Plusieurs langues (bien choisies) valent mieux qu'une : traduction statistique multi-source par renforcement lexical

Josep Maria Crego[1], Aurélien Max[1,2] et François Yvon[1,2]
{jmcrego,amax,yvon}@limsi.fr
(1) LIMSI-CNRS, Orsay
(2) Université Paris-Sud 11, Orsay

Résumé. Les systèmes de traduction statistiques intègrent différents types de modèles dont les prédictions sont combinées, lors du décodage, afin de produire les meilleures traductions possibles. Traduire correctement des mots polysémiques, comme, par exemple, le mot *avocat* du français vers l'anglais (*lawyer* ou *avocado*), requiert l'utilisation de modèles supplémentaires, dont l'estimation et l'intégration s'avèrent complexes. Une alternative consiste à tirer parti de l'observation selon laquelle les ambiguïtés liées à la polysémie ne sont pas les mêmes selon les langues source considérées. Si l'on dispose, par exemple, d'une traduction vers l'espagnol dans laquelle *avocat* a été traduit par *aguacate*, alors la traduction de ce mot vers l'anglais n'est plus ambiguë. Ainsi, la connaissance d'une traduction français→espagnol permet de renforcer la sélection de la traduction *avocado* pour le système français→anglais. Dans cet article, nous proposons d'utiliser des documents en plusieurs langues pour renforcer les choix lexicaux effectués par un système de traduction automatique. En particulier, nous montrons une amélioration des performances sur plusieurs métriques lorsque les traductions auxiliaires utilisées sont obtenues manuellement.

Abstract. Statistical Machine Translation (SMT) systems integrate various models that exploit all available features during decoding to produce the best possible translation hypotheses. Correctly translating polysemous words, such as the French word *avocat* into English (*lawyer* or *avocado*) requires integrating complex models. Such translation lexical ambiguities, however, depend on the language pair considered. If one knows, for instance, that *avocat* was translated into Spanish as *aguacate*, then translating it into English is no longer ambiguous (*avocado*). Thus, in this example, the knowledge of the Spanish translation allows to reinforce the choice of the appropriate English word for the French→English system. In this article, we present an approach in which documents available in several languages are used to reinforce the lexical choices made by a SMT system. In particular, we show that gains can be obtained on several metrics when using auxiliary translations produced by human translators.

Mots-clés : Traduction automatique statistique, désambiguïsation lexicale, réévaluation de listes d'hypothèses.

Keywords: Statistical Machine Translation, Word Sense Disambiguation, *N*-best list rescoring.

1 Introduction

Les systèmes de traduction statistiques actuels intègrent différents modèles qui mettent en jeu, lors du décodage, le plus d'informations disponibles afin de produire les meilleures traductions possibles. En particulier, dans leur version standard, ces systèmes embarquent un modèle de traduction qui probabilise la correspondance entre des séquences de taille variable en source et en cible, un modèle de réordonnancement, qui évalue les distortions entre l'ordre des mots en source et en cible, et un modèle de langage, qui détermine la probabilité des phrases cible (Koehn *et al.*, 2003). Les scores combinés de ces trois modèles permettent de déterminer la meilleure traduction pour le système. Une déficience, soulignée dans de nombreux travaux, de cette approche est l'absence d'un modèle permettant de résoudre explicitement les ambiguïtés sémantiques en source (Carpuat & Wu, 2005).

Étendre les systèmes standard, afin de pouvoir traduire correctement des mots polysémiques, comme par exemple le mot *avocat* du français vers l'anglais (*lawyer* ou *avocado*), requiert l'intégration de modèles complexes (voir par ex. (Max *et al.*, 2009)). Or, cette difficulté inhérente à la polysémie n'est pas la même en fonction des langues sources considérées. Si l'on dispose, par exemple, d'un document en espagnol dans lequel *avocat* a été traduit par *aguacate*, alors la traduction de ce mot vers l'anglais n'est pas ambiguë et permet donc de renforcer la sélection de la traduction *avocado* pour le système français→ anglais. Dans cet article, nous proposons d'utiliser des documents en plusieurs langues pour renforcer les choix lexicaux opérés par un système de traduction automatique. L'objectif général est d'améliorer un système pour une paire de langues $\mathrm{L}_1 \rightarrow \mathrm{L}_2$ en exploitant conjointement des traductions disponibles dans d'autres langues L_i (avec $i > 2$) et les sorties de systèmes automatiques $\mathrm{L}_i \rightarrow \mathrm{L}_2$. Nous présentons deux manières d'aborder ce problème : 1) en exploitant des traductions humaines disponibles entre les langues L_1 et L_i, et 2) en exploitant des traductions automatiques entre les langues L_1 et L_i.

