TALN 2010, Montréal, 19–23 juillet 2010

Adaptation d'un Système de Traduction Automatique Statistique avec des Ressources monolingues

Holger Schwenk
LIUM, Université du Maine,
72085 Le Mans cedex, France
Holger.Schwenk@lium.univ-lemans.fr

Résumé. Les performances d'un système de traduction statistique dépendent beaucoup de la qualité et de la quantité des données d'apprentissage disponibles. La plupart des textes parallèles librement disponibles proviennent d'organisations internationales. Le jargon observé dans ces textes n'est pas très adapté pour construire un système de traduction pour d'autres domaines. Nous présentons dans cet article une technique pour adapter le modèle de traduction à un domaine différent en utilisant des textes dans la langue source uniquement. Nous obtenons des améliorations significatives du score BLEU dans des systèmes de traduction de l'arabe vers le français et vers l'anglais.

Abstract. The performance of a statistical machine translation system depends a lot on the quality and quantity of the available training data. Most of the existing, easily available parallel texts come from international organizations and the jargon observed in those texts is not very appropriate to build a machine translation system for other domains. In this paper, we present a technique to automatically adapt the translation model to a new domain using monolingual data in the source language only. We observe significant improvements in the BLEU score in statistical machine translation systems from Arabic to French and English respectively.

Mots-clés : Traduction statistique, adaptation du modèle de traduction, corpus monolingue, apprentissage non-supervisé.

Keywords: Statistical machine translation, translation model adaptation, monolingual data, unsupervised training.

1 Introduction

La traduction automatique statistique est aujourd'hui considérée comme une alternative sérieuse aux systèmes de traductions à base de règles. Ces derniers effectuent d'abord une analyse de la phrase source, puis une étape de transfert et ensuite la génération de la phrase dans la langue cible. Le développement et le maintien d'un tel système nécessite généralement un travail humain important par des spécialistes (bilingues). Un système statistique, en revanche, peut en principe être développé sans connaissance des langues traitées. Considérons la traduction d'une phrase en français f vers l'anglais e :

$$e^* = \arg\max_e p(e|f) = \arg\max_e p(f|e)P(e) \quad (1)$$

Le modèle de traduction $p(f|e)$ est appris à partir d'exemples de traductions, c'est-à-dire des textes en langue source et les traductions correspondantes, alignés au niveau de la phrase. Ces textes sont communément appelés « textes parallèles » ou « bitextes » . Le modèle de langue $P(e)$ est construit à partir de textes dans la langue cible. Cet apprentissage automatique à partir d'exemples est généralement avancé comme un grand avantage des systèmes de traduction statistique. Ceci a notamment permis de construire rapidement des systèmes de traduction pour toutes les combinaisons de 22 langues européennes (Koehn *et al.*, 2009), grâce à l'utilisation des textes traduits par la Commission Européenne. Ce corpus parallèle est connu sous le nom d'Europarl.

En même temps, il est clair que les performances de toute approche d'apprentissage automatique dépendent largement de la quantité et de la qualité des données d'apprentissage disponibles. On constate souvent que les performances s'améliorent lorsque l'on utilise davantage de données d'apprentissage bien que cet effet s'accentue rapidement. D'autre part, il s'avère souvent que le domaine des données d'apprentissage correspond peu ou pas au domaine d'utilisation prévu du système de traduction. Pour citer un exemple, on comprend aisément que les traductions apprises automatiquement à partir des exemples de traductions dans le domaine de la finance conviennent mal pour traduire des textes médicaux. En effet, les vocabulaires risquent d'être différents et il y a des mots qui se traduisent différemment en fonction du domaine.

