PRES LUNAM
Ecole Doctorale STIM
Sciences et Technologies de l'Information et Mathématiques

Spécialité : Informatique
Laboratoire : Laboratoire d'informatique de l'université de maine (LIUM)
Equipe : Language and Speech Technology (LST)

Traduction automatique statistique des documents multimodaux

AFLI, Haithem
Mél : haithem.afli@lium.univ-leman.fr

Résumé : La taille des données textuelles et sonores augmente continuellement au fil du temps. Ces ressources peuvent être exploitées pour l'amélioration des systèmes de traduction automatique statistique. Dans ce contexte multimédia, nous visons en particulier à utiliser les documents audio et textes comme ressources pour construire un système spécialisé ou adapter un système générique à un domaine spécifique. Ce papier porte sur l'extraction de données parallèles à partir de corpus multimodaux dans le but d'améliorer des systèmes de TAS. Plusieurs expériences ont été menées sur les données de la campagne IWSLT'11 (TED) qui montrent la faisabilité de notre approche. Nous étudions l'influence de la qualité des données et les différences des tâches dans l'amélioration des résultats.

Mots clés : *traduction automatique statistique, corpus multimodal bilingue, extraction de phrases parallèles.*

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 des textes n'est pas approprié. Il y a relativement peu de paires de langues pour lesquelles des corpus parallèles de tailles raisonnables sont disponibles comme l'anglais, le français, l'espagnol, l'arabe, le chinois et quelques langues européennes [1]. De plus, ces corpus disponibles viennent principalement des 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 [2]. 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 [3]. Ces corpus comparables peuvent couvrir différents sujets.

Dans notre contexte de travail, nous nous intéressons à l'exploitation des corpus comparables multimodaux avec différents niveaux de similitude. La multimodalité dans notre cas concernera l'utilisation de documents textuels et audio.

Contexte

Ces travaux s'inscrivent dans le cadre du projet DEPART [1] (Documents Écrits et PAroles - Reconnaissance et Traduction) dont l'un des objectifs est l'exploitation de données multimodales et multilingues pour la traduction automatique. Nous considérons le cas, assez fréquent pour des domaines spécifiques, où un manque de données textuelles peut être pallié par l'exploitation de données audio, et où le manque de données parallèles peut être amendé par des corpus comparables. La question que nous nous posons

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

alors est de savoir si un corpus comparable multimodal comme celui de la figure 1 permet d'apporter des solutions au problème du manque de données parallèles. Dans ce travail nous proposons une méthode pour l'utilisation des corpus comparables multimodaux, en se limitant aux modalités texte et audio, pour l'extraction de données parallèles comme présenté dans la figure 2.

FIGURE 1 – Exemple de ressource de données comparables multimodaux en domaine du sport à partir du site Euronews

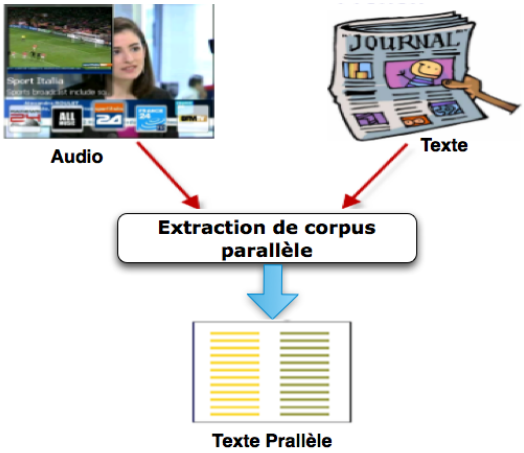

FIGURE 2 – Objectif de l'exploitation des corpus comparables multimodaux.

2 Méthode proposée

2.1 Architecture générale

Notre but est de trouver une méthode pour 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 3.

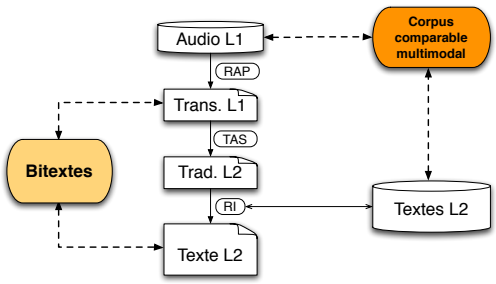

FIGURE 3 – Architecture générale du système

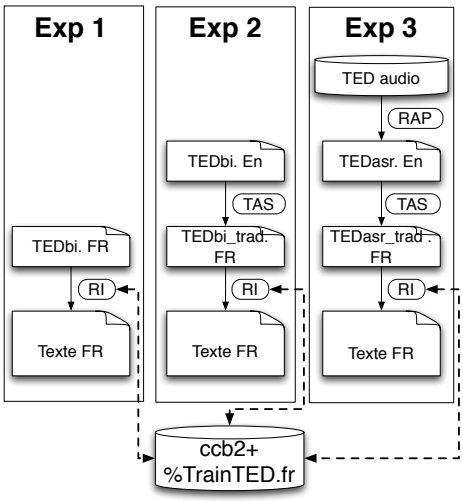

FIGURE 4 – Expériences permettant de mesurer l'impact des différents modules mis en jeu sur le corpus bilingue extrait.

