Traduction de la parole dans le projet RAPMAT

Hélène Bonneau-Maynard[1] Natalia Segal[1] Eric Bilinski[1] Jean-Luc Gauvain[1]
Li Gong[1] Lori Lamel[1] Antoine Laurent[1] François Yvon[1]
Julien Despres[2] Yvan Josse[2] Viet Bac Le[2]
(1) LIMSI-CNRS B.P. 133 91403 Orsay Cedex
(2) Vocapia Research 28, rue Jean Rostand Parc Orsay Université 91400 Orsay
{maynard,segal,bilinski,laurent,gauvain,gong,lamel,yvon}@limsi.fr,
{despres,joss,levb}@vocapia.com

RÉSUMÉ

Le projet RAPMAT vise à développer des systèmes de traduction de la parole en s'intéressant aux deux traitements constitutifs de la chaîne complète : la reconnaissance de la parole (RAP) et la traduction (TA). Dans la situation classique, les modèles statistiques utilisés par les deux systèmes sont estimés indépendemment, à partir de données de différentes natures (transcriptions manuelles de données de parole pour la RAP et corpus bilingues issus de données textuelles pour la TA). Nous proposons une approche semi-supervisée pour l'adaptation des modèles de traduction à la traduction de parole, dans laquelle les modèles de TA sont entraînés en intégrant des transcriptions manuelles et automatiques de la parole traduites automatiquement. L'approche est expérimentée sur la direction de traduction français vers anglais. Un prototype de démonstration sur smartphones, incluant notamment la traduction de parole pour les paires de langues français/anglais et français/chinois a été développé pour permettre la collecte de données.

ABSTRACT

Speech translation in the RAPMAT project

The goal of the RAPMAT project is to develop automatic speech-to-speecj translation systems focusing on the interconnection of its two main processing steps : automatic speech recognition (ASR) and machine translation (MT). The basic approach consists of training independent statistical models for each of the systems, using data of different nature : manual transcriptions of speech are used for ASR, whereas the MT models are trained using bilingual corpora of written texts. In this paper we present a semi-supervised approach for adaptation of the translation models to data produced by ASR, where the training corpus includes machine translated automatic and manual transcriptions of speech. This approach was tested on translation from French to English. An application prototype for smartphones has been developed for speech translation of French/English and French/Chinese language pairs that will also serve to collect additional data.

MOTS-CLÉS : reconnaissance de la parole, traduction automatique.

KEYWORDS: speech recognition, automatic translation.

1 Introduction

La traduction de la parole est un des défis sociétaux majeurs dont l'objectif est de permettre la communication parlée entre personnes s'exprimant dans des langues différentes. Le défi est également scientifique puisqu'il se situe à la jonction de deux domaines de recherche que sont la reconnaissance automatique de la parole (RAP) (Gauvain et Lamel, 2008) et la traduction automatique (TA) (Allauzen et Yvon, 2012). Si ces deux domaines ont connu dans les dernières décennies des progrès considérables grâce à l'introduction de modèles statistiques, il n'existe pas encore de solution satisfaisante pour leur couplage (Stüker *et al.*, 2012).

Le projet DGA RAPMAT (Reconnaissance Automatique de la Parole Multilingue Accentuée pour la Traduction) vise la réalisation de systèmes de traduction de la parole pour rendre cette technologie accessible à des usagers en déplacement, via un smartphone. Un prototype de démonstration été développé sur les plate-formes iOS et Androïd permettant d'expérimenter les fonctionnalités proposées. Concernant la traduction, c'est l'anglais, le chinois et le vietnamien qui sont traités, l'objectif étant d'offrir leur traduction depuis et vers le français. Trois autres langues sont traitées dans le projet.

Après une brève présentation du projet RAPMAT, l'article se focalise sur les aspects traduction de la parole. La section 3 situe les travaux dans le contexte scientifique. Les caractéristiques du système de reconnaissance de la parole sont décrites dans la section 4. La section 5 décrit les approches proposées pour l'adaptation des modèles de traduction à la traduction de la parole. La description des données et les aspects expérimentaux sont décrits dans les parties 6.1 et 6.

