Traduction automatique de la parole arabe/anglais par segmentations multiples

Fethi Bougares

Laboratoire d'Informatique de Grenoble, équipe GETALP
BP 53, 38041 Grenoble Cedex 9, FRANCE
Courriel : fethi.bougares@imag.fr

ABSTRACT

Machine translation seems to have made an impressive progress in the last few years. Today a number of systems available produce translation outputs of sufficient quality to be useful for certain specific applications.
In this work, we are interested in Arabic to English Statistical Machine Translation (SMT). Arabic is a morphologically rich language, and recent experiments in our laboratory have shown that the performance of Arabic to English speech translation systems varies greatly with Arabic morphological analyses applied. First, we try to show the relation between translation quality and Arabic morphological analyses, by constructing our baseline systems using different morphological analyzers and measuring their impact on machine translation quality. Then, we present our method to integrate multi-segmentation in the translation process, which aims to minimize the translation errors, using different alternative segmentations instead of a single segmentation and integrate the segmentation process with the search for the best translation. Hence, the segmentation decision is only made during the generation of the final translation.

1. INTRODUCTION

La traduction de la parole est un problème très intéressant qui met en jeu plusieurs défis scientifiques importants. Ce problème peut se traiter par l'enchaînement de trois composantes développées indépendamment : une composante de reconnaissance de la parole, une composante de traduction et enfin une composante de synthèse.
La tâche considérée dans ce travail est la traduction de sortie de système de reconnaissance de l'arabe fournie par la campagne d'évaluation IWSLT [1] sous forme de treillis d'hypothèses pour chaque phrase prononcée. En effet, notre travail consiste à garantir la compatibilité entre la partie reconnaissance et la partie traduction, à étudier la relation entre la segmentation d'un texte arabe et la qualité de traduction et construire un système à multi-segmentations.
Ce papier est organisé comme suit : la section 2 résume les principes de la traduction automatique probabiliste, la section 3 décrit les systèmes de traduction à mono-segmentation, la section 4 présente l'architecture du nouveau système à multi-segmentations et ses performances. Une discussion de résultats obtenus et de la perspective de ce travail fera l'objet de la dernière section.

1. International Workshop for Spoken Language Translation http ://mastarpj.nict.go.jp/IWSLT2008/

2. TRADUCTION STATISTIQUE

La littérature en traduction automatique fondée sur les approches statistiques est relativement récente. Un des premiers articles concernant cette approche est celui de Brown dans [Bro93]. Depuis, de nombreuses recherches se sont inspirées de cette approche et de ses variantes statistiques.
La traduction automatique probabiliste se base sur le principe d'utilisation des corpus annotés pour apprendre les associations les plus statistiquement significatives entre deux langues. Soient deux ensembles de phrases *S* (phrases sources) et *C* (phrases cibles), pour chaque phrase f dans *S* l'approche consiste à choisir, parmi toutes les traductions possibles, la phrase *e* dans *C* qui en constitue sa traduction la plus probable.

Le problème se décompose de la manière suivante :

$$e^* = argmax_e Pr(e/f) \quad (1)$$

$$e^* = argmax_e Pr(f/e)Pr(e) \quad (2)$$

où Pr(f/e) représente le modèle de traduction et Pr(e) représente le modèle de langue cible.

3. SYSTÈME DE RÉFÉRENCE POUR LA TRADUCTION DE PAROLE ARABE/ANGLAIS

La méthode la plus simple pour construction un système de traduction de parole est de prendre la sortie du système de reconnaissance comme entrée à la traduction. Cependant la sortie du système de reconnaissance n'est pas nécessairement compatible avec l'entrée de notre système de traduction. Ainsi dans ce qui suit nous décrivons les transformations appliquées sur la sortie du système de reconnaissance afin de garantir la compatibilité entre les deux modules reconnaissance-traduction et de construire nos systèmes de référence (baseline).
La sortie de reconnaissance fournie par la campagne IWSLT a la forme d'un treillis des mots pour chaque phrase, alors que le système de traduction est basé sur les morphèmes et accepte comme entrée un réseau de confusion [Ber07]. Les transformations consistent en une étape de décomposition de treillis de mots pour construire un treillis de morphèmes suivi d'une étape de génération du réseau de confusion à partir de ce treillis.
Etant donnée un treillis de mots à transformer nous procédons tout d'abord par l'extraction de son vocabulaire (ensemble des mots), suivi par la segmentation de ce vocabulaire et la génération de nouveau treillis basée sur les

morphèmes [2]. L'algorithme de décomposition de treillis de mots est décrit dans [Bes07]. Pour la génération des réseaux de confusion, nous utilisons la bibliothèque *lattice-tool* de la boite à outils SRILM [3] et pour la recherche de la meilleur traduction le décodeur *moses* [4] , Le système de référence est décrit en détail dans [Bes07] et son architecture est donnée dans la figure 1.

