Spécialisation de modèles neuronaux pour la transcription phonémique : premiers pas vers la reconnaissance de mots pour les langues rares

Cécile Macaire, Guillaume Wisniewski, Séverine Guillaume, Benjamin Galliot, Guillaume Jacques, Alexis Michaud, Solange Rossato, Minh-Châu Nguyễn, Maxime Fily

▶ **To cite this version:**

Cécile Macaire, Guillaume Wisniewski, Séverine Guillaume, Benjamin Galliot, Guillaume Jacques, et al.. Spécialisation de modèles neuronaux pour la transcription phonémique : premiers pas vers la reconnaissance de mots pour les langues rares. Journées scientifiques du Groupement de recherche "Linguistique informatique, formelle et de terrain" (GDR LIFT), Dec 2021, Grenoble, France. halshs-03475443

HAL Id: halshs-03475443
https://shs.hal.science/halshs-03475443

Submitted on 10 Dec 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial - ShareAlike 4.0 International License

Spécialisation de modèles neuronaux pour la transcription phonémique : premiers pas vers la reconnaissance de mots pour les langues rares

Cécile Macaire[1, 3] Guillaume Wisniewski[2] Séverine Guillaume[1]
Benjamin Galliot[1] Guillaume Jacques[4] Alexis Michaud[1] Solange Rossato[3]
Minh-Châu Nguyên[1, 3] Maxime Fily[1]

(1) Langues et Civilisations à Tradition Orale (LACITO), Unité Mixte de Recherche 7107 CNRS - Sorbonne Nouvelle - Institut National des Langues et Civilisations Orientales (INALCO)

(2) Laboratoire de Linguistique Formelle (LLF), Unité Mixte de Recherche 7110 CNRS - Université de Paris

(3) Laboratoire d'Informatique de Grenoble (LIG), Unité Mixte de Recherche 5217 CNRS - Université Grenoble Alpes - Grenoble INP - Institut national de recherche en informatique et en automatique (INRIA)

(4) Centre de Recherches Linguistiques sur l'Asie Orientale (CRLAO), Unité Mixte de Recherche 8563 CNRS - École des Hautes Études en Sciences Sociales - Institut National des Langues et Civilisations Orientales

cecile.macaire@univ-grenoble-alpes.fr, Guillaume.Wisniewski@u-paris.fr, severine.guillaume@cnrs.fr, b.g01lyon@gmail.com, rgyalrongskad@gmail.com, alexis.michaud@cnrs.fr, Solange.Rossato@univ-grenoble-alpes.fr, minhchau.ntm@gmail.com, maxime.fily@gmail.com

RÉSUMÉ

Nous décrivons les résultats les plus récents que nous avons obtenus dans le cadre du développement d'outils de Traitement Automatique des Langues (TAL) pour réduire l'effort de transcription et d'annotation que doivent fournir les linguistes « de terrain » au fil de leur travail de documentation et description de langues rares. En particulier, nous montrons comment une nouvelle approche neuronale fondée sur la spécialisation d'un modèle de représentation générique permet d'améliorer significativement la qualité de la transcription phonémique automatique, et surtout d'envisager la reconnaissance automatique de mots, approchant ainsi du stade de la *reconnaissance automatique de la parole* au sens plein du terme.

ABSTRACT

A first step towards automatic word recognition for low-resource languages

We describe the latest results we have obtained in the development of NLP (Natural Language Processing) tools to reduce the transcription and annotation workload of field linguists, as part of workflows to document and describe the world's languages. We show how a new deep learning approach based on the fine-tuning of a generic representation model allows to significantly improve the quality of automatic phonemic transcription, and, more significantly, to take a first step towards automatic word recognition for low-resource languages.

MOTS-CLÉS : documentation linguistique assistée par ordinateur, reconnaissance automatique de la parole, modèles neuronaux, science ouverte, linguistique de terrain.

KEYWORDS: Computational Language Documentation, Automatic Speech Recognition, Open Science, Deep Learning, linguistic fieldwork.