2 Le projet RAPMAT

Le projet DGA RAPMAT (24 mois, mai 2012 - mai 2014), dont les partenaires sont le LIMSI-CNRS et la société Vocapia Research [1], vise à développer des systèmes de traduction de la parole en s'intéressant aux deux traitements constitutifs de la chaîne complète : la reconnaissance de la parole puis la traduction de transcriptions. Par ailleurs, la participation du LIMSI au consortium U-STAR [2] permet de bénéficier de ses infrastructures de collecte de données offrant l'accès à des corpus dans de nombreuses langues.

En reconnaissance de la parole, l'accent est mis sur le développement de systèmes en vue de la traduction. Les langues pour lesquelles il est prévu de réaliser un système de transcription automatique seront dans un premier temps le français et l'anglais (avec adaptation pour la prise en compte d'accents non natifs, essentiellement asiatiques). Au total 7 langues sont traitées dans ce projet, Concernant la traduction, l'anglais et le chinois ont été traités en priorité, l'objectif étant d'offrir leur traduction depuis et vers le français. La construction des systèmes dépend de la quantité de données disponibles. Afin de mettre à jour les modèles de transcription et de traduction, deux ensembles de corpus multilingues de type "guide de conversation et requêtes touristiques" ont été utilisés au sein du projet : U-STAR (corpus HIT) et BTEC. Ces données ont servi à l'adaptation des modèles existants. Pour les langues pour lesquelles Vocapia Research et le LIMSI ne disposent pas encore de système, des données d'apprentissage ont été collectées

1. http ://www.vocapia.com
2. http ://www.ustar-consortium.com

sur internet (transcriptions approximatives ou news). Ces données ont permis de faire un apprentissage peu supervisé des modèles, évitant ainsi des coûts prohibitifs de transcription manuelle.

3 Traduction de la parole : État de l'art

Il existe plusieurs façons de coupler le système de reconnaissance de la parole avec le système de traduction automatique pour faire la traduction de la parole. L'approche découplée la plus élémentaire consiste à donner en entrée du système de traduction la séquence de mots la plus probable issue du système de reconnaissance de la parole, sans se soucier de l'absence d'adéquation entre les modèles qu'utilisent ces deux systèmes.

Plusieurs manières d'améliorer ce couplage simpliste ont été explorées dans la littérature, notamment à travers les évaluations IWSLT (Mauro *et al.*, 2013) :

– L'approche semi-découplée : permet au module de RAP de produire plusieurs hypothèses (n-meilleures solutions, treillis, réseaux de confusions) laissant ainsi au système de traduction la possibilité de choisir l'hypothèse de transcription qui produira la traduction la plus probable (Matusov et Ney, 2011). Cette approche, qui permet souvent d'obtenir des améliorations globales de performances, est une extension de l'approche découplée, car le système de traduction et le système de reconnaissance sont entraînés indépendamment et prennent leurs décisions d'une façon indépendante.
– L'approche intégrée : compose le modèle acoustique et le modèle de traduction en un système de décodage à un seul passage (Perez *et al.*, 2012). Cette approche impose une contrainte sur le modèle de traduction qui doit être implanté sous la forme d'un transducteur fini pour pouvoir réaliser la composition des modèles.
– L'adaptation du modèle de traduction qui consiste à modifier la partie source des corpus d'entraînement des modèles de traduction pour la rendre compatible avec la sortie produite par le système de reconnaissance de la parole.

La dernière de ces solutions, s'est avérée efficace pour améliorer la qualité de la traduction (Peitz *et al.*, 2012) ; elle nécessite toutefois une grande quantité de données orales pour lesquelles on disposerait également de la traduction manuelle de référence. Ces données sont coûteuses à obtenir, surtout pour un type d'application particulier tel que le dialogue. En revanche, il est plus facile d'obtenir, d'un côté, des données de parole pour lesquelles la transcription manuelle de référence est disponible, et de l'autre côté, des traductions manuelles de textes écrits. Nous pouvons ainsi entraîner un système de traduction de base sur des textes écrits (apprentissage supervisé) et ensuite enrichir ce système en utilisant les traductions automatiques faites à partir des transcriptions de référence de la parole (apprentissage semi-supervisé). Des approches utilisant l'apprentissage semi-supervisé ont déjà été proposées pour adapter un système de traduction automatique à un domaine en utilisant comme corpus d'adaptation les traductions automatiques faites par le système de base (Schwenk, 2008; Gahbiche-Braham *et al.*, 2011) et pour la portabilité d'un système de dialogue d'une langue vers une autre (Jabaian *et al.*, 2013). Nous présentons une adaptation similaire pour les données issues de la reconnaissance de la parole.

