TALN 2010, Montréal, 19-23 juillet 2010

Apprentissage non supervisé pour la traduction automatique : application à un couple de langues peu doté

Do Thi Ngoc Diep[1,2], 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 une méthode non-supervisée pour extraire des paires de phrases parallèles à partir d'un corpus comparable. Un système de traduction automatique est utilisé pour exploiter le corpus comparable et détecter les paires de phrases parallèles. Un processus itératif est exécuté non seulement pour augmenter le nombre de paires de phrases parallèles extraites, mais aussi pour améliorer la qualité globale du système de traduction. Une comparaison avec une méthode semi-supervisée est présentée également. Les expériences montrent que la méthode non-supervisée peut être réellement appliquée dans le cas où on manque de données parallèles. Bien que les expériences préliminaires soient menées sur la traduction français-anglais, cette méthode non-supervisée est également appliquée avec succès à un couple de langues peu doté : vietnamien-français.

Abstract This paper presents an unsupervised method for extracting parallel sentence pairs from a comparable corpus. A translation system is used to mine and detect the parallel sentence pairs from the comparable corpus. An iterative process is implemented not only to increase the number of extracted parallel sentence pairs but also to improve the overall quality of the translation system. A comparison between this unsupervised method and a semi-supervised method is also presented. The experiments conducted show that the unsupervised method can be really applied in cases where parallel data are not available. While preliminary experiments are conducted on French-English translation, this unsupervised method is also applied successfully to a low e-resourced language pair (Vietnamese-French).

Mots-clés : apprentissage non-supervisé, système de traduction automatique, corpus comparable, paires de phrases parallèles

Keywords: unsupervised training, machine translation, comparable corpus, parallel sentence pairs

1 Introduction

Les systèmes de traduction automatique (TA) obtiennent aujourd'hui de bons résultats sur certains couples de langues comme anglais-français, anglais-italien, anglais-espagnol, etc. 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. Toutefois, la recherche sur la TA pour des couples de langues dits « peu dotés » doit faire face au défi du manque de données. La TA probabiliste est fondée sur l'apprentissage de modèles à partir de grands corpus parallèles bilingues pour les langues source et cible. Ensuite, ces 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). Ainsi un grand corpus parallèle bilingue est préalablement nécessaire. Un tel corpus n'est pas toujours disponible en grande quantité, surtout pour les langues peu dotées. Les méthodes les plus communes pour construire des corpus parallèles consistent en des méthodes automatiques qui collectent des paires de phrases parallèles à partir du Web (Resnik, Smith, 2003 ; Kilgarriff, Grefenstette, 2003), ou des méthodes d'alignement qui extraient des documents/phrases parallèles à partir des deux corpus monolingues (Gale, Church, 1993 ; Patry, Langlais, 2005). Il y a aussi les méthodes d'extraction de paires de phrases parallèles à partir d'un corpus comparable (Zhao and Vogel, 2002; Fung and Cheung, 2004; Munteanu and Marcu, 2006). Ces méthodes nécessitent un corpus parallèle initial pour construire le premier système de TA qui sera utilisé dans le processus d'extraction (voir plus de détails dans la section 2.1). Nous supposons que dans le cas d'un couple de langues peu doté, même un petit corpus parallèle n'est pas forcément disponible pour développer le système de TA initial. La question que nous nous posons alors est : est-ce qu'un processus totalement non-supervisé, initialisé à partir d'un corpus comparable particulièrement bruité, permet d'apporter des solutions au problème du manque de données parallèles ?

Cet article présente une méthode d'extraction entièrement non-supervisée, qui est comparée avec une méthode semi-supervisée. Les premiers résultats montrent que la méthode non-supervisée peut être réellement appliquée dans le cas du manque de données parallèles. La section 2 définit les deux méthodes d'extraction de paires de phrases parallèles à partir d'un corpus comparable : la méthode semi-supervisée et la méthode totalement non-supervisée. La section 3 présente nos expériences et nos résultats obtenus préalablement pour la méthode non-supervisée sur des données parallèles bruitées anglais - français. La section suivante présente une application de cette méthode sur un corpus comparable d'un couple de langues peu doté : vietnamien - français. La dernière section donne quelques conclusions et perspectives.

2 Apprentissage semi-supervisé et non-supervisé

2.1 Méthode d'apprentissage semi-supervisée

Un corpus comparable contient des données qui ne sont pas parallèles (des phrases non-alignées), mais « étroitement liés par les mêmes contenus » (Zhao, Vogel, 2002 ; Fung, Cheung, 2004). Il contient « des niveaux de parallélisme différents, tels que des mots, des chaînes de mots, des phrases... » (Kumano et al., 2007). Pour extraire des données parallèles à partir d'un corpus comparable, (Zhao, Vogel, 2002) proposent un critère de maximum de vraisemblance qui combine des modèles de longueur des phrases et un modèle de lexique extrait d'un corpus parallèle aligné existant. Un processus itératif est appliqué pour ré-apprendre le modèle de lexique en utilisant les données extraites. (Munteanu, Marcu, 2006) présentent une méthode d'extraction des fragments parallèles de phrases. Chaque document en langue source est traduit vers la langue cible, en utilisant un dictionnaire bilingue. Le document dans la langue cible qui correspond à cette traduction est extrait. Des paires de phrases et de fragments parallèles sont extraits à partir de cette paire de document en utilisant un lexique de traduction. (Abdul-Rauf, Schwenk, 2009) présentent une technique similaire à celle de (Munteanu, Marcu, 2006), mais un système de TA statistique est utilisé au lieu

