décodeur PHARAOH sur le corpus de test d'EUROPARL avec les tables de traduction étudiées jusqu'ici. Avec la table de traduction Triggers m-n dont les probabilités ont été calculées suivant l'Information Mutuelle seulement (Triggers m-n NORM1), le score BLEU associé aux traductions automatiques étaient de $30,83$. Avec la table de traduction dont les probabilités dépendent de la valeur de l'IM mais aussi du rang des triggers inter-langues (Trigger m-n NORM2) , les traductions automatiques produites par le décodeur PHARAOH obtiennent cette fois un score BLEU de $34,41$. L'ajout des triggers inter-langues de séquences a donc un réel impact sur la qualité des traductions comparées à celles produites à l'aide des triggers de mots. Malgré cette nette amélioration, les performances obtenues avec notre table de traduction de séquences demeurent inférieures à celles obtenues avec la

Séquence anglaise	**IM(e,f)**$\times 10^{-}4$	$p_{NORM1}(e\|f) \times 10^{-1}$	$p_{NORM2}(e\|f) \times 10^{-1}$
free competition	1,60	1,90250	7,39323
of free competition	1,02	1,21284	1,73388
free	0,94	1,11772	5,8783
competition	0,83	0,98692	0,19095
of free	0,75	0,89180	0,06347
free competition and	0,73	0,86801	0,02273
competition and	0,50	0,59453	0,00573
free competition which	0,34	0,40428	0,00143
having to ensure	0,33	0,39239	0,00051
dogmas	0,28	0,33294	0,00016
ensure the	0,26	0,30916	0,00005
market	0,23	0,27348	0,00002
competition which	0,22	0,26159	0,00001
equality of	0,19	0,22592	0,00000
free market	0,19	0,22592	0,00000
Entropie		3, 58	1, 19

TAB. 6.19 – Traductions de la séquence française “de la libre concurrence” dans la table Triggers m-n

Table de traduction	BLEU
Triggers 1-1	30,97
Triggers m-n RS	32,75
Triggers m-n NORM1	30,83
Triggers m-n NORM2	34,41
Référence	37,15

TAB. 6.20 – Evaluations des tables de traduction en terme de score BLEU sur le corpus de test de EUROPARL

table de traduction Référence. Cette différence de 2, 74 points en score BLEU est statistiquement significative[16].

6.5 Discussion

Dans ce chapitre, nous avons proposé une méthode originale de construction d'une table de traduction de séquences. La première originalité de notre approche tient au fait que nous sélectionnons a priori les séquences sources de notre table de traduction sur la partie source du corpus bilingue d'apprentissage et ce, totalement indépendamment de la partie cible. Ceci nous permet de constituer un ensemble de séquences sources pertinentes. La deuxième originalité est que nous utilisons les triggers inter-langues pour déterminer les traductions de ces séquences sources et non l'alignement des mots au sein du corpus d'apprentissage. L'ensemble des séquences sources

[16]test de Wilcoxon

associées à leurs traductions potentielles est appelé "Triggers m-n". Notre objectif était d'injecter ces triggers m-n dans notre table de traduction de mots (Triggers 1-1) et ainsi améliorer la qualité des traductions produites par PHARAOH.

Nous avons mené deux grandes séries d'expériences afin d'évaluer la qualité des triggers inter-langues de séquences dans un système de Traduction Automatique. A chaque fois, nous nous sommes comparés à l'approche suivie par Koehn [Koehn 03] qui consiste à extraire la table de traduction de séquences à partir de l'alignement bidirectionnel des mots. Nous avons appelé cette table de traduction Référence.

Dans la première série, nous avons utilisé l'algorithme du Recuit Simulé afin d'ajouter un sous-ensemble optimal des Triggers m-n dans la table de traduction. Sur le corpus SSTITRES, le manque de données a fait que l'utilisation des triggers de séquences ne s'est pas avéré concluante. La qualité des traductions a en effet baissé par rapport à celle des traductions produites avec la table de traductions Triggers 1-1. Toutefois, le même phénomène a été observé avec la table de traduction de séquences Référence construite à partir de l'alignement des mots. Plus important encore, l'utilisation des paires de séquences construites à partir des triggers inter-langues a engendré une perte moins importante en terme de score BLEU que l'utilisation de paires de séquences extraites d'après l'alignement des mots. Un tel alignement est une tâche automatique qui devient difficile face à des corpus peu fournis et complexes comme le corpus SSTITRES. Du fait de ne pas s'appuyer sur l'alignement automatique des mots, notre méthode s'avère plus robuste.
Sur le corpus EUROPARL en revanche, l'introduction de triggers m-n dans la table de traduction a engendré un gain de $1,78$ points BLEU sur le corpus de test Une analyse approfondie des résultats nous a conduit à la deuxième série d'expériences. L'étude a révélé dans un premier temps que notre table de traduction couvrait beaucoup moins de séquences sources que la table Référence, elle était donc beaucoup plus petites en terme de nombre d'entrées. Par conséquent, dans la seconde série d'expériences, nous avons sélectionné un nombre plus important de séquences sources au préalable. Pour chacune des séquences sources, nous avons sélectionné un ensemble de triggers m-n sur le corpus d'apprentissage et nous avons ajouté tous les triggers m-n à la table de traduction de mots sans passer par l'étape de sélection menée par l'algorithme du Recuit Simulé. En effet, nous avons constaté que cette étape causait un effet de sur-apprentissage sur le corpus de développement et que par conséquent les bénéfices qu'apportaient les triggers m-n sélectionnés ne se répercutaient pas sur le corpus de test.

Après optimisation de notre table de traduction Triggers m-n, les traductions produites par PHARAOH atteignent un score BLEU de $34,41$ sur le corpus de test d'EUROPARL, soit un gain de $3,44$ points par rapport au fait d'utiliser la table de traduction de mots Triggers 1-1. L'influence positive de l'utilisation de séquences sur la qualité des traductions automatiques est donc confirmée au sein de nos modèles basés sur les triggers inter-langues. Toutefois, les traductions produites avec notre table de traduction Triggers m-n demeurent de moins bonne qualité en terme de score BLEU que celles obtenues avec la table de traduction de séquences faisant état de référence et construite à partir de l'alignement des mots.

Bien qu'aboutissant à un score BLEU inférieur, notre approche présente certains avantages par rapport à l'approche suivie par Koehn [Koehn 03] que nous avons choisie comme référence. Tout d'abord, la sélection *a priori* de la liste des séquences sources de notre table de traduction sur le corpus d'apprentissage indépendamment de la partie cible nous garantit de contrôler la taille de la table ainsi que la pertinence des couples de traduction qu'elle contient. Nous avons vu dans la table de référence construite à partir de l'alignement des mots que $50,24\%$ des entrées

avaient une probabilité de traduction de 1 et que si elles étaient écartées durant la phase de décodage, la qualité des traductions produites par PHARAOH restait inchangée. Dans ce sens, les travaux de Zettlemoyer [Zettlemoyer 07] ont également montré qu'il était possible de diminuer de moitié la taille d'une telle table de traduction en sélectionnant un sous-ensemble seulement des couples extraits à partir de l'alignement des mots tout en améliorant le score BLEU. Dans son approche, Koehn extrait autant de paires de séquences que possible pourvu qu'elles soient consistantes avec l'alignement des mots avec le risque de sélectionner des couples non pertinents. Ceci conduit à des tables de traduction de taille démesurée de plusieurs millions d'entrées (la table Référence en compte plus de 24 millions) et à des aberrations. Par exemple, dans la table de traduction Référence construite sur le corpus EUROPARL, le mot français "de" est associé à 29507 traductions possibles et 140 séquences parmi les 13 millions de séquences françaises couvertes par la table sont associées à plus de 1000 traductions. Il s'agit très souvent des mots outils de la langue tels que "de, à, le, la" ... Un autre avantage de notre approche tient au fait que nous n'utilisons pas l'alignement des mots pour sélectionner les couples de séquences de la table de traduction, contrairement à la méthode de Koehn. Bien que cette dernière extrait les couples de séquences par de simples heuristiques, elle nécessite toutefois l'alignement bi-directionnel des mots du corpus d'apprentissage. Chacun des alignements Anglais-Français et Français-Anglais est calculé automatiquement par un processus itératif complexe qui peut prendre dans le cas de grand corpus comme EUROPARL des heures, voire des jours de calcul. De plus, les alignements résultants ne sont pas dénués d'erreurs, ce qui peut conduire à de mauvais couples de traduction. Dans notre cas, une fois les séquences sources sélectionnées, nous identifions leurs traductions potentielles à l'aide des triggers inter-langues. La sélection des triggers s'opère simplement et instantanément en une seule itération. Notre table de traduction basée sur les triggers se construit par conséquent beaucoup plus rapidement qu'une table mise en place à partir de l'alignement des mots.

