Traduction automatique à partir de corpus comparables: extraction de phrases parallèles à partir de données comparables multimodales

Haithem AFLI Loïc BARRAULT Holger SCHWENK
Laboratoire d'Informatique de l'Université du Maine
prénom.nom@lium.univ-lemans.fr

RÉSUMÉ

Les performances des systèmes de traduction automatique statistique dépendent de la disponibilité de textes parallèles bilingues, appelés aussi *bitextes*. Cependant, les corpus parallèles sont des ressources limitées et parfois indisponibles pour certains couples de langues ou domaines. Nous présentons une technique pour l'extraction de phrases parallèles à partir d'un corpus comparable multimodal (audio et texte). Ces enregistrements sont transcrits avec un système de reconnaissance automatique de la parole et traduits avec un système de traduction automatique. Ces traductions sont ensuite utilisées comme requêtes d'un système de recherche d'information pour sélectionner des phrases parallèles sans erreur et générer un bitexte. Plusieurs expériences ont été menées sur les données de la campagne IWSLT'11 (TED) qui montrent la faisabilité de notre approche.

ABSTRACT

Automatic Translation from Comparable corpora : extracting parallel sentences from multimodal comparable corpora

Statistical Machine Translation (SMT) systems depend on the availability of bilingual parallel text, also called bitext. However parallel corpora are a limited resource and are often not available for some domains or language pairs. We present an alternative method for extracting parallel sentences from multimodal comparable corpora. This work extends the use of comparable corpora, in using audio instead of text on the source side. The audio is transcribed by an automatic speech recognition system and translated with a base-line SMT system. We then use information retrieval in a large text corpus of the target language to extract parallel sentences. We have performed a series of experiments on data of the IWSLT'11 speech translation task (TED) that shows the feasibility of our approach.

MOTS-CLÉS : Reconnaissance de la parole, traduction automatique statistique, corpus comparables multimodaux, extraction de phrases parallèles.

KEYWORDS: Automatic speech recognition, statistical machine translation, multimodal comparable corpora, extraction of parallel sentences.

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN, pages 447–454,
Grenoble, 4 au 8 juin 2012.

1 Introduction

La construction d'un système de traduction automatique statistique (TAS) nécessite un corpus dit parallèle pour l'apprentissage du modèle de traduction et des données monolingues pour construire le modèle de langue cible. Un corpus parallèle est une collection de textes bilingues alignés au niveau de la phrase, c'est-à-dire des textes en langue source avec leurs traductions.

Malheureusement, les textes parallèles librement disponibles sont aussi des ressources rares : la taille est souvent limitée, la couverture linguistique insuffisante ou le domaine n'est pas approprié. Il y a relativement peu de paires de langues pour lesquelles des corpus parallèles de taille raisonnable sont disponibles comme l'anglais, le français, l'espagnol, l'arabe, le chinois et quelques langues européennes (Hewavitharana et Vogel, 2011). De plus, ces corpus proviennent principalement de sources gouvernementales, comme le parlement canadien ou européen, ou de l'Organisation des Nations Unies. Ceci est problématique en TAS, parce que les systèmes de traduction appris sur des données provenant, par exemple, d'un domaine politique ne donnent pas de bons résultats lorsqu'ils sont utilisés pour traduire des articles scientifiques.

Une façon de pallier ce manque de données parallèles est d'exploiter les corpus comparables qui sont plus abondants. Un corpus comparable est un ensemble de textes dans deux langues différentes, qui ne sont pas parallèles au sens strict du terme, mais qui contiennent les mêmes informations. On peut par exemple citer les actualités multilingues produites par des organismes de presse tels que l'Agence France Presse (AFP), Xinhua, l'agence Reuters, CNN, BBC, etc. Ces textes sont largement disponibles sur le Web pour de nombreuses paires de langues (Resnik et Smith, 2003). Le degré de parallélisme peut varier considérablement, en allant de documents peu parallèles, aux documents quasi parallèles ou « parallèles bruités » qui contiennent de nombreuses phrases parallèles (Fung et Cheung, 2004). Ces corpus comparables peuvent couvrir différents sujets.

