Modèles statistiques enrichis par la syntaxe pour la traduction automatique

Holger Schwenk, Daniel Déchelotte
Hélène Bonneau-Maynard, Alexandre Allauzen
LIMSI-CNRS, B.P. 133, 91403 Orsay cedex
{schwenk,dechelot,hbm,allauzen}@limsi.fr

Résumé. La traduction automatique statistique par séquences de mots est une voie prometteuse. Nous présentons dans cet article deux évolutions complémentaires. La première permet une modélisation de la langue cible dans un espace continu. La seconde intègre des catégories morpho-syntaxiques aux unités manipulées par le modèle de traduction. Ces deux approches sont évaluées sur la tâche Tc-Star. Les résultats les plus intéressants sont obtenus par la combinaison de ces deux méthodes.

Abstract. Statistical phrase-based translation models are very efficient. In this paper, we present two complementary methods. The first one consists in a a statistical language model that is based on a continuous representation of the words in the vocabulary. By these means we expect to take better advantage of the limited amount of training data. In the second method, morpho-syntactic information is incorporated into the translation model in order to obtain lexical disambiguation. Both approaches are evaluated on the Tc-Star task. Most promising results are obtained by combining both methods.

Mots-clés : traduction automatique, approche statistique, modélisation linguistique dans un espace continu, analyse morpho-syntaxique, désambiguïsation lexicale.

Keywords: statistical machine translation, continuous space language model, POS tagging, lexical disambiguation.

1 Introduction

La traduction automatique est un thème de recherche depuis plusieurs décennies et différentes approches ont été proposées, telles que la traduction par règles, la traduction à base d'exemples ou la traduction statistique. Les travaux récents en traduction statistique confirment que les modèles fondés sur des séquences de mots (Och *et al.*, 1999; Koehn *et al.*, 2003) obtiennent des performances significativement meilleures que ceux fondés sur des mots (Brown *et al.*, 1993). En utilisant des séquences de mots, les systèmes de traduction parviennent à préserver certaines contraintes locales sur l'ordre des mots. L'entraînement d'un tel modèle nécessite l'alignement d'un corpus parallèle. Les régularités du langage naturel comme celles de la syntaxe, ou, encore à un niveau supérieur, celles de la sémantique sont ainsi, en principe, implicitement capturées par les modèles.

Depuis les débuts de l'approche statistique en traduction automatique, les efforts de modélisation se sont principalement concentrés sur les modèles de traduction et d'alignement, comme en témoignent les nombreuses publications sur ces sujets. Dans cet article, nous explorons deux pistes complémentaires pour l'amélioration des modèles de traduction statistique : d'une part, l'exploration d'une modélisation statistique du langage dans un espace continu, et d'autre part l'intégration d'informations syntaxiques dans le modèle de traduction.

Traditionnellement, les systèmes de traduction statistiques utilisent des modèles de langage trigramme à repli. Dans ces modèles classiques, les mots sont représentés par un indice dans un espace discret, le vocabulaire. Ceci ne permet pas de faire de véritables interpolations des probabilités d'un n-gramme non observé puisqu'un changement dans l'espace des mots peut entraîner un changement arbitraire de la probabilité. Nous proposons ici d'appréhender *dans un domaine continu* le problème de l'estimation d'un modèle linguistique. L'idée consiste à projeter les indices des mots dans une représentation continue (un espace vectoriel) et d'estimer les probabilités dans cet espace (Bengio *et al.*, 2003). Actuellement, un réseau de neurones multi-couches complètement connecté est utilisé pour apprendre conjointement la projection des mots sur un espace continu et l'estimation des probabilités n-grammes.

