Overlap-based Vocabulary Generation Improves Cross-lingual Transfer Among Related Languages

Vaidehi Patil [*1]   **Partha Talukdar**[†2]   **Sunita Sarawagi**[‡1]

[1]Indian Institute of Technology Bombay, India
[2]Google Research, India

Abstract

Pre-trained multilingual language models such as mBERT and XLM-R have demonstrated great potential for zero-shot cross-lingual transfer to low web-resource languages (LRL). However, due to limited model capacity, the large difference in the sizes of available monolingual corpora between high web-resource languages (HRL) and LRLs does not provide enough scope of co-embedding the LRL with the HRL, thereby affecting the downstream task performance of LRLs. In this paper, we argue that *relatedness* among languages in a language family along the dimension of lexical overlap may be leveraged to overcome some of the corpora limitations of LRLs. We propose **Overlap BPE** (**OBPE**), a simple yet effective modification to the BPE vocabulary generation algorithm which enhances overlap across related languages. Through extensive experiments on multiple NLP tasks and datasets, we observe that OBPE generates a vocabulary that increases the representation of LRLs via tokens shared with HRLs. This results in improved zero-shot transfer from related HRLs to LRLs without reducing HRL representation and accuracy. Unlike previous studies that dismissed the importance of token-overlap, we show that in the low-resource related language setting, token overlap matters. Synthetically reducing the overlap to zero can cause as much as a four-fold drop in zero-shot transfer accuracy.

1 Introduction

Zero-shot cross-lingual transfer is the ability of a model to learn from labeled data in one language and transfer the learning to another language without any labeled data. Transformer (Vaswani et al., 2017) based multilingual models pre-trained on unlabeled data from multiple languages are the state-of-the-art means for cross-lingual transfer (Ruder et al., 2019; Devlin et al., 2019a). While pre-training based cross-lingual transfer holds great promise for low web-resource languages (LRLs), such techniques are found to be more effective for transfer within high web-resource languages (HRLs) (Wu and Dredze, 2020).

Vocabulary generation is an important step in multilingual model training, where vocabulary size directly impacts model capacity. Usually, the vocabulary is generated from a union of HRL and LRL data. This often results in under-allocation of vocabulary bandwidth to LRLs, as LRL data is significantly smaller in size compared to HRL. This under-allocation of model capacity results in lower LRL performance (Wu and Dredze, 2020), as mentioned previously. In response, prior research has explored development of region-specific models (Antoun et al.; Khanuja et al., 2021), generating vocabulary specific to language clusters (Chung et al., 2020), and exploring relatedness among languages to build better LMs for LRLs (Khemchandani et al., 2021). However, none of these methods have utilized relatedness among languages for better vocabulary generation during multilingual pre-training.

In this paper, we hypothesize that exploiting language relatedness can result in an overall more effective vocabulary, which is also better representative of LRLs. Closely related languages (e.g., languages belonging to a single family) have common origins for words with similar meanings. We show some examples across three different families of related languages in Table 10. Morphological inflections of the root word lead to lexically overlapping tokens across languages. Learning representations for such subwords in lexically overlapping words shared across HRL and its related LRLs can enable better transfer of supervision from HRL to LRLs. During Masked Language Modelling (MLM) pre-training (Devlin et al., 2019a), the shared tokens can serve as anchors in learning contextual representations of neighboring tokens. However, choos-

*vaidehipatil16@gmail.com
†partha@google.com
‡sunita@iitb.ac.in

Language and Token frequencies	English: University (10), versity (6); German: Universitaten (2); Dutch: Universiteit (1); Western Frisian: Universiteiten (1)
Starting Vocab	Uni, versit, U,n,i,v,e,r,s,i,t,y,a
BPE Vocab	versity, Uni, versit, U,n,i,v,e,r,s,i,t,y,a
OBPE Vocab	Universit, Uni, versit, U,n,i,v,e,r,s,i,t,y,a

Table 1: First row shows lexically overlapping tokens in four different languages with their corpus frequencies (in brackets), with English (En) as the High Web-Resource Language (HRL). From a starting vocabulary shown in the second row, BPE merges tokens based on greater overall frequency, adding new vocabulary item *versity* as it has the highest overall frequency (16). OBPE instead adds *Universit* since it also rewards cross-lingual overlap, even though *Universit* has lower overall frequency (15).

ing the correct granularity of sharing automatically is tricky. On one extreme, we can choose a vocabulary which favours longer units frequent in HRL without regard for sharing, thereby leading to better semantic representation of the tokens but no cross-lingual transfer. On the other extreme, we can choose character-level vocabulary (Ma et al., 2020), where every token is shared across languages but have no semantic significance.

Given text from a mix of high and low Web-resource languages (HRL and LRL, respectively), Byte Pair Encoding (BPE) (Sennrich et al., 2016) and its variants like Wordpiece (Schuster and Nakajima, 2012) and Sentencepiece (Kudo and Richardson, 2018) prefer frequent tokens, most of those from the HRLs. This would cause most long HRL tokens to get included, leaving only a limited budget of short tokens for the LRL. Any sub-token level overlap between HRL and LRL could get lost in this process. In a zero-shot setting, since available supervision is HRL based, this creates a bottleneck when transferring supervision from HRL to LRLs. Oversampling LRLs is a common strategy to offset this imbalance but that hurts HRL performance as shown in (Conneau et al., 2020a).

In this paper, we propose Overlap BPE (OBPE). OBPE chooses a vocabulary by giving token overlap among HRL and LRLs a primary consideration. OBPE prefers vocabulary units which are shared across multiple languages, while also encoding the input corpora compactly. Thus, OBPE tries to balance the trade-off between cross-lingual subword sharing and the need for robust representation of individual languages in the vocabulary. This re-

sults in a more balanced vocabulary, resulting in improved performance for LRLs without hurting HRL accuracy. Table 1 shows an example to highlight this difference between OBPE and BPE.

Recently K et al. (2020); Conneau et al. (2020b) concluded that token overlap is unimportant for cross-lingual transfer. However, they studied language pairs where either both languages had a large corpus, or where the languages were not sufficiently related. We focus on related languages within a family and observe drastic drop in zero-shot accuracy when we synthetically reduce the overlap to zero (58% F1 drops to 17% for NER, 71% drops to 30% for text classification).

This paper offers the following contributions

- We present OBPE, a simple yet effective modification to the popular BPE algorithm to promote overlap between LRLs and a related HRL during vocabulary generation. OBPE uses a generalized mean based formulation to quantify token overlap among languages.

- We evaluate OBPE on twelve languages across three related families, and show consistent improvement in zero-shot transfer over state-of-the art baselines on four NLP tasks. We analyse the reasons behind the gains obtained by OBPE and show that OBPE increases the percentage of LRL tokens in the vocabulary without reducing HRL tokens. This is unlike over-sampling strategies where increasing one reduces the other.

- Through controlled experiments on the amount of token overlap on a related HRL-LRL pair, we show that token overlap is extremely important in the low-resource, related language setting. Recent literature which conclude that token overlap is unimportant may have overlooked this important setting.

The source code for our experiments is available at https://github.com/Vaidehi99/OBPE.