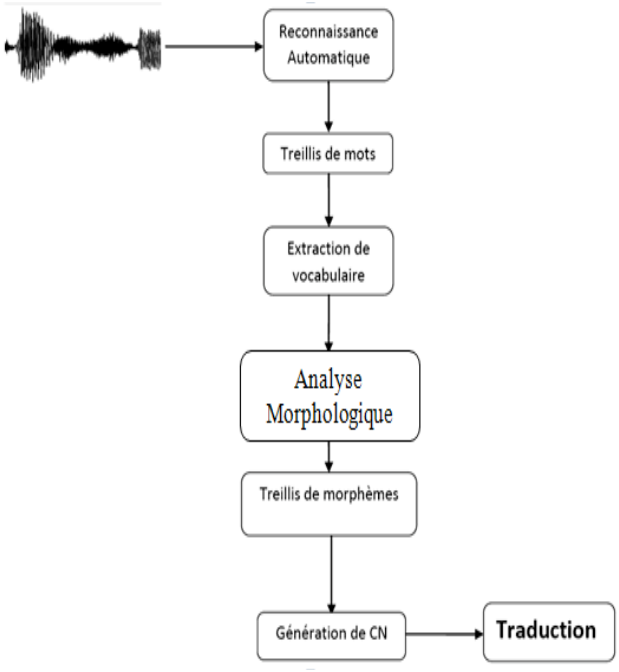

Figure 1: Architecture des systèmes de référence.

Comme référence nous avons construit deux systèmes de référence de traduction de la parole appris sur la même quantité de données, avec les mêmes outils et différenciés par la méthode d'analyse morphologique utilisée.
Les corpus d'entraînement, de développement et de test sont distribués par la campagne IWSLT. L'optimisation de tous les systèmes est faite avec l'algorithme MERT1 sur le corpus de développement dev06 qui contient 489 phrases.

3.1. Système ASVM

Le premier système de référence est construit en utilisant l'analyseur morphologique ASVM [Dia04] appris sur un corpus segmenté manuellement. Les performances de ce système en terme de score BLEU [Pap01] multi-références sont données dans le tableau suivant :

Table 1: Performances du système de traduction de parole fondé sur la segmentation ASVM.

	dev06	tst06	tst07
ASVM	28.27	22.78	41.04

2. la plus petite unité porteuse de sens qu'il soit possible d'isoler dans un enoncé
3. http ://www.speech.sri.com/projects/srilm/
4. http ://www.statmt.org/moses

3.2. Système Buckwalter

Pour la construction du deuxième système de référence nous avons utilisé l'analyseur morphologique à base de règles de T. Buckwalter [Buc02] et les performances obtenues sont données dans le tableau suivant :

Table 2: Performances du système de traduction de parole fondé sur la segmentation Buckwalter.

	dev06	tst06	tst07
Buckwalter	27.69	24.19	42.54

3.3. Segmentions vs qualité de traduction

Appris sur les mêmes données et avec les mêmes outils les deux systèmes de référence donnent une qualité de traduction variable, le système *asvm* est significativement meilleur sur les corpus Dev06 et tst07 alors que le système *Buckwalter* traduit mieux le corpus Tst07. Pour montrer la différence de qualité nous avons réalisé une analyse plus profonde en regardant les sorties de deux systèmes et nous constatons que la qualité de traduction fournie par les systèmes probabilistes pour le couple de langue arabe-anglais est fortement sensible à la segmentation de texte arabe comme le montre le tableau 3 où les traductions correctes sont misent en gras.