La lecture humaine des sorties d'un système statistique de traduction, même basé sur des séquences de mots, nécessite parfois un difficile exercice de réordonnancement et de restructuration syntaxique pour restituer le sens de l'énoncé d'origine. La modélisation du langage comme une source markovienne (modèle de langage n-gramme), avec comme unité le mot ou la séquence de mots, ne permet pas de prendre en compte les contraintes syntaxiques ou les dépendances à long terme entre les mots. Il apparaît donc nécessaire d'utiliser des méthodes dans lesquelles les propriétés structurelles des langues sont explicitement représentées. Plusieurs tentatives sur l'utilisation d'informations morpho-syntaxiques dans la traduction statistique ont déjà été menées. (Och *et al.*, 2004) ont exploré de nombreuses fonctions caractéristiques, dont certaines d'ordre syntaxique. La réévaluation des n meilleures hypothèses avec des étiquettes morpho-syntaxiques a également été étudiée par (Hasan *et al.*, 2006). Dans (Kirchhoff & Yang, 2005), un modèle de langage factorisé quadrigramme utilisant des informations syntaxiques n'a pas montré des performances meilleures qu'un modèle n-gramme de mots. Les modèles de langage fondés sur la syntaxe ont enfin été explorés par (Charniak *et al.*, 2003). Tous ces travaux ont en commun d'utiliser des séquences de mots comme unités du système de traduction et de n'introduire les catégories morpho-syntaxiques que dans une seconde passe de traitement.

Dans ce travail, nous proposons d'intégrer les informations syntaxiques *dans* le modèle de traduction lui-même. De plus, nous proposons de combiner cette approche avec les méthodes classiques de réévaluation de listes de n meilleures hypothèses. À notre connaissance, cette approche n'a pas été évaluée sur une large tâche (elle a été appliquée par (Hwang *et al.*, 2007) à la tâche BTEC (Basic Travel Expression Corpus) beaucoup plus réduite). Nous présentons ici des résultats sur la tâche Tc-Star (traduction des transcriptions des sessions plénières du Parlement européen).

Cet article est organisé comme suit. Dans la section suivante, nous présentons d'abord la structure du système de traduction automatique et ses différentes extensions. Les résultats expérimentaux sont résumés et discutés dans la section 3. La dernière section conclue cet article et suggère des extensions et travaux futurs.

2 Description du système

L'objectif d'un système de traduction automatique est de proposer pour une phrase $\mathbf{f}$ en langue « source » sa traduction en une phrase $\mathbf{e}$ dans la langue « cible ». L'approche statistique consiste à choisir, parmi les phrases possibles, la plus probable. Le problème se décompose de la manière suivante :

$$\mathbf{e}^* = \arg\max_{\mathbf{e}} \Pr(\mathbf{e}|\mathbf{f}) = \arg\max_{\mathbf{e}} \Pr(\mathbf{f}|\mathbf{e}) \Pr(\mathbf{e}),$$

où la probabilité $\Pr(\mathbf{f}|\mathbf{e})$ est estimée par le modèle de traduction et $\Pr(\mathbf{e})$ par le modèle de langage de la langue cible. Cette équation résume l'approche *source/canal* historique (Brown *et al.*, 1993) qui considère le mot comme unité et la phrase comme une séquence de mots. Le modèle de traduction peut être estimé automatiquement à partir de textes parallèles alignés au niveau de la phrase. Ce calcul est effectué par le logiciel libre GIZA++.

Ces dernières années, les travaux en traduction statistique ont étendu avec succès l'unité qu'était le mot à la séquence de mots (Och *et al.*, 1999; Koehn *et al.*, 2003). Cette nouvelle unité se définit alors comme un groupe de mots successifs $\tilde{\mathbf{f}}$ de la langue source. Sa traduction est également une séquence de mots $\tilde{\mathbf{e}}$ dans la phrase cible. Les séquences de mots peuvent être extraites automatiquement à partir de données bilingues alignées au niveau du mot dans les deux sens. L'utilisation du principe du maximum d'entropie permet de décomposer le problème de la manière suivante (Och & Ney, 2002) :

$$\mathbf{e}^* = \arg\max p(\mathbf{e}|\mathbf{f}) = \arg\max_{e} \{\exp \sum_i \lambda_i h_i(\mathbf{e}, \mathbf{f})\} \quad (1)$$

où chaque fonction h_i quantifie l'adéquation des phrases $\mathbf{f}$ et $\mathbf{e}$[1]. Les coefficients λ_i pondèrent l'importance relative de ces fonctions.