4 Reconnaissance de la parole

Le système de transcription de la parole pour le français, développé conjointement par le LIMSI et Vocapia Research, repose sur deux composants principaux : un partitioneur parole/non-parole et un décodeur de mots. Le décodeur de mots utilise un modèle acoustique HMM et un modèle de langage bi-gramme pour construire un treillis de mots qui est ensuite réévalué par modèle de langue quadri-gramme. Le système produit alors une transcription automatique, alignée temporellement au niveau des mots, dans un fichier XML.

Modèles acoustiques : L'apprentissage des modèles acoustiques est réalisé en alignant une transcription orthographique exacte sur le signal de parole au moyen d'un jeu de modèles phonétiques et d'un lexique de prononciation. Ces modèles sont entraînés à partir de paramètres MLP (Multi-Layer Perceptron) concaténés avec des paramètres PLP (Perceptual Linear Prediction) et la fréquence fondamentale. Ils ont bénéficié d'un apprentissage discriminant (MMIE) et adapté aux locuteurs (SAT). Une adaptation par blocs diagonaux a été utilisé pour les paramètres MLP et PLP+F0. Le vecteur MLP+PLP+f0 est composé de 81 paramètres ($39MLP + 39PLP + F0 + \Delta F0 + \Delta\Delta F0$). Les modèles acoustiques sont des HMMs (Hidden Markov Models) composés d'un mélange de gaussiennes. Ils sont à états liés, dépendants du contexte et du sexe du locuteur. Le silence est modélisé par un seul état composé de 1024 Gaussiennes. Au final, à partir d'un corpus de 866 heures de parole, 19679 modèles contextuels ont été entraînés, partageant 11517 états, pour un total de 370k gaussiennes.

Modèles de langue : 2.9 milliards de mots ont été utilisés pour entraîner des modèles de langue génériques. Les modèles sont construits autour d'un vocabulaire de 200K mots. Les modèles génériques ont été adaptés pour le projet RAPMAT en les interpolant avec des modèles spécifiques construits à partir des données des corpora BTEC et HIT décrits dans le paragraphe 6.1.

Modèle de prononciations : Les prononciations du dictionnaire sont obtenues par une phonétisation automatique (plus de 1000 règles), plusieurs possibilités de phonétisation sont générées pour les noms propres. Une passe de correction manuelle a été faite pour réduire les cas de prononciations multiple ce qui permet d'obtenir un gain en vitesse lors du décodage. Les probabilités de prononciations, obtenues par l'observation des prononciations, sont ajoutées au dictionnaire lors de l'apprentissage acoustique (Gauvain *et al.*, 2005).

5 Adaptation des modèles de traduction

Comme exposé ci-dessus (section 3) la traduction de parole peut bénéficier d'un couplage plus étroit entre les deux modules, en particulier pour que la traduction prenne d'une part en compte certaines spécificités du langage parlé, et d'autre part les erreurs produites par la transcription automatique. Notre approche pour réaliser ce couplage s'inspire de techniques d'adaptation au domaine selon le principe suivant.

Dans un premier temps, nous rapprochons les données d'entraînement pour le modèle de traduction des données en sortie de reconnaissance en enlevant toute la ponctuation de la source

dans le corpus d'apprentissage. Par la suite, l'adaptation du modèle de traduction aux données RAP est effectuée de deux manières différentes :

– Un petit corpus de parole, pour lequel la traduction manuelle de référence est disponible, est utilisé pour faire le réglage du modèle de traduction de base.
– Un grand corpus de parole, pour lequel uniquement la transcription manuelle de référence est disponible, est traduit automatiquement avec le système de traduction de base. Les traductions automatiques ainsi obtenues, alignées à la fois avec la transcription de référence et avec les sorties de RAP, sont utilisées pour l'entraînement d'un nouveau modèle de traduction plus adapté aux données de la parole.

6 Expériences et résultats

6.1 Données

Les corpus parallèles du tableau 1 ont été utilisés pour entraîner et développer le modèle de traduction de base. Notons que ces corpus sont de natures très différentes, le corpus HIT avec des phrases de dialogues courtes étant le mieux adapté à l'application visée. Les corpus parallèles contiennent des signes de ponctuation produits manuellement.

