Exploitation d'un corpus bilingue pour la création d'un système de traduction probabiliste Vietnamien - Français

Thi-Ngoc-Diep Do (1,2), Viet-Bac Le (1), Brigitte Bigi(1),
Laurent Besacier (1), Eric Castelli (2)

(1) Laboratoire LIG, GETALP, Grenoble, France
(2) Centre MICA, CNRS/UMI-2954, Hanoi, Vietnam
thi-ngoc-diep.do@imag.fr

Résumé Cet article présente nos premiers travaux en vue de la construction d'un système de traduction probabiliste pour le couple de langue vietnamien-français. La langue vietnamienne étant considérée comme une langue peu dotée, une des difficultés réside dans la constitution des corpus parallèles, indispensable à l'apprentissage des modèles. Nous nous concentrons sur la constitution d'un grand corpus parallèle vietnamien-français. La méthode d'identification automatique des paires de documents parallèles fondée sur la date de publication, les mots spéciaux et les scores d'alignements des phrases est appliquée. Cet article présente également la construction d'un premier système de traduction automatique probabiliste vietnamien-français et français-vietnamien à partir de ce corpus et discute l'opportunité d'utiliser des unités lexicales ou sous-lexicales pour le vietnamien (syllabes, mots, ou leurs combinaisons). Les performances du système sont encourageantes et se comparent avantageusement à celles du système de Google.

Abstract This paper presents our first attempt at constructing a Vietnamese-French statistical machine translation system. Since Vietnamese is considered as an under-resourced language, one of the difficulties is building a large Vietnamese-French parallel corpus, which is indispensable to train the models. We concentrate on building a large Vietnamese-French parallel corpus. The document alignment method based on publication date, special words and sentence alignment result is applied. The paper also presents an application of the obtained parallel corpus to the construction of a Vietnamese-French statistical machine translation system, where the use of different units for Vietnamese (syllables, words, or their combinations) is discussed. The performance of the system is encouraging and it compares favourably to that of Google Translate.

Mots-clés : traduction probabiliste, corpus bilingue, alignement de documents, table de traduction

Keywords: statistical machine translation, bilingual corpus, document alignment, phrase table

1 Introduction

Les systèmes de traduction automatique (TA) obtiennent aujourd'hui de bons résultats sur certains couples de langue comme anglais-français, anglais-allemand, anglais-japonais, etc. Toutefois, pour les langues peu dotées, on trouve encore peu de systèmes performants disponibles. Par exemple, bien que la langue vietnamienne soit parlée par environ 85 millions de personnes dans le monde[1], il n'existe pas beaucoup de recherches sur la TA de la langue vietnamienne.

Le premier système de TA de la langue vietnamienne est le système de « Logos Corporation » des années 1970. Ce système a été développé pour traduire des manuels d'utilisation en aéronautique de l'anglais vers le vietnamien (Hutchins, 2001). Au Vietnam, jusqu'à présent, on compte peu de groupes de recherche travaillant sur la TA vietnamien - anglais (Ho, 2005) et les résultats obtenus par les systèmes sont modestes. La recherche sur la TA vietnamien-français est encore plus rare. (Doan, 2001) a proposé un module de traduction pour le vietnamien dans ITS3, un système de TA multilingue basé sur l'approche par analyse-transfert-génération. (Nguyen, 2006) a étudié la langue vietnamienne et l'alignement des textes vietnamien-français. Mais, aucun système de TA n'a été proposé à ce jour.

Il existe de nombreuses approches de TA : des approches expertes fondées sur des règles linguistiques, des approches empiriques fondées sur l'apprentissage automatique à partir de corpus bilingues, ainsi que des approches hybrides. Nous nous concentrons dans cet article sur la construction d'un système de traduction probabiliste pour le couple de langue vietnamien-français. Une telle approche nécessite un corpus parallèle bilingue pour les langues source et cible. Ce corpus sert à construire les modèles statistiques (modèle de traduction et modèle de langage). Ensuite, ces deux modèles et un module de recherche sont utilisés pour décoder la meilleure hypothèse de traduction à partir d'un texte inconnu (Brown et al., 1993 ; Koehn et al., 2003).

La tâche de construction d'un grand corpus bilingue parallèle est très importante pour la TA probabiliste. Ce corpus est décrit comme un ensemble de paires de phrases bilingues. Aujourd'hui, un tel corpus parallèle vietnamien-français n'est pas disponible. (Nguyen, 2006) a collecté un corpus parallèle vietnamien-français de documents de droit et d'économie. Cependant, l'objectif de notre travail étant de construire un système de TA probabiliste dans le domaine des nouvelles journalistiques (news), nous nous concentrons à exploiter un corpus bilingue d'actualités vietnamien-français collecté à partir du Web.

