Modèles discriminants pour l'alignement mot à mot

Alexandre Allauzen — Guillaume Wisniewski

Univ Paris-Sud 11, Orsay, F-91405
LIMSI-CNRS, B.P. 133, 91403 Orsay cedex, France

{allauzen , wisniews }@limsi.fr

RÉSUMÉ. *Un alignement mot à mot entre une phrase et sa traduction consiste à extraire des relations d'appariement entre les mots de la phrase source et les mots de sa traduction. Aujourd'hui le principal système d'alignement état de l'art, Giza++, repose sur une combinaison des modèles génératifs IBM. Bien que Giza++ soit utilisé par la plupart des systèmes statistiques de traduction automatique, la qualité des alignements qu'il prédit n'est pas satisfaisante. Nous proposons d'aborder ce problème avec des modèles discriminants (maximum d'entropie et champs conditionnels aléatoires) afin d'intégrer des caractéristiques plus riches et robustes. Les différents modèles sont évalués en termes de taux d'erreur d'alignement (AER) sur deux paires de langues (français/anglais et arabe/anglais). Nos résultats montrent le gain et le potentiel des modèles discriminants pour la tâche d'alignement.*

ABSTRACT. *Word alignment aims to link each word of a translated sentence to its related words in the source sentence. Nowadays, Giza++ is the most used word alignment system. This toolkit implements the generative IBM models. Despite its popularity, several limitations remain. We thus propose to address this task using discriminative models (Maximum Entropy and Conditional Random Fields) which can easily make use of additional features. These models are evaluated in terms of Alignment Error Rate (AER) using two language pairs (French/english and Arabic English). Our results show that discriminant models are well suited for this task and that they can outperform IBM models.*

MOTS-CLÉS : *modèles d'alignement mot à mot, maximum d'entropie, champs conditionnels aléatoires.*

KEYWORDS: *Word alignment models, Maximum entropy, Conditionnal Random Fields.*

1. Introduction

Un alignement mot à mot entre une phrase et sa traduction consiste à extraire des relations d'appariement entre les mots de la phrase source et les mots de sa traduction. Prédire automatiquement ces *relations de traduction* est une tâche qui a de nombreuses applications, par exemple, en recherche d'information multilingue (Rogati et Yang, 2003), en aide à la traduction (Gaussier *et al.*, 2000 ; Deléger *et al.*, 2009) ou en extraction de lexiques bilingues (Fung, 2000). Dans cet article, nous nous intéressons plus particulièrement à l'une des applications de l'alignement mot à mot : les systèmes de traduction statistique.

Les outils de traduction statistique (par exemple (Lopez, 2008) pour un état de l'art complet) s'appuient sur des modèles probabilistes pour traduire automatiquement des textes. Les paramètres de ces modèles sont appris automatiquement à partir de corpus de textes parallèles regroupant un grand ensemble de phrases et leur traduction. Deux principaux types de paramètres sont mis en jeu : les paramètres du *modèle de langage* décrivant à quel point la traduction est une phrase « correcte » et les paramètres du *modèle de traduction* évaluant si le contenu de la phrase source est conservé lors de la traduction. Le modèle de traduction à base de segments repose sur une *table de traduction* qui décrit la probabilité de traduire un ou plusieurs mots de la langue source en un groupe de mots de la langue cible. Cette table de traduction est construite directement à partir des alignements mot à mot appris sur l'ensemble du corpus parallèle (Och *et al.*, 1999).

Aujourd'hui, le système d'alignement le plus utilisé est `Giza++` (Och et Ney, 2003). Deux raisons peuvent expliquer l'importance prise par `Giza++` dans la communauté. D'une part, il repose sur une combinaison de modèles génératifs, les modèles IBM (Brown *et al.*, 1993), qui sont des modèles appris de manière non supervisée à partir des grands corpus alignés phrase à phrase. D'autre part, l'implémentation de `Giza++` est librement disponible, ce qui en fait une boîte noire facile d'utilisation et robuste. Cependant, les performances de `Giza++`, aussi bien du point de vue de la qualité des alignements prédits que des temps de calcul, ne sont pas satisfaisantes et l'étape d'alignement mot à mot constitue, aujourd'hui, un facteur limitant dans de nombreux systèmes de traduction automatique (Fraser et Marcu, 2007).

Dans cet article, nous proposons d'utiliser des modèles discriminants pour améliorer le traitement de ces deux problèmes. Contrairement aux modèles génératifs, les modèles discriminants permettent de prendre en compte facilement des caractéristiques arbitraires et offrent de meilleures garanties sur les performances en généralisation (Collins, 2004). La motivation de ce travail est double :

– en prenant en compte des caractéristiques plus riches que celles considérées par les modèles génératifs, nous espérons améliorer la qualité des alignements prédits ;

– en incluant les différentes caractéristiques directement, et non pas en combinant différents modèles génératifs à l'aide d'approximations, comme le fait `Giza++`, nous espérons que les modèles proposés seront plus rapides en apprentissage et en inférence, mais aussi plus simples à utiliser.

Toutefois, l'utilisation de modèles discriminants pour l'alignement se heurte à une difficulté majeure : leur apprentissage nécessite des corpus alignés mot à mot alors que la quasi-totalité des corpus disponibles aujourd'hui sont alignés phrase à phrase et que les rares corpus alignés mot à mot ne comportent généralement que peu d'exemples : l'un des corpus que nous utiliserons dans nos expériences, le Hansard, est composé de plus d'un million de phrases alignées mais de seulement quelques centaines de phrases alignées mot à mot. Il faudra donc que les modèles développés puissent être appris à partir de peu d'exemples.

Dans ce travail, nous présentons deux modèles discriminants d'alignement mot à mot. Le premier modèle formalise la tâche d'alignement comme une tâche de classification multiclasse et traite celle-ci avec un classifieur à maximum d'entropie (Berger *et al.*, 1996 ; Carpenter, 2008). Ce modèle permet d'introduire aisément des caractéristiques arbitraires tout en présentant une complexité faible aussi bien en apprentissage qu'en inférence. Il prédit les alignements indépendamment les uns des autres, bien qu'il soit intuitivement plus pertinent de le faire conjointement afin de pouvoir choisir l'alignement d'un mot en tenant compte des alignements de ses voisins. C'est pourquoi nous avons considéré ensuite un modèle fondé sur les champs conditionnels aléatoires (CRF) (Lafferty *et al.*, 2001 ; Blunsom et Cohn, 2006 ; Sutton et McCallum, 2006) afin d'introduire des dépendances entre alignements au prix d'une complexité accrue.

La contribution de ce travail est triple : *i)* nous effectuons une synthèse de la problématique de l'alignement mot à mot ; *ii)* nous présentons un nouveau modèle d'alignement discriminant reposant sur un classifieur *MaxEnt* et généralisons un modèle de l'état de l'art (le modèle CRF de (Blunsom et Cohn, 2006)) ; *iii)* nous évaluons les performances de ces deux modèles sur deux paires de langues (anglais/français et arabe/anglais). Les résultats expérimentaux que nous obtenons montrent d'une part l'importance de modéliser explicitement les dépendances entre les prédictions d'alignement, et d'autre part qu'il est possible d'obtenir de meilleures performances que celles des modèles génératifs en apprenant à partir de plus petits corpus.

Cet article est organisé comme suit. Nous introduisons en section 2 la tâche de prédiction d'alignement et détaillons les difficultés qu'elle présente. Nous introduisons également, dans cette section, les corpus et les critères d'évaluation que nous utiliserons dans nos expériences. Nous présentons ensuite, dans la section 3, les modèles génératifs qui sont les plus couramment utilisés, avant de détailler les deux modèles que nous proposons dans les sections 4 et 5. Finalement, nous analysons les autres modèles discriminants proposés dans la littérature, dans la section 6.

2. L'alignement mot à mot : définition, corpus et évaluation

Dans cette section, nous allons formaliser le concept d'alignement mot à mot (sous-section 2.1) et présenter ses principales caractéristiques (sous-section 2.2). Nous décrirons ensuite les deux principaux types de corpus existants (sous-section 2.3) ainsi

que les méthodes d'évaluation (sous-section 2.4) et la manière dont nous évaluerons les différents modèles.

2.1. *Définition*

Un alignement mot à mot entre une phrase et sa traduction associe à chaque mot de cette phrase un mot de la traduction. Un alignement au niveau d'une paire de phrases regroupe donc un ensemble de liens décrivant une relation de traduction entre mots. Il est possible qu'un mot n'ait pas de traduction directe, il est alors aligné sur un symbole spécial noté `null`.

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 $\mathbf{e} = e_1, ..., e_i, ..., e_I$ est une séquence de I mots et $\mathbf{f} = f_1, ..., f_j, ..., f_J$ est une séquence de J mots. Un alignement mot à mot d'un couple de phrases est représenté par une *matrice d'alignement*. L'élément (i, j) de la matrice est 1 si le i-ème mot de $\mathbf{e}$ est aligné avec le j-ème mot de $\mathbf{f}$ et 0 sinon. La figure 1 donne un exemple de matrice d'alignement (et des liens qui lui sont associés).

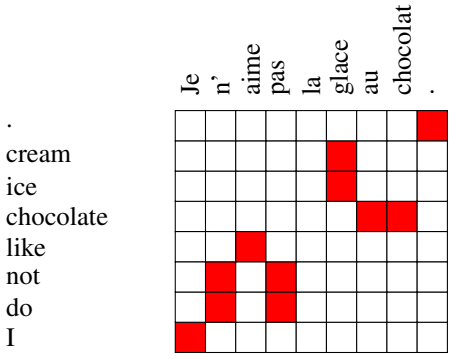

Figure 1. *Exemple de matrice d'alignement entre une phrase anglaise et une phrase française. Les termes non nuls de la matrice sont représentés par des carrés pleins. L'ensemble des liens associés à cette matrice est donc* $\{(1,1), (2, 2), (2, 3), (3, 4), (4,2), (4,3), (6, 6), (6, 7), (7, 5), (8, 5), (9, 8)\}$

Nous pouvons d'ores et déjà remarquer qu'une matrice d'alignement comporte majoritairement des termes nuls : en première approximation, chaque mot de la phrase $\mathbf{e}$ est aligné avec un mot de la phrase $\mathbf{f}$; la matrice d'alignement comporte donc environ $\min(I,J)$ éléments non nuls parmi les $I \times J$ valeurs (où I et J sont les tailles des deux phrases à aligner).

2.2. *Caractéristiques des alignements mot à mot*

Les alignements possèdent deux caractéristiques fondamentales. Premièrement, comme le montre l'exemple de la figure 1, les alignements ne sont pas nécessairement contigus : deux mots consécutifs dans la phrase $\mathbf{e}$ peuvent être alignés avec deux mots arbitrairement distants de la phrase $\mathbf{f}$. On appelle communément *distorsion* ce réagencement syntagmatique qui a lieu durant une traduction. Suivant la paire de langues considérée, la distorsion peut être plus ou moins importante : le français et l'anglais sont, par exemple, deux langues « proches » pour lesquelles l'ordre des mots est généralement conservé (on parle alors d'alignements monotones), alors que dans une traduction de l'arabe vers l'anglais, l'ordre des mots peut être modifié de manière importante.