– BTEC (Kikui *et al.*, 2006) : corpus de dialogues autour des transports et du tourisme.
– HIT (Yang *et al.*, 2006) (Harbin Institute of Technology) : corpus de dialogues couvrant 5 domaines (voyages, restauration, sports, transports et business). Le corpus initial trilingue (anglais, chinois, japonais) est enrichi dans le cadre du projet U-STAR. Le LIMSI en a produit la traduction en français. Les langues disponibles à ce jour sont : français, anglais, chinois, japonais, vietnamien et hindi.
– Tatoeba : corpus multilingue couvrant des domaines divers provenant de traductions collaboratives en ligne [3], utilisé uniquement pour la paire français/anglais.
– TedTalk [4] : corpus de transcriptions et de traductions d'exposés vidéos, utilisé uniquement pour la paire français/chinois.

Pour l'adaptation semi-supervisée des modèles de traduction, un corpus d'actualités radio-télévisées monolingue en français (BN (Lamel et al, 2011)) a été transcrit automatiquement par le système de RAP (dernière ligne du tableau 1), avec un taux d'erreur mots de 12%. Le corpus BN contient des signes de ponctuation produits automatiquement par le système de reconnaissance de la parole.

Deux types de corpus ont été utilisés pour le développement et les tests : les corpus dev 2004 (500 phrases) et test 2009 BTEC (469 phrases) des évaluations IWSLT, et les corpus dev (436 phrases) et test (437 phrases) Tatoeba pour laquels on dispose du signal de parole en français de sa transcription manuelle, et de sa traduction en anglais.

3. http ://tatoeba.org
4. https ://wit3.fbk.eu/archive/2012-02//texts/zh-cn/fr/zh-cn-fr.tgz

Corpus	Lignes	Tokens FR	Tokens EN	Tokens ZH
BTEC	19 972	202 559	177 370	N/A
HIT	62 226	671 694	597 528	588 501
Tatoeba	111 168	1 038 548	928 936	N/A
TedTalk	142 296	2 925 803	N/A	2 614 332
BN	487 856	9 843 627	N/A	N/A

TABLE 1 – Statistiques des corpus.

6.2 Performances des systèmes de traduction de base

Pour l'ensemble des systèmes développés, les données parallèles sont alignées mot à mot avec MGgiza++ [5]. Le décodeur Moses [6] est ensuite utilisé pour symétriser les alignements en utilisant l'heuristique grow-diag-final-and et pour extraire les phrases avec une longueur maximale de 7 mots. Les poids des paramètres sont optimisés par MERT sur le corpus de développement. La traduction est réalisée par le décodeur Moses fonctionnant en mode "serveur".

Les systèmes de traduction de base ont été développés pour les paires français/anglais et français/chinois. L'ensemble des textes parallèles a été tokenisé côté source pour les rendre totalement compatibles avec la tokenisation des sorties du système de RAP.

Dans les textes chinois, toutes les ponctuations sont normalisées en ponctuations chinoises. Les chiffres et les lettres alphabétiques sont codés en ASCII. Au niveau de l'écriture, le texte chinois n'est pas segmenté : il n'y a pas de délimiteur explicite entre les mots. Pour construire les modèles de langage et les modèles de traduction où le mot joue un rôle primordial, une segmentation du texte chinois en mots est utilisée (Luo *et al.*, 2009). Ce module est basé sur un modèle N-gramme de mots.

Le tableau 2 résume les performances des systèmes de traduction de base réglés et testés sur les dev/test BTEC. Les corpus HIT, Tatoeba et BTEC ont été utilisés pour l'entraînement des modèles de traduction pour la paire français/anglais et les corpus HIT et TED pour la paire français/chinois.

Direction de traduction	Score BLEU
FR->EN	62,8 (6 références)
EN->FR	53,0
ZH->FR	25,0
FR->ZH	10,2

TABLE 2 – Performances sur le test 2009 BTEC des systèmes de traduction de base avec ponctuation en source.

5. http ://sourceforge.net/projects/mgizapp/
6. http ://www.statmt.org/moses/

6.3 Adaptation de la traduction automatique aux données RAP

Les expériences d'adaptation des modèles de traduction à la traduction de parole présentées ci-dessous ont été menées uniquement pour la direction du français vers anglais pour le moment, l'objectif étant de définir les techniques d'adaptation les plus prometteuses qui pourront par la suite être appliquées aux autres directions de traduction pertinentes pour le projet RAPMAT.

