Traduction Automatique Statistique à partir de corpus comparables : Application au couple de langues arabe-français

Rahma Sellami* — Fatiha Sadat — Lamia Hadrich Belguith***

**ANLP Research Group – Laboratoire MIRACL*
Faculté des Sciences Economiques et de Gestion de Sfax
B.P. 1088, 3018 - Sfax – TUNIESIE
rahma.sellami@fsegs.rnu.tn
l.belguith@fsegs.rnu.tn
***Université du Québec à Montréal, 201 av. President Kennedy,*
Montréal, QC, H3X 2Y3, Canada
sadat.fatiha@uqam.ca

RÉSUMÉ. *Dans cet article, nous nous intéressons à l'exploitation de corpus comparables pour la Traduction Automatique Statistique (TAS). Dans ce contexte, nous proposons deux approches.*

En premier lieu, une approche hybride basée sur des techniques statistiques et linguistiques est proposée afin d'extraire un lexique de terminologie bilingue à partir de Wikipédia. En second lieu, une approche hybride basée sur la longueur des phrases et un dictionnaire est proposée pour l'alignement du corpus des Nations Unies (UN), au niveau des phrases.

Nous avons intégrer les ressources crées dans un système de Traduction Automatique Statistique pour la paire de langues arabe-français. Nous avons obtenu des améliorations significatives du score BLEU, en utilisant ces deux approches en plus d'une technique de prétraitement des corpus en langue source (arabe).

ABSTRACT. *The present research aims to exploit comparable corpora for Statistical Machine Translation (SMT).*

First, a hybrid approach based on statistical and linguistics-based information is proposed for bilingual terminology extraction from Wikipedia documents. Then, we propose a hybrid approach based on length and dictionary model for the alignment of the United Nations (UN) corpus at the sentence level.

In order to validate the proposed approaches, we conducted evaluations on Arabic-French SMT. We evaluation showed significant improvement in term of BLEU scores when using these two approaches as well as a pre-processing technique, on the source language (Arabic).

MOTS-CLÉS : *Traduction Automatique Statistique (TAS), corpus comparable, Wikipédia, arabe-français.*

KEYWORDS: *Statistical Machine Translation (SMT), comparable corpora, Wikipedia, Arabic-French.*

1. Introduction

L'approche de Traduction Automatique Statistique (TAS) nous permet de construire rapidement un système de traduction si des données parallèles pour les langues source et cible sont disponibles. Toutefois, la recherche sur la TA pour certaines paires de langues comme l'arabe-français doit faire face au défi du manque de données parallèles. En effet, ces ressources ne sont pas toujours disponibles, pour certaines paires de langues qui ne font pas intervenir l'anglais (Morin *et al.,* 2011).

En outre, la performance de l'approche de TAS dépend de la façon de représenter les données (Sadat *et al*., 2006). La langue arabe est une langue à morphologie riche. Un simple prétraitement ne peut pas être efficace.

Cette étude concerne l'utilisation de corpus comparables afin d'améliorer la qualité de la TAS. 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 (Schwenk, 2010). C'est une ressource beaucoup plus disponible que les corpus parallèles. Dans ce type de corpus, même si les documents de la langue cible ne sont pas l'exacte traduction de ceux en langue source, on peut y retrouver des segments en relation de traduction.

Dans cet article, nous proposons d'exploiter deux sources de corpus comparables : l'encyclopédie en ligne Wikipédia et le corpus des Nations Unies (UN) pour la paire de langues arabe-français. En premier lieu, nous adoptons une approche hybride (statistique et linguistique) pour la construction d'un lexique de terminologie bilingue à partir de Wikipédia. En second lieu, nous proposons une approche d'alignement du corpus comparable UN. Aussi, nous proposons une technique de prétraitement des données arabe. Les ressources obtenues et les prétraitements effectués améliorent significativement notre système de TAS de base.

Dans la suite de ce papier, nous dressons en section 2 quelques travaux antérieurs qui ont exploité les corpus comparables pour la TA. Les sections 3 et 4 présentent nos approches d'exploitation de corpus comparables pour la TAS. La section 5 décrit nos évaluations en TAS. Nous concluons ce papier dans la section 6.

