Estimation d'un modèle de traduction à partir d'alignements mot-à-mot non-déterministes

Nadi Tomeh Alexandre Allauzen François Yvon
Université Paris Sud et LIMSI/CNRS
BP 133 91 403 Orsay
{nadi,allauzen,yvon}@limsi.fr

Résumé. Dans les systèmes de traduction statistique à base de segments, le modèle de traduction est estimé à partir d'alignements mot-à-mot grâce à des heuristiques d'extraction et de valuation. Bien que ces alignements mot-à-mot soient construits par des modèles probabilistes, les processus d'extraction et de valuation utilisent ces modèles en faisant l'hypothèse que ces alignements sont déterministes. Dans cet article, nous proposons de lever cette hypothèse en considérant l'ensemble de la *matrice d'alignement*, d'une paire de phrases, chaque association étant valuée par sa probabilité. En comparaison avec les travaux antérieurs, nous montrons qu'en utilisant un modèle exponentiel pour estimer de manière discriminante ces probabilités, il est possible d'obtenir des améliorations significatives des performances de traduction. Ces améliorations sont mesurées à l'aide de la métrique BLEU sur la tâche de traduction de l'arabe vers l'anglais de l'évaluation *NIST MT'09*, en considérant deux types de conditions selon la taille du corpus de données parallèles utilisées.

Abstract. In extant phrase-based statistical translation systems, the translation model relies on word-to-word alignments, which serve as constraints for further heuristic extraction and scoring processes. These word alignments are infered in a probabilistic framework ; yet, only one single best word alignment is used as if alignments were deterministically produced. In this paper, we propose to take the full probabilistic alignment matrix into account, where each alignment link is scored by its probability score. By comparison with previous attempts, we show that using an exponential model to compute these probabilities is an effective way to achieve significant improvements in translation accuracy on the *NIST MT'09* Arabic to English translation task, where the accuracy is measured in terms of BLEU scores.

Mots-clés : traduction statistique, modèles de traduction à base de segments, modèles d'alignement mot-à-mot.

Keywords: statistical machine translation, phrase based translation models, word alignment models.

1 Introduction

Dans les systèmes de traduction statistique à base de segments (*phrase-based systems*), le *modèle de traduction* sert de pont entre les langues source et cible. Sur la base d'hypothèses de segmentation de la phrase source à traduire, il permet de proposer, pour chacun des segments, des traductions candidates en langue cible. Ces hypothèses de traduction sont sélectionnées dans un inventaire qui enregistre des appariements valués entre segments de longueur variable. Ces associations et les scores qui les accompagnent constituent la table de traductions (*phrase-table*).

Ce modèle est estimé en deux temps à partir d'un corpus parallèle : (i) extraction d'un ensemble de couples de segments candidats, (ii) valuation des couples retenus dans la phase (i). Faute de disposer de méthodes d'estimation théoriquement bien fondées, chacune de ces deux étapes repose sur un ensemble d'heuristiques. Il s'avère en effet impossible d'estimer directement les valuations calculées en (ii), ni même de recencer tous les appariements possibles en (i). En effet, estimer de façon non-supervisée un modèle probabiliste des alignements de segments demanderait de pouvoir calculer des sommes sur tous les alignements de segments possibles, à défaut, de savoir calculer un alignement optimal utilisant des segments de taille variable. Ces deux procédures posent des problèmes combinatoires NP-difficiles (DeNero & Klein, 2008) et ne peuvent être effectuées de manière exacte. De manière plus subtile, construire des modèles d'alignements de segments demande de mettre en compétition des segmentations conjointes de taille variable des phrases source et cible, au risque de toujours préférer les alignements impliquant des segments longs. Enfin, ne considérer qu'une seule segmentation lors de l'apprentissage semble avoir un effet négatif sur la capacité de généralisation du modèle (DeNero *et al.*, 2006).

La solution pratique qui s'est progressivement imposée contourne le problème en considérant en premier lieu une segmentation

maximale et en effectuant un alignement préalable au niveau des mots ; des procédures efficaces fondées sur l'algorithme EM (*Expectation-Maximisation*) pour effectuer cet alignement de manière efficace existent depuis le début des années 90 (Brown *et al.*, 1993; Och & Ney, 2003). Ces alignements de mots sont ensuite ré-analysés pour en déduire des alignements de segments, la méthode la plus répandue consistant à extraire les alignements de segments *compatibles* avec les contraintes posées par les alignements de mots.

Dans un troisième temps, les statistiques d'occurrence de ces alignements de segments sont collectées et utilisées pour attribuer des scores de confiance à ces groupes bilingues. Ces trois étapes successives de la construction du modèle de traduction sont usuellement abordées et optimisées séparément les unes des autres. Le risque est naturellement que les erreurs s'accumulent le long de cette séquence de traitements. Ainsi, des erreurs précoces dans les calculs des alignements mot-à-mot viennent bruiter le processus d'extraction des couples de segments appariés et biaiser les calculs de scores afférents.

Pour pallier ce problème, les auteurs de (Liu *et al.*, 2009) proposent d'extraire davantage d'informations de la phase d'alignement des mots, sous la forme d'une *matrice d'alignements pondérés*, qui représente de manière compacte un ensemble d'alignements de mots potentiels. Cette matrice est utilisée dans les étapes ultérieures de l'apprentissage. Dans une matrice pondérée, chaque lien d'alignement potentiel est nanti d'une probabilité qui mesure la confiance dans l'alignement de ces deux mots. Dans (Liu *et al.*, 2009), ces probabilités sont estimées à partir du calcul des n-meilleurs alignements de mots tels que produits par les modèles d'alignement standards. À l'aide de cette technique, ces auteurs parviennent à améliorer de façon modeste leurs systèmes de traduction automatique. Une des limites de cette approche est toutefois l'utilisation d'une liste de n-meilleurs, qui ne représente que très imparfaitement la diversité et la variabilité des alignements de mots potentiels, et conduit à des mauvais estimateurs des probabilités *a posteriori* des liens d'alignement.

Dans ce travail, nous soutenons qu'une meilleure estimation des probabilités des liens d'alignement est susceptible de donner lieu à de meilleurs modèles. Nous étudions donc une méthode alternative pour réaliser cette estimation, fondée sur des modèles discriminants pour l'alignement de mots (Ayan & Dorr, 2006; Tomeh *et al.*, 2010, 2011) et analysons les performances qu'elles permettent d'obtenir. La principale contribution de ce travail est donc de nature empirique : en comparant différentes manières de calculer et d'exploiter ces matrices d'alignement pondérées, nous montrons qu'il peut être bénéfique, en particulier quand les données d'apprentissage du modèle de traduction sont réduites, de prendre en compte l'information contenue dans ces matrices, au-delà du meilleur alignement mot-à-mot.

Cet article est organisé comme suit. Après avoir brièvement posé le cadre de la construction du modèle de traduction dans l'approche standard, nous présentons à la section 2 les principes de construction et d'exploitation de matrices d'alignements pondérées. Nous introduisons, à la section 3 une approche alternative permettant d'estimer directement la matrice d'alignement pondérée. Les résultats expérimentaux sont ensuite décrits à la section 4. Enfin, nous explicitons le positionnement de notre approche par rapport aux travaux existants, avant de conclure et d'évoquer diverses pistes vers lesquelles nous comptons nous orienter dans le futur.