2.1 Décodeur Moses

Moses[2] est un système de traduction automatique à base de séquences de mots à l'état de l'art. Il est distribué librement avec les scripts nécessaires à l'entraînement d'un système de traduction complet, ainsi qu'une mise en œuvre efficace d'un algorithme de recherche de type *recherche en faisceau* pour produire les traductions. Le décodeur Moses peut également générer une liste des n hypothèses envisagées les plus probables. Cette liste des n meilleures hypothèses contient en général plusieurs fois la même phrase, avec des probabilités différentes, puisque plusieurs segmentations de la phrase source en séquences de mots peuvent aboutir à une même phrase cible. Comme effectué dans les expériences ci-dessous, il est possible de contraindre le décodeur pour que cette liste contienne n hypothèses distinctes.

Dans sa version standard, Moses utilise huit fonctions caractéristiques modélisant le processus de traduction. Ces fonctions permettent d'intégrer à la recherche de la phrase cible les contraintes suivantes : les probabilités de traduction des séquences de mots dans les

[1] Cette « adéquation » est à prendre au sens large, puisqu'un système de traduction inclut toujours un modèle de langage cible $h_i(\mathbf{e}, \mathbf{f}) = p(\mathbf{e})$.

[2] `http://www.statmt.org/moses/`

deux sens, les probabilités de traduction des mots dans les deux sens, une mesure de distorsion, deux pénalités d'insertion de mots et de séquences de mots, et la probabilité calculée par le modèle de langage de la langue cible.

L'approche couramment employée pour optimiser les poids λ_i des fonctions caractéristiques est la maximisation sur un corpus de développement de la mesure BLEU (Papineni *et al.*, 2002). Pour cela, l'outil d'optimisation numérique *Condor* (Berghen & Bersini, 2005) est intégré à l'algorithme itératif suivant :

1. Partant d'un jeu de poids initial, les listes des $n = 1000$ meilleures hypothèses sont générées avec Moses (une liste par phrase source).
2. Ces listes sont réévaluées en utilisant le jeu de poids courant.
3. Les meilleures hypothèses sont extraites et évaluées.
4. À partir du score BLEU aisni calculé, *Condor* calcule un nouveau jeu de poids (l'algorithme retourne alors à l'étape 2), sauf si un maximum local est détecté ce qui met fin à l'algorithme.

Le jeu de poids solution est en général trouvé après une centaine d'itérations. Remarquons que les listes des 1000 meilleures hypothèses sont générées une seule fois lors de l'initialisation et que les itérations réévaluent les listes des 1000 meilleures hypothèses en fonction des poids proposés par *Condor*.

2.2 Désambiguïsation lexicale par catégories syntaxiques

D'une langue à l'autre, les structures et les propriétés syntaxiques diffèrent, par exemple l'espagnol est une langue fortement fléchie alors que l'anglais l'est peu. Or ces structures syntaxiques induisent des ambiguïtés lexicales qui ne sont pas explicitement prises en compte par la modélisation statistique du processus de traduction décrit dans la section ci-dessus.

Il est toujours possible d'utiliser des modèles de langage n-grammes de catégories morpho-syntaxiques pour réévaluer les listes des n meilleures hypothèses de mots générées par un système de traduction. Ce processus nécessite alors d'étiqueter les hypothèses contenues dans les listes. Cependant, les étiqueteurs morpho-syntaxiques ont été appris sur des énoncés correctement formés, ce qui n'est pas toujours le cas des hypothèses provenant d'un système de traduction automatique. Cette étape peut ainsi être une source d'erreurs qui limite les performances de la réévaluation. Nous proposons donc d'intégrer les catégories morpho-syntaxiques *au cœur* du modèle de traduction, ce qui permet d'éviter cet écueil. L'étiqueteur est alors utilisé sur des énoncés syntaxiquement corrects (en tout cas, des énoncés réellement produits), ici sur les corpus parallèles. Par ailleurs, utiliser lors de l'apprentissage des corpus étiquetés morpho-syntaxiquement dans les deux langues permet de prendre en compte les spécificités syntaxiques des deux langues et leur interaction, alors que dans le cas de la réévaluation des listes de meilleures hypothèses, seules les spécificités de la langue cible interviennent.

Nous proposons d'utiliser dans le modèle de traduction des **unités enrichies** constituées des formes de surface des mots, auxquelles sont agglutinées leurs catégories morpho-syntaxiques respectives. Cette méthode permet une désambiguïsation des mots tenant

compte de leurs rôles et de leurs contextes grammaticaux. Un exemple d'énoncé, avec les unités enrichies, est donné à la Figure 1 en anglais et en espagnol.

