Combining Pretrained High-Resource Embeddings and Subword Representations for Low-Resource Languages

Machel Reid, Edison Marrese-Taylor, Yutaka Matsuo

東京大学 THE UNIVERSITY OF TOKYO

Motivation

The contrast between the need for large amounts of data for current Natural Language Processing (NLP) techniques, and the lack thereof, is accentuated in the case of African languages, most of which are considered low-resource.

To help circumvent this issue, we explore techniques for **combining qualities of morphologically rich languages (MRLs) and leveraging pretrained word vectors in well-resourced languages (such as English)**.

We evaluate our ideas Xhosa-English translation as downstream task for evaluation.

Proposed Approach

We assume the existence the following:

1. Low Resource Vocabulary: $\mathbb{V} = \{v_1, \ldots, v_T\}$
2. The corresponding translations of the words in $\mathbb{V}$ in a high-resource language: $\mathbb{D} = \{d_1, \ldots, d_T\}$[1]
3. A pretrained embedding matrix in the high-resource language: E_{HR}

[1]Note that d_i can be comprised of a sequence of words in the case that v_i cannot be appropriately expressed in one word in the high resource language.

- To leverage the high-resource language, we embed the atomic elements *(e.g. indoda → man)* of $\mathbb{D}$ in E_{HR} and map the resulting vectors to the corresponding word in $\mathbb{V}$.
- In the case of d_i being a sequence *(e.g. bethuna [listen, everyone])*, we take a similar approach to Lazaridou et al. 2017 and sum the normalized word vectors for each word in d_i to produce a word representation for the word v_i. **We refer to the resulting embedding matrix as E_V.**
- **We pretrain another embedding matrix E_M on a corpus in our low resource language** using subword information to capture similarity correlated with the morphological nature of words

Data

For use on our downstream task: we use two datasets:

1. The Multilingual Parallel Bible Corpus (Christodoulopoulos and Steedman 2015)
2. *XhOxExamples* - New dataset created from examples collected from the isiXhosa-English Oxford Dictionary

Embedding Configurations

- **Random Initialization**
- ***Xh*Pre** - Initialization with E_V. Words not present in E_V are initialized with E_M.
- ***Xh*Sub** - Initialization with E_M only.
- **VecMap** - We learn cross-lingual word embedding mappings by taking two sets of monolingual word embeddings, E_V and E_M, and mapping them to a common space following Artexte et al., 2018.
- ***Xh*Meta** - We compute meta-embeddings for every word w_i by taking the mean of $E_V(w_i)$ and $E_M(w_i)$, following Coates & Bollegala 2019. Words not present in E_V are associated with an UNK token and its corresponding vector.

Evaluation

- On our collected datasets, we use Xhosa-English translation as a downstream task
- The generative task is evaluated using sentence level BLEU, BLEU-4 and BEER.

Results

	Bible		XhOxExamples	
Model	BLEU	BEER	BLEU-4	BEER
Random Initializaton	21.79	21.84	16.08	25.30
VecMap	22.46	22.03	16.38	25.42
*Xh*Sub	24.65	22.79	17.37	26.04
*Xh*Pre	27.67	22.40	17.06	25.70
*Xh*Meta	**29.09**	**23.33**	**17.77**	**26.44**

We show that both types of representations are important in the context of the low-resourced MRL, Xhosa, and hope that this research is able to assist others when doing NLP for African languages.

machelreid2004@gmail.com, {emarrese,matsuo}@weblab.t.u-tokyo.ac.jp