d'un dictionnaire bilingue, et une métrique d'évaluation (TER) est utilisée pour décider du degré de parallélisme entre les phrases. (Sarikaya et al, 2009) présente une méthode semi-supervisée avec des itérations lors desquelles les données extraites sont ajoutées au corpus initial parallèle pour re-construire le système de traduction. Toutes ces méthodes sont présentées comme des méthodes efficaces pour extraire des phrases/fragments parallèles à partir d'un corpus comparable.

Dans le contexte de notre travail, on considère que les termes corpus « comparable » et corpus « parallèle bruité » ont un sens équivalent, car un corpus « parallèle bruité » peut être extrait à partir d'un corpus « comparable » en utilisant un module de recherche d'information (RI) fondé sur des caractéristiques de base comme la date de publication, la longueur de phrases, etc. La proposition d'approches de RI avancées pour l'exploitation le corpus comparable est en dehors du champ de cet article dont le but est précisément de proposer un processus itératif permettant de s'affranchir de méthodes de RI avancées pour l'extraction de corpus parallèles.

2.2 Méthode d'apprentissage non-supervisée

Les méthodes presentées ci-dessus peuvent être considérées comme des méthodes semi-supervisées, qui ont besoin d'un corpus parallèle initial (ou au moins un dictionnaire bilingue) pour construire le système d'extraction. Nous supposons que dans le cas des langues peu dotées, ce corpus parallèle, même de petite taille, n'est pas disponible. Donc, nous proposons ici une méthode totalement non-supervisée où le point de départ est un corpus comparable bruité. Un des objectifs de ce travail est donc de voir si on peut construire un système de TA acceptable à partir d'un tel point de départ (corpus comparable bruité, versus corpus vraiment parallèle). La figure 1 illustre la différence entre notre définition des deux méthodes. Dans le cas de l'apprentissage non-supervisé, un système de TA statistique S_0 (le système de référence) est construit à partir d'un corpus comparable (C_2) (à l'inverse, dans la méthode semi-supervisée, le système S_0 est construit à partir d'un corpus parallèle (C_1)). Bien sûr, la qualité de S_0 n'est pas bonne dans le cas non-supervisé. Nous proposons d'utiliser ce système S_0 pour exploiter un autre corpus comparable bruité (D), et aussi pour améliorer la qualité globale du système de traduction. Le côté source du corpus D est traduit par le système S_0. La sortie est en suite comparée avec le côté cible du corpus D. Une métrique d'évaluation est calculée pour chaque paire de phrases. Plusieurs métriques d'évaluation sont envisagées et comparées pour déterminer laquelle est le plus appropriée : BLEU (Papineni et al., 2002), NIST (Doddington, 2002), TER (Snover et al., 2006) et une modification de PER : PER* (voir détails dans la section 3.3). Une paire est considérée comme une paire parallèle si la métrique d'évaluation est plus grande (avec les métriques BLEU, NIST, PER*) ou moins grande (avec la métrique TER) qu'un certain seuil.

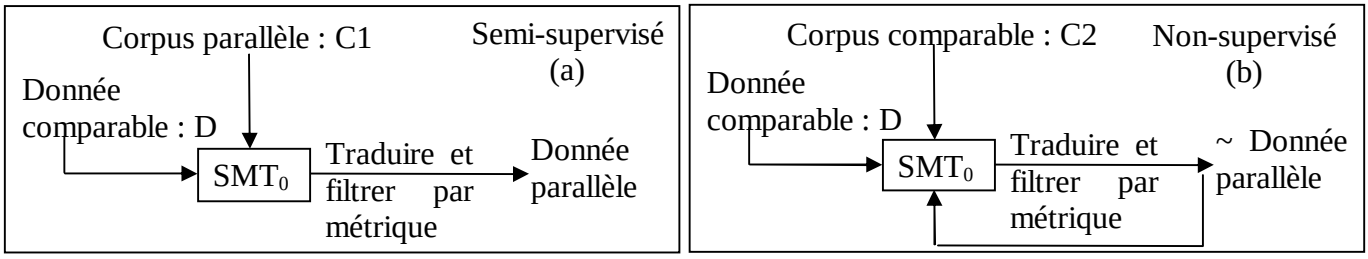

Figure 1 : Méthode de l'apprentissage semi supervisée (a) et non-supervisée (b)