Anglais : I_{PP} declare_{VVP} resumed_{VVD} the_{DT} session_{NN} of_{IN} the_{DT} European_{NP} Parliament_{NP}

Espagnol : declaro_{VLfin} reanudado_{VLadj} el_{ART} período_{NC} de_{PREP} sesiones_{NC} del_{PDEL} Parlamento_{NC} Europeo_{ADJ}

FIG. 1 – Exemple d'un texte parallèle composé d'unités enrichies utilisé pour entraîner le modèle de traduction.

Lorsque les modèles de traduction et de langage sont fondés sur les unités enrichies, le système de traduction attend en entrée et produit en sortie des séquences d'unités enrichies. Ainsi les phrases à traduire doivent être préalablement étiquetées. Réciproquement, si une traduction classique est requise en sortie, il est nécessaire de retirer les catégories morpho-syntaxiques de l'hypothèse proposée.

Par ailleurs, il devient possible, sur la base des n meilleures hypothèses enrichies, d'effectuer une réévaluation en utilisant un modèle n-gramme de catégories morpho-syntaxiques, sans avoir à utiliser *a posteriori* un étiqueteur sur ces hypothèses.

Pour les expériences présentées dans cet article, nous avons utilisé *TreeTagger* (Schmid, 1994), un étiqueteur markovien utilisant des arbres de décision pour estimer les probabilités trigramme de transition. Ce logiciel est librement disponible pour les deux langues considérées dans cet article. La version anglaise a été entraînée sur le corpus *PENN treebank*[3], et la version espagnole sur le corpus *CRATER*[4]. Le nombre de catégories est assez restreint : 59 pour l'anglais et 69 pour l'espagnol. Notons que les catégories espagnoles ne contiennent pas de distinction en genre et en nombre.

2.3 Modèle de langage neuronal

L'architecture du modèle de langage neuronal est résumée à la Figure 2. Un réseau de neurones multi-couches complètement connecté est utilisé pour apprendre conjointement la projection des mots dans un espace continu et l'estimation des probabilités n-grammes.

Les entrées du réseau sont les $n-1$ mots précédents du vocabulaire et les sorties sont les probabilités a-posteriori pour *tous* les mots du vocabulaire :

$$P(w_j = i | w_{j-n+1}, ..., w_{j-2}, w_{j-1}) = P(w_j = i | h_j) \qquad \forall i \in [1, N] \tag{2}$$

où N est la taille du vocabulaire et h_j le contexte $w_{j-n+1}, ..., w_{j-1}$. Ces entrées sont projetées sur un espace continu (couche P dans la Figure 2). Les autres couches servent à l'estimation non-linéaire des probabilités. La valeur de la i-ème sortie correspond à la probabilité du n-gramme $P(w_j = i | h_j)$. Le réseau calcule donc directement les probabilités de *tous* les mots du vocabulaire pour le même contexte. L'apprentissage se fait par rétropropagation du gradient, en utilisant la cross-entropie comme fonction d'erreur.

[3] `http://www.cis.upenn.edu/~treebank`
[4] `http://www.comp.lancs.ac.uk/linguistics/crater/corpus.html`

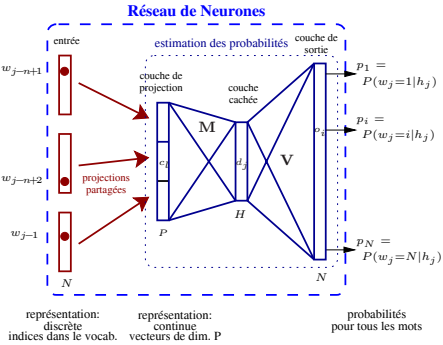

Fig. 2 : Architecture du modèle de langage neuronal. h_j dénomme le contexte $w_{j-n+1}, \ldots, w_{j-1}$. P est la taille d'une projection, et H et N correspondent à la dimension de la couche cachée et de sortie, respectivement.

Dans ce modèle, la complexité est dominée par la taille importante de la couche de sortie. Ainsi, nous proposons de limiter l'estimation des probabilités aux 8 192 mots les plus fréquents, les autres mots étant traités par le modèle à repli standard. Dans nos expériences, environ 90% des requêtes de probabilités sont traitées par le réseau de neurones. Il est important de noter que tous les mots du vocabulaire sont considérés à l'entrée du réseau.