L'organisation de cet article est la suivante : la section 2 présente la méthodologie générale d'exploitation d'un corpus bilingue de texte pour obtenir un corpus parallèle. Nous présentons une vue d'ensemble des méthodes d'alignement de document, d'alignement de phrases et discutons de la méthode d'identification automatique des paires de documents parallèles basée sur la date de publication, les mots spéciaux et les scores d'alignements des phrases. La section 3 décrit nos expériences sur l'exploitation automatique d'un site Web multilingue d'actualités vietnamien-français. La section 4 présente la construction de notre système de traduction vietnamien-français à partir du corpus parallèle obtenu. Nous discutons également l'utilisation de différentes unités lexicales côté vietnamien (mots, syllabes, ou combinaison de deux unités). Finalement, la section 5 donne quelques conclusions et perspectives.

[1] Source : Bureau national de statistique du Vietnam - http://www.gso.gov.vn

2 Exploitation d'un corpus bilingue de texte

(Munteanu et Marcu, 2006) présentent une méthode pour l'extraction des fragments parallèles de phrases à partir de corpus comparables. Cependant, leur méthode nécessite un corpus bilingue parallèle initial pour construire le système, qui n'est pas disponible pour le couple de langue français-vietnamien, en particulier dans le domaine des news.

Généralement, le processus d'exploitation d'un corpus de texte bilingue pour la traduction automatique se décompose en cinq étapes (Koehn, 2005) : la collection de données brutes, l'alignement des documents, la segmentation en phrases, la normalisation des écritures, l'alignement de phrases. Cette section présente les deux étapes principales : l'alignement de documents et l'alignement de phrases. Nous proposons et discutons également dans cette section une méthode qui combine l'alignement de phrases dans l'alignement de documents.

2.1 Alignement de documents

Soient *S1*, l'ensemble des documents dans la langue *L1*, et *S2* l'ensemble des documents dans la langue *L2*. L'extraction des paires de documents parallèles (PDPs) ou l'alignement de documents à partir de deux ensembles *S1*, *S2* consiste à trouver la traduction *D2*, dans l'ensemble *S2*, d'un document *D1*, dans l'ensemble *S1*. Après cette phase d'alignement, on obtient la paire de documents parallèle (PDP) *D1-D2*.

Pour collecter les données de texte bilingue, le Web est une source possible (Kilgarriff et Grefenstette, 2003). Pour ce type de données, certaines méthodes d'alignement de documents ont été proposées. Les documents peuvent être alignés simplement en utilisant les liens, l'information dans les URLs (Ma et Liberman, 1999) ou la structure de la page (Resnik et Smith, 2003). Toutefois, ces informations ne sont pas toujours disponibles ou fidèles. (Rosińska, 2007) a cité que ces méthodes ne sont pas vraiment efficaces dans la cas où le couple de langue n'est pas bien représenté sur Internet ou si les URLs changent. On peut également utiliser les titres des documents (Yang et Li, 2002), mais ils sont parfois totalement différents. Une autre source d'information utile est la présence de mots ou symboles invariants d'une langue à l'autre, comme par exemple les entités nommées, les dates et les nombres, qui sont souvent présents dans des données de type « news ». Ces mots invariants seront dénommés mots spéciaux dans cet article. (Patry et Langlais, 2005) ont utilisé des nombres, des marques de ponctuation, et des noms d'entités nommées pour mesurer le degré de parallélisme entre deux documents.

2.2 Alignement de phrases

A partir d'une paire *D1-D2*, le processus d'alignement de paires de phrases parallèles (PPPs) entre deux documents *D1* et *D2* est appliqué. Pour chaque paire *D1-D2*, nous avons un ensemble $SenAlignment_{D1\text{-}D2}$ qui contient toutes les PPPs.

$SenAlignment_{D1\text{-}D2}$ = {« *sen1-sen2* » | *sen1* est zéro, une ou plusieurs phrase(s) dans le document *D1*, *sen2* est zéro, une ou plusieurs phrase(s) dans le document *D2*, « *sen1-sen2* » est considéré comme une PPP}.