2 Matrices pondérées pour la construction de modèles de traduction

Pour un système de traduction à base de segments (Zens *et al.*, 2002), le modèle de traduction est la source de connaissance principale qui établit une correspondance entre les deux langues (source et cible). Son rôle est de guider la construction, pour chaque phrase source, d'un ensemble d'hypothèses de traduction en langue cible. L'unité de traduction est le segment, qui correspond à un groupe de mots contigus. L'association entre un segment source et une traduction possible en cible forme un bi-segment. Notons qu'il est possible qu'un segment admette plusieurs traductions alternatives, donnant lieu à plusieurs bi-segments partageant le même segment source. Afin de faire un bon usage de ces bi-segments, il est nécessaire de leur associer des mesures, par exemple statistiques, qui quantifient la confiance en l'association ainsi réalisée.

Dans la suite de cet article, nous utilisons les notations suivantes : un couple de phrases est désigné par $(\mathbf{e}, \mathbf{f})$, où la phrase source $\mathbf{f} = f_1, ..., f_i, ..., f_I$ est une séquence de I mots et la phrase cible $\mathbf{e} = e_1, ..., e_j, ..., e_J$ est une séquence de J mots. De plus, une sous-séquence de mots extraite d'une phrase sera notée $f_{i_1}^{i_2} = f_{i_1} \dots f_{i_2}$ et donc $\mathbf{f} = f_1^I$.

2.1 Cadre général

Les méthodes décrites dans la littérature pour construire le modèle de traduction peuvent se résumer par l'algorithme présenté dans la partie gauche de la figure 1. Le point de départ est un couple de phrases accompagné d'un alignement mot-à-mot représenté par une *matrice d'alignement*. Chaque cellule de cette matrice booléenne $\mathbf{A} = \{a_{i,j} : 1 \leq i \leq I, 1 \leq j \leq J\}$ représente un lien

1: **POUR** toutes les paires de phrases (f_1^J, e_1^I) **FAIRE**
2: **POUR** tous les segments f_{j1}^{j2} **FAIRE**
3: Construire l'ensemble des bi-segments $E_A = \{f_{j1}^{j2}, e_{i1}^{i2}\}$ satisfaisant le jeu de contraintes $\mathcal{C}_\mathcal{A}$
4: Trier E_A selon la fonction f_R
5: Appliquer le critère de sélection $\mathcal{C}_\mathcal{S}$ définissant l'ensemble E_{AS} des bi-segments à extraire
6: Assigner une fonction de compte f_C à chaque bi-segments $(f_{j1}^{j2}, e_{i1}^{i2})$ de E_{AS}
7: **end POUR**
8: **end POUR**
9: **POUR** chaque bi-segments extraite $\{(e, f)\}$ **FAIRE**
10: Calcul des scores :

$$\phi(e|f) = \frac{f_C(e,f)}{\sum_{f_i} f_C(e,f_i)},$$

$$lex(e|f, \mathbf{A}) = \prod_{i=1}^{length(e)} \frac{1}{|\{j : (i,j) \in \mathbf{A}\}|} \sum_{\forall (i,j) \in a} w(e_i|f_j),$$

où $\mathbf{A}$ désigne la matrice d'alignement, et w une probabilité de traduction lexicale (IBM1 ou fréquence relative).
11: **end POUR**

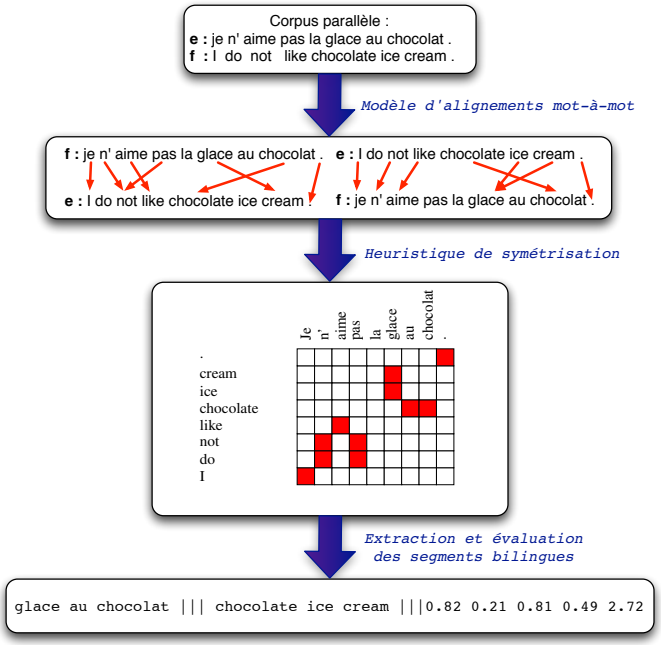

FIGURE 1 – Algorithme générique pour la construction du modèle de traduction et un exemple de son application fréquement utilisé

d'alignement potentiel ; la variable binaire $a_{i,j}$ vaut 1 si le lien entre le $i^{\text{ème}}$ mot de **f** et le $j^{\text{ème}}$ mot de **e** est actif, et 0 sinon.

Un jeu de contraintes $\mathcal{C}_\mathcal{A}$ permet de définir, parmi tous les bi-segments potentiellement contenus dans une paire de phrases, ceux qui sont « acceptables » ou cohérents avec la matrice d'alignement. Les contraintes apportées par les alignements de mots permettent l'énumération conjointe de toutes les segmentations de la paire de phrases avec tous les alignements de segments autorisés. Une fois cet ensemble de bi-segments identifié, il est possible de le trier (f_R) et de lui appliquer un critère de séléction $\mathcal{C}_\mathcal{S}$ afin d'éliminer les bi-segments qui semblent *a priori* les moins plausibles. La dernière étape concerne la valuation des bi-segments ainsi extraits. Les fonctions de valuation les plus communément utilisées sont :
– la fréquence d'observation du segment e connaissant le segment f notée $\phi(e|f)$ ainsi que le terme symétrique $\phi(f|e)$;
– les poids lexicaux ou *lexical weights* dans les deux directions ($lex(e|f, \mathbf{A})$ et $lex(f|e, \mathbf{A})$), qui utilisent, le plus souvent, les probabilités de traduction lexicale du modèle IBM1.
Ces fonctions sont définies dans l'algorithme détaillé sur la figure 1 (ligne 10).

L'instanciation standard de cet algorithme correspond aux travaux de (Zens *et al.*, 2002; Koehn *et al.*, 2003) (voir partie droite de la figure 1), qui se déduit du cadre général en utilisant les définitions suivantes :
– $\mathcal{C}_\mathcal{A}$ représente des contraintes de cohérence qui s'appliquent à un alignement mots-à-mots symétrisé d'une paire de phrases. Ces alignements se déduisent des deux meilleures hypothèses données par le modèle *IBM4* (une pour chaque direction de traduction), symétrisées par l'heuristique *grow-diag-final-and* (Koehn *et al.*, 2003).
– La fonction de compte et celle de tri sont les mêmes : $f_R = f_C = 1$
– la contrainte $\mathcal{C}_\mathcal{S}$ est définie par un seuil portant sur la longueur relative des segments source et cible et permet de filtrer les bi-segments trop longs.
Les hypothèses simplificatrices utilisées dans l'approche standard permettent d'obtenir une procédure efficace et robuste ; elles soulèvent néanmoins quelques critiques. Tout d'abord, le choix du modèle IBM4 pose problème puisque sa complexité interdit d'utiliser des procédures exactes lors de l'inférence et du calcul des probabilités *a posteriori* de chacun des liens d'alignement. Ainsi, les contraintes de cohérence des bi-segments portent sur des alignements qui ne sont pas forcément les meilleurs et pour lesquels les approximations des probabilités *a posteriori* ne reflètent qu'imparfaitement la confiance du modèle. Ce dernier point implique naturellement le choix des fonctions de compte et de tri $f_C = f_R = 1$, puisqu'en l'absence de mesure de confiance, une décision binaire s'impose. Enfin, ces simplifications entraînent que l'exploration de la matrice d'alignement est restreinte à la sous-partie sélectionnée par les alignements IBM4 et ne prend pas en considération la plus grande partie de la matrice d'alignement.