Table 3: Exemple de performance de systèmes selon la segmentation

Systeme ASVM	Systeme Buckwalter
خمسة دولارات ل الواحد **five dollars each**	خمس ة دولار ات ل ال واحد five dollars for one
نعم دولارين ل ال نسخة ال واحدة yes two dollars for a one	نعم دولار ين ل ال نسخ ة ال واحد ة **yes two dollars for one copy**
حوالي عشرة كيلومتر about a ten	حوالي عشرة كيلومتر **about ten kilometers**
ما هذه ال اغنية **what's this song**	ما هذه ال اغني ة what's this sing a song

4. SYSTÈME À MULTI-SEGMENTATIONS

La segmentation de l'arabe présente plusieurs problèmes d'ambiguïté ; en plus le choix à priori d'une méthode de segmentation peut pénaliser les performances de système de traduction. L'idée est alors de créer un système à multi-segmentations où l'ambiguïté segmentale de texte arabe est gardée à l'entrée du système et le choix de segmentation est guidé par les modèles de traductions lors du processus de décodage. Dans cette section nous présentons quelques problèmes liés à la segmentation de texte arabe suivi de la formalisation de la multi-segmentation et les résultats obtenus.

4.1. Problèmes de segmentation de l'arabe

La structuration des mots constitue une source d'information intéressante pour des langues morphologiquement riches telles que l'arabe. Par conséquent, la segmentation représente une étape fondamentale dans le traitement automatique d'un texte, son rôle est de découper un texte en unités d'un certain type que l'on aura définies et repérées préalablement. La segmentation d'un texte informatisé est donc l'opération de délimitation des segments de ses éléments de base qui sont les caractères, en éléments constituants de différents niveaux structurels : paragraphe, phrase, syntagme [5], mot graphique [6], morphème, etc. [Mou08]. L'opération d'analyse morphologique automatique pour la langue arabe est fondée sur la notion de mot graphique. De ce fait, l'analyse morphologique d'un texte arabe revient à identifier les constituants du mot en le décomposant en proclitiques, préfixes, base, suffixes et enclitiques et à associer à chaque constituant sa ou ses catégories grammaticales. La représentation suivante schématise une structure possible d'un mot arabe. Le problème

Proclitique	Préfixe	Base	Sufffixe	Enclitique

Figure 2: schéma d'un mot arabe.

de segmentation pour l'arabe réside de la richesse grammaticale de cette langue. En effet, un mot peut changer de sens selon sa segmentation. Par exemple, le tableau suivant montre les trois segmentations possibles du même mot arabe et l'interprétation de chaque segmentation.

Table 4: Exemple de variation de sens d'un mot arabe selon la segmentation.

Segmentation possible	Traduction en français
أ + لم + هم	les a-t-il ramassées
ألم + هم	leur douleurs
أل + مهم	l'important

4.2. Formalisation de la multi-segmentation

La formalisation de l'intégration des multi-segmentations consiste à l'ajout d'un modèle de segmentation dans l'équation (1) comme suit : Soient P la phrase arabe composée de mots $f_1 \dots f_j \dots f_J$ et une suite de morphèmes qui représente une possibilité de segmentation de P, la fonction de recherche de la meilleure traduction est réécrite comme :

$$e^* = argmax_e Pr(e/f) \quad (3)$$

$$e^* = argmax_e \sum_{a_1^k} Pr(e, a_1^k/f) \quad (4)$$

$$e^* = argmax_e \sum_{a_1^k} Pr(a_1^k/f).Pr(e/f, a_1^k) \quad (5)$$

$$e^* = argmax_e \{max_{a_1^k} Pr(a_1^k/f).Pr(e/a_1^k)\} \quad (6)$$

Le modèle de traduction dans l'équation (3) est basé sur les morphèmes, on approxime donc que la phrase cible ne dépend que de la phrase source au niveau morphèmes a_1^k et on remplace aussi la somme sur a_1^k par la maximisation. La nouvelle équation à résoudre par le décodeur pour la recherche de la traduction la plus probable est :

$$e^* = argmax_e \{argmax_{a_1^k} Pr(a_1^k/f) Pr(a_1^k/e) Pr(e)\} \quad (7)$$

cette équation fait intervenir trois modèles :

- $Pr(e)$: modèle de langue cible.
- $Pr(a_1^k/f)$: modèle de segmentation.
- $Pr(a_1^k/e)$: modèle de traduction pour une segmentation donnée.