Malheureusement, il se trouve que pratiquement tous les bitextes librement disponibles proviennent du domaine parlementaire ou politique : le corpus Europarl, les comptes rendus en français et en anglais du parlement canadien (« Hansard ») ou des textes des Nations Unies. Le modèle de traduction appris sur ces données risque donc de favoriser des traductions spécifiques de ce domaine. On constate aussi que la première personne n'est pas fréquemment utilisée dans ces textes. Cependant, on peut supposer qu'il y ait suffisamment de *textes monolingues* pour une grande variété de langues et domaines. Ces textes peuvent souvent être trouvés sur Internet ou sont disponibles auprès de l'utilisateur du système de traduction. Il est donc nettement plus facile de construire un modèle de langue spécifique à un domaine.

Dans ce travail, nous proposons une méthode qui permet d'adapter un modèle de traduction générique à un domaine particulier en utilisant des données monolingues dans la langue source. Cet article est organisé comme suit. Dans la section suivante, nous résumons d'abord d'autres recherches qui abordent le problème de ressources insuffisantes. La section 3 présente les systèmes de traduction de référence et la section 4 résume nos expériences. L'article termine avec une conclusion et une discussion de futures directions de recherche.

2 Recherches précédentes

Plusieurs techniques ont été proposées dans la littérature pour aborder le problème de ressources bilingues insuffisantes. On pourrait notamment essayer d'extraire des textes parallèles à partir de corpora comparables. Un corpus comparable bilingue peut être défini comme une collection de textes dans deux langues qui traitent le même sujet sans être des traductions parfaites. Wikipedia constitue un exemple bien connu d'un grand corpus comparable.

Une autre piste consiste à adapter le modèle de traduction à la tâche sans utiliser des ressources bilingues supplémentaires. On peut distinguer deux façons d'effectuer cette adaptation : premièrement, on ajoute de nouveaux mots en langue source ou de nouvelles traductions ; et deuxièmement, on modifie les distribu-

tions de probabilité du modèle existant pour qu'elles conviennent mieux au domaine. Ces deux directions sont complémentaires et peuvent être effectuées simultanément.

Une technique classique pour adapter un modèle statistique consiste à utiliser un mélange de plusieurs modèles et à optimiser les coefficients d'interpolation à la tâche. Ceci a été étudié par plusieurs auteurs dans le cadre de la traduction statistique, par exemple pour l'alignement des mots (Civera & Juan, 2007), pour la modélisation linguistique (Zhao *et al.*, 2004; Koehn & Schroeder, 2007), et pour le modèle de traduction (Foster & Kuhn, 2007; Chen *et al.*, 2008). L'avantage de cette approche consiste dans le fait que peu de paramètres sont modifiés, *i.e.* les coefficients des mélanges. Cependant, beaucoup de probabilités sont modifiées en même temps et il n'est pas possible de modifier sélectivement la probabilité d'une traduction particulière.

L'extraction de textes alignés à partir de corpora comparables se fait souvent avec des techniques de recherche d'information, voir par exemple (Hildebrand *et al.*, 2005). Récemment, une technique similaire a été mise en œuvre pour adapter le modèle de traduction et de langage avec des textes monolingues dans la *langue source* (Snover *et al.*, 2008). Les auteurs ont utilisé une recherche d'information interlingue pour trouver des textes dans la langue cible qui correspondent au domaine des textes dans la langue source. Cependant, il est difficile de trouver les alignements entre les phrases en langue source et cible, et un simple modèle de type IBM-1 a été utilisé.

Une autre direction de recherche consiste dans l'auto-amélioration du modèle de traduction. Ceci a été proposé la première fois par (Ueffing, 2006). L'idée consiste à traduire les données de test, à filtrer les traductions avec une mesure de confiance et à utiliser les meilleures traductions pour entraîner un nouveau (petit) modèle de traduction qui est utilisé conjointement avec la table de traduction générique. Ceci est en fait une approche par mélange de modèles dont un modèle est construit explicitement pour chaque jeu de test. En pratique, ceci est uniquement possible lorsqu'une certaine quantité de données est disponible pour être traduite en une seule fois. Ceci est typiquement le cas lors des évaluations du style NIST ou WMT avec des jeux de test d'environ 50.000 mots, mais l'utilisation de cette méthode semble être plus difficile dans le cadre d'un service de traduction sur Internet. Dans une telle application, on ne demande habituellement que la traduction de quelques phrases. Cette approche a été améliorée par la suite (Ueffing, 2007) et appliquée aux autres modèles statistiques dans un système de traduction (Chen *et al.*, 2008).