Ce modèle de langage a été utilisé avec succès dans un système de reconnaissance de la parole à grand vocabulaire (Schwenk, 2007), et dans un système de traduction statistique pour la tâche Btec avec un nombre très limité de données d'apprentissage (Schwenk *et al.*, 2006). Cet article décrit la première application du modèle de langage neuronal dans un système de traduction statistique avec plusieurs milliers d'exemples d'apprentissage.

3 Résultats expérimentaux

Les expériences décrites dans cet article ont été effectuées dans le cadre des évaluations internationales organisées par le projet européen Tc-Star[5]. L'objectif de ce projet est de motiver, fédérer, et promouvoir les recherches sur la traduction automatique de la parole. La tâche principale de ce projet est la traduction des transcriptions des sessions plénières du Parlement européen (SPPE). La communauté européenne met à disposition les minutes de ces sessions en plusieurs langues, aussi connues sous le nom « Éditions du texte final » (ETF). Ces textes, alignés au niveau des phrases, sont utilisés pour apprendre les modèles statistiques. Nous disposons également d'environ 100 heures d'enregistrement des sessions plénières du Parlement européen. Ces données audio ont été transcrites manuellement et servent principalement au développement des systèmes de reconnaissance de la parole, mais elles sont aussi utilisées pour entraîner les modèles de langage cible dans le système de traduction.

Trois conditions sont considérées dans les évaluations Tc-Star : la traduction des minutes ETF (*texte*), la traduction des transcriptions des données acoustiques (*verbatim*) et la traduction des hypothèses du système de reconnaissance de la parole (*parole*). Dans ce travail, nous ne considérons que la condition *verbatim*, pour la paire de langues espagnol/anglais. Nous donnons des résultats sur les données de développement et de test de

[5] http://www.tc-star.org/

l'évaluation organisée en 2007. Deux traductions de référence sont disponibles pour les deux jeux de test. Plusieurs étapes de normalisation ont été appliquées aux minutes des sessions plénières afin d'approcher la condition *verbatim* ou *parole*, notamment la transformation en mots des nombres. Les modèles de traduction sont estimés sur les données SPPE qui représentent 1,2M de phrases parallèles, soit environ 35M de mots en anglais.

3.1 Apprentissage des modèles de langage

Pour l'apprentissage des modèles de langage, nous avons utilisé la partie monolingue des données parallèles SPPE ainsi que les transcriptions des données acoustiques. Des données extérieures ont également été utilisées pour une estimation plus robuste des modèles : deux corpus de textes provenant des parlements espagnol (49M mots) et britannique (55M mots). Ainsi, pour chaque langue, nous disposons de trois sources de texte donnant lieu à l'estimation de trois modèles indépendants. Ces trois modèles sont *in fine* interpolés linéairement pour créer un modèle de la langue cible. Les coefficients d'interpolation sont estimés via l'algorithme E.M. de manière à minimiser la perplexité sur les données de développement. Les coefficients obtenus sont 0,81 pour le modèle SPPE, 0,12 pour le modèle estimé sur les données additionnelles du parlement et 0,07 pour celui utilisant les transcriptions acoustiques.

Tous les modèles de langage n-grammes utilisés, hormis le modèle neuronal, sont des modèles classiques avec repli utilisant le lissage de Kneser-Ney modifié. Le SRI LM-toolkit (Stolcke, 2002) a été utilisé pour leur construction.

Les caractéristiques des données et les perplexités des modèles de langage sont résumées dans le Tableau 1. Les modèles trigrammes interviennent pendant le décodage, alors que les modèles quadrigrammes sont utilisés pour réévaluer les listes de n meilleures hypothèses. Le modèle de langage neuronal obtient une réduction de la perplexité de 15% environ. Il est à noter que les données de développement en anglais, donc la traduction de l'espagnol vers l'anglais, proviennent de deux sources différentes (parlements européen et espagnol). Cette différence explique les perplexités relativement élevées. Les perplexités sur les données du Parlement européen uniquement sont plus basses : 85.0, 77.8 et 64.3 pour le tri-, quadrigramme à repli et le quadrigramme neuronal respectivement.