2 Related Work

Transformer-based multilingual language models such as mBERT (Devlin et al., 2019b) and XLM-R (Conneau et al., 2020a) are now established as the de-facto method for zero-shot cross-lingual transferability, and thus hold promise for low resource domains. However, recent studies have indicated that even the current state-of-the-art models such as XLM-R (Large) do not yield reasonable transfer performance across low resource target languages

with limited data (Wu and Dredze, 2020). This has led to a surge of interest in enhancing cross-lingual transfer of multilingual models to the low-resource setting. We categorize existing work based on the stage of the pre-training pipeline where it is relevant:

Input Data In the data creation stage, Conneau et al. (2020a) propose over-sampling of LRL documents to improve LRL representation in the vocabulary and pre-training steps. Khemchandani et al. (2021) specifically target related languages and propose transliteration of LRL documents to the script of related HRL for greater lexical overlap. We deploy both these tricks in this paper.

Tokenization Rust et al. (2021) study that even the tokenization step could have a crucial impact on performance accrued to each language in a multilingual models. They propose the use of dedicated tokenizer for each language instead of the automatically generated multilingual mBERT tokenizer. However, they continue to use the default mBERT vocabulary generator.

Vocabulary Generation Sennrich et al. (2016) highlighted the importance of subword tokens in the vocabulary and proposed use of the BPE algorithm (Gage, 1994) for efficiently growing such a vocabulary incrementally. Variants like Wordpiece (Schuster and Nakajima, 2012) and Sentencepiece (Kudo and Richardson, 2018) either build on top of BPE or follow a very similar process. Kudo (2018) is a variant method that chooses tokens based on unigram LM score. We obtained better results with BPE and continued with that. All these BPE variants incrementally add subwords based on overall frequency in the combined corpus, and they all ignore language boundaries. Chung et al. (2020) observed that such a combined approach could under-represent several languages, and proposed instead to separately create vocabularies for clusters of related languages and take a union of each cluster-specific vocabulary. However, within each cluster they continue to use the default vocabulary generator. Our approach can be used as a drop-in replacement to further enhance the quality of the cluster-specific vocabulary that they obtain. Wang et al. (2019); Gao et al. (2020) propose a soft-decoupled encoding approach for exploiting subword overlap between LRLs and HRLs. However, their focus is NMT models and does not easily integrate in existing multilingual models such as mBERT. (Maronikolakis et al., 2021) targets tok-

enization compatibility based purely on vocabulary size and does not focus on choosing the tokens that go in the vocabulary.

Pre-Training and Adaptation Several previous works have proposed to include additional alignment loss between parallel (Cao et al., 2020) or pseudo-parallel (Khemchandani et al., 2021) sentences to co-embed HRLs and LRLs. Another approach is to design language-specific Adapter layers (Pfeiffer et al., 2020a,b; Artetxe et al., 2020; Üstün et al., 2020) that can be easily fine-tuned for each new language. Pfeiffer et al. (2021) leverages the pre-trained embeddings of lexically overlapping tokens between the vocabulary of pre-trained model and that of unseen target language to initialize the corresponding embeddings of target language. However, they did not attempt to increase the fraction of such tokens in the vocabulary.

We are not aware of any prior work that explicitly promotes overlapping tokens between LRLs and HRLs in the vocabulary of multilingual models.

3 Overlap-based Vocabulary Generation

We are given monolingual data $D_1, ..., D_n$ in a set of n languages $\mathcal{L} = \{L_1, ..., L_n\}$ and a vocabulary budget V. Our goal is to generate a vocabulary $\mathcal{V}$ that when used to tokenize each D_i in a multilingual model would provide cross-lingual transfer to LRLs from related HRLs. We use $\mathcal{L}_{LRL}$ to denote the subset of the n languages that are low-resource, the remaining languages $\mathcal{L} - \mathcal{L}_{LRL}$ are denoted as the set $\mathcal{L}_{HRL}$ of high resource languages.

Existing methods of vocabulary creation start with a union D of monolingual data $D_1, ..., D_n$, and choose a vocabulary $\mathcal{V}$ that most compactly represents D. We first present an overview of BPE, a popular algorithm for vocabulary generation.

3.1 Background: BPE

Byte Pair Encoding (BPE) (Gage, 1994) is a simple data compression technique that chooses a vocabulary $\mathcal{V}$ that minimizes total size of $D = \cup_i D_i$ when encoded using $\mathcal{V}$.

$$\mathcal{V} = \operatorname*{argmin}_{S:|S|=V} \sum_{i=1}^{n} |\text{encode}(D_i, S)| \qquad (1)$$

The size of the encoding $|\text{encode}(D_i, S)|$ can be alternately expressed as the sum of frequency of tokens in S when D_i is tokenized using S. This motivates the following efficient greedy algorithm to implement the above optimization (Sennrich et al.,

Algorithm 1 Overlap based BPE (OBPE)

---

for $i \in \{1, 2, ..., n\}$ **do**
    Split words in D_i into characters C_i with a special marker after every word
end for
$\mathcal{V} = \cup_{i=1}^{n} C_i$
while $|\mathcal{V}| < \text{V}$ **do**
    Update token and pair frequency on $\{D_i\}, \mathcal{V}$
    Add to $\mathcal{V}$ token k formed by merging pairs $u, v \in \mathcal{V}$ with the largest value of

$$(1-\alpha) \sum_j f_{kj} + \alpha \sum_{i \in \mathcal{L}_{\text{LRL}}} \max_{h \in \mathcal{L}_{\text{HRL}}} \left(\frac{f_{ki}^p + f_{kh}^p}{2} \right)^{\frac{1}{p}}$$

end while

2016). Let f_{ki} denote the frequency of a candidate token k in the corpus D_i of language L_i. The BPE algorithm grows $\mathcal{V}$ incrementally. Initially, $\mathcal{V}$ comprises of characters in D. Then, until $|\mathcal{V}| \leq \text{V}$, it chooses the token k obtained by merging two existing tokens in $\mathcal{V}$ for which the frequency in D is maximum.

$$\mathcal{V} = \mathcal{V} \cup \operatorname*{argmax}_{k=[u,v]:u,v \in \mathcal{V}} \sum_i f_{ki} \qquad (2)$$

A limitation of BPE on multilingual data is that tokens that appear largely in low-resource D_i may not get added to $\mathcal{V}$, leading to sentences in L_i being over-tokenized. For a low resource language, the available monolingual data D_i is often orders of magnitude smaller than another high-resource language. Models like mBERT and XLM-R address this limitation by over-sampling documents of low-resource languages. However, over-sampling LRLs might compromise learned representation of HRLs where task-specific labeled data is available. We propose an alternative strategy of vocabulary generation called OBPE that seeks to maximize transfer from HRL to LRL.