Une autre approche comparable est l'apprentissage légèrement supervisé (Schwenk, 2008). Dans ce travail, un système de traduction statistique est utilisé pour traduire de grandes quantités de données en langue source. Ces traductions sont ensuite filtrées et les meilleures sont ajoutées aux bitextes existants. Cette technique semble être très similaire à l'auto-amélioration telle que proposée par (Ueffing, 2006), mais il y a plusieurs différences conceptuelles. Premièrement, nous n'utilisons à aucun moment le jeu de test pour adapter le modèle de traduction, mais un grand corpus monolingue. Deuxièmement, nous créons un tout nouveau modèle qui peut être appliqué sur tout corpus de test sans modification supplémentaire. Ainsi, il est possible d'utiliser un système adapté de cette façon dans un service de traduction sur Internet.

Dans cet article, nous étudions l'utilité de cette approche pour adapter des systèmes de traduction de l'arabe vers l'anglais et vers le français. La traduction de l'arabe est intéressante puisqu'il s'agit d'une langue morphologiquement riche. Ainsi, le texte en arabe est habituellement décomposé pour séparer les affixes et les suffixes d'un mot ce qui permet de diminuer considérablement la taille du vocabulaire de traduction. Plusieurs auteurs signalent une amélioration de la qualité des traductions grâce à cette décomposition morphologique, par exemple (Habash & Sadat, 2006). Elle donne également beaucoup de groupes de mots peu fréquents, ce qui peut entraîner une mauvaise estimation des probabilités de

traduction par fréquence relative. Notre but est d'améliorer l'estimation de ces probabilités par l'utilisation de textes monolingues.

3 Systèmes de traduction de référence

Dans cet article, un système de traduction statistique basé sur les segments est utilisé (en anglais « *phrase-based statistical machine translation system* ») pour les deux paires de langues, en utilisant le logiciel libre Moses (Koehn *et al.*, 2007). L'équation 1 peut être réécrite afin de faire apparaître des fonctions caractéristiques $f_i(e, f)$:

$$e^* = \arg\max_e p(f|e)P(e) = \arg\max_e \prod_i f_i(e,f)^{\lambda_i} = \arg\max_e \sum_i \lambda_i \log f_i(e,f) \quad (2)$$

Nous utilisons quatorze fonctions caractéristiques : les probabilités de traduction et lexicales dans les deux directions, sept fonctions pour le modèle de distorsion lexicalisé, une pénalité sur les mots et les groupes de mots, et une fonction pour le modèle de langue. Les systèmes sont construits de la façon suivante : d'abord le logiciel GIZA++ est utilisé afin d'obtenir les alignements mot à mot dans les deux directions. Il existe une version qui permet d'accélérer le calcul sur des machines multi-cœurs (Gao & Vogel, 2008).[1] Ensuite les groupes de mots et les réordonnements sont extraits, avec les valeurs de défaut de l'outil Moses. Finalement, les coefficients des fonctions caractéristiques sont optimisés par l'outil CMERT.

Les modèles de langage sont des quadri-grammes à repli, construits avec l'outil SRILM (Stolcke, 2002). Les données d'entraînement correspondent au côté anglais des bitextes plus une importante collection de textes de journaux. Ces textes sont disponibles auprès du LDC sous le nom *corpus Gigaword*.