Le système de RAP utilisé offre la possibilité de produire une ponctuation automatique de ses sorties. Le tableau 3 résume les résultats de comparaison des modèles de traduction de base avec et sans ponctuation dans la partie source des données d'entraînement et de test. La ponctuation en cible a été gardée pour pouvoir viser directement la traduction correcte et complète sans post-traitement. Les performances ont été évaluées à la fois sur le test BTEC 2009 et sur les transcriptions exactes et automatiques (WER 22,5%) du test Tatoeba. Si une dégradation du score BLEU peut être observée pour les textes propres, ce n'est pas le cas sur les transcriptions automatiques (légère amélioration), où la ponctuation est produite automatiquement et peut contenir des erreurs. Il est important de noter que le corpus test BTEC contenait uniquement 3% de ponctuations (ce qui peut expliquer les performances très proches avec et sans ponctuation), tandis que le corpus test de Tatoeba contenait approximativement 12% de ponctuations dans la transcription manuelle et dans la transcription automatique. Dans les expériences suivantes nous avons donc utilisé les versions des corpus sans ponctuation en source pour l'entraînement, le réglage et les tests des modèles de traduction.

	test BTEC (6 références)	test Tatoeba manuel	test Tatoeba auto
Base, avec ponctuation en source	62,8	45,7	32,3
Base, sans ponctuation en source	62,9	45,1	32,5

TABLE 3 – Comparaison des scores BLEU des modèles de traduction avec et sans ponctuation en source, sur le test BTEC et les transcriptions manuelles du test Tatoeba

Adaptation par réglage : La première tentative consiste à utiliser des transcriptions automatiques lors de la phase de réglage des systèmes. Nous comparons ainsi le modèle de base (réglé sur le corpus de développement BTEC), à un modèle réglé en utilisant en plus les transcriptions manuelles et les transcriptions automatiques du corpus de développement Tatoeba. Les tests sont menés sur le test BTEC et sur les transcriptions manuelles et automatiques du test Tatoeba (première et deuxième ligne du tableau 4). On observe, comme on pouvait s'y attendre, que les performances baissent sur les données écrites (corpus de test BTEC) avec le réglage sur les données transcrites automatiquement. En revanche, les résultats de cette technique d'adaptation montrent une amélioration du score BLEU sur le corpus de test Tatoeba. Cette amélioration est plus prononcée pour les transcriptions automatiques que pour les transcriptions manuelles du même corpus, et nous pouvons donc conclure que l'adaptation est plus liée aux particularités de données produites par la RAP qu'au corpus lui-même.

Adaptation par production de données parallèles : La seconde approche consiste à intégrer un corpus de données transcrites automatiquement dans les données parallèles d'entraînement du système de traduction. Par manque de données parallèles qui correspondraient à des transcriptions automatiques de parole et à leur traduction manuelle, l'approche consiste à produire

Corpus Entraînement	Corpus Développement	Corpus Tests		
		BTEC	Tatoeba manuel	Tatoeba auto
Base	BTEC	62,9	45,1	32,5
Base	BTEC + Tatoeba (manuel + auto.)	61,8	45,3	33,3
Base + adaptation BN	BTEC	62,3	45,3	33,0
Base + adaptation BN	BTEC + Tatoeba (manuel+auto)	61,2	46,4	34,0

TABLE 4 – Comparaison des scores BLEU entre le modèle de traduction de base et le modèle entraîné avec les traductions automatiques des données RAP, avec réglage sur les données écrits et sur les données provenant de RAP.

artificiellement ce type de données en traduisant automatiquement le corpus monolingue français BN avec le modèle de traduction de base. Pour éviter des dégradations liées à la qualité médiocre de certaines traductions automatiques, un filtrage des traductions selon leur score de confiance (normalisé par rapport à la longueur de chaque traduction en mots) a été réalisé en éliminant les traductions dont le score normalisé est inférieur à la moyenne pondérée. Les traductions automatiques sélectionnées ont été alignées parallèlement à la fois à leurs transcriptions manuelles et à leurs transcriptions automatiques. Ces données parallèles, adaptées au système de RAP, sont utilisées comme un corpus supplémentaire pour l'entraînement du modèle de traduction. Le tableau 4 donne la comparaison entre le modèle de base et le modèle entraîné en utilisant en plus les données adaptées au système de RAP (avec les deux types de réglage présentés ci-dessus : 3ème et 4ème lignes). Le meilleur résultat est obtenu avec le modèle de traduction entraîné et réglé en utilisant les données provenant de la reconnaissance automatique.

