Les modèles pré-entraînés à l'épreuve des langues rares : expériences de reconnaissance de mots sur la langue japhug (sino-tibétain)

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

Contexte / Objectifs à long terme

Faire bénéficier la documentation et l'étude des langue rares des outils de pointe en TAL
Travail interdisciplinaire associant linguistes de terrain, chercheurs en informatique…
Enjeu pour l'étude de la parole : disponibilité de données fiables et abondantes sur la diversité des langues

Une approche de reconnaissance en deux étapes

► Entraînement non-supervisé par XLSR-53[1] : reconnaissance du signal de parole sur 53 langues (56 000 h)
▹ Modèle de représentation du signal
► Pré-entraînement sur un sous-ensemble transcrit du corpus de japhug
▹ utilisation de données de terrain déposées dans la collection Pangloss (audio + transcription phonémique)
Puis application à la reconnaissance des phonèmes (& des espaces) sur de l'audio.
➪Évaluation à l'aune de deux métriques classiques : CER (character error rate) et WER (word error rate)

Résultats

Audio en langue japhug transcrit automatiquement :

tɕe kɯɕɯŋgɯ tɕe iɕqha @mingchao(u→ˌ)ɯraŋg nɯ-tɕu pjɤŋu tɕendɤre iɕqha nɤki @yanguo kɤti rjɤlkhɤβ ɣɯ nɯrjɤlpu nɯ kɯ, iɕqha nɯ, iɕqha nɯ ɯftsa nɯnɯ rjɤlpu lusɯndɤm pjɤsɯso. tɕe nɯ rjɤlpu lusɯndɤm pjɤsɯso tɕe, tɕendɤre nɤkinɯ, sɤtɕha ra tosɤtʂoʁloʁnɯ ʑo ɕti tɕe, tɕendɤre iɕqha nɯ, @shandong nɯtɕu ɯrmi @zhangxiaobing kɯrmi ci tɯtsɣe ɯkɯβzu ci pjɤtu, tɤtɕɯ. ɕendɤre ɯrʑaβ nɯ ɯskhrɯ mɯjɤβdi χsɯsla ma mɯtoβzu ri […]

Sortie d'un modèle entraîné sans ponctuation.
En rouge : corrections manuelles du linguiste

tɕe kɯɕɯŋgɯ tɕe iɕqha-, @mingchao ɯraŋ nɯtɕu pjɤŋu-ˌ tɕendɤre iɕqha, nɤki-; @yanguo kɤti rjɤlkhɤβ ɣɯ-, nɯrjɤlpu nɯ kɯ, iɕqha nɯ(.→,) iɕqha nɯ, ɯftsa nɯnɯ rjɤlpu lusɯndɤm pjɤsɯso. tɕe nɯ rjɤlpu lusɯndɤm pjɤsɯso tɕe, tɕendɤre, nɤkinɯ, sɤtɕha ra tosɤtʂoʁloʁnɯ ʑo ɕti tɕe, tɕendɤre iɕqha nɯ, @shandong nɯtɕu, ɯrmi @zhangxiaobing kɯrmi ci, tɯtsɣe ɯkɯβzu ci pjɤtu, tɤtɕɯ. ɕendɤre ɯrʑaβ nɯ ɯskhrɯˌ mɯjɤβd(er,→i) χsɯslaˌ ma mɯtoβzu ri […]

Sortie d'un modèle entraîné avec ponctuation.
En rouge : corrections manuelles du linguiste

De bons résultats dès 2h à 3h de données :

