Phone-ing it in: Towards Flexible, Multi-Modal Language Model Training using Phonetic Representations of Data

Colin Leong [1] Daniel Whitenack [2]

[1]University of Dayton, Dayton, OH, USA [2]SIL International, Dallas, TX USA

Abstract

Multi-modal techniques offer significant untapped potential to unlock improved NLP technology for local languages. However, many advances in language model pre-training are focused on text, a fact that only increases systematic inequalities in the performance of NLP tasks across the world's languages. In this work, we propose a multi-modal approach to train language models using whatever text and/or audio data might be available in a language. Initial experiments using Swahili and Kinyarwanda data suggest the viability of the approach for downstream Named Entity Recognition (NER) tasks, with models pre-trained on phone data showing an improvement of up to 6% F1-score above models that are trained from scratch. Preprocessing and training code will be uploaded to https://github.com/sil-ai/phone-it-in.

Introduction

Only a negligible fraction of the 7000+ currently spoken languages [3] have sufficient text corpora to train state-of-the-art language models. This data scarcity results in systematic inequalities in the performance of NLP tasks across the world's languages [2].

Local language communities that are working to develop and preserve their languages are producing diverse sets of data beyond pure text.

Thus, we propose **a multi-modal approach to train both language models and models for downstream NLP tasks using whatever text and/or audio data might be available in a language (or even in a related language).**

1. Utilize recent advances in phone recognition and text/grapheme-to-phone transliteration
2. Pre-train character-based language models in this phone-space
3. Fine-tune models for downstream tasks by mapping text-based training data into the phonetic representation

We demonstrate our phonetic approach by training Named Entity Recognition (NER) models for Swahili [swh] using various combinations of Swahili text data, Swahili audio data, Kinyarwanda [kin] text data, and Kinyarwanda audio data.

Methodology

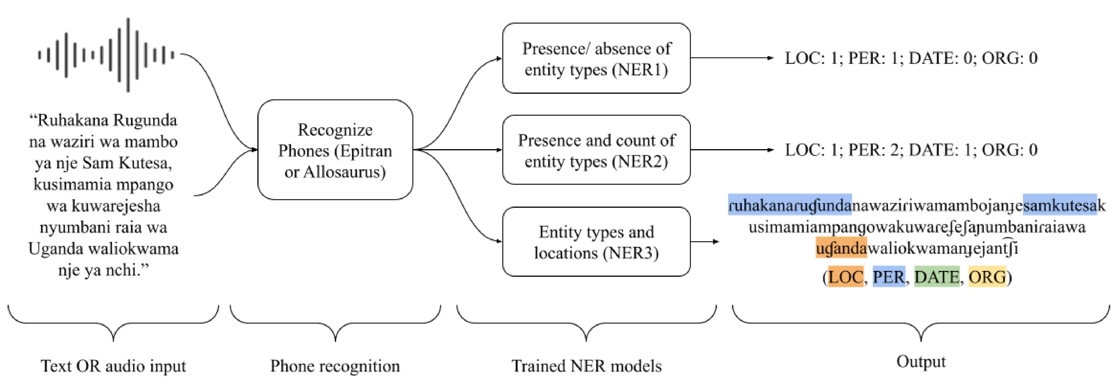

Figure 1. Our approach: input from either modality can be converted by phone recognition, e.g. Epitran for text, Allosaurus for speech. Then we test on several downstream tasks which we designate NER1, NER2, NER3.

