How Linguistically Fair Are Multilingual Pre-Trained Language Models?

Monojit Choudhury, Amit Deshpande
Microsoft Research Lab India
Email: {*monojitc, amitdesh*}*@microsoft.com*

Abstract

Massively multilingual pre-trained language models, such as mBERT and XLM-RoBERTa, have received significant attention in the recent NLP literature for their excellent capability towards crosslingual zero-shot transfer of NLP tasks. This is especially promising because a large number of languages have no or very little labeled data for supervised learning. Moreover, a substantially improved performance on low resource languages without any significant degradation of accuracy for high resource languages lead us to believe that these models will help attain a fairer distribution of language technologies despite the prevalent unfair and extremely skewed distribution of resources across the world's languages.

Nevertheless, these models, and the experimental approaches adopted by the researchers to arrive at those, have been criticised by some for lacking a nuanced and thorough comparison of benefits across languages and tasks. A related and important question that has received little attention is how to choose from a set of models, when no single model significantly outperforms the others on all tasks and languages. As we discuss in this paper, this is often the case, and the choices are usually made without a clear articulation of reasons or underlying fairness assumptions. In this work, we scrutinize the choices made in previous work, and propose a few different strategies for fair and efficient model selection based on the principles of fairness in economics and social choice theory. In particular, we emphasize Rawlsian fairness, which provides an appropriate framework for making fair (with respect to languages, or tasks, or both) choices while selecting multilingual pre-trained language models for a practical or scientific set-up.

1 Introduction

NLP technologies require large amount of labeled and/or unlabeled data for training. The distribution of resources across languages, on the other hand, is extremely skewed (Joshi et al. 2020; Bender 2011), because data annotation is an expensive and effort-intensive affair. This digital divide is further widened by modern technologies, especially the deep learning approaches to NLP, which require even more data that all but a handful of the world's languages possess.

Since their introduction in 2019, massively Multilingual pre-trained language models (MultiLMs), such as Multilingual BERT (mBERT) (Devlin et al. 2019), XLM-RoBERTa (XLM-R) (Conneau et al. 2020), Massively Multilingual Translation Encoder (MMTE) (Arivazhagan et al. 2019) and Unicoder (Liang et al. 2020), have received significant attention from the NLP researchers. Their surprisingly well crosslingual zero-shot transfer capabilities (Pires, Schlinger, and Garrette 2019; Wu and Dredze 2020) have revolutionized our approach to multilingual NLP, and offer a promise of sophisticated and efficient NLP systems for all languages regardless of the availability of labeled data. This is because once a MultiLM is pre-trained on unlabeled corpora of a large number (typically 100+) of languages, task-specific labeled data for fine-tuning is required only for one language, called the pivot (usually a high-resource language such as English). The fine-tuned model works across all the languages that the MultiLM was pre-trained on, albeit with varying degrees of success.

This crosslingual zero-shot setting also creates the following interesting and important dilemma which researchers and engineers working in this field are often faced with. Imagine that a researcher has come up with two MultiLMs, A and B. She decides to test her models on a standard benchmark, say XNLI (Conneau et al. 2018), which has training data for Natural Language Inferencing task in English and test data for 15 languages including English. She observes that A performs better than B on 10 languages, B performs better than A on 3 languages, and on the remaining two, the models perform equally well. Should she declare that model A is better than B because it outperforms the latter on *most* of the languages in XNLI? Or should she declare the one with higher average accuracy as the winner? Or should it be some other statistic, say the median or geometric mean of the performances, according to which the models should be compared?

We will refer to this dilemma as the *Multilingual Model Selection Problem* or MMSP for short. The problem is not only philosophically interesting, but also has great practical significance for two reasons. First, before our researcher can resolve the dilemma, she has to decide what does "better" mean in the case of MultiLMs. Intuitively, *average performance* over the set of tested languages appears to be a sensible choice, and has been adopted by several researchers (e.g., Liang et al. (2020); Aharoni, Johnson, and Firat (2019)). However, the *average performance* on a limited set of tasks

and languages does not capture practically important factors such as the set of languages and tasks the model is expected to work well on, and the amount of data across languages that the MultiLM was pre-trained and fine-tuned on (Hu et al. 2020; Wu and Dredze 2020; Lauscher et al. 2020). More importantly, as far as we have seen, authors do not provide any formal justification or articulate the guiding principles behind their choice of the statistic.

Second, one should keep in mind that one of the primary advantages as well as the reason behind the popularity of MultiLMs is their excellent crosslingual zero-shot transfer ability. As Wu and Dredze (2020) point out, "*top high resource languages are slightly hurt by massively multilingual joint training*"; yet these are preferred because they offer a solution for low-resource languages which do not have sufficient labeled data. Indeed, equal or equitable accuracy of MultiLMs across languages and tasks has been one of the critical points of scrutiny in the recent times (Hu et al. 2020). If this is the central tenet behind the conception and construction of MultiLMs, then MMSP must be resolved in a manner that conforms to this normative principle of "fairness" or "distributive justice" across languages.