	Anglais	Espagnol
Textes du Parlement européen	35,3M	36,6M
Textes parlementaires supplémentaires	55,1M	48,9M
Transcriptions acoustiques	1,5M	777k
Vocabulaire	82,6k	132,5k
Perplexité trigramme	134,5	69,7
Quadrigramme à repli	123,4	64,0
Quadrigramme neuronal	102,8	54,6

Tab. 1 – Données d'apprentissage (en nombre de mots) utilisées pour l'estimation des modèles de langage et perplexités obtenues sur les données de développement.

3.2 Résultats sur les données de développement

Nous avons effectué de nombreuses études comparatives sur les données de développement pour évaluer les apports des différentes techniques. Les résultats principaux sont résumés dans le Tableau 2. En ce qui concerne la désambiguïsation lexicale, seul le sens de traduction de l'anglais vers l'espagnol (vers la langue la plus infléchie) a été évalué à ce jour. Pour chaque sens de traduction, le score Bleu du modèle de base avec un trigramme est donné, ainsi qu'après la réévaluation avec un quadrigramme à repli et neuronal.

L'utilisation d'un quadrigramme permet d'augmenter le score Bleu d'environ 0,4 points pour la traduction vers l'anglais et de 0,6 points vers l'espagnol. Nous avons également essayé de réévaluer les n meilleures hypothèses avec des modèles de langage n-grammes de catégories morpho-syntaxiques, mais sans effet sur les performances du système. L'utilisation du modèle de langage neuronal, par ailleurs, produit une amélioration du score BLEU de plus de 0,6 points pour les deux directions.

	Espagnol → anglais			Anglais → espagnol					
	Sans désambiguïsation			Sans désambiguïsation			Avec désambiguïsation		
	base	4-gram	NNLM	base	4-gram	NNLM	base	4-gram	NNLM
BLEU	47,20	47,64	48,26	48,78	49,39	50,15	48,92	49,45	50,30

Tab. 2 – Scores BLEU sur les données de développement. NNLM dénomme le modèle de langage neuronal.

Les gains apportés par la désambiguïsation lexicale par catégories syntaxiques sont relativement faibles lorsqu'on considère les systèmes avec un tri- ou quadrigramme à repli. Là encore, une réévaluation avec des modèles n-grammes de catégories syntaxique n'est pas efficace. Cependant, les résultats sont intéressants lorsqu'on combine la modélisation de langage neuronal et la désambiguïsation lexicale : le score Bleu passe de 49,39 à 50,30. Ceci montre bien l'intérêt de travailler conjointement sur une amélioration des techniques statistiques et sur l'incorporation de connaissances lexicales ou syntaxiques. En effet, la réévaluation des n meilleures hypothèses avec un modèle de langage semble être plus efficace si les mots proposés par le modèle de traduction sont mieux choisis.

3.3 Résultats sur les données de test

Les performances sur les données de test de l'évaluation Tc-Star 2007 sont résumées dans le Tableau 3. Les coefficients λ_i des fonctions caractéristiques sont les mêmes que ceux du système optimisé sur les données de développement. Le système n'a donc pas été adapté sur les données de test. Sept centres de recherche publiques et industriels ont participé à l'évaluation qui s'est déroulée en février 2007. Les scores BLEU varient entre 42.95 et 49.60 (espagnol/anglais) et entre 37.39 et 51.04 (anglais/espagnol). Les performances du système avec désambiguïsation lexicale sont très légèrement au-dessous du système de base, dans le cas de l'utilisation d'un modèle de langage à repli. Cependant la combinaison avec un modèle de langage neuronal donne de bons résultats, sans pour autant pouvoir dépasser le système sans désambiguïsation.

	Espagnol → anglais			Anglais → espagnol					
	Sans désambiguïsation			Sans désambiguïsation			Avec désambiguïsation		
	base	4-gram	NNLM	base	4-gram	NNLM	base	4-gram	NNLM
BLEU	48,42	48,67	49,19	49,19	50,17	51,04	49,13	49,91	51,04

TAB. 3 – Scores BLEU sur les données de test.

4 Conclusion

Nous avons présenté et évalué deux évolutions d'un système de traduction statistique. L'une propose une modélisation linguistique dans un espace continu et la seconde intègre les catégories morpho-syntaxiques des mots dans le modèle de traduction. La combinaison des deux méthodes donne des résultats intéressants. Notre système a obtenu de très bons résultats à l'évaluation TC-STAR organisée début 2007.