Deuxièmement, les alignements peuvent décrire aussi bien des correspondances entre mots qu'entre *blocs* de mots. En effet, il n'y a généralement pas de correspondance une à une entre les mots des phrases $\mathbf{e}$ et $\mathbf{f}$. C'est notamment le cas lorsque l'on traduit des expressions idiomatiques (par exemple, l'expression « manger les pissenlits par la racine » se traduit, en anglais, par *« push up daisies*[1] *»*.) ou, lorsque les deux langues présentent des différences grammaticales (absence d'article partitif, utilisation d'un auxiliaire pour marquer certains temps...). Ainsi, dans l'exemple de la figure 1, le mot français « glace » est traduit, en anglais, par les deux mots *« ice cream »* et le mot anglais *« chocolate »* par le groupe prépositionnel « au chocolat ».

Ces alignements entre groupes de mots compliquent la définition d'un alignement mot à mot de référence. Ainsi, dans l'exemple de la figure 1, plusieurs alignements de l'auxiliaire « *do* » sont effectivement justifiables :

– comme sa présence est requise par la forme négative de la phrase, il peut être aligné avec l'un des marqueurs de négation (« ne » ou « pas ») ;

– il peut également être aligné avec « aime », puisqu'il apporte une information sur le temps, le mode et la personne du verbe « *like* » ;

– un autre choix est de n'aligner « *do* » avec aucun mot de la phrase $\mathbf{f}$ en considérant que la présence de cet auxiliaire est une particularité de la langue anglaise qui n'a pas d'équivalent direct en français.

Ces deux caractéristiques doivent être prises en compte lors de la construction d'un système d'alignement et donc en amont, lors de l'annotation d'un corpus.

2.3. *Constitution de corpus*

Face à la difficulté de constituer des alignements de référence, deux solutions ont été imaginées. Certaines campagnes d'annotation proposent de distinguer des alignements *sûrs* pour lesquels les différents annotateurs sont d'accord, des alignements

1. Littéralement « pousser des marguerites »

probables qui indiquent les alignements ambigus et notamment les expressions figées et les traductions libres. Les alignements probables sont généralement utilisés pour aligner des *blocs* de mots sans préciser les alignements mot à mot à l'intérieur de ceux-ci. La figure 2 montre un exemple d'un alignement issu du corpus Hansard (Och et Ney, 2003) qui distingue alignements sûrs et alignements probables.

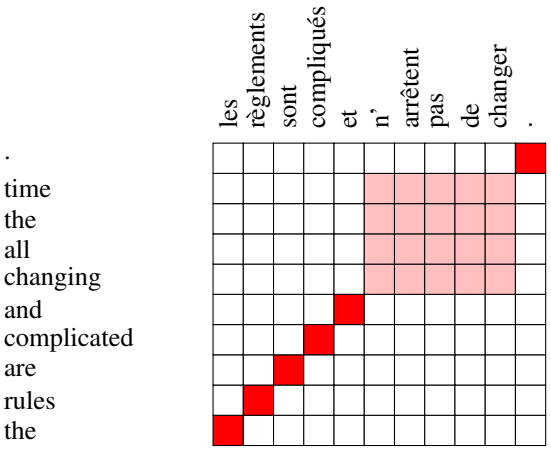

Figure 2. *Exemple d'un alignement comportant des alignements sûrs (carrés rouge) et des alignements probables (carrés rose)*

Une autre solution consiste à fixer des spécifications décrivant précisément quel alignement choisir pour chaque situation ambiguë. C'est, par exemple, la solution qui a été choisie pour constituer le corpus d'alignement arabe/anglais que nous utiliserons dans nos expériences : un guide d'annotation (Ittycheriah *et al.*, 2006) regroupe plusieurs dizaines de règles indiquant comment aligner les mots problématiques.

La première méthode a l'avantage d'être indépendante de la paire de langues considérée, mais, comme nous le verrons à la section 2.4, elle nécessite de définir des mesures d'évaluation *ad hoc*. De plus, elle réduit la quantité de données disponibles pour l'apprentissage : il est difficile d'apprendre à partir de données imprécises et les alignements probables sont généralement ignorés pendant l'apprentissage. La seconde méthode complique le travail d'annotation puisqu'il faut redéfinir les règles pour chaque nouvelle paire de langues et s'assurer que les annotateurs comprennent et respectent celles-ci. Mais elle permet d'évaluer un système d'alignement de manière fiable.

Si l'annotation de corpus alignés mot à mot est coûteuse, la constitution de corpus parallèles, dont les documents sont alignés phrase à phrase, est plus aisée. Il y a donc,

aujourd'hui, beaucoup plus de corpus alignés au niveau des phrases que de corpus alignés au niveau des mots.

2.4. *Méthodes d'évaluation*

Nous présentons dans cette partie les corpus et les métriques d'évaluation que nous utilisons lors de nos expériences.

2.4.1. *Les corpus*

Afin de montrer le degré de dépendance des méthodes proposées par rapport aux paires de langues considérées, nous réaliserons nos expériences sur un jeu de données français/anglais et sur un jeu arabe/anglais. Pour chaque paire de langues, nous distinguons quatre corpus :

– un corpus parallèle, aligné phrase à phrase, et considéré comme non étiqueté. Ce corpus est utilisé pour l'entraînement des modèles génératifs grâce à l'outil `Giza++` ;

– un corpus d'entraînement aligné mot à mot dédié à l'apprentissage des modèles discriminants ;

– un corpus de développement aligné mot à mot permettant de choisir les méta-paramètres de nos modèles et notamment le paramètre de régularisation des modèles *MaxEnt* et CRF ;

– un corpus de test permettant d'estimer l'erreur en généralisation des différents modèles.

Détaillons ces quatre corpus pour les deux paires de langues.

2.4.1.1. Données français/anglais

Pour la paire de langues français/anglais, nous avons utilisé le corpus Hansard (Och et Ney, 2003 ; Mihalcea et Pedersen, 2003) constitué de transcriptions des débats du parlement Canadien. Ce corpus contient 884 couples de phrases alignées mot à mot. Elles sont réparties en un corpus d'apprentissage de 424 phrases, un corpus de test de 423 phrases et un corpus de développement de 37 phrases. Les textes ont subi des prétraitements (filtrage et normalisation) et l'annotation humaine distingue les alignements sûrs et probables.

Un corpus nettement plus important est également utilisé pour entraîner les modèles génératifs d'alignement et pour collecter des statistiques bilingues. Il s'agit de 1 130 104 couples de phrases extraits du Hansard, mais alignés au niveau de la phrase[2]. Les prétraitements sont identiques à ceux appliqués aux données alignées mot à mot.

2. Ces données ont été collectées par Ulrich Germann et sont librement accessibles sur le site `http://www.isi.edu/natural-language/download/hansard/index.html`

2.4.1.2. Données arabe/anglais

Pour la paire de langues arabe/anglais, nous avons utilisé les données décrites dans (Ittycheriah *et al.*, 2006) et distribuées par le *Linguistic Data Consortium*[3]. Suivant (Elming *et al.*, 2009) nous avons utilisé la partie de ce corpus issue de l'*Arabic Treebank* pour l'entraînement et le développement et le reste des phrases pour l'évaluation. Ces données ont été filtrées afin d'écarter les couples de phrases contenant, par exemple, des URL ou des problèmes d'encodage et nous n'avons gardé que les phrases de moins de 25 mots afin d'accélérer l'apprentissage et l'inférence des modèles[4]. Le corpus final comporte 119 871 exemples alignés au niveau des phrases.

Les données obtenues se répartissent en 800 phrases pour l'apprentissage et 300 phrases pour le test. Pour entraîner les modèles génératifs et évaluer certaines caractéristiques, nous avons considéré des corpus parallèles alignés au niveau des phrases.[5] Toutes les données utilisées sont prétraitées de manière à conserver la segmentation en mots. Cette contrainte est indispensable pour permettre l'exploitation des alignements manuels. Ainsi les données arabes ont été translittérées selon le schéma de Buckwalter puis normalisées. Les données anglaises ont été simplement converties en minuscules puis normalisées.

2.4.2. *Les mesures d'évaluation*

L'alignement mot à mot est une tâche intermédiaire dont le seul objectif est d'extraire des ressources pour une tâche « de plus haut niveau » (système de traduction automatique, de recherche d'information...). L'évaluation la plus pertinente devrait donc se faire par rapport à la tâche de plus haut niveau. Toutefois, lors du développement de systèmes d'alignement, il est important de disposer d'une mesure permettant d'évaluer directement la qualité d'un alignement proposé par rapport à un alignement de référence, plutôt que de répéter intégralement la tâche de haut niveau (ce qui peut être particulièrement long).

La prédiction d'un alignement consiste à prédire pour chaque paire de mots si elle est, ou non, alignée. C'est, par conséquent, une tâche de classification binaire dans laquelle les deux classes sont fortement déséquilibrées. Il est donc naturel d'estimer la qualité d'un alignement par le rappel et la précision qui mesurent la capacité d'un classifieur binaire à prédire correctement la classe positive lorsque les classes sont déséquilibrées (Makhoul *et al.*, 1999).

Suivant les travaux de (Och et Ney, 2003 ; Fraser et Marcu, 2007), définissons un alignement hypothétique A (c'est-à-dire un ensemble de liens entre les mots de la phrase $\mathbf{e}$ et les mots de la phrase $\mathbf{f}$) et un alignement de référence constitué de liens

3. sous la référence LDC2006G09

4. Les expériences en inférence sans ce filtrage sur la longueur des phrases montrent que cela ne change pas les résultats.

5. Ces données sont également distribuées par le LDC sous les références LDC2004T18, LDC2005E46, LDC2004T17 et LDC2004E72.

d'alignement sûrs S et probables P. Le rappel R et la précision P_r sont respectivement définis par :

$$R = \frac{|A \cap S|}{|S|} \qquad P_r = \frac{|A \cap P|}{|A|}, \qquad [1]$$

où $|X|$ désigne le cardinal de l'ensemble X[6]. Ces deux mesures sont usuellement combinées pour donner la F-mesure :

$$F = \frac{2 \cdot P_r \cdot R}{P_r + R} \qquad [2]$$

Plusieurs travaux (Fraser et Marcu, 2007) ont montré qu'il y avait une forte corrélation entre la qualité d'un alignement mot à mot évaluée par la F-mesure et la qualité d'une traduction évaluée par le score BLEU[7]. Il est donc possible d'optimiser le score de la traduction en optimisant le score d'alignement.