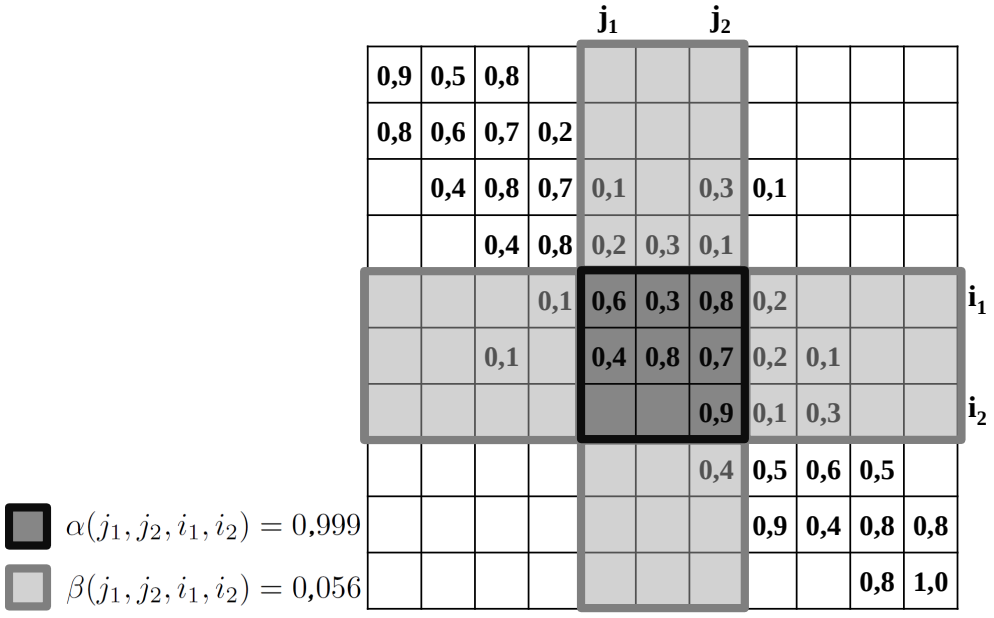

FIGURE 2 – Illustration du calcul des comptes fractionnaire pour un bi-segment donné. Dans cet exemple, le calcul des comptes fractionnaires se fait de la manière suivante : $f_C(f_{j_1}^{j_2}, e_{i_1}^{i_2}) = \alpha(j_1, j_2, i_1, i_2) \times \beta(j_1, j_2, i_1, i_2)$.

2.2 La matrice d'alignement pondérée

Dans (Liu *et al.*, 2009), les auteurs proposent d'augmenter le nombre des alignements mot-à-mot qui sont impliqués dans l'estimation des modèles de traduction et introduisent, à cet effet, la notion de *matrice d'alignement pondérée* : $\mathbf{A_p} = \{p(a_{i,j}|\mathbf{e}, \mathbf{f}) : 1 \leq i \leq I, 1 \leq j \leq J\}$. Dans cette matrice, chaque lien d'alignement est pondéré par sa probabilité *a posteriori* $p(a_{i,j}|\mathbf{e}, \mathbf{f})$. Ces probabilités sont calculées à partir des n-meilleurs alignements symétrisés proposés par le modèle IBM4. Partant de cette matrice, l'algorithme représente à la figure 1 est modifié de la manière suivante :

– Les contraintes de cohérence $\mathcal{C}_\mathcal{A}$ stipulent qu'un bi-segment est acceptable si au moins un lien d'alignement $a_{i,j}$ à l'intérieur du bi-segment est tel que $p(a_{i,j}|\mathbf{e}, \mathbf{f})$ est supérieur à un certain seuil.
– Les fonctions de compte $f_C = f_R$ prennent en compte le caractère non-déterministe des liens d'alignement de la manière suivante. Pour un bi-segment $f_C(f_{j_1}^{j_2}, e_{i_1}^{i_2})$:

$$f_C(f_{j_1}^{j_2}, e_{i_1}^{i_2}) = \alpha(j_1, j_2, i_1, i_2) \times \beta(j_1, j_2, i_1, i_2) \text{ avec} \tag{1}$$

$$\alpha(j_1, j_2, i_1, i_2) = 1 - \prod_{(j,i) \in in(j_1, j_2, i_1, i_2)} \bar{p}(a_{i,j}|\mathbf{e}, \mathbf{f}), \tag{2}$$

$$\beta(j_1, j_2, i_1, i_2) = \prod_{(j,i) \in out(j_1, j_2, i_1, i_2)} \bar{p}(a_{i,j}|\mathbf{e}, \mathbf{f}) \tag{3}$$

où $\bar{p}(a_{i,j}|\mathbf{e}, \mathbf{f}) = (1 - p(a_{i,j}|\mathbf{e}, \mathbf{f}))$, le coefficient α correspond à la confiance accordée au lien à l'intérieur (in) du bi-segment et β quantifie la masse totale de probabilité des liens situés à l'extérieur (out) de ce bi-segment. L'estimation de cette fonction est illustrée à la figure 2.

Avec ces nouvelles définitions, l'évaluation des bi-segments doit être modifiée pour également prendre en compte les probabilités des alignements. La fonction ϕ ne nécessite pas de modification, puisqu'elle utilise la fonction f_C, qui a été redéfinie. En revanche, les poids lexicaux sont maintenant définis comme suit :

$$lex(e|f, \mathbf{A_p}) = \prod_{i=1}^{|e|} \left(\left(\frac{1}{\{j|p(a_{i,j}|\mathbf{e}, \mathbf{f}) > 0\}} \sum_{\forall j: p(a_{i,j}|\mathbf{e}, \mathbf{f}) > 0} w(e_i|f_j) p(a_{i,j}|\mathbf{e}, \mathbf{f}) \right) + w(e_i|f_0) \prod_{j=1}^{|f|} \bar{p}(a_{i,j}|\mathbf{e}, \mathbf{f}) \right). \tag{4}$$

L'une des hypothèses explorée dans notre travail est que les gains modestes obtenus par (Liu *et al.*, 2009) sont dus à la méthode utilisée pour estimer cette matrice pondérée, qui s'appuie sur un petit ensemble d'alignements calculés par le modèle IBM4. En

effet l'échantillonnage des alignements en ne considérant que les n-meilleures hypothèses des modèles IBM4 ($n = 10$ en pratique) revient à considérer qu'un sous-ensemble qui ne contient que peu de variation et beaucoup de redondance. Ainsi, l'exploration de la matrice d'alignement est par construction très limitée et l'estimation approximative. Par ailleurs, le calcul de la matrice d'alignement s'appuie sur une procédure *ad hoc* de recombinaisons des probabilités *a posteriori* des alignements initialement calculés séparément pour chaque direction de traduction.

L'alternative que nous proposons d'explorer consiste à estimer cette matrice en utilisant une modélisation directe de la probabilité d'un lien d'alignement en utilisant des modèles conditionnels exponentiels qui seront décrits à la section 3.

