Développement et Évaluation d'un Système de Traduction Automatique de la Parole en Pashto vers le Français

Fethi Bougares[1] Anthony Rousseau[1] Paul Deléglise[1]
Yannick Estève[1] Loïc Barrault[1] Holger Schwenk[1]
Sylvie Brunessaux[2] Khaled Khelif[2] Mathieu Manta[3]
Laboratoire d'Informatique de l'Université du Maine[1] prénom.nom@lium.univ-lemans.fr
Airbus Defence and Space[2] prénom.nom@cassidian.com
Direction Générale de l'Armement[3] mathieu.manta@intradef.gouv.fr

RÉSUMÉ

Le pashto fait partie des langues peu dotées caractérisées par un manque remarquable d'outils de traitement automatique. Ce travail présente une première expérience dans le LIUM avec la langue pashto. Dans cet article, nous présentons les résultats de nos premières expériences de traduction de la parole en pashto/français. La traduction de la parole pashto est réalisée en deux étapes : dans un premier temps, un système de transcription de la parole est utilisé pour transcrire de la parole. La sortie de la transcription est par la suite traduite avec un système de traduction statistique. Nous décrivons dans cet article les développements des différents modèles de chaque système. Ces travaux ont été réalisés en collaboration avec Airbus Defence and Space (anciennement Cassidian) dans le cadre du projet TRAD financé par la DGA.

ABSTRACT

Development and Evaluation of a Pashto/French Spoken Language Translation System

This paper presents an overview of the development of a Pashto spoken language translation system. We present a number of challenges encounter the development of this system highlighting the lack of Pashto resources and tools of all sorts. Based on open source tools and in collaboration with Airbus Defence and Space (formerly Cassidian) under the TRAD project funded by the French Ministry of Defence (DGA), a speech recognition and machine translation systems are developed, described and evaluated in this work.

MOTS-CLÉS : Traduction automatique de la Parole, Langue peu dotée, Pashto.

KEYWORDS: Spoken Language Translation, Under-resourced Language, Pashto.

1 Introduction

L'article 40 de la déclaration universelle des droits linguistiques signé à Barcelone en juin 1996 par l'UNESCO et plusieurs organismes gouvernementaux affirme depuis 1996 que «*Toute communauté linguistique a le droit de disposer, dans le domaine de l'informatique, d'équipements adaptés à son système linguistique et d'outils de production dans sa langue, afin de profiter pleinement du potentiel qu'offrent ces technologies pour l'auto-expression, l' éducation, la communication, l'édition, la traduction, et en général le traitement de l'information et de la diffusion culturelle.*»

Malgré cette affirmation, et malgré les nombreuses initiatives organisées pour informatiser des langues peu dotées, toutes ne sont pas égales devant le processus d'informatisation et beaucoup de langues restent aujourd'hui mal dotées ou avec peu de ressources linguistiques informatisées. Le manque de telles ressources et les caractéristiques méconnues de ces langues rendent le développement de systèmes de traitement automatique difficile.

Dans cet article, nous présentons le développement et l'analyse d'un système de traduction automatique de la parole pour une langue particulièrement peu dotée : le pashto. Le système consiste à transcrire, dans un premier temps, automatiquement des enregistrements audio des conversations en Pashto et de les traduire, ensuite vers le français avec un système de traduction statistique. Dans la section 2, nous décrivons les particularités de la langue pashto ainsi que les difficultés liées à son traitement automatique. La section 3 est réservée à la description du développement de système de transcription et son développement. Avant de conclure, la section 4 présente le système de traduction statistique et les résultats obtenus.

2 Particularités et traitement automatique du pashto

Le pashto est une langue parlée par environ quarante-cinq à cinquante millions de personnes dans le monde. Elle constitue l'une des deux langues officielles de l'Afghanistan, particulièrement parlée par le peuple *pachtoune* dans le sud, l'est et le sud-ouest du pays, ainsi que dans le nord et le nord-ouest du Pakistan mais également en Iran, au Tadjikistan et en Inde (Uttar Pradesh et Cachemire).

2.1 Particularités linguistiques

Le pashto appartient à la famille des langues indo-iraniennes, basée sur les lettres de l'alphabet arabe, augmenté par des lettres supplémentaires empruntées aux alphabets perse et urdu. Son alphabet est composé de quarante-quatre lettres, auxquelles s'ajoutent quatre marques diacritiques. Le pashto s'écrit de droite à gauche et un mot peut se composer de plusieurs éléments (proclitique, préfixe, base, suffixe et enclitique). De plus, cette langue a la particularité de se décliner en quatre dialectes, selon la région d'origine du locuteur, ajoutant une difficulté supplémentaire à son traitement automatique. Cependant, la majeure difficulté provient du fait que le pashto est une langue très peu dotée en ressources linguistiques : par exemple, il n'existe pas de corpus pashto dans les catalogues de LDC [1].