Nous appelons *sen1-sen2* un alignement du type *m : n* quand *sen1* contient *m* phrases consécutives et *sen2* contient *n* phrases consécutives. Plusieurs approches pour l'alignement automatique au niveau de la phrase ont été proposées. Elles utilisent la longueur de la phrase (Brown et al., 1991) ou des informations lexicales (Kay et Roscheisen, 1993). Une approche statistique est présentée dans (Gale et Church, 1993), dont l'idée principale est que les phrases plus longues (resp. courtes) dans une langue ont tendance à être traduites en phrases longues

(resp. courtes) dans l'autre langue. Certains outils tels que Hunalign[2] et Vanille[3] sont basés sur ces approches. Toutefois, ils sont plus efficaces lorsque les documents *D1* et *D2* contiennent peu de suppressions ou d'insertions, et contiennent principalement des PPPs du type 1 : 1. (Ma, 2006) a proposé un logiciel appelé Champollion[4] pour résoudre cette limitation. Champollion permet l'alignement de type *m : n* (*m, n* = 0, 1, 2, 3, 4). Champollion peut utiliser également des informations lexicales (lexèmes, mots vides, dictionnaire bilingue, etc.) pour aligner les phrases. Il peut aussi être adapté facilement à de nouveaux couples de langues. A l'heure actuelle, les couples de langues disponibles dans Champollion sont anglais-arabe et anglais-chinois (Ma, 2006).

2.3 Méthode de combinaison

La figure 1 décrit notre méthode d'alignement de documents en combinant avec l'alignement de phrases. Pour chaque document *D1* dans l'ensemble *S1*, on trouve le document aligné *D2* dans l'ensemble *S2*. Nous proposons d'utiliser la date de publication, les mots spéciaux et les scores d'alignements des phrases pour découvrir des PDPs. D'abord, la date de publication est utilisée pour limiter le nombre de documents possibles *D2*. Ensuite, nous utilisons un filtrage basé sur des mots spéciaux contenus dans le document pour déterminer les candidats *D2*. Enfin, les informations de longueur et de lexique du document, qui sont extraites à partir des scores d'alignements des phrases, sont utilisées pour extraire des documents candidats *D2*. Le détail de chaque étape est donné dans les sous-sections suivantes.

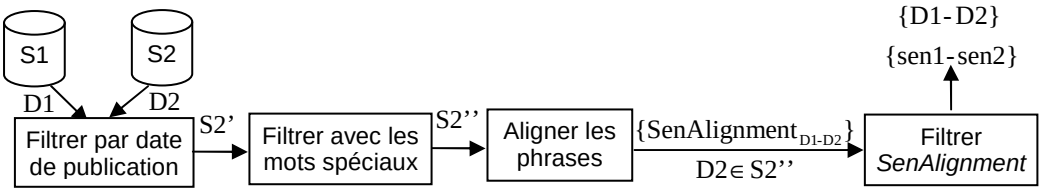

Figure 1 : Méthode d'alignement de documents

2.3.1 Premier filtrage : la date de publication

Nous supposons que le document est traduit et publié au plus *d* jours après la date de publication du document original. Dans la mesure où nous ignorons si *D1* ou *D2* est le document original, nous supposons que le document *D2* est publié *d* jours avant ou après le document *D1*. Après ce filtrage, on obtient un ensemble *S2'* contenant un ensemble de documents possibles de *D2*.

2.3.2 Deuxième filtrage : les mots spéciaux

Nous définissons les mots spéciaux comme étant des nombres ou des entités nommées. Non seulement les nombres, mais aussi les symboles joints ('$', '%', '‰', '.', ',' ...) sont extraits à partir des documents, par exemple: « 12.000 $ », « 13,45 », « 50% »... Les entités nommées sont pour l'instant précisées par une chaîne de mots dans laquelle tout mot commence par une lettre majuscule, par exemple, « Paris », « Nations Unies » en français. Nous envisageons de remplacer cette approche rudimentaire par une véritable détection d'entités nommées dans le futur. Bien que les entités nommées dans la langue *L1* sont généralement traduites dans les entités correspondantes dans la langue *L2*, dans certains cas, les entités nommées dans la langue *L1* (tels que les noms de personnes ou les noms de organisations) ne changent pas dans

[2] http://mokk.bme.hu/resources/hunalign
[3] http://nl.ijs.si/telri/Vanilla
[4] http://champollion.sourceforge.net

la langue *L2*. En particulier, les noms en vietnamien sont souvent traduits dans l'autre langue en supprimant juste les diacritiques. Par exemple, le nom en vietnamien « Nông Đức Mạnh » est traduit en français « Nong Duc Manh », le nom « Điện Biên » est traduit « Dien Bien »). À partir du document *D1*, tous les mots spéciaux sont extraits et on obtient une liste de mots spéciaux $w_1, w_2,...,w_n$. Pour chaque mot w_i, on recherche dans l'ensemble *S2'* les documents *D2* qui contiennent ce mot. Pour chaque mot, on obtient encore une liste de documents *D2*. Le document *D2* qui apparaît le plus dans toutes les listes est choisi. C'est le document contenant le plus grand nombre de mots spéciaux. Nous pouvons trouver zéro, un ou plusieurs documents qui satisfont cette condition. Nous appelons cet ensemble de documents *S2''* (voir la figure 2).