3 Modélisation de la matrice d'alignement

Un alignement mot à mot entre une phrase source, et sa traduction (la phrase cible) regroupe un ensemble de liens décrivant une relation de traduction entre mots. Ainsi, prédire la matrice d'alignement peut être envisagé comme un problème de classification supervisé pour des données structurées. Lorsque des données étiquettées sont disponibles, la solution proposée dans (Ayan & Dorr, 2006; Tomeh *et al.*, 2010, 2011) consiste à estimer de manière indépendante la probabilité de chaque lien dans la matrice à l'aide d'un modèle de régression logistique défini par :

$$p\left(y|\boldsymbol{x}\right) = \frac{1}{Z(\boldsymbol{x})} \exp\left(\sum_{k=1}^{K} \lambda_k f_k\left(y, \boldsymbol{x}\right)\right), \tag{5}$$

où y désigne la variable aléatoire binaire qui indique si un lien est actif, $\boldsymbol{x}$ l'observation, $Z(\boldsymbol{x})$ le facteur de normalisation, $(f_k)_{k=1}^K$ définit un ensemble de fonctions caractéristiques, et $(\lambda_k)_{k=1}^K$ les poids associés. Dans l'équation (5), l'observation $\boldsymbol{x}$ désigne la paire de phrases augmentée de son étiquetage morphosyntaxique et des liens d'alignement produits par les modèles génératifs.

Cette formulation du problème permet de modéliser directement chaque cellule de la matrice d'alignement. Mais elle peut être également perçue comme une manière de fusionner différents alignements d'une paire de phrases. Cette approche permet donc également de remplacer l'étape heuristique de symétrisation, nommée *grow-diag-final-and* (Koehn *et al.*, 2003) dans l'approche standard, par un modèle d'apprentissage statistique pouvant prendre en compte un nombre arbitraire d'alignements en entrée.

Estimer ce modèle à partir d'exemples demande néanmoins de prendre en considération le caractère très creux de la matrice d'alignement, conséquence du fait qu'une forte majorité de liens sont inactifs. La tâche de classification considérée est donc très déséquilibrée. Afin d'éviter d'apprendre un classifieur trop biaisé en faveur de la prédiction de liens inactifs, l'ensemble des liens à étiqueter peut être au préalable réduit à un sous-ensemble de la matrice. Pour définir ce sous-ensemble, les modèles génératifs classiques sont utilisés (modèles de Markov cachés et/ou IBM4 dans les deux directions) : tout lien qui n'apparait pas dans un des alignements génératifs est considéré comme inactif ; les autres liens sont évalués par le modèle de classification. Dans ce cadre, les alignements génératifs sont utilisés pour réduire l'espace de recherche et permettent de limiter l'effet potentiellement néfaste de données déséquilibrées (Ayan & Dorr, 2006; Elming & Habash, 2007).

Ce modèle est utilisé pour estimer la matrice pondérée d'alignement $\mathbf{A_p}$ décrite à la section 2.2. Le classifieur supervisé estime donc la probabilité $p(a_{i,j}|\mathbf{e}, \mathbf{f})$ pour chaque cellule de la matrice.

Apprentissage L'estimation des paramètres du modèle (les λ_k dans l'équation (5)) est faite de manière à maximiser la vraisemblance conditionnelle régularisée à partir d'un corpus d'entraînement. La régularisation utilisée est connue sous le nom d' *elastic-net* (Zou & Hastie, 2005) et combine un terme de régularisation ℓ^1, qui aide à sélectionner les fonctions caractéristiques les plus utiles et ainsi réduire la taille du modèle, et un terme de régularisation ℓ^2, qui garantit que le Hessien de la fonction objectif n'est jamais trop proche de zéro, et permet ainsi d'éviter les problèmes d'instabilité numérique. Ce choix de régularisation nous permet d'envisager de nombreuses fonctions caractéristiques, sachant que certaines d'entre elles seront éliminées lors de l'entraînement car jugées inutiles.

Les caractéristiques Les fonctions caractéristiques utilisées pour le classifieur sont décrites en détail dans (Tomeh *et al.*, 2010) et reprennent en partie celles proposées par (Ayan & Dorr, 2006). Elles prennent en compte les multiples sources d'informations : la paire de phrases augmentée de son étiquetage morphosyntaxique et les liens d'alignement produits par les différents modèles génératifs considérés. Ainsi, pour un lien d'alignement donné, ces fonctions binaires indiquent par exemple : l'association entre les mots source/cible et de même pour les étiquettes morphosyntaxiques associées ; quel modèle génératif propose ce lien comme actif

ainsi que le nombre total de modèles génératifs proposant ce lien comme actif ; combien de liens sont proposés par les modèles génératifs dans le voisinage ; la fertilité du mot source (et resp. du mot cible) considérant l'ensemble des alignements d'entrée ; l'écart du lien à la diagonale afin de favoriser ou non les alignements monotones ; la distance du lien avec le mot aligné le plus proche (en source et en cible) afin de caractériser si ce lien est isolé des autres.

À ces caractéristiques s'ajoutent celles que nous allons décrire. Une première famille de fonctions caractéristiques décrit les mots source et cible relatifs à un lien d'alignement (*i.e* une case de la matrice) :
– Probabilité de traduction lexicale pour le couple de mots utilisé : $p(f_i|e_j)$ et $p(e_i|f_j)$ estimées par le modèle IBM1.
– La fréquence des mots source et cible ainsi que leur ratio.
– Un test sur tous les préfixes et suffixes de longueur 3.
– La similarité entre les mots source et cible calculée par la distance d'édition. Cette caractéristique tente de capturer la propension qu'ont les noms propres à être traduits de manière similaire, comme par exemple : *SdAm Hsyn* et *Saddam Hussein*.
– Un test portant sur l'égalité entre les mots source et cible.
– Un test indiquant si l'un des mots est une ponctuation associé avec un mot qui n'est pas une ponctuation.
Nous avons également définit un ensemble de fonctions caractéristiques permettant de décrire la structure de la matrice et les liens qui la composent. En plus des fonctions décrites dans (Tomeh *et al.*, 2010), nous ajoutons la fonction qui indique si un lien d'alignement implique un mot dupliqué dans l'une des phrases. Cette caractéristique permet de pallier une faible modélisation de la distorsion. Par exemple le mot arabe *fy* peut apparaître plusieurs fois dans une même phrase et être ainsi toujours aligné avec le même mot *in* en cible. Cette fonction retourne la distance du lien considéré à la diagonale.

4 Expériences

Pour évaluer les différentes approches, nous utilisons la tâche de traduction de l'arabe vers l'anglais de l'évaluation *NIST MT'09*. Nous comparons quatre méthodes d'estimation de la matrice pondérée : l'approche standard qui utilise les modèles d'alignement IBM4 et les heuristiques d'extraction et de valuation usuelles ; la méthode décrite dans le premier article sur les matrices pondérées (Liu *et al.*, 2009) ; le système *PostCAT* (Graça *et al.*, 2010) (décrit brièvement à la section 4.1) ; et l'estimation directe de la matrice via un modèle de régression logistique. Le système de traduction utilisé est MOSES(Koehn *et al.*, 2007), un outil sous licence libre.

