Character Level ASR for 3 low resource African Languages

Aissatou Ndoye
Lloyd Acquaye Thompson
Jama Hussein Mohamud

July 6, 2020

1 Introduction

Supervised learning in Automatic Speech Recognition(ASR) systems require volumes of data to train, that is a requirement that low-resource languages lack. Therefore, in this project, we trained three different African languages in an unsupervised fashion. We learn representations with Contrastive Predictive Coding - an unsupervised learning approach that learned to extract useful features from high-dimensional data but totally different domain.

1.1 Data and preprocessing

In this project, we considered three different languages(Somali, Wolof and Ga). We learned representations both multilingually and monolingually. 1) Somali is an *Afroasiatic* language belonging to the Cushitic branch. It is an official language of Somalia and Somaliland, a national language in Djibouti, and a working language in the Somali Region of Ethiopia and also in North Eastern Kenya. It is written officially with the Latin alphabets and spoken approximately by 36.6 million speakers spread in Greater Somalia [1]. It has 22 consonants and five vowel phonemes 2) The natives of Accra speak the Ga-Adangme language, which has approximately 2 million people. The Ga language has 31-consonant phonemes, 7-oral vowels, five nasal vowels with different vowel lengths. 3) The Wolof language is from the family of Niger-Congo, which belongs to the Atlantic branch. It is spoken in Senegal, Gambia, and Mauritania, and is the native language of the ethnic group of the Wolof people. Approximately 10 million people speak Wolof. It has 29-consonant phonemes and 9-vowel phonemes. We collected public corpus of these languages and recorded via Lig-Aikuma Mobile App.

During the data collection process, we did encounter a few challenges; The Lig-Aikuma Mobile App did not offer a smooth path. It did crash on several occasions which then increased the number of corrupted files. For some reason, we noticed the app seems to work better on some phones as compared to others depending on the android version. The app is unable to accommodate longer sentences and by so doing slows down the reading pace of the reader since tokenized sentences lose the natural flow. On average Lig-Aikuma comes in handy considering the quality of audio and its ability to save the speech with linker files.

After obtaining all the audios from the app, we had to filter the audio folder by deleting all the corrupted files and making sure the linker addresses match their corresponding lines in the text files. To create a numerical representation of our text for the models, we created a character generator for all the alphabets that are recognizable by the language to generate a map-up file which contains a unique number for each alphabet in the sentence.

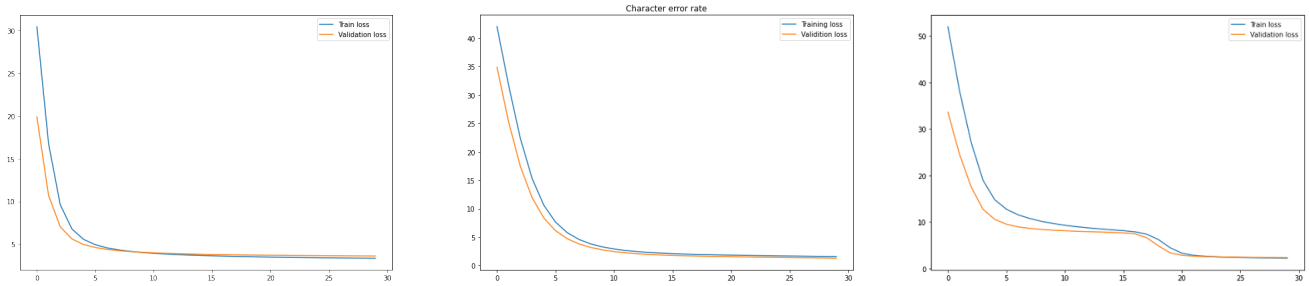
2 Contrastive Predictive Coding

Contrastive Predictive Coding(CPC) is an unsupervised representation learning method that predicts future samples via Autoregressive Models(RNN or Linear Models). The model, instead of directly using the output context from RNN or Linear layers, it learns to discriminate the real output representation at time $t+k$ from several other negative features taken elsewhere in the batch [2]. To examine the features extracted by the CPC, we measured the Character Error Rate(CER) with a linear classifier trained on

top of these features, which demonstrates how linearly separable the relevant classes are under these features.