Notre utilisation des mots spéciaux est différente de celle présentés dans (Patry et Langlais, 2005). Par exemple, nous n'utilisons pas de ponctuation ; nous utilisons par ailleurs les symboles joints avec les nombres. Enfin, dans (Patry et Langlais, 2005), l'ordre des mots spéciaux dans les documents est utilisé comme critère important. Cependant, cet ordre n'est pas toujours respecté dans une PDP, particulièrement dans le domaine des nouvelles journalistiques. Donc, notre méthode ne pris pas en compte l'ordre des mots spéciaux dans les documents, et si un mot apparaît plusieurs fois dans un document, ceci n'affecte pas le résultat.

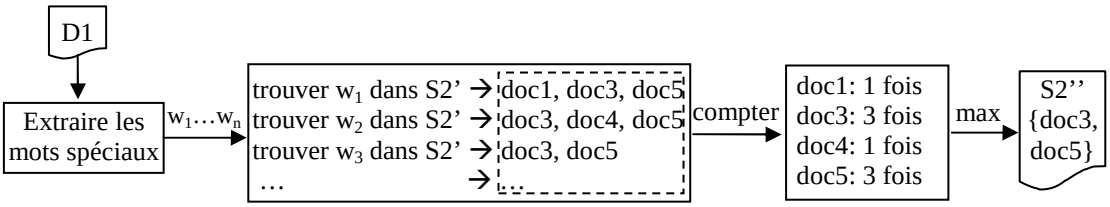

Figure 2 : Utiliser les mots spéciaux pour filtrer les documents D2

2.3.3 Troisième filtrage : alignement en phrases

Pour chaque document *D1*, nous découvrons un ensemble *S2''*, qui contient zéro, un ou plusieurs documents *D2*. Si certains couples de documents sont mal appariés (i.e. s'ils ne correspondent pas au même contenu), lorsque nous alignons ces documents en phrases, nous obtenons des PPPs de faible qualité. Les résultats de l'alignement en phrases nous permettent donc de filtrer les documents de *D2*. Après avoir aligné les documents en phrases, nous avons donc un ensemble de PPPs, $SenAlignment_{D1\text{-}D2}$, pour chaque PDP *D1-D2*. Nous ajoutons alors deux seuils α et β pour filtrer les documents *D2*.

Dans le premier seuil, le nombre de PPPs dans l'ensemble $SenAlignment_{D1\text{-}D2}$ est nommé $card(SenAlignment_{D1\text{-}D2})$. Le nombre de phrases qui ne peuvent pas trouver leur partenaire (quand *sen1* ou *sen2* est « null ») est noté $nbr\_omitted\ (SenAlignment_{D1\text{-}D2})$. Lorsque le rapport $[nbr\_omitted\ (SenAlignment_{D1\text{-}D2}) / card(SenAlignment_{D1\text{-}D2})]$ est supérieur à un seuil α , le candidat PDP *D1-D2* est éliminé.

Pour chaque PPP, nous ajoutons deux scores x_{L1} et x_{L2} pour les phrases *sen1* et *sen2* avec $x_{Li} = nombre\_de\_mots\_traduits\_dans\_sen_i / nombre\_de\_mots\_dans\_sen_i$. Les mots traduits sont les mots invariants ou les mots qui ont des équivalents dans l'autre phrase (nous utilisons un dictionnaire bilingue pour déterminer ces mots). Lorsque tous les PPPs dans $SenAlignment_{D1\text{-}D2}$ ont les deux scores x_{L1} et x_{L2} qui sont plus petits que β, le PDP *D1-D2* est éliminé. Ce seuil élimine les PDPs de mauvaise qualité qui génèrent des PPPs de faible qualité. Après avoir utilisé ces trois filtrages, nous obtenons un corpus de PDPs, et aussi un corpus de PPPs correspondant.