Dans la plupart des études des outils tels que l'analyseur de Buckwalter et les outils MADA et TOKAN de l'université de Columbia sont utilisés pour effectuer la décomposition morphologique des textes en arabe (Habash & Sadat, 2006). Dans le présent travail, nous utilisons le module d'analyse du système de traduction de l'entreprise SYSTRAN pour effectuer ce travail. Des règles de décomposition sont d'abord appliquées, assistées par un dictionnaire. La décomposition la plus probable des mots absents du dictionnaire est effectuée. De façon générale, toutes les décompositions possibles sont envisagées et puis filtrées en utilisant le contexte dans la phrase. Cette étape se base sur une analyse globale de la phrase ainsi que des connaissances lexicales. Les textes français ont été tokénisés avec les outils de Moses. La casse et les ponctuations sont préservées.

3.1 Traduction arabe/anglais

Le *National Institute of Standards and Technology (NIST)* organise depuis quelques années des campagnes d'évaluations internationales des systèmes de traduction automatique. Ces évaluations sont communément considérées comme la référence dans le domaine. Le système de traduction arabe/anglais décrit ici fait partie des meilleurs systèmes de l'évaluation organisée en 2009. Les conditions et résultats détaillés sont disponibles sur le site Internet de NIST.[2]

[1] Les sources sont disponibles à `http://www.cs.cmu.edu/~qing/`

[2] `http://www.itl.nist.gov/iad/mig/tests/mt/2009/`

Le modèle de traduction est appris sur divers textes parallèles disponibles auprès du LDC dans le cadre de l'évaluation NIST pour un total d'environ 56 millions de mots arabes. Nous avons également ajouté 133M de mots du corpus des Nations Unies. Le modèle de langue est appris sur un total de plus de 3 milliards de mots. Ces ressources étaient les mêmes pour tous les participants à cette évaluation (« condition contrainte »).

L'optimisation des paramètres a été effectuée sur les données d'évaluation de 2006. Nous donnons également des résultats sur les données d'évaluation de 2008 qui ont été utilisées comme jeu de test interne. Dans les deux cas, il s'agit de données du domaine des actualités radio et télévisées et des discussions sur Internet. Quatre références de traduction sont disponibles et la casse et la ponctuation sont préservées.

Bitextes	Taille des bitextes	Dev Nist06	Test Nist08
News + ISI bitextes	56M	42,69	42,06
+ données ONU	189M	43,51	42,19

TAB. 1 – Scores BLEU du système de référence arabe/anglais.

Les scores BLEU de ces systèmes de référence sont donnés dans le tableau 1. On note que les données de l'ONU apportent un faible gain du score BLEU malgré une taille considérable. Ceci s'explique par le fait qu'il s'agit de données hors domaine. Bien que ce bitexte apporte beaucoup de traductions, il entraîne également une modification des probabilités de traduction calculées par fréquence relative. Les traductions du domaine « ONU » semblent donc dominer les traductions plus adaptées des bitextes du domaine. Nous montrerons dans cet article que ceci peut être « corrigé » en adaptant le modèle de traduction avec des données monolingues.

3.2 Traduction arabe/français

Nous considérons également la traduction de l'arabe vers le français. Cette paire de langues nous semble intéressante pour plusieurs raisons. Premièrement, il s'agit de deux langues morphologiquement riches, par rapport à la traduction habituelle vers l'anglais. Deuxièmement, il y a peu de bitextes bien adaptés au domaine de traduction et un grand corpus hors domaine. Ce sont exactement les conditions qui ont motivé notre approche d'adaptation du modèle de traduction. Finalement, on peut facilement identifier des applications d'un système de traduction de l'arabe vers le français.

Le développement des premiers systèmes de traduction statistiques pour cette paire de langues a probablement débuté avec le projet DGA TRAMES[3] dont le but était la traduction de la parole arabe vers le français. Dans le cadre de ce projet, environ 90 heures de discours radio et télévisés ont été enregistrés, transcrits et ensuite traduits en français. La DGA nous a donné accès à ces textes parallèles d'environ 260 mille mots. Ces données sont parfaitement adaptées au domaine mais sont bien sûr de taille trop limitée pour entraîner un modèle de traduction statistique performant. Ainsi, nous les avons complétés par 1,1 million mots de textes téléchargés du site Internet du projet *Syndicate*[4] et par environ 200 millions de mots de données de l'ONU. Ce dernier corpus a été collecté par l'entreprise SYSTRAN.