Ces travaux s'inscrivent dans le cadre du projet DEPART (Documents Écrits et PAroles – Reconnaissance et Traduction) dont l'un des objectifs est l'exploitation de données multimodales et multilingues pour la TAS. Nous considérons le cas, assez fréquent pour des domaines de spécialité, où un manque de données textuelles peut être pallié par l'exploitation de données audio. Un domaine de spécialité est un sous-domaine possédant un vocabulaire spécifique, tel que la chirurgie dans le domaine plus large de la médecine. Nous pouvons également considérer les conférences ou séminaires scientifiques et leurs articles associés pour un domaine de recherche spécifique.

La question que nous nous posons alors est la suivante : un corpus comparable multimodal permet-il d'apporter des solutions au problème du manque de données parallèles ? Dans ce travail nous proposons une méthode pour l'utilisation de corpus comparables multimodaux, en se limitant aux modalités texte et audio, pour l'extraction de données parallèles.

2 Recherches précédentes

Plusieurs travaux ont traité de l'extraction des données parallèles à partir d'un corpus comparable bilingue. Un critère de maximum de vraisemblance est proposé par Zhao et Vogel (2002) qui ont combiné des modèles de longueur de phrases avec un lexique extrait d'un corpus parallèle

aligné existant. Le lexique est itérativement adapté avec un processus de réapprentissage en utilisant les données extraites. Resnik et Smith (2003) ont montré qu'ils peuvent générer un grand nombre de documents parallèles à partir du WEB en utilisant leur système d'extraction de textes parallèles, « STRAND ». Do *et al.* (2010) ont utilisé une méthode non-supervisée pour extraire des paires de phrases parallèles à partir d'un corpus comparable et ont montré que cette approche est intéressante surtout pour les langues peu dotées. La détection des paires de phrases parallèles est faite en utilisant un système de traduction automatique de base qui est amélioré avec un processus itératif.

Afin de construire un corpus parallèle anglais/japonais, Utiyama et Isahara (2003) utilisent la recherche d'information cross-langue et la programmation dynamique pour l'extraction de phrases parallèles à partir d'un corpus comparable dans le domaine des actualités. Les paires d'articles similaires sont identifiées et traitées comme des textes parallèles afin d'aligner leurs phrases. La procédure d'alignement commence par la traduction mot à mot des textes japonais en utilisant un dictionnaire bilingue, qui sont ensuite pris comme requêtes de recherche d'information dans la partie anglaise des textes. L'approche de Fung et Cheung (2004) utilise la mesure « cosinus » pour calculer le degré de similarité des phrases. Toutes les paires de phrases possibles d'un corpus « non-parallèle »ont été considérées, et celles ayant un niveau de similarité supérieur à un certain seuil sont conservées pour construire un dictionnaire qui sera réappris itérativement.

Une méthode d'extraction des segments de phrases parallèles est présentée par Munteanu et Marcu (2005). Un dictionnaire bilingue existant est utilisé pour traduire chaque document en langue source vers la langue cible afin d'extraire le document cible qui correspond à cette traduction. Pour chaque paire de documents, des paires de phrases et de segments parallèles sont extraites en utilisant un lexique de traduction et un classifieur à maximum d'entropie pour le choix final des phrases parallèles. Rauf et Schwenk (2011) présentent une technique similaire à celle de Munteanu et Marcu (2005). Les différences majeures résident dans l'utilisation d'un système de TA statistique à la place du dictionnaire bilingue, et dans l'utilisation de mesures d'évaluation, comme le taux d'erreur mot (WER) ou le taux d'édition de la traduction (TER), pour évaluer le degré de parallélisme des phrases extraites.

Toutes ces méthodes sont présentées comme des techniques efficaces pour extraire des données parallèles à partir d'un corpus comparable. Certains travaux exploitent la modalité audio pour l'extraction de données parallèles. Paulik et Waibel (2009) ont montré que les modèles de traductions statistiques peuvent être appris automatiquement d'une manière non-supervisée à partir des données parallèles audio. Dans notre contexte de travail, nous nous intéressons à l'exploitation des corpus comparables multimodaux avec différents niveaux de similitude. La multimodalité concernera l'utilisation de documents textuels et audio.