2. Travaux antérieurs

Différents travaux s'intéressent à l'extraction de lexiques bilingues à partir de corpus comparables. (Morin *et al*., 2004) extrait les termes composés dans chaque langue et les aligne au niveau mot en utilisant une méthode statistique exploitant le contexte des termes. Aussi, (Morin *et al*., 2011) présente une approche pour l'extraction de lexique bilingue spécialisé à partir d'un corpus comparable augmenté d'un corpus parallèle. (Hazem *et al*., 2011) propose un modèle inspirée des métamoteurs de recherche d'information. Récemment, (Hazem *et al*., 2012) représente ce problème par analogie au système Question-Réponse. Il considère le

mot comme une partie d'une question et cherche sa traduction dans la réponse de cette question dans la langue cible.

Plusieurs travaux ont exploité Wikipédia comme corpus comparable, pour l'extraction de terminologie bilingue. (Patry *et al.*, 2011), (Adafre *et al.*, 2006) et (Decklerck *et al*., 2006) ont exploité les liens inter-langues de Wikipédia. (Erdmann *et al.*, 2008) analyse non seulement les liens inter-langues de Wikipédia, mais exploite aussi les redirections et les liens inter-wiki. Les auteurs ont montré l'apport de l'utilisation de Wikipédia par rapport aux corpus parallèles pour l'extraction d'un dictionnaire bilingue. Cet apport apparait surtout au niveau de la large couverture des termes. (Ivanova, 2012) a exploré les liens inter-langues et les redirections. L'approche proposée par (Sadat *et al.*, 2010) tend à extraire d'abord des paires de termes et leurs traductions à partir des différents types d'informations de Wikipédia. Elle se base, ensuite, sur l'information linguistique pour réordonner les termes et leurs traductions pertinentes et éliminer les termes cibles inutiles. (Sellami *et al*, 2012) propose deux approches pour l'extraction de terminologie bilingue à partir de Wikipédia. Il s'agit, d'abord, d'extraire des paires de termes et traductions à partir des liens inter-langues de Wikipédia. Puis, appliquer soit une approche purement statistique, soit une approche hybride statistique et linguistique, afin de ré-ordonner les termes et leurs traductions pertinentes. Les auteurs ont montré que l'approche hybride est plus performante en termes de précision et de rappel que l'approche purement statistique.

L'extraction des phrases ou segments sous-phrastiques à partir des corpus comparables constitue un autre volet de recherche. Un critère de maximum de vraisemblance est proposé par (Zhao *et al*., 2002). Celui-ci combine des modèles basés sur la longueur de phrases avec un lexique extrait d'un corpus parallèle aligné existant. (Utiyama *et al*., 2003) utilise les techniques d'apprentissage en recherche d'information inter-langue et la programmation dynamique pour l'extraction des phrases parallèles à partir d'un corpus comparable. (Fung *et al*., 2004) utilise la mesure de similarité cosinus pour extraire les phrases parallèles. Un dictionnaire bilingue est crée à partir de ces phrases et enrichit itérativement pour actualiser la liste des phrases parallèles. Cette approche permet non seulement de confirmer le parallélisme des premiers couples extraits mais aussi de découvrir de nouvelles phrases parallèles. (Do *et al*., 2010) présente une approche non-supervisée pour l'extraction des paires de phrases parallèles à partir d'un corpus comparable. Il utilise un système de TA de base pour la détection des paires de phrases parallèles. Ce système de traduction est amélioré itérativement. (Munteanu *et al*., 2005) entraine un classifieur à maximum d'entropie permettant de discriminer les couples de phrases parallèles de ceux non parallèles, en se basant sur une série de traits généraux (longueur des phrases, pourcentage de mots traduits, …) et de traits d'alignement basés sur un lexique de terminologie (extrait à partir d'un corpus parallèle). Une approche similaire est présentée par (Rauf *et al*., 2011).