4.1 Corpus et outils

Pour entraîner le modèle logistique, nous avons utilisé *Wapiti* (Lavergne *et al.*, 2010) [1], avec comme corpus d'apprentissage et de développement les données alignées manuellement du corpus IBMAC (Ittycheriah *et al.*, 2006), contenant respectivement 10 000 et 663 paires de phrases. Nous avons construit 2 sous-ensembles, de taille différente, de données parallèles pour entraîner le système de traduction, afin d'évaluer l'impact du volume de données disponibles sur les résultats obtenus. Ces deux corpus ont été constitués à partir des données autorisées dans la tâche *contrainte* de l'évaluation *NIST MT'09*. Elles sont toutes disponibles via le *Linguistic Data Consortium* [2].

Nous avons ainsi défini 2 tâches, l'une avec un corpus parallèle de 30 000 phrases (30k) et l'autre avec 130 000 phrases (130k). Les systèmes de traduction sont construits avec MOSES [3] en utilisant la configuration par défaut. Les paramètres de ces systèmes sont optimisés de manière usuelle avec l'outil MERT (*Minimum Error Rate Training*) avec comme données de développement le corpus *NIST MT'06* contenant 1 800 phrases arabes et 4 traductions anglaises. Les traductions produites sont évaluées avec la métrique BLEU (Papineni *et al.*, 2002) sur les données d'évaluation *NIST MT'08*, qui contiennent 1 400 phrases et 53k mots.

Pour le système *PostCAT* [4] et l'extraction des unités de traduction [5], nous avons utilisé les outils libres disponibles sur la toile. Enfin les modèles de langue cible ont été appris avec la boîte à outils *SRILM* [6] en utilisant toutes les données monolingues autorisées dans le cadre de l'évaluation *NIST MT'09* (pour plus de détails, on se reportera à (Allauzen *et al.*, 2009)).

La partie anglaise des données est pré-traitée de manière classique (les méthodes utilisées sont décrites dans (Allauzen *et al.*, 2009)).

1. http ://wapiti.limsi.fr/
2. La description complète est disponible à l'adresse http ://www.itl.nist.gov/iad/mig/tests/mt/2009/
3. http ://www.statmt.org/moses/
4. `http://www.seas.upenn.edu/~strctlrn/CAT/CAT.html`
5. `http://www.nlp.org.cn/~liuyang/wam/wam.html`.
6. `http://www-speech.sri.com/projects/srilm/`.

Pour la partie arabe, toutes les phrases sont analysées et segmentées avec l'outil MADA+TOKAN [7]. Nous avons utilisé le schéma de segmentation D2 afin de tenir compte de la morphologie riche de l'arabe et ainsi segmenter les mots arabes en des unités qui correspondent approximativement aux mots anglais.

4.2 Construction des modèles de traduction

Dans la section 2, nous avons décrit un algorithme générique pour la construction d'un modèle de traduction. Cet algorithme fonctionne en trois étapes séparées : construction des matrices d'alignement pondérées, extraction puis évaluation des bi-segments. Nous allons maintenant évaluer l'impact de ces trois étapes sur les résultats en traduction automatique.

Pour la première étape, nous expérimentons deux manières de construire les matrices pondérées :
(i) la méthode standard qui ne considère que les meilleurs alignements
(ii) la matrice pondérée par les probabilités qui est utilisée dans le processus d'extraction et de valuation.
Notons qu'il est possible de passer de la configuration (ii) à (i) par un simple seuillage sur les probabilités. Dans toutes nos expériences, nous utilisons un seuil de $0,5$. Ainsi, pour chaque modèle d'alignement, deux types de systèmes sont construits : *standard* (configuration (i)) et *WAM* pour la matrice pondérée (configuration (ii)). Le corpus de référence *IBMAC* contient également un jeu de test qui est utilisé pour calculer le taux d'erreur d'alignement (ou AER, pour *Alignment Error Rate*).

Les deux autres étapes (extraction et valuation des bi-segments) dépendent du mode de construction de la matrice d'alignement. Dans le cas *standard*, les bi-segments sont extraits et évalués selon les heuristiques décrites à la section 2.1. Lorsque l'on utilise des matrices pondérées, nous utilisons les méthodes d'extraction et de valuation décrites à la section 2.2, qui prennent en compte la probabilité des liens d'alignement. Pour cette dernière approche, seuls les bi-segments dont la probabilité est supérieure à un seuil sont conservés. Ceci permet, comme le préconisent les auteurs de (Liu *et al.*, 2009), de restreindre le nombre de bi-segments qui sont extraits. De plus, comme cela est fait dans l'approche standard, les bi-segments comprenant un segment de longueur supérieur à 7 sont également rejetés. Comme il est d'usage, les performances en traduction automatique sont évaluées par la métrique BLEU (Papineni *et al.*, 2002).

4.3 Modèles d'alignement mot-à-mot

En plus des deux méthodes de construction du modèle de traduction, nous avons également considéré plusieurs modèles d'alignement mot-à-mot, que nous allons décrire brièvement.

MGIZA++ [8] propose une implémentation efficace et parallèle (Gao & Vogel, 2008) des modèles génératifs les plus utilisés : les modèles IBM1 à IBM4 (Brown *et al.*, 1993) et HMM (Vogel *et al.*, 1996). Ces modèles sont utilisés par la suite pour construire des modèles de traduction selon la configuration *standard* et pour entraîner notre système d'alignement discriminant (voir section 3).

N-best WAM construit la matrice pondérée en effectuant une moyenne des occurences des liens d'alignement à partir des n-meilleures séquences d'alignement produites par le modèle IBM4. Cette méthode correspond à l'article original sur les matrices pondérées (Liu *et al.*, 2009). Comme ces auteurs, nous avons utilisé la valeur $n = 10$.

PostCAT (Posterior Constrained Alignment Toolkit) propose une implémentation des modèles HMM permettant d'injecter des contraintes lors de l'apprentissage via l'algorithme EM. Ces contraintes portent sur les probabilités *a posteriori* des variables latentes (Graça *et al.*, 2010) et permettent de corréler les deux directions d'alignement. Deux types de contraintes simples (*symmétrie* et *bijectivité*) permettent au modèle HMM d'atteindre des performances comparables au modèle IBM4. Le fait d'utiliser des modèles HMM permet de pouvoir calculer de manière exacte et efficace les probabilités *a posteriori* et ainsi construire la matrice pondérée en considérant l'ensemble des liens d'alignement. Dans cet article, nous avons utilisé la boite à outils Geppetto [9] (Ling *et al.*, 2010), une implémentation de PostCAT et des matrices d'alignement pondérées.

7. http ://www1.ccls.columbia.edu/ cadim/MADA.html
8. http ://geek.kyloo.net/
9. http ://code.google.com/p/geppetto/

MaxEntWA est le système décrit à la section 3. Il s'agit d'un classifieur *MaxEnt* qui prédit pour chaque lien de la matrice sa probabilité *a posteriori*.

Exception faite du modèle noté *MGIZA++*, il est possible pour tous les modèles d'extraire et de valuer les bi-segments selon les deux méthodes. Pour appliquer la méthode (i), nous avons appliqué pour toutes les expériences un seuil de $0,1$ comme les auteurs de (Liu *et al.*, 2009).

4.4 Résultats

Les résultats expérimentaux pour les différentes configurations et les différents modèles d'alignement sont rassemblés dans le tableau 1. Examinons pour commencer, la partie *30k* du tableau qui correspond aux expériences où MOSES a été entraîné sur un corpus de 30 000 phrases parallèles. La partie *MGIZA++* présente les résultats obtenus en utilisant l'approche standard : utilisation des meilleures hypothèses d'alignement IBM4 symmetrisés pour extraire et valuer les bi-segments via les heuristiques usuelles (Koehn *et al.*, 2003). Ainsi sur la tâche 30k, le système standard obtient un score BLEU de 35,9. La partie 10-*best WAM* correspond au matrice pondérée où les probabibilités *a posteriori* sont estimées à partir des 10 meilleurs alignements de IBM4. Cette approche permet d'obtenir un faible gain de 0,3 points BLEU par rapport à l'approche standard, soit (36,2). Ce résultat est cohérent avec ceux publiés dans (Liu *et al.*, 2009).