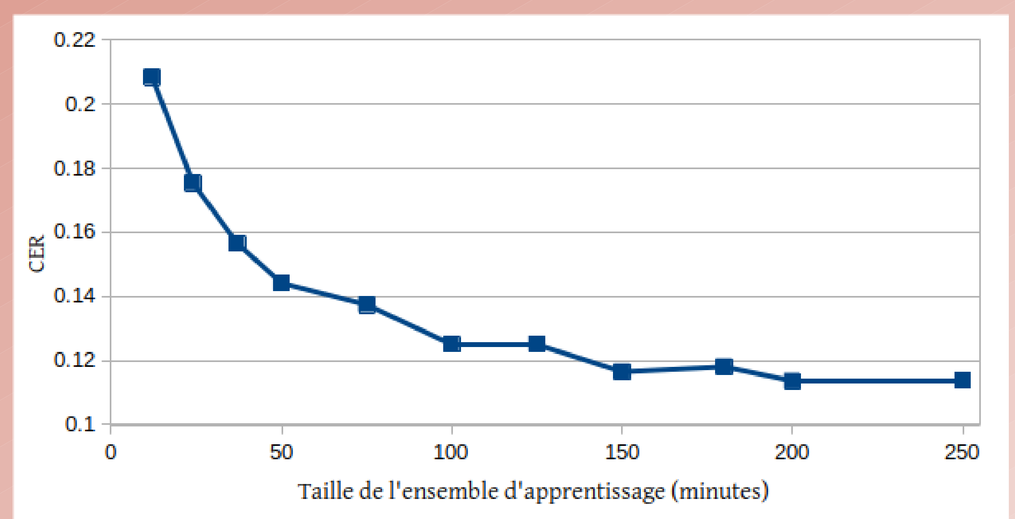

Figure 1 – Évolution des performances en fonction de la taille du corpus d'apprentissage

Les mystères de l'apprentissage machine :

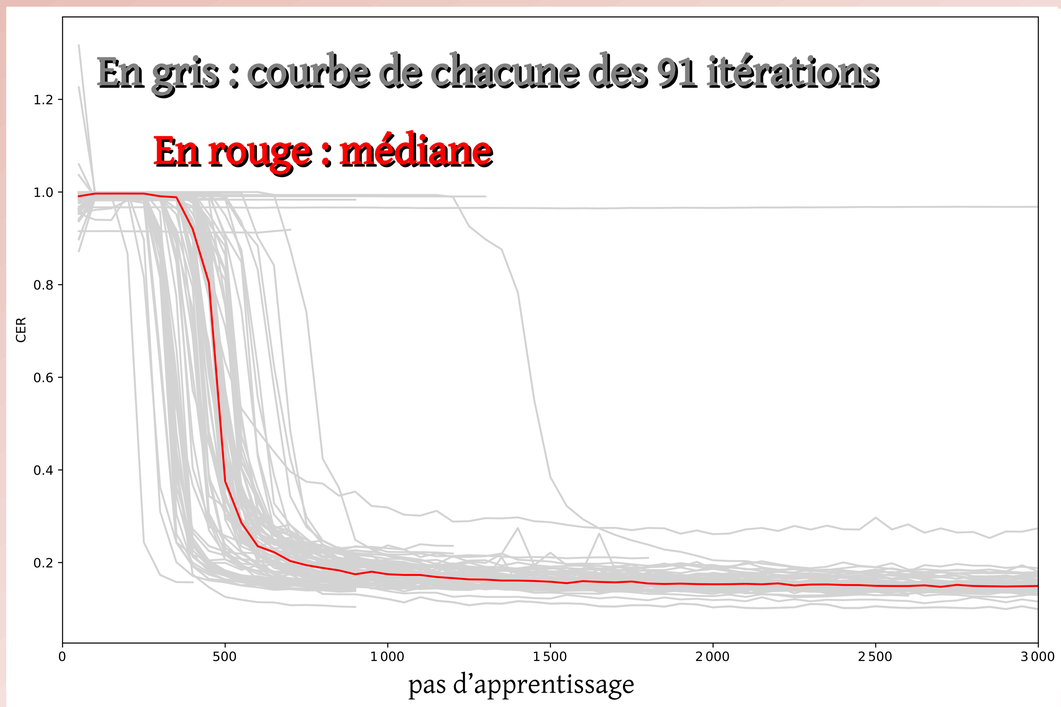

Figure 2 - CER sur l'ensemble de validation au cours de différentes itérations d'optimisation

Perspectives

Gains en temps de transcription : après production "manuelle" d'un corpus de 2 à 3h, le modèle statistique produit une transcription qui sert de point de départ pour la suite.

