Apport de l'adaptation automatique des modèles de langage pour la reconnaissance de la parole: évaluation qualitative extrinsèque dans un contexte de traitement de cours magistraux

Salima Mdhaffar[1] Yannick Estève[2] Nicolas Hernandez[3] Antoine Laurent[1]
Solen Quiniou[3]
(1) LIUM, Avenue Olivier Messiaen ,72085 Cedex 9 Le Mans , France
(2) LIA, AGROPARC, BP 1228, 84911 Cedex 9 Avignon, France
(3) LS2N, 2 rue de la houssinière, BP 92208, 44322 Cedex 3 Nantes, France
firstname.lastname@{univ-lemans,univ-avignon,univ-nantes}.fr

RÉSUMÉ

Malgré les faiblesses connues de cette métrique, les performances de différents systèmes de reconnaissance automatique de la parole sont généralement comparées à l'aide du taux d'erreur sur les mots. Les transcriptions automatiques de ces systèmes sont de plus en plus exploitables et utilisées dans des systèmes complexes de traitement automatique du langage naturel, par exemple pour la traduction automatique, l'indexation, la recherche documentaire... Des études récentes ont proposé des métriques permettant de comparer la qualité des transcriptions automatiques de différents systèmes en fonction de la tâche visée. Dans cette étude nous souhaitons mesurer, qualitativement, l'apport de l'adaptation automatique des modèles de langage au domaine visé par un cours magistral. Les transcriptions du discours de l'enseignant peuvent servir de support à la navigation dans le document vidéo du cours magistral ou permettre l'enrichissement de son contenu pédagogique. C'est à-travers le prisme de ces deux tâches que nous évaluons l'apport de l'adaptation du modèle de langage. Les expériences ont été menées sur un corpus de cours magistraux et montrent combien le taux d'erreur sur les mots est une métrique insuffisante qui masque les apports effectifs de l'adaptation des modèles de langage.

ABSTRACT

Contribution of automatic adaptation of language models for speech recognition : extrinsic qualitative evaluation in a context of educational courses

Despite the known weaknesses of this metric, the performance of various automatic speech recognition systems is generally compared using the word error rate. The automatic transcriptions of these systems are usually used in complex natural language processing systems, for example for machine translation, indexation, document retrieval... Recent studies have proposed metrics to compare the quality of automatic transcriptions of different systems according to the target task. In this study, we investigated to qualitatively measure the contribution of the automatic adaptation of language models to the domain of a lecture. The transcriptions of the teacher's speech can serve as a basis for navigating in the video of the lecture or or for allowing the enrichment of its pedagogical content. By taking these two tasks into account, we propose to evaluate the contribution of the language model adaptation. Experiments were conducted on an educational corpus, and show how the word error rate is an insufficient metric that masks the strength contributions of the adaptation of language models.

MOTS-CLÉS : reconnaissance automatique de la parole, adaptation de modèles de langage, mesure

d'indexibilité, recherche d'information, éducation.

KEYWORDS: Automatic Speech Recognition, Language Model Adaptation, Word Error Rate, Indexability, Information Retrieval, Transcription, Educational Applications.

1 Introduction

La transcription automatique de cours magistraux convertit automatiquement le discours de l'enseignant (audio) en texte. Même si ces dernières années la technologie de reconnaissance automatique de la parole a considérablement progressé, principalement grâce aux architectures neuronales pour la modélisation acoustique, un système de reconnaissance automatique de la parole (SRAP) reste sensible aux mots hors vocabulaire et à la précision de ses modèles de langage (ML). Un tel système doit par exemple être bien préparé pour traiter des documents spécialisés. Or, dans le cadre de la transcription de cours magistraux, chaque cours nécessite une terminologie précise liée à son domaine. L'adaptation des modèle de langage est une technique indispensable pour résoudre ce problème d'inadéquation entre les données d'apprentissage et de test.