1 Introduction

L'amélioration significative de la qualité des outils de Traitement Automatique des Langues (TAL) ouvre de nouvelles perspectives pour faciliter le travail des linguistes de terrain (Anastasopoulos et al., 2020; Partanen et al., 2020; Hjortnaes et al., 2021). Dans nos travaux précédents (Wisniewski et al., 2020a; Michaud et al., 2018), nous avons montré que les méthodes fondées sur les réseaux de neurones permettent de développer des systèmes de reconnaissance phonémique qui aident à la transcription. Ces méthodes sont désormais intégrées dans l'outil ELPIS (Foley et al., 2018), doté d'une interface graphique conviviale (Wisniewski et al., 2020b).

Nous décrivons ici les résultats les plus récents que nous avons obtenus dans le cadre du développement d'outils de TAL pour réduire l'effort d'annotation des linguistes de terrain. Nous montrons comment une nouvelle approche neuronale fondée sur la spécialisation d'un modèle de représentation générique *(fine-tuning)* permet d'améliorer encore la qualité de la transcription phonémique, et surtout de passer à la reconnaissance automatique d'entités de plus haut niveau, à savoir des mots.

2 Spécialisation de modèles pour la transcription phonémique

Principe L'approche mise en œuvre repose sur la spécialisation d'un modèle de représentation multilingue du signal, une méthode introduite par Conneau et al. (2020) pour développer des modèles de reconnaissance de la parole à partir de peu de données.

Dans une première étape, `XLSR-53`, un modèle multilingue appris de manière non supervisée sur un corpus regroupant 56 000 heures d'enregistrements en 53 langues, est utilisé pour construire automatiquement une représentation du signal. Dans une seconde étape, ces représentations sont utilisées en entrée d'un système de reconnaissance phonémique, entraîné à partir de données associées à une transcription manuelle fournie par le/la linguiste.

Utilisation pour la prédiction de phonème Nous avons simplement défini un jeu d'étiquettes correspondant à l'ensemble des caractères composant les phonèmes. Nos expériences passées (Wisniewski et al., 2020a) ont en effet montré que la prédiction des caractères composant les phonèmes (et non pas directement des phonèmes) permettait d'obtenir de bonnes prédictions tout en faisant l'économie de l'étape qui consiste à lister explicitement les phonèmes de la langue. À ce jeu d'étiquettes s'ajoute l'espace, pour délimiter les mots, et par là, s'approcher un peu plus du développement d'un véritable système de reconnaissance de la parole pour les langues rares.

3 Résultats expérimentaux

Nous avons testé la méthode décrite ci-dessus sur deux langues minoritaires de Chine : le na et le japhug. Cela soulève plusieurs défis. Tout d'abord, la quantité de données disponible est très faible. Certes, parmi les langues rares, ces deux langues ne sont pas les moins bien documentées, loin de là. Les corpus transcrits, disponibles dans Zenodo (Galliot et al., 2021) et dans la collection Pangloss (Michaud et al., 2016), sont conséquents : de l'ordre de 3h30 pour le na et 32h pour le japhug. Il faut toutefois mettre ces chiffres en rapport avec les tailles de corpus utilisées pour les langues « courantes » : le corpus libre COMMONVOICE contenait, en 2019, 173h d'audio annoté

pour le français et 780h pour l'anglais (Ardila et al., 2020) ; le modèle de représentation de la parole le plus récent de Facebook, XSL-R, est appris sur 436 000h (quatre cent trente-six mille heures !) d'audio, regroupant 128 langues (Babu et al., 2021), soit une moyenne de plus de 3 000h par langue (moyenne qui cache certes de grandes disparités, mais fournit un ordre de grandeur). En outre, le japhug et le na possèdent des caractéristiques structurelles propres. Par exemple, le système tonal du na (Michaud, 2017) a une organisation fondamentalement différente de celui des 2 langues tonales (sur 53) du corpus multilingue utilisé (`XLSR-53`) : le mandarin et le vietnamien ; et le japhug présente un degré de complexité morphosyntaxique particulièrement impressionnant au vu de son contexte aréal (Jacques, 2021).

La qualité de notre système est évalué en utilisant deux métriques usuelles : le taux d'erreur sur les caractères, *character error rate* (CER), distance d'édition entre la référence et la prédiction calculée au niveau des caractères, et le taux d'erreur sur les mots, *word error rate* (WER), une métrique similaire calculée au niveau des mots.