[3]Traduction Automatique par Méthodes Statistiques

[4]`http://www.project-syndicate.org`

La DGA a également produit un jeu de test avec 4 traductions de référence. Ce corpus a été aléatoirement divisé en jeu de développement et test de 10 mille mots chacun environ. Les performances des systèmes de référence sont données dans le tableau 2. Le modèle de langue est un quadri-grammes entraîné sur un peu plus de 1,3 milliard de mots (côté français des bitextes, corpus Gigaword français et d'autres journaux).

Bitexts	#mots	Dev	Test
TRAMES + Syndicate	858k	36,68	35,45
ONU	203M	40,02	37,91
TRAMES + Syndicate + ONU	204M	41,88	40,04

TAB. 2 – Scores BLEU du système de référence arabe/français.

A notre connaissance, un seul autre système de traduction statistique arabe/français a été développé, précisément dans le cadre du projet TRAMES (Hasan & Ney, 2008). Dans ce travail, le même jeu de test a été utilisé, mais les bitextes sont différents : les textes parallèles du projet TRAMES, des données de l'ONU de la période 2001 à avril 2007, les archives de Amnesty International et des articles du Monde Diplomatique. Les auteurs donnent un score BLEU de 41,1 sur le jeu de test complet d'environ vingt milles mots. Ce système utilise donc d'autres ressources que le nôtre et il n'est pas possible de comparer directement les performances. Cependant, on peut probablement conclure que des scores BLEU supérieurs à 40 points semblent correspondre à l'état de l'art pour cette paire de langues. Ceci correspond également aux résultats observés dans les évaluations NIST pour la paire de langues arabe/anglais (cf. tableau 1). Dans les deux cas il s'agit de la traduction de textes radio et télévisés et quatre références de traductions sont disponibles.

4 Adaptation du modèle de traduction

Le but de ce travail est l'adaptation du modèle de traduction sans bitextes supplémentaires, mais avec des données monolingues dans la langue source. Habituellement, il est bien plus facile de trouver de tels textes, en particulier lorsqu'il s'agit de textes du domaine des actualités comme dans ce travail. Nous utilisons ici des parties du corpus Gigaword en arabe du LDC.

Ces textes sont traduits par les systèmes de référence décrits ci-dessus. Ensuite, les traductions automatiques sont filtrées afin de ne garder que les « meilleures » . Cette sélection pourrait être basée sur des scores de confiance au niveau des mots (Ueffing, 2007). Dans notre cas, nous avons utilisé le logarithme de la vraisemblance fourni par le décodeur, normalisé par le nombre de mots dans chaque phrase. Les traductions filtrées sont ajoutées aux bitextes existants et la procédure complète de construction d'un système de traduction statistique est effectuée, c'est-à-dire l'alignement des mots par GIZA++, l'extraction des groupes de mots et l'optimisation des coefficients λ_i. Alternativement, on pourrait réutiliser les alignements déterminés par le décodeur Moses. Ceci pourrait accélérer le processus puisque nous omettons l'étape effectuée par GIZA++.

Les caractéristiques des corpora Gigaword de LDC sont données dans le tableau 3. Le système arabe/ français n'a été adapté que sur le corpus AFP alors que nous avons utilisé les corpora AFP, XIN et NHR pour le système arabe/anglais. Notons que les textes de LDC en anglais et français sont utilisés lors de la construction du modèle de langue $P(e)$. On peut supposer que ces textes contiennent les traductions de quelques phrases des textes en arabe, ce qui devrait aider à produire de bonnes traductions automatiques. Ainsi nous parlons d'un apprentissage légèrement supervisé par le modèle de langue (Schwenk, 2008).

source	arabe	anglais	français
AFP	145M	527M	570M
APW	-	1011M	200M
ASB	7M	-	-
HYT	175M	-	-
NHR	188M	-	-
UMH	1M	-	-
XIN	58M	280M	-

TAB. 3 – Caractéristiques des corpora Gigaword de LDC (nombre de mots).