Généralement, pour la reconnaissance de la parole, le gain en performance de l'adaptation des modèles de langage est mesurée à l'aide du taux d'erreur mots (WER) (Pallett, 2003) : cette métrique d'évaluation est couramment utilisée dans la littérature pour l'analyse des performances des systèmes de reconnaissance automatique de la parole. Cette mesure s'appuie sur une comparaison entre la phrase produite par le SRAP et la phrase correspondante transcrite manuellement. Un alignement mot à mot utilisant la distance de Levenshtein est réalisé entre la transcription manuelle (référence) et la transcription automatique (hypothèse). Ensuite, une comparaison est effectuée selon les différents types d'erreurs sur les mots que peut commettre le système : insertions, suppressions et substitutions. Le calcul du WER s'effectue selon la formule suivante :

$$WER = \frac{S + I + D}{N} \tag{1}$$

où S est le nombre de mots substitués par le système, I est le nombre de mots insérés par le système, D est le nombre de mots supprimés par le système et N est le nombre total de mots dans la phrase.

WER attribue un score d'erreur en pourcentage pour la transcription globale. Cela est très utile lorsqu'il s'agit d'évaluer la performance du SRAP isolément. Cependant, les systèmes SRAP sont souvent conçus comme une brique dans d'autres applications de traitement de langage naturel qui utilisent les transcriptions de sortie pour effectuer d'autres tâches. Ces transcriptions constituent en effet une ressource précieuse pour d'autres modules technologiques appliquant des traitements tels que la recherche d'informations, la traduction, l'indexation de documents... La qualité des transcriptions de sortie affecte ainsi directement les performances de ces modules.

L'adaptation des MLs pour des cours magistraux a suscité beaucoup d'attention dans la littérature (Cerva *et al.*, 2012; Bell *et al.*, 2013; Yamazaki *et al.*, 2007; Kawahara *et al.*, 2008; Martínez-Villaronga *et al.*, 2013). La performance de ces travaux a été évaluée en utilisant WER ou la perplexité. Cependant, cette mesure ne prend pas en compte la gravité de l'erreur en fonction de la tâche finale (Luzzati *et al.*, 2014). En 2002, les auteurs de (Hürst *et al.*, 2002) ont déjà démontré que l'utilisation d'un vocabulaire thématique améliorait la reconnaissance de la parole et l'indexation des cours magistraux. Dans leurs analyses, ils se sont intéressés à l'analyse de l'impact du locuteur et du vocabulaire et ils ont démontré que l'utilisation d'un vocabulaire du domaine du cours améliorait les

performances du SRAP.

Considérant que le taux d'erreur de mot (WER) n'est pas suffisamment pertinent pour comparer la performance du SRAP pour certaines tâches spécifiques (Jannet *et al.*, 2015; Favre *et al.*, 2013) et que les erreurs du SRAP peuvent avoir une forte incidence sur la précision de plusieurs tâches de TALN, nous explorons, dans cet article, l'utilisation de deux mesures d'évaluation plus pertinentes pour comparer les apports de l'adaptation du modèle de langage pour un SRAP.

2 Données expérimentales

2.1 Description des données

Les expériences ont étés effectuées sur des données du domaine de l'éducation : elles sont constituées d'environ 10 heures d'enregistrements de cours magistraux en français. Toutes ces données ont été transcrites manuellement et annotées en segments thématiques par des experts. Les annotations ont été réalisées conformément au guide d'annotation développé durant le projet PASTEL (Mdhaffar *et al.*, 2018).
Ce corpus contient (1) des vidéos filmées durant des séances de cours magistraux, (2) les transcriptions manuelles de ces cours magistraux, (3) les diapositives de ces cours et (4) des annotations de segmentation thématique avec deux niveaux de granularité. « Granularité 1 » représente une segmentation fine du cours : chaque nouveau concept abordé durant le cours constitue un segment. « Granularité 2 » regroupe les segments de type « Granularité 1 » : elle est utilisée lorsqu'il y a un changement de sous-thème plus général qui permettrait d'arrêter l'apprentissage humain à ce moment-là et de reprendre plus tard l'apprentissage d'autres concepts. Le tableau 1 présente quelques statistiques du corpus (les cours pour lesquelles on dispose des diapositives). La deuxième et la troisième colonnes du tableau représentent, respectivement, le nombre de segments de « Granularité 1 » (G1) et le nombre de segments de « Granularité 2 » (G2). Le Nombre de locuteurs du corpus présenté dans le tableau est égale à 4.