Les paires de phrases extraites sont ensuite combinées avec le système de référence S_0 selon plusieurs manières pour créer un nouveau système de traduction. Un processus itératif est effectué qui va re-traduire le côté source par le nouveau système, re-calculer les métriques d'évaluation et re-filtrer les paires de phrases parallèles. Nous espérons que chaque itération augmente non seulement le nombre de paires de phrases parallèles extraites, mais améliore également la qualité du système de traduction. Pour utiliser les données extraites, quatre combinaisons différentes sont proposées :

- W1 : le système de TA à la i^{eme} étape est entraîné par un corpus consistant en C2 et E_{i-1} (les données extraites à la dernière itération) ; E_0 étant les données extraites lorsque le système de TA est entraîné par C2 seulement (S_0).
- W2 : le système de TA à la i^{eme} étape est entraîné par un corpus comprenant C2 et $E_0 + E_1 + ... + E_{i-1}$ (les données sont extraites aux itérations précédentes).
- W3 : à la i^{eme} itération, une nouvelle table de traduction est construite basée sur des données extraites E_{i-1}. Le système de TA décode en utilisant deux tables de traduction combinée dans un modèle log-linéaire : S_0 et cette nouvelle table. Les poids associés à chacune des tables sont les mêmes.
- W4 : la même combinaison que W3, mais la table de traduction de S_0 et la nouvelle table sont combinées en donnant plus d'importance aux données extraites E_{i-1} (par exemple 1:2).

3 Expériences préliminaires sur des données français-anglais

Dans cette section, nous présentons des expériences préliminaires concernant la méthode non-supervisée, qui est comparée à une approche semi-supervisée. Deux systèmes ont été construits, un fondé sur la méthode semi-supervisée (Sys1), un autre basé sur la méthode non-supervisée (Sys2).

3.1 Préparation des données

Afin de contrôler la précision et le rappel de la méthode d'extraction, un corpus comparable « artificiel » a été simulé en assemblant des paires de phrases parallèles et non parallèles, ainsi ces paires sont marquées en vue de l'estimation de la précision et du rappel du processus d'extraction. Nous avons choisi le couple de langues français-anglais pour ces expériences préliminaires. Les paires de phrases parallèles ont été choisies dans le corpus Europarl, version 3 (Koehn, 2005). Le corpus comparable « artificiel » a été construit par l'introduction d'un grand nombre de paires de phrases non-parallèles dans les données (environ 50%) (donc il peut être considéré comme un corpus parallèle bruité). Pour être comparable avec le cas réel traité dans la section 4 (pour les langues peu dotées), la taille des données expérimentales a été choisie relativement petite. Ainsi, le corpus C1 (cas semi-supervisé, voir fig.1a) ne contient que 50K paires de phrases parallèles correctes. Le corpus C2 (cas non supervisé, voir fig.1b) contient 25K paires de phrases parallèles correctes (retirées à partir de C1) et 25K paires de phrases non-parallèles. Le corpus D, données d'entrée pour le processus d'extraction, a été construit, quant à lui, avec 10K paires de phrases parallèles correctes et 10K paires de phrases non-parallèles, différentes des paires de phrases de C1 et C2.

3.2 Construction du système

Les deux systèmes Sys1 et Sys2 ont été construits en utilisant l'outil Moses (Koehn et al., 2007). Nous avons utilisé les paramètres par défaut de Moses et le paramétrage peut être résumé comme suit :

- L'outil GIZA++ (Och, Ney, 2003) a été utilisé pour l'alignement au niveau des mots, l'option pour l'extraction des séquences est de type « *grow-diag-final-and* »
- 14 caractéristiques ont été utilisées dans le modèle log-linéaire : le modèle de distorsion (6 caractéristiques), un modèle de langage, les probabilités de traduction bidirectionnelle au niveau séquence (2 caractéristiques), les probabilités associées au lexique bilingue mots (2 caractéristiques), une pénalité de phrase, une pénalité de mot et une pénalité de distance de distorsion.
- Le modèle de langage cible (3-gramme) a été construit à partir seulement de la partie anglaise du corpus Europarl entier en utilisant l'outil SRILM (Stolcke, 2002).
- Les modèles de traduction initiaux ont été construits à partir des corpus C1 et C2, pour les méthodes semi-supervisée et non supervisée respectivement.

3.3 Commencer à partir d'un corpus parallèle ou comparable?

La première question à laquelle nous voulons répondre tout d'abord est de savoir si le système de TA basé sur un corpus parallèle bruité ou comparable peut être utilisé pour filtrer les données d'entrée aussi efficacement que le système de TA basé sur un corpus parallèle. Pour répondre à cette question, le côté français du corpus D a été traduit par les Sys1 et Sys2. Ensuite, les traductions ont été comparées avec le côté anglais du corpus D. Quatre métriques d'évaluation ont été utilisées pour cette comparaison : BLEU, NIST, TER et PER*. Notre métrique PER* (la modification de PER - *position-independent word error rate* (Tillmann et al., 1997)) est calculée en se fondant sur la similitude, alors que le PER mesure une erreur (les mots différents) entre les hypothèses et la référence. Ainsi la formule de notre PER* est la suivante :

$$PER^* = \frac{2 * \text{nombre de mots identiques}}{\text{longueur de la hypothèse + longueur de la référence}}$$

Ensuite, les distributions des scores d'évaluation pour les paires de phrases parallèles correctes et les paires de phrases non-parallèles sont calculées et présentées dans la figure 2.