4.1 Adaptation du système arabe/anglais

Les scores BLEU après adaptation du système arabe/anglais aux textes en arabe de l'AFP, XIN et HYT respectivement sont donnés dans le tableau 4. Bien que les scores BLEU sur les données de développement ne changent que peu, on constate une nette amélioration des performances sur les données de test. On note aussi que les systèmes adaptés utilisent moins de bitextes que le système de référence. Ceci s'explique par le fait que les données de l'ONU ne sont plus utilisées dans les systèmes adaptés puisque ces données hors-domaine sont remplacées par les traductions automatiques des corpora arabes du domaine. Pour chaque corpus, nous avons essayé différents seuils sur la vraisemblance normalisée du décodeur. Le choix du meilleur seuil était bien sûr basé uniquement sur les performances obtenues sur les données de développement.

Adaptation	Taille des bitextes	Dev Nist06	Test Nist08
Aucune	189M	43,51	42,19
AFP	81M	43,64	43,10
XIN	48M	43,36	43,06
HYT	49M	43,77	43,00
Combinaison	-	43,98	43,28

TAB. 4 – Adaptation du système arabe/anglais.

Finalement, nous avons effectué une simple combinaison des trois systèmes adaptés indépendamment : les listes des n meilleures hypothèses sont concaténées et la meilleure hypothèse est extraite. Ceci a permis d'obtenir un faible gain supplémentaire (dernière ligne du tableau 4). Ce système a été très bien placé lors des évaluations NIST de 2009.[5] Le système officiel inclut une autre composante que nous avons omise ici par manque de place (la modélisation linguistique dans l'espace continu (Schwenk, 2010)).

4.2 Adaptation du système arabe/français

Les performances du système arabe/français adapté sont résumées dans le tableau 5. Ici, nous constatons un gain en score BLEU très appréciable de plus de 3,5 points BLEU sur les données de test. Cette amélioration importante pourrait s'expliquer par le faible nombre de bitextes du domaine par rapport aux données

[5] http://www.itl.nist.gov/iad/mig/tests/mt/2009/

	#mots arabe	Dev	Test
Référence	217M	41,88	40,04
Adapté	48M	45,44	43,68

TAB. 5 – Adaptation du système de traductions arabe/français.

de l'ONU, très volumineuses mais hors domaine. Ce rapport était plus équilibré pour la paire de langues arabe/anglais.

Nous avons analysé la table de traduction pour ce système adapté et le système de référence qui a été entraîné sur plus de 200M de mots. Ceci est résumé dans le tableau 6. La table de traduction initiale avait 329M de lignes dont 22,9M pouvaient être potentiellement appliquées aux données de test. La table de traduction du système adapté, d'autre part, n'utilise que 700k d'un total de 8,6M d'entrées. Il paraît clair que la table de traduction obtenue en entraînant sur les données de l'ONU contienne beaucoup d'entrées qui ne sont pas utilisées, voire même fausses. Il est surprenant de voir que la table de traduction du système adapté soit plus petite et qu'elle contienne 11% de segments de la langue source en plus (18029 par rapport à 16263). Toutes ces entrées correspondent aux nouvelles *séquences de mots* puisque l'apprentissage non-supervisé ne permet pas d'augmenter le vocabulaire des mots source.

	Référence	Adapté
Nombre d'entrées	22,9M	700k
Nombre d'entrées différentes côté source	16263	18029
Nombre moyen de traductions	1406,4	38,8
Longueur moyenne d'une entrée côté source	2,65	2,81

TAB. 6 – Caractéristiques de la table de traduction des deux systèmes. Dans les deux cas, la table a été filtrée pour ne contenir que les groupes de mots qui peuvent être appliqués aux données de test.