2.2 Annotation en mots clés

Les mots du domaine ont été extraits manuellement à partir des transcriptions manuelles et des diapositives des cours. Nous considérons, comme mots du domaine, les expressions linguistiques faisant référence à des concepts, des objets ou des entités essentielles pour la compréhension de la diapositive actuelle ou d'une transcription donnée. Nous avons inclus tous les termes scientifiques et techniques ainsi que les acronymes et expressions permettant d'aller plus loin dans le sujet du cours. Ces annotations ont été réalisées pour les cours pour lesquels nous disposons des diapositives (6 cours du corpus). Les colonnes 4 et 5 du tableau 1 représentent, respectivement, le nombre de mots-clés annotés pour la transcription (Kw_t) et pour les diapositives (Kw_s). La dernière colonne contient la durée de chaque cours.

3 Description du SRAP et de l'adaptation du ML

3.1 SRAP

Le système SRAP est basé sur la boîte à outils Kaldi (Povey *et al.*, 2011). Des modèles acoustiques de type chain-TDNN (Povey *et al.*, 2016) ont été entraînés sur environ 300 heures de parole, princi-

Cours	G1	G2	Kw_t	Kw_s	Durée
Introduction à l'informatique	31	2	65	59	1h 04m
Introduction à l'algorithmique	38	10	30	37	1h 17m
Les fonctions	35	3	121	79	1h 14m
Réseaux sociaux et graphes	43	7	74	97	1h 05m
Algorithmique distribuée	72	5	314	158	1h 16m
Langage naturel	52	5	130	107	1h 09m
Total	271	32	734	537	7h 05m

TABLE 1 – Statistiques du corpus

palement de type journaux d'actualités radio- ou télé-diffusés en français. Les modèles de langage génériques (n-grammes) ont été entraînés sur les transcriptions manuelles de la parole utilisée pour l'apprentissage des modèles acoustiques mais également sur des articles de journaux, pour un total de 1,6 milliard de mots. Le vocabulaire du modèle de langage générique contient environ 160 000 mots. Les détails sur les modèles de langage peuvent être trouvés dans (Rousseau *et al.*, 2014).

3.2 Adaptation du ML

Dans cette étude, nous émettons l'hypothèse que l'enseignant collabore *a minima* en fournissant les diapositives de son cours. Les titres des diapositives sont importants pour donner aux étudiants une idée rapide du contenu des parties d'un cours. Ils représentent ainsi souvent l'information principale sur laquelle l'étudiant s'appuie pour chercher et naviguer dans le cours. L'idée est donc d'utiliser les titres des diapositives comme des requêtes. En se basant sur le travail de (Lecorvé *et al.*, 2008), les requêtes sont soumises à un moteur de recherche (Google) et les pages pointées par les liens renvoyés sont téléchargées. Nous avons limité la recherche à 400 pages Web pour chaque requête. Le contenu textuel principal de ces pages est extrait pour construire un modèle de langage du domaine. Une adaptation du modèle de langage est faite par interpolation linéaire de deux modèles : le modèle générique et le modèle du domaine. Les mots les plus fréquents parmi les 400 pages Web récupérées sont ajoutés au vocabulaire du SRAP pour le traitement d'un cours magistral particulier.

4 Évaluation du ML

4.1 Méthodologie d'évaluation : tâche de recherche d'information

L'une des tâches qui peuvent être utiles pour des applications pédagogiques est l'enrichissement automatique (ou sous forme de recommandation) des transcriptions avec des ressources externes. C'est l'un des objectifs des applications pédagogiques qui vise à offrir aux étudiants des liens externes utiles qui peuvent servir pour réviser ou avoir plus d'explications détaillées concernant les concepts du cours. Dans ce cas, il est important d'évaluer l'impact de la transcription sur une tâche de récupération de documents. Notre évaluation consiste à comparer les résultats de requêtes de recherche pour chaque segment de « Granularité 1 ». Les requêtes sont construites en utilisant l'approche TF-IDF sur les transcriptions de chacun de ces segments. Ces requêtes sont soumises à un moteur de recherche (ici, Google). Notre but est de déterminer la pertinence des documents récupérés. Nous avons défini comme documents pertinents (référence) les documents extraits de requêtes basées sur les transcriptions

manuelles. Sur la base de cette référence, une comparaison avec les documents récupérés à partir de requêtes construites sur des transcriptions résultant d'une reconnaissance automatique sans et avec adaptation des modèles de langage a été effectuée, en calculant un taux de couverture.