Experiments

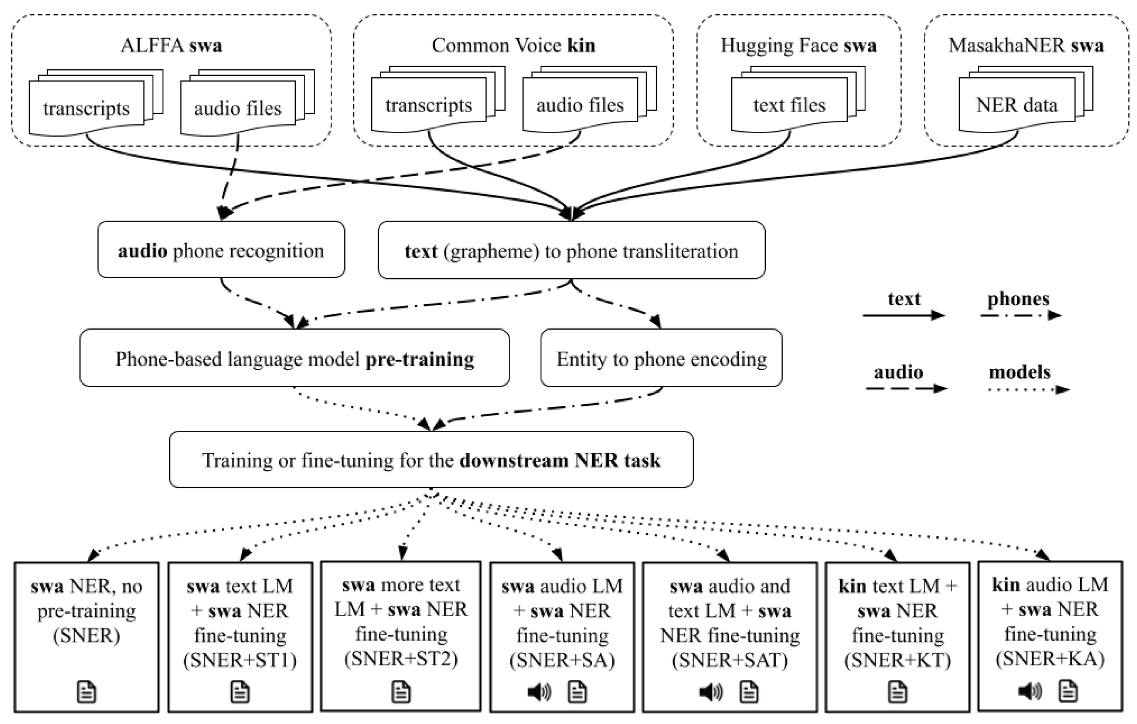

Figure 2. Training scenarios: we pre-train on various combinations of phonemized datasets, evaluating on the downstream NER task. SNER-ST denotes "**S**wahili **T**ext (ST) pre-training, **S**wahili **NER** (SNER) fine-tuning", SNER-SAT denotes Swahili NER with **S**wahili **A**udio and **T**ext (SAT) pre-training, SNER-**KA** uses **K**inyarwanda **A**udio (KA), etc.

In order to evaluate the quality of learned phonetic representations, we **transliterate several text and audio data sets in the Swahili [swh] language**.

We **pre-train phonetic language models** on various combinations of these data sets and evaluate downstream performance on NER tasks.

- The NER1 task tries to determine the presence or absence of certain kinds of entities within an input. For our task we use PER, ORG, DATE, and LOC entities.
- The NER2 task additionally requires models to predict the correct numbers of these entities within an input.
- Finally, the NER3 task requires models to determine entities at the correct locations with an input sequence of phones.

Data

- swh - The "Language Modeling Data for Swahili" dataset [6] and the ALFFA speech dataset [5]
- kin - Common Voice (CV) Kinyarwanda 6.1 subset [1]
- NER - MasakhaNER [4]

Model Architecture and Training

- All models use the SHIBA implementation of CANINE [7], a character-based model not reliant on word boundaries/spaces.
- Training runs were performed multiple times
- Results included below are averages of these multiple runs

Results and Conclusions

Model	F1 NER1	F1 NER2	F1 NER3	F1 NER3 (s)
SNER	0.829	0.753	0.357	0.161
SNER+ST1	0.827	0.770	0.401	0.213
SNER+ST2	0.824	0.747	0.394	0.166
SNER+SA	0.817	0.751	0.363	0.163
SNER+SAT	0.818	0.763	0.405	0.203
SNER+KT	0.823	**0.771**	**0.408**	**0.217**
SNER+KA	**0.846**	0.763	0.397	0.197

Table 1. Mean results for presence/absence of entity types (NER1), presence and *count* of entity types (NER2), and entity types and precise locations (NER3). Average of at least three trials per experiment. Calculated with seqeval, (s) denotes "strict" setting.

- Interestingly, the best results for [swh] NER tasks occurred when we pre-trained phone-based LMs using [kin] data.
- **Pre-training LMs does help boost performance in downstream tasks** (like NER).
- This phone-based pre-training **makes more of a difference as the difficulty of the downstream task increases**.

Future Directions

- Language specific phone recognizers for improved accuracy (we would only need 100's of samples, see Luhya case study by Siminyu et al.)
- Word segmentation to recover word boundaries
- Other datasets and languages (Common Voice or audio Bibles)
- Subwords instead of characters

References

[1] Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. *CoRR*, abs/1912.06670, 2019.

[2] Damián E. Blasi, Antonios Anastasopoulos, and Graham Neubig. Systematic inequalities in language technology performance across the world's languages. *ArXiv*, abs/2110.06733, 2021.

[3] David M. Eberhard, Gary F. Simons, and Charles D. Fennig. *Ethnologue: Languages of the World*. SIL International, Dallas, Texas, twenty-fourth edition, 2021.

[4] D. Adelani et al. MasakhaNER: Named Entity Recognition for African Languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131, 2021.

[5] Hadrien Gelas, Laurent Besacier, and Francois Pellegrino. Developments of Swahili resources for an automatic speech recognition system. In *SLTU - Workshop on Spoken Language Technologies for Under-Resourced Languages*, Cape-Town, Afrique Du Sud, 2012.

[6] Shivachi Casper Shikali and Mokhosi Refuoe. Language modeling data for Swahili, November 2019. Type: dataset.

[7] Joshua Tanner and Masato Hagiwara. SHIBA: Japanese CANINE model, 2021. Publication Title: GitHub repository.

 cleong1@udayton.edu dan_whitenack@sil.org