Nous supposons que ceci est particulièrement important avec la décomposition morphologique de l'arabe. Cette décomposition permet de réduire considérablement le vocabulaire, mais produit également beaucoup de séquences de tokens. Il semble être important d'inclure dans la table de traduction des séquences qui apparaissent dans les textes du domaine. Comme effet de bord, la plus petite table de traduction entraîne un gain de vitesse de la traduction d'environ 40%.

Nous avons comparé les traductions du système avant et après adaptation : le TER[6] est à environ 30. Les deux traductions diffèrent donc significativement. Quelques exemples de traductions sont reproduits dans la figure 1. Le système adapté produit manifestement de meilleures traductions pour ces exemples. Il reste bien sûr quelques erreurs dans ces phrases, mais la qualité des traductions permet largement de comprendre le sens des phrases.

[6]Translation Edit Rate (Snover *et al.*, 2006)

Source:	اَلمَحكَمَةُ العِرَاقِيَّة بَدَأَت مُنذُ قَلِيل بِتَوجِيهِ لَائِحَةِ تُهَم ضِدَّ اَلرَئِيس اَلعِرَاقِيِّ السَّابِق.
Base:	le tribunal irakien a commencé depuis peu par la direction du règlement des accusations contre l'ancien président irakien.
Adapt:	le tribunal irakien a commencé depuis peu une liste d'accusations contre l'ancien président irakien.
Ref:	La Cour irakienne a commencé à dresser la liste des inculpations de l'ancien président irakien.
Source:	أَفَادَت مَصَادِرُ عَسكَرِيَّةٌ إِسرَائِيلِيَّة أَنَّ الجَيشَ الإِسرَائِيلِيّ اِعتَقَلَ لَيلاً نَاشِطاً فِي رَام اَللَّه فِي الضِّفَّةِ الغَربِيَّة كَمَا تَمَّ اعتِقَالُ نَاشِطَينِ آخَرَين كَانُوا يُحَضِّرُونَ
Base:	De source militaire israélienne a indiqué que l'armée israélienne a arrêté dans la nuit militants à Ramallah en Cisjordanie ont été arrêtés autres militants qui ...
Adapt:	Selon des sources militaires israéliennes, l'armée israélienne a arrêté dans la nuit de militants à Ramallah, en Cisjordanie, a également été arrêté deux autres activistes qui ...
Ref:	Des sources militaires israéliennes ont indiqué que l'armée israélienne a arrêté de nuit un activiste à Ramallah en Cisjordanie, ainsi que deux autres activistes qui ...
Source:	مُحَمَّدُ الغُبَارِي، جَولَةُ الصَّحَافَة، اَليَمَن.
Base:	Mohammed du brouillard, le cycle de la presse, au Yémen.
Adapt:	Mohammed, une tournée de la presse le Yémen.
Ref:	Mohamed Al-Ghobari, tour de la presse, Yémen.
Source:	مِن جِهَةٍ أُخرَى شَرَعَت تَايلَاند أَيضاً فِي سَحبِ جُنُودِهَا مِنَ العِرَاق.
Base:	d'autre part commencé aussi embarrassée à retirer ses troupes d'Irak.
Adapt:	D'autre part, la Thaïlande a commencé à retirer ses troupes d'Irak.
Ref:	D'autre part, la Thaïlande a également commencé à retirer ses troupes d'Irak.

FIG. 1 – Exemples de traductions automatiques tirés du jeux de test (système de référence, système adapté et référence de traduction humaine).

5 Conclusion

L'approche statistique à la traduction automatique est aujourd'hui utilisée pour construire rapidement des systèmes de traduction pour de nombreuses paires de langues. En général, on se contente de prendre tous les textes parallèles disponibles pour entraîner le modèle de traduction. La plupart de ces textes proviennent cependant d'un domaine bien spécifique – les discours parlementaires – ce qui les rend peu appropriés pour d'autres domaines. D'autre part, des textes monolingues existent généralement dans la plupart des domaines d'intérêt.