3.2 Our Proposal: OBPE

The key idea in OBPE is to maximize the overlap between an LRL and a closely related HRL while simultaneously encoding the input corpora compactly as in BPE. When labeled data D_h^T for a task T is available in an HRL L_h, then a multilingual model fine-tuned with D_h^T is likely to transfer better to a related LRL L_i when L_i and L_h share several tokens in common. Thus, the objective that OBPE

seeks to optimize when creating a vocabulary is:

$$\mathcal{V} = \operatorname*{argmin}_{S:|S|=\text{V}} \left[(1-\alpha) \sum_{i=1}^{n} |\text{encode}(D_i, S)| \right.$$
$$\left. - \alpha \sum_{i \in \mathcal{L}_{\text{LRL}}} \max_{j \in \mathcal{L}_{\text{HRL}}} \text{overlap}(L_i, L_j, S) \right] \qquad (3)$$

where $0 \leq \alpha \leq 1$ determines importance of the two terms. The first term in the objective compactly represents the total corpus, as in BPE's (Eq (1)). The second term additionally biases towards vocabulary with greater overlap of each LRL to one HRL where we expect task-specific labeled data to be present. There are several ways in which we can measure the overlap between two languages with respect to a current vocabulary. First, we encode each of D_i and D_j using the vocabulary S, which then yields a multiset of tokens in each corpus. Inspired by the literature on fair allocation (Barman et al., 2021), we explore a continuously parameterized function that expresses overlap between two languages' encoding as a generalized mean function as follows:

$$\text{overlap}(L_i, L_h, S) = \sum_{k \in S} \left(\frac{f_{ki}^p + f_{kh}^p}{2} \right)^{\frac{1}{p}}, \;\; p \leq 1 \qquad (4)$$

where f_{ki} denotes the frequency of token k when D_i is encoded with S. For different values of p, we get different tradeoffs between fairness to each language and overall goodness. When $p = -\infty$, generalized mean reduces to the minimum function, and we get the most egalitarian allocation. However, this ignores the larger of the two frequencies. When $p = 1$, we get a simple average which is what the first term in Equation (3) already covers. For $p = 0, -1$, we get the geometric and harmonic means respectively. Due to smaller size of LRL monolingual data, the frequency of a token which is shared across languages is likely to be much higher in HRL monolingual data as compared to that in LRL monolingual data, Hence, setting p to large negative values will increase the weight given to LRLs and thus increase overlap. We will present an exploration of the effect of p on zero-shot transfer in the experiment section.

The greedy version of the above objective that controls the candidate vocabulary item to be in-

Family	HRL	LRLs	Number of HRL Docs	
			BALANCED	SKEWED
West Germanic	English (en)	German (de), Dutch (nl), Western Frisian (fy)	0.16M	1.00M
Romance	French (fr)	Spanish (es), Portuguese (pt), Italian (it)	0.16M	0.50M
Indo-Aryan	Hindi (hi)	Marathi (mr), Punjabi (pa), Gujarati (gu)	0.16M	0.16M

Table 2: Twelve Languages *simulated* as HRLs and LRLs across with two different corpus distribution: BALANCED and SKEWED. Number of documents in languages simulated as LRLs is 20K.

ducted in each iteration of OBPE is thus:

$$
\mathcal{V} = \mathcal{V} \cup \operatorname*{argmax}_{k=[u,v]:u,v\in\mathcal{V}} (1-\alpha) \sum_j f_{kj}
$$
$$
+ \alpha \sum_{i\in\mathcal{L}_{\text{LRL}}} \max_{h\in\mathcal{L}_{\text{HRL}}} \left(\frac{f_{ki}^p + f_{kh}^p}{2} \right)^{\frac{1}{p}} \quad (5)
$$

The data structure maintained by BPE to efficiently conduct such merges can be applied with little changes to the OBPE algorithm. The only difference is that we need to separately maintain the frequency in each language in addition to overall frequency. Since the time and resources used to create the vocabulary is significantly smaller than the model pre-training time, this additional overhead to the pre-training step is negligible.

4 Experiments

We evaluate by measuring the efficacy of zero-shot transfer from the HRL on four different tasks: named entity recognition (NER), part of speech tagging (POS), text classification(TC), and Cross-lingual Natural Language Inference (XNLI). Through our experiments, we evaluate the following questions:

1. Is OBPE more effective than BPE for zero-shot transfer? (Section 4.2)
2. What is the effect of token overlap on overall accuracy? (Section 4.3)
3. How does increased LRL representation in the vocabulary impact accuracy? (Section 4.4)

We report additional ablation and analysis experiments in Section 4.5.

4.1 Setup

Pre-training Data and Languages As our pre-training dataset $\{D_i\}$, we use the Wikipedia dumps of all the languages as used in mBERT. We pre-train with 12 languages grouped into three families of four related languages as shown in Table 2. In each family, we simulate as HRL the most populous language, and call the remaining as LRLs. The number of documents for languages simulated as

	Lang	Number of sentences			
Dataset split		NER	POS	TC	XNLI
Train:HRL	hi	5.0	53.0	25.0	
	en	10.5	18.0	10.0	393.0
	fr	7.5	16.5	10.0	393.0
Validation:HRL	hi	1.0	3.0	4.0	
	en	6.0	4.0	10.0	2.5
	fr	4.0	2.0	10.0	2.5
Test data	hi	0.2	12.0	7.0	
	en	6.0	4.6	10.0	5.0
	fr	4.0	4.1	10	5.0
	mr	0.8	9.5	6.5	-
	pa	0.2	13.4	7.9	-
	gu	0.3	14.0	8.0	-
	de	12.0	19.3	10.0	5.0
	nl	8.0	1.0	-	-
	fy	0.8	-	-	-
	es	5.0	3.1	10.0	5.0
	pt	4.0	2.5	-	-
	it	5.0	3.4	-	-

Table 3: Task-specific data sizes. Number of sentences in thousands.

LRLs is set to 20K. For the HRLs, we consider two corpus distributions:

- BALANCED : all three HRLs get 160K documents each
- SKEWED : English gets one million, French half million, and Hindi 160K documents

We evaluate twelve-language models in each of these settings, and present results for separate four language models per family in Table 12 in the Appendix. For the Indo-Aryan languages set, the monolingual data of Punjabi and Gujarati is transliterated to Devanagari, the script of Hindi and Marathi. We use libindic's indictrans library (Bhat et al., 2015) for transliteration. Languages in the other two sets do not require transliteration as they have a common script. Thus, all four languages in each set are in the same script so their lexical overlap can be leveraged.

Pre-Training Details To ensure that LRLs are not under-represented, we over-sample using exponentially smoothed weighting similar to multilingual BERT (Devlin et al., 2019b) with exponentiation

factor 0.7. We perform MLM pretraining on a BERT base model with 110M parameters from scratch. We generate a vocabulary of size of 30k. We chose batch size as 2048, learning rate as 3e-5 and maximum sequence length as 128. Pre-training of BERT was done with duplication factor 5 for for 64k iterations for HRLs. For all LRLs, duplication factor was 20 and training was done for 24K iterations. MLM pre-training was done on Google v3-8 Cloud TPUs where 10K iterations required 2.1 TPU hours.