3 Expériences

3.1 Quelques caractéristiques de la langue vietnamienne

L'unité de base de la langue vietnamienne est la syllabe. A l'écrit, les syllabes sont séparées par un espace. Un mot correspond à une ou plusieurs syllabes (Nguyen, 2006). Par exemple, la phrase vietnamienne en syllable « *Thành phố hy vọng sẽ đón nhận khoảng 3 triệu khách du lịch nước ngoài trong năm nay* » peut être segmentée en mots comme suit : « *Thành_phố| hy_vọng| sẽ| đón_nhận| khoảng| 3| triệu| khách_du_lịch| nước_ngoài| trong| năm| nay* ». En ce qui concerne la morphologie du vietnamien, les verbes, noms et adjectifs peuvent être modifiés par d'autres unités (syllabe ou mots) qu'on leur juxtapose comme « *những* », « *các* » pour exprimer le pluriel, « *đã* » et « *sẽ* » pour exprimer le passé et l'avenir. Les fonctions syntaxiques sont également déterminées par l'ordre des mots dans la phrase (Nguyen, 2006).

3.2 Collection et prétraitement des données

En vue de la construction d'un corpus de texte parallèle vietnamien-français, nous avons appliqué notre méthodologie d'alignement pour exploiter un corpus de textes à partir d'un site Web multilingue d'actualités, le Vietnam News Agency[5] (VNA). Ce site contient des articles de presse écrits en quatre langues (vietnamien, anglais, français et espagnol). Cependant, tous les articles vietnamiens n'ont pas été traduits dans les trois autres langues. Il n'y a pas de lien ou d'information dans les URLs pour découvrir la relation entre les articles traduits. En plus, les articles traduits sont partiellement parallèles, ils contiennent le texte non identifié. La répartition de la quantité de données dans les quatre langues est différente (40% en vietnamien, 27% en anglais, 20% en français et 13% en espagnol). Chaque document est obtenu via un lien URL depuis le site Web VNA. À ce jour, nous avons obtenu environ 121000 documents dans les quatre langues, qui ont été récupérés entre le 12 avril 2006 et le 14 août 2008 ; chaque document contient en moyenne 10 phrases, avec environ 30-35 mots par phrase.

Nous avons séparé les données en deux ensembles. Le premier ensemble contient 1000 documents, nommé E_{1k}, dont l'appariement correct français-vietnamien est connu, suite à une annotation manuelle. Il a été utilisé pour régler les paramètres du système de filtrage décrit précédemment. Le reste des données est appelé E_{all}, dans ce cas les paramètres du système de filtrage, réglés sur E_{1k} ont été appliqués pour construire le corpus parallèle entier. Nous avons appliqué le processus de traitement ci-dessous pour chaque ensemble E_{1k} et E_{all} : (1) Extraire le contenu des documents, (2) Classer automatiquement les documents par langue, en utilisant l'outil TextCat[6], un outil d'identification des langues fondé sur les n-grammes de mots, (3) Traiter et nettoyer les documents vietnamiens et français en utilisant l'outil CLIPS-Text-Tk (LE et al., 2003) : convertir les documents html en documents textes, convertir le code des caractères, segmenter en phrases, segmenter en mots. Les corpus obtenus sont nommés *S1*, pour le français et *S2*, pour le vietnamien.

3.3 Estimation des paramètres de filtrage sur E_{1k}

Notre méthode d'alignement a été appliquée sur les corpus *S1* et *S2* qui sont extraits à partir de l'ensemble E_{1k}. Selon de notre corpus, nous avons supposé que $d = 2$. Le deuxième filtrage a été réalisé sur l'ensemble *S1* et l'ensemble *S2** qui a été créé en supprimant les diacritiques depuis l'ensemble *S2* (dans le cas du vietnamien). Après l'utilisation des deux filtrages, on obtient les données indiquées dans le tableau 1.

[5] http://www.vnagency.com.vn/
[6] http://www.let.rug.nl/~vannoord/TextCat/

Le processus d'alignement en phrases a été réalisé en utilisant les ensembles *S1, S2* et l'outil Champollion. Nous avons adapté les paramètres de Champollion au couple de langue vietnamien-français. Le troisième filtrage a ensuite été appliqué en faisant varier le paramètre α (0,4 ; 0,5 ; 0,6 ; 0,7) et le paramètre β (0,1 ; 0,15 ; 0,2 ; 0,25 ; 0,3 ; 0,35 ; 0,4). La F-mesure (F1 score) est donnée dans le tableau 2.