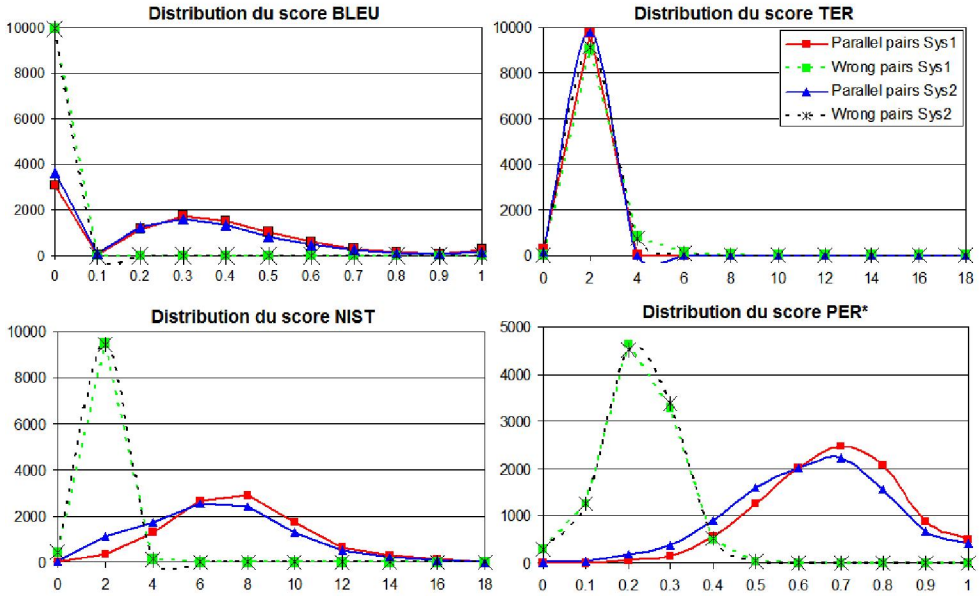

Figure 2 : Les distributions des scores pour la méthode semi-supervisée (Sys1) et non-supervisée (Sys2)

Nous pouvons faire les observations suivantes : les distributions des scores ont la même forme entre Sys1 et Sys2. En particulier, les distributions des scores pour les paires non-parallèles sont presque identiques selon les deux systèmes. Ainsi, un corpus parallèle bruité ou un corpus comparable peut remplacer un corpus parallèle dans la construction du système de TA initial. Par conséquent, cette méthode non-supervisée peut être réellement appliquée dans le cas du manque de données parallèles. Un autre résultat important est que le PER*, un score simple et facile à calculer, peut être considéré comme le meilleur score pour filtrer les paires de phrases parallèles correctes. Le tableau 1 présente la précision et le rappel du filtrage des paires de phrases parallèles des deux systèmes : Sys1 et Sys2.

Sys1 – méthode semi-supervisée						Sys2 – méthode non-supervisée					
Filtré par	Trouvé	Correct	Précision	Rappel	F1-mesure	Filtré par	Trouvé	Correct	Précision	Rappel	F1-mesure
Bleu=0.1	6908	6892	99.76	68.92	81.52	Bleu=0.1	6233	6218	99.75	62.18	76.61
Nist=0.4	8350	8347	99.96	83.47	90.97	Nist=0.4	7110	7108	99.97	71.08	83.08
Per*=0.3	10342	9785	94.61	97.85	96.20	Per*=0.3	10110	9468	93.65	94.68	94.16
Per*=0.4	9390	9333	99.39	93.33	96.27	Per*=0.4	8682	8629	99.38	86.29	92.37

Tableau 1 : Précision et rappel du filtrage des paires de phrases parallèles (avec 10K paires des phrases parallèles correctes)

3.4 Itérations de la méthode non-supervisée

La section 3.3 a montré qu'une méthode non-supervisée peut être utilisée aussi pour filtrer/extraire les paires de phrases parallèles à partir d'un corpus comparable. Toutefois, le résultat du filtrage dans le Sys2 est plus faible que celui dans Sys1 (par exemple, le nombre de paires de phrases correctes extraites est réduit (Tableau 1)). Ainsi nous proposons, dans cette section, un processus itératif, afin d'améliorer la qualité du système de traduction, puis d'augmenter le nombre de paires de phrases extraites correctement.