Cependant, il existe une autre mesure *ad hoc* qui est utilisée dans la grande majorité des travaux portant sur les modèles d'alignement. Cette mesure, proposée par (Och et Ney, 2003) est l'*Alignment Error Rate* (AER). Elle est définie par :

$$\text{AER}(A, S, P) = 1 - \frac{|A \cap P| + |A \cap S|}{|A| + |S|} \qquad [3]$$

L'AER est un taux d'erreur qu'il faut chercher à minimiser. Il est minimal lorsque l'on a prédit tous les liens sûrs ; si l'on ne prédit pas l'ensemble des liens sûrs, le score sera d'autant plus faible que l'on a prédit correctement des liens sûrs ou probables.

La principale différence entre ces deux mesures est que la F-mesure pénalise un déséquilibre éventuel entre le rappel et la précision (Fraser et Marcu, 2007). Remarquons également que s'il n'y a pas de distinction dans les références entre alignements sûrs et probables, comme c'est le cas pour les données arabe/anglais, alors ces deux mesures sont en relation déterministe : $AER = 1 - F$. Ainsi, les résultats sur la paire de langues français/anglais utiliseront les deux mesures (la F-mesure et l'AER), alors que pour la paire arabe/anglais seul l'AER sera utilisé.

3. Modèles génératifs pour l'alignement mot à mot

Nous allons présenter rapidement dans cette section la famille de modèles génératifs « historiques » et leurs principales limites qui ont motivé notre travail. Ces modèles reposent sur une même modélisation du processus d'alignement que nous introduisons dans la section 3.1. Nous détaillerons ensuite les deux principaux modèles génératifs

6. Pour assurer la cohérence des mesures, l'ensemble des liens probables est défini comme l'union de l'ensemble des liens étiquetés sûrs et des liens étiquetés probables.

7. Le score BLEU (Papineni *et al.*, 2002) est la mesure traditionnellement utilisée pour évaluer la qualité d'une traduction automatique. C'est une mesure de similarité entre une hypothèse de traduction et une traduction de référence.

utilisés : les modèles IBM 1 (section 3.2) et les modèles IBM d'ordre supérieur (section 3.3). Finalement, nous évaluerons dans la section 3.4 les performances de ces deux modèles sur les corpus que nous avons introduits dans la section précédente. Ces résultats constitueront les points de comparaison pour évaluer les performances de nos modèles discriminants.

3.1. *Principe*

Quatre modèles génératifs, appelés IBM 1, 2, 3 et 4 en fonction de la richesse des caractéristiques qu'ils prennent en compte, ont été introduits par (Brown *et al.*, 1993) dans le cadre des systèmes statistiques de traduction à base de mots. Ils modélisent la traduction complète d'une phrase source en une phrase cible. Dans ces modèles, l'alignement mot à mot est introduit sous la forme de variables latentes et peut être vu comme un « produit dérivé » de la traduction.

Depuis l'introduction des systèmes de traduction à base de segments (Zens *et al.*, 2002 ; Koehn *et al.*, 2003), ces modèles ne constituent plus l'état de l'art en traduction. La prédiction d'alignements mot à mot reste cependant un domaine de recherche à part entière particulièrement actif notamment parce que les alignements prédits par ces systèmes continuent à jouer un rôle essentiel dans la construction des *tables de traduction* (Lopez, 2008) au cœur aussi bien des systèmes de traduction à base de segments (*phrase based*) comme `Moses` (Koehn *et al.*, 2007) que des modèles hiérarchiques de traduction comme `Hiero` (Chiang, 2007).

Comme introduit dans la section 2.1, un alignement mot à mot se représente par une matrice binaire. Plutôt que de modéliser directement cette matrice, les modèles IBM formalisent la tâche d'alignement comme une tâche d'étiquetage de séquences. Pour cela, ils distinguent une phrase source et une phrase cible[8] et prédisent pour chacun des mots de la phrase source le mot de la phrase cible avec lequel il est aligné, ce qui revient à étiqueter chaque mot de la phrase source avec un des mots de la phrase cible.

Plus formellement, si nous choisissons pour phrase source $\mathbf{e} = e_1, ..., e_I$ et pour phrase cible $\mathbf{f} = f_1, ..., f_J$, la prédiction d'un alignement consiste à déterminer la séquence d'étiquettes $\boldsymbol{a} = a_1, ..., a_I$ associée à $\mathbf{e}$ où chaque étiquette a_i est une variable aléatoire ayant comme espace de réalisation les indices des mots dans la phrase cible $0, ..., J$, l'indice 0 représentant le symbole `null` et signifiant l'absence d'alignement pour ce mot. L'ensemble des étiquettes a_i permet de représenter la matrice d'alignement sous la forme d'une fonction a qui associe une position de la phrase cible à une position de la phrase source. Cette fonction a pour domaine $[\![1, I]\!]$ et $[\![0, J]\!]$ pour codomaine.

8. La distinction entre phrase source et phrase cible est liée à l'utilisation des modèles IBM à la traduction et n'est pas pertinente dans le cadre de l'alignement.

Cette formalisation permet de simplifier le paramétrage des modèles IBM et la complexité de l'estimation de leurs paramètres. Mais elle rend le modèle d'alignement asymétrique : par construction, les mots de la phrase source sont alignés avec, au plus, un mot de la phrase cible (alignement dit `1:n`). Ce type d'alignement ne décrit que partiellement les alignements réels. En pratique, il est nécessaire de *symétriser* les sorties des modèles IBM pour reconstituer la matrice d'alignement. Il faut pour cela apprendre un modèle d'alignement dans les deux directions (les deux phrases jouent successivement le rôle de source) puis utiliser des heuristiques pour combiner les deux alignements prédits et retrouver la matrice d'alignement.

Dans (Och et Ney, 2003 ; Koehn *et al.*, 2003), les auteurs proposent plusieurs heuristiques de symétrisation qui ont pour point commun de partir de l'intersection des deux sens d'alignement puis de compléter cet ensemble avec des points de l'union des alignements. Dans cette article, nous utilisons, pour `Giza++` et les modèles discriminants, l'heuristique *grow-diag-final-and* (Koehn *et al.*, 2003) qui est la plus employée dans la communauté.

3.2. *Le modèle IBM 1*

3.2.1. *Description*

Le modèle IBM 1 est un modèle génératif de traduction mot à mot. Ce modèle met en œuvre deux intuitions qui fondent la découverte de traductions à partir d'un corpus bilingue : *i)* un mot donné n'a qu'un petit nombre de traductions ; *ii)* un mot et sa traduction sont fréquemment en cooccurrence. Par exemple, sans même connaître le grec, il est relativement intuitif d'extraire, du corpus décrit dans le tableau 1 les traductions de mots suivantes : maison → σπίτι, une → ένα, la → το et vague → κύμα.

Français	Grec
une maison	ένα σπίτι
la maison	το σπίτι
une vague	ένα κύμα

Tableau 1. *Exemple de corpus bilingue*

L'apprentissage et l'inférence du modèle IBM 1 reposent sur deux intuitions : lors de l'apprentissage, la probabilité que deux mots soient traduction l'un de l'autre est renforcée itérativement lorsqu'ils cooccurrent (intuition *i)*) ; lors de l'inférence, chaque mot source est aligné avec le mot cible dont la probabilité de traduction est la plus élevée (intuition *ii)*).

Plus formellement, à partir d'un *lexique* donnant la probabilité $t(f|e)$ de traduire un mot de la langue source e en un mot de la langue cible f, le modèle IBM 1 pa-

ramètre la probabilité $p(\mathbf{f}, a|\mathbf{e})$ de traduire une phrase $\mathbf{e}$ en une phrase $\mathbf{f}$ selon un alignement a par :

$$p(\mathbf{f}, a|\mathbf{e}) \propto \prod_{i=1}^{I} t(f_{a_i}|e_i) \quad [4]$$

Ce paramétrage revient à traduire chaque mot de la phrase source indépendamment des autres mots puis à choisir sa position dans la phrase cible sans tenir compte ni de la traduction ni de la position des autres mots.

Le modèle IBM 1 est paramétré par les probabilités de traduction mot à mot $t(f|e)$ qui décrivent la fréquence avec laquelle un mot de la langue source e est traduit en un mot de la langue cible f. Grâce à l'algorithme EM (Dempster *et al.*, 1977), ces paramètres peuvent être estimés facilement à partir de corpus alignés phrase à phrase. La fonction objective de ce problème d'apprentissage étant concave (Brown *et al.*, 1993) l'algorithme EM converge vers un optimum global, ce qui simplifie la mise en œuvre de l'apprentissage.

Le modèle IBM 1, tel que nous venons de le présenter, permet de traduire une phrase $\mathbf{e}$ en une phrase $\mathbf{f}$[9]. Il permet également de déterminer l'alignement le plus probable d'une paire de phrases données :

$$\hat{a} = \underset{a \in \mathcal{A}}{\operatorname{argmax}}\, p(a|\mathbf{f}, \mathbf{e}) \quad [5]$$

$$= \underset{a \in \mathcal{A}}{\operatorname{argmax}}\, p(\mathbf{f}, a|\mathbf{e}) \quad [6]$$

où $\mathcal{A}$ est l'ensemble des alignements possibles. L'équation 5 correspond à l'estimateur du *maximum a posteriori* habituellement utilisé en classification ; l'équation 6 permet de reformuler cet estimateur en fonction du modèle IBM 1. La recherche du meilleur alignement est particulièrement efficace puisque l'évaluation de chaque alignement a_i est indépendante des autres.

3.2.2. *Limites*

Le modèle IBM 1 est un modèle d'alignement très simple qui utilise comme unique information les propriétés lexicales des mots (sous la forme des cooccurrences d'un mot et de sa traduction). La « pauvreté » de ces informations complique l'estimation : comme le modèle IBM 1 ne considère que les formes fléchies des mots, le risque de trouver lors de la prédiction un mot hors vocabulaire est grand. La méthode d'apprentissage introduit également un biais dans l'estimation des statistiques : pour un mot f donné, la probabilité de traduction mot à mot $t(f|e)$ est d'autant plus petite que le mot e apparaît fréquemment dans le corpus d'apprentissage. Lors de la prédiction, un mot peu fréquent dans le corpus d'apprentissage aura donc tendance à être aligné avec

9. Depuis le développement des modèles de traduction à base de segments, les modèles IBM ne sont plus utilisés que comme modèles d'alignement et non plus comme modèles de traduction. Ils sont utilisés dans les systèmes à base de segments lors de l'extraction des segments et parfois lors du décodage.

davantage de mots e candidats qu'un mot fréquent. Cet effet a été décrit comme l'effet « *garbage collector* » des mots rares (Moore, 2004).

Pour éviter ces deux problèmes, il est nécessaire d'apprendre le modèle IBM 1 à partir d'une grande quantité de données (plusieurs millions de paires de phrase en général). Plusieurs travaux (Ganchev *et al.*, 2008 ; Taskar *et al.*, 2005) ont toutefois montré qu'il était possible d'améliorer les résultats de modèles appris à partir de moins de données en considérant d'autres caractéristiques.