La partie *PostCAT* introduit par rapport aux travaux de (Liu *et al.*, 2009) l'utilisation des modèles HMM pour les alignements de mot et donc la possibilité d'estimer les probabilités *a posteriori* de manière exacte pour l'ensemble de la matrice. Cet apport permet d'augmenter le BLEU de manière significative : de 35,9 à 36,9 ou 37,0 selon la variante du modèle HMM utilisée. Enfin la partie *MaxEntWA* présente les résultats obtenus en utilisant un modèle exponentiel pour prédire la matrice d'alignement. Les résultats montrent un gain en BLEU supplémentaire et conséquent : 1,5 points par rapport à l'approche standard et 0,5 points par rapport à l'approche *PostCAT*. Notons également, que même si les méthodes standard d'extraction et de valuation sont utilisées, les matrices d'alignements engendrées par *PostCAT* et *MaxEntWA* permettent d'obtenir de meilleurs résultats et que *MaxEntWA* est à nouveau la méthode donnant le meilleur résultat.

Sur la tâche *130k* (MOSES est entrainé sur 130 000 phrases parallèles), nous observons les mêmes tendances, avec cependant des gains en BLEU moindres. Notons que le gain modeste obtenu avec la méthode 10-*best* pour estimer la matrice pondérée est similaire à celui obtenu sur la tâche *30k*. Pour les autres méthodes de calcul de la matrice pondérée, les gains restent significatifs, bien que moins importants. De nouveau, nous pouvons observer que le calcul de la matrice d'alignement avec le modèle de régression logisitique (*MaxEntWA*) permet d'obtenir de meilleurs résultats en termes de score BLEU.

La colonne *PT* du tableau 1 indique la taille du modèle de traduction en nombre de bi-segments extraits. Nous observons, tout naturellement, que quand on considère l'intégralité de la matrice pondérée (*PostCAT* et *MaxEntWA*), la taille du modèle de traduction augmente considérablement, puisqu'elle se trouve multipliée par plus de 4, alors même que le seuil de filtrage est resté constant à 0,1. Le risque était, en multipliant les entrées du modèle de traduction, d'ajouter un bruit pouvant affecter le comportement global du système. Toutefois, il apparaît que la valuation des bi-segments par les probabilités (voir la section 2.2) est un moyen effectif pour filtrer les bi-segments les moins utiles lors de l'étape de traduction.

Ainsi, l'amélioration de la valuation des bi-segments a un impact significatif sur les résultats en BLEU. Si cette amélioration peut être imputée en partie à l'utilisation des matrice pondérée, la colonne *AER (Alignment Error Rate)* montre que cette amélioration peut provenir également d'alignements mot-à-mot de meilleure qualité. Partant d'un l'*AER* obtenu avec les modèles IBM4 symétrisés d'une valeur de 25,0%, on note tout d'abord que l'usage des 10-meilleurs alignements ne permet pas d'améliorer la qualité intrinsèque des alignements. En revanche, l'utilisation d'un modèle plus approprié tel que *PostCAT* entraîne une amélioration sensible des alignements, avec un *AER* de 22,5%. Cette tendance est encore plus affirmée avec la méthode *MaxEntWA*, qui introduit dans le processus des alignements de qualité nettement accrue, puisque la réduction absolue de l'*AER* est de plus de 10 points.

Globalement, les résultats expérimentaux montrent que l'utilisation de la matrice pondérée pour extraire et valuer les bi-segments permet d'améliorer les performances des systèmes de traduction, quand cette méthode est associée à un mode de calcul pertinent pour les valuations de la matrice pondérée. Ce dernier point recouvre d'une part la manière dont sont calculées les probabilités d'alignement, et d'autre part la fraction de cette matrice qui est effectivement explorée. La différence de résultats entre les deux tâches (*30k* et *130k*) suggère que l'utilisation d'un modèle de régression logistique pour estimer la matrice pondérée conduit à des gains bien plus importants sur la petite tâche (*30k*). Une explication de cette différence est que cette approche permet, lorsque l'on dispose de peu de données parallèles, d'extraire plus de bi-segments : lorsque les données manquent pour estimer le modèle de traduction, il est en effet important de pouvoir malgré tout engendrer un grand nombre de bi-segments potentiels. De surcroît, on note que la valuation par des probabilités permet effectivement de limiter, au moment du décodage, les effets de l'introduction d'entrées bruitées dans la table de traduction.

Tâche de traduction :		**30K**					**130K**				
Construction du MT :			**Standard(i)**		**WAM(ii)**			**Standard(i)**		**WAM(ii)**	
Alignement		**AER**	**BLEU**	**PT**	**BLEU**	**PT**	**AER**	**BLEU**	**PT**	**BLEU**	**PT**
MGIZA++	HMM	28,4	35,0	3,6	-	-	26,8	39,2	9,7	-	-
	IBM4	25,0	35,9	2,4	-	-	23,3	40,2	6,5	-	-
10**-best**	IBM4	24,9	35,8	2,4	36,2	3,0	23,3	40,0	6,6	40,4	8,5
PostCAT	Bijective	22,5	36,6	3,3	36,9	10,2	20,5	40,1	9,1	40,6	29,5
	Symmetric	22,5	36,7	2,9	37,0	10,7	20,8	40,2	8,5	40,4	30,2
MaxEntWA	HMM	17,6	36,9	6,7	37,5	11,7	16,4	40,5	17,7	40,8	30,0
	IBM4	15,6	37,2	5,5	37,5	9,6	14,3	41,0	14,5	41,1	25,0
	HMM+IBM 1,3,4	14,7	37,1	5,2	37,9	8,6	13,9	40,8	13,4	41,1	22,2

TABLE 1 – Comparaison de 4 modèles d'alignement (MGIZA++, 10-best, PostCAT and MaxEntWA) et de leurs interactions avec la méthode d'extraction et de valuation de la table de traduction en termes de taux d'erreur d'alignement (*AER*), de score BLEU et de la taille de la table de traduction exprimée en millions de bi-segments (*PT*). Les deux méthodes de construction du modèle de traduction (*MT*) sont l'approche standard (*standard*) et celle utilisant les matrices pondérées (*WAM*). Deux tailles de données parallèles d'apprentissage sont considérées (*30K* et *130K*).

5 Discussion

De nombreux travaux récents se sont intéressés aux méthodes d'extraction d'unités de traduction à partir de corpus parallèles. Que ce soit dans le cadre des systèmes hiérarchiques ou à base de segments, le processus d'extraction (Koehn *et al.*, 2003; Chiang, 2007) repose sur les matrices d'alignement mot-à-mot construites à partir des modèles d'alignement IBM4 (Brown *et al.*, 1993) symétrisés. Comme nous l'avons évoqué à la section 2.1, ce choix de la première étape se justifie par un souci d'efficacité puisqu'il restreint considérablement l'espace des unités qui sont explorées, puis sélectionnées. Néanmoins, ce choix favorise la propagation d'erreurs dues à des décisions (d'accepter ou de rejeter des liens d'alignement) qui sont prises trop tôt dans le processus, sans qu'il soit de surcroit possible d'affecter de réels scores de confiance à ces décisions.