Augmenter le nombre de paires de phrases correctes extraites : Les paires de phrases extraites sont combinées avec le système de référence S_0 de quatre manière différentes (comme mentionné dans la section 2.2). L'expérience avec les itérations a été effectuée pour le Sys2 (non-supervisé). Afin de recevoir le nombre maximal de paires de phrases correctes extraites, pour toutes les itérations, on a choisi le score d'évaluation PER* avec le seuil = 0,3, qui a donné le rappel maximum = 94,68% pour le système de référence. La figure 3 présente le nombre de paires de phrases extraites correctement après 6 itérations pour les quatre combinaisons différentes : W1, W2, W3 et W4 décrites dans la section 2.2. Le nombre de paires correctes extraites est augmenté dans tous les cas, mais la combinaison W2 introduit le plus grand nombre de paires de phrases correctes extraites.

Augmenter la précision et le rappel du processus de filtrage : La précision et le rappel de ces quatre combinaisons sont présentés dans la figure 4. Parce que le processus de filtrage se concentre sur l'extraction du plus grand nombre de paires de phrases correctes extraites, la précision diminue. Toutefois, en utilisant la combinaison W2, le rappel après 6 itérations (97,77) atteint presque le rappel du système semi-supervisé Sys1 (97,85).

Évaluation du système de TA : La qualité du système de TA est évaluée également. Un ensemble de test contenant 400 paires de phrases parallèles Français-Anglais qui ont été extraites du corpus Europarl, est utilisé. Chaque phrase française n'a qu'une seule référence en anglais. La qualité est calculée selon BLEU et TER. La figure 5 donne les scores d'évaluation pour les systèmes après chaque itération.

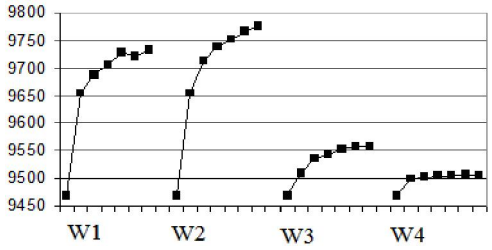

Figure 3 : Nombre de paires de phrases extraites correctement après 6 itérations pour quatre combinaisons différentes

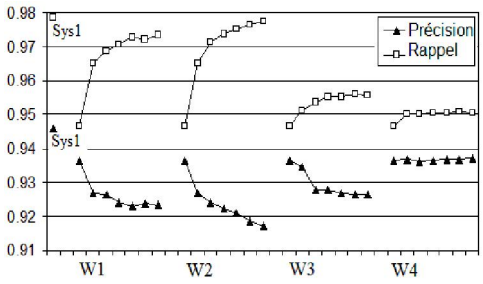

Figure 4 : Précision et rappel du filtrage en utilisant des combinaisons différentes

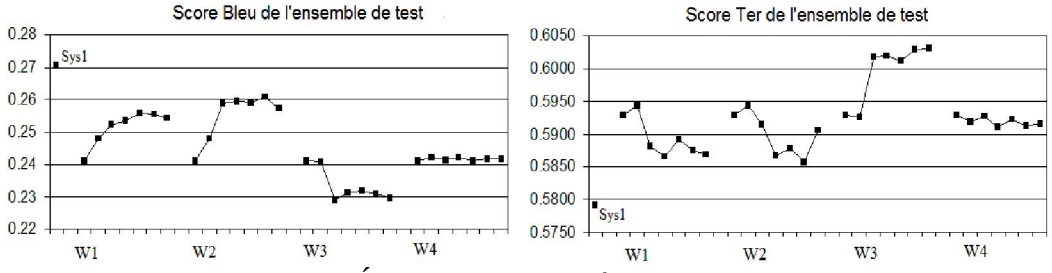

Figure 5 : Évaluation des systèmes de traduction

L'évaluation du système de TA révèle un résultat important. La qualité du système de TA augmente rapidement au cours des premières itérations, mais diminue après. On peut expliquer que, dans les premières itérations, un grand nombre de paires de phrases parallèles nouvelles sont extraites et sont incluses dans le modèle de traduction. Toutefois, dans les itérations suivantes, lorsque la précision du processus d'extraction

diminue, des paires de phrases non parallèles sont ajoutées au système ; le modèle de traduction est alors dégradé et la qualité du système de TA est réduite. Après environ 3 itérations, le score BLEU peut augmenter d'environ 2 points. On notera qu'il n'y a ici aucun réglage des paramètres du modèle log-linéaire à chaque itération (pas de données de développement utilisées, etc.).

(Sarikaya et al, 2009) présente une méthode semi-supervisée avec des itérations mais le système de TA initial est fondé sur un corpus parallèle. Il utilise la métrique d'évaluation Bleu pour le filtrage, et une combinaison semblable à notre combinaison W2. Cependant, leur recherche ne fournit pas une explication complète sur la façon dont ils choisissent la métrique d'évaluation, ou la méthode de combinaison (une seule est proposée à chaque fois), et en plus, la diminution de la qualité du système de TA après plusieurs itérations n'est pas mentionnée.