3 Architecture générale

Notre but est d'exploiter les données comparables multimodales afin d'en extraire des données parallèles nécessaires pour construire, adapter et améliorer nos systèmes de traduction automatique statistique. L'architecture générale de notre approche, qui se résume en 3 étapes, est décrite dans la figure 1.
Notre corpus comparable multimodal est constitué de données audio en langue source L1 et de données textuelles en langue cible L2. Les données audio sont tout d'abord transcrites par un système de Reconnaissance Automatique de la Parole (RAP). Ce système produit une hypothèse

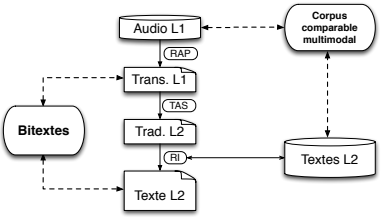

FIGURE 1 – Architecture générale du système

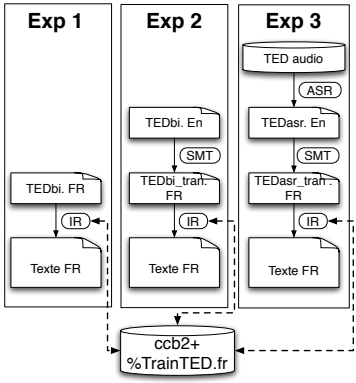

FIGURE 2 – Expériences permettant de mesurer l'impact des différents modules mis en jeu sur le corpus bilingue extrait.

de transcription qui est ensuite traduite par le système TAS. La meilleure hypothèse de traduction est utilisée comme requête dans le système de recherche d'information (RI), dont le corpus indexé correspond à la partie textuelle en langue cible du corpus comparable multimodal. Dans cette approche, qui se base sur les travaux de Rauf et Schwenk (2011), nous utilisons le logiciel libre Lemur (Ogilvie et Callan, 2001) pour effectuer la RI. Au final, nous obtenons un bitexte constitué d'une part de la transcription automatique et d'autre part du résultat de la RI, qui pourra être réinjecté dans le système de base.

Ce cadre de travail soulève toutefois plusieurs problèmes. Chaque module mis en jeu pour la traduction de la parole introduit un certain nombre d'erreurs. Il est donc important de mettre en évidence la faisabilité de l'approche ainsi que l'impact de chaque module sur les données générées. Pour cela, nous avons effectué 3 types d'expérience différents, décrits dans la figure 2. Le premier type d'expérience (*Exp 1*) consiste à utiliser la référence de traduction comme requête pour la RI. Ce cas est le plus favorable, cela simule le fait que les modules de RAP et de TAS ne commettent aucune erreur. Le second type d'expérience (*Exp 2*) utilise la référence de transcription pour alimenter le système de traduction automatique. Cela permet de mettre en évidence l'impact des erreurs de traduction. Enfin, le troisième type d'expérience (*Exp 3*) met en œuvre l'architecture complète décrite ci-dessus. Cela correspond au cas réel auquel nous sommes confrontés.

Une autre problématique concerne l'importance du degré de similitude (*comparabilité*) des corpus comparables utilisés. Nous avons donc artificiellement créé des corpus comparables plus ou moins ressemblants en intégrant une quantité plus ou moins grande (*25%, 50%, 75% et 100%*) de données du domaine dans le corpus indexé par la RI.

Les résultats de la RI ne sont pas toujours satisfaisants, il est donc nécessaire de filtrer ces résultats afin de ne pas ajouter de phrases non parallèles dans le bitexte final. Nous considérons le Taux d'Edition de la Traduction (*Translation Edit Rate* - TER) calculé entre les phrases retournées par la RI et la requête, comme mesure de filtrage des phrases trouvées. Les phrases ayant un TER

bitextes	# de mots	du domaine ?
nc7	3,7M	non
eparl7	56,4M	non
ccb2_px70	1,3M	non
TEDasr	1,8M	oui
TEDbi	1,9M	oui

TABLE 1 – Données utilisées pour l'apprentissage des systèmes de traduction automatique.