Lorsqu'il s'agit d'étendre l'espace des unités qui sont explorées, la première difficulté est la complexité qui résulte de l'énumération puis de la valuation de toutes les unités de traduction possible. Ainsi, une partie des travaux récents s'intéresse à l'élaboration d'une représentation efficace. Dans (Mi & Huang, 2008), le processus d'extraction des règles pour un système hiérarchique est étendu en considérant l'ensemble composé des n-meilleurs arbres d'analyse syntaxique au lieu de tenir compte uniquement du meilleur. Afin de représenter puis de manipuler efficacement ces n-meilleurs arbres, les auteurs utilisent une représentation efficace (*packed forest*) (Billot & Lang, 1989) ayant également démontré son utilité (Galley *et al.*, 2006; Wang *et al.*, 2007) en traduction automatique.

De manière similaire, les n-meilleurs alignements peuvent être utilisés afin d'enrichir la matrice d'alignement, que ce soit pour extraire les bi-segments (Xue *et al.*, 2006), ou les règles d'un système hiérarchique (Venugopal *et al.*, 2008). Dans ce dernier article comme dans (Mi & Huang, 2008), les auteurs définissent une distribution de probabilité sur les alignements à partir des n-meilleurs alignements et des n-meilleurs arbres d'analyse syntaxique. Cette approche par échantillonage permet aux auteurs d'introduire des comptes fractionnaires pour les règles extraites et ainsi de pouvoir estimer le modèle de traduction.

Ce recours à l'échantillonnage pour l'inférence des probabilités *a posteriori* des d'alignement se justifie par la complexité d'inférence du modèle IBM4. Il existe en revanche, pour les modèles plus simples, tels que ceux qui s'inspirent des modèles de Markov cachés (souvent désignés de manière générique sous le nom de « modèle *HMM* ») (Vogel *et al.*, 1996) ou pour le modèle IBM1 (Brown *et al.*, 1993), des algorithmes d'inférence exacts et efficaces (Venugopal *et al.*, 2003; Deng & Byrne, 2005). Une des limitations du modèle HMM est son absence de modélisation de la fertilité. Pour pallier cette limitation, les auteurs de (Deng & Byrne, 2005) définissent un HMM permettant d'aligner des mots avec des segments qui rivalise en termes de performances avec le modèleIBM4, tout en préservant la possibilité d'un calcul exact des probabilités *a posteriori* des alignements de mots et qui s'étend au calcul de distributions *a posteriori* des segments ou des règles. Les expériences montrent que cette approche améliore significativement le processus d'extraction d'unités de traductions pour les systèmes à base de segments (Deng & Byrne, 2005) et hiérarchiques (de Gispert *et al.*, 2010).

L'introduction des matrices pondérées (Liu *et al.*, 2009) que nous décrivons à la section 2 peut être considérée comme l'adaptation

des *packed forests* des systèmes hiérarchiques au systèmes à base de segments : une exploration plus exhaustive de la matrice d'alignement, l'usage des probabilités des alignements de mots pour dériver des scores de confiance sur les bi-segments extraits. Pour ce dernier point, les auteurs s'inspirent d'ailleurs des travaux de (Mi & Huang, 2008).

Comme mentionné à la section 2, un des problème des matrices pondérées est l'estimation des probabilités *a posteriori* des alignements. Dans (Liu *et al.*, 2009), cette estimation est faite en échantillonnant les n-meilleurs alignements des modèles IBM4, alors que dans (Deng & Byrne, 2005; de Gispert *et al.*, 2010; Ling *et al.*, 2010) le modèle HMM ou une de ses variante est utilisé pour les estimer de manière exacte. Cependant, dans ce dernier type d'approche, il est encore nécessaire de fusionner les alignements correspondant aux deux directions (un modèle d'alignement de source vers cible et réciproquement). Les solutions envisagées semblent peu satisfaisantes : soit la fusion est heuristique et consiste simplement à prendre la moyenne arithmétique des distributions *a posteriori* (Graça *et al.*, 2010; Ling *et al.*, 2010) ; soit de manière beaucoup plus coûteuse, deux systèmes de traduction indépendants sont utilisés utilisant chaque modèle HMM, la fusion se fait alors sur les treillis engendrés par chaque système (de Gispert *et al.*, 2010).

Dans cet article, nous introduisons donc une extension du travail de (Liu *et al.*, 2009) en proposant une nouvelle méthode de construction de la matrice d'alignement. Pour cela, nous proposons d'utiliser un classifieur au maximum d'entropie décrit dans (Ayan & Dorr, 2006; Tomeh *et al.*, 2010, 2011). Cette approche permet en effet de calculer directement la matrice pondérée sans avoir recours ni à une fusion heuristique des distributions *a posteriori*, ni à une coûteuse étape de fusion de système. Faute de données étiquettées permettant de mettre en œuvre cette démarche, l'approche de (Graça *et al.*, 2010) semble fournir des performances proches de nos meilleurs résultats.

6 Conclusion

Dans cet article, nous avons abordé le problème de l'estimation des modèles de traduction à partir d'alignements mot-à-mot non-déterministes. En effet, dans l'approche considérée comme standard, les modèles de traduction sont estimés à partir d'alignements mot-à-mot grâce à des heuristiques d'extraction et de valuation. Bien que ces alignements mot-à-mot soient construits par des modèles probabilistes, les processus d'extraction et de valuation utilisent ces modèles comme produisant des alignements déterministes. À la suite (Liu *et al.*, 2009), la solution que nous avons envisagée lève cette limitation en considérant une matrice d'alignement pondérée, dans laquelle chaque lien d'alignement est valué par sa probabilité. Les premiers travaux dans cette direction étaient, selon nos hypothèses, limités par la méthode d'estimation de la matrice pondérée, et nous avons proposé une méthode permettant d'estimer directement cette matrice à l'aide d'une méthode de classification supervisée.

Afin de valider cette approche, nous avons effectué des expériences sur la tâche de traduction automatique de l'Arabe vers l'Anglais de l'évaluation *NIST MT'09*. Dans ce cadre expérimental, nous avons comparé 4 méthodes de construction du modèle de traduction, contrastant ainsi l'approche standard avec l'usage des matrices pondérées, et évaluant différents estimateurs de cette matrice. Les résultats ont montré que l'usage des matrices pondérées impliquait une extraction plus importante de bi-segments et que leur valuation adaptée permettait au système de traduction d'obtenir de meilleurs résultats mesurés en terme de BLEU. En particulier, des gains significatifs (entre 2 et 0,9 point BLEU, selon la tâche considérée) ont été obtenus par notre méthode, qui semble la mieux à même de produire des alignements de bonne qualité (au sens de l'*AER*). Ces résultats nous ont permis de conclure que le choix de l'estimateur des matrices pondérés a un impact net sur les performances en traduction et que notre méthode est nettement plus pertinente que celles proposées dans les travaux antérieurs.

Contrairement aux heuristiques standard, notre méthode permet de contrôler et d'adapter le nombre de bi-segments extraits à la taille des données parallèles d'entraînement. Nous souhaitons donc à l'avenir explorer cet aspect. L'approche envisagée consiste à extraire le plus de bi-segments possibles et à travailler sur leur filtrage. L'intérêt de cette approche est que nous pensons ainsi limiter l'impact des erreurs commises par les modèles d'alignement. De plus, l'étape de filtrage peut se faire en prenant en compte l'utilité des bi-segments lors de l'étape de traduction et ainsi ne pas se limiter à des tests statistiques qui ne prennent pas en compte la finalité des modèles de traduction. Des articles récents comme (Wuebker *et al.*, 2010) montrent l'importance d'une valuation des bi-segments qui améliorerait les simples calculs de fréquences, et qui serait plus directement en rapport avec la finalité des modèles de traduction.