E_{1k}	Nbr. de doc. : 1000 Nbr. de doc. en français : 173 Nbr. de doc. en vietnamien : 348 Nbr. de PDPs vrais : 129
S2''	Nbr. de PDP découverts : 379 Nbr de PDP corrects : 129 Précision = 34,04% Rappel = 100%

Tableau 1 : Précision/Rappel après utilisation des deux premiers filtrages

β \ α	0,4	0,5	0,6	0,7
0,1	0,69	0,76	0,77	0,75
0,15	0,71	0,79	0,83	**0,84**
0,2	0,71	0,77	0,82	0,83
0,25	0,60	0,65	0,70	0,73
0,3	0,48	0,52	0,56	0,59
0,35	0,36	0,39	0,41	0,44
0,4	0,21	0,23	0,26	0,27

Tableau 2 : F-mesure pour des valeurs différentes de α et β

3.4 Application sur le corpus entier E_{all}

Nous avons appliqué la méthode avec les paramètres estimés dans la section 3.3 sur l'ensemble E_{all}. Au vu des résultats obtenus sur E_{1k}, nous choisissons les paramètres α=0,7 et β=0,15. Les caractéristiques du corpus entier traité sont présentées dans le tableau 3. Nous obtenons un total de 50k phrases parallèles qui constitue le corpus parallèle pour l'apprentissage d'un modèle de traduction français-vietnamien.

E_{all}	Nombre de documents : 120218 Nombre de documents en français : 20884 Nombre de documents en vietnamien : 54406
Corpus parallèle obtenu	Nombre de PDPs : 12108 Nombre de PPPs : 50322

Tableau 3 : Caractéristiques du corpus parallèle obtenu

4 Application : système de traduction probabiliste pour le couple de langue vietnamien-français

Nous avons construit un système de TA probabiliste vietnamien-français à partir du corpus parallèle obtenu en utilisant les outils libres GIZA++ (Och et Ney, 2003) et Moses (Koehn et al., 2007). Les scripts fournis avec Moses nous permettent de construire un modèle de traduction fondé sur des séquences de mots (*phrase table*). Il contient également des outils pour régler les paramètres des modèles et pour évaluer la qualité de traduction avec le score BLEU.

4.1 Préparation des données

A partir du corpus entier, nous avons choisi 50 PDPs pour le développement (Dev : 351 PPPs) et le réglage des paramètres du système de traduction (MERT), 50 PDPs pour le test (Tst : 384 PPPs), et le reste est réservé pour l'apprentissage (Trn : 49587 PPPs). En ce qui concerne les ensembles de développement et de test, ceux-cis ont été vérifiés manuellement et nous avons éliminé les PPPs de faible qualité. Au final, on obtient 198 PPPs de bonne qualité pour le développement et 210 PPPs de bonne qualité pour le test. Les données qui ont servi pour créer les modèles de traduction et de langage ont été extraites automatiquement depuis les 49587 PPPs de l'ensemble d'apprentissage, non vérifiées manuellement.

4.2 Systèmes de référence

Nous avons construit des systèmes de TA dans les deux sens : français vers vietnamien (F→V) et vietnamien vers français (V→F). Les données en vietnamien ont été segmentées en syllabes et en mots. La segmentation en mots est fondée sur l'algorithme de « longest matching» à partir d'un dictionnaire monolingue de mots vietnamiens. Nous avons au total quatre systèmes de TA. Nous avons supprimé les phrases qui sont plus longues que 100 mots/syllabes dans l'ensemble Trn et l'ensemble Dev, ainsi le nombre de PPPs utilisés pour chaque ensemble est légèrement différent entre les systèmes. Tous les mots trouvés sont implicitement ajoutés au vocabulaire du modèle de langage.

Segmentation du vietnamien	**Nombre de PPPs par ensemble**	**Langue**	**Taille du vocabulaire (K)**	**Nombre de mots/syllabes (K)**
Syllabe Système S1FV (F→V) Système S1VF (V→F)	Trn : 47081	Fr	38,6	1783,6
		Vn	21,9	2190,2
	Dev : 198	Fr	1,8	6,3
		Vn	1,2	6,9
	Tst : 210	Fr	1,9	6,4
		Vn	1,3	7,1
Mot Système S2FV (F→V) Système S2VF (V→F)	Trn : 48864	Fr	39,7	1893,0
		Vn	33,4	1629,0
	Dev : 198	Fr	1,8	6,3
		Vn	1,5	4,8
	Tst : 210	Fr	1,9	6,3
		Vn	1,6	4,9

Tableau 4 : Nos quatre systèmes de TA

	BLEU (%)
S1FV	40,09
S1VF	31,73
S2FV	40,59
S2VF	30,58

Tableau 5 : Evaluation des systèmes de TA sur l'ensemble de test

Nous obtenons les performances présentées dans le tableau 5 pour tous ces systèmes. Dans le cas de systèmes où le texte vietnamien a été segmenté en mots, les phrases traduites en vietnamien sont re-segmentées en syllabes avant de calculer le score BLEU, ceci afin que tous les scores BLEU évalués soient comparables. Les scores BLEU pour le sens de traduction français vers vietnamien sont environ 40% et pour le sens vietnamien vers français environ 31%, ce qui est encourageant pour un premier résultat. De plus, une seule référence a été utilisée pour estimer les scores BLEU dans nos expériences. Il est également intéressant de noter que la segmentation des phrases vietnamiennes en syllabes ou en mots ne modifie pas sensiblement la performance des deux sens de traduction.