3. Approche d'extraction du lexique bilingue à partir de Wikipédia

Le processus d'extraction de terminologie bilingue à partir des documents de Wikipédia se base sur une approche hybride statistique et linguistique. La figure 1 montre ce processus. Il s'agit d'exploiter les liens inter-langues entre les articles arabes et français pour extraire les termes (simples ou composés) arabes et leurs traductions en français et les termes français et leurs traductions en arabe. Nous avons extrait 180 237 paires de titres.

Dans le but d'avoir un lexique composé uniquement des termes simples, nous avons procédé à une étape d'alignement des mots. Cette étape présente plusieurs défis dont : Premièrement, les alignements ne sont pas nécessairement contigus. On appelle ce phénomène *distorsion*. Deuxièmement, un mot en langue source peut être aligné à plusieurs mots en langue cible ; ce qui est défini en tant que *fertilité*.

Une méthode statistique basée sur les modèles IBM (Brown *et al*., 1993) et HMM (Vogel *et al*.,1996) est utilisée. En effet, ces modèles standards se sont avérés efficaces dans de nombreux travaux d'alignement de mots. Pour cela, l'outil GIZA++ (Och *et al*., 2003) a été utilisé dans les deux sens des langues du corpus. Nous avons obtenu deux alignements au niveau mot de type [1:N].

Afin de combiner ces alignements, nous avons procédé à un filtrage basé sur la correspondance des étiquettes morphosyntaxiques arabes et français[1]. À partir de ces alignements, nous avons extrait un lexique composé de 722 776 paires de termes.

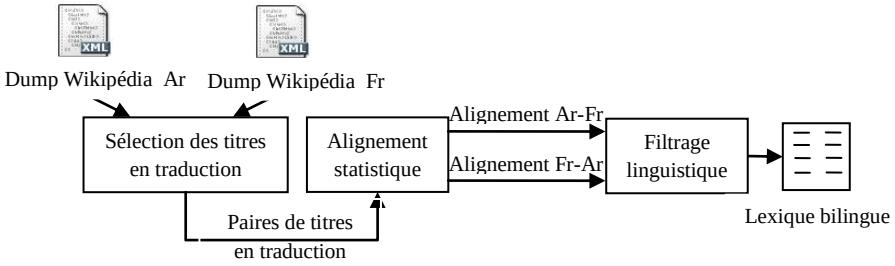

Figure 1. *Processus d'extraction du lexique*

4. Approche d'alignement du corpus comparable

Le corpus comparable UN contient les documents du ODS du site des Nations Unies pour les années 2000 à 2009. Ces documents sont nettoyés du contenu multimédia, segmentés en phrases et convertis en format XML (Eisele *et al*., 2010). Nous avons utilisé les documents arabes et français des années 2001, 2004, 2006 et

1 Nous avons utilisé l'étiqueteur morphosyntaxique Stanford pour les deux langues : http://nlp.stanford.edu/software/tagger.shtml

2008 pour notre travail. Un total de 27 072 documents arabes et 38 617 documents français est récupéré.

La figure 2 montre le processus d'alignement du corpus UN. Au début, nous avons aligné les documents en se basant sur leur nom. En effet, le nom d'un document du corpus UN est un identifiant unique, partagé entre les documents comparables. Nous avons obtenu 26 274 paires de documents comparables. Ensuite, nous avons aligné les phrases des documents comparables en utilisant une approche hybride basée sur la longueur des phrases et sur un dictionnaire. Le processus d'alignement est le suivant : les phrases d'une paire de documents comparables sont d'abord alignées en se basant sur leur longueur. Un dictionnaire bilingue est généré à partir de ces alignements. En se basant sur ce dictionnaire, les phrases sont réalignées pour construire l'alignement final. Nous avons réalisé ce processus avec l'outil Hunalign (Varga *et al.*, 2005). Au total, un nombre de 2 769 361 de phrases parallèles a été récupéré pour être considéré dans nos évaluations.