Remerciements

Ces travaux ont été en partie financé par l'agence OSEO dans le cadre du programme Quaero. Les auteurs tiennent à remercier Thomas Lavergne pour son aide précieuse concernant la mise en œuvre de *Wapiti*.

Références

ALLAUZEN A., CREGO J., MAX A. & YVON F. (2009). LIMSI's statistical translation systems for WMT'09. In *Proc. of the 4th Workshop on Statistical Machine Translation*, p. 100–104, Athens, Greece : Association for Computational Linguistics.

AYAN N. F. & DORR B. J. (2006). A maximum entropy approach to combining word alignments. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, p. 96–103 : Association for Computational Linguistics.

BILLOT S. & LANG B. (1989). The structure of shared forests in ambiguous parsing. In *Proceedings of the 27th annual meeting on Association for Computational Linguistics*, ACL '89, p. 143–151.

BROWN P. F., PIETRA V. J. D., PIETRA S. A. D. & MERCER R. L. (1993). The mathematics of statistical machine translation : parameter estimation. *Comput. Linguist.*, **19**, 263–311.

CHIANG D. (2007). Hierarchical phrase-based translation. *Comput. Linguist.*, **33**(2), 201–228.

DE GISPERT A., PINO J. & BYRNE W. (2010). Hierarchical phrase-based translation grammars extracted from alignment posterior probabilities. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, p. 545–554, Morristown, NJ, USA : Association for Computational Linguistics.

DENERO J., GILLICK D., ZHANG J. & KLEIN D. (2006). Why generative phrase models underperform surface heuristics. In *Proceedings on the Workshop on Statistical Machine Translation*, p. 31–38, New York City : Association for Computational Linguistics.

DENERO J. & KLEIN D. (2008). The complexity of phrase alignment problems. In *Proceedings of ACL-08 : HLT, Short Papers*, p. 25–28, Columbus, Ohio : Association for Computational Linguistics.

DENG Y. & BYRNE W. (2005). Hmm word and phrase alignment for statistical machine translation. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, p. 169–176, Morristown, NJ, USA : Association for Computational Linguistics.

ELMING J. & HABASH N. (2007). Combination of statistical word alignments based on multiple preprocessing schemes. In *Human Language Technologies 2007 : The Conference of the North American Chapter of the Association for Computational Linguistics ; Companion Volume, Short Papers*, NAACL-Short '07, p. 25–28, Stroudsburg, PA, USA : Association for Computational Linguistics.

GALLEY M., GRAEHL J., KNIGHT K., MARCU D., DENEEFE S., WANG W. & THAYER I. (2006). Scalable inference and training of context-rich syntactic translation models. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, p. 961–968, Sydney, Australia : Association for Computational Linguistics.

GAO Q. & VOGEL S. (2008). Parallel implementations of word alignment tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, SETQA-NLP '08, p. 49–57, Stroudsburg, PA, USA : Association for Computational Linguistics.

GRAÇA J. A. V., GANCHEV K. & TASKAR B. (2010). Learning tractable word alignment models with complex constraints. *Comput. Linguist.*, **36**, 481–504.

ITTYCHERIAH A., AL-ONAIZAN Y. & ROUKOS S. (2006). *The IBM Arabic-English Word Alignment Corpus*. Rapport interne RC24024, IBM.

KOEHN P., HOANG H., BIRCH A., CALLISON-BURCH C., FEDERICO M., BERTOLDI N., COWAN B., SHEN W., MORAN C., ZENS R., DYER C., BOJAR O., CONSTANTIN A. & HERBST E. (2007). Moses : Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, p. 177–180, Prague, Czech Republic : Association for Computational Linguistics.

KOEHN P., OCH F. J. & MARCU D. (2003). Statistical phrase-based translation. In *NAACL '03 : Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, p. 48–54 : Association for Computational Linguistics.

LAVERGNE T., CAPPÉ O. & YVON F. (2010). Practical very large scale CRFs. In *Proceedings the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, p. 504–513 : Association for Computational Linguistics.

LING W., LUÍS T., GRAÇA J., COHEUR L. & TRANCOSO I. (2010). Towards a General and Extensible Phrase-Extraction Algorithm. In M. FEDERICO, I. LANE, M. PAUL & F. YVON, Eds., *Proceedings of the seventh International Workshop on Spoken Language Translation (IWSLT)*, p. 313–320.

LIU Y., XIA T., XIAO X. & LIU Q. (2009). Weighted alignment matrices for statistical machine translation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing : Volume 2 - Volume 2*, EMNLP '09, p. 1017–1026, Morristown, NJ, USA : Association for Computational Linguistics.

MI H. & HUANG L. (2008). Forest-based translation rule extraction. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, p. 206–214, Honolulu, Hawaii : Association for Computational Linguistics.

OCH F. J. & NEY H. (2003). A systematic comparison of various statistical alignment models. *Comput. Linguist.*, **29**, 19–51.

PAPINENI K., ROUKOS S., WARD T. & ZHU W.-J. (2002). Bleu : a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, p. 311–318, Stroudsburg, PA, USA : Association for Computational Linguistics.

TOMEH N., ALLAUZEN A., WISNIEWSKI G. & YVON F. (2010). Refining word alignment with discriminative training. In *Proceedings of the ninth Conference of the Association for Machine Translation in the America (AMTA)*, Denver, CO.

TOMEH N., LAVERGNE T., ALLAUZEN A. & YVON F. (2011). Designing an improved discriminative word aligner. In *Proceedings of the 12th International Conference on Intelligent Text Processing and Computational Linguistics (CICLING)*, Tokyo, Japan.

VENUGOPAL A., VOGEL S. & WAIBEL A. (2003). Effective phrase translation extraction from alignment models. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, p. 319–326, Stroudsburg, PA, USA : Association for Computational Linguistics.

VENUGOPAL A., ZOLLMANN A., SMITH N. A. & VOGEL S. (2008). Wider pipelines : N-best alignments and parses in MT training. In *Proceedings of the Association for Machine Translation in the Americas (AMTA)*.

VOGEL S., NEY H. & TILLMANN C. (1996). Hmm-based word alignment in statistical translation. In *Proceedings of the 16th conference on Computational linguistics*, p. 836–841 : Association for Computational Linguistics.

WANG W., KNIGHT K. & MARCU D. (2007). Binarizing syntax trees to improve syntax-based machine translation accuracy. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, p. 746–754, Prague, Czech Republic : Association for Computational Linguistics.

WUEBKER J., MAUSER A. & NEY H. (2010). Training phrase translation models with leaving-one-out. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, p. 475–484, Uppsala, Sweden : Association for Computational Linguistics.

XUE Y.-Z., LI S., ZHAO T., YANG M. & LI J. (2006). Bilingual phrase extraction from n-best alignments. In *ICICIC (3)*, p. 410–414.

ZENS R., OCH F. J. & NEY H. (2002). Phrase-based statistical machine translation. In *KI '02 : Proceedings of the 25th Annual German Conference on AI*, p. 18–32, London, UK : Springer-Verlag.

ZOU H. & HASTIE T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, **67**, 301–320.