Intuitivement, plusieurs caractéristiques peuvent faciliter le choix de l'alignement d'un mot :

– la *distorsion* : dans le modèle IBM 1, la position d'un mot dans la phrase cible ne dépend ni de la position du mot avec lequel il est aligné ni des alignements des autres mots de la phrase : ce modèle est donc incapable de modéliser le fait que des mots adjacents dans la phrase source ont tendance à être traduits par des mots adjacents dans la phrase cible ;

– la *fertilité* : en outre, lors de la prédiction de l'alignement, ce modèle ne prend pas en compte le nombre de mots de la phrase cible avec lesquels un mot de la phrase source est aligné.

La prise en compte de ces caractéristiques nécessite une nouvelle modélisation, plus complexe, de l'alignement et a motivé le développement des modèles IBM d'ordre supérieur.

3.3. *Les modèles IBM d'ordre supérieur*

Dans (Brown *et al.*, 1993), les auteurs proposent une succession de modèles de complexité croissante permettant d'intégrer notamment les notions de distorsion et de fertilité. L'histoire générative de chaque modèle devient alors plus complexe.

Une des critiques forte à l'encontre du modèle IBM 1 est l'hypothèse selon laquelle toutes les valeurs d'un alignement a_i sont équiprobables. Le modèle IBM 2 (Brown *et al.*, 1993) introduit une dépendance entre la valeur de l'alignement a_i et la position i du mot dans la phrase source. Cependant, ce modèle n'est guère utilisé car il néglige, entre autres, la monotonicité des alignements. Le modèle de Markov caché (nommé usuellement HMM) (Vogel *et al.*, 1996) introduit une notion de « proximité » dans le modèle IBM 1 en modélisant explicitement la distance entre l'alignement du mot courant et l'alignement du mot précédent. Cependant, les histoires génératives des modèles IBM 1 et HMM impliquent qu'un mot ne peut en générer qu'un seul. Le simple exemple de *potatoes* qui se traduit avec les trois mots *pommes de terre* en français montre les limites de tels scénarios. Les modèles IBM 3 et IBM 4 généralisent les modèles précédents en introduisant successivement la notion de distorsion et de fertilité afin de modéliser ce phénomène.

Cet empilement de modèles et d'hypothèses montre les difficultés inhérentes aux modèles génératifs. L'introduction de nouvelles caractéristiques passe par la réécriture

de l'histoire générative et une complexité fortement accrue. Un exemple emblématique en est le modèle IBM 4. Sa complexité est telle que la recherche de la séquence d'alignements la plus probable devient NP-difficile (Brown *et al.*, 1993). Pour les modèles à base de fertilité, l'inférence n'est donc plus exacte et nécessite des hypothèses simplificatrices. De même, l'apprentissage est rendu particulièrement complexe par l'existence de minima locaux. En pratique, l'entraînement du modèle IBM 4 nécessite d'entraîner au préalable les modèles IBM 1, HMM et IBM 3 en guise d'initialisation.

3.4. *Évaluation des modèles génératifs*

Nous allons comparer, dans cette sous-section, les résultats expérimentaux de ces trois principaux modèles génératifs. Tous les résultats expérimentaux présentés dans cette sous-section ont été obtenus avec `Giza++` (Och et Ney, 2003). Conformément à ce qui se fait généralement dans la communauté, les paramètres par défaut de `Giza++` ont été utilisés.

Les résultats de l'alignement des données français/anglais sont regroupés dans le tableau 2 pour les deux directions d'alignement et les alignements ont été symétrisés par l'heuristique *grow-diag-final-and*. Le tableau 2 indique également les résultats obtenus avec cette heuristique. Nous observons une amélioration importante de l'AER pour le passage du modèle IBM 1 à HMM due à l'introduction des dépendances markoviennes. Ce gain s'explique par le fait que l'ordre des mots est très similaire entre le français et l'anglais. Le passage à IBM 4 ajoute un gain plus modeste en AER par rapport au modèle HMM. Les mêmes tendances s'observent pour la F-mesure.

Modèle	$e \rightarrow f$		$f \rightarrow e$		sym.	
	AER	F	*AER*	F	*AER*	F
IBM 1	30,5 %	71,1 %	26,7 %	74,4 %	19,3 %	81,7 %
HMM	11,3 %	89,3 %	11,0 %	89,4 %	8,7 %	91,9 %
IBM 4	**9,0** %	91,2 %	**9,6** %	90,1 %	**7,1** %	93,0 %

Tableau 2. *AER des modèles génératifs sur les données français/anglais*

Modèle	$a \rightarrow e$	$e \rightarrow a$	sym.
IBM 1	37,3 %	51,0 %	34,3 %
HMM	34,3 %	53,0 %	35,1 %
IBM 4	**30,5** %	**48,2** %	**31,6** %

Tableau 3. *AER des modèles génératifs sur les données arabe/anglais*

Le tableau 3 présente les résultats obtenus par les différents modèles pour la paire de langues arabe/anglais[10]. Ces deux langues sont linguistiquement très différentes,

10. Ce tableau ne donne que l'AER puisque, ce corpus ne distinguant pas les alignements sûrs des alignements probables, la F-mesure se déduit directement de l'AER.

ce qui se reflète dans les résultats : l'AER est nettement plus élevé que pour la paire français/anglais. Deux raisons peuvent être avancées pour expliquer cette différence de performance importante.

Tout d'abord, la difficulté d'aligner de l'anglais vers l'arabe peut être due au caractère agglutinant de la langue arabe : si, comme dans le modèle IBM 1, seules les propriétés lexicales des mots sont prises en compte (sans décomposition ou informations sur les affixes), on ne pourra jamais capturer, à l'aide de cooccurences, des associations entre affixes et morphèmes. Une solution à ce problème serait de travailler sur la décomposition des phrases arabes, afin que les deux langues aient une segmentation similaire (Elming *et al.*, 2009). Toutefois, cette approche complique l'évaluation puisqu'il n'est plus possible d'exploiter les alignements de référence une fois le texte resegmenté.

L'absence de gain entre les modèles IBM 1 et HMM montre également que la modélisation d'alignements monotones s'avère moins efficace, ce qui n'est pas surprenant puisque la distorsion entre les deux langues est nettement plus importante que la distorsion entre le français et l'anglais. En revanche, le passage à IBM 4 améliore l'AER, reflétant l'intérêt de modéliser la distorsion et la fertilité en particulier pour la direction anglais vers arabe.

Ces résultats montrent l'importance de prendre en compte des caractéristiques riches puisque les scores du modèle IBM 4 sont systématiquement meilleurs que ceux du modèle IBM 1. Toutefois, les performances faibles obtenues sur l'alignement arabe/anglais nous incitent à penser que les modèles IBM, initialement développés pour la traduction du français vers l'anglais, ne sont pas adaptés à tous les couples de langues : chaque paire de langues possède des difficultés linguistiques qui lui sont propres et qu'il est nécessaire de capturer pour obtenir de bonnes performances.

Il est donc important que les caractéristiques considérées soient motivées linguistiquement. Or, en général, l'histoire générative sur laquelle repose un modèle génératif complique l'introduction de caractéristiques suffisamment riches et réduit leur aptitude à s'adapter aux différences linguistiques inhérentes à certains couples de langues. Cette observation nous conforte dans notre choix d'utiliser des modèles discriminants pouvant intégrer facilement des caractéristiques arbitraires.

4. Modèle *MaxEnt* pour l'alignement mot à mot

Les modèles IBM que nous avons introduits dans la section précédente sont des modèles *génératifs* qui reposent sur une modélisation de la probabilité jointe des observations (les phrases source et cible $\mathbf{e}$ et $\mathbf{f}$) et des étiquettes (les alignements). Il existe une deuxième famille de modèles d'apprentissage statistique, les modèles *discriminants* (Klein et Manning, 2002 ; Collins, 2004).

Les modèles discriminants proposent de modéliser directement la probabilité sur laquelle la prédiction est fondée, c'est-à-dire la probabilité conditionnelle des éti-

quettes connaissant les observations. Ils traitent donc une tâche plus simple que l'estimation de la probabilité jointe, puisqu'il n'est plus nécessaire d'estimer la probabilité de l'observation qui n'intervient pas directement dans la prise de décision. De plus, ces modèles permettent de prendre en compte des caractéristiques arbitraires, alors que les modèles génératifs, pour des questions d'efficacité, imposent des hypothèses d'indépendance fortes entre les caractéristiques et n'ont donc qu'une expressivité réduite.

Ainsi, dans les modèles IBM, la décision d'aligner deux mots est prise en ne considérant qu'un petit nombre d'observations (essentiellement la forme fléchie des mots). Pourtant, il existe plusieurs autres caractéristiques décrivant les mots (lemme, étiquette morphosyntaxique...), ainsi que leur position dans la phrase ou le contexte dans lequel ils sont utilisés. Ces caractéristiques sont pertinentes pour construire un alignement mot à mot. Utiliser des modèles discriminants pour l'alignement est un moyen simple de les prendre en compte .

Le premier modèle discriminant que nous proposons peut être considéré comme la version discriminante du modèle IBM 1. Nous allons commencer par présenter le modèle *MaxEnt* dans le cas général, puis nous détaillerons l'application de ce modèle à la tâche d'alignement mot à mot (section 4.1). Nous introduirons ensuite les caractéristiques que nous avons utilisées (section 4.2) avant de présenter, dans la section 4.3, les résultats expérimentaux de cette approche.

4.1. *Description du modèle MaxEnt*

4.1.1. *Les classifieurs à maximum d'entropie*

Le classifieur à maximum d'entropie (*MaxEnt*) est un classifieur discriminant dans lequel la probabilité conditionnelle d'une étiquette y connaissant une observation $\mathbf{x}$ est paramétrée par un modèle exponentiel :

$$p(y|\mathbf{x}\,;\theta_y) = \frac{1}{Z(\mathbf{x})} \cdot \exp\langle \theta_y, f(\mathbf{x})\rangle \qquad [7]$$

où $\langle \mathbf{x}, \theta_y\rangle = \sum_k x_k \cdot \theta_{y,k}$ est le produit scalaire usuel, $Z(\mathbf{x}) = \sum_y \exp\langle \theta_y, f(\mathbf{x})\rangle$ un facteur de normalisation, $f(\mathbf{x})$ est une fonction permettant de décrire l'observation $\mathbf{x}$ par un ensemble de caractéristiques et θ_y est le vecteur de paramètres de la classe y. Il y a, dans un classifieur *MaxEnt*, autant de vecteurs de paramètres que de classes possibles. Chaque paramètre $\theta_{y,k}$ décrit l'importance de la k-ième caractéristique dans la prédiction de la classe y, c'est-à-dire à quel point le fait d'observer cette caractéristique est *représentatif* de la classe y. Nous noterons θ l'ensemble des paramètres du modèle. θ est défini par la concaténation de l'ensemble des θ_y.