Avec une méthode moins complexe et par conséquent plus rapide, nous aboutissons à des traductions automatiques ayant un score BLEU inférieur de $2,74$ points au score BLEU associé aux traductions produites grâce à une table de traduction construite avec la méthode de Koehn. Notre travail ne s'achève bien évidemment pas avec ce manuscrit. Nous explorons actuellement des pistes dans le but d'améliorer le score BLEU associé aux traductions automatiques produites par le décodeur PHARAOH avec les informations apportées par les triggers inter-langues et montrer qu'il est possible d'utiliser des alternatives aux modèles de traduction actuels.

Conclusion et perspectives

L'objectif de cette thèse était d'apporter quelques pierres au grand édifice en devenir qu'est la Traduction Automatique Statistique. Nous avons concentré nos efforts sur deux principaux composants caractérisant cette approche, à savoir : les corpus parallèles alignés nécessaires à l'apprentissage des modèles de traduction et la table de traduction qui est une des ressources primordiales dont le décodeur a besoin pour calculer les traductions automatiques les plus probables de phrases d'entrée.

La première difficulté à laquelle se heurte toute personne souhaitant mettre en place un système de TA statistique est la collecte de corpus bilingues alignés à partir desquels la machine va tirer ses connaissances. Recueillir de tels corpus est une tâche difficile et coûteuse. Il faut en effet trouver des textes qui soient disponibles en plusieurs langues et ensuite constituer le corpus de telle sorte que chaque phrase soit mise en correspondance avec sa traduction. Dans nos travaux, nous avons proposé une méthode originale basée sur la programmation dynamique capable de constituer automatiquement des corpus parallèles pour la Traduction Automatique à partir de sous-titres de films. L'utilisation de sous-titres de films présente plusieurs avantages : le premier est que les sous-titres sont des données très abondantes et libres d'accès. Il est donc possible de constituer de grands corpus pour l'apprentissage des modèles de traduction. Un autre avantage, est que la parole transcrite dans les sous-titres se rapproche d'avantage de la parole spontanée que celle retranscrite dans les corpus classiquement utilisés dans le domaine de la TA statistique. En effet, les corpus généralement employés sont des transcriptions des actes du Parlement Européen. La parole transcrite est celle émise par les membres de la Commission du Parlement Européen, qui utilisent un langage très formel ne rappelant guère la parole spontanée. Les discours tenus sont par ailleurs ciblés sur des sujets bien spécifiques tels que l'environnement ou les affaires étrangères et concernent donc un domaine assez restreint. A l'inverse, les sous-titres de films offrent une variété de discours très importante et un vocabulaire diversifié composé aussi bien de termes soutenus que de termes familiers en fonction du genre des films. Les sous-titres constituent par conséquent un matériel plus adéquat pour l'apprentissage de modèles de traduction destinés à un système grand public qui ne serait pas dédié à un domaine spécifique mais qui traduirait de la parole spontanée. L'algorithme que nous avons proposé permet d'aligner automatiquement les sous-titres d'un même film en deux langues différentes. Il nécessite simplement pour cela un dictionnaire de mots bilingues grâce auquel il établit un score de similarité entre les sous-titres et calcule ensuite le meilleur alignement de sous-titres pour le film. Les tests menés sur une quarantaine de films ont montré que notre méthode permet de retrouver jusqu'à 93% des alignements de référence établis manuellement. Nous avons ainsi constitué un corpus parallèle aligné original de près de $30K$ paires de sous-titres Français et Anglais dont nous nous sommes servies par la suite pour l'apprentissage et l'évaluation des différents systèmes de traduction.

Une fois les corpus parallèles alignés à disposition, ils sont utilisés pour l'apprentissage des

modèles de traduction. Parmi les nombreux paramètres pouvant exister dans un modèle de traduction, un décodeur tel que PHARAOH ou MOSES, pour trouver la traduction la plus probable d'une phrase source, a essentiellement besoin d'une table de traduction qui associe des probabilités de traduction à des couples de mots ou de séquences et d'un modèle de langage cible qui estime la probabilité d'une suite de mots dans la langue cible. La table de traduction permet au décodeur de trouver les correspondances les plus probables dans la langue cible des différents mots ou groupes de mots de la phrase source. Que l'unité de traduction soit le mot ou le groupe de mots, la plupart des méthodes d'apprentissage de tables de traduction utilise les modèles IBM.

Concernant les tables de traduction de mots, elles sont directement des paramètres de ces modèles. Les tables de traduction de séquences, quant à elles, peuvent être extraites à partir d'alignements des mots au sein des paires de traductions du corpus d'apprentissage. Ces alignements sont calculés automatiquement à partir des modèles d'alignement qui sont des paramètres des modèles IBM. Autrement dit, les modèles IBM sont omniprésents dans le domaine de la Traduction Automatique Statistique. Il s'agit de modèles dont la complexité augmente du premier au cinquième et qu'il n'est pas toujours évident de maîtriser. Ils combinent en effet différents sous-modèles comme les tables de traduction, les modèles d'alignement qui définissent les probabilités d'alignement entre une position dans la phrase source et une position dans la phrase cible, les modèles de fertilité qui définissent la probabilités qu'un mot source soit traduit en un ou plusieurs mots cibles. Chacun de ces sous-modèles interfère sur l'estimation des paramètres des autres sous-modèles. Toutefois, il n'est guère besoin de maîtriser parfaitement les modèles IBM pour les entraîner puisqu'il existe des outils tels que Giza++ permettant leur apprentissage de façon totalement opaque. Ces outils font office de boîtes noires. Ils utilisent en entrée les corpus parallèles alignés et fournissent en sortie les paramètres des modèles IBM tels que les modèles d'alignement ou les tables de traduction de mots. L'apprentissage des modèles IBM est un processus complexe pouvant prendre plus d'une dizaine d'heures de calcul en fonction de la taille du corpus d'apprentissage, il nécessite également une puissance de calcul importante. Notre deuxième objectif était de proposer une alternative aux modèles IBM pour l'apprentissage des tables de traduction de mots ou de séquences de mots. Nous souhaitions prospecter de nouvelles voies afin de proposer des méthodes moins coûteuses et plus performantes. Pour cela, nous avons proposé un nouveau modèle de traduction qui ne fait appel à aucun moment aux modèles IBM et qui ne tient pas compte de la probabilité d'alignement des mots au sein des paires de traduction du corpus d'apprentissage. Nous nous sommes pour cela reposés sur le concept de triggers inter-langues.

Le concept de triggers inter-langues reprend celui de triggers utilisé en modélisation du langage non pas dans un contexte monolingue mais dans un contexte bilingue. Les triggers sont des couples de mots d'une même langue fortement corrélés au sens de l'Information Mutuelle. Les triggers inter-langues, quant à eux, sont des couples de mots de deux langues différentes fortement corrélés au sens l'Information Mutuelle. Ils sont, eux, calculés sur un corpus bilingue aligné.
Dans nos travaux, nous avons proposé des méthodes originales utilisant le concept de triggers inter-langues pour établir des correspondances lexicales entre le Français et l'Anglais d'abord au niveau du mot puis au niveau du groupe de mots. L'évaluation des correspondances lexicales ainsi établies par les triggers inter-langues s'est faite de deux façons.

Dans un premier temps, nous avons tiré profit des correspondances révélées au niveau du mot par les triggers inter-langues pour mettre en place un dictionnaire bilingue Français-Anglais que

nous avons ensuite comparé à deux dictionnaires de référence. Dans l'étude que nous avons menée, nous avons établi que, jusqu'à 65% des entrées de notre dictionnaire bilingue étaient associées aux mêmes traductions que celles proposées dans les dictionnaires de référence. De plus, nous avons montré que les 35% des entrées de notre dictionnaire pour lesquelles nous n'avons trouvé aucune traduction commune avec les dictionnaires de référence relevaient plus d'un choix inapproprié des dictionnaires de référence que de mauvaises traductions établies par les triggers inter-langues.