4.2 Méthodologie d'évaluation : tâche d'indexation

Dans cette seconde tâche, notre objectif est d'évaluer l'indexabilité des transcriptions. En d'autres termes, nous souhaitons déterminer si la qualité des transcriptions joue un rôle dans l'indexation et la récupération des transcriptions, qui sont utiles pour naviguer dans la vidéo du cours et pour atteindre rapidement ce que cherche l'apprenant. Les segments de type « Granularité 1 » ont été indexés en utilisant le moteur de recherche lemur [1]. Trois ensembles de segments ont été considérés : ceux des transcriptions manuelles, ceux des transcriptions automatiques sans adaptation et ceux des transcriptions automatiques avec adaptation. Des requêtes vont être présentés au moteur de recherche (lemur) pour récupérer les segments pertinents à partir de chaque ensemble distinct de segments. Les requêtes utilisées sont les 40 mots les plus fréquents dans la transcription manuelle, les mots des titres des diapositives et les mots clés annotés à partir de la transcription manuelle. Chaque requête renvoie une liste ordonnée de segments. Pour estimer la qualité de l'indexabilité, nous avons utilisé le coefficient de Spearman qui mesure la corrélation de rang entre les segments récupérés à partir d'une recherche dans l'ensemble de segments de la transcription manuelle et les segments récupérés à partir d'une recherche dans l'ensemble de segments de la transcription automatique (sans et avec adaptation).

5 Résultats et discussions

5.1 Performances en WER

Le tableau 2 présente les résultats d'adaptation du modèle de langage en utilisant la métrique WER. Nous remarquons une réduction absolue de 3,04% en WER avec un système adapté au domaine.

	SRAP sans adaptation	**SRAP avec adaptation**
Taux d'erreurs sur les mots (WER)	19,46	16,42

TABLE 2 – Résultats d'adaptation du modèle de langage en WER

5.2 Performances pour la tâche de recherche d'information

La figure 1 présente le taux de couverture des documents récupérés à partir de requêtes construites sur des transcriptions automatiques, avec (lignes continues) ou sans (pointillés) adaptation des modèles de langage, par rapport aux documents récupérés à partir de requêtes construites sur des transcriptions manuelles. Le taux de couverture est calculé en fonction du nombre de documents visés (de 1 à 20). Nous avons également expérimenté différents types de requêtes composées de 1 à 5 mots (k dans la figure 1) extraits par TF-IDF. Les résultats montrent que la transcription avec adaptation surpasse la

1. https ://www.lemurproject.org

transcription sans adaptation en termes de récupération des ressources pertinentes, pour toutes les tailles de requêtes.

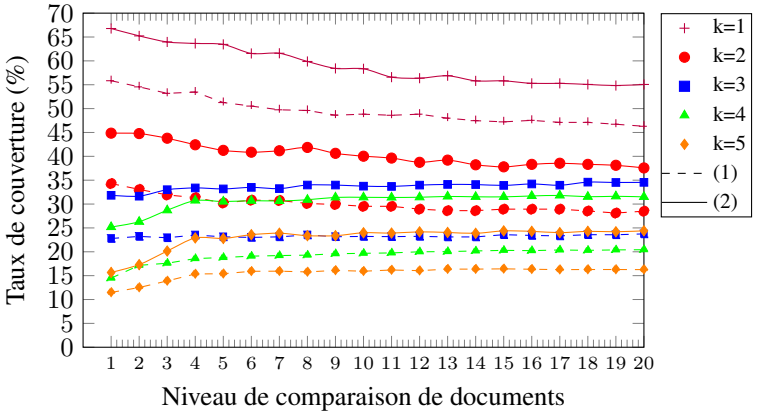

FIGURE 1 – Tâche de recherche d'information : comparaison du taux de couverture entre les requêtes construites à partir de segments de transcriptions manuelles et de transcriptions automatiques (1) sans et (2) avec adaptation des modèles de langage.