Task-specific Data We evaluate on four downstream tasks: (1) NER: data from WikiANN (Pan et al., 2017) and XTREME (Hu et al., 2020), (2) XNLI: data from (Conneau et al., 2018), (3) POS: data from XTREME (Hu et al., 2020) and TDIL[1], and (4) Text Classification (TC): data from TDIL and XGLUE (Liang et al., 2020). We downsampled the TDIL data for each language to make them class-balanced. The POS tagset for Indo-Aryan languages used was the BIS Tagset (Sardesai et al., 2012). Table 3 presents a summary. The test set to compute LRL perplexity was formed by sampling 10K sentences from Samanantar corpus(Ramesh et al., 2021) for Indic languages and from Tatoeba corpus[2] for other languages. The perplexity reported for a language is the average of sentence perplexity over all the sentences sampled from that language's corpus.

Task-specific fine-tuning details We perform task-specific fine-tuning of pre-trained BERT on the task-specific training data of HRL and evaluate on all languages in the same family. Here we used learning-rate 2e-5 and batch size 32, with training duration as 16 epochs for NER, 8 epochs for POS and 3200 iterations for Text Classification and XNLI. The models were evaluated on a separate validation dataset of the HRL and the model with the minimum validation loss, maximum F1-score, accuracy and minimum validation loss was selected for final evaluation for XNLI, NER, POS and Text Classification respectively. All fine-tuning experiments were performed on Google Colaboratory. The results reported for all the experiments are an average of 3 independent runs.

4.2 Effectiveness of OBPE

We evaluate the impact of OBPE on improving zero-shot transfer from HRLs to LRLs within the

[1]Technology Development for Indian Languages (TDIL), https://www.tdil-dc.in
[2]Tatoeba , https://tatoeba.org

same family across four different tasks. We compare with four existing methods that represent different methods of vocabulary creation and allocation of budget across languages:

Methods compared

1. **BPE** (Sennrich et al., 2016), the existing default method of vocabulary generation.

2. **Clustered vocabulary (CV)** (Chung et al., 2020) Since the paper uses a SentencePiece unigram for vocabulary, we followed the same approach for this comparison. We allocate each family equal number of vocabulary tokens which is V/3.

3. **BPE-dropout (BPE-dp)** (Provilkov et al., 2020) uses the vocabulary generated by BPE but tokenizes the text using a dropout rate of 0.1. This allows the training of tokens that are subsumed by larger tokens in the vocabulary.

4. **Compatibility of Tokenizations (Tok-Comp)** (Maronikolakis et al., 2021) uses a method to select meaningful vocabulary sizes in an automated manner for all language using compression rates. Since their best performances are found, when the compression rates are similar, we choose a size for each language corresponding to compression rate of 0.5. The tokenizer used in this method is WordPiece. .

5. **OBPE** (Ours) with default $\alpha = 0.5, p = -\infty$. We also do ablation on these.

In Table 4 we observe that across all four tasks, zero-shot LRL accuracy improves compared to BPE. For example, the average accuracy on XNLI for the LRL languages improves from 55.6 to 58.1 just by changing the set of tokens in the vocabulary. These gains are obtained without compromising HRL performance on the tasks. The Clustered Vocabulary (CV) approach is much worse than BPE. These experiments are on the Balanced-12 model. In the supplementary section, we report the results on the Skewed-12 (Table 5) and Balanced-4 models (Table 12) and show similar gains even with these models. In this table, we averaged the gains over nine LRLs, and in the Supplementary Table 14 we show consistent gains for individual languages.

In addition to improving zero-shot transfer from HRLs to LRLs on downstream tasks, OBPE also leads to better intrinsic representation of LRLs. We validate that by measuring the pseudo-perplexity (Salazar et al., 2020) of a test set of LRL sentences. We find that average perplexity of LRL

Method	LRL Performance (↑)				HRL Performance (↑)			
	NER	TC	XNLI	POS	NER	TC	XNLI	POS
BPE (Sennrich et al., 2016)	64.48	65.52	52.07	84.64	83.26	**82.07**	62.71	**95.20**
BPE-dp (Provilkov et al., 2020)	63.92	64.15	52.66	84.75	81.73	81.07	63.74	94.61
CV (Chung et al., 2020)	59.58	61.91	49.30	81.68	81.15	80.93	64.51	94.47
TokComp (Maronikolakis et al., 2021)	63.79	65.77	53.94	**85.49**	82.43	80.93	66.10	94.86
OBPE (This paper)	**65.72**	**68.02**	**54.03**	85.26	**83.98**	81.91	**66.27**	95.09

Table 4: Zero-shot performance of models in the Balanced-12 setting trained on 9 LRL and 3 HRL languages. Performance is measured on four tasks: NER (F1), Text Classification (Accuracy), POS (Accuracy), and XNLI (Accuracy). For all metrics, higher is better (↑). Zero-shot transfer to LRL improves without hurting HRL accuracy. P-value of paired-t-test between BPE and OBPE LRL gains has values 0.01, 0.04, 0.02, 0.01 for each of the 4 tasks establishing statistical significance. Detailed results for each language is pesented in Table 14. Section 4.2 has further discussion.

Method	LRL Performance (↑)				HRL Performance (↑)			
	NER	TC	XNLI	POS	NER	TC	XNLI	POS
BPE (Sennrich et al., 2016)	52.91	51.68	48.57	74.79	81.78	80.04	64.96	95.03
CV (Chung et al., 2020)	52.73	54.40	44.28	76.70	79.84	77.74	57.18	94.60
OBPE (This paper)	**55.09**	**55.37**	**50.01**	75.05	**82.94**	**80.31**	**65.57**	**95.09**

Table 5: Zero-shot performance of models in the Skewed-12 setting of Table 2 on same four tasks as Table 4. OBPE shows gains here too. Detailed numbers in Table 11 of Supplementary. Section 4.2 has further discussion.
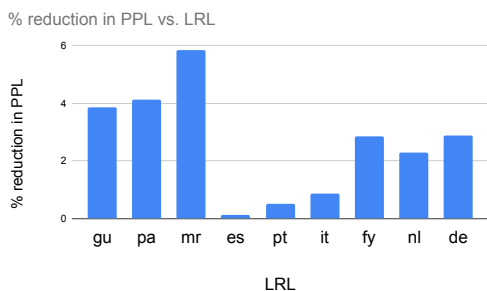

Figure 1: Percentage reduction in Pseudo perplexity (Salazar et al., 2020) for different LRLs as we go from BPE to OBPE vocabulary. (Section 4.2)

	en-es	
	High (es: 1 GB)	Low: (es: 20K)
NER	-1.4	-11.7
XNLI	0.7	-1.3
	hi-mr	
	High (mr: 110K)	Low (mr: 20K)
NER	-12.2	-41.6
TC	-2.7	-41.3
POS	-6.6	-7.8

Table 6: Drop in Accuracy of Zero-shot transfer when we synthetically reduce token overlap to zero. Transfer is from English (en) as HRL to Spanish (es) and from Hindi (hi) as HRL to Marathi (mr) in two settings: (1) High where es, mr have sizes comparable to the HRL and (2) Low where their sizes are only 20K. Token overlap is important in the low-resource and related language setting (Section 4.3)

sentences drops by 2.6% when we go from the BPE to OBPE vocabulary. More details on this experiment appear in Figure 1.

