Sequence Mixup for Zero-Shot Cross-Lingual Part-Of-Speech Tagging

Megh Thakkar
BITS, Pilani
megh.1211@gmail.com

Vishwa Shah
BITS Pilani, K. K. Birla Goa Campus
f20180109@goa.bits-pilani.ac.in

Ramit Sawhney
ShareChat
ramitsawhney@sharechat.co

Debdoot Mukherjee
ShareChat
debdoot.iitd@gmail.com

Abstract

There have been efforts in cross-lingual transfer learning for various tasks. We present an approach utilizing an interpolative data augmentation method, Mixup, to improve the generalizability of models for part-of-speech tagging trained on a source language, improving its performance on unseen target languages. Through experiments on ten languages with diverse structures and language roots, we put forward its applicability for downstream zero-shot cross-lingual tasks.

1 Introduction

Recently, neural network models have obtained state-of-the-art results in part-of-speech (POS) tagging tasks across multiple languages. Since numerous languages lack suitable corpora annotated with POS labels, there have been efforts to design models for cross-lingual transfer learning. Cross-lingual learning enables us to utilize the annotated corpora of a source language to train models that are effective over a different target language. Interpolative data augmentation methods have been proposed to mitigate overfitting in models in the absence of enough training data. Sequence-based mixup (Chen et al., 2020) is an interpolative data augmentation method for named entity recognition. However, these methods have not been explored for cross-lingual transferability.

Interpolative data augmentation methods are aimed at increasing the diversity of the training distribution and, as a result, improving the generalizability of underlying models. We leverage this capability of sequence Mixup (Seq. Mixup) to capture rich linguistic information for cross-lingual transferability of POS tagging tasks for ten languages with different structures and language roots. To this end, we first measure the dataset level cosine similarity across languages, defined as the average of sentence-level embedding of dataset samples. This gives an overview of the syntactical and semantical relationship among the different languages. We then finetune multilingual models over a source language using sequence Mixup and evaluate it across a target language for varying language similarities, probing sequence-based interpolative data augmentation. We also evaluate sequence Mixup on combinations of similar and dissimilar languages and verify its transferability on various target languages.

2 Methodology

Mixup (Zhang et al., 2018) is a data augmentation technique that generates virtual training samples from convex combinations of individual inputs and labels. For a pair of data points (x, y) and (x', y'), Mixup creates a new sample $(\widetilde{x}, \widetilde{y})$ by interpolating the data points using a ratio λ, sampled from a Beta distribution, where $\widetilde{x} = \lambda \cdot x + (1-\lambda) \cdot x'$ and corresponding mixed label $\widetilde{y} = \lambda \cdot y + (1-\lambda) \cdot y'$.

We perform Mixup over the latent space representations for interpolating sequences. Pair of sentences (x, y) and (x', y') are randomly sampled and interpolated in the hidden space using a L-layer encoder $f(:, \theta)$. The hidden layer representations for x and x' upto the k^{th} layer are given as,

$$\begin{aligned} h_l &= f_l(h_{l-1}; \theta), \quad l \in [1, k] \\ h'_l &= f_l(h'_{l-1}; \theta), \quad l \in [1, k] \end{aligned} \tag{1}$$

At the k^{th} layer, the hidden representations of each token in x are linearly interpolated with each token in x'. After this, $\widetilde{h}_k$ is fed to the upper layers,

$$\begin{aligned} \widetilde{h}_k &= \lambda h_k + (1-\lambda) h'_k \\ \widetilde{h}_l &= f_l(\widetilde{h}_{l-1}; \theta), \quad l \in [k+1, L] \end{aligned} \tag{2}$$

We evaluate the performance of zero-shot learning on sequence Mixup, where the model is trained on one source language or a set of source languages and evaluated on a target language. To choose the source and target, for each language we average sentence level embeddings of dataset instances and use the average embeddings as language representation. Using these representations, we find the cosine similarity among the languages as shown in

Proceedings of the 1st Workshop on Multilingual Representation Learning, pages 245–247
November 11, 2021.

Figure 1. This approach can be extended for tasks on under-resourced languages, where models can be trained on a similar high-resourced language or a set of languages.