5.3 Performances pour la tâche d'indexation

Le tableau 3 présente les scores de corrélation moyens pour l'ensemble du corpus en utilisant trois ensembles de jeux de requête : les 40 mots les plus fréquents de la transcription, les mots des titres des diapositives, les mots clés de la transcription. Le coefficient de Spearman varie entre -1 et +1. Une valeur proche de +1 indique une forte corrélation entre les deux listes de documents renvoyés par la recherche alors qu'une valeur proche de 0 indique une faible corrélation (-1 indique une forte corrélation mais dans un sens opposé). Le tableau 3 présente le score moyen de corrélation pour l'ensemble du corpus. Ici également, les résultats indiquent une meilleure indexabilité obtenue après adaptation du modèle de langage du SRAP, pour tous les jeux de requêtes.

Jeux de requêtes	**SRAP sans adaptation**	**SRAP avec adaptation**
Les 40 mots les plus fréquents de la transcription manuelle	0,367	0,498
Les mots des titres des diapositives	0,458	0,588
Les mots clés annotés à partir de la transcription manuelle	0,288	0,516

TABLE 3 – Évaluation de l'indexabilité des transcriptions : comparaison des résultats d'extraction avec le coefficient de corrélation de rang de Spearman, en utilisant différents jeux de requêtes

5.4 Discussions

Comme nous l'avons vu dans notre cadre expérimental, l'adaptation automatique de modèles de langage pour la reconnaissance de la parole permet de réaliser environ trois erreurs de moins pour

cent mots transcrits (WER passant de 19,46% à 16,42%), ce qui correspond à une réduction du WER d'environ 15,6%. Ces valeurs, bien qu'intéressantes, ne mettent pas en avant certains phénomènes très intéressants liés aux tâches finales pour lesquelles les transcriptions automatiques sont générées.

En termes de recherche d'information, par exemple, nous constatons une augmentation du taux de couverture des documents retrouvés (par rapport aux documents qui auraient été trouvés à partir de requêtes extraites de transcriptions manuelles) qui peut dépasser 28,5% (k=1, niveau 1, taux de couverture passant de 56% à 67%). Enfin, en termes d'indexabilité, nous montrons que, dans cette étude, le taux de corrélation de Spearman (par rapport à l'indexation obtenue par des transcriptions manuelles) peut augmenter de plus de 79% (de 0,288 à 0,516) pour les termes les plus importants du document grâce à l'adaptation des modèles de langage.

Les résultats présentés montrent que le WER ne permet pas de bien mesurer les performances pour les tâches de recherche d'information et d'indexation considérées, puisqu'il masque les gains réels apportés par l'adaptation des modèles de langage sur les tâches visées : cela prouve la nécessité d'utiliser de nouvelles mesures, telles que celles présentées, pour évaluer l'apport réel de l'adaptation des modèles de langage.

6 Conclusion

Le taux d'erreurs sur les mots WER n'est pas toujours la meilleure mesure à utiliser pour évaluer les systèmes de reconnaissance de la parole. Une meilleure compréhension de l'impact des erreurs de transcription automatique implique nécessairement le développement de meilleures mesures (ou métriques) pour l'évaluation, afin de prendre en compte le contexte applicatif dans lequel les SRAP sont utilisés. Dans cet article, nous avons proposé l'utilisation de deux mesures d'évaluation extrinsèque pour évaluer la capacité de créer des requêtes pertinentes pour l'extraction d'information et l'indexabilité de transcriptions automatiques. Nous avons appliqué ces méthodes d'évaluation pour mesurer l'impact de l'adaptation des ML dans le contexte de cours magistraux. Les résultats obtenus montrent que le taux d'erreur sur les mots est une métrique insuffisante qui masque les apports effectifs de l'adaptation des modèles de langage.

Remerciements

Nous remercions l'agence ANR pour son financement à travers le projet PASTEL sous le numéro de contrat ANR-16-CE33-0007.

Références

BELL P., YAMAMOTO H., SWIETOJANSKI P., WU Y., MCINNES F., HORI C. & RENALS S. (2013). A lecture transcription system combining neural network acoustic and language models. In *Proc. of the Annual Conference of the International Speech Communication Association (Interspeech'13)*, p. 3087–3091.