In order to investigate the reasons behind the OBPE gains, we first inspected the percentage of tokens in the vocabulary that belong to LRLs, HRLs, and in their overlap. We find that with OBPE both LRL tokens and overlapping tokens increase. Either of these could have led to the observed gains. We analyze the effect of each of these factors in the following two sections.

4.3 Effect of Token Overlap

We present the impact of token overlap via two sets of experiments: first, a controlled setup where we synthetically vary the fraction of overlap and second where we measure correlation between overlap and gains of OBPE on the data as-is.

For the controlled setup we follow (K et al., 2020) for synthetically controlling the amount of overlap between HRL and LRL. We trained a bilingual model between Hindi (HRL 160K) and Marathi (LRL 20K) — two closely related languages in the Indo-Aryan family. To find the set of overlapping tokens between Hindi and Marathi, we first run OBPE on Hindi-Marathi language pair to generate a vocabulary and label all tokens present
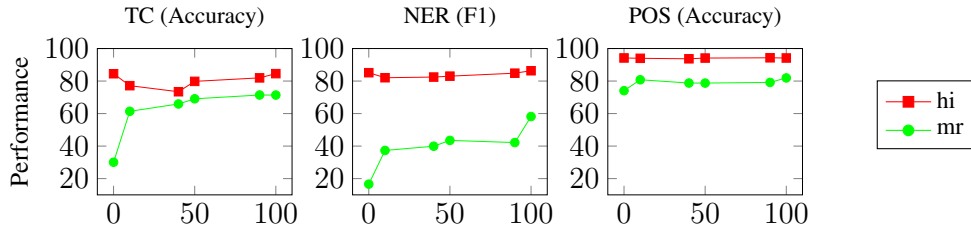
Figure 2: Zero-shot performance vs Overlap of models trained on unicode shifted HRL data to simulate increasing overlap between HRL (SynthHindi) and LRL (mr). Performance is measured on three tasks: Text Classification (Accuracy), NER (F1) and POS (Accuracy). On TC and NER observe the huge drop in LRL accuracy as we decrease overlap from 100 down to 0. Further discussions in Section 4.3.

in both languages as *overlapping tokens*. We then incrementally sample 10%, 40%, 50%, 90% of the tokens from this set. We shift the Unicode of the entire Hindi monolingual data except the set of sampled tokens so that there are no overlapping tokens between Hindi (hi) and Marathi (mr) monolingual data other than the sampled tokens. Let us call this Hindi data **SynthHindi**. We then run OBPE on SynthHindi-Marathi language pair to generate a vocabulary to pretrain the model. The task-specific Hindi data is also converted to SynthHindi during fine-tuning and testing of the model.

Figure 2 shows results with increasing overlap. We observe increasing gains in LRL accuracy as we go from no overlap to full overlap on all three tasks. NER accuracy increases from 17% to 58% for the LRL (mr) even while the HRL (hi) accuracy stays unchanged. For TC we observe similar gains. For POS, even without token overlap, we get good cross-lingual transfer because POS tags are more driven by structural similarity, and Hindi and Marathi follow similar structure.

Our results contradict the conclusions of (K et al., 2020) which claimed that token overlap is unimportant for cross-lingual transfer. However, there are two key differences with our setting: (1) unlike (K et al., 2020), we explore low-resource settings, and (2) except for English-Spanish, the other language pairs they considered are not linguistically related. To explain the importance of both these factors, in Table 6 we present accuracy of English-Spanish in a simulated low-resource setting where we sample 20K Spanish documents and 160K English documents. Also, we repeat our Hindi-Marathi experiments where Marathi is not low-resource. We observe that (1) Spanish as LRL benefits significantly on overlap with English. (2) Marathi gains from token overlap with Hindi even in the high resource setting.

Thus, we conclude that as long as languages are related, token overlap is important and the benefit from overlap is higher in the low resource setting.

Overlap Vs Gain: Real data setup We further substantiate our hypothesis that the shared tokens across languages favoured by OBPE enable transfer of supervision from HRL to LRL via statistics on real-data. In Table 9 we show the Pearson product-moment correlation coefficient between overlap gain and performance gain within LRLs of the same family and task. We get a high positive correlation coefficient, with an average of 0.644.

4.4 Effect of Increased LRL representation

We next investigate the impact of increased representation of LRL tokens in the vocabulary. OBPE increases LRL representation by favoring overlapping tokens, but LRL tokens can also be increased by just over-sampling LRL documents. We train another BALANCED12 model but with further over-sampling LRLs with exponentiation factor of 0.5 instead of 0.7. We observe in Figure 8 that this increases LRL fraction but reduces HRL tokens in the vocabulary. Table 7 also shows the comparison of zero-shot transfer accuracy with over-sampled BPE against over-sampled OBPE. We find that OBPE even with default exponentiation factor achieves highest LRL gains, whereas aggressively over-sampled BPE hurts HRL accuracy. Within the same sampling setting, OBPE is better than corresponding BPE.

4.5 Ablation study

We conducted experiments for different values of p that controls the amount of overlap in the generalized mean function (Equation (5)). Figure 3 and Table 14 show the results for various p. Setting $p = 1$ gives the original BPE algorithm. Setting $p = 0, -1$ gives geometric and harmonic mean respectively, setting $p = -\infty$ gives minimum. We

Method	LRL Performance (↑)				HRL Performance (↑)			
	NER	TC	XNLI	POS	NER	TC	XNLI	POS
BPE	64.5	65.5	52.1	84.6	83.3	**82.1**	62.7	**95.2**
+overSample	64.4	67.6	52.1	84.6	82.4	82.0	62.0	95.2
OBPE	**65.7**	**68.0**	**54.0**	**85.3**	**84.0**	81.9	**66.3**	95.1
+overSample	64.6	67.9	53.5	85.1	82.7	81.7	65.7	94.8

Table 7: Zero-shot performance of models in the same setting as Table 4 but comparing default sampling with oversampling (exponentiation factor S=0.5). Note, even if BPE_overSamp improves LRL somewhat, it causes HRL to drop. OBPE with default sampling is best for both LRLs and HRLs. Also OBPE_overSampled is better than BPE_overSampled (Section 4.4).

Lang family	Task	Pearson Correlation
Indo-Aryan	NER	0.835
	POS	0.690
West Germanic	NER	0.387
	POS	0.348
Romance	NER	0.946
	POS	0.595

Table 9: Correlation coefficient between performance gain and overlap gain within languages in a family for various tasks. (Section 4.3).

compare the task-specific results for different values of p as shown in Table 14 and find that the gains we obtain are highest in the $p = -\infty$ (minimum) setting (Figure 3).
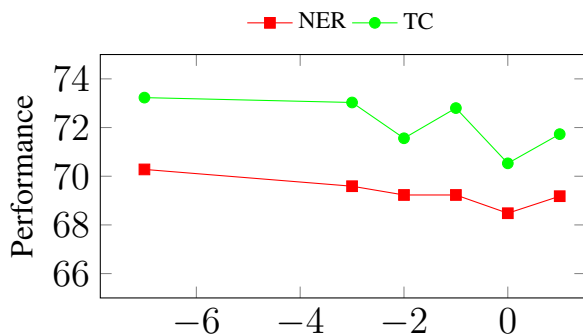

Figure 3: Zero-shot LRL performance of models in the same setting as Table 4 for different values of p evaluated on NER and Text Classification. Best results at $p = -\infty$.(Section 4.5)

We also experiment with $\alpha = 0.7$, and find that for most languages the results were not better than our default $\alpha = 0.5$.