Dans un second temps, nous avons utilisé les correspondances lexicales de mots et de groupes de mots révélées par les triggers inter-langues dans un processus original afin d'établir des tables de traduction que nous avons ensuite évaluées dans un processus de Traduction Automatique. Les expériences ont été parallèlement menées sur le corpus EUROPARL extrait des actes du Parlement Européen et sur le corpus SSTITRES que nous avons construit automatiquement à partir de sous-titres de films.
Nous avons proposé trois méthodes permettant la mise en place de tables de traduction de mots, toutes emploient les triggers inter-langues pour sélectionner les couples de traduction de mots mais utilisent des heuristiques de sélection différentes et de plus en plus fines pour choisir les triggers inter-langues à intégrer aux tables de traduction. Les heuristiques avaient pour but d'écarter les triggers qui ne représentaient pas réellement des couples de traduction de mots. Pour évaluer les différentes tables de traduction ainsi produites, nous les avons intégré au décodeur PHARAOH et nous avons comparé la qualité des traductions produites par le décodeur en terme de score BLEU avec ces différentes tables. Nous avons constaté que les traductions automatiques produites obtiennent un meilleur score BLEU lorsque le décodeur a un plus large choix de traductions possibles pour un même mot. En effet, plus le nombre de triggers inter-langues à intégrer aux tables de traduction était restreint par des heuristiques de filtrage, moins bonne était la qualité des traductions produites par le décodeur PHARAOH en terme de score BLEU. En d'autres termes, certains couples de mots même s'ils ne constituent pas de réelles traductions sont utiles au décodeur PHARAOH pour obtenir des performances optimales.
Nous nous sommes comparés ensuite avec la table de traduction de mots extraite du modèle IBM3. Nous avons montré que les traductions automatiques produites par le décodeur PHARAOH avec la table de traduction extraite à partir des triggers inter-langues obtiennent un score BLEU similaire sur le corpus SSTITRES et légèrement meilleur sur le corpus EUROPARL, que le score BLEU associé aux traductions produites par le même décodeur avec la table de traduction extraite du modèle IBM3. Une table de traduction mise en place grâce aux triggers inter-langues est donc comparable à la table de traduction extraite du modèle IBM3. Il est important de souligner qu'à score BLEU équivalent, notre méthode présente l'avantage d'être simple et rapide puisqu'elle ne tient pas compte de l'alignement des mots au sein du corpus d'apprentissage et que les triggers sont sélectionnés en une seule itération sur le corpus d'apprentissage. A l'inverse, le modèle IBM3 est entraîné suite à un processus itératif complexe et long dépendant entre autre de l'estimation de l'alignement des mots sur le corpus d'apprentissage.

Les triggers inter-langues ayant montré leur potentiel à établir une table de traduction de mots, nous avons ensuite étendu le concept aux groupes de mots ou séquences. Au même titre qu'ils sont capables d'associer un mot cible et un mot source fortement corrélés, nous avons sélectionné les triggers non plus en calculant l'Information Mutuelle entre un mot source et un mot cible mais entre un groupe de mots sources et un groupe de mots cibles. Ainsi, les triggers inter-langues nous ont également permis d'établir des correspondances lexicales au niveau du groupe de mots. Partant de ce concept de triggers inter-langues de séquences, nous avons proposé une méthode originale de construction d'une table de traduction de séquences que nous

avons évalué dans un processus de Traduction Automatique textuelle de l'Anglais vers le Français, tout comme nous l'avons fait pour les tables de traduction de mots. Notre approche présente une double originalité. D'une part, nous sélectionnons *a priori* les séquences sources de notre table de traduction sur la partie source du corpus bilingue d'apprentissage et ce, totalement indépendamment de la partie cible. Pour cela, nous avons utilisé un algorithme qui procède par agglomération locale successive des mots en fonction de leur Information Mutuelle et qui ne retient que les séquences permettant de minimiser la perplexité. La segmentation du corpus source en séquences nous permet d'avoir une table de traduction de taille raisonnable et, qui plus est, avec des séquences sources pertinentes. D'autre part, à l'inverse de plusieurs méthodes d'extraction de table de traduction de séquences, nous n'utilisons pas l'alignement des mots pour trouver les traductions possibles de nos séquences sources. En effet pour cela, nous faisons appel aux triggers inter-langues de séquences.

Dans une première série d'expériences, nous avons utilisé l'algorithme du Recuit Simulé afin de sélectionner un sous-ensemble optimal des triggers de séquences à ajouter à la table de traduction. Une fois de plus, les expériences ont été menées parallèlement sur le corpus EUROPARL et sur le corpus SSTITRES. L'évaluation des tables de traduction s'est faite en comparant la qualité des traductions automatiques produites, en terme de score BLEU, par le décodeur PHARAOH avec les différentes tables de traduction testées. Nous avons étudié l'impact de l'utilisation de séquences de mots plutôt que des mots seuls dans la tables de traduction. Nous nous sommes également comparés avec une méthode état-de-l'art qui extrait la table de traduction de séquences à partir d'heuristiques sur l'alignement bi-directionnel des mots sur le corpus d'apprentissage.
Sur le corpus SSTITRES, de par le manque de données mais aussi de par leur complexité, l'utilisation des séquences dans la table de traduction a engendré une diminution du score BLEU associé aux traductions produites, que ce soit pour la table de traduction utilisant les triggers inter-langues ou pour la table de traduction état-de-l'art. Toutefois, la perte en score BLEU a été moins importante pour les tables de traduction utilisant les triggers inter-langues. Ne s'appuyant pas sur l'alignement des mots, notre méthode est plus robuste face à des corpus bilingues peu dotés.
Sur le corpus EUROPARL, l'introduction de séquences dans la table de traduction a permis au décodeur PHARAOH de produire des traductions automatiques ayant un score BLEU supérieur à celle produite avec une table de traduction de mots. Les triggers de séquences ont ainsi engendré un gain de $1,78$ points BLEU par rapport au score BLEU obtenu avec la table de traduction basée sur les triggers de mots. L'amélioration apportée par la table de traduction obtenue d'après l'alignement des mots sur la table de traduction de mots extraite du modèle IBM3 était, quant à elle, de $7,58$ points pour atteindre un score BLEU de $37,15$. Nous en avons conclu que l'algorithme du Recuit Simulé dont la tâche était de sélectionner sur un corpus de développement un sous-ensemble optimal des triggers de séquences à intégrer dans la table de traduction provoquait un effet de sur-apprentissage, ce qui expliquait le gain minime en score BLEU entre l'utilisation des triggers de mots et ceux de séquences. Ceci nous a mené à la deuxième série d'expériences.
L'étude comparative que nous avons menée afin de comprendre les différences de scores BLEU entre notre approche et l'approche Référence suivie par Koehn nous a amené, dans la seconde série d'expériences conduites sur le corpus EUROPARL, à agir sur différents paramètres de notre méthode. Nous avons d'une part sélectionné davantage de séquences sources sur le corpus d'apprentissage afin que notre table de traduction couvre plus de suites de mots sources. Nous avons ensuite construit notre table de traduction de séquences en utilisant tous les triggers de séquences sélectionnés au préalable sur le corpus d'apprentissage. Nous n'avons pas effectué de filtrage sur les triggers par le biais de l'algorithme du Recuit Simulé. Enfin, nous avons normalisé les valeurs

de l'Information Mutuelle associée aux triggers inter-langues selon une décroissance exponentielle afin de discrimer plus amplement les couples de traduction présents dans la table et aider ainsi le décodeur à prendre les bonnes décisions. Après optimisation, les traductions produites par PHARAOH avec la table de traduction fondée sur les triggers de séquences atteignent un score de 34, 41 points BLEU, soit un gain de 3, 44 points par rapport à la table de traduction utilisant les triggers de mots.