2.2 Particularités de traitement

Pour l'écriture du pashto, les lettres changent de glyphe selon leur position dans le mot. Chaque lettre a donc une représentation graphique différente en fonction de sa position. Dans le cadre d'encodage électronique de texte, comme pour l'arabe, chaque lettre doit être idéalement encodée

1. www.ldc.upenn.edu

par la valeur principale Unicode correspondante, indépendante de la position de la lettre et donc du glyphe affiché. Par exemple, le bloc de valeurs principales Unicode pour l'écriture de l'arabe est compris entre U+FB50 et U+FDFF et entre U+FE70 et U+FEFF. Dans ces conditions, c'est à la charge du moteur de rendu graphique de gérer l'affichage du bon glyphe correspondant au contexte d'un caractère.

Ce point est très important car la pluralité des sources de texte rend ces conditions difficiles à respecter. Ainsi, il est très fréquent d'observer dans les données rencontrées des lettres encodées non pas par leur valeur Unicode principale, indépendante du glyphe, mais directement par la valeur du glyphe affiché. Ainsi, un même mot d'un même corpus peut être représenté électroniquement par plusieurs séquences de valeurs différentes. Il s'agit là d'un problème crucial car fréquent dans les données, entraînant du bruit dans les modèles de langage, perturbant le processus de phonétisation et biaisant les évaluations.

Une difficulté se présente également, similaire à celle que l'on rencontre dans le traitement de l'arabe : un mot n'est pas transcrit avec ses voyelles et, pour une même représentation orthographique, peut se prononcer différemment en fonction du contexte lexical dans lequel apparaît la suite de consonnes le représentant. Cette représentation orthographique peut aussi faire référence à des mots ayant des sens différents.

3 Reconnaissance automatique de la parole en pashto

3.1 Ressources linguistiques

Les données accessibles au LIUM pour l'apprentissage des différents modèles du système de reconnaissance automatique de la parole (SRAP) en pashto se composent de corpus produits dans le cadre du projet TRAD par ELDA (Mostefa *et al.*, 2012) et par Appen Butler Hill. Dans l'objectif de représenter au mieux la variabilité dialectale entre locuteurs, le corpus utilisé pour le SRAP est composé des enregistrements de locuteurs (hommes et femmes) provenant de trois différentes régions (Afghan du sud, Afghan du nord et pakistanais).

Ces données sont de plusieurs types et sources :

- des données textuelles issues de pages Web pour l'apprentissage des modèles de langage (ensemble web produit par ELDA) ;
- des enregistrements audio de journaux télévisés et leur transcriptions manuelles (ensemble H4 produit par ELDA) ;
- des enregistrements audio de la parole conversationnelle et leur transcriptions manuelles (ensemble H5 produit par Appen Butler Hill).

L'objectif étant de créer un système de traduction de la parole conversationnelle, nous avons extrait un corpus de développement (dev-H5) du corpus H5. Le tableau 1 présente les caractéristiques de ces données.

En plus de ces données d'apprentissage et de développement, afin de construire notre dictionnaire, nous disposons de la phonétisation manuelle de 9000 des mots les plus fréquents en pashto. Cette fréquence a été mesurée sur les transcriptions de parole à notre disposition. Néanmoins, ce dictionnaire ne contient que la prononciation canonique d'un mot, représenté sous la forme

Corpus	Durée de parole (h)	Nombre de segments/phrases	Nombre de mots	Durée moyenne d'un segment (s)
H4	86,06	42 095	967,3k	7,36
H5	31,67	31 038	407,5k	3,67
web	N/A	1 202 839	108,97M	N/A
dev-H5	0,97	877	N/A	3,98

TABLE 1 – Caractéristiques des corpus utilisés pour l'apprentissage et l'optimisation des SRAP en pashto.

d'une séquence de consonnes.