Dev	# de mots
dev.outASR	36k
dev.refSMT	38k

Test	# de mots
tst.outASR	8,7k
tst.refSMT	9,1 k

TABLE 2 – Données de développement (Dev) et de Test.

supérieur à un certain seuil (déterminé empiriquement) sont exclues.

Dans tous les cas, l'évaluation de l'approche est nécessaire. Ainsi, les données parallèles extraites sont réinjectées dans le système de base, qui est ensuite utilisé pour traduire les données de test à nouveau. L'évaluation peut ensuite se faire avec une mesure automatique comme BLEU (Papineni *et al.*, 2002).

4 Expériences et résultats

Pour nos expériences, nous exploitons les données de la campagne d'évaluation IWSLT'11 dans laquelle des données bilingues multimodales sont disponibles. Cette tâche, détaillée dans Rousseau *et al.* (2011), consiste à traduire des discours de TED [1] de l'anglais vers le français. Le système de RAP est appris sur 773 discours représentant 118 heures de parole. Les données de développement et de test officielles sont utilisées pour évaluer notre approche.

Le corpus de développement est composé de 19 discours représentant un peu plus de 4 heures de parole. Les corpus bilingues suivants sont utilisés pour l'apprentissage des modèles de traduction : News-Commentary version 7(*nc7*), le corpus des actes du parlement européen (*eparl7*) et le corpus Gigaword_EnFr (ccb2_px70). [2] De ce dernier, ne sont conservées que les paires de phrases dont la perplexité du côté cible (calculée avec le modèle de langue utilisé pour le système TAS) est inférieure à un seuil (ici 70). Le détail des données disponibles est présenté dans le tableau 1.

Le système de reconnaissance de la parole utilisé est basé sur le système libre CMU Sphinx (*version 3 et 4*), modifié et amélioré. Le système anglais qui été développé pour transcrire les données audio de TED utilise cinq passes similaires à celui du français décrit dans Deléglise *et al.* (2009). Les systèmes de traduction mis en œuvre sont fondés sur Moses (Koehn *et al.*, 2007), approche par segments (*phrase-based*). Le modèle de langue est un modèle 4-gramme construit avec l'outil SRILM (Stolke, 2002). Nous avons utilisé toutes les données monolingues disponibles et le côté cible des bitextes.

Comme mentionné précédemment, le score *TER* est utilisé comme métrique de filtrage des phrases résultantes de la RI, c'est-à-dire que les phrases ayant un TER supérieur à un certain seuil ne sont pas conservées. Ce seuil est déterminé expérimentalement. Pour cela, nous avons filtré les corpus extraits dans les différentes conditions d'expérimentation avec différents seuils *TER* (de 0 à 100). Pour chaque seuil TER nous obtenons un nombre de phrases parallèles. Le corpus obtenu est ajouté aux données d'entraînement du système de base (*eparl7* et *nc7*) pour

1. http ://www.ted.com/
2. Ces corpus sont librement disponibles sur le site de la campagne IWSLT'11 et WMT'11.

obtenir le système adapté. Les résultats en terme de score *BLEU* sur le corpus de développement obtenus avec les différents systèmes adaptés sont présentés dans la figure 3.

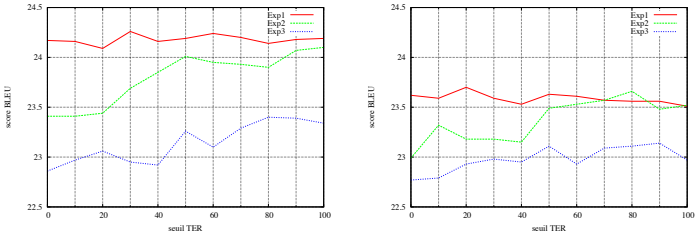

FIGURE 3 – Score BLEU de la traduction du Dev en utilisant les systèmes adaptés avec les bitextes correspondant à différents seuils TER, extraits d'un corpus d'index constitué par *ccb2 + 100% TEDbi* (à gauche) et *ccb2+ 25% TEDbi* (à droite).