5 Conclusion

In this paper, we address the problem of cross-lingual transfer from HRLs to LRLs by exploiting
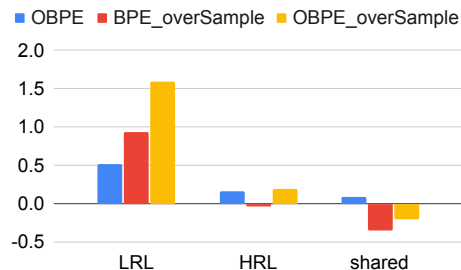

Table 8: Percentage rise over BPE in representation of LRL, HRL and Shared (percentage of tokens shared between HRL and LRL weighted by frequency) in vocabulary generated by OBPE and BPE_overSample and OBPE_overSample (Section 4.4).

relatedness among them. We focus on lexical overlap during the vocabulary generation stage of multilingual pre-training. We propose Overlap BPE (OBPE), a simple yet effective modification to the BPE algorithm, which chooses a vocabulary that maximizes overlap across languages. OBPE encodes input corpora compactly while also balancing the trade-off between cross-lingual subword sharing and language-specific vocabularies. We focus on three sets of closely related languages from diverse language families. Our experiments provide evidence that OBPE is effective in leveraging overlap across related languages to improve LRL performance. In contrast to prior work, through controlled experiments on the amount of token overlap between two related HRL-LRL language pairs, we establish that token overlap is important when a LRL is paired with a related HRL.

References

Wissam Antoun, Fady Baly, and Hazem Hajj. Arabert: Transformer-based model for arabic language understanding. In *LREC 2020 Workshop Language*

Resources and Evaluation Conference 11–16 May 2020, page 9.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Siddharth Barman, Arindam Khan, and Arnab Maiti. 2021. Universal and tight online algorithms for generalized-mean welfare.

Irshad Ahmad Bhat, Vandan Mujadia, Aniruddha Tammewar, Riyaz Ahmad Bhat, and Manish Shrivastava. 2015. Iiit-h system submission for fire2014 shared task on transliterated search. In *Proceedings of the Forum for Information Retrieval Evaluation*, FIRE '14, pages 48–53, New York, NY, USA. ACM.

Steven Cao, Nikita Kitaev, and Dan Klein. 2020. Multilingual alignment of contextual word representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Hyung Won Chung, Dan Garrette, Kiat Chuan Tan, and Jason Riesa. 2020. Improving multilingual models with language-clustered vocabularies. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4536–4546, Online. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020a. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020b. Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019a. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL 2019*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019b. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Philip Gage. 1994. A new algorithm for data compression. *C Users Journal*, 12(2):23–38.

Luyu Gao, Xinyi Wang, and Graham Neubig. 2020. Improving target-side lexical transfer in multilingual neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.

Christoph Hube. 2017. Bias in wikipedia. In *Proceedings of the 26th International Conference on World Wide Web Companion*, WWW '17 Companion, page 717–721, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-lingual ability of multilingual bert: An empirical study. In *International Conference on Learning Representations*.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, et al. 2021. Muril: Multilingual representations for indian languages. *arXiv preprint arXiv:2103.10730*.

Yash Khemchandani, Sarvesh Mehtani, Vaidehi Patil, Abhijeet Awasthi, Partha Talukdar, and Sunita Sarawagi. 2021. Exploiting language relatedness for low web-resource language model adaptation: An Indic languages study. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1312–1323, Online. Association for Computational Linguistics.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. 2020. XGLUE: A new benchmark datasetfor cross-lingual pre-training, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6008–6018, Online. Association for Computational Linguistics.

Wentao Ma, Yiming Cui, Chenglei Si, Ting Liu, Shijin Wang, and Guoping Hu. 2020. CharBERT: Character-aware pre-trained language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 39–50, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Antonis Maronikolakis, Philipp Dufter, and Hinrich Schütze. 2021. Wine is not v i n. on the compatibility of tokenizations across languages. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2382–2399, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958, Vancouver, Canada. Association for Computational Linguistics.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020a. AdapterHub: A framework for adapting transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 46–54, Online. Association for Computational Linguistics.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020b. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online. Association for Computational Linguistics.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2021. UNKs everywhere: Adapting multilingual language models to new scripts. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10186–10203, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ivan Provilkov, Dmitrii Emelianenko, and Elena Voita. 2020. BPE-dropout: Simple and effective subword regularization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1882–1892, Online. Association for Computational Linguistics.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Kumar Deepak, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Shantadevi Khapra. 2021. Samanantar: The largest publicly available parallel corpora collection for 11 indic languages.

Sebastian Ruder, Matthew E. Peters, Swabha Swayamdipta, and Thomas Wolf. 2019. Transfer learning in natural language processing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, pages 15–18, Minneapolis, Minnesota. Association for Computational Linguistics.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How good is your tokenizer? on the monolingual performance of multilingual language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3118–3135, Online. Association for Computational Linguistics.

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712, Online. Association for Computational Linguistics.

Madhavi Sardesai, Jyoti Pawar, Shantaram Walawalikar, and Edna Vaz. 2012. BIS annotation standards with reference to Konkani language. In *Proceedings of the 3rd Workshop on South and Southeast Asian Natural Language Processing*, pages 145–152, Mumbai, India. The COLING 2012 Organizing Committee.

Mike Schuster and Kaisuke Nakajima. 2012. Japanese and korean voice search. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5149–5152. IEEE.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational*

Linguistics (Volume 1: Long Papers), pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Ahmet Üstün, Arianna Bisazza, Gosse Bouma, and Gertjan van Noord. 2020. UDapter: Language adaptation for truly Universal Dependency parsing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2302–2315, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Xinyi Wang, Hieu Pham, Philip Arthur, and Graham Neubig. 2019. Multilingual neural machine translation with soft decoupled encoding. In *International Conference on Learning Representations*.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130, Online. Association for Computational Linguistics.

A Appendix

A.1 Examples of Token Overlap within Language Families

Table 10 shows examples of overlapping tokens within three different language families, and Figure 4 shows a real example of how OBPE chooses shared tokens.

A.2 Limitations

- Our approach is expected to improve cross-lingual transfer from HRL to LRL only when the HRL and LRL are related linguistically since it relies on the presence of lexically overlapping tokens
- It requires the transliteration of LRL data to the script of its related HRL if LRL does not have the same script.

A.3 Potential risks

Language models may amplify bias in data and also introduce new ones. Multilingual models explored in the paper are not immune to such issues. Detecting such biases and mitigating them is a topic of ongoing research. We are hopeful that our focus on better representation of LRLs in the vocabulary is a step towards more inclusive models.