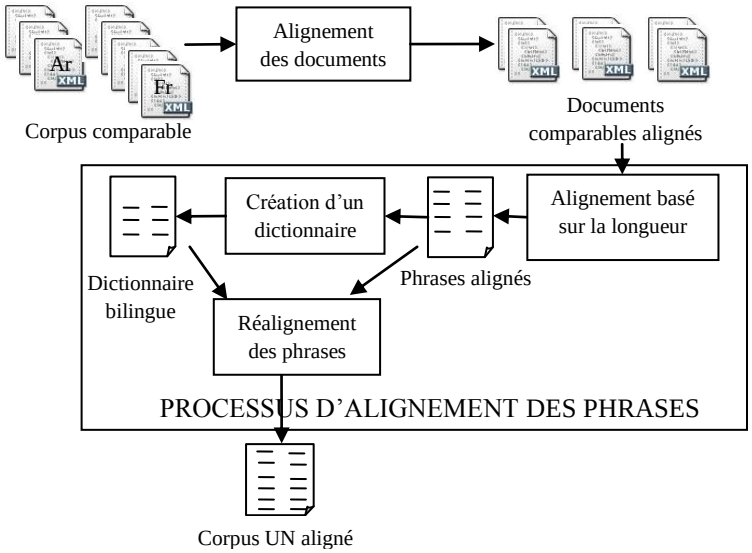

Figure 2. *Processus d'alignement du corpus UN*

5. Expérimentations et évaluations

Afin de montrer l'apport de nos ressources linguistiques : le corpus comparable aligné et le lexique de terminologie bilingue, nous les avons intégrés dans un système de TAS à base de segments.

Nous avons utilisé le bitexte « news commentary » que nous avons obtenu lors de notre participation à la compagne d'évaluation TRAD 2012[2] pour l'apprentissage des modèles de langue et de traduction du système de base. Ce corpus est de taille limité, il comporte 90 753 paires de phrases. Le corpus de développement est le corpus de test de nist08 de taille 813 paires phrases. Le corpus de test comporte 423 phrases en langue source avec 4 références fournies par la compagne TRAD 2012.

Nous avons utilisé l'outil open source Moses (Koehn *et al*., 2007) avec ses paramètres par défaut pour construire notre système de base pour la TAS. L'outil GIZA++ (Och *et al*., 2003) a été utilisé pour aligner les mots du corpus d'apprentissage. Un modèle 5-gramme de langue a été implémenté en utilisant le programme SRILM (Stolcke, 2002).

Le tableau 2 montre les résultats des méthodes de traduction suivantes pour la paire de langues arabe-français :

1. La méthode N utilise le corpus « news commentary » pour apprendre un modèle de langue et un modèle de traduction.

2. La méthode N_U utilise les corpus « news commentary » et « UN » pour apprendre deux modèles de langue et deux modèles de traduction.

3. La méthode N_U_TK utilise les mêmes ressources que la méthode N_U. De plus, une technique de tokenization du corpus arabe est faite selon le schème de déclitisation D3. La tokenization des mots du corpus français est simple ; elle se base sur les espaces pour séparer les mots, les nombres et la ponctuation.

4. La méthode N_U_TK_L utilise le lexique de terminologie bilingue extrait de Wikipédia , en plus du corpus « news commentary » et « UN », pour apprendre trois modèles de langues et trois modèles de traduction.

Dans les méthodes N et N_U, nous avons introduit une simple tokenization des mots du corpus arabe et français qui prend en compte les espaces pour séparer les mots, les nombres et la ponctuation.

Dans les méthodes N_U_TK et N_U_TK_L, la tokenization du corpus arabe est faite avec l'outil d'analyse morphologique et de désambiguïsation MADA+TOKEN V.3.2 (Habash et al., 2009) en appliquant le schème de déclitisation D3. (Sadat et al., 2006) a montré l'efficacité de ce schème pour la TAS pour la paire de langue arabe-anglais, surtout en utilisant un corpus d'apprentissage de taille limité. Ce schème segmente le mot en conjonction, particule, déterminant, racine et enclitique.

D3 : CONJ+ PART+ ART+ BASE +ENC.

Pour évaluer les différentes méthodes citées, nous avons utilisé la métrique BLEU et le taux des mots hors-vocabulaire (OOV, Out Of Vocabulary).

2 http://www.trad-campaign.org/

Le tableau 1 montre l'amélioration du système de TAS pour chaque méthode proposée et une comparaison avec l'outil « Google Translate »[3], en termes de score BLEU et de taux OOV.