Dans ce travail, nous avons proposé une approche qui utilise des textes monolingues en langue source pour adapter un modèle de traduction générique. Pour cela les textes sont traduits par un système générique initial, filtrés et les plus fiables sont ajoutés aux textes parallèles. Après un nouveau cycle d'apprentissage, nous obtenons un système adapté. Cette technique a permis d'obtenir des améliorations significatives du score BLEU dans des systèmes de traduction arabe/anglais et arabe/français.

Plusieurs extensions de l'approche sont actuellement étudiées, notamment d'autres scores de confiance pour filtrer les traductions automatiques, le traitement des n meilleures hypothèses au lieu de la traduction la plus probable, et l'utilisation des alignements fournis par le décodeur au lieu de relancer GIZA++.

Remerciements

Ces recherches ont été partiellement financées par le gouvernement français sous le projet INSTAR (ANR JCJC06_143038) et la Commission Européenne sous le projet EuromatrixPlus. Le corpus parallèle arabe/ français de données radio et télévisées ainsi que les données de test correspondantes ont été mises à disposition par la DGA. Une partie de ces travaux a été effectuée en collaboration avec l'entreprise SYSTRAN.

Références

CHEN B., ZHANG M., AW A. & LI H. (2008). Exploiting n-best hypotheses for SMT self-enhancement. In *ACL*, p. 157–160.

CIVERA J. & JUAN A. (2007). Domain adaptation in statistical machine translation with mixture modelling. In *Second Workshop on SMT*, p. 177–180.

FOSTER G. & KUHN R. (2007). Mixture-model adaptation for SMT. In *EMNLP*, p. 128–135.

GAO Q. & VOGEL S. (2008). Parallel implementations of word alignment tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, p. 49–57, Columbus, Ohio : Association for Computational Linguistics.

HABASH N. & SADAT F. (2006). Arabic preprocessing schemes for statistical machine translation. In *NAACL*, p. 49–52.

HASAN S. & NEY H. (2008). A multi-genre SMT system for Arabic to French. In *LREC*, p. 2167–2170.

HILDEBRAND A. S., ECK M., VOGEL S. & WAIBEL A. (2005). Adaptation of the translation model for statistical machine translation based on information retrieval. In *EAMT*, p. 133–142.

KOEHN P., BIRCH A. & STEINBERGER R. (2009). 462 machine translation systems for Europe. In *MT Summit*.

KOEHN P., HOANG H., BIRCH A., CALLISON-BURCH C., FEDERICO M., BERTOLDI N., COWAN B., SHEN W., MORAN C., ZENS R., DYER C., BOJAR O., CONSTANTIN A. & HERBST E. (2007). Moses : Open source toolkit for statistical machine translation. In *ACL, demonstration session*.

KOEHN P. & SCHROEDER J. (2007). Experiments in domain adaptation for statistical machine translation. In *Second Workshop on SMT*, p. 224–227.

SCHWENK H. (2008). Investigations on large-scale lightly-supervised training for statistical machine translation. In *IWSLT*, p. 182–189.

SCHWENK H. (2010). Continuous space language models for statistical machine translation. *The Prague Bulletin of Mathematical Linguistics*, (93).

SNOVER M., DORR B. & SCHWARTZ R. (2008). Language and translation model adaptation using comparable corpora. In *EMNLP*.

SNOVER M., DORR B., SCHWARTZ R., MICCIULLA L. & MAKHOUL J. (2006). A study of translation edit rate with targeted human annotation. In *ACL*.

STOLCKE A. (2002). SRILM - an extensible language modeling toolkit. In *ICSLP*, p. II : 901–904.

UEFFING N. (2006). Using monolingual source-language data to improve MT performance. In *IWSLT*, p. 174–181.

UEFFING N. (2007). Transductive learning for statistical machine translation. In *ACL*, p. 25–32.

ZHAO B., ECK M. & VOGEL S. (2004). Language model adaptation for statistical machine translation with structured query models. In *Coling*.