A l'issue de ces travaux, nous pouvons dire que le concept de triggers inter-langues a un fort potentiel pour la Traduction Automatique. Il nous a permis d'établir des correspondances lexicales au sein de tables de traduction de mots et de séquences de mots. Ces tables ont conduit à des traductions automatiques ayant un score BLEU supérieur à 30, en d'autres termes, des traductions automatiques compréhensibles et lisibles. Le recours aux triggers inter-langues pour la mise en place de tables de traduction présente plusieurs avantages. Le premier point fort est que la sélection des triggers est un processus simple qui ne nécessite qu'une seule passe sur le corpus d'apprentissage. Ce processus de sélection en une passe garantit un gain de temps important sur de gros corpus d'apprentissage pour la mise en place de tables de traduction comparé aux modèles itératifs IBM qui requièrent plusieurs passage sur le corpus pour ré-estimer les paramètres. Le second point fort est que la sélection des triggers de mots, tout comme la sélection des triggers de séquences ne dépend pas de l'alignement des mots au sein des paires de traduction qui constituent le corpus d'apprentissage. Au niveau des correspondances lexicales de séquences de mots, nous avons vu que cela permet de construire rapidement des tables de traduction de taille raisonnable. Nous avons également montré que le fait de ne pas s'appuyer sur l'alignement des mots rendait notre méthode plus robuste sur des corpus peu dotés et complexes tel que le corpus SSTITRES qu'une méthode fondée sur l'alignement bi-directionnel des mots.

Pour conclure sur ces travaux, nous dirons que la Traduction Automatique est un domaine de recherche "jeune" et que de nombreuses voies restent encore à explorer. Traduire automatiquement un texte relève encore de l'utopie. Même avec les meilleurs systèmes actuels, les scores BLEU associés aux traductions automatiques demeurent faibles et rarement supérieur à 50 dans des domaines non restreints. Nous avons montré qu'il est possible de mettre en place un système de Traduction Automatique sans pour autant utiliser les modèles IBM. Nous sommes convaincus que le concept de triggers inter-langues peut aider à améliorer les performances des systèmes de Traduction Automatique. Il s'agit d'un concept simple à mettre en place et parfaitement modulable. Dans ces travaux, nous nous sommes limités à sélectionner des mots ou des groupes de mots fortement corrélés. De la même manière, ce concept pourrait également nous permettre d'établir des correspondances non seulement lexicales mais aussi grammaticales tel que le couple Français-Anglais (*racine + terminaison du futur*, *will + Base Verbale*) ou encore syntaxiques tel que le couple Français-Anglais(*NOM+ADJ*,*ADJ+NOM*). Le premier couple indiquerait au décodeur que si la structure (will + Base Verbale) a été rencontrée dans la phrase anglaise à traduire alors il est probable que le verbe de la phrase française correspondante soit de la forme (racine + terminaison du futur). Les triggers inter-langues peuvent donc apporter de nombreuses informations, qui une fois combinées de façon optimale permettraient d'aboutir à des traductions automatiques de meilleur qualité en terme de score BLEU, au même titre que l'ajout de fonctions caractéristiques dans la formule log-linéaire de la Traduction Automatique Statistique.

Par la suite, nous allons concentrer nos travaux à différents niveaux. La stratégie que nous avons adoptée en optant pour un corpus parallèle aligné originale construit à partir de sous-titres de films nous parait importante à poursuivre. Les sous-titres se rapprochent en effet plus de la

parole spontanée. Il serait donc intéressant de s'investir davantage dans cette voie. Nous devons poursuivre nos efforts pour la construction d'un corpus parallèle aligné de sous-titres de films de taille plus conséquente que celle dont nous disposons jusqu'à maintenant et trouver également des solutions aux problèmes liés à ce type de données.
Il nous paraît également intéressant de nous pencher sur le problème de l'évaluation automatique des traductions, qui comme nous l'avons évoqué constitue une réelle difficulté dans le domaine de la Traduction Automatique. A court terme, nous allons dans un premier temps évaluer les traductions automatiques produites avec nos tables de traduction en utilisant d'autres mesures automatiques que le score BLEU, telles que METEOR, ROUGE, PER, TER mais aussi avec une évaluation manuelle. Cette étude nous permettra dans un premier temps de nous positionner par rapport à l'état de l'art autrement que par le score BLEU et nous révélerait sans doute quelques points importants sur lesquels réfléchir et travailler.
Pour finir, nous nous intéresserons aux triggers inter-langues pour la mise en évidence de correspondances grammaticales et syntaxiques. Nous étudierons un moyen de combiner tous les types de triggers au sein d'un décodeur original pour la Traduction Automatique utilisant notamment un algorithme génétique pour trouver la meilleure traduction automatique d'une phrase source. Nous pensons également nous pencher sur l'utilisation de la co-information qui généralise l'Information Mutuelle à n variables aléatoires (également appelé en Anglais *Interaction Information*) afin d'établir des associations non plus seulement binaires entre une unité source et une unité cible mais n-aires entre plusieurs unités cibles et plusieurs unités sources et ce dans le but d'aider le décodeur à faire les bons choix lors de la recherche des meilleures traductions.

Annexe A

Un modèle de langage pour la prise en compte de l'accord en genre et en nombre des mots

A.1 Introduction

L'accord en genre et en nombre des mots est un réel problème en modélisation statistique du langage. Dans les modèles de langage n-gramme, celui-ci est implicitement pris en compte dans la probabilité assignée à chaque suite de mots. Mais quant est-il lorsque les mots à accorder ne sont pas contigüs. Cela pose un réel soucis. Dans ce chapitre nous proposons un modèle de langage original, appelé Cache-Trait (CT) et inspiré du modèle Cache de Kuhn et DeMori [Kuhn 90]. Il permet de pallier ce problème en estimant la probabilité du genre et du nombre des mots à prédire. Ce modèle pourrait nous permettre à court terme de ré-estimer les scores des n meilleures hypothèses de traductions produites par un décodeur et ainsi permettre de corriger les fautes d'accord souvent présentes dans les sorties de systèle de Traduction Automatique. Avant de décrire notre modèle original ainsi que les expériences menées, nous rappelons dans la section suivante les fondements du modèle Cache classique.

A.1.1 Le modèle Cache

Le modèle Cache part du fait que certains mots sont représentatifs d'un texte. Un article politique, par exemple, contiendra certainement plusieurs occurrences des mots "président", "ministre", "asssemblée", "Sarkozy" .... Mais il suppose en plus qu'un mot récemment apparu a une probabilité d'occurrence plus importante que celle indiquée par sa fréquence dans la langue, c'est-à-dire celle indiquée par un modèle ngram. Ainsi, si le mot "Président" est rencontré dans un passé récent, sa probabilité d'occurrence se verra augmenter. Le modèle Cache permet de prendre en compte cette distribution particulière des mots et estime la probabilité d'un mot en fonction de sa fréquence d'apparition dans un passé récent, appelé Cache, selon l'équation suivante :

$$P(w_i) = \frac{1}{N} \sum_{j=1}^{n} \delta(w_i, w_j) \tag{A.1}$$

où N est la taille du Cache et $\delta(w_i, w_j)$ est la fonction de Kronecker égale à 1 si $w_i = w_j$ et 0 sinon. Le modèle Cache permet donc de prendre en compte des dépendances entre mots de longue distante. Il est généralement combiné à un modèle ngram pour renforcer certaines probabilités.