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 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'informations (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 [4], nous utilisons le logiciel libre Lemur [5] 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.

2.2 Problématique

Cette méthode soulève plusieurs problèmes. Chaque module mis en place pour la traduction de la parole introduit un certain nombre d'erreurs. Il est 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 trois types d'expériences, décrits dans la figure 4. 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 car 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 oeuvre l'architecture complète décrite ci-dessus. Cela correspond au cas réel auquel nous avons été 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, puisque nous pouvons avoir de fausses traductions retournées par le système de RI compte tenu du fait que la recherche se fait dans un corpus comparable qui contient probablement des traductions des requêtes originales. Il est donc nécessaire de filtrer ces résultats afin de ne pas ajouter de phrases non parallèles dans le bitexte final. J'ai ensuite considéré, comme métrique de filtrage des phrases trouvées, le Taux d'Edition de la Traduction (*Translation Edit Rate* - TER) , calculé entre les phrases retournées par la RI et la requête. Les phrases ayant un TER supérieur à un certain seuil, qui est déterminé empiriquement, sont exclues.

L'évaluation de l'approche était nécessaire. Les données parallèles extraites sont donc réinjectées dans le système de traduction de base, qui est ensuite utilisé pour traduire les données de test. L'évaluation peut alors se faire avec une mesure automatique comme le score *BLEU* [6].

3 Expériences et résultats

Dans nos expériences, nous avons exploité les données de la campagne d'évaluation *IWSLT'11*[2] au sein de laquelle des données bilingues multimodales sont disponibles. Cette tâche, détaillée dans [7], consiste à traduire des discours de *TED*[3] de l'anglais vers le français.

Les résultats détaillés dans [8] montrent des améliorations en qualité de traduction de notre système TAS. Ce système est fondé sur Moses [9], approche par segments (*phrase-based*), et est construit suivant la méthode décrite dans [7]. Ces résultats, comme présentés dans le tableau 1, ont montré l'intérêt de l'approche pour exploiter des documents multimodaux dans le contexte de traduction automatique. Nous pouvons remarquer une amélioration du système de base dans tous les cas d'expériences, notamment dans *Exp3* où tous les modules sont enchaînés. Le tableau 2 présente les résultats des systèmes adaptés en fonction du degré de similitude du corpus comparable, dans les conditions d'expérimentation *Exp3*. Nous remarquons que lorsque nous augmentons la proportion de corpus du 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 conservées est réduit drastiquement par le filtrage. L'impact de l'adaptation est donc plus faible.

4 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

2. http://www.iwslt2011.org/
3. Technology Entertainment Design : http://www.ted.com/

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 1 – % BLEU obtenus sur le Dev et Test de IWSLT'11 après l'ajout des bitextes extraits au système de base dans les conditions *Exp1, Exp2 et Exp3.*

Expérience	Dev	Test	# mots injectés
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 2 – Résultats (%BLEU) obtenus avec les systèmes adaptés lorsque le degré de similitude du corpus comparable varie.

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.

Plusieurs extensions de l'approche sont actuellement étudiées, comme l'approfondissement de l'analyse des meilleures techniques de filtrage, l'expérimentation de l'influence du système SMT initial utilisé dans l'extraction notamment le domaine des données d'apprentissage et l'évaluation du degré de similitude des corpus comparables initiales.

Références

[1] Sanjika Hewavitharana and Stephan Vogel. 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, 2011.

[2] Philip Resnik and Noah A. Smith. The web as a parallel corpus. *Comput. Linguist.*, 29 :349–380, September 2003.

[3] Pascale Fung and Percy Cheung. Multi-level bootstrapping for extracting parallel sentences from a quasi-comparable corpus. In *Proceedings of the 20th international conference on Computational Linguistics*, COLING '04, 2004.

[4] Sadaf Abdul-Rauf and Holger Schwenk. Parallel sentence generation from comparable corpora for improved smt. *Machine Translation*, 2011.

[5] Paul Ogilvie and Jamie Callan. Experiments using the lemur toolkit. *Procedding of the Trenth Text Retrieval Conference (TREC-10)*, 2001.

[6] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu : a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, 2002.

[7] Anthony Rousseau, Fethi Bougares, Paul Deléglise, Holger Schwenk, and Yannick Estève. LIUM's systems for the IWSLT 2011 speech translation tasks. *International Workshop on Spoken Language Translation 2011*, 2011.

[8] Haithem Afli, Loic Barrault, and Holger Schwenk. Parallel texts extraction from multimodal comparable corpora. In *JapTAL*, volume 7614 of *Lecture Notes in Computer Science*, pages 40–51. Springer, 2012.

[9] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. Moses : open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180, 2007.