4.3. Architecture du nouveau système proposé

Le système à multi-segmentation à construire contient deux modèles de traduction (un par segmentation), ces deux modèles étant utilisés lors de la recherche de la meilleure traduction. Pour intégrer la multi-segmentation et puisque le décodeur utilisé (moses) décode les réseaux de confusion, nous construisons un réseau de confusion par phrase en alignant les deux segmentations possibles de chaque phrase à traduire et en donnant la même chance pour tous les chemins dans le réseau (score équiprobable).

La figure 3 montre un exemple de réseau de confusion pour deux segmentations différentes de la même phrase arabe qui se traduit en français par « j'ai besoin d'elle aujourd'hui. »

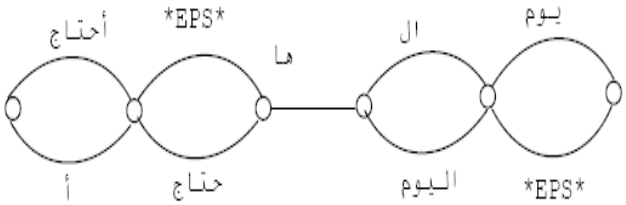

Figure 3: Exemple de réseaux de confusion de deux segmentations de la même phrase arabe.

5. Unité syntaxique plus ou moins complexe située entre la limite supérieure de la syntaxe, constituée par la phrase, et la limite inférieure, constituée par la catégorie simple (unité de base indissociable).

6. Une suite de constituants immédiats ; préfixes, base, suffixes et enclitiques.

L'architecture du nouveau système est la suivante :

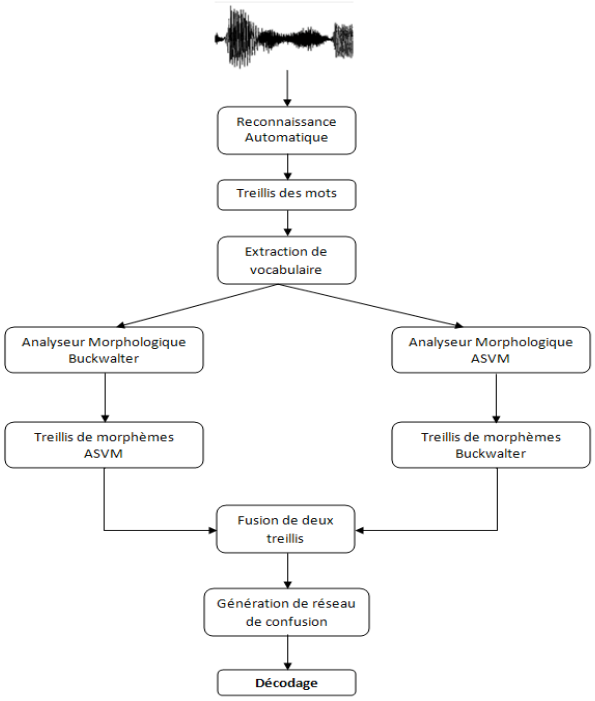

Figure 4: Architecture du système de traduction de parole à multi-segmentation

4.4. Performances des systèmes

La figure 5 montre les performances du nouveau système (score BLEU) par rapport aux systèmes de référence, pour chaque corpus le meilleur score est mis en gras.

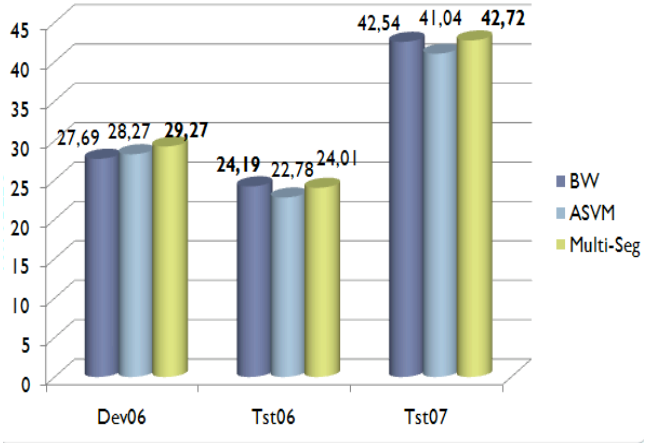

Figure 5: Performances de système à multi-segmentation par rapport aux systèmes de références.