Cet article est organisé comme suit. Nous commençons par analyser brièvement les approches permettant d'exploiter plusieurs systèmes et des entrées multiples (section 2). Nous décrivons ensuite les particularités de notre approche (section 3.1), et les architectures correspondant aux deux contextes d'application (section 3.2). Nous présentons une évaluation de notre approche sur une tâche de traduction français→ anglais (section 4.1) et les scores de réévaluation des hypothèses de traduction utilisés (section 4.2). Les résultats obtenus en utilisant simultanément les neuf langues auxiliaires disponibles sont présentés (section 4.3), puis nous décrivons deux stratégies de recherche heuristique permettant de trouver des ensembles de langues minimaux menant aux meilleurs résultats possibles (section 4.4). Nous discutons enfin nos résultats et concluons (section 5).

2 Travaux antérieurs

La réévaluation des meilleures hypothèses (*N-best list reranking*) produites par un système de traduction automatique statistique est fréquemment opérée comme un post-traitement en sortie d'un décodeur (ex. (Shen *et al.*, 2004)), car elle permet d'appliquer des modèles plus fins pour la sélection de la meilleure hypothèse sur la liste des N meilleures traductions proposées en première passe[1]. Il est ainsi possible de calculer les scores de modèles difficiles à intégrer lors

[1]Une évaluation de type oracle sur une sortie constituée de 1000 meilleures hypothèses pour nos expériences décrites dans la section 4 montre un potentiel important d'un gain de 8.8 points BLEU entre la meilleure hypothèse

du décodage (par ex. utilisation d'un modèle de langue à grand empan) ou nécessitant des hypothèses correspondant à des phrases complètes, lorsqu'il s'agit par exemple d'exploiter des traits syntaxiques (Hasan *et al.*, 2006).

Un nombre important de travaux ont également porté sur l'exploitation conjointe des sorties proposées par plusieurs systèmes distincts, afin notamment de faire bénéficier le système combiné des forces de chacune des approches implémentées. L'étude présentée par (Rosti *et al.*, 2007) porte sur la combinaison de sorties de systèmes à différents niveaux et montre que les meilleurs gains sont obtenus en combinant les hypothèses à la fois au niveau des mots, des segments et des phrases.

Si la majorité des travaux dans le domaine portent sur la combinaison de systèmes implémentant des approches différentes pour une même paire de langues, (Och & Ney, 2001) ont proposé d'utiliser des traductions disponibles en plusieurs langues et de les traduire vers une même langue, puis de sélectionner pour chaque phrase la traduction obtenue menant au meilleur score, ce qui revient implicitement à sélectionner la meilleure langue source pour chaque phrase à traduire. En procédant de cette manière, ils observent des améliorations notables pour la métrique WER. (Nomoto, 2004) s'inscrit dans la même logique, en reclassant les différentes hypothèses proposées par un modèle de langue cible. Les expériences plus récentes de Schwartz (Schwartz, 2008), si elles mettent clairement en évidence les potentialités de l'approche multi-source, soulignent les limites de la démarche de Och et Ney, dont les gains (pour la métrique BLEU) s'avèrent plus faibles qu'espéré, et discute d'alternatives. Parmi celles-ci, l'utilisation de réseaux de concensus construits à partir de traductions de langues différentes, est conceptuellement simple à implémenter, et conduit effectivement à des améliorations importantes (Callison-Burch *et al.*, 2008; Leusch *et al.*, 2009). Notre méthode, qui s'appuie sur les mêmes intuitions que ces travaux, utilise des moyens sensiblement différents : à la manière de (Hildebrand & Vogel, 2008) (qui ne manipulent qu'une langue source) la combinaison de systèmes s'opère dans une étape de réévaluation, durant laquelle les différentes hypothèses sont renforcées explicitement via des scores qui favorisent les traductions concensuelles.