4 Application pour le couple de langues vietnamien-français

Le vietnamien est la 14eme langue la plus parlée dans le monde. Cependant, les recherches sur la TA pour le Vietnamien sont rares. Le plus ancien système de TA pour le Vietnamien 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. Nous nous concentrons sur la construction d'un système de TA statistique vietnamien-français. Le corpus d'apprentissage a été créé par l'exploitation d'un corpus des nouvelles journalistiques (news) collecté à partir du Web.

Une des méthodes d'exploitation a déjà été présentée par les auteurs de cet article dans (Do et al., 2009). Cette méthode est basée sur l'utilisation de la date de publication, des mots spéciaux et des scores d'alignement des phrases. D'abord, les paires de documents parallèles possibles sont filtrés utilisant la date de publication et les mots spéciaux (les numéros, les symboles joints, les entités nommées). Deuxièmement, des phrases dans une paire de documents parallèles possibles sont alignées en utilisant l'outil Champollion (Ma, 2006), qui utilise des informations lexicales (lexèmes, mots vides, dictionnaire bilingue, etc.). Enfin, des paires de phrases parallèles sont extraites en utilisant des informations d'alignement, des informations de longueur et un lexique du document. Cette méthode a été appliquée pour exploiter un corpus de texte à partir d'un site Web multilingue d'actualités, le Vietnam News Agency[1] (VNA) (contenant 20.884 documents français et 54.406 documents vietnamiens) qui est un corpus véritablement comparable : mêmes sujets traités, mais pas d'alignement évident en phrases. 50.322 paires de phrases « parallèles » ont été extraites. Un système de TA statistique pour le vietnamien-français a ensuite été construit en utilisant l'outil Moses avec les mêmes paramètres par défaut que ceux décrits dans la section 3.2 (voir plus dans (Do et al., 2009)). Ici, nous souhaitons comparer notre méthode non-supervisée, avec la méthode d'extraction présentée dans (Do et al., 2009). La méthode proposée a été appliquée sur le même corpus de VNA. Au lieu d'aligner les phrases et de les filtrer par des informations d'alignement de phrases, nous créons un corpus comparable et appliquons la méthode non-supervisée proposée directement sur ce corpus bruité.

4.1 Préparation des données

Chaque phrase dans un document vietnamien a été fusionnée avec toutes les phrases dans le document français correspondant. Ainsi une paire de documents vietnamien (contenant *m* phrases*)* et français (contenant *n* phrases*)* produit *mxn* paires de phrases. A partir du corpus VNA, nous avons obtenu un

[1] http://www.vnagency.com.vn/

corpus comparable de 1.442.448 de paires de phrases. Nous avons gardé seulement les paires avec le rapport de longueur (compté en mots) de la phrase française et de la phrase vietnamienne entre 0,8 et 1,3. Nous avons obtenu ainsi un corpus comparable de 345.575 paires de phrases (nommé C_{all}).

4.2 Création du système de traduction initial

Afin d'appliquer la méthode non-supervisée proposée, nous avons divisé le corpus C_{all} en deux ensembles : un corpus d'apprentissage initial C2 et un corpus à « fouiller » D (C2 et D sont indiquées dans la figure 1). Pour garantir une qualité minimale de C2 (et par conséquent pour la système de TA initial S_0), nous proposons ci-dessous un processus de filtrage croisé pour extraire le corpus C2 :

- Diviser le corpus C_{all} en 4 sous-corpus contenant des paires de phrases différentes : SC_1 (85.011 paires de phrases), SC_2 (85.008 paires de phrases), SC_3 (86.529 paires de phrases), SC_4 (89.027 paires de phrases).
- Construire 4 systèmes de TA différents : $SC_1 \rightarrow SMT_{SC1}$, $SC_2 \rightarrow SMT_{SC2}$, $SC_3 \rightarrow SMT_{SC3}$, $SC_4 \rightarrow SMT_{SC4}$.
- Appliquer la méthode non-supervisée pour chaque paire (SC_1, SMT_{SC2}), (SC_2, SMT_{SC1}), (SC_3, SMT_{SC4}), (SC_4, SMT_{SC3}), (une itération seulement ; seuil de PER* = 0,45 pour assurer la fiabilité des paires de phrases extraites (selon la figure 2) et un nombre acceptable de paires pour construire le système de traduction). Nous obtenons les paires de phrases extraites $C2_1$, $C2_2$, $C2_3$, $C2_4$, et leur union est considérée comme suffisamment fiable pour servir comme corpus comparable initial C2. Le reste est traité comme le corpus D.

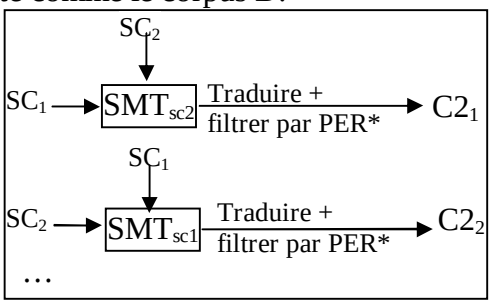

Figure 6 : Procesus d'extraction du corpus C2, pour la paire (SC_1, SMT_{SC2}), (SC_2, SMT_{SC1}), etc.