A.1.2 Les modèles Cache-Trait et Cache-Trait-Partiel

Nous avons vu que le modèle Cache est efficace pour renforcer la probabilité ngram d'un mot récemment apparu dans le passé. Nous avons étendu cette idée en l'appliquant non plus sur les mots mais sur leurs caractérisques et plus particulièrement sur le genre et le nombre. Nous proposons un nouveau modèle, appelé Cache-Trait, qui permet de renforcer la probabilité d'un mot en fonction de la compatibilité de ses caractéristiques avec les caractéristiques des mots récemment apparus. Si une caractéristique est dominante dans le passé, alors sa probabilité d'occurence est plus importante. Prenons l'exemple suivant :

les pommes croquantes sont souvent vertes

Un modèle trigram ne permet généralement pas de favoriser la suite "sont souvent vertes" par rapport à la suite "sont souvent verts". Toutefois, la dominance du genre féminin dans le contexte gauche devrait pouvoir renforcer la probabilité du mot "vertes" également de genre féminin. En effet "vertes" est compatible avec "pommes" et "croquantes" en termes de genre alors que "verts" ne l'est pas. Le modèle Cache-Trait va permettre d'assigner à chaque mot une probabilité qui va rendre compte de sa compatibilité en termes de genre et de nombre avec son contexte gauche. Ainsi il permettra aux modèles de langage classiques de discrimer davantage les ngrams.
La probabilité Cache-Trait d'un mot est calculée de la façon suivante :

$$P_{CT}(w_i) = \alpha \frac{N(Genre(w_i))}{\sum_{w_j \in V} N(Genre(w_j))} + \beta \frac{N(Nombre(w_i))}{\sum_{w_j \in V} N(Nombre(w_j))} \tag{A.2}$$

où $N(f(x))$ est le nombre d'occurrences du trait f du mot x dans le Cache, $Genre(x)$ est le genre du mot x, $Nombre(x)$ est le nombre du mot x, et V est le vocabulaire utilisé. Plus le genre et le nombre du mot w_i seront apparus dans le passé, appelé aussi le Cache, plus sa probabilité sera importante. Nous avons constitué le modèle CT en combinant deux modèles : un pour le genre et un pour le nombre. En effet certains mots sont invariables en genre (c'est le cas de "égoïste", "tranquille" etc.) et/ou en nombre (c'est le cas de "corps", "souris"). La distribution de ces mots n'étant pas équilibrée par rapport aux mots variables, il paraît plus correct de dissocier les deux traits.
Le tableau A.1 présente les traits que nous utilisons dans notre modèle. Le modèle CT va donc

Trait		Exemple
FS	(Féminin-Singulier)	porte
MS	(Masculin-Singulier)	stylo
FP	(Féminin-Pluriel)	portes
MP	(Masculin-Pluriel)	stylos
Fi	(Féminin-invariant en nombre)	souris
Mi	(Masculin-invariant en nombre)	tapis
iS	(invariant en genre-Singulier)	égoïste
iP	(invariant en genre-Pluriel)	ces
ii	(invariant en genre et en nombre)	beaucoup

TAB. A.1 – Liste des traits utilisés par le modèle Cache-Trait

nous permettre de renforcer ou au contraire de réduire la probabilité des mots assignée par des

modèles de langage classique. Nous l'utilisons donc en le combinant linérairement avec d'autres modèles. Ainsi, la combinaison avec un modèle ngram se fait de la façon suivante :

$$P(w_i|h) = \alpha P_n gram(w_i|h) + \beta P_{CT}(w_i|Cache) \quad \text{(A.3)}$$

où h est l'historique du mot w_i et $Cache$ est une séquence de m traits. α et β sont les poids respectifs des modèles ngram et CT qu'il nous faudra déterminer.

A.1.3 Expérimentations

A.1.3.1 Description des données

Les expériences ont été menées sur un corpus du journal Le Monde. Les corpus d'apprentissage, de développement et de test sont respectivement composés de 32, 8 et 1,8 millions de mots. Les 57 000 mots les plus fréquents constituent le vocabulaire utilisé. Le genre et nombre de chaque mot sont extraits d'une base de données lexicales de plus de 430 000 mots Français (appelée BDLEX1). Chaque entrée de cette base correspond à un mot et contient sa forme fléchie dérivée de sa forme canonique, son orthographe, ses attributs morphosyntaxiques, un indicateur de fréquence ou encore sa phonétisation.
Les modèles bigramme et trigramme, interpolés successivement avec le modèle CT, ont été appris sur le corpus d'apprentissage (avec la méthode de lissage de GoodTuring) et les paramètres de l'interpolation linéaire ont été optimisés en utilisant l'algorithme EM sur le corpus de développement. Deux ensembles de paramètres ont été mis en place pour α et β . Dans le premier ensemble (1parM), des valeurs statiques du poids des modèles ont été choisies quelque soit l'historique du mot à prédire. Dans le second ensemble (1parH), le poids des modèles est variable et dépend du contexte gauche du mot à prédire.
Les résultats sont présentés dans la section suivante.

A.1.3.2 Résultats et discussion

Nous avons évalué les performances de notre modèle CT interpolé avec un modèle ngramme sur le corpus de test et les avons comparées aux performances du modèle ngramme seul en termes de perplexité. Le tableu A.2 montre les résultats obtenus en terme de perplexité. Ces premières ex-

	Baseline	Modèle CT interpolé	
ordre ngram	ngram	1parM	1parH
$PPL_{n=2}$	212,83	206,85	**204,75**
$PPL_{n=3}$	165,35	**159,42**	-

Tab. A.2 – Perplexités évaluées sur le corpus de test pour les modèles ngrammes et CT interpolés

périences sont encourageantes. La combinaison du modèle CT avec un modèle bigramme montre une amélioration de plus de 8 points de perplexité par rapport au modèle bigramme seul. La combinaison avec un modèle trigramme améliore la perplexité de presque 6 points par rapport au trigramme seul. Les tests ont été réalisés sur différents corpus de test et pour chacun une baisse de la perplexité a été observée. Ce qui confirme le potentiel de notre approche à réduire la perplexité. Nous sommes donc parvenus, grâce au modèle CT, à mieux prendre en compte l'accord en genre et en nombre des mots et donc à améliorer les performances en modélisation du langage sans utiliser ni d'analyse grammaticale ni de règles linguistiques.

A.1.3.3 Les limites

Les résultats obtenus jusqu'à présent ont montré l'intérêt de l'utilisation du modèle Cache-Trait pour la prise en compte du genre et du nombre des mots en modélisation du langage. Toutefois, dans certains cas, la prédominance d'un trait dans le contexte gauche ne signifie pas pour autant que ce même trait à une plus forte probabilité d'apparaître. Prenons l'exemple suivant :

le beau portefeuille bleu de ma grande soeur

La probabilité du mot "soeur" va ici être diminuée par le modèle CT. En effet, le genre dominant dans le contexte gauche est masculin alors que "soeur" porte le trait féminin. Notre modèle peut donc induire certaines erreurs. Nous n'utilisons aucune analyse grammaticale ou syntaxique pour établir le Cache de notre modèle. Il s'agit simplement d'une suite de mots apparus. Pourtant les mots d'une phrase sont souvent regroupés en syntagme. Et l'accord en genre et en nombre se fait le plus souvent au sein d'un même syntagme. Dans notre exemple, nous devons considérer deux groupes ou syntagmes différents délimités par la préposition "de". Nous avons d'une part "le beau portefeuille bleu" et d'autre part "ma grande soeur". La compatibilité des traits de "soeur " doit être vérifiée avec les traits du groupe "ma grande". Cette décomposition du Cache doit se faire de façon dynamique en fonction justement des mots qu'il comporte. Ceci nous amène donc amène à définir un nouveau modèle appelé Cache-Trait Partiel (CTP). Cette fois le Cache n'est plus simplement les k mots du contexte gauche, mais il va dépendre de la présence ou non de mots délimiteurs de syntagmes.

A.1.3.4 L'apport du modèle CTP

Nous avons vu dans l'exemple précédent que la préposition « de» permettait de délimiter deux syntagmes. Nous avons testé une première fois le modèle CTP en prenant comme délimiteur du Cache de traits les séparateurs « de» et « du». Le calcul des probabilités CTP a été fait de la même façon que pour les probabilités CT. Les résultats obtenus sont présentés dans le tableau A.3. Le modèle CTP est combiné avec un modèle bigramme ou trigramme. Malgré la

	Baseline	Modèle CTP interpolé	
ordre ngram	ngram	1parM	1parH
$PPL_{n=2}$	212,83	206,57	**204,43**
$PPL_{n=3}$	165,35	**159,19**	-

Tab. A.3 – Perplexités des modèles ngrammes et CTP interpolés avec "de" et "du" comme délimiteurs de syntagmes

faible amélioration de la perplexité du modèle CTP interpolé par rapport au modèle CT, nous sommes convaincus que l'accord en genre et en nombre doit être pris en compte au sein d'un même syntagme délimité par des mots séparateurs. Pour confirmer notre intuition, nous avons introduit d'autres mots séparateurs pour délimiter les syntagmes au sein du Cache de traits. La liste des séparateurs utilisés est la suivante :

de, du, mais, ou, et, donc, or, ni, car, dans, avant, depuis, que, qui

L'introduction de ces nouveaux séparateurs a conduit à une perplexité de 203, 95 pour le modèle CTP interpolé avec un modèle bigramme et une perplexité de 159, 18 lorsqu'il est interpolé avec un modèle trigramme comme le montre le tableau A.4.