3.2 Choix du système principal de reconnaissance automatique de la parole

Le LIUM a choisi de travailler principalement sur la base du décodeur Kaldi [2] pour développer ses SRAP. Deux raisons nous ont amenés à faire ce choix :

— Kaldi est un système très performant et à l'état de l'art quant à la modélisation acoustique. Le LIUM l'a déjà utilisé avec succès, combiné avec Sphinx, dans le cadre de la campagne d'évaluation interne au défi ANR REPERE [3] en janvier 2013 ;
— dans le cadre de nos expériences, il était nécessaire de développer un SRAP fonctionnant en conservant une latence faible. Le système Kaldi a ceci de particulier puisqu'il est basé sur une approche par transducteurs qui implique que l'espace de recherche est calculé avant le décodage : cela lui permet d'être plus rapide que la plupart des SRAP d'une autre nature. Le prix de cette vitesse est une consommation mémoire importante, qui peut toutefois être maitrisée par un vocabulaire de taille limitée. La vitesse de traitement de Kaldi a été un critère déterminant ;

3.3 Phonétisation du Pashto

Afin de disposer d'un dictionnaire contenant un nombre suffisant d'entrées pour l'apprentissage d'un SRAP, nous avons utilisé les 9000 mots phonétisés manuellement comme corpus d'apprentissage pour un système de phonétisation automatique, basé sur une approche probabiliste. Le système utilise l'outil Moses pour apprendre un système de traduction statistique avec comme corpus d'apprentissage bilingue les mots avec leur phonétisation. L'idée est d'utiliser le paradigme de la traduction automatique pour traduire un mot vers sa phonétisation, la méthode est détaillée dans (Laurent *et al.*, 2009). Dans nos expérimentations, sur un sous-ensemble des 9000 mots de départ, nous avons relevé que le taux de mots correctement phonétisés était de 83.29% (qui représente le score BLEU de la traduction mots-phonèmes).

À l'aide cet outil, le dictionnaire que nous avons constitué est composé des 40700 mots les plus fréquents rencontrés dans nos transcriptions de parole : nous avons donc phonétisé automatiquement

2. http ://kaldi.sourceforge.net/about.html
3. http ://www.defi-repere.fr

près de 32000 mots.

Pour pallier la difficulté induite par le manque de voyelles dans les transcriptions, nous avons relâché les contraintes utilisées lors d'un alignement forcé des corpus d'apprentissage en autorisant plusieurs voyelles à être alignées dans certains contextes. Puis, en analysant le résultat du processus d'alignement automatique, nous avons injecté des variantes de prononciation observées dans le corpus d'apprentissage dans notre dictionnaire.

3.4 Modélisation du langage

Les modèles de langage ont été estimés sur les données H4 et H5 à notre disposition, en plus des données textuelles extraites depuis le Web.

3.4.1 Modèles de langage H4

Pour le type de données H4, le modèle de langage principal est un modèle trigramme. Nous avons expérimenté l'utilisation d'un modèle quadrigramme, mais il dégradait les performances mesurées par un calcul de perplexité sur le corpus de développement. Nous pensons que cela provient du manque de données proches des données H4 dans les données provenant d'Internet.

Pour l'estimation de ce modèle, nous avons d'abord calculé un modèle de langage par source (H4, H5 et web) avec le vocabulaire de 40700 mots. Ensuite, une interpolation linéaire a été réalisée à l'aide de coefficients calculés de manière à minimiser la valeur de perplexité du modèle interpolé sur le corpus de développement H4. La perplexité du modèle final calculée sur ce corpus de développement est de 176 (185 pour le modèle quadrigramme). Un modèle de langage bigramme, pour le décodage principal, a également été estimé de manière similaire. Il obtient une valeur de perplexité égale à 236.

3.4.2 Modèles de langage H5

La même procédure a été suivie pour le modèle de langage H5 qui est lui aussi un modèle trigramme. Le vocabulaire utilisé contient 13244 mots : nous avons restreint la taille du vocabulaire en raison de la faible taille du corpus d'apprentissage et de l'application beaucoup plus contrainte d'un point de vue sémantique. Ce vocabulaire est composé de tous les mots présents dans le corpus d'apprentissage et le corpus de développement H5, plus les 5000 mots les plus fréquents du corpus d'apprentissage H4 absents du corpus d'apprentissage H5. Les trois sources (H4, H5 et web) ont été utilisées pour le modèle trigramme. Sur le corpus de développement H5, la valeur de perplexité obtenue est de 128. Un modèle de langage bigramme a également été estimé qui obtient une valeur de perplexité égale à 134.

3.5 Modélisation acoustique et décodage

Au cœur des SRAP utilisés pour la reconnaissance du pashto, les modèles acoustiques se présentent sous la forme classique de modèles de Markov cachés (HMM), dont chaque état est associé à une

mixture de gaussiennes. Chaque HMM modélise un phonème en contexte (contexte phonétique gauche, contexte phonétique droit, position dans le mot).