Sous-corpus	Traduit par	Nombre de paires C2	Nombre de paires D
SC_1	SMT_{SC2}	$C2_1$: 2916	82095
SC_2	SMT_{SC1}	$C2_2$: 3495	81513
SC_3	SMT_{SC4}	$C2_3$: 3820	82709
SC_4	SMT_{SC3}	$C2_4$: 3892	85135

Tableau 2 : Données extraites pour C2 et D

Après cette étape, nous avons obtenu un corpus C2 contenant 14.123 paires de phrases, et un corpus D contenant 331.452 paires de phrases. La méthode non-supervisée décrite dans la section 2.2 est ensuite appliquée sur C2 et D pour extraire plus de paires de phrases parallèles.

4.3 Application de la méthode non-supervisée

Le premier système de TA vietnamien-français S_0 a été construit à partir du corpus d'apprentissage C2 de 14.123 paires de phrases. Le corpus D contient 331.452 paires de phrases. La méthode non-supervisée a été appliquée avec le type de combinaison W2 et la métrique d'évaluation PER*. Il n'y a pas de processus de réglage des poids des modèles log-linéaires de traduction. Le nombre de paires de phrases extraites après chaque itération est indiqué dans la figure 7. Après 5 itérations, nous avons obtenu 39.758 paires. La qualité du système de TA est évaluée également sur un ensemble de test de 400 paires de phrases parallèles extraites manuellement (le même ensemble de test que dans (Do et al., 2009)). Les phrases en vietnamien sont segmentées en syllabes (pas de segmentation en mots). Chaque phrase vietnamienne n'a qu'une seule référence en français. Les scores d'évaluation après chaque itération sont reportés dans le tableau 3. Les résultats dans ce cas sont similaires à ceux obtenus lors des expériences préliminaires : le nombre de paires de phrases extraites augmente après quelques itérations, la qualité du système de TA augmente également lors des premières itérations et diminue par la suite.

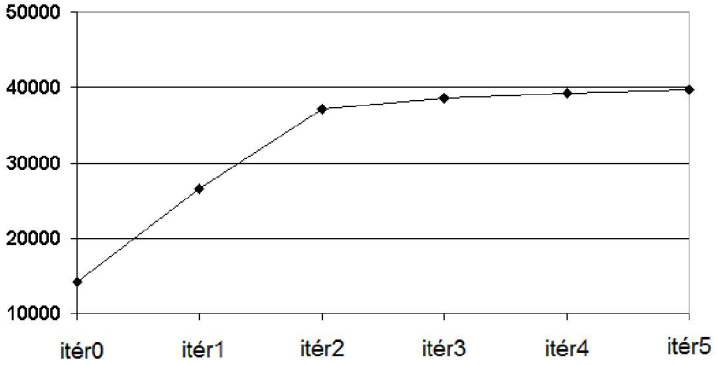

Figure 7 : Nombre de paires de phrases extraites après chaque itération dans le système de TA VN-FR

iter.	Données d'apprentissage	Bleu	Nist	Ter
0	14.123	30,67	6,45	0,59
1	26.517	32,18	6,70	0,57
2	37.210	32,42	6,75	0,56
3	38.530	32,45	6,77	0,55
4	39.254	32,14	6,73	0,56
5	39.758	31,85	6,68	0,56

Tableau 3 : Scores d'évaluation après chaque itération pour le système de TA VN-FR

Bien que le nombre de paires de phrases d'apprentissage ait augmenté d'environ deux fois de l'itération 0 à l'itération 1, le score d'évaluation n'augmente que de 2 points pour BLEU. Une raison est peut être que le système initial (S_0) a déjà une bonne performance grâce à notre filtrage croisé décrit dans la section 4.2. En outre, l'évaluation est conduite exclusivement avec des métriques automatiques en utilisant une seule référence, donc une analyse plus approfondie devrait être menée avec des évaluations subjectives. Pour comparer avec la méthode d'exploitation présentée dans (Do et al., 2009), la qualité des systèmes de TA des deux méthodes (évaluée sur le même corpus) est donnée dans le tableau 4. Bien que le nombre de paires de phrases extraites dans notre méthode soit plus faible que celui dans (Do et al., 2009), la qualité du système de TA est comparable. La méthode proposée dans (Do et al., 2009) dépend cependant de données / informations supplémentaires telles que la qualité du dictionnaire bilingue ou des règles heuristiques.