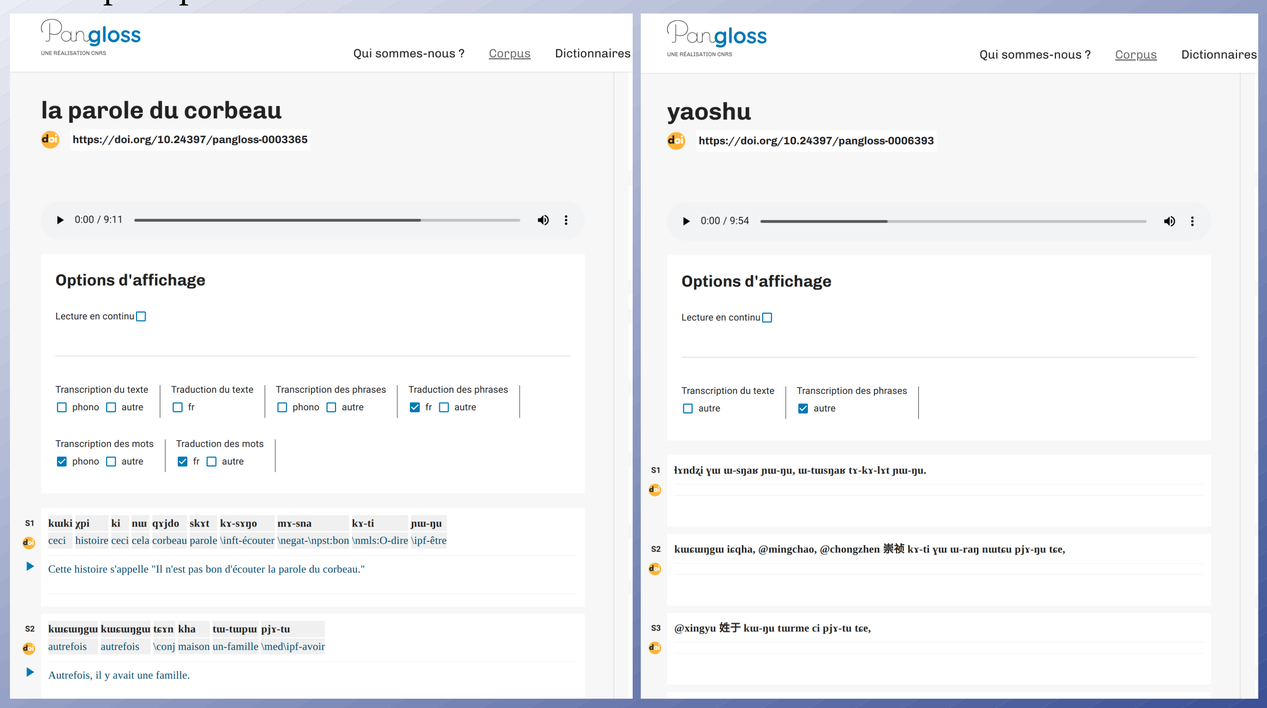

Figure 3 – Extraits Pangloss du corpus Pangloss[2] de langue japhug de Guillaume Jacques.
À gauche : un travail complet de transcription-glose-traduction. À droite : transcription seule..

Conclusions

➢Un résultat très satisfaisant au plan qualitatif
➢ Des perspectives applicatives prometteuses pour les langues peu dotées
➢ Des thèmes porteurs pour la recherche interdisciplinaire en parole

1 https://huggingface.co/transformers/model_doc/xlsr_wav2vec2.html
2 https://pangloss.cnrs.fr