Les classifieurs *MaxEnt*, comme tous les classifieurs discriminants, n'imposent aucune restriction sur les caractéristiques qui peuvent être prises en compte : la fonction $f(\mathbf{x})$ peut décrire n'importe quelle caractéristique de l'observation que nous jugeons pertinente pour la tâche considérée. La possibilité de considérer facilement des

caractéristiques arbitraires est l'un des principaux avantages des modèles discriminants par rapport aux modèles génératifs (Klein et Manning, 2002).

Les paramètres d'un classifieurs *MaxEnt* sont appris en maximisant la vraisemblance conditionnelle pénalisée du modèle sur un corpus d'apprentissage $(\mathbf{x}_i, y_i)_{i=1}^n$:

$$\theta^* = \underset{\theta \in \Theta}{\operatorname{argmax}} \left\{ \prod_{i=1}^{n} p(y_i|\mathbf{x_i}\,;\theta_{y_i}) - \lambda \cdot \sum_y ||\theta_y||^2 \right\}$$

où $||\cdot||$ est la norme euclidienne usuelle, $\sum_y ||\theta_y||^2$ est un facteur de régularisation (Schölkopf et Smola, 2002) qui permet de limiter les problèmes de surapprentissage[11] ; λ est un métaparamètre, généralement choisi de manière à minimiser l'erreur de classification sur le corpus de développement (section 2.4.1), qui permet de fixer l'importance du terme de régularisation.

4.1.2. *Application à la tâche d'alignement, partage des paramètres*

Le modèle *MaxEnt* décrit dans le paragraphe précédent ne peut pas être utilisé directement pour prédire des alignements mot à mot. En effet, dans sa formulation classique, *MaxEnt* définit un vecteur de paramètres différent pour chaque étiquette à prédire. Or, la tâche d'alignement, telle que nous l'avons introduite dans la section 3.1, consiste à prédire, pour chaque position de la phrase source, la position correspondante dans la phrase cible. Il y a donc autant d'étiquettes et, par conséquent, autant de vecteurs de paramètres que de positions possibles dans les phrases cibles, ce qui n'est pas sans poser problème. En effet, un vecteur de paramètres différent étant utilisé pour chaque étiquette, le poids associé à chaque caractéristique lors de l'inférence ne sera choisi qu'en fonction de la position du mot cible. Ainsi, lors de l'alignement des deux paires de phrases suivantes :

Le$_1$ chocolat$_2$ est$_3$ bon.$_4$	J'$_1$ aime$_2$ le$_3$ chocolat.$_4$
Chocolate$_1$ is$_2$ good.$_3$	I$_1$ love$_2$ chocolate.$_3$

l'existence d'un lien entre *chocolat* et *chocolate* dans la première paire de phrases, déterminée par $p(a_2 = 1|\mathbf{e},\mathbf{f}\,;\theta_1)$, reposera sur les mêmes critères (même vecteur de poids θ_1) que l'existence d'un lien dans la seconde phrase entre *aime* et *I*. Un choix plus pertinent pour prendre ces décisions serait de considérer les mots plutôt que leur position : il est plus sensé d'aligner le mot anglais *chocolate* avec le mot français *chocolat* même si la position de ceux-ci à l'intérieur d'une phrase peut changer.

Pour résoudre ce problème, nous proposons de suivre et de généraliser les travaux de (Blunsom et Cohn, 2006) en proposant d'autres critères pour le partage des paramètres. Dans le modèle *MaxEnt* classique les paramètres sont partagés selon la valeur de l'étiquette. Pour l'alignement, nous envisageons six schémas de partage différents :

11. Notons qu'il est possible d'utiliser d'autres facteurs de régularisation.

– partage selon le mot source ;

– partage selon le mot cible ;

– partage selon le mot source et le mot cible ;

– partage selon l'étiquette morphosyntaxique du mot source ;

– partage selon l'étiquette morphosyntaxique du mot cible ;

– partage selon les étiquettes morphosyntaxiques du mot source et du mot cible.

Dans le premier cas (partage selon le mot source), le vecteur de poids utilisé est déterminé par le mot à étiqueter quelle que que soit l'étiquette et donc la position du mot cible. Le modèle d'alignement *MaxEnt* s'écrit alors :

$$p(a_i = j|\mathbf{e},\mathbf{f}\,;\theta_{e_i}) = \frac{1}{Z(\mathbf{e},\mathbf{f},i)} \cdot \exp\langle\theta_{e_i}, f(\mathbf{e},\mathbf{f},a_i,i)\rangle \qquad [8]$$

où θ_{e_i} est le vecteur de paramètres caractérisant les alignements mettant en jeu le mot source e_i, et $f(\mathbf{e},\mathbf{f},i)$ est un vecteur de caractéristiques arbitraires décrivant la paire de mots à aligner. Les autres schémas de partage sont construits selon le même principe.

Pour expliciter l'impact de ces schémas de partage des paramètres, le choix du mot source e_i comme critère de partage équivaut à construire un modèle *MaxEnt* avec un vecteur de paramètres distinct par mot source observé lors de l'apprentissage. Ainsi chaque caractéristique sera pondérée différemment selon le mot source à aligner. De même, choisir le mot cible comme critère de partage des paramètres revient à pondérer différemment les caractéristiques pour chaque mot cible. Le modèle *MaxEnt* est dans ces deux cas totalement lexicalisé, ce qui pose le problème classique, en traitement automatique des langues, des données éparses.

Une manière de généraliser les observations effectuées lors de l'apprentissage est d'utiliser les catégories morphosyntaxiques des mots comme critère de partage. Cependant, un compromis est à trouver entre un partage trop généraliste et un partage trop précis. Nous abordons ce problème en combinant différents critères de partage des paramètres. Prenons par exemple la combinaison de deux critères de partage des paramètres : le mot source et la catégorie morphosyntaxique du mot source. Le modèle *MaxEnt* s'écrit alors de la manière suivante :

$$p(a_i|\mathbf{e},\mathbf{f}\,;\theta_{e_i},\theta_{P(e_i)}) = \frac{1}{Z(\mathbf{e},\mathbf{f},i)} \cdot \exp\left(\left\langle\theta_{e_i} + \theta_{P(e_i)}, f(\mathbf{e},\mathbf{f},a_i,i)\right\rangle\right), \qquad [9]$$

où $P(e_i)$ représente la catégorie morphosyntaxique du mot e_i. La combinaison des critères de partage s'apparente donc à une forme de lissage permettant une meilleure exploitation des données d'apprentissage (Sutton et McCallum, 2006). Enfin, nous utilisons un vecteur $\theta_{\text{défaut}}$ utilisé systématiquement en sus des critères de partage énoncés précédemment. Étant donné la faible quantité de données d'apprentissage, ce vecteur permet d'éviter les cas où aucune fonction caractéristique ne serait activée lors de l'inférence.

4.2. *Caractéristiques*

Pour paramétrer le modèle présenté dans la section précédente, nous avons utilisé plusieurs types de caractéristiques. Celles-ci sont nommées *unigrammes* car elles ne portent que sur une seule étiquette.

Le premier type de caractéristiques que nous considérons décrit directement les mots à aligner par leur position relative et leur étiquette morphosyntaxique. La position relative de deux mots source et cible est définie par :

$$rsp(e_i, f_{a_i}) = \left| \frac{i}{|\mathbf{e}|} - \frac{a_i}{|\mathbf{f}|} \right| \quad [10]$$

Cette caractéristique permet de favoriser les alignements proches de la diagonale et est intuitivement pertinente dans le cas de paires de langues dans lesquelles l'ordre des mots est globalement conservé lors de la traduction.

Les étiquettes morphosyntaxiques apportent également une information importante sur l'existence d'un alignement entre deux mots : en première approximation, la catégorie morphosyntaxique d'un mot est conservée lors de la traduction (un nom est traduit en un nom et un verbe en un verbe) et nous souhaitons donc arriver à décrire le fait que les deux mots que l'on cherche à aligner ont la même étiquette morphosyntaxique ou non. L'information sur les étiquettes morphosyntaxiques est introduite par trois caractéristiques : une caractéristique décrivant l'étiquette morphosyntaxique du mot source et du mot cible ; une caractéristique décrivant le mot source et l'étiquette morphosyntaxique du mot cible ; une caractéristique décrivant le mot cible et l'étiquette morphosyntaxique du mot source. Pour les données anglaises et françaises, nous avons utilisé le TreeTagger[12] et MADA[13] pour les données arabes.

Nous considérons également un second type de caractéristiques décrivant la tendance du mot source à être aligné avec un mot cible :

– le coefficient Dice (Dice, 1945) qui quantifie la proportion de co-occurrences d'un couple de mots dans des corpus parallèles ;

– la probabilité IBM 1 qui permet d'affiner l'estimation faite par le coefficient Dice (Melamed, 2000).

Ces deux caractéristiques sont estimées à partir des corpus alignées phrase à phrase parmi ceux que nous utilisons. Il peut paraître paradoxal d'utiliser le modèle IBM 1 pour construire des caractéristiques alors que la principale motivation de notre travail est de remplacer les modèles IBM. Cependant l'utilisation de ces caractéristiques est le moyen le plus simple de prendre en compte une information particulièrement pertinente contenue dans les corpus alignés phrase à phrase. Notons également que le modèle IBM 1 est le modèle IBM le plus simple et dont les justifications mathé-

12. http ://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/
13. http ://www1.ccls.columbia.edu/ cadim/MADA.html

matiques sont les plus solides. Sa mise en œuvre est donc aisée et son estimation est rapide, même sur de gros corpus.

4.3. *Évaluation expérimentale*

Dans cette section, nous proposons d'évaluer le modèle *MaxEnt* que nous venons de présenter et notamment l'influence des différents schémas de partage sur les résultats.

4.3.1. *Expérimentation sur le corpus français/anglais*

Les résultats sont rassemblés dans le tableau 4. Pour l'ensemble de ces expériences, la valeur de la pénalité est estimée de manière à minimiser l'AER sur les données de développement.