	Baseline	Modèle CTP interpolé	
ordre ngram	ngram	1parM	1parH
$PPL_{n=2}$	212,83	206,56	203,95
$PPL_{n=3}$	165,35	159,18	-

Tab. A.4 – Perplexités des modèles ngrammes et CTP interpolés avec une liste étendue de délimiteurs de syntagmes

Globalement, notre nouvelle méthode permet de diminuer la perplexité d'un modèle bigramme d'environ 8 points. Nous devons poursuivre nos efforts dans le but de délimiter de façon optimale les syntagmes dans le Cache de traits. Nous avons analysé les poids assignés aux probabilités CT dans l'interpolation linéaire avec un modèle ngramme. Cette étude nous a montré que la probabilité CT a un poids supérieur à 0,3 pour seulement 3655 historiques sur les 57 000 possibles et un poids supérieur à 0,6 pour seulement 736. Le tableau A.5 donne quelques exemples d'historiques. La colonne de gauche montre des historiques pour lesquels le modèle

Historique	poids CT t	Historique	poids CT
Derrick	0	apportée	1
Tiozzo	0,06	conceptuels	1
increvable	0,19	concertation	0,99
fuseaux	0,21	concurrente	0,99
défaits	0,26	voies	0,90
Sun	0,27	restaient	0,78
votre	0,31	ressentons	0,76
arrière-goût	0,36	verdeur	0,61

Tab. A.5 – Historiques de mots et poids du modèle CT résultant dans la combinaison linéaire

n-gramme a un poids beaucoup plus fort dans l'interpolation linéaire que le modèle CT. La colonne de droite montre, au contraire, des cas dans lesquels la probabilité CT a une importance beaucoup plus significative que la probabilité n-gramme. Malgré le faible nombre d'historiques amenant une dominance du modèle CT dans l'interpolation linéaire avec un modèle n-gramme, la contribution des traits entraîne tout de même une amélioration de la perplexité ngramme.

A.1.4 Conclusion

Nous avons proposé un modèle de langage statistique original faisant appel aux traits linguistiques des mots. L'idée est de considérer un mot non plus seulement comme une forme orthographique mais comme une unité linguistique portant plusieurs attributs et notamment le genre et le nombre. Plusieurs modèles se basant sur les traits ont été proposés afin de trouver le meilleur. Ainsi, la compatibilité des traits a été considérée entre le trait du mot à prédire et les traits de son contexte gauche récent (Cache de traits). Les meilleures performances ont été obtenues avec un Cache de traits de taille variable (modèle Cache-Trait Partiel). La combinaison

du modèle CTP avec un modèle bigramme a permis de diminuer la perplexité bigramme sur le corpus de test de plus de 8 points. Interpolé avec un modèle trigramme et en dépit du fait d'utiliser une liste de séparateurs de syntagmes sous-optimale, le modèle CTP nous a permis d'obtenir une amélioration de 6 points de la perplexité trigramme. Au vu de ces résultats, nous avons montré la faisabilité du concept des modèles de langage basés sur des traits linguistiques et non plus uniquement sur les formes orthographiques ainsi que la simplicité à les mettre en place. Il serait intéressant de tester ce modèle en deuxième passe d'un décodeur dans un système de Traduction Statistique afin d'attribuer un nouveau score aux n meilleures hypothèses de traduction et ainsi espérer favoriser les traductions automatiques avec les bons accords en genre et en nombre.

Annexe B

Exemple de traductions automatiques produites par PHARAOH sur le corpus de test EUROPARL

REF :	le programme de travail avance conformément au calendrier prévu .
KOEHN :	le programme de travail avance comme dans les temps .
TRIGGERS m-n :	le plan de travail progresse et selon le calendrier .
REF :	le vote aura lieu demain à CARD heures
KOEHN :	le vote aura lieu demain à CARD heures .
TRIGGERS m-n :	le vote aura lieu demain à CARD heures .
REF :	en réalité elles se félicitent des clarifications que ce document apporte au marché .
KOEHN :	en fait ils thereby elle met en lumière le marché .
TRIGGERS m-n :	en fait je me félicite des éclaircissements qu il apporte à l intérieur .
REF :	j ai écouté attentivement les utilisateurs et j ai procédé à de larges consultations .
KOEHN :	j ai écouté attentivement les usagers . j ai largement consulté .
TRIGGERS m-n :	j ai écouté avec attention à la utilisateurs et j ai largement consulté .
REF :	permettez moi de relier la problématique qui nous occupe à un événement personnel .
KOEHN :	arriveronsnous . je voudrais faire une déclaration d intérêt à ce stade .
TRIGGERS m-n :	peut qu être d intérêt . à cet endroit .
REF :	juste avant noël ma mère a inopinément perdu la vue .
KOEHN :	juste avant noël . ma mère subitement perdu son boisseau .
TRIGGERS m-n :	juste avant noël . ma mère improviste . je perdu de vue
REF :	j aurais honte si cet amendement CARD était adopté .
KOEHN :	président j aurais honte que l amendement CARD est adopté .
TRIGGERS m-n :	ce serait une honte pour l amendement CARD en une seule fois .
REF :	m. miller l a très bien fait dans son intervention .
KOEHN :	m.tomlinson a effectué du bon travail sur ce point .
TRIGGERS m-n :	il a accompli un bon travail . en ce

REF indique la phrase d'entrée du décodeur PHARAOH, KOEHN correspond à la meilleure hypothèse de traduction fournie par le décodeur avec la table de traduction de séquences construite d'après l'alignement bidirectionnel des mots comme proposé dans l'approche de Koehn [Koehn 03].

Enfin, TRIGGERS m-n donne la meilleure hypothèse de traduction fournie par le décodeur avec la table de traduction de séquences construites à partir des triggers inter-langues et ayant menée aux meilleures performances sur le corpus de test EUROPARL.

Bibliographie

[A. Lavie 04] K. Sagae A. Lavie et S. Jayaraman. *The Significance of Recall in Automatic Metrics for MT Evaluation.* In proceedings of the sixth Conference of the Association for Machine Translation in the Americas, pages 134–143, Washington, September 2004.

[Aarts 85] E. H. L. Aarts et V. Laarhoven. *Statistical cooling : A general approach to combinatorial optimization problems.* Philips J. Res., vol. 40, no. 4, pages 193–226, 1985.

[Agarwal 08] A. Agarwal et A. Lavie. *METEOR, M-BLEU and M-TER : Evalutaion Metrics for High-Correlation with Human Rankings of Machine Translation Output.* In Proceedings of the Third Workshop on Statistical Machine Translation, pages 115–118, Colombus, Ohio, Juin 2008.

[Babych 03] B. Babych, T. Hartley et E. Atwell. *Statistical Modelling of MT output corpora for Information Extraction.* In Proceedings of the Corpus Linguistics 2003 conference, pages 62–70, Lancaster University (UK), March 2003.

[Babych 04] B. Babych et T. Hartley. *Extending the BLEU MT Evaluation Method with Frequency Weightings.* In Proceedings of ACL 2004 (42nd Annual Meeting of the Association for Computational Linguistics), pages 621–628, Barcelone, Espagne, 2004.

[Banerjee 05] S. Banerjee et A. Lavie. *METEOR : An Automatic Metric for MT Evaluation with Improved Correlation with HUman Judgments.* In Proceedings of the ACL Workshop on Intrinic and Extrinsic Evaluation Measures for MT and/or Summarization, 2005.

[Black 98] E. Black, A. Finch et H. Kashioka. *Trigger-pair predictors in parsing and tagging.* In Proc. of ACL, 1998.

[Brants 07] T. Brants, A.C. Popat, P. Xu, F.J. Och et J. Dean. *Large Language Models in Machine Translation.* In Proc. of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natual Language Learning, Prague, June 2007.

[Brown 91] Peter F. Brown, Jennifer C. Lai et Robert L. Mercer. *Aligning Sentences in Parallel Corpora.* In Meeting of the Association for Computational Linguistics, pages 169–176, 1991.

[Brown 93] P. F. Brown et al. *The mathematics of statistical machine translation : parameter estimation.* Computational Linguistics, vol. 19, pages 263–311, 1993.

[Brown 96] R. Brown. *Example-Based Machine Translation in the Pangloss System.* In Proceedings of the 16th International Conference on Computational Linguistics, Copenhagen, Denmark, 1996.