Le tableau 1 présente les principaux résultats (pour toutes précisions, on se permet de renvoyer à Macaire 2021). Il en ressort que l'approche proposée permet d'obtenir des transcriptions phonémiques de bonne qualité. Le CER pour les deux langues est inférieur à 8%, soit une réduction de 4 points pour le japhug et de 6 points pour le na par rapport aux précédents résultats (Wisniewski et al., 2020b), lesquels reposaient sur une méthode de transcription phonémique qui était également fondée sur un réseau de neurones, mais qui apprenait une représentation du signal uniquement à partir des données d'apprentissage, sans utiliser un modèle pré-entraîné. Il faut toutefois noter que l'erreur au niveau des mots est bien plus élevée que l'erreur au niveau des caractères, mais cette différence est essentiellement liée à la manière dont les deux mesures d'évaluation sont définies : dans la mesure où il y a nettement moins de mots dans une phrase que de caractères, une erreur au niveau d'un caractère (qui se traduit naturellement par une erreur au niveau du mot le contenant) aura un impact beaucoup plus fort sur le WER que sur le CER. Une analyse plus fine des résultats montre que nos systèmes ne font que très peu d'erreurs sur les frontières de mots, aussi bien pour le na que pour le japhug (près de 90% des espaces sont correctement prédits).

	taille corpus apprentissage (mn)	taille corpus test (mots)	WER (%)	CER (%)
na				
évaluation	180	—	41.5	7.9
correction	180	71	38.5	5.7
japhug				
évaluation	600	—	18.5	7.4
correction	600	236	5.4	1.3

TABLE 1 – Résultats obtenus en spécialisant les représentations construites par `XLSR-53` pour la transcription phonémique. Les hypothèses sont évaluées soit par rapport à une référence pré-existante (condition *évaluation*) soit par rapport à une référence obtenue en corrigeant les prédictions du système (colonne *correction*).

Les linguistes de l'équipe ont corrigé quelques transcriptions automatiques. Cette expérience-pilote n'était pas systématisée comme celle de Sperber et al. (2017) (ou d'autres études des processus de *post-édition* en traduction automatique). Elle ne concerne que 71 mots pour le na, et 236 mots pour le japhug (voir, à nouveau, le tableau 1). Elle débouche néanmoins sur une observation claire : le

nombre de corrections à effectuer est beaucoup plus faible que ne le suggère le taux d'erreurs (CER).

Cette observation, bien que peu surprenante (des résultats similaires ont été observés dans le cas de l'évaluation de la traduction automatique), est particulièrement importante : elle suggère que la qualité « réelle » des systèmes est plus élevée que ne le suggéraient les métriques d'évaluation employées jusqu'ici. Au moins dans le cas du na et du japhug, l'effort demandé pour corriger des transcriptions automatiques est considéré comme très faible par les cinquième et sixième auteurs du présent travail (qui sont les linguistes ayant recueilli les corpus na et japhug).

La qualité des prédictions au niveau des mots reste nettement en-deçà de celle obtenue sur des langues « bien dotées ». Ces résultats nous paraissent néanmoins remarquables, et tout à fait encourageants pour l'avenir des efforts conjoints associant TAListes et linguistes de terrain.

Références

Anastasopoulos, A., Cox, C., Neubig, G., and Cruz, H. (2020). Endangered languages meet modern NLP. In *Proceedings of the 28th International Conference on Computational Linguistics : Tutorial Abstracts*, pages 39–45, Barcelona, Spain (Online). International Committee for Computational Linguistics. https://aclanthology.org/2020.coling-tutorials.7/.

Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F. M., and Weber, G. (2020). Common voice : a massively-multilingual speech corpus. *arXiv preprint arXiv :1912.06670*.

Babu, A., Wang, C., Tjandra, A., Lakhotia, K., Xu, Q., Goyal, N., Singh, K., von Platen, P., Saraf, Y., Pino, J., Baevski, A., Conneau, A., and Auli, M. (2021). XLS-R : Self-supervised cross-lingual speech representation learning at scale. *arXiv preprint arXiv :2111.09296*.