Schéma de partage	$e \rightarrow f$		$f \rightarrow e$		sym.	
	AER	*F*	*AER*	*F*	*AER*	*F*
Mot source	**16,1** %	83,5	**16,5** %	83,1	**12,2** %	87,4
Mot cible	18,9 %	80,6	22,3 %	77,2	14,5 %	85,2
Mots source et cible	17,8 %	81,8	18,3 %	81,3	14,3 %	85,4
POS source	18,2 %	81,3	20,3 %	79,0	13,9 %	85,6
POS cible	19,5 %	79,9	22,1 %	77,3	15,0 %	84,6
IBM 1 (rappel)	30,5 %	71,1 %	26,7 %	74,4 %	19,3 %	81,7 %
IBM 4 (rappel)	9,0 %	91,2 %	9,6 %	90,17 %	7,1 %	93 %

Tableau 4. *AER et F-mesure sur les données français/anglais pour les différentes manières de partager les paramètres avec un modèle MaxEnt. La dernière ligne rappelle les résultats obtenus avec le modèle IBM 1*

Les résultats montrent que le modèle *MaxEnt* obtient de meilleurs résultats que le modèle IBM 1 pour tous les schémas de partage. En particulier, en partageant les paramètres selon le mot source, l'AER des alignements est de 12,2 % contre 19,3 % pour le modèle IBM 1, alors que l'apprentissage du modèle discriminant ne prend que quelques minutes contre plusieurs heures pour l'apprentissage du modèle IBM 1[14]. Il s'agit du meilleur résultat que nous ayons obtenu avec le modèle *MaxEnt*. La combinaison des différents schémas de partage n'apporte ici aucun gain.

4.3.2. *Expérimentation sur le corpus arabe/anglais*

Les résultats sur les données arabe/anglais sont très différents de ceux obtenus sur les données français/anglais : aucun schéma de partage ne parvient au niveau de per-

14. Les deux programmes ont été codés dans des langages différents et par des personnes différentes. Une comparaison précise de leurs performances n'a pas de sens et nous ne donnons donc que des ordres de grandeur des durées d'apprentissage.

formance du modèle IBM 1 (tableau 3). Remarquons cependant que les deux meilleurs schémas de partage sont sur les mots source/cible et sur l'étiquette morphosyntaxique.

Schéma de partage	$a \rightarrow e$	$e \rightarrow a$	sym.
Mot source	56,6 %	53,5 %	50,1 %
Mot cible	54,6 %	59,4 %	52,2 %
Mots source/cible	43,8 %	47,8 %	40,8 %
POS source	39,9 %	42,5 %	41,5 %
POS cible	41,5 %	43,7 %	43,2 %
Combinaison	**39,0** %	**36,5** %	**34,7** %
IBM 1 (rappel)	37,3 %	51,0 %	34,3 %

Tableau 5. *AER sur les données arabe/anglais pour les différents schémas de partage avec un modèle MaxEnt*

Lorsque tous les schémas de partage sont combinés l'AER diminue de manière significative, sans pour autant atteindre les performances du modèle IBM 1. Ce résultat montre que, pour une tâche complexe comme l'alignement, le choix du schéma de partage revêt une réelle importance et qu'il doit donc être adapté à la paire de langues traitée.

5. CRF pour l'alignement

Il est possible, grâce au modèle *MaxEnt*, d'inclure facilement beaucoup de caractéristiques décrivant le couple de phrases à aligner, ce qui permet d'apprendre, à partir de peu de données, des modèles ayant de bonnes performances en prédiction (section 4.3). Toutefois, dans le modèle décrit dans la section précédente, les alignements sont également prédits indépendamment les uns des autres, comme dans le cas du modèle IBM 1 : les caractéristiques décrivant les alignements des mots voisins ne peuvent pas être prises en compte, même si celles-ci semblent très pertinentes. De plus les performances du modèle *MaxEnt* restent inférieures à celles du modèle IBM 4.

Dans cette section, nous proposons, en nous inspirant de (Blunsom et Cohn, 2006), d'utiliser des modèles issus de l'apprentissage structuré afin d'apprendre et de prédire les alignements de manière conjointe, c'est-à-dire de choisir l'alignement d'un mot en considérant l'alignement de ses voisins. Plus précisément, nous proposons d'utiliser des champs conditionnels aléatoires (CRF) linéaires (Lafferty *et al.*, 2001). Comme pour le modèle *MaxEnt*, les CRF ne peuvent être appliqués directement à la tâche d'alignement et différents schémas de partage des paramètres devront être mis en place.

5.1. *Champs conditionnels aléatoires linéaires*

Dans leur formulation générale, les CRF linéaires proposent de paramétrer la probabilité conditionnelle d'une *séquence* d'étiquettes $\boldsymbol{y}$ connaissant une *séquence* d'observations $\boldsymbol{x}$ de la manière suivante :

$$p(\boldsymbol{y}|\boldsymbol{x}\,;\theta) = \frac{1}{Z(\mathbf{x})} \cdot \exp\left\{\sum_i \left\langle \theta_{y_i,y_{i-1}}, f(y_i, y_{i-1}, \boldsymbol{x})\right\rangle\right\} \qquad [11]$$

où $Z(\boldsymbol{x})$ est un coefficient de normalisation, x_i et y_i les i-ème éléments des séquences $\boldsymbol{x}$ et $\boldsymbol{y}$ et θ la concaténation de l'ensemble des vecteurs de paramètres.

La principale différence entre le classifieur CRF (équation 11) et le classifieur *MaxEnt* (équation 7) réside dans la définition des fonctions caractéristiques : celles-ci dépendent désormais de l'étiquette précédente y_{i-1} en plus de l'étiquette courante y_i et de l'observation $\boldsymbol{x}$. De manière similaire, les paramètres sont choisis en fonction du couple de deux étiquettes consécutives. Ces deux éléments traduisent l'existence d'une dépendance statistique entre deux étiquettes consécutives.

L'estimation des paramètres est effectuée de manière similaire au modèle *MaxEnt* en maximisant la vraisemblance conditionnelle du corpus d'apprentissage. La dépendance entre y_i et y_{i-1} nécessite toutefois d'utiliser des algorithmes de programmation dynamique pour pouvoir calculer efficacement les mises à jour des paramètres ainsi que pour inférer les étiquettes de manière conjointe.

L'inférence jointe permet de prendre en compte une information supplémentaire lors de la prédiction des étiquettes (l'étiquette de l'observation précédente) mais a un coût : la complexité de l'algorithme Viterbi (Forney Jr, 1973) est en $\mathcal{O}\left(n \times m^2\right)$ où n est le nombre d'observations et m le nombre d'étiquettes. De même, lors de l'apprentissage, une adaptation de l'algorithme *forward-backward* est utilisée afin de réduire la complexité des calculs (Rabiner, 1989 ; Sutton et McCallum, 2006).

Il existe aujourd'hui de nombreuses bibliothèques fournissant une implémentation des CRF. Dans nos expériences, nous avons utilisé une version modifiée par nos soins[15] d'une de ces bibliothèques, *GRMM* (Sutton, 2006).

5.2. *Caractéristiques et partage des paramètres*

Les CRF utilisent deux types de caractéristiques : des caractéristiques unigrammes qui ne dépendent que d'une étiquette et des caractéristiques bigrammes ou markoviennes qui dépendent de deux étiquettes. Les caractéristiques unigrammes que

15. Les modifications apportées concernent l'utilisation de caractéristiques réelles et non binaires, les différents schémas de partage des paramètres, et le calcul « à la volée » des caractéristiques lors de l'inférence. En effet, les caractéristiques ne peuvent être calculées « à l'avance » de manière statique, puisqu'elles dépendent, *via* le partage des paramètres, de l'étiquetage en cours.

nous considérons sont les mêmes que celles que nous avons utilisées dans le modèle *MaxEnt* (sous-section 4.2). Nous introduisons également les caractéristiques bigrammes suivantes :

– la largeur de saut qui reflète la tendance des alignements à être monotone. Elle se définit par :

$$f_k(a_i, a_{i-1}, \mathbf{e},\mathbf{f}) = |a_i - a_{i-1} - 1| \text{ si } (a_i \neq 0 \text{ et } a_{i-1} \neq 0) \quad [12]$$

Si les deux alignements adjacents suivent la diagonale, cette caractéristique est nulle, sinon elle est positive ;

– la transition avec un alignement nul qui permet de modéliser les transitions impliquant un alignement nul (car, dans ce cas, la largeur de saut n'est pas définie). Cette caractéristique est représentée par trois fonctions binaires décrivant le cas où la transition s'effectue depuis un alignement nul, vers un alignement nul, ou entre deux alignements nuls.

Comme pour le modèle *MaxEnt*, il est nécessaire d'introduire un schéma de partage des paramètres si l'on veut utiliser les CRF pour la tâche d'alignement de séquences. Nous proposons de partager les paramètres associés aux fonctions caractéristiques unigrammes de la même manière que dans le modèle *MaxEnt* en considérant toutes les combinaisons de caractéristiques que nous avons introduites dans la sous-section 4.2. Pour les fonctions caractéristiques bigrammes, nous considérons les schémas de partage suivants :

– les mots sources e_{i-1}, e_i

– les mots cibles $f_{a_{i-1}}, f_{a_i}$

– le couple de paires de mots sources et cibles, $e_{i-1}, e_i, f_{a_{i-1}}, f_{a_i}$

où (e_{i-1}, e_i) sont deux mots de la source, (a_{i-1}, a_i) les deux alignements dont le score est évalué et $(f_{a_{i-1}}, f_{a_i})$ les deux mots de la cible alignés avec e_i et e_{i-1}.

Ainsi, pour un modèle CRF utilisant un partage des paramètres selon le mot source pour les fonctions caractéristiques unigrammes et bigrammes, l'équation du modèle s'écrit en étendant au CRF l'équation 8 introduite pour le *MaxEnt* :

$$p(\boldsymbol{a}|\mathbf{e},\mathbf{f}\,;\theta) = \frac{1}{Z(\mathbf{e},\mathbf{f})} \times \quad [13]$$
$$\exp\left\{\sum_i \langle \theta_{e_i}, f_u(\mathbf{e},\mathbf{f}, a_i, i)\rangle + \left\langle \theta_{e_i,e_{i-1}}, f_b(\mathbf{e},\mathbf{f}, a_i, a_{i-1}, i\right\rangle)\right\}$$

où $\theta_{e_i,e_{i-1}}$ est le vecteur de paramètres caractérisant les alignements mettant en jeu le couple de mots sources consécutifs e_i, e_{i-1}, $f_u(\mathbf{e},\mathbf{f}, i)$ est un vecteur de caractéristiques unigrammes et $f_b(\mathbf{e},\mathbf{f}, i)$ le vecteur de caractéristiques bigrammes. Les autres schémas de partage sont construits selon le même principe.

5.3. *Résultats expérimentaux*

Nous proposons, dans cette section, d'évaluer les performances des modèles d'alignement utilisant un classifieur CRF. Ces expériences utilisent toutes le même jeu de données et sont réalisées dans les mêmes conditions que les expériences du modèle *MaxEnt*. Les résultats sur le corpus français/anglais sont résumés dans la première partie du tableau 6. En complément, le tableau 7 indique le nombre de paramètres qu'implique chaque schéma de partage.

Le partage selon le couple de mots source et cible obtient de mauvaises performances. Ces résultats peuvent s'expliquer par le nombre important de paramètres de ce modèle par rapport à la taille du corpus d'apprentissage : le modèle est très certainement « surappris » et n'est donc pas capable de généraliser. Les résultats obtenus en partageant les paramètres selon les mots sources constituent une situation presque opposée : les résultats sont mauvais, bien que le nombre de paramètres soit faible, ce qui suggère que le modèle n'est pas assez expressif pour discriminer les mots alignés des mots non alignés.