4.3 Combinaison des systèmes fondés sur mot et syllabe en vietnamien

Nous avons effectué un autre test sur la combinaison des unités lexicales (syllabes et mots) sur le vietnamien. Nous avons réalisé le test dans le sens de traduction vietnamien vers français. En fait, l'outil Moses permet la combinaison des tables de traduction. Les tables de traduction du système S1VF (T_{syl}) et du système S2VF (T_{mot}) ont été utilisées. Une autre table (T_{mot*}) a été créée à partir de la table T_{mot}, dans laquelle tous les mots ont été re-transformés en syllabe (dans ce dernier cas, la segmentation en mot a été utilisée durant le processus d'alignement et de construction de la table de traduction, mais la partie en vietnamien de la table finale est re-segmentée en syllabes). Les combinaisons de ces trois tables de traduction ont également été créées. Les entrées en vietnamien, pour cette expérience, étaient segmentées soit en mot soit en syllabe. Comme précédemment, l'ensemble de développement a été utilisé pour régler des paramètres et l'ensemble de test a été utilisé pour estimer le score BLEU. Les résultats obtenus sont présentés dans le tableau 6. Des cellules sont marquées par X car certaines combinaisons n'ont pas de sens (par exemple la combinaison entre l'entrée en mots et la table de traduction en syllabes). Ces résultats montrent que la performance peut être améliorée en combinant les informations de mots et de syllabes du côté vietnamien. Le score BLEU est

amélioré de 35,30% à 38,02% sur l'ensemble Dev et de 31,73% à 32,08% sur l'ensemble Tst. À l'avenir, nous allons analyser plus en détails la combinaison entre les unités lexicales syllabes et mots pour les systèmes de TA vietnamiens et nous allons étudier l'utilisation d'un réseau de confusion comme entrée du système de TA, qui offre l'avantage de conserver les deux segmentations (en mots et en syllabes) dans une même structure.

	Tables de traductions utilisées	Entrée en syllabe		Entrée en mots	
		Dev	Tst	Dev	Tst
VN vers FR	Tsyl	**35,30**	**31,73**	X	X
	Tmot	X	X	35,70	30,58
	Tmot*	37,31	31,76	X	X
	Tsyl + Tmot	35,30	31,43	36,80	30,68
	Tsyl + Tmot*	**38,02**	**32,08**	X	X
	Tmot + Tmot*	37,42	30,23	36,67	30,21

Tableau 6 : Scores BLEU (%) obtenus depuis les combinaisons entre les tables de traduction (calculés sur l'ensemble de développement et l'ensemble de test)

4.4 Comparaison avec le système de TA de Google[7]

Le système de TA Google Translate a récemment ajouté la langue vietnamienne à sa liste de langues traitées. Dans la plupart des cas, il utilise l'anglais comme une langue intermédiaire. Pour la première évaluation comparative, un test simple a été réalisé. Deux ensembles de données ont été utilisés : un dans le domaine des actualités (l'ensemble Tst de la section 4.2), et un hors du domaine des actualités. Ce dernier a été obtenu à partir du site Web bilingue vietnamien-français de l'Ambassade de France au Vietnam[8]. Après avoir traité préalablement et aligné manuellement, nous avons obtenu 100 PPPs pour l'ensemble hors du domaine de données. Les phrases vietnamiennes ont été segmentées en syllabes. Les deux ensembles de données ont été traités par nos systèmes de TA (S1FV, S1VF) et le système de TA de Google. Les résultats des systèmes ont été post-traités (passage en minuscule) et les scores BLEU ont été estimés. Le tableau 7 présente les résultats de ce test. Bien que notre système soit logiquement meilleur sur l'ensemble de données dans le domaine, il est également légèrement meilleur que le système de TA de Google sur l'ensemble de données hors domaine (pour le couple de langue vietnamien-français).