7 Conclusion

Le projet RAPMAT a permis, entre autres objectifs, d'explorer les spécificités de la traduction de la parole. Nous avons proposé une approche d'apprentissage semi-supervisé permettant d'adapter les systèmes de traduction automatique à la traduction de la parole sans utiliser de grandes quantités des données de parole traduites manuellement. Le système de traduction ainsi adapté améliore les performances de traduction sur les transcriptions automatiques par rapport au système de base non adapté. Nous envisageons par la suite d'étendre cette approche aux autres directions de traduction du projet et de l'affiner en utilisant des modèles de traduction plus sophistiqués (notamment à multiples tables de traduction). D'autres approches d'adaptation et de création de données artificielles pour la traduction de la parole, vont également être explorées à court terme, telles que la génération de pseudo-transcriptions automatiques à partir de données textuelles.

Remerciements

Ces travaux ont été menés grâce au soutien de la DGA (Direction Générale de L'Armement) et de la DGCIS (Direction Générale de la Compétitivité, de l'Industrie et des Services).

Références

ALLAUZEN, A. et YVON, F. (2012). Statistical methods for machine translation. *In* GAUSSIER, E. et YVON, F., éditeurs : *Textual Information Access*, pages 223–304. ISTE/Wiley, London.

GAHBICHE-BRAHAM, S., BONNEAU-MAYNARD, H. et YVON, F. (2011). Two ways to use a noisy parallel news corpus for improving statistical machine translation. *In Proceedings of the 4th Workshop on Building and Using Comparable Corpora*, Portland, Oregon, USA.

GAUVAIN, J., ADDA, G., LAMEL, L., LEFÈVRE, F. et SCHWENK, H. (2005). Transcription de la parole conversationnelle. *Traitement Automatique des Langues (TAL)*, 45.

GAUVAIN, J.-L. et LAMEL, L. (2008). Speech recognition systems. *In* MARIANI, J., éditeur : *Spoken Language Processing*.

JABAIAN, B., BESACIER, L. et LEFEVRE, F. (2013). Comparison and combination of lightly supervised approaches for language portability of a spoken language understanding system. *IEEE Transactions on Audio, Speech and Language Processing*, 21:636–648.

KIKUI, G., YAMAMOTO, S., TAKEZAWA, T. et SUMITA, E. (2006). Comparative study on corpora for speech translation. *IEEE Transactions on Audio, Speech and Language Processing*, 14:1674–1682.

LAMEL, L. et AL (2011). Speech recognition for machine translation in quaero. *In Proceeding of the 11th International Workshop on Spoken Language Translation*, San Francisco, USA.

LUO, J., LAMEL, L. et GAUVAIN, J. (2009). Modeling characters versus words for mandarin speech recognition. *In International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.

MATUSOV, E. et NEY, H. (2011). Lattice-based ASR-MT interface for speech translation. *IEEE Transactions on Audio, Speech and Language Processing*, 19:721–732.

MAURO, C., JAN, N., SEBASTIAN, S., LUISA, B. et MARCELLO, F. (2013). Report on the 10th iwslt evaluation campaign. *In IWSLT*, Eidelberg, Allemagne.

PEITZ, S., WIESLER, S., NUSSBAUM-THOM, M. et NEY, H. (2012). Spoken language translation using automatically transcribed text in training. *In Proceeding of the 9th International Workshop on Spoken Language Translation*, Hong Kong, China.

PEREZ, A., TORRES, I. et CASACUBERTA, F. (2012). Finite-state acoustic and translation model composition in statistical speech translation : empirical assessment. *In Proceedings of the 10th International Workshop on Finite State Methods and Natural Language Processing*, Donostia-San Sebastian, Spain.

SCHWENK, H. (2008). Investigations on largescale lightly-supervised training for statistical machine translation. *In Proceedings of the International Workshop on Spoken Language Translation*, Hawaii, USA.

STÜKER, S., HERRMANN, T., KOLSS, M., NIEHUES, J. et WOLFEL, M. (2012). Research opportunities in automatic speech-to-speech translation. *Potentials, IEEEe*, 3:26–33.

YANG, M., JIANG, H., ZHAO, T. et LI, S. (2006). Construc trilingual parallel corpus on demand. *Chinese Spoken Language Processing, Lecture Notes in Computer Science*, 4274:760–767.