Google Translate est un outil de traduction basé sur les modèles statistiques. Sa mémoire de traduction contient environ 200 milliards de mots provenant des documents des Nations Unies[4].

Méthode	OOV	Score BLEU
N	18.23	25.84
N_U	9.7	34.26
N_U_TK	1.56	38.46
N_U_TK_L	1.39	39.09
Google Translate	0.2	36.15

Tableau 1. *Résultats en termes de score BLEU sur le corpus de test*

Les meilleures performances de notre système de traduction sont atteintes lors de l'utilisation de toutes les données d'apprentissages proposées et la technique de prétraitement des données arabes (la méthode N_U_TK_L).

Une bonne amélioration du score BLEU et une baisse du taux des mots OOV sont remarquées lors de l'introduction de la tokenization des données arabes (la méthode N_U_TK). Ceci montre clairement l'intérêt de l'étape de prétraitement pour la langue arabe. En effet, une tokenization approfondie enlève l'ambiguïté engendrée par le phénomène de l'agglutination dans les mots arabes.

L'utilisation du corpus UN aligné améliore le score BLEU du système de traduction de neuf points et baisse le taux des mots OOV de neuf points. Ceci peut s'expliquer par la nouvelle quantité importante des données d'apprentissage et le large vocabulaire introduit (la méthode N_U). Aussi, l'introduction du lexique, extrait à partir de Wikipédia, élargit le vocabulaire d'apprentissage. Ce qui améliore les résultats en termes de score BLEU et de taux des mots OOV.

Une comparaison de nos méthodes avec l'outil « Google Translate » montre que nos résultats sont meilleurs en terme de score BLEU. « Google Translate » produit moins de mots hors vocabulaire (OOV) avec une différence d'environ trois points avec la méthode N_U_TK_L. Ceci peut être dû à la différence de taille des données d'apprentissage. En effet, nous n'avons utilisé que les documents des années 2001, 2004, 2006 et 2008 de l'ODS du site des Nations Unies alors que les données d'apprentissage de « Google Translate » sont en mise à jour progressive.

3 Ces expérimentations ont été faites en ligne en date du 15 décembre 2012

4 http://edouard-lopez.com/fac/SciCo%20-%20S5/TAL/projet/TAL%20-%20systran%20vs.%20google%20translate.pdf

D'après l'analyse des mots non traduits (OOV) par la méthode N_U_TK_L, nous envisageons la lemmatisation : enrichir des données arabes par le lemme de chaque mot pour mieux représenter les données arabes.

6. Conclusion

Nous avons proposé dans ce papier plusieurs améliorations du score BLEU d'un système de TAS arabe-français en exploitant deux corpus comparables et en intégrant une technique de prétraitement de la langue arabe.

Dans le futur, nous souhaitons explorer la translittération des entités nommées et la lemmatisation afin d'améliorer la performance de notre système de traduction. Aussi, une analyse des mots appartenant à différents dialectes arabes sera très avantageuse pour la TAS utilisant la langue arabe comme langue source.

Bibliographie

Adafre, S. F., De Rijke, M., « Finding Similar Sentences across Multiple Languages in Wikipedia », *Proceedings of the EACL Workshop on NEW TEXT Wikis and blogs and other dynamic text sources*, 2006, p. 62–69.

Brown P., Della Pietra S., Della Pietra V., Mercer R., « The Mathematics of Machine Translation: Parameter Estimation ». *Computational Linguistics* 19(2) Juin 1993, 263-312.

Do, T. N. D., Besacier, L. Et Castelli, E., « Apprentissage non supervisé pour la TA : application à un couple de langues peu doté », *TALN*, 2010.

Eisele A., Chen Y., « MultiUN : A multilingual corpus from United Nation documents », *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May. *European Language Resources Association (ELRA).* 2010.

Erdmann, M., Nakayama, K., Hara, T., Nishio, S. « A bilingual dictionary extracted from the wikipedia link structure », *In Proceedings of International Conference on Database Systems for Advanced Applications (DASFAA) Demonstration Track*, 2008, p. 380-392.