3 Traduction multisource par renforcement lexical

3.1 Description de l'approche

Notre approche vise à améliorer les performances d'un système pour une paire de langues particulière en exploitant des textes source disponibles en plusieurs langues. Elle se distingue toutefois des approches comparables dans la mesure où elle privilégie une direction de traduction (correspondant au système que nous désignerons dans la suite comme le système *principal*) ; les autres sources disponibles (alimentant des systèmes *auxiliaires*) fournissent des informations supplémentaires susceptibles d'aider le système principal.

L'objectif poursuivi est de renforcer les choix lexicaux qui sont présents dans les hypothèses multiples du système (*N*-best lists) et également proposés par des systèmes traduisant depuis d'autres langues vers la même langue cible. Par exemple, lorsque la traduction d'un mot polysémique est ambiguë, si l'on dispose d'une traduction du texte à traduire dans une autre langue pour laquelle la traduction n'est pas ambiguë, alors cette traduction peut être préférée. Le mot

du système initial et la meilleure hypothèse pour chaque phrase relativement au score BLEU.

français *avocat*, dont deux sens peuvent se traduire en anglais par *lawyer* et *avocado*, possède également deux traductions couvrant les mêmes sens en espagnol, respectivement *abogado* et *aguacate*, mais la traduction de chacune d'elles vers l'anglais n'est plus ambiguë. La connaissance de la traduction en espagnol permet donc de choisir la traduction en anglais. En général, cependant, il est difficile de garantir qu'une traduction n'est pas ambiguë. Les traductions proposées à partir d'autres langues pourront donc jouer le rôle d'indices venant renforcer les choix du système principal.

Une caractéristique importante de cette approche est la dépendance des performances du système principal aux systèmes auxiliaires utilisés. En effet, si ces systèmes tendent à produire de mauvaises traductions, ils peuvent renforcer de mauvaises hypothèses. À l'inverse, une amélioration sensible d'un système auxiliaire permettra d'améliorer le renforcement des choix du système principal, sans que celui-ci n'ait eu à connaître d'amélioration directe. Une implémentation performante de cette approche permettra donc un cercle vertueux, où les améliorations d'un système profiteront également aux autres.

En outre, l'approche proposée se distingue de l'approche *pivot* plus traditionnelle (ex., (Wu & Wang, 2007)) dans laquelle un système est construit pour une paire de langues en traduisant successivement de la langue source vers une langue intermédiaire, puis de cette langue intermédiaire vers la langue cible. Si ce type d'approche est intéressant lorsque les ressources parallèles disponibles sont trop limitées pour constuire un système de traduction direct, elle a comme inconvénient que les erreurs commises lors de la traduction de la langue source vers la langue pivot sont difficilement réparables. La traduction multisource par renforcement lexical permet de renforcer certaines hypothèses d'un système, et offre donc des perspectives de correction de la meilleure hypothèse produite par un système par sélection d'une autre hypothèse. Il est, par ailleurs, possible de n'utiliser qu'une seule langue auxiliaire, et les performances du système seront d'autant améliorées que cette langue sera bien choisie.

3.2 Contextes d'utilisation et architectures des systèmes

Il existe de nombreux contextes dans lesquels des traductions existent dans plusieurs langues et où l'on peut souhaiter traduire vers de nouvelles langues, comme dans le cas de traductions de manuels techniques, où des traductions sont tout d'abord effectuées vers des langues principales puis vers des langues à impact commercial moins important. Un tel contexte de traduction en série permet l'exploitation conjointe, par un système automatique, de textes déjà traduits en plusieurs langues. L'architecture de l'expérience MultiRef correspondante est présentée dans la partie gauche de la figure 1. Dans cet exemple, on traduit un texte du français vers l'anglais, en utilisant les traductions vers l'anglais des textes disponibles en espagnol, allemand et néerlandais. Ces traductions sont utilisées par un module de réévaluation des meilleures hypothèses du système principal, qui sélectionne de nouvelles hypothèses sur la base d'un renforcement lexical. Ce type de réévaluation est particulièrement utile lorsqu'il permet de calculer les scores de modèles qui requièrent des phrases cibles complètes. Bien que cette contrainte ne soit pas imposée par notre approche, la réévaluation permet ici de ne pas avoir à modifier les décodeurs des systèmes de traduction.