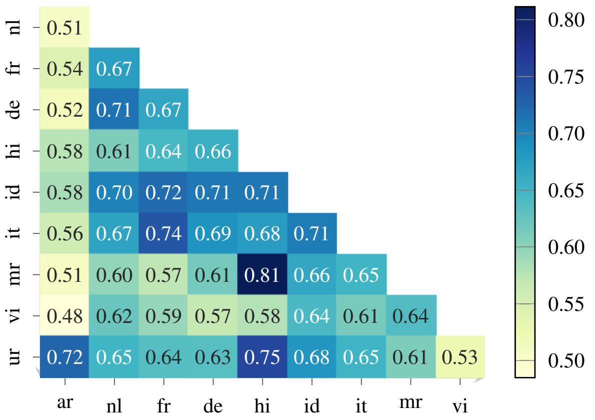

Figure 1: Cosine Similarity of the ten languages.

3 Experiments

We evaluate our approach on POS tagging with datasets from the Universal Dependencies (UD) dataset [1] for ten different languages - Arabic (ar), Dutch (nl), French (fr), German (de), Hindi (hi), Indonesian (id), Italian (it), Marathi (mr), Vietnamese (vi) and Urdu (ur). For each experiment, we use 800 sentences for each source language for training, 100 sentences for validation and test each from the target language for evaluation. BERT-base-multilingual-cased (mBERT) has been used as an encoder in sequence Mixup and for obtaining the embeddings to evaluate cosine similarity.

Training Setup: The learning rate is 5e-5 with Adam optimizer and batch size 16. All hyperparameters are selected based on validation F1-score.

3.1 Single Language Transfer

We train the model on a single source language and a different target language dataset to evaluate its performance, as shown in Table 1. As sequence Mixup trains on interpolated sequences, it regularizes the model and prevents overfitting, outperforming mBERT. For the target language Italian, we observe higher F1 when the source language is French and lower scores for source languages Hindi and Arabic. This is in line with the trend observed in Fig 1, where the cosine similarity for the language pair (French, Italian) is highest, lower for (Hindi, Italian) and lowest for (Arabic, Italian). We observe large improvements when sequence Mixup is applied over dissimilar languages, validating that Mixup is able to generate more diverse input samples which intersect with the target language structure and semantics.

Source	Target	mBERT	Seq. Mixup
	High Similarity		
French	Italian	94.52	94.75
Indonesian	Vietnamese	56.08	56.34
German	Dutch	85.32	85.48
Hindi	Marathi	64.41	64.97
Urdu	Arabic	44.61	47.38
	Low Similarity		
Hindi	Italian	58.63	63.71
French	Arabic	39.63	40.55
Arabic	Italian	25.41	28.42

Table 1: F1-scores for POS tagging on Seq. Mixup and mBERT (mean of 10 runs). Improvements are shown with blue (↑) over mBERT.

3.2 Multi Language Transfer

To extend our experiments, we choose a pair of languages on which the model is trained and present the results in Table 2. This helps to infer in what manner additional language data impacts the performance. Languages Dutch, German and French have high cosine similarity, leading to larger improvement for the Dutch language compared to single language transfer. For target language Italian, F1-score decreases when trained on both Arabic and French data; this can be reasoned by the low cosine similarity of Arabic and Italian language.

Source	Target	Single Source	Dual Source
German+French	Italian	94.75	94.85
French+German	Dutch	85.48	86.11
Arabic+French	Italian	94.75	28.00
Arabic+Hindi	Marathi	64.97	30.07

Table 2: F1-scores for POS tagging on Seq. Mixup and mBERT (mean of 10 runs) when trained on two source languages (New+Original). Improvements are shown with blue (↑) and poorer performance with red (↓).

4 Conclusion

We analyze interpolative regularization based data augmentation over tokens for zero-shot cross-lingual transfer of part-of-speech tagging across ten languages. Through extensive experiments over languages with varying syntactic and semantic structures on single and pair of languages, we pave the way for using interpolative data augmentation to improve the generalizability of neural networks for zero-shot transfer learning on downstream tasks.

[1] https://universaldependencies.org/

References

Jiaao Chen, Zhenghui Wang, Ran Tian, Zichao Yang, and Diyi Yang. 2020. Local additivity based data augmentation for semi-supervised NER. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1241–1251, Online. Association for Computational Linguistics.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. 2018. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*.