5. DISCUSSION

Avec le nouveau système, on arrive à améliorer les résultats sur le corpus de développement (dev06) et le corpus de test (tst07) tout en gardant un score équivalent au meilleur système sur le corpus de test (tst06). L'amélioration obtenue sur le corpus dev06 est due au fait que les paramètres de notre système ont été optimisés sur ce corpus. Le corpus tst06 contient de nombreuses phrases longues. De plus lorsqu'il s'agit de traduire la sortie d'un système de reconnaissance appliqué sur ce corpus les phrases deviennent plus longues à cause des différentes hypothèses proposées, ce qui fait que le système n'arrive pas à choisir les meilleurs chemins dans les réseaux de confusion en entrées. Contrairement à tst06, les phrases du corpus tst07 sont courtes et la multi-segmentation permet au décodeur d'avoir plus de choix pour produire la meilleure traduction en se basant sur les tables de traduction. Indépendamment de longueur des phrases, l'utilisation de réseaux de confusion introduit des nouveaux chemins et donc des nouvelles segmentations incorrectes.
Le tableau 5 montre un exemple où le nouveau système choisi le bon chemin dans le réseau de confusion et fourni une traduction correcte, et inversement.

Table 5: Exemple de variation de sens d'un mot arabe selon la segmentation.

Phrase arabe	اود استئجار هذه السيارة لمدة اسبوع تقريبا
Buckwalter	I'd like this car for about a week
Asvm	to rent this car for long old almost
Multi-segmentation	**I'd like to rent this car for about a week**
Phrase arabe	ايمكنني استعمال هاتفك
Buckwalter	**can i use your phone**
Asvm	**can i use your phone**
Multi-segmentation	i use your phone

6. CONCLUSION ET PERSPECTIVES

Ce travail a porté sur la paire de langues arabe/anglais. Nous nous sommes intéressés à l'étude de la relation entre la qualité de sortie d'un système de traduction et la méthode de segmentation d'un texte arabe. Nous avons montré que segmentation et qualité de traduction sont fortement corrélées, ce qui nous a conduit vers l'idée de construire un système à segmentations multiples où on ne prend aucune décision à priori sur la façon de segmenter le texte arabe ; la décision est laissée au processus de décodage lui-même. L'intégration de la multi-segmentation a été faite en utilisant des réseaux de confusion (graphes simples), qui malgré leurs capacités à représenter l'ambiguïté segmentale d'un texte arabe, ont le défaut d'introduire du bruit sous la forme de nouveaux chemins dans le graphe, ne correspondant à aucune segmentation valide. Malgré ce défaut, nos expérimentations et notre évaluation montrent l'efficacité de notre approche appliquée sur la traduction de textes ou sur la traduction de sorties d'un système de reconnaissance vocale. Comme continuation de ce travail nous proposons de travailler plus sur les poids des réseaux de confusion pour donner « plus de chance » aux chemins (c'est-à-dire aux segmentations) les plus pertinents. On propose aussi de rajouter d'autres informations à notre modèle de segmentation en ajoutant des étiquettes morphosyntaxique de mots sur les arcs des réseaux de confusion, on se retrouve alors dans un paradigme de modèles factoriels.

RÉFÉRENCES

[Ber07] Bertoldi, N., Zens, R. and Frederico (April 2007), "Speech translation by confusion network decoding", in *ICASSP*, vol. 4, pp. 1297–1300.

[Bes07] Besacier, L. and Mahdhaoui, A. (2007), "The lig arabic/english speech translation system at iwslt07", in *IWSLT*.

[Bro93] Brown, P., Pietra, S.D., Pietra, V.D. and Mercer, R. (1993), "The mathematics of machine translation : Parameter estimation.", in *Computational Linguistics,*, pp. 263–312.

[Buc02] Buckwalter (2002), "Buckwalter arabic morphological analyser version 1.0", .

[Dia04] Diab, M., Hacioglu, K. and Jurafsky, D. (2004), "Automatic tagging of arabic text : From raw text to base phrase chunks", in *5th Meeting of the North American Chapter of the Association for Computational Linguistics/Human Language Technologies Conference (HLT-NAACL04), Boston, MA. USA*.

[Mou08] Moulehi, Z. (2008), "Araseg un segmenteur semi-automatique des textes arabes", in *9 ieme Journée Internationale d'Analyse Statistique des données Textuelles*.

[Pap01] Papineni, K., Roukos, S., Ward, T. and Zhu, W.J. (2001), "Bleu : a method for automatic evaluation of machine translation", in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics ACL*, pp. 311–318.