La sélection d'hypothèses par renforcement lexical se base cependant sur l'hypothèse forte que les différents textes disponibles seront des traductions assez directes les uns des autres, et qu'ils n'auront pas subi une « localisation » trop importante. Les traductions automatiques, qui sont souvent plus littérales, peuvent ici être utilisées de façon avantageuse. Dans ce contexte, on

pourrait penser que le renforcement lexical pourrait plus simplement être effectué par consultation des modèles de traduction. Or, un système de traduction met en jeu plusieurs modèles, dont en particulier un modèle de langue cible, qui ont pour but de mieux choisir les traductions en contexte. La partie droite de la figure 1 présente l'architecture de l'expérience MultiAuto, dans laquelle plusieurs langues auxiliaires sont utilisées. Les langues utilisées sont les mêmes que pour l'exemple précédent, mais les traductions du texte à traduire en espagnol, allemand et hollandais sont ici obtenues de façon automatique. Ce contexte correspond à des situations beaucoup plus communes dans lesquelles un seul texte source est disponible.

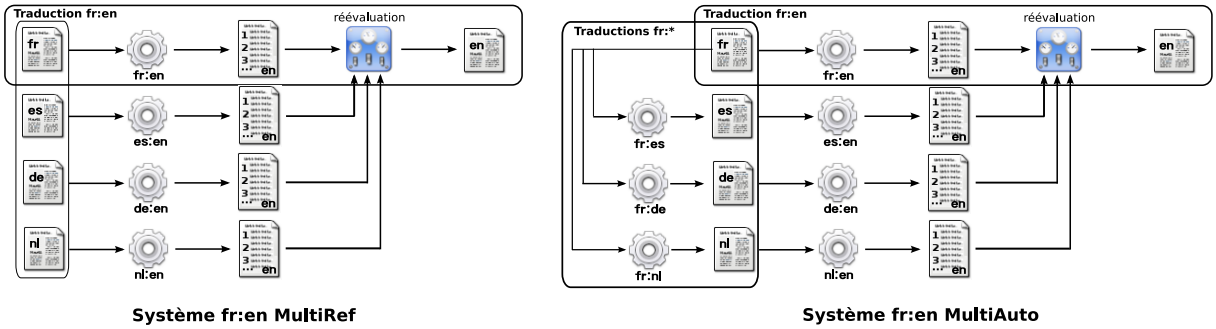

FIG. 1 – Architecture de deux systèmes MultiRef et MultiAuto pour la paire français→anglais avec l'espagnol, l'allemand et le néerlandais comme langues auxiliaires.

4 Expériences et résultats

4.1 Données et systèmes utilisés

Comme source de textes parallèles fortement multilingues, nous avons utilisé le corpus de débats parlementaires européens Europarl (Koehn, 2005) pour les 11 langues suivantes : allemand (de), anglais (en), danois (da), espagnol (es), finlandais (fi), français (fr), grec (el), italien (it), néerlandais (nl), portugais (pt) et suèdois (sv). Afin d'utiliser des systèmes aux performances comparables, nous avons retenu la partie commune à toutes les langues du corpus, pour un total de 318 804 lignes (soit environ 10,3 millions de mots pour le français). La paire de langues principale pour nos expériences est la paire français→anglais.

Tous nos systèmes sont des systèmes statistiques basés sur les tuples (Crego & Mariño, 2007), qui combinent linéairement plusieurs scores. Un modèle de traduction est estimé comme un modèle de langue *n*-gram basé sur les tuples, qui définit une probabilité jointe entre les langues d'une paire (Mariño *et al.*, 2006). Lors du décodage, seuls les réordonnancements encodés dans un treillis de mots sont considérés. Le modèle de réordonnancement en source est appris automatiquement depuis un corpus parallèle bilingue aligné au niveau des mots, et est appliqué aux phrases à traduire avant leur traduction. Les catégories morphosyntaxiques sont utilisées pour généraliser le modèle de réordonnancement. Une règle telle que $NN\ JJ \rightsquigarrow JJ\ NN$ permet par exemple d'exprimer l'inversion adjectif-nom entre le français et l'anglais. Les phrases en français, espagnol et allemand sont analysées avec le TreeTagger[2] pour obtenir les catégories morphosyntaxiques ; pour les autres langues, les règles de réordonnancement sont apprises

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

directement sur des séquences de formes, ce qui conduit, en pratique, à des systèmes moins performants. Une implémentation maison de la recherche du simplexe (Nelder & Mead, 1965) est utilisée pour déterminer les pondérations des différents modèles, en optimisant les scores BLEU (Papineni *et al.*, 2002) obtenus sur un corpus de développement.