[Brown 00] R. Brown. *Automated Generalization of Translation Examples.* In Proceedings of the Eighteenth International Conference on Computational Linguistics (COLING-2000), Saarbrücken, Germany, 2000.

[Callison-Burch 08] C. Callison-Burch, C. Fordyce, P. Koehn, C. Monz et J. schroeder. *Further Meta-Evaluation of Machine Evaluation.* In Proceedings of the Third Workshop on Statistical Machine Translation, pages 70–106, Colombus, Juin 2008.

[Callison-Burch 09] C. Callison-Burch, P. Koehn, C. Monz et J. Schroeder, editeurs. Proceedings of the fourth workshop on statistical machine translation. Association for Computational Linguistics, Athens, Greece, March 2009.

[Chandioux 96] J. Chandioux et A. Grimaila. *Specialized machine translation.* In In 2nd Conf. of the Association for Machine Translation in the Americas (AMTA 96), pages 206–212, Montreal, Canada, 1996.

[Dempster 77] A. Dempster, N. Laird et D. Rubin. *Maximum likelihood from incomplete data via the EM algorithm.* Journal of the Royal Statistical Society B, vol. 39, pages 1–38, 1977.

[DeNero 06] J. DeNero, D. Gillick, J. Zhang et D. Klein. *Why generative Phrase Models Underperform Surface Heuristics.* In Proceedings of the Workshop on Statistical Machine Translation, pages 31–38, New York City, June 2006.

[Deniz 08] N. Deniz et C. Turhan. *English to Turkish Example-Based Machine Translation with Synchronous SSTC.* In Proceedings of the Fifth International Conference on Information Technology : New Generations, 2008.

[Doddington 02] G. Doddington. *Automatic Evaluation of Machine Translation Quality Using N-gram Co-occurrence Statistics.* In Proceedings of HLT 2002 (Second Conference on Human Language Technology), pages 128–132, San Diago, CA, 2002.

[Fung 95] P. Fung. *A pattern matching method for finding noun and proper noun translations from noisy parallel corpora.* In Proceedings of the 33rd Annual Meeting, pages 236–243, Boston, MA, 1995. Association for Computational Linguistics.

[Gale 91a] W. A. Gale et K. W. Church. *Identifying word correspondence in parallel texts.* In HLT '91 : Proceedings of the workshop on Speech and Natural Language, pages 152–157, 1991.

[Gale 91b] William A. Gale et Kenneth Ward Church. *A Program for Aligning Sentences in Bilingual Corpora.* In Meeting of the Association for Computational Linguistics, pages 177–184, 1991.

[Gangadharaiah 06] R. Gangadharaiah, R. Brown et J. Carbonell. *Spectral Clustering for Example Based Machine Translation.* In Proceedings of the Human Language Technology Conference of the North American Chapter of ACL, pages 41–44, New-York, USA, June 2006.

[Germann 03] U. Germann. *Greedy Decoding for Statistical Machine Translation in Almost Linear Time.* In Proceedings of HLT-NAACL-2003, Edmonton, AB, Canada, 2003.

[Gimenez 08] J. Gimenez et L. Marquez. *A smorgasbord of Features for Automatic MT Evaluation.* In Proceedings of the Third Workshop on Statistical Machine Translation, pages 195–198, Colombus, Juin 2008.

[Giménez 06] J. Giménez et E. Amigó. *IQMT : A Framework for Automatic Machine Translation Evaluation.* In Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06), 2006.

[Hamon 08] O. Hamon, A. Popescu-Belis, A. Hartley, W. Mustafa El Hadi et M. Rajman. L'évaluation des technologies de traitement de la langue : les campagnes technolangue, chapitre CESTA : la Campagne d'Évaluation de Systèmes de Traduction Automatique, pages 93–116. Hermès, Paris, 2008.

[Hutchins 05] J. Hutchins. *Towards a definition of example-based machine translation.* In MT Summit X - Proceedings of Workshop on Example-Based Machine Translation, pages 63–70, Phuket, Thailand, September 2005.

[Jelinek 80] F. Jelinek et R. L. Mercer. Pattern recognition in practice, chapitre Interpolated Estimation of Markov Source Parameters from Sparse Data, pages 381–397. Amsterdam, Holland, 1980.

[Jur 06] Speech and language processing : An introduction to natural language processing, computational linguistics, and speech recognition, chapitre Machine Translation. 2006.

[Katz 87] S. M. Katz. *Estimation of probabilities from sparse data for the language model component of a speech recognizer.* IEEE Trans. ASSP, vol. 35, pages 400–401, 1987.

[Kim 04] W. Kim et S. Khudanpur. *Lexical triggers and latent semantic analysis for cross-lingual language model adaptation.* ACM Transactions on Asian Language Information Processing (TALIP), vol. 3, no. 2, pages 94–112, 2004.

[Kneser 95] R. Kneser et H. Ney. *Improved backing-off for m-gram language modeling stochastic language models.* In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, pages 181–184, 1995.

[Knight 99] K. Knight. *Decoding complexity in word-replacement translation models.* Computational linguistics, 1999.

[Koehn 03] P. Koehn, F. Och et D. Marcu. *Statistical Phrase-Based Translation.* In Proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference, pages 48–54, Edmonton, May-June 2003.

[Koehn 04] P. Koehn. *Pharaoh : A Beam Search Decoder for Phrase-Based Statistical Machine Translation Models.* In 6th Conference Of The Association For Machine Translation In The Americas, pages 115–224, Washington, DC, USA, 2004.

[Koehn 05] P. Koehn. *Europarl : A Multilingual Corpus for Evaluation of Machine Translation.* In MT Summit, Thailand, 2005.

[Koehn 07a] P. Koehn et H. Hoang. *Factored Translation Models.* In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Prague, Juin 2007.

[Koehn 07b] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin et E. Herbst. *Moses : Open Source Toolkit for Statistical Machine Translation.* In proceedings of ACL 2007, Prague, Juin 2007.

[Kuhn 90] R. Kuhn et R. DeMori. *A cache-based natural language model for speech recognition.* IEEE Trans. PAMI, vol. 12, no. 6, pages 570–582, 1990.

[Langlais 06] P. Langlais et F. Gotti. *EBMT by Tree-Phrasing.* Journal of Machine Translation, vol. 20, no. 1, pages 1–23, 2006.

[Lau 93] R. Lau, R. Rosenfeld et S. Roukos. *Trigger-based language models : a maximum entropy approach.* In Proc. ICASSP-93, 1993.

[Levenshtein 66] Vladimir Iosifovich Levenshtein. *Binary Codes Capable of Correcting Deletions, Insertions and Reversals.* Soviet Physics Doklady, vol. 10, no. 8, pages 707–710, February 1966.

[Li 09] Z. Li, C. Callison-Burch, S. Khudanpur et W. Thornton. *Decoding in Joshua : Open Source, Parsing-Based Machine Translation.* The Prague Bulletin of Mathematical Linguistics, no. 91, pages 47–56, 2009.

[Mangeot 05] M. Mangeot et E. Giguet. *Multilingual Aligned Corpora From Movie Subtitles.* Rapport technique, LISTIC, 2005.

[Marcu 02] Daniel Marcu et William Wong. *A phrase-based joint probability model for statistical machine translation.* In Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP, 2002.

[Melamed 96] I. Dan Melamed. *A Geometric Approach to Mapping Bitext Correspondence.* In Eric Brill et Kenneth Church, editeurs, Proceedings of the Conference on Empirical Methods in Natural Language Processing, pages 1–12. Association for Computational Linguistics, Somerset, New Jersey, 1996.

[Melamed 00] I. Dan Melamed. *Models of Translational Equivalence among Words.* Computational Linguistics, June 2000.

[Moore 02] Robert C. Moore. *Fast and Accurate Sentence Alignment of Bilingual Corpora.* In Proceedings of the Association for Machine Translation in the Americas Conference, pages 135–144, 2002.

[Nagao 84] M. Nagao. Artificial and human intelligence, chapitre A framework of a mechanical translation between japanese and english by analogy principle. Elsevier Science Publishers. B. V., 1984.

[Ney 94] H. Ney, U. Essen et R. Kneser. *On structuring probabilistic dependences in stochastic language modelling.* Computer Speech and Language, vol. 8, pages 1–38, 1994.