Enfin, le schéma de partage qui semble être le plus approprié est celui sur les mots cibles. Ce modèle obtient un AER de $15{,}4\ \%$ une fois les alignements symétrisés. Ce résultat reste toutefois moins bon que ceux obtenus avec le meilleur paramétrage du modèle *MaxEnt*.

Schéma de partage	$e \rightarrow f$		$f \rightarrow e$		sym.	
	AER	F	*AER*	F	*AER*	F
Mot source	26,9 %	72,4 %	33,9 %	65,2 %	20,3 %	79,1 %
Mot cible	21,5 %	78,4 %	21,4 %	78,5 %	15,4 %	84,5 %
Mot source et cible	29,1 %	70,2 %	41,3 %	57,7 %	22,0 %	77,4 %
Mot cible + défaut	12,1 %	87,8 %	11,9 %	89,2 %	10,2 %	89,8 %
Meilleure combinaison	**8,9** %	91,0 %	**7,2** %	92,8 %	7,4 %	92,6 %
IBM 4 (rappel)	9,0 %	91,2 %	9,6 %	90,17 %	**7,1** %	93 %

Tableau 6. *AER et F-mesure sur les données français/anglais pour les différents schémas de partage des paramètres avec un modèle CRF*

Schéma de partage	#paramètres
Mot source	8 280
Mot cible	44 128
Mot source et cible	1 178 527
Mot cible + défaut	44 130
Meilleure combinaison	128 180

Tableau 7. *Nombres de paramètres d'un modèle CRF pour chaque modèle selon les schémas de partage*

Pour améliorer ces résultats il est possible, à l'instar de ce qui a été proposé pour les modèles *MaxEnt*, de définir un jeu de paramètres par *défaut* associé aux caracté-

ristiques markoviennes. Cet ajout permet d'améliorer nettement les résultats avec un AER de $10{,}2$ % une fois les alignements symétrisés.

Un gain supplémentaire en terme d'AER est obtenu en ajoutant à ce dernier modèle toutes les caractéristiques unigrammes avec les schémas de partage suivant : mot source, POS cible, mot source/mot cible. L'AER symétrisé est alors de $7{,}4$ %. Une expérience incrémentale *contrastive* montre que l'usage des caractéristiques fondées sur les catégories morphosyntaxiques apporte un gain absolu de $1{,}6$ % en AER. Ce résultat est proche de celui obtenu avec le modèle IBM 4.

Caractéristique	$e \rightarrow f$		$f \rightarrow e$		sym.	
	AER	F	*AER*	F	*AER*	F
Complet	8,9 %	91,0 %	7,20 %	92,7 %	7,4 %	92,6 %
-IBM 1	10,2 %	89,7 %	8,0 %	91,9 %	8,8 %	91,1 %
-Dice	9,1 %	90,8 %	7,4 %	92,6 %	7,4 %	92,5 %
-Dice -IBM 1	17,8 %	81,9 %	14,9 %	84,9 %	17,3 %	82,5 %

Tableau 8. *Expériences incrémentales avec les CRF sur les données anglais/français, étude de l'apport du modèle IBM 1 et du coefficient Dice*

Afin d'utiliser la totalité des données disponibles, nous avons effectué une validation croisée 10 fois, au lieu d'utiliser le découpage en corpus de test/corpus d'apprentissage fait par les créateurs des corpus utilisés. Cette expérience nous permet d'obtenir un AER de $6{,}1$ %, soit un gain absolu de 1 % par rapport au modèle IBM 4. Les résultats en terme de F-mesure sont comparables.

Ces résultats montrent qu'avec peu de données annotées (quelques centaines de paires de phrases), les CRF permettent d'obtenir des résultats comparables, voire meilleurs que ceux des meilleurs modèles génératifs. Les performances des modèles discriminants sont d'autant plus remarquables que les modèles génératifs nécessitent d'être appris sur de grand corpus de données (de l'ordre du million de paires de phrases) non annotées.

Ces résultats doivent cependant être modérés : les modèles discriminants utilisent des informations (par exemple les étiquettes morphosyntaxiques) qui ont été apprises sur d'autres corpus, voire des informations extraites des mêmes corpus que ceux utilisés pour apprendre les modèles génératifs. C'est le cas, dans nos modèles, pour des caractéristiques comme les probabilités IBM 1 et les coefficients Dice.

Afin de quantifier l'impact de ces deux caractéristiques, le tableau 8 rassemble les résultats d'une expérience incrémentale. La première ligne du tableau reprend les résultats obtenus avec le modèle CRF le plus performant du tableau 6 appelé pour l'occasion « complet » puisqu'il utilise toutes les caractéristiques. Le tableau propose ensuite les résultats obtenus en retirant successivement du modèle complet : les caractéristiques IBM 1, le coefficient Dice, puis les deux. Les résultats montrent tout

d'abord que ces deux caractéristiques apportent une information proche[16] et que le modèle IBM 1 utilise un mécanisme d'estimation plus fin. Dans le cas où seules les caractéristiques Dice sont retirées, les performances restent inchangées alors que dans le cas où seules les caractéristiques IBM 1 sont retirées, l'AER se dégrade de $1{,}4$ points. De plus, la dernière ligne du tableau montre clairement qu'il est important pour le modèle CRF qu'une de ces deux caractéristiques soit présente.

Ainsi, les modèles discriminants d'alignement ne peuvent faire l'économie de statistiques simples estimées sur de vastes corpus parallèles[17].

Sur le corpus arabe/anglais, le meilleur modèle est celui combinant tous les modes de conditionnement comme pour le modèle *MaxEnt*. L'AER est de $32{,}9\ \%$ pour l'alignement de l'arabe vers l'anglais, $30{,}4\ \%$ pour l'alignement de l'anglais vers l'arabe et de $28{,}8\ \%$ après symétrisation. L'usage des CRF permet donc d'obtenir un gain significatif par rapport au modèle IBM 4. En utilisant toutes les phrases de test disponibles (même celles ayant une longueur de plus de 25 mots), l'AER obtenu avec les CRF est de $28.4\ \%$.

6. Modèles discriminants pour l'alignement mot à mot, travaux récents

Plusieurs modèles discriminants pour l'alignement mot à mot ont été proposés dans la littérature. Ainsi, dans (Ittycheriah et Roukos, 2005), les auteurs proposent un modèle inspiré du modèle HMM : les probabilités de transition sont fixes et sont fondées uniquement sur la distorsion ; la probabilité d'observation est estimée par un modèle à maximum d'entropie. Le problème de cette approche est qu'il n'y a pas d'inférence exacte possible et que la « qualité » des solutions n'est donc pas garantie. D'autres modèles ont été proposés (Moore, 2005 ; Moore *et al.*, 2006 ; Liu *et al.*, 2005), qui ont en commun de reposer sur une modélisation complexe et ne permettent qu'une inférence approchée requérant de nombreuses heuristiques.

Le modèle proposé par (Taskar *et al.*, 2005) diffère des précédents. Les auteurs posent le problème de l'alignement comme la recherche d'un appariement deux à deux optimal dans un graphe biparti. L'avantage de cette approche est que la recherche de la meilleure solution peut être résolue de manière exacte. Cet algorithme ne permet d'envisager que des alignements `1:1`. De plus, aucune caractéristique n'est définie sur la séquence d'étiquettes. Cette approche se compare donc à IBM 1 et ne modélise pas la monotonicité des alignements.

Les champs aléatoires conditionnels (CRF) ont été utilisés dans les travaux de (Blunsom et Cohn, 2006) et plus récemment de (Niehues et Vogel, 2008). Dans (Blunsom et Cohn, 2006), les auteurs proposent d'utiliser un CRF linéaire de

16. Des statistiques de cooccurrence de paire de mots dans un corpus parallèle.

17. Remarquons tout de même que l'estimation des probabilités IBM 1 prend moins de 1 demi-heure pour les données français/anglais, alors que l'estimation du modèle IBM 4 nécessite 15 heures.

manière similaire au modèle HMM en introduisant des dépendances entre deux étiquettes consécutives. Leur approche est similaire à celle que nous avons présentée dans la section 5. L'utilisation d'un CRF linéaire pose deux problèmes : seules les dépendances entre deux alignements consécutifs sont prises en compte et, comme pour les modèles IBM, le modèle n'est pas symétrique et doit donc être appris dans les deux sens.

(Niehues et Vogel, 2008) proposent de généraliser cette approche pour résoudre ces deux problèmes : dans ce travail, les CRF sont utilisés pour modéliser directement la matrice d'alignement. Cette modélisation permet d'inclure explicitement, et non par le biais de caractéristiques, les notions de fertilité et de distorsion dans le modèle. Mais, comme pour les modèles précédents, la structure complexe du modèle rend impossible l'inférence exacte. De plus, le nombre de paramètres à estimer contraint les auteurs à une optimisation complexe difficilement reproductible.

Un dernier type de modèle discriminant pour l'alignement est introduit par (Ayan *et al.*, 2005). Contrairement aux autres approches, les auteurs ne cherchent pas à remplacer les modèles IBM, mais les heuristiques utilisées lors de la symétrisation. Ils proposent pour cela un modèle discriminant capable d'apprendre à fusionner les sorties de plusieurs systèmes d'alignement.

7. Conclusion

Dans cet article nous avons présenté deux modèles discriminants de la famille exponentielle pour l'alignement mot à mot, afin d'améliorer la qualité des alignements prédits par les modèles de l'état de l'art et de réduire le temps d'apprentissage et d'inférence de ces modèles.

Le premier modèle que nous avons proposé repose sur un classifieur à maximum d'entropie (*MaxEnt*), le second sur les champs conditionnels aléatoires (CRF). Ces deux modèles formalisent l'alignement mot à mot comme une tâche d'étiquetage de séquence dans laquelle chaque mot de la phrase source est associé à l'index d'un mot de la phrase cible. Cette formalisation nous oblige à adapter la formulation classique du classifieur *MaxEnt* et des CRF, dans laquelle le partage des paramètres s'appuie sur la valeur des étiquettes. En effet, contrairement aux tâches de classification habituelles, il n'y a pas, dans cette formalisation de l'alignement, de sémantique claire attachée aux étiquettes prédites : il est intuitivement toujours sensé d'aligner le mot anglais *chocolate* avec le mot français *chocolat* même si la position de ceux-ci à l'intérieur d'une phrase peut changer. Pour résoudre ce problème, nous avons proposé différents schémas de partage des paramètres en utilisant, par exemple, les mots ou les catégories morphosyntaxiques impliqués par les étiquettes plutôt que les étiquettes elles-mêmes.