Le système de référence correspond à un système de traduction à base de tuples standard. Les systèmes MultiRef et MultiAuto, correspondant respectivement à un système multisource utilisant des traductions de référence et à un système multisource utilisant des traductions automatiques, exploitent les neuf langues auxiliaires restantes.

4.2 Réévaluation de listes d'hypothèses

Nous fondons notre réévaluation des hypothèses du système principal sur le score donné par le décodeur pour chaque hypothèse, ainsi que sur le nombre de *n*-grams d'une traduction candidate qui sont également présents dans une ou plusieurs traductions de « référence », à la manière de la métrique automatique BLEU, qui privilégie les traductions qui sont de taille comparable avec une traduction servant de référence et partageant le plus de *n*-grams communs avec celle-ci. Pour les scores portant sur les unigrammes, seuls les mots n'appartenant pas une liste de mots vides (obtenue par seuillage sur les fréquences à partir du corpus d'apprentissage) ont été retenus, afin de ne pas donner trop d'importance à la présence de mots outils pour lesquels notre approche n'est sûrement pas adaptée. Cette limitation ne pouvait cependant pas s'appliquer de façon naturelle aux n-grams plus grands. Le score calculé par un modèle (interprétable comme un coût) pour chaque langue auxiliaire l correspond à[3] :

$$score(l) = \sum_{n=1}^{4}(1 - np_n(l)) \quad (1)$$

np est la précision n-gram définie comme : $np_n = \frac{|N_n^{hyp} \cap N_n^{ref}|}{|N_n^{hyp}|}$, où N_1^{hyp} et N_1^{ref} correspondent respectivement à l'ensemble des mots pleins de l'hypothèse et de la référence, et N_n^{hyp} et N_n^{ref} ($2 \geq n \geq 4$) correspondent aux n-grams de mots.

Une seconde étape de réglage des poids associés à ces différents scores est effectuée sur un corpus de développement, de nouveau grâce à la méthode du simplexe, qui permet de trouver la combinaison de poids maximisant le score BLEU sur ce corpus.

4.3 Résultats exploitant toutes les langues disponibles

La figure 2 présente les scores automatiques BLEU, TER (Translation Edit Rate) et WER (Word Error Rate) obtenus en ajoutant au score du décodeur un score de précision *n*-gram pour chacune

[3]Nous avons expérimenté avec d'autres scores, tels que la précision unigram seule ou la moyenne des précisions *n*-grams, dont les moins bons résultats, que nous imputons en grande partie au paramètrage utilisé pour l'optimisation par le simplexe, n'ont pas été rapportés ici. En outre, le score BLEU ne pouvait pas être utilisé car celui-ci retourne des scores nuls lorsqu'une valeur de précision *n*-gram (avec typiquement $1 \geq n \geq 4$) est nulle, ce qui arrive fréquemment avec notre approche. L'utilisation de scores indépendants correspondant aux différentes précisions *n*-gram pour les différentes langues nous a permis d'obtenir de bons résultats, mais l'optimisation du nombre de modèles correspondants (jusqu'à $4 * 9 = 36$ avec neuf langues auxiliaires) n'a pas été possible avec la technique d'optimisation utilisée.

des neuf langues ajoutées. La condition MultiRef mène à des gains sur le corpus de test de +1.29 en BLEU, -2.95 en TER[4] et -3.19 en WER, qui montrent la capacité de la méthode présentée à exploiter à bon escient le renforcement lexical implicitement fourni par l'utilisation de plusieurs langues. La condition MultiAuto mène elle à des gains bien moins importants (+0.09 BLEU, -0.53 TER et -0.57 WER), observés cependant sur toutes les métriques, ce qui semble traduire une amélioration du système principal. Les raisons principales pour expliquer les faibles gains obtenus incluent le fait que les systèmes utilisés ne sont pas très performants car appris sur peu de données et n'intègrent pas de modèles de désambiguïsation lexicale. De plus, la dépendance aux seules meilleures sorties de ces systèmes ne permet pas de renforcer des choix lexicaux proposés par les hypothèses suivantes.

	Baseline	**MultiRef**	**MultiAuto**
BLEU	30.47	31.76	30.54
TER	53.73	50.78	53.18
WER	58.08	54.89	57.32

FIG. 2 – Résultats pour les métriques automatiques BLEU, TER et WER en utilisant les neuf langues disponibles pour les trois systèmes.