Ces résultats montrent que le choix du seuil de *TER* adéquat dépend de la nature des données. En effet, pour la condition de l'*Exp1* où les requêtes de la RI sont sans erreur, nous remarquons que le meilleur résultat est obtenu pour un seuil proche de 0. Dans les deux autres conditions (*Exp2* et *Exp3*), le meilleur seuil est dans l'intervalle [80-90]. Dans nos expériences, nous retiendrons le seuil de 80 pour le filtrage des résultats de la RI.

Dans l'expérience *Exp2*, les traductions automatiques sont utilisées en tant que requêtes pour la RI. On peut espérer que la RI elle-même n'est pas trop affectée par les erreurs de traduction, mais ceci influence bien sûr le filtrage basé sur le score TER. Nous n'avons pas observé un maximum du score BLEU en fonction du seuil sur le score TER - dans nos expériences les performances semblent augmenter de façon continue. Néanmoins, afin de limiter l'impact des phrases bruitées, nous avons choisi un seuil de 70. On peut observer que le score BLEU du système adapté est très proche de celui de *Exp1*. Ainsi, nous pouvons conclure que les erreurs commises par la TAS n'ont pas une influence importante sur l'algorithme d'extraction des phrases parallèles. Ceci confirme l'analyse de (Rauf et Schwenk, 2011).

Notre système de base entraîné avec des données génériques obtient un score BLEU de 22,93. Dans *Exp1*, nous utilisons les traductions de référence en tant que requêtes et la RI devrait en principe trouver toutes les phrases avec un TER de zéro. Les figures montrent que la RI fonctionne comme attendu : l'amélioration du score BLEU ne dépend pas du seuil sur le score TER puisque la plupart des phrases ont effectivement un score TER de zéro. L'amélioration du score BLEU dépend bien sûr de la quantité de données extraites : le score BLEU augmente de 22,93 à 24,14 lorsque 100% des données ont été injectées, alors que nous n'obtenons que 23,62 avec 20% des données TED. Ces résultats nous donnent une borne supérieure des résultats envisageables avec l'utilisation d'un corpus multimodal.

Finalement, dans *Exp3*, la RAP est utilisée dont le taux d'erreur est d'environ 18%. Les phrases extraites du corpus multimodal permettent d'améliorer le système de traduction : le score BLEU n'est que 0,5 points en dessous de celui obtenu dans *Exp1* ou *Exp2*. Les résultats obtenus après adaptation du système de base sont présentés dans le tableau 4. Dans ce cas, le corpus indexé par la RI est constitué des corpus *ccb2_px70* et *TEDbi* (100%).

	Phrase extraite
Français	vous allez chez ibm et vous prenez un superordinateur ...
Anglais	you get a supercomputer because they know ...
	Test audio
Sortie ASR	a supercomputer has calculated that humans and only ...
Référence	a supercomputer has calculated that humans have only ...
	Traductions de la sortie ASR
Système de base	un supercomputer a calculé que les humains et seulement ...
Système adapté	un superordinateur a calculé que les humains et seulement ...
Référence	un superordinateur a calculé que les humains n' avaient plus que ...
	Traductions améliorées
Sys de base	j'ai écrit un article sur la nourriture génétiquement modifiée
Sys adapté	j'ai écrit un article sur les produits alimentaires génétiquement modifiés
Sys de base	yeah tu as raison de réparer
Sys adapté	euh oui tu as raison il faut réparer

TABLE 3 – Exemples d'amélioration du système de base : vocabulaire enrichi à partir des phrases parallèles extraites dans la condition *Exp3*.

Le tableau 5 présente les résultats des systèmes adaptés en fonction du degré de similitude du corpus comparable, dans les conditions d'expérimentation *Exp3*. Des exemples d'adaptation sont présentés dans le tableau 3. Nous pouvons remarquer que le degré de similitude est un facteur important. Un résultat attendu est que lorsque nous augmentons la proportion de corpus du

Expérience	Dev	Test
Système de base	22,93	23,96
Exp1	24,14	25,14
Exp2	23,90	25,15
Exp3	23,40	24,69

TABLE 4 – % BLEU obtenus sur le Dev et Test après l'ajout des bitextes extraits au système de base, dans les conditions *Exp1, Exp2 et Exp3*.