Habash, N., Owen R., Ryan., « MADA+TOKAN: A Toolkit for Arabic Tokenization, Diacritization, Morphological Disambiguation, POS Tagging, Stemming and Lemmatization », *In Proceedings of the 2nd International Conference on Arabic Language Resources and Tools (MEDAR)*, Cairo, Egypt, 2009.

Hazem, A., Morin, E., Sebastian P. S., « Bilingual Lexicon Extraction from Comparable Corpora as Metasearch », *Proceedings of the 4th Workshop on Building and Using Comparable Corpora*, 2011, pages 35–43, 49th Annual Meeting of the Association for Computational Linguistics, Portland, Oregon.

Hazem, A. et Morin, E., (2012). « A New Method for Bilingual Lexicon Extraction from Comparable Corpora », *Proceedings of the 13th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing)*, New Delhi, India.

Morin, E., Daille, B., « Extraction de terminologies bilingues à partir de corpus comparables d'un domaine spécialisé », *Traitement Automatique des Langues (TAL)*, 2004, p. 103–122.

Morin, E., Prochasson E. « Bilingual Lexicon Extraction from Comparable Corpora Enhanced with Parallel Corpora ». *4th Workshop on Building and Using Comparable Corpora*, 2011.

Munteanu, D. S., Marcu, D., « Improving Machine Translation Performance by Exploiting Non-Parallel Corpora ». *Computational Linguistics*, 2005, 31(4):477–504.

Och, F.J. et Ney, H.. « A systematic comparison of various statistical alignment models », *Computational Linguistics*, March 2003, p. 19–51.

Patry Alexandre et Langlais Philippe, « Identifying Parallel Documents from a Large Bilingual Collection of Texts: Application to Parallel Article Extraction in Wikipedia », *Proceedings of the 4th Workshop on Building and Using Comparable Corpora*, Portland, OR. , 2011.

Rauf, S. A. et Schwenk, H., « Parallel sentence generation from comparable corpora for improved SMT », *Machine Translation*, 2011, 25(4):341–375.

Rohit G. Bharadwaj, Vasudeva Varma, « Language independent identication of parallel sentences using wikipedia », *Proceedings of the 20th international conference companion on World wide web, WWW '11*, pages 11-12, New York, NY, USA, 2011. ACM. 16, 45.

Sadat, F. et Habash N., « Combination of Arabic preprocessing schemes for statistical machine translation », *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, 2006.

Sadat, F. et Terrassa, A., « Exploitation de Wikipédia pour l'Enrichissement et la Construction des Ressources Linguistiques », *TALN 2010*, Montréal.

Sadat, F. et Yoshikawa, M., Uemura, S., « Bilingual terminology acquisition from comparable corpora and phrasal translation to cross-language information retrieval », *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume*, 2003, p. 141–144

Sellami R., Sadat F. et Hadrich Belguith L., « Exploiting Wikipedia as a Knowledge Base for the Extraction of Linguistic resources: Application on Arabic-French Comparable Corpora and Bilingual Lexicons », *Proc. of the CAASL4 Workshop at AMTA 2012 (Fourth Workshop on Computational Approaches to Arabic Script-based Languages)*, San Diego, CA, 2012.

Schwenk Holger. « Adaptation d'un Système de Traduction Automatique Statistique avec des Ressources monolingues », *TALN 2010*, Montréal, 19–23.

Utiyama, M. et Isahara, H., « Reliable measures for aligning japanese-english news articles and sentences ». *In Proceedings of ACL'03*, 2003, volume 1, p. 72–79.

Varga D., Németh L., Halacsy P., Kornai A., Tron V., and Nagy V., « Parallel corpora for medium density languages », *In RANLP*, 2005, Borovets, Bulgaria, p. 560–596.

Vogel, S., Ney H., C. Tillmann, « HMM-based word alignment in statistical translation ». *In Preceding of the Conference on Computational Linguistics*, Morristown, NJ, USA, 1996.

Zhao, B., Vogel, S. « Adaptive parallel sentences mining from web bilingual news collection », *In Proceedings of IEEE International Conference on Data Mining*, 2002, p. 745.