4.4 Recherche d'un ensemble minimal de langues

Les expériences rapportées en 4.3 considèrent le cas où toutes les langues disponibles sont utilisées. Il est toutefois intéressant d'essayer d'obtenir des performances comparables avec le moins de langues possible, car cela autorise davantage de contextes d'utilisation, et permet notamment d'aborder une traduction fortement multicible par traductions successives. En outre, il est également important de voir si l'utilisation d'un nombre de langues auxiliaires plus petit permet d'améliorer les résultats d'un système, ce qui correspondrait à des cas où une langue auxiliaire serait moins désambiguïsatrice que d'autres relativement à une paire de langues principale, et/ou aurait davantage tendance à renforcer de mauvais choix lexicaux et donc à dégrader les performances d'un système.

La recherche d'une combinaison optimale de langues nécessite la mise en place et l'optimisation des systèmes pour l'ensemble des configurations possibles pour les langues auxiliaires. Avec neuf langues comme précédemment, il y aurait donc $\sum_{k=1}^{9} \mathrm{C}_9^k = 511$ combinaisons à essayer. Afin de proposer une solution plus générale s'appliquant quel que soit le nombre de langues impliquées, nous avons implémenté une recherche heuristique gloutone. L'ensemble des combinaisons impliquant une seule langue auxiliaire est tout d'abord évalué, puis l'ensemble des combinaisons impliquant la meilleure langue et une seconde langue, et successivement les combinaisons impliquant jusqu'à neuf langues. Le nombre maximal de configurations est ainsi réduit à $\sum_{i=9}^{1} i = 45$; la recherche peut être interrompue dès que des pertes supérieures à un seuil sont observées.

Une approche alternative pour diriger la recherche consiste à estimer les contributions en renforcement lexical positif, ainsi qu'en renforcement négatif. On considère pour cela l'ensemble $\mathcal{I}$ constitué de l'intersection, pour chaque phrase à traduire, des *n*-grams présents dans une traduction de « référence » pour la langue cible (ensemble $\mathcal{T}$) et dans les hypothèses du système

[4]TER et WER sont des taux d'erreurs : il est souhaitable de les faire diminuer.

principal (ensemble $\mathcal{P}$). L'ensemble ordonné $\mathcal{L}$, initialement vide, correspond à la séquence des langues par contribution décroissante, où la contribution d'une langue s'entend relativement à l'ensemble des langues précédemment ajoutées. L'ensemble $\mathcal{R}$ (pour « renforcés ») contient l'ensemble des *n*-grams issus de $\mathcal{I}$ qui ont été proposés par l'hypothèse d'au moins une langue ajoutée, et est donc initialement vide. Pour évaluer la contribution d'une nouvelle langue étant donné un état pour $\{\mathcal{L}, \mathcal{I}, \mathcal{R}\}$, on considère les trois valeurs suivantes :

– la quantité de *n*-grams proposés par l'hypothèse de la langue considérée appartenant à $\mathcal{I} \cap \overline{\mathcal{R}}$, notée a ; cela correspond au fait de renforcer des *n*-grams qui n'avaient pas encore été renforcés par d'autres langues, ce qui est donc une contribution très souhaitable ;
– la quantité de *n*-grams proposés par l'hypothèse de la langue considérée appartenant déjà à $\mathcal{R}$, notée b ; cela correspond au fait de renforcer des *n*-grams déjà renforcés par au moins une autre langue, ce qui est une contribution souhaitable ;
– la quantité de *n*-grams proposés par l'hypothèse de la langue considérée appartenant à $\mathcal{P} \cap \overline{\mathcal{I}}$, notée c ; cela correspond au fait de renforcer des *n*-grams proposés par le système principal mais n'appartenant pas à la traduction de référence, ce qui est interprété ici comme une contribution non souhaitable[5].

Dans cette étude, la fonction d'évaluation utilisée par notre seconde recherche heuristique est : $h(l, \mathcal{I}, \mathcal{R}, \mathcal{P}) = 4 * a + 2 * b - c$. Le tableau de la figure 1 donne les résultats pour les 3 métriques automatiques précédentes, tout d'abord en ajoutant une seule langue, puis après chaque ajout de langue en suivant les deux méthodes heuristiques présentées.

On constate tout d'abord que les deux langues ayant l'impact le plus fort individuellement sont une langue proche de la langue source (espagnol) et une langue proche de la langue cible (suèdois). Ces deux langues réalisent la contribution collective la plus marquée (par ex., 75% du gain en BLEU) sur MultiRef.