Expérience	Dev	Test	# mots
Système de base	22,93	23,96	-
25% TEDbi	23,11	24,40	∼110k
50% TEDbi	23,27	24,58	∼215k
75% TEDbi	23,43	24,42	∼293k
100% TEDbi	23,40	24,69	∼393k

TABLE 5 – Résultats (%BLEU) obtenus avec les systèmes adaptés lorsque le degré de similitude du corpus comparable varie.

domaine dans le corpus indexé, les performances sont meilleures. Il est important de noter que lorsque les corpus sont moins similaires, le nombre de phrases conservé est réduit drastiquement par le filtrage, et donc l'impact de l'adaptation est plus faible. Sans filtrage, les performances du système de base peuvent être dégradées.

5 Conclusion

Dans ce travail nous avons proposé une méthode permettant d'extraire des textes parallèles à partir de corpus comparables multimodaux (audio et texte) pour adapter et améliorer des systèmes de traduction automatique statistique. Plusieurs modules sont utilisés pour extraire du

texte parallèle : reconnaissance automatique de la parole, traduction automatique et recherche d'information. Nous validons notre méthode en injectant les données produites dans l'apprentissage de nouveaux systèmes de TAS. Des améliorations en termes de BLEU sont obtenues pour différents cadres expérimentaux. Il en ressort que l'enchaînement des modules ne dégrade que faiblement les résultats, mais le filtrage des résultats de la RI est nécessaire. Le degré de similitude du corpus comparable est un facteur important qu'il faudra prendre en compte lorsque cette architecture sera exploitée dans des conditions réelles.

Remerciements

Ces recherches ont été financées par la région des Pays de la Loire sous le projet DEPART[3].

Références

Deléglise, P., Estève, Y., Meignier, S. et Merlin, T. (2009). Improvements to the LIUM french ASR system based on CMU Sphinx : what helps to significantly reduce the word error rate ? *In Interspeech 2009*.

Do, T. N. D., Besacier, L. et Castelli, E. (2010). Apprentissage non supervisé pour la traduction automatique : application à un couple de langues peu doté. *TALN 2010*.

Fung, P. et Cheung, P. (2004). Multi-level bootstrapping for extracting parallel sentences from a quasi-comparable corpus. *In Proceedings of COLING '04*.

Hewavitharana, S. et Vogel, S. (2011). Extracting parallel phrases from comparable data. *In Proceedings of the 4th Workshop on Building and Using Comparable Corpora : Comparable Corpora and the Web*, BUCC '11, pages 61–68.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A. et Herbst, E. (2007). Moses : open source toolkit for statistical machine translation. *In Proceedings of ACL'07*, pages 177–180.

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

Ogilvie, P. et Callan, J. (2001). Experiments using the lemur toolkit. *Procedding of the Trenth Text Retrieval Conference (TREC-10)*.

Papineni, K., Roukos, S., Ward, T. et Zhu, W.-J. (2002). Bleu : a method for automatic evaluation of machine translation. *In Proceedings of ACL '02*, pages 311–318.

Paulik, M. et Waibel, A. (2009). Automatic translation from parallel speech : Simultaneous interpretation as mt training data. *ASRU*.

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

Resnik, P. et Smith, N. A. (2003). The web as a parallel corpus. *Comput. Linguist.*, 29:349–380.

Rousseau, A., Bougares, F., Deléglise, P., Schwenk, H. et Estève, Y. (2011). LIUM's systems for the IWSLT 2011 speech translation tasks. *In Proceedings of IWSLT'11*.

Stolke, A. (2002). Srilm - an extensible language modeling toolkit. *ICSLP*, pages 901–904.

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

Zhao, B. et Vogel, S. (2002). Adaptive parallel sentences mining from web bilingual news collection. *In Proceedings of IEEE International Conference on Data Mining*, page 745.

3. http://www.projet-depart.org/