Méthodes	Donnée d'apprentissage	Bleu	Nist	Ter
Informations lexicales + heuristiques (Do et al., 2009)	50.322	32,74	6,78	0,55
Non-supervisée (it. 3)	38.530	32,45	6,77	0,55

Tableau 4 : Comparaison entre la méthode d'exploitation de (Do et al., 2009) et la méthode non-supervisée

A partir de ces résultats, nous pouvons dire que la méthode non-supervisée a été appliquée avec succès pour une couple de langues peu doté : vietnamien-français. Le résultat montre que cette méthode peut être réellement appliquée dans le cas d'un manque de données parallèles. En outre, la qualité du système de TA construit à partir des données extraites est comparable avec celle du système de TA d'une autre méthode utilisant des informations lexicales et des filtrages heuristiques. Cette méthode proposée ne nécessite pas de données supplémentaires. Nous avons l'intention d'appliquer cette méthode à une plus grande échelle pour exploiter un plus grand flux de données comparables extraites du Web.

5 Conclusions et perspectives

Cet article présente une méthode non-supervisée pour l'extraction de paires de phrases parallèles à partir d'un corpus comparable. Un système de TA initial a été construit, fondé sur un corpus parallèle bruité ou comparable, au lieu d'un corpus parallèle. Le système de TA initial a été ensuite utilisé pour traduire un autre corpus comparable. Un processus itératif a été évalué pour augmenter le nombre de paires de phrases parallèles extraites et pour améliorer la qualité du système de traduction. Les expériences montrent que cette méthode peut être réellement appliquée, notamment dans le cas du manque de données parallèles. Plusieurs méthodes de filtrage et utilisant les données extraites ont également été présentées. Un résultat intéressant est que la qualité du système de TA peut être améliorée au cours des premières itérations, mais elle est dégradée plus tard en raison de l'ajout de données bruitées dans les modèles statistiques. En outre, la qualité du système de TA construit avec cette méthode est comparable à celle d'une autre méthode qui nécessite des données de meilleure qualité comme un dictionnaire bilingue, des heuristiques etc. Dans

l'avenir, nous nous concentrerons sur l'approfondissement de l'analyse des meilleures techniques de filtrage, sur l'expérimentation à plus grande échelle, et sur des évaluations subjectives pour confirmer notre méthode non-supervisée.

6 Références

ABDUL-RAUF S., SCHWENK H. (2009). On the use of comparable corpora to improve SMT performance, *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*.

BROWN P.F., PIETRA S.A.D., PIETRA V.J.D., MERCER R.L. (1993). The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*. Vol. 19, no. 2.

DO T.N.D., LE V.B., BIGI B., BESACIER L., CASTELLI E. (2009). Exploitation d'un corpus bilingue pour la création d'un système de traduction probabiliste Vietnamien - Français. *TALN 2009*.

DODDINGTON G. (2002). Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. *In Human Language Technology Proceedings.*

FUNG P., CHEUNG P. (2004). Mining very-non-parallel corpora: parallel sentence and lexicon extraction via bootstrapping and EM. *Conference on Empirical Methods on Natural Language Processing.*

GALE W.A., 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.* ISSN 0750-8069.

KILGARRIFF A, GREFENSTETTE G. (2003). Introduction to the special issue on the Web as corpus. *Computational Linguistics, volume 29.*

KOEHN P. (2005). Europarl: a parallel corpus for statistical machine translation. *Machine Translation Summit*.

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

KOEHN P., HOANG H., BIRCH A., CALLISON-BURCH C., ZENS R., FEDERICO M., BERTOLDI N., COWAN B., SHEN W., MORAN C. (2007). Moses: open source tool-kit for statistical machine translation. *Proceedings of the Association for Computational Linguistics*.

KUMANO, T., TANAKA H., TOKUNAGA T. (2007). Extracting phrasal alignments from comparable corpora by using joint probability SMT model. *Conference on Theoretical and Methodological Issues in Machine Translation.*

MA X. (2006). Champollion: A robust parallel text sentence aligner. LREC: Fifth International Conference on Language Resources and Evaluation.

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

OCH F.J., NEY H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics 29.1*

PAPINENI K., ROUKOS S., WARD T., ZHU W. (2002). BLEU:a method for automatic evaluation of machine translation. *In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*.

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

RESNIK P., SMITH N.A. (2003). The Web as a parallel corpus. *Computational Linguistics.*

SARIKAYA R., MASKEY S., ZHANG R., JAN E., WANG D., RAMABHADRAN B., ROUKOS S. (2009). Iterative sentence–pair extraction from quasi–parallel corpora for machine translation. *Interspeech*.

SNOVER M., DORR B., SCHWARTZ R., MICCIULLA L., MAKHOUL J. (2006). A study of translation edit rate with targeted human annotation. *Proceedings of Association for Machine Translation in the Americas*.

STOLCKE A. (2002). SRILM an extensible language modeling toolkit. *Intl. Conf. on Spoken Language Processing.*

TILLMANN C., VOGEL S., NEY H., ZUBIAGA A., SAWAF H. (1997). Accelerated DP based search for statistical translation. *In 5th European Conf. on Speech Communication and Technology*.

ZHAO B., VOGEL S. (2002). Adaptive parallel sentences mining from Web bilingual news collection. *International Conference on Data Mining*.