Nous étudions aussi l'application des mêmes techniques à la traduction automatique d'autres paires de langues, notamment la traduction entre l'anglais et le français. Pour cela le corpus Europarl est utilisé (Koehn, 2006). Nous sommes en train de produire une deuxième référence de traduction qui sera librement disponible pour d'autres laboratoires de recherche intéressés dans la traduction automatique du français[6].

Plusieurs extensions du système décrit dans cet article sont actuellement à l'étude. Nous travaillons sur une meilleure incorporation des connaissances linguistiques, notamment sur l'utilisation d'étiqueteurs prenant en compte le genre et le nombre, voire le sens des mots, afin d'améliorer la désambiguïsation dans le modèle de traduction. Un logiciel de visualisation des erreurs de traduction est en cours de développement afin de permettre une analyse qualitative des erreurs pour affiner le choix des étiquettes, notamment pour le français. En ce qui concerne l'amélioration des techniques statistiques, nous sommes très intéressés par une représentation factorisée des mots, incluant notamment des informations morpho-syntaxiques et linguistiques, aussi bien pour le modèle de traduction que pour le modèle de la langue cible.

Remerciements

Ces recherches ont été partiellement financées par le projet européen TC-STAR et par le projet ANR Instar, JCJC06_143038.

Références

BENGIO Y., DUCHARME R., VINCENT P. & JAUVIN C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research*, **3**(2), 1137–1155.

BERGHEN F. V. & BERSINI H. (2005). CONDOR, a new parallel, constrained extension of powell's UOBYQA algorithm : Experimental results and comparison with the DFO algorithm. *Journal of Computational and Applied Mathematics*, **181**, 157–175.

[6]Données disponibles à partir de la page internet `http://instar.limsi.fr`

Brown P., Della Pietra S., Della Pietra V. J. & Mercer R. (1993). The mathematics of statistical machine translation. *Computational Linguistics*, **19**(2), 263–311.

Charniak E., Knight K. & Yamada K. (2003). Syntax-based language models for machine translation. In *MT Summit*.

Hasan S., Bender O. & Ney H. (2006). Reranking translation hypothesis using structural properties. In *EACL Workshop on Learning Structured Information in Natural Language Applications*.

Hwang Y., Finch A. & Sasaki Y. (2007). Improving statistical machine translation using shallow linguistic knowledge. *Computer Speech & Language*, **21**(2), 350–372.

Kirchhoff K. & Yang M. (2005). Improved language modeling for statistical machine translation. In *ACL'05 workshop on Building and Using Parallel Text*, p. 125–128.

Koehn P. (2006). Europarl : A parallel corpus for statistical machine translation. In *MT Summit*.

Koehn P., Och F. J. & Marcu D. (2003). Statistical phrased-based machine translation. In *Joint Conference on Human Language Technology and of the North American Chapter of the Asociation for Computational Lingustics*, p. 127–133.

Och F.-J., Gildea D., Khudanpur S., Sarkar A., Yamada K., Fraser A., Kumar S., Shen L., Smith D., Eng K., Jain V., Jin Z. & Radev D. (2004). A smorgasbord of features for statistical machine translation. In *Proceedings of the North American Chapter of the Asociation for Computational Lingustics*, p. 161–168.

Och F. J. & Ney H. (2002). Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of ACL*, p. 295–302.

Och F. J., Tillmann C. & Ney H. (1999). Improved alignment models for statistical machine translation. In *Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Copora*, p. 20–28.

Papineni K., Roukos S., Ward T. & Zhu W. (2002). BLEU : a method for automatic evaluation of machine translation. In *Proceedings of ACL*, p. 311–318.

Schmid H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*.

Schwenk H. (2007). Continuous space language models. *Computer Speech and Language*, **21**, 492–518.

Schwenk H., Costa-jussà M. R. & Fonollosa J. A. R. (2006). Continuous space language models for the IWSLT 2006 task. In *International Workshop on Spoken Language Translation*, p. 166–173.

Stolcke A. (2002). SRILM - an extensible language modeling toolkit. In *International Conference on Speech and Language Processing*, p. II : 901–904.