	Tokens
Hindi	नियुक्ति(Niyukti), नियुक्त(Niyukta)
Panjabi	नियुकत(Niyukata), नियुकतीआं(Niyukatiaa), नियुकती(Niyukati)
Marathi	नियुक्ति(Niyukti), नियुक्त(Niyukta)
Gujarati	नियुक्ति(Niyukti), नियुक्त(Niyukta)
BPE vocab	नियुक्ति(Niyukti), नियुक्त(Niyukta)
OBPE vocab	नियुक्ति(Niyukti), नियुक्त(Niyukta), **नियुक्(Niyuk)**

Figure 4: Similar meaning words with shared root forms across related Indo-Aryan languages. BPE vocabulary does not capture the tokens corresponding to Punjabi as it is a LRL and will thus tokenize Niyukata into multiple tokens which do not captures its meaning whereas Niyukata when tokenized by OBPE tokenizer will contain Niyuk which captures most of the meaning of the token Niyukata whose representation will be learnt when pretraining using Punjabi monolingual data

A.4 Replicability

BERT configuration parameters used in our experiments are as follows:

"attention_probs_dropout_prob": 0.1, "hidden_act": "gelu", "hidden_dropout_prob": 0.1, "hidden_size": 768, "initializer_range": 0.02, "intermediate_size": 3072, "max_position_embeddings": 512, "num_attention_heads": 12, "num_hidden_layers": 12, "type_vocab_size": 2, "vocab_size": 30000 All the task-specific fine-tuning experiments are done using GPUs on Google Colaboratory where each fine-tuning experiment requires 2 GPU hours.

A.5 License

Tatoeba data, GLUE data, Wikipedia dumps use the Creative Commons licenses. TDIL data used for Indic languages uses Research license type and Xtreme dataset uses Apache License 2.0. To the best of our knowledge, the use of scientific artifacts in this work is consistent with their intended use.

Indo-Aryan	Hindi:Vaapariyo, Marathi:Vaapartat , Punjabi:Vaaparan, Gujarati:Vaaparvana
	Hindi:Jaate, Marathi:Jaaoon , Punjabi:Jaana, Gujarati:Jaao
West-Germanic	English:Category, German:Kategorie, Dutch:Categorie, Western Frisian:Kategory
	English:University, German:Universitaten, Dutch:Universiteit, Western Frisian:Universiteiten
Romance	French:Association, Spanish:Associacion, Portuguese:Associacao, Italian:Associazione
	French:Certifie, Spanish:Certificar, Portuguese:Certificado, Italian:Certificato

Table 10: Lexically overlapping tokens with similar meanings across four languages in each of three families. OBPE, our proposed method, exploits such meaning-preserving overlap among related languages to induce vocabulary for multilingual learning.

Lang	hi HRL	mr	pa	gu	en HRL	de	nl	fy	fr HRL	es	pt	it	LRL avg	HRL avg
						NER								
BPE	83.66	45.03	25.85	24.25	75.94	52.42	62.83	62.63	85.75	70.53	68.34	64.34	52.91	81.78
CV	83.83	47.67	32.69	33.43	72.35	46.89	55.13	57.88	83.34	71.78	66.45	62.61	52.73	79.84
OBPE	85.92	47.55	26.05	32.79	77.15	52.72	62.87	65.55	85.76	73.35	70.25	64.69	55.09	82.94
						TC								
BPE	75.8	51.46	49.88	51.9	88.27	49.5			76.05	55.64			51.68	80.04
CV	76.46	54.37	55.49	56.33	81.94	51.5			74.81	54.31			54.40	77.74
OBPE	76.58	55.38	53.98	54.06	88.3	57.85			76.06	55.59			55.37	80.31
						POS								
BPE	93.96	74.84	59.34	65.87	94.81	69.18	74.96		96.33	86.66	84.67	82.81	74.79	95.03
CV	93.67	77.68	71.28	75.81	94.1	67.68	72.75		96.04	84.33	82.44	81.65	76.70	94.60
OBPE	94.11	75.46	58.84	68.5	94.94	68.1	75.18		96.22	86.54	84.3	83.46	75.05	95.09
						XNLI								
BPE					67.05	45.51			62.87	51.62			48.57	64.96
CV					54.87	39.87			59.48	48.68			44.28	57.18
OBPE					67.71	47.33			63.43	52.69			50.01	65.57

Table 11: Zero-shot performance of models in the Skewed-12 setting trained on 9 LRL and 3 HRL languages. Performance is measured on four tasks: NER (F1), Text Classification (Accuracy), POS (Accuracy), and XNLI (Accuracy). For all metrics, higher is better . Zero-shot transfer to LRL improves without hurting HRL accuracy. Averages results across HRLs and LRLs are presented in Table 5. OBPE shows gains here too. Section 4.2 has further discussion.

A.6 Data bias

We have used standard Wikipedia corpus, and there have been some studies on bias in such corpus.(Hube, 2017)

Lang	hi HRL	mr	pa	gu	en HRL	de	nl	fy	fr HRL	es	pt	it	LRL avg	HRL avg
NER														
BPE	85.49	54.88	75.35	40.5	74.99	53.16	62.91	66.54	84.24	70.14	70.2	63.86	61.95	81.57
OBPE($\alpha=0.5$)	86.59	59.23	76.15	41.84	74.74	56.95	63.19	67.92	83.73	69.99	69.76	64.91	63.33	81.69
OBPE($\alpha=0.7$)	85.99	59.54	75.59	41.37	75.36	54.6	63.43	66.86	83.95	71.77	69.27	66.29	63.19	81.77
TC														
bpe	83.97	68.01	74.24	77.1	88.2	57.6			77.45	53.45			66.08	83.21
OBPE($\alpha=0.5$)	83	71.78	75.21	78.28	88.28	62.41			76.88	54.19			68.37	82.72
OBPE($\alpha=0.7$)	83.56	69.3	74.84	77.09	87.93	57.9			77.11	57.84			67.39	82.87
POS														
bpe	94.14	81.7	86.57	86.86	94.5	69.2	80.39		95.79	88.62	84.8	85.74	82.99	94.81
OBPE($\alpha=0.5$)	94.18	82.79	86.63	86.5	94.6	70.53	79.49		95.94	88.79	86.62	86.41	83.47	94.91
OBPE($\alpha=0.7$)	94.1	81.56	87.04	86.55	94.38	70.67	79.99		96.17	89.8	87.77	86.19	83.70	94.88
XNLI														
bpe					65.79	48.3			63.21	54.93			51.62	64.50
OBPE($\alpha=0.5$)					66.77	50.84			66.77	53.27			52.06	66.77
OBPE($\alpha=0.7$)					66.37	48.54			63.57	54.85			51.70	64.97

Table 12: Zero-shot performance of three different models each trained on 3 LRLs and 1 HRL in the respective families 2 in the SMALLCAPS BALANCED-4 setting . Performance is measured on four tasks: NER (F1), Text Classification (Accuracy), POS (Accuracy), and XNLI (Accuracy). For all metrics, higher is better . Zero-shot transfer to LRL improves without hurting HRL accuracy. OBPE shows gains here too. Languages in Romance family show some improvements in $\alpha = 0.7$ setting as compared to $\alpha = 0.5$. (Section sec:ablation)