Il est par ailleurs intéressant de constater que, hormi pour les trois premières langues ajoutées, les deux heuristiques ne se contentent pas d'ajouter les langues par impact individuel décroissant, ce qui laisse entendre que la complémentarité au niveau de la désambiguïsation opérée relève de mécanismes assez complexes. Le cas de l'allemand est à ce titre assez intéressant, puisque, dans les séquences d'ajout incrémental de langues, cette langue semble toujours apporter un complément, et notamment dans le cas de la condition MultiAuto. Enfin, notre deuxième heuristique, qui obtient sur MultiRef des performances comparables à la recherche gloutonne, permet de définir à moindre coût un ordonnancement des langues compétitif.

5 Discussion et perspectives

Dans cet article, nous avons présenté une approche permettant d'améliorer un système de traduction statistique en utilisant des traductions dans d'autres langues du texte à traduire. Des améliorations significatives sont obtenues lorsque ces traductions sont révisées par des humains. Nos résultats actuels ne montrent cependant pas d'amélioration marquée lorsque ces traduc-

[5] Le fait de considérer cette dernière situation comme non souhaitable est discutable : il pourrait en effet s'agir de *n*-grams participant à des traductions correctes bien qu'absents de la traduction de référence. Cela suggère la prise en compte de traductions de références multiples, comme dans les mesures d'évaluation du type de BLEU. Par ailleurs, on ne considère pas ici les *n*-grams proposés qui n'appartiennent pas à $\mathcal{P}$, bien que cette valeur pourrait être utilisée, en particulier pour juger qu'une traduction auxiliaire est trop différente des hypothèses du système principal et ne peut donc pas être utilisée par notre approche.

Ajout d'une seule langue (MultiRef)

Language	-	es	sv	da	fi	it	de	nl	el	pt
BLEU	30.47	**31.05**	30.90	30.76	30.48	30.43	30.59	30.29	30.51	30.62
TER	53.73	**52.45**	52.83	52.67	53.36	53.25	53.24	53.12	52.87	53.22
WER	58.08	**56.67**	57.06	56.91	57.61	57.48	57.55	57.34	56.93	57.53

Recherche gloutone (MultiRef)

Languages	-	+es	+sv	+da	+fi	+it	+de	+nl	+el	+pt
BLEU	30.47	31.05	31.49	31.52	31.54	31.54	31.60	**31.79**	31.78	**31.76**
TER	53.73	52.45	51.51	51.59	51.38	51.40	51.09	50.88	50.87	**50.78**
WER	58.08	56.67	55.69	55.80	55.54	55.53	55.20	55.02	55.00	**54.89**

Recherche par ajout de langue par complémentarité décroissante (MultiRef)

Languages	-	+es	+sv	+da	+pt	+el	+de	+it	+nl	+fi
BLEU	30.47	31.05	31.49	31.52	31.50	31.57	**31.73**	**31.73**	31.54	**31.76**
TER	53.73	52.45	51.51	51.59	51.58	51.55	51.24	51.17	51.12	**50.78**
WER	58.08	56.67	55.69	55.80	55.71	55.65	55.37	55.35	55.23	**54.89**

Recherche par ajout de langue par complémentarité décroissante (MultiAuto)

Languages	-	+es	+sv	+da	+pt	+el	+de	+it	+nl	+fi
BLEU	30.47	30.50	30.47	30.41	30.53	30.45	**30.66**	30.48	30.57	30.54
TER	53.73	53.72	53.73	53.63	53.49	53.48	53.35	53.35	53.25	**53.18**
WER	58.08	58.06	58.08	57.92	57.66	57.76	57.57	57.57	57.40	**57.32**

TAB. 1 – Résultats sur le corpus de test obtenus pour chaque langue auxiliaire et par ajout successif de langues auxiliaires avec les deux méthodes heuristiques

tions auxiliaires sont produites automatiquement. Nos travaux à venir porteront notamment sur l'étude de l'impact de l'amélioration des systèmes auxiliaires utilisés sur la performance du système principal, et sur l'application à d'autres paires de langues. En particulier, nous souhaiterons valider l'hypothèse que la prise en compte du contexte source pour des systèmes auxiliaires, à la manière de (Max *et al.*, 2009), permettra d'améliorer notre approche. Nous avons également proposé dans cet article une approche permettant d'identifier de façon heuristique un ensemble minimal de langues menant aux gains les plus importants. Un résultat particulier de notre étude est que les langues les plus utiles parmi neuf langues européennes pour améliorer un système français→anglais sont l'espagnol et le suèdois.