Les modèles acoustiques ont été estimés sur les données audio en pashto avec transcriptions manuelles disponibles. Pour leur apprentissage, nous avons appliqué plusieurs approches successives : d'abord une estimation LDA (Linear Discriminant Analysis) (Haeb-Umbach et Ney, 1992) avec transformation MLLT (Maximum Likelihood Linear Transform), puis une approche SAT (Speaker Adaptative Training) (McDonough *et al.*, 2002) avec transformation fMLLR (Maximum Likelihood Linear Regression) et enfin une estimation fMMI (Minimum Mutual Information).

Le SRAP développé est basé sur un noyau issu du système Kaldi (Povey *et al.*, 2011) pour le décodage principal qui génère un graphe de mots. Ce décodage principal s'effectue en deux étapes : un premier décodage, indépendant du locuteur, qui permet d'obtenir une première transcription. Cette transcription est utilisée pour calculer des matrices de transformation fMLLR qui permettent d'adapter le décodage acoustique à la voix du locuteur et aux conditions acoustiques. Le second décodage exploite ces matrices et génère des graphes de mots. Ces deux décodages utilisent un modèle bigramme plutôt qu'un modèle trigramme afin d'accélérer le traitement en réduisant la taille de l'espace de recherche. Chaque graphe de mots généré par le noyau Kaldi est ensuite traité pour réévaluer les scores linguistiques à l'aide d'un modèle trigramme.

4 Traduction automatique de la parole pashto

La présence d'une grande quantité de données d'apprentissage adéquates ainsi que l'utilisation de logiciels libres permettent la construction rapide de systèmes de traduction automatique produisant de bons résultats. Toutefois, la traduction vers ou à partir de langues peu dotées doit faire face au manque de données linguistiques disponibles. De plus, la majorité des langues peu dotées ne disposent pas d'outils de traitement linguistique dédiés comme par exemple les analyseurs morphologiques qui se montrent utiles dans le cadre de la traduction de langues morphologiquement riches. Dans l'objectif de surmonter ce manque de ressources linguistiques, plusieurs travaux s'orientent vers la création de données de manière non supervisée en exploitant par exemple des données comparables. Dans cette section, nous présentons notre premier système de traduction pour la paire de langues pashto-français.

4.1 Données d'apprentissage

La construction manuelle d'un corpus bilingue est généralement réalisée par des locuteurs natifs de la langue cible (la langue française dans notre cas). Dans notre contexte, il est très difficile de trouver un locuteur français natif maitrisant la langue pashto. C'est pour cette raison que les corpus d'apprentissage ont été traduits du pashto par des locuteurs natifs (langue source) et révisés ensuite par des locuteurs natifs français (langue cible) (Mostefa *et al.*, 2012).

Le système de traduction a été entraîné à partir de données distribuées par ELDA et Appen Butler Hill. Deux corpus bilingues ont été utilisés. Ces données représentent la traduction des corpus pashto H4 et H5 transcrits par des locuteurs natifs et présentés dans la section 3.1.

Dans l'objectif d'augmenter la couverture des données d'apprentissage, nous avons ajouté deux corpus bilingues. Le premier compte 5737 traductions d'entités nommées (nom de personnes

et de villes) et le second contient 737 conversations usuelles en langue pashto ainsi que leurs traductions françaises. Ces deux corpus sont téléchargeables gratuitement sur le site du LIUM [4].

La Table 2 présente la taille, en nombre de phrases, de l'ensemble des données bilingues utilisées pour l'apprentissage et le développement du système de traduction.

Corpus	Taille (#phrases)
H4	33k
H5	29K
names	5737
Web-unsup	737

TABLE 2 – Corpus bilingues utilisés pour construire et développer le système de traduction pashto-français

La Table 3 présente les données monolingues utilisées pour apprendre le modèle de langage en langue cible (la langue française en l'occurrence). En plus de la partie cible des corpus bilingues, nous avons utilisé deux corpus libres et téléchargeables sur Internet : le corpus OpenSubtitles2012 [5] et le corpus de Wikipédia [6].

Corpus	(#phrases)
H4-fr	33k
H5-fr	29K
wiki-extract	13M
OpenSubtitles-2012	4,3M

TABLE 3 – Corpus monolingues utilisés pour le système de traduction pashto-français

Les données d'apprentissage (bilingues et monolingues) ont fait l'objet d'un pré-traitement pour s'adapter au style de la sortie du SRAP. Ce pré-traitement comprend, entre autres, la suppression de la ponctuation et de la casse, ainsi que la conversion des chiffres en lettres.