	Sens de traduction	Le score BLEU (%)	
		Notre système	Google
Dans le domaine news (210 PPPs)	F→V	40,09	24,82
	V→F	32,08	15,63
Hors du domaine (100 PPPs)	F→V	25,00	24,38
	V→F	20,22	15,82

Tableau 7 : Comparer avec le système de TA de Google

5 Conclusions et perspectives

Dans cet article, nous avons présenté notre travail sur l'exploitation d'un corpus bilingue pour construire des systèmes de traduction probabilistes pour le couple de langue vietnamien-français. Nous avons décrit la méthode d'alignement de documents, qui est basée sur la date de publication, des mots spéciaux et l'utilisation des résultats de l'alignement en phrases. La méthode proposée est appliquée aux données vietnamiennes et françaises récupérés depuis un site Web multilingue d'actualités. Nous avons obtenu près de 12100 paires de documents parallèles et 50300 paires de phrases bilingues. Nous avons construit des systèmes de TA

[7] http://translate.google.com
[8] http://www.ambafrance-vn.org

utilisant l'outil Moses. Les scores BLEU pour le système de traduction français vers vietnamien est de 40,09% et vietnamien vers français est de 32,08%. De plus, la combinaison des informations entre les mots et les syllabes vietnamiennes peut être utile pour améliorer les performances du systèmes de TA vietnamien. Dans l'avenir, nous allons augmenter la taille du corpus et étudier l'utilisation d'unités lexicales différentes de la langue vietnamienne (syllabes, mots) dans un système de TA, ainsi que proposer des méthodes d'apprentissage non-supervisé pour améliorer notre système.

Références

BROWN P.F., LAI J.C. ET MERCER R.L. (1991). Aligning sentences in parallel corpora. *Proceedings of 47th Annual Meeting of the Association for Computational Linguistics*.

BROWN P.F., PIETRA S.A.D., PIETRA V.J.D. ET MERCER R.L. (1993). The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics. Vol. 19, no. 2.*

DOAN N.H. (2001). Generation of Vietnamese for French-Vietnamese and English-Vietnamese Machine Translation. *ACL, Proceedings of the 8th European workshop on Natural Language Generation*.

GALE W.A. ET CHURCH K.W. (1993). A program for aligning sentences in bilingual corpora. *Proceedings of the 29th annual meeting on Association for Computational Linguistics*.

HO T.B. (2005). Current Status of Machine Translation Research in Vietnam Towards Asian wide multi language machine translation project. *Vietnamese Language and Speech Processing Workshop*.

HUTCHINS W.J. (2001). Machine translation over fifty years. Histoire, epistemologie, langage: *HEL, ISSN 0750-8069, Vol. 23, Nº 1, 2001.*

KAY M. ET ROSCHEISEN M. (1993). Text - translation alignment. *Association for Computational Linguistics.*

KILGARRIFF A. ET GREFENSTETTE G. (2003). Introduction to the Special Issue on the Web as Corpus. *Computational Linguistics, volume 29*.

KOEHN P., OCH F.J. ET MARCU D. (2003). Statistical phrase-based translation. *Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*.

KOEHN P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. *Machine Translation Summit.*

KOEHN P., HOANG H., BIRCH A., CALLISON-BURCH C., ZENS R., FEDERICO M., BERTOLDI N., COWAN B., SHEN W. ET MORAN C. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. *Proceedings of the ACL.*

LE V.B., BIGI B., BESACIER L. ET CASTELLI E. (2003). Using the Web for fast language model construction in minority languages. *Eurospeech'03*.

MA X. (2006). Champollion: A Robust Parallel Text Sentence Aligner. *LREC: Fifth International Conference on Language Resources and Evaluation*.

MA X. ET LIBERMAN M. (1999). Bits: A method for bilingual text search over the web. *Machine Translation Summit VII.*

MUNTEANU D.S. ET MARCU D. (2006). Extracting parallel sub-sentential fragments from non-parallel corpora . *44th annual meeting of the Association for Computational Linguistics*

NGUYEN T.M.H. (2006). Outils et ressources linguistiques pour l'alignement de textes multilingues français-vietnamiens. *Thèse présentée pour l'obtention du titre de Docteur de l'Université Henri Poincaré, Nancy 1 en Informatique*.

OCH F. J. ET NEY H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics. Vol. 29.*

PATRY A. ET LANGLAIS P. (2005). Paradocs: un système d'identification automatique de documents parallèles. *12e Conference sur le Traitement Automatique des Langues Naturelles.* Dourdan, France.

RESNIK P. ET SMITH N.A. (2003). The Web as a Parallel Corpus. *Computational Linguistics.*

ROSIŃSKA M. (2007). Collecting polish-german parallel corpora in the internet. *Proc. of the International Multiconference on Computer Science and Information Technology*.

YANG C.C. ET LI P.W. (2002). Mining English/Chinese Parallel Documents from the World Wide Web. *Proceedings of the 11th International World Wide Web Conference*, Honolulu, USA.