	% overlap retained	En-Es				Hi-Mr		
		High		Low		High		Low
		es	en	es	hi	mr	hi	mr
NER	100	72.3	75.1	63.4	85.9	55.6	86.3	58.2
	0	70.9	67.7	51.7	82.7	43.4	85.1	16.6
TC	100		88.2	63.7	84.4	75.1	84.6	71.4
	0		82.6	53.8	78.9	72.4	84.5	30.1
POS	100		94.7	82.9	94.2	83.3	94.2	81.9
	0		92.8	60.4	94.0	76.7	94.2	74.1
XNLI	100	61.9	66.6	55.2				
	0	62.6	61.5	53.9				

Table 13: Accuracy of Zero-shot transfer from English (En) as HRL to Spanish (Es) and from Hindi(Hi) as HRL to Marathi(Mr) in two settings: (1) High where Es,Mr have sizes comparable to the HRL and (2) Low where their sizes are only 20K. As the percentage of overlapping tokens retained is decreased from 100% to 0%, the accuracy drops but the drop is higher in the low-resource setting. Task-specific accuracy numbers in first column(En-Es-High-es) have been taken from (K et al., 2020). Table 6 contains the reduction in accuracy on decreasing overlap from 100% to 0 % i.e. the difference between the rows corresponding to 100% and 0%

Lang	hi	mr	pa	gu	en	de	nl	fy	fr	es	pt	it	avg
Method(p)							NER						
OBPE(1)=BPE	86.57	59.71	69.71	41.89	77.42	60.14	67.87	69.73	85.79	72.96	71.02	67.31	69.18
OBPE(0)	87.29	61.86	67.21	41.46	76.08	59.50	67.30	66.86	86.02	69.80	70.89	67.54	68.48
OBPE(-1)	86.67	64.19	72.38	39.93	77.17	58.25	67.09	69.86	85.83	72.99	70.43	66.02	69.23
OBPE(-2)	86.17	60.91	67.30	44.43	76.47	59.66	67.13	70.03	85.25	75.02	71.82	66.53	69.23
OBPE(-3)	87.14	62.68	72.25	44.73	77.24	61.41	67.38	69.87	86.15	69.82	71.06	65.37	69.59
OBPE(-∞)	87.09	62.96	72.17	44.25	77.93	60.44	68.65	70.23	86.92	74.14	72.55	66.05	70.28
BPE-dp	85.54	62.64	71.46	39.75	75.51	59.29	67.76	70.42	84.15	67.43	68.82	67.74	68.38
TokComp	86.43	61.12	72.82	45.88	76.57	55.25	65.28	67.85	84.22	71.04	68.87	66.00	68.44
CV	84.27	55.66	43.37	50.19	74.99	53.51	65.36	65.39	84.20	73.05	66.16	63.49	64.97
Bsamp	84.68	59.73	67.31	40.62	76.76	61.34	67.29	71.80	85.89	73.71	71.20	66.69	68.92
Osamp	84.71	63.22	67.82	42.03	77.83	62.35	68.08	71.59	85.50	69.16	70.06	66.70	69.09
							TC						
OBPE(1)=BPE	80.35	61.45	69.00	72.32	88.63	62.27			77.23	62.58			71.73
OBPE(0)	80.11	64.07	68.26	70.48	87.61	54.96			76.53	62.23			70.53
OBPE(-1)	80.00	64.37	69.10	72.10	87.89	66.25			77.33	65.36			72.80
OBPE(-2)	79.21	64.83	68.58	70.41	88.17	65.76			76.78	58.71			71.56
OBPE(-3)	81.00	62.79	68.17	73.20	89.38	68.34			77.50	63.84			73.03
OBPE(−∞)	80.68	68.90	70.03	72.14	87.92	66.05			77.14	63.00			73.23
BPE-dp	79.68	63.45	69.43	70.36	87.39	59.75			76.15	57.76			70.50
TokComp	82.06	67.17	70.42	72.48	88.02	58.47			77.22	60.29			72.02
CV	79.90	61.33	65.68	68.96	87.98	55.79			74.92	57.79			69.04
Bsamp	81.00	65.29	70.97	72.30	88.05	66.11			76.92	63.51			73.02
Osamp	80.11	66.08	70.11	72.38	88.39	66.25			76.57	64.58			73.06
							POS						
OBPE(1)=BPE	94.22	79.60	86.83	86.21	94.91	77.70	82.00		96.47	89.74	87.79	87.27	87.52
OBPE(0)	94.13	76.26	86.53	85.03	94.85	76.09	82.48		96.31	88.78	87.01	86.62	86.74
OBPE(-1)	94.20	79.13	86.23	85.14	94.87	78.22	82.56		96.32	89.62	87.25	87.27	87.34
OBPE(-2)	93.98	81.07	86.54	85.86	94.68	76.80	82.08		96.23	89.14	86.31	86.45	87.19
OBPE(-3)	94.31	79.55	86.67	86.65	95.03	76.34	83.63		96.30	89.97	87.76	88.00	87.66
OBPE(−∞)	94.18	81.55	87.01	86.76	94.98	79.28	82.38		96.40	90.04	88.01	88.21	87.94
BPE-dp	93.26	80.03	86.31	85.23	94.49	77.90	83.07		96.10	90.01	87.84	87.63	87.44
TokComp	93.99	80.38	86.75	86.79	94.79	79.50	84.91		95.80	89.87	87.05	88.70	88.05
CV	93.12	74.82	84.62	81.56	94.28	74.74	79.41		96.01	87.51	85.52	85.27	85.17
Bsamp	94.3	77.81	86.35	85.68	94.93	77.2	82.94		96.34	89.93	88.05	88.54	87.46
Osamp	94.01	81.22	86.66	86.36	94.35	76.99	82.93		96.04	89.75	88.24	88.43	87.73
							XNLI						
OBPE(1)=BPE					64.35	50.36			61.06	53.77			57.39
OBPE(0)					64.33	49.06			59.96	54.71			57.02
OBPE(-1)					64.35	48.62			61.40	53.51			56.97
OBPE(-2)					65.05	50.36			64.45	55.31			58.79
OBPE(-3)					67.86	50.64			64.85	57.11			60.11
OBPE(−∞)					67.41	50.76			65.13	57.29			60.15
BPE-dp					64.31	50.16			63.17	55.17			58.20
TokComp					67.98	53.05			64.21	54.83			60.02
CV					65.19	47.43			63.83	51.16			56.90
Bsamp					63.41	51.02			60.58	53.09			57.03
Osamp					67.13	50.38			64.25	56.64			59.60

Table 14: Zero-shot performance of models in the Balanced-12 setting trained on 9 LRL and 3 HRL languages. Performance is measured on four tasks: NER (F1), Text Classification (Accuracy), POS (Accuracy), and XNLI (Accuracy). For all metrics, higher is better . Zero-shot transfer to LRL improves without hurting HRL accuracy. Averages results across HRLs and LRLs are presented in Table 4. Section 4.2 has further discussion.Table 4 contains the values corresponding to rows BPE, BPE-dp, CV, TokComp, OBPE(−∞) averaged over LRLs and HRLs, Table 7 contains the values corresponding to rows Bsamp, Osamp averaged over LRLs and HRLs, , Figure 3 plots the rows correponding to varying p values. (Section 4.5)