CERVA P., SILOVSKY J., ZDANSKY J., NOUZA J. & MALEK J. (2012). Real-time lecture transcription using asr for czech hearing impaired or deaf students. In *Proc. of the Annual Conference of the International Speech Communication Association (Interspeech'12)*.

FAVRE B., CHEUNG K., KAZEMIAN S., LEE A., LIU Y., MUNTEANU C., NENKOVA A., OCHEI D., PENN G., TRATZ S. *et al.* (2013). Automatic human utility evaluation of ASR systems : Does WER really predict performance ? In *Proc. of the Annual Conference of the International Speech Communication Association (Interspeech'13)*, p. 3463–3467.

HÜRST W., KREUZER T. & WIESENHÜTTER M. (2002). A qualitative study towards using large vocabulary automatic speech recognition to index recorded presentations for search and access over the web. In *Proc. of the International Conference on WWW/Internet (ICWI'02)*, p. 135–143.

JANNET M. A. B., GALIBERT O., ADDA-DECKER M. & ROSSET S. (2015). How to evaluate ASR output for named entity recognition ? In *Proc. of the Annual Conference of the International Speech Communication Association (Interspeech'15*.

KAWAHARA T., NEMOTO Y. & AKITA Y. (2008). Automatic lecture transcription by exploiting presentation slide information for language model adaptation. In *Proc. of the International Conference on Acoustics, Speech and Signal Processing (ICASSP'08)*, p. 4929–4932.

LECORVÉ G., GRAVIER G. & SÉBILLOT P. (2008). An unsupervised web-based topic language model adaptation method. In *Proc. of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP'08)*, p. 5081–5084.

LUZZATI D., GROUIN C., VASILESCU I., ADDA-DECKER M., BILINSKI E., CAMELIN N., KAHN J., LAILLER C., LAMEL L. & ROSSET S. (2014). Human annotation of ASR error regions : Is" gravity" a sharable concept for human annotators ? In *Proc. of the International Conference on Language Resources and Evaluation (LREC'14)*, p. 3050–3056.

MARTÍNEZ-VILLARONGA A., MIGUEL A., ANDRÉS-FERRER J. & JUAN A. (2013). Language model adaptation for video lectures transcription. In *Proc. of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP'13)*, p. 8450–8454.

MDHAFFAR S., LAURENT A. & ESTÈVE Y. (2018). Le corpus PASTEL pour le traitement automatique de cours magistraux. In *Actes de la Conférence sur le Traitement Automatique des Langues Naturelles (TALN'18)*.

PALLETT D. S. (2003). A look at NIST's benchmark ASR tests : Past, Present, and Future. In *Proc. of the Workshop on Automatic Speech Recognition and Understanding (ASRU'03)*, p. 483–488.

POVEY D., GHOSHAL A., BOULIANNE G., BURGET L., GLEMBEK O., GOEL N., HANNEMANN M., MOTLICEK P., QIAN Y., SCHWARZ P. *et al.* (2011). The Kaldi speech recognition toolkit. In *Proc. of the Workshop on Automatic Speech Recognition and Understanding (ASRU'11)*.

POVEY D., PEDDINTI V., GALVEZ D., GHAHREMANI P., MANOHAR V., NA X., WANG Y. & KHUDANPUR S. (2016). Purely sequence-trained neural networks for ASR based on lattice-free mmi. In *Proc. of the Annual Conference of the International Speech Communication Association (Interspeech'16)*, p. 2751–2755.

ROUSSEAU A., BOULIANNE G., DELÉGLISE P., ESTÈVE Y., GUPTA V. & MEIGNIER S. (2014). LIUM and CRIM ASR system combination for the REPERE evaluation campaign. In *Proc. of the International Conference on Text, Speech, and Dialogue (TSD'14)*, p. 441–448.

YAMAZAKI H., IWANO K., SHINODA K., FURUI S. & YOKOTA H. (2007). Dynamic language model adaptation using presentation slides for lecture speech recognition. In *Proc. of the Annual Conference of the International Speech Communication Association (ICSLP'07)*.