[Och 99] F. J. Och, C. Tillmann et H. Ney. *Improved alignment models for statistical Machine Translation.* In the joint conference of Empirical Methods in Natural Language Processing and Very Large Corpora, pages 20–28, University of Maryland, College Park, MD, 1999.

[Och 01] F.J. Och, N. Ueffing et H. Ney. *An efficient A* search algorithm for statistical machine translation.* In Proceedings of the ACL 2001 Workshop on data-Driven Methods in Machine Translation, Toulouse, France, 2001.

[Och 02] F. J. Och et H. Ney. *Discriminative Training and Maximum Entropy Models for Statistical Machine Translation.* In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), pages 295–302, Philadelphia, 2002.

[Och 03a] F. J. Och. *Minimum error rate training in statistical machine translation.* In Proceedings of ACL, Morristown, NJ, USA, 2003.

[Och 03b] F. J. Och et H. Ney. *A Systematic Comparison of Various Statistical Alignment Models.* Computational Linguistics, vol. 29, no. 1, pages 19–51, 2003.

[Papineni 01] K. Papineni, S. Roukos, T. Ward et W-J. Zhu. *Bleu : a Method for Automatic Evaluation of Machine Translation.* In Proceedings of the 40th Annual of the Association for Computational linguistics, pages 311–318, Philadelphia, USA, 2001.

[Rosenfeld 95] R. Rosenfeld. *The CMU statistical language modeling toolkit and its use in the 1994 ARPA CSR evaluation.* In Proceeding of the Spoken Language Systems Technology Workshop, pages 47–50, Austin, 1995.

[Rosenfeld 96] R. Rosenfeld. *A Maximum Entropy Approach to Adaptative Statistical Language Modeling.* Computer Speech and Language, vol. 10, pages 187–228, 1996.

[Salton 68] G. Salton et M.E. Lesk. *Computer evaluation of indexing and text processing.* Journal of ACM, vol. 15, no. 1, pages 8–36, 1968.

[Schwenk 09] H. Schwenk, S. Abdul Rauf, L. Barrault et J. Senellart. *SMT and SPE Machine Translation Systems for WMT'09.* In Proceedings of the Fourth Workshop on Statistical Machine Translation, pages 130–134, Athens, Greece, March 2009. Association for Computational Linguistics.

[Senellart 01] J. Senellart, P. Dienes et T. Varadi. *New Generation Systran Translation System.* In MT Summit VIII, Santiago de Compostela, Spain, September 2001.

[Singh 05] Ami Kumar Singh et Samar Husain. *Comparison, Selection and Use of Sentence Alignment Algorithms for New language pairs.* In Proceedings of the ACL Workshop on Building and using Parallel texts, pages 99–106, 2005.

[Snover 06] M. Snover, B. Dorr, R. Schwartz, L. Micciulla et J. Makhoul. *A Study of Translation Edit Rate with Targeted Human Annotation.* In In Proceedings of Association for Machine Translation in the Americas, 2006.

[Snover 09] M. Snover, N. Madnani, B. Dorr et R. Schwartz. *Fluency, Adequacy, or HTER ? Exploring Different Human Judgments with a Tunable MT Metric.* In Proceedings of the Fourth Workshop on Statistical Machine Translation at the 12th Meeting of the European Chapter of the Association for Computational Linguistics (EACL-2009), Athens, Greece, March 2009.

[T. Mitamura 91] E. H. Nyberg T. Mitamura et J. G. Carbonell. *An Efficient Interlingua Translation System for Multi-lingual Document Production.* In Proceedings of Machine Translation Summit III, pages 2–4, 1991.

[Tillmann 96] C. Tillmann et H. Ney. Selection criteria for word trigger pairs in language modeling, pages 98–106. Lecture Notes in Artificial Intelligence 1147, Springer Verlag, 1996.

[Tillmann 97a] C. Tillmann et H. Ney. *Word Trigger and the EM Algorithm.* In Proceedings of the Conference on Computational Natural Language Learning, pages 117–124, Madrid, Spain, 1997.

[Tillmann 97b] C. Tillmann, S. Vogel, H. Ney, A. Zubiaga et H. Sawaf. *Accelerated DP based search for statistical translation.* In Fifth European Conf. on Speech Communication ans Technology, pages 2667–2670, Rhodos, Greece, September 1997.

[Tillmann 00] C. Tillmann et H. Ney. *Word reordering and DP-based search on statistical machine Translation.* In Proceedings of the 18th COLING, pages 850–856, Allemagne, 2000.

[Tillmann 03] C. Tillmann. *A projection extension algorithm for statistical machine translation.* In Proceedings of the 2003 conference on Empirical methods in natural language processing, pages 1–8, Morristown, NJ, USA, 2003. Association for Computational Linguistics.

[Tinsley 08] J. Tinsley, Y. Ma, S. Ozdowska et A. Way. *Matrex : the DCU MT system for WMT 2008.* In Proceedings of the Third Workshop on Statistical Machine Translation, pages 171–174, Colombus, Ohio, Juin 2008.

[Turain 03] J.P. Turain, L. Shen et I.D. Melamed. *Evaluation of Machine Translation and its Evaluation.* In Proceedings of MT Summit IX, New Orleans, U.S.A., 2003.

[Uwe 06] M. Uwe. *Fully Automatic High Quality Machine Translation of Restricted Text : A Case Study.* In Proceedings of the Twenty-eighth International Conference on Translating and the Computer, London, 2006.

[Vandeghinste 04] V. Vandeghinste et E.Tjong Kim Sang. *Using a Parallel Transcript/Subtitle Corpus for Sentence Compression.* In LREC, Lisbon, Portugal, 2004.

[Vasconcellos 85] M. Vasconcellos et M. Leon. *SPANAM and ENGSPAN : Machine Translation at the Pan American Health Organization.* Computational Linguistics, vol. 11, no. 2-3, pages 122–136, 1985.

[Veale 97] T. Veale et A. Way. *Gaijin : A Template-Driven Bootstrapping Approach to Example-Based Machine Translation.* In Proceedings of NeMNLP97, Sofia, Bulgaria, September 1997.

[Vogel 96] S. Vogel, H. Ney et C. Tillmann. *HMM-based word alignment in statistical translation.* In Proc. of the Conference on Computational Linguistics, pages 836–841, Morristown, NJ, USA, 1996.

[Wang 97] Y. Wang et A ; Waibel. *Decoding algorithm in statistical machine translation.* In Proceedings of the 35th ACL, pages 366–372, Madrid, Esapgne, 1997.

[Wheeler 84] P.J. Wheeler. *Changes and improvements to the european commission s systran MT system 1976/84.* Terminologie (Luxembourg), 1984.

[Witkam 88] T. Witkam. *DLT An Industrial R&D Project for Multilingual MT*, 1988.

[Yamada 01] K. Yamada et K. Knight. *A syntax-based statistical translation model.* In Proc. of ACL, 2001.

[Yu 04] Jian-Heng Yu. *Alignment of Bilingual Web pages based on the MT evaluation method of BLEU.* Rapport technique, National Tsing Hua University, 2004.

[Zens 02] R. Zens, F.J. Och et H. Ney. *Phrase-based Machine Translation*. In G. Lakemeyer (Eds.) M. Jarke J. Koehler, editeur, Advances in Artificial Intelligence. 25. Annual German Conference on AI, KI 2002, September 2002.

[Zettlemoyer 07] L. Zettlemoyer et R. C. Moore. *Selective Phrase Pair Extraction for Improved Statistical Machine Translation*. In Proceedings of HLT-NAACL, Rochester, NY, 2007.

[Zhang 03] Ying Zhang, Stephan Vogel et Alex Waibel. *Integrated Phrase Segmentation and Alignment Model for Statistical Machine Translation*. In Proceedings of International Conference on Natural Language Processing and Knowledge Engineering, Beijing, China, 2003.

[Zhang 06] Y. Zhang, A.S. Hildebrand et S. Vogel. *Distributed Language Modeling for N-best List Re-ranking*. In Proc. of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006), Sydney, July 2006.

[Zitouni 03] I. Zitouni, K. Smaïli et J.-P. Haton. *Statistical language modeling based on variable-length sequences*. Computer Speech and Language, vol. 17, pages 27–41, 2003.