Les deux modèles que nous avons proposés ont été évalués sur deux paires de langues proposant des difficultés linguistiques différentes : français/anglais et arabe/anglais. Les résultats montrent que, bien choisi, le mode de partage des paramètres a un impact important sur la qualité des alignements prédits et qu'il confère

aux modèles discriminants la souplesse nécessaire pour s'adapter aux difficultés inhérentes à la paire de langues considérée. De plus, l'introduction de dépendances grâce à l'utilisation de CRF permet d'améliorer la qualité des alignements de manière significative, en particulier pour la paire de langues français/anglais pour laquelle la relation de monotonicité est plus importante que pour la paire arabe/anglais.

Les résultats que nous avons obtenus améliorent significativement les résultats du modèle IBM 4 qui est, aujourd'hui, le modèle le plus utilisé : sur la paire français/anglais, le modèle IBM 4 obtient un taux d'erreur (AER) de $7{,}1$ % alors que notre meilleur modèle a un taux d'erreur de $6{,}1$ %. L'amélioration est également significative pour la paire arabe/anglais : le score d'IBM 4 est de $31{,}6$ %, alors que les CRF ont un AER de $28{,}8$ %. Ces résultats peuvent être améliorés par un travail complémentaire sur les caractéristiques utilisées. En particulier, pour la langue arabe, il est envisagé d'intégrer les aspects liés à la décomposition morphologique.

Les modèles discriminants que nous avons présentés peuvent être appris à partir de très peu d'exemples (de l'ordre de quelques centaines de paires de phrases), alors que les modèles de l'état de l'art nécessitent, en général, des corpus beaucoup plus volumineux. Leur temps d'apprentissage et d'inférence sont nettement inférieurs. Les modèles discriminants offrent donc une alternative intéressante aux modèles génératifs aussi bien du point de vue de la qualité des alignements prédits que du point de vue de l'efficacité (calculatoire) des méthodes.

Les travaux futurs visent à intégrer et évaluer ces modèles dans un système de traduction statistique complet. En effet, de nombreux articles récents abordent le lien entre la qualité des alignements (ou du moins des métriques comme l'AER) mais avec des conclusions divergentes. Pourtant, intuitivement, nous pensons que ce lien existe et il nous paraît important de le qualifier et de le quantifier.

Remerciements

Ce travail a été en partie financé par l'ANR dans le cadre du projet CroTal (ANR-07-MDCO-003) . Les auteurs tiennent à remercier Nadi Tomeh pour son aide et son travail indispensable sur les données arabe/anglais, ainsi que Charles Sutton pour ses précieux conseils et éclaircissements sur les graphes de facteurs et l'utilisation de GRMM. Les auteurs remercient également François Yvon, Aurélien Max et les relecteurs anonymes pour leurs remarques constructives qui ont fortement contribué à améliorer cet article.

8. Bibliographie

Ayan N. F., Dorr B. J., Monz C., « NeurAlign : combining word alignments using neural networks », *HLT '05 : Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Morristown, NJ, USA, p. 65-72, 2005.

Berger A. L., Della Pietra S. A., Della Pietra V. J., « A Maximum Entropy Approach to Natural Language Processing », *Computational Linguistics*, vol. 22, n° 1, p. 39-71, 1996.

Blunsom P., Cohn T., « Discriminative word alignment with conditional random fields », *ACL-44 : Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meet ing of the Association for Computational Linguistics*, Association for Computational Linguistics, Morristown, NJ, USA, p. 65-72, 2006.

Brown P. F., Pietra V. J. D., Pietra S. A. D., Mercer R. L., « The mathematics of statistical machine translation : parameter estimation », *Comput. Linguist.*, vol. 19, n° 2, p. 263-311, 1993.

Carpenter B., Lazy Sparse Stochastic Gradient Descent for Regularized Multinomial Logistic Regression, Technical report, Alias-i INC, 2008.

Chiang D., « Hierarchical Phrase-Based Translation », *Comput. Linguist.*, vol. 33, n° 2, p. 201-228, 2007.

Collins M., « Parameter Estimation for Statistical Parsing Models : Theory and Practice of Distribution-Free Methods. », *New Developments in Parsing Technology, Kluwer*, Harry Bunt, John Carroll and Giorgio Satta, 2004.

Deléger L., Merkel M., Zweigenbaum. P., « Translating medical terminologies through word alignment in parallel text corpora », *Journal of Biomedical Informatics*, vol. 42, p. 692-701, 2009.

Dempster A. P., Laird N. M., Rubin D. B., « Maximum likelihood from incomplete data via the em algorithm », *Journal of the Royal Statistical Society*, vol. 39, p. 1-38, 1977.

Dice L. R., « Measures of the amount of ecologic association between species », *Journal of Ecology*, vol. 26, p. 297-302, 1945.

Elming J., Habash N., Crego J., *Learning Machine Translation*, MIT Press, chapter Combination of Statistical Word Alignments Based on Multiple Preprocessing Schemes, p. 93-110, 2009.

Forney Jr G., « The Viterbi Algorithm », *Proceedings of the IEEE*, vol. 61, n° 3, p. 268-278, 1973.

Fraser A., Marcu D., « Measuring Word Alignment Quality for Statistical Machine Translation », *Comput. Linguist.*, vol. 33, n° 3, p. 293-303, 2007.

Fung P., *A statistical view on bilingual lexicon extraction - From parallel corpora to non-parallel corpora*, Kluwer Academic, p. 1-17, 2000.

Ganchev K., Graça J. a. V., Taskar B., « Better Alignments = Better Translations ? », *Proceedings of ACL-08 : HLT*, Association for Computational Linguistics, Columbus, Ohio, p. 986-993, June, 2008.

Gaussier E., Hull D., Ait-Mokhtar S., « Term alignment in use : Machine-aided human translation », *in* J. Véronis (ed.), *Parallel Text Processing*, Kluwer Academic, chapter 13, 2000.

Ittycheriah A., Al-Onaizan Y., Roukos S., The IBM Arabic-English Word Alignment Corpus, Technical Report n° RC24024, IBM, 2006.

Ittycheriah A., Roukos S., « A maximum entropy word aligner for Arabic-English machine translation », *HLT '05 : Proceedings of the conference on Human Language Technology and Empirical Methods in Natur al Language Processing*, Association for Computational Linguistics, Morristown, NJ, USA, p. 89-96, 2005.

Klein D., Manning C. D., « Conditional structure versus conditional estimation in NLP models », *EMNLP '02 : Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, Association for Computational Linguistics, Morristown, NJ, USA, p. 9-16, 2002.

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., « Moses : Open Source Toolkit for Statistical Machine Translation », *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, Association for Computational Linguistics, Prague, Czech Republic, p. 177-180, June, 2007.

Koehn P., Och F. J., Marcu D., « Statistical phrase-based translation », *NAACL '03 : Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, Association for Computational Linguistics, Morristown, NJ, USA, p. 48-54, 2003.

Lafferty J., McCallum, A. Pereira F., « Conditional random fields : Probabilistic models for segmenting and labeling sequence data », *in* M. Kaufmann (ed.), *Proc. 18th International Conf. on Machine Learning*, San Francisco, CA, p. 282-289, 2001.

Liu Y., Liu Q., Lin S., « Log-linear models for word alignment », *ACL '05 : Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, Association for Computational Linguistics, Morristown, NJ, USA, p. 459-466, 2005.

Lopez A., « Statistical machine translation », *ACM Comput. Surv.*, vol. 40, n° 3, p. 1-49, 2008.

Makhoul J., Kubala F., Schwartz R., Weischedel R., « Performance measures for information extraction », *Proceedings of DARPA Broadcast News Workshop*, Herndon, VA, p. 249-252, 1999.

Melamed I. D., « Models of translational equivalence among words », *Comput. Linguist.*, vol. 26, n° 2, p. 221-249, 2000.

Mihalcea R., Pedersen T., « An evaluation exercise for word alignment », *Proceedings of the HLT-NAACL 2003 Workshop on Building and using parallel texts*, Association for Computational Linguistics, Morristown, NJ, USA, p. 1-10, 2003.

Moore R. C., « Improving IBM Word Alignment Model 1 », *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04), Main Volume*, Barcelona, Spain, p. 518-525, July, 2004.

Moore R. C., « A discriminative framework for bilingual word alignment », *HLT '05 : Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Morristown, NJ, USA, p. 81-88, 2005.

Moore R. C., Yih W.-t., Bode A., « Improved discriminative bilingual word alignment », *ACL-44 : Proceedings of the 21st International Conference on Computational Linguistics and*

the 44th annual meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Morristown, NJ, USA, p. 513-520, 2006.

Niehues J., Vogel S., « Discriminative Word Alignment via Alignment Matrix Modeling », *Proceedings of the Third Workshop on Statistical Machine Translation*, Association for Computational Linguistics, Columbus, Ohio, p. 18-25, June, 2008.

Och F. J., Ney H., « A systematic comparison of various statistical alignment models », *Comput. Linguist.*, vol. 29, n° 1, p. 19-51, 2003.

Och F. J., Tillmann C., Ney H., Vi L. F. I., « Improved Alignment Models for Statistical Machine Translation », *University of Maryland, College Park, MD*, p. 20-28, 1999.

Papineni K., Roukos S., Ward T., Zhu W. J., « BLEU : a method for automatic evaluation of machine translation », *in Proceedings of the 40th Annual meeting of the Association for Computational Linguistics*, p. 311-318, 2002.

Rabiner L. R., « A tutorial on hidden markov models and selected applications in speech recognition », *Proceedings of the IEEE*, p. 257-286, 1989.

Rogati M., Yang Y., « Multilingual Information Retrieval Using Open, Transparent Resources in CLEF 2003 », *in* C. Peters, J. Gonzalo, M. Braschler, M. Kluck (eds), *CLEF*, vol. 3237 of *Lecture Notes in Computer Science*, Springer, p. 133-139, 2003.

Schölkopf B., Smola A., *Learning with Kernels : Support Vector Machines, Regularization, Optimization, and Beyond*, MIT Press, 2002.

Sutton C., « GRMM : A Graphical Models Toolkit », 2006, http ://mallet.cs.umass.edu.

Sutton C., McCallum A., « An Introduction to Conditional Random Fields for Relational Learning », *in* L. Getoor, B. Taskar (eds), *Introduction to Statistical Relational Learning*, MIT Press, 2006.

Taskar B., Lacoste-Julien S., Klein D., « A discriminative matching approach to word alignment », *HLT '05 : Proceedings of the conference on Human Language Technology and Empirical Methods in Natur al Language Processing*, Association for Computational Linguistics, Morristown, NJ, USA, p. 73-80, 2005.

Vogel S., Ney H., Tillmann C., « HMM-based word alignment in statistical translation », *Proceedings of the 16th conference on Computational linguistics*, Association for Computational Linguistics, Morristown, NJ, USA, p. 836-841, 1996.

Zens R., Och F. J., Ney H., « Phrase-Based Statistical Machine Translation », *KI '02 : Proceedings of the 25th Annual German Conference on AI*, Springer-Verlag, London, UK, p. 18-32, 2002.