4.2 Système de traduction

Le système de traduction développé est un système à base de segments (*phrase-based*). La table de traduction et le modèle de réordonnement sont estimés à l'aide des outils issus du décodeur Moses [7] (Koehn *et al.*, 2007). Le modèle de langage est un modèle quadrigramme de repli Kneyser-Ney modifié obtenu après une interpolation des différents modèles présentés dans la table 3 à l'aide des outils SRILM (Stolcke, 2002). La perplexité de chaque modèle sur le corpus de développement ainsi que les cœfficients d'interpolation sont présentés dans la Table 4. La tokénisation des corpus est réalisée via l'outil inclus dans le décodeur Moses et les textes parallèles ont été alignés au niveau des mots avec l'outil mGiza. Enfin, les paramètres du système (14 paramètres du modèle log-linéaire) ont été optimisés avec l'algorithme MERT (Och, 2003) sur le corpus de développement.

4. http://www-lium.univ-lemans.fr/~bougares/personnel.php
5. http://www.opensubtitles.org/fr
6. http://dumps.wikimedia.org
7. http://www.statmt.org/moses/

Corpus	Perplexité	Cœfficient
H5-fr	61,88	0,66
H4-fr	130,23	0,09
wiki-extract	102,65	0,19
OpenSubtitles2012	149,37	0,06
Interpolation	49,57	n/a

TABLE 4 – Perplexité sur le corpus de développement et interpolation des modèles de langage estimés sur chacun des corpus monolingues.

4.3 Performances du système

La table 5 présente le taux d'erreur mots de système de transcription et les scores BLEU obtenus par le système de traduction selon le type de l'entrée de traduction.

Entrée de traduction	WER	BLEU
Transcription de référence	n/a	17,78
Transcription automatique+audio-dev	25,9	n/a
Transcription automatique	36,0	13,22

TABLE 5 – Score BLEU de la traduction de différent type de texte

Le système a été utilisé dans un premier temps pour traduire la transcription manuelle du corpus de DEV. Cette première expérience montre le score maximum qu'on pourrait avoir avec une transcription parfaite. La traduction de la transcription de référence donne une sortie avec un score BLEU de 17.78 points. Le système a été utilisé par la suite pour traduire la sortie de système de transcription. Le tableau montre bien la corrélation entre la qualité de la transcription et celle de la traduction. En effet, la traduction d'une transcription automatique de 36% de WER réduit le score BLEU de 4.56 points absolus. Dans l'objectif de mesurer l'effet de manque de données sur la qualité de la transcription, nous avons augmenté le corpus d'apprentissage acoustique en ajoutant le corpus de DEV (le modèle de langage n'a pas été modifié). L'intégration de ces données permet une réduction absolue de WER 10,1 point ce qui montre le manque évident de données d'apprentissage de modèle acoustiques.

Conclusion

Dans cet article nous avons présenté un premier investissement du LIUM dans le domaine de traitement automatique des langues peu dotées. Nous avons développé un système de traduction automatique de la parole pour le pashto dans le cadre du projet TRAD en collaboration avec Airbus Defence and Space pour le compte de la DGA. Ce travail présente les premiers résultats qui montrent que l'application des technologies de traitement des langues à l'état de l'art aux langues peu dotées ne permet pas d'avoir des performances équivalentes à leur utilisation avec les langues ayant beaucoup de ressources numériques.

Références

HAEB-UMBACH, R. et NEY, H. (1992). Linear discriminant analysis for improved large vocabulary continuous speech recognition. *In Proceedings of the 1992 IEEE International Conference on Acoustics, Speech and Signal Processing - Volume 1*, ICASSP'92, pages 13–16, Washington, DC, USA. IEEE Computer Society.

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 the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

LAURENT, A., DELÉGLISE, P. et MEIGNIER, S. (2009). Grapheme to phoneme conversion using an SMT system. *In INTERSPEECH*, pages 708–711.

MCDONOUGH, J. W., SCHAAF, T. et WAIBEL, A. (2002). On Maximum Mutual Information speaker-adapted training. *In ICASSP*, pages 601–604. IEEE.

MOSTEFA, D., CHOUKRI, K., BRUNESSAUX, S. et BOUDAHMANE, K. (2012). New language resources for the pashto language. *In LREC 2012*.

OCH, F. J. (July 2003). Minimum error rate training in statistical machine translation. *In Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167.

POVEY, D., GHOSHAL, A., BOULIANNE, G., BURGET, L., GLEMBEK, O., GOEL, N., HANNEMANN, M., MOTLICEK, P., QIAN, Y., SCHWARZ, P., SILOVSKY, J., STEMMER, G. et VESELY, K. (2011). The kaldi speech recognition toolkit. *In IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society.

STOLCKE, A. (2002). SRILM — an extensible language modeling toolkit. *In in Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP)*, Denver, CO, USA,.