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.

Experiments

In order to evaluate the quality of learned phonetic representations, we transiterate 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 NER2 task requires models to determine entities at the correct locations with an input sequence of phones.

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

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 strict (b) denotes "strict" setting.

<table>
<thead>
<tr>
<th>Model</th>
<th>F1 NER1</th>
<th>F1 NER2</th>
<th>F1 NER3 (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNER</td>
<td>0.127</td>
<td>0.397</td>
<td>0.161</td>
</tr>
<tr>
<td>SNER+ST1</td>
<td>0.287</td>
<td>0.770</td>
<td>0.213</td>
</tr>
<tr>
<td>SNER+ST2</td>
<td>0.284</td>
<td>0.747</td>
<td>0.394</td>
</tr>
<tr>
<td>SNER+ST3</td>
<td>0.297</td>
<td>0.753</td>
<td>0.363</td>
</tr>
<tr>
<td>SNER+KA</td>
<td>0.798</td>
<td>0.814</td>
<td>0.203</td>
</tr>
<tr>
<td>SNER+KA+ST</td>
<td>0.798</td>
<td>0.814</td>
<td>0.203</td>
</tr>
<tr>
<td>SNER+KA+ST2</td>
<td>0.798</td>
<td>0.814</td>
<td>0.203</td>
</tr>
<tr>
<td>SNER+KA+ST3</td>
<td>0.846</td>
<td>0.763</td>
<td>0.397</td>
</tr>
</tbody>
</table>

References