Conneau, A., Baevski, A., Collobert, R., Mohamed, A., and Auli, M. (2020). Unsupervised cross-lingual representation learning for speech recognition. *CoRR*, abs/2006.13979. https://arxiv.org/abs/2006.13979.

Foley, B., Arnold, J., Coto-Solano, R., Durantin, G., and Ellison, T. M. (2018). Building speech recognition systems for language documentation : the CoEDL Endangered Language Pipeline and Inference System (ELPIS). In *Proceedings of the 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU), 29-31 August 2018*, pages 200–204, Gurugram, India. ISCA. https://www.isca-speech.org/archive/SLTU_2018/pdfs/Ben.pdf.

Galliot, B., Wisniewski, G., Guillaume, S., Besacier, L., Jacques, G., Michaud, A., Rossato, S., Nguyên, M.-C., and Fily, M. (2021). Deux corpus audio transcrits de langues rares (japhug et na) normalisés en vue d'expériences en traitement du signal. In *Journées scientifiques du Groupement de recherche "Linguistique informatique, formelle et de terrain" (GDR LIFT)*, Grenoble.

Hjortnaes, N., Partanen, N., and Tyers, F. M. (2021). Keyword spotting for audiovisual archival search in uralic languages. In *Proceedings of the Seventh International Workshop on Computational Linguistics of Uralic Languages*, pages 1–7, Syktyvkar, Russia (Online). Association for Computational Linguistics.

Jacques, G. (2021). *A grammar of Japhug*. Number 1 in Comprehensive Grammar Library. Language Science Press, Berlin.

Macaire, C. (2021). Recognizing lexical units in low-resource language contexts with supervised and unsupervised neural networks. Research report, LACITO (UMR 7107). https://hal.archives-ouvertes.fr/hal-03429051.

Michaud, A. (2017). *Tone in Yongning Na : lexical tones and morphotonology*. Number 13 in Studies in Diversity Linguistics. Language Science Press, Berlin. 10.5281/zenodo.439004.

Michaud, A., Adams, O., Cohn, T., Neubig, G., and Guillaume, S. (2018). Integrating automatic transcription into the language documentation workflow : experiments with Na data and the Persephone toolkit. *Language Documentation and Conservation*, 12 :393–429. http://hdl.handle.net/10125/24793.

Michaud, A., Guillaume, S., Jacques, G., Mac, D.-K., Jacobson, M., Pham, T.-H., and Deo, M. (2016). Contribuer au progrès solidaire des recherches et de la documentation : la Collection Pangloss et la Collection AuCo. In *Journées d'Etude de la Parole 2016*, volume 1, pages 155–163. https://halshs.archives-ouvertes.fr/halshs-01341631/.

Partanen, N., Hämäläinen, M., and Klooster, T. (2020). Speech recognition for endangered and extinct samoyedic languages. *CoRR*, abs/2012.05331. https://arxiv.org/abs/2012.05331.

Sperber, M., Neubig, G., Niehues, J., Nakamura, S., and Waibel, A. (2017). Transcribing against time. *Speech Communication*, 93 :20–30.

Wisniewski, G., Guillaume, S., and Michaud, A. (2020a). Phonemic transcription of low-resource languages : To what extent can preprocessing be automated ? In Beermann, D., Besacier, L., Sakti, S., and Soria, C., editors, *Proceedings of the 1st Joint SLTU (Spoken Language Technologies for Under-resourced languages) and CCURL (Collaboration and Computing for Under-Resourced Languages) Workshop*, pages 306–315, Marseille, France. European Language Resources Association (ELRA). https://halshs.archives-ouvertes.fr/hal-02513914.

Wisniewski, G., Michaud, A., Galliot, B., Besacier, L., Guillaume, S., Aplonova, K., and Jacques, G. (2020b). Ouvrir aux linguistes « de terrain » un accès à la transcription automatique. In Poibeau, T., Parmentier, Y., and Schang, E., editors, *2èmes journées scientifiques du Groupement de Recherche Linguistique Informatique Formelle et de Terrain (LIFT)*, pages 83–94, Montrouge, France. CNRS. https://hal.archives-ouvertes.fr/hal-03047148.