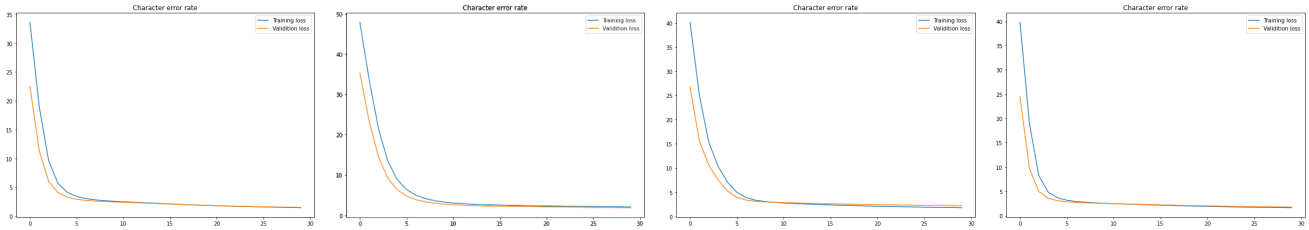
3 Experiments and Results

Initially, we independently trained models on three languages namely; Somali(Somalia), Wolof(Senegal) and Ga(Ghana) using 40 minutes as the training set, 20 minutes as the validation set and 60 minutes as the test set from each language. Then afterwards, we did a bi-lingual training by picking two languages at a time, i.e. Somali/Wolof, Wolof/Ga and Ga/Somali. In the multi-lingual training, we combined all the three languages by taking 30 minutes of the training set, 10 minutes of validation set and 30 minutes of the testing set which accumulated to 210 minutes (90 mins-Train, 30 mins-Validation, 90 mins-Test). Thus, we trained all our models for 30 epochs. Based on the of set parameters, all the curves did converge upon visualization (see figure 1 and 2)



(a) *train val loss for Somali* (b) *Train-Val loss for Wolof* (c) *Train-Valid loss for Ga*

Figure 1: *Train-Validation For monolingual Models*



(a) *Somali-Wolof* (b) *Wolof-Ga* (c) *Somali-Ga* (d) *Somali-WOL-GA*

Figure 2: *Train-Validation For multilingual Models*

From table 1 and 2, we can deduce that combining languages gives a lower CER than when trained in isolation. Somali language alone obtained test CER of 0.87, but after adding Wolof to it, the CER drops drastically to 0.54. Wolof/Ga recorded an impressive CER of 0.45 on the test. Somali/Ga performed poorly with CER of 0.92 on validation and CER of 0.91 during test time. Finally, we combined all the three languages and trained with an equal amount of data from each language, which achieved an impressive CER of 0.78 on validation and CER of 0.36 upon testing. In summary, we noticed training multiple languages outperforms cases where we had to train separately.

Languages	Validation	Test
Somali_CER	0.83	0.87
Wolof_CER	**0.71**	**0.55**
Ga_CER	0.94	0.92

Table 1: Character Error Rate (CER) for the three languages and their combination

Languages	Validation	Test
Somali_wolof_CER	0.73	0.54
Somali_Ga_CER	0.92	0.91
Wolof_Ga_CER	0.81	0.45
Somali_wolof_Ga_CER	**0.78**	**0.36**

Table 2: Character Error Rate (CER) for the combination of the three languages

4 Conclusion

Unsupervised learning have been very succesful in natural language processing (NLP) tasks where unsupervised learning methods have been trained to predict future, missing or contextual information. The learned representation with these approaches have been very succesfull in most NLP down-stream tasks even in the case where have even limited text. Similarly, in this study, we learned that similar approaches can be done with speech too. The learned representation by CPC show that we can build ASR systems even when we have limited (speech-text) data.

Details of the implementation of this project can be found on this Github-Repo

References

[1] Tillman Roder. Constitutionalism in islamic countries: Between upheaval and continuity. oup usa.

[2] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748, 2018.