Parmi les perspectives de ce travail, nous envisageons également d'intégrer un travail spécifique sur l'optimisation du système de réévaluation. Par ailleurs, nous porterons notre attention sur les niveaux de correspondance utilisés pour le renforcement lexical, en passant du niveau des phrases complètes au niveau des tuples, ainsi que sur l'amélioration de la robustesse de notre approche, en considérant plusieurs références plutôt qu'une seule ou en ayant recours aux lemmes, synonymes ou paraphrases locales des *n*-grams impliqués.

Références

CALLISON-BURCH C., FORDYCE C. S., KOEHN P., MONZ C. & SCHROEDER J. (2008). Further meta-evaluation of machine translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, p. 70–106, Columbus, Ohio.

CARPUAT M. & WU D. (2005). Word sense disambiguation vs statistical machine translation. In *Proceedings of ACL*, p. 387–394, Ann Arbor, USA.

CREGO J. & MARIÑO J. (2007). Improving statistical mt by coupling reordering and decoding. *Machine Translation*, **20**(3), 199–215.

HASAN S., BENDER O. & NEY H. (2006). Reranking translation hypotheses using structural properties. In *Proceedings of the EACL06 Workshop on Learning Structured Information in Natural Language Applications*, p. 41–48, Trento, Italy.

HILDEBRAND A. S. & VOGEL S. (2008). Combination of machine translation systems via hypothesis selection from combined n-best lists. In *Proceedings of the Eighth Conference of the Association for Machine Translation in the Americas*, p. 254–261, Waikiki, Hawaï.

KOEHN P. (2005). Europarl : A parallel corpus for statistical machine translation. In *Proceedings of MT Summit*, Phuket, Thailand.

KOEHN P., OCH F. J., & MARCU D. (2003). Statistical phrase-based translation. In *Proceedings of NAACL/HLT*, p. 127–133, Edmonton, Canada.

LEUSCH G., MATUSOV E. & NEY H. (2009). The RWTH system combination system for WMT 2009. In *Proceedings of the ACL workshop on Statistical Machine Translation*, p. 51–55, Athens, Greece.

MARIÑO J., BANCHS R., CREGO J., DE GISPERT A., LAMBERT P., FONOLLOSA J. & COSTAJUSSÀ M. (2006). N-gram based machine translation. *Computational Linguistics*, **32**(4), 527–549.

MAX A., MAKHLOUFI R. & LANGLAIS P. (2009). Prise en compte de dépendances syntaxiques pour la traduction contextuelle de segments. In *Actes de TALN 2009*, Senlis, France.

NELDER J. & MEAD R. (1965). A simplex method for function minimization. *The Computer Journal*, **7**, 308–313.

NOMOTO T. (2004). Multi-engine machine translation with voted language model. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04), Main Volume*, p. 494–501, Barcelona, Spain.

OCH F. J. & NEY H. (2001). Statistical multi-source translation. In *Proceedings of MT Summit*, Santiago de Compostela, Spain.

PAPINENI K., ROUKOS S., WARD T. & ZHU W.-J. (2002). Bleu : a method for automatic evaluation of machine translation. In *Proceedings of ACL*, p. 127–133, Philadelphia, USA.

ROSTI A.-V., AYAN N. F., XIANG B., MATSOUKAS S., SCHWATZ R. & DORR B. J. (2007). Combining outputs from multiple machine translation systems. In *Proceedings of NAACL-HTL*, p. 127–133, Rochester, USA.

SCHWARTZ L. (2008). Multi-source translation methods. In *MT at work : Proceedings of the Eighth Conference of the Association for Machine Translation in the Americas*, p. 279–288, Waikiki, Hawaï.

SHEN L., SARKAR A. & OCH F. J. (2004). Discriminative reranking for machine translation. In D. M. SUSAN DUMAIS & S. ROUKOS, Eds., *HLT-NAACL 2004 : Main Proceedings*, p. 177–184, Boston, Massachusetts, USA : Association for Computational Linguistics.

WU H. & WANG H. (2007). Pivot language approach for phrase-based statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, p. 856–863, Prague, Czech Republic.