In this paper, drawing inspiration and ideas from the discourse on fairness in machine learning, ethics, social choice theory, economics and decision theory, we provide several possible resolution of the MMSP, which are driven by different choice of the normative principles of efficiency and fairness. We do not give any new results but instead scrutinize a few popular and important works on MultiLMs by explicitly calling out the principles of distributive justice entailed by the choices made by the researchers while resolving the MMSP, but not stated as such. As we shall see, most of the work, whenever possible, follow the *Pareto-efficiency* principle, i.e., choose the model which does as good or better than all others on *all languages* that were tested for; otherwise, a *utilitarian* approach is adopted, where a simple unweighted average performance across languages is used as the model selection criterion. We further argue that the utilitarian approach does not conform to the *egalitarian* principles on which MultiLMs are founded. Instead, given the skewed resource distribution across languages and other limitations of the current technology, a practical trade-off between the egalitarian and utilitarian ideologies could be the *prioritarian* or *Rawlsian principle* (Rawls 1999) of distributive justice, where one should select the model which maximizes the minimum performance over all languages. Rawlsian fairness is based on the *principle of least difference*, and proposes to narrow the gap between unequal accuracies or utilities, instead of equalizing them. Interestingly, Rawlsian fairness based resolution of MMSP is also the robust utilitarian choice under an adversarial assumption, and it also ensures a progressively more egalitarian distribution under the assumption that language resources for all languages will grow over time, without being severely unfair to high performing languages in the short run.

Recently several researchers have critiqued the utilitarian approach to MMSP (Wu and Dredze 2020; Hu et al. 2020); the objective of those studies have been to empirically show that in most cases a single Pareto-efficient model does not exist. These critiques, however, do not propose any mechanism to resolve the MMSP under such a situation. Instead, they suggest that given the current state of limited understanding and testing of the models on a handful of languages and tasks, the resolution of MMSP should be deferred till we have a clearer understanding of the model performances. While we agree with the criticisms, and the necessity of wide-scale experimentation, it is unlikely that in the near future we will have a MultiLM that is Pareto-efficient across languages and tasks. Therefore, it is useful and important to resolve the MMSP under partial knowledge in such a manner that the solutions hold irrespective of the state of technology and resource availability.

It is important to note that the ongoing discourse on fairness in ML (Mehrabi et al. 2019; Barocas, Hardt, and Narayanan 2017; Leben 2020), and more specifically NLP (Blodgett et al. 2020), argue for individual and group fairness (Binns 2018). We are not aware of any work that discuss the specific issues of distributive justice when languages, instead of individuals, are viewed as *entities*. While a language can be equated to a group of individuals, as we discuss in Sec 4, there are important philosophical and practical distinctions between these notions, such that they merit independent treatments.

The rest of the paper is organized as follows. In Sec 2, we formally introduce the MMSP and discuss its potential resolutions under various normative principles of distributive justice. This section will also introduce the reader to the basic concepts and some well established results of distributive justice in social choice theory. Sec 3 reinterprets the model selection decisions taken in the recent papers. We discuss certain philosophical, technical and practical aspects of MMSP and its resolution in Sec 4, and conclude in Sec 5 with some practical recommendations.

2 Principles of optimal model selection

In this section, we formally define the *Multilingual Model Selection Problem* (MMSP) and lay down fundamental guiding principles from economics and ethics for optimal model selection. To apply these principles, languages need to be considered as individual entities instead of representing a group of users (see details in Sec 4).

Multilingual Model Selection Problem (MMSP): Given two MultiLMs A and B along with m different tasks $t_1, t_2, \ldots, t_m$ and n different languages $L_1, L_2, \ldots, L_n$, let a_{ij} and b_{ij} respectively denote the accuracy of model A and B for task t_i in language L_j. This subsumes the case $m = 1$, where we are given the accuracy of two MultiLMs on different languages but for the same task. For simplicity, we keep aside the pros and cons of various modeling and training choices, and focus only on their performance as black-box models. Then the problem of picking the better model between A and B boils down to comparing their performance vectors $\mathbf{a} = (a_{11}, \ldots, a_{mn})$ and $\mathbf{b} = (b_{11}, \ldots, b_{mn})$ according to an objective that quantifies efficiency and fairness. Such objectives are well-studied in economics, ethics, social choice theory, decision theory, and are guided by certain normative principles of efficiency and fairness (Yaari 1981; Sen 1985; Young 1994). There is a recent, renewed

interest to revisit these principles in the context of fairness in machine learning (Binns 2018; Heidari et al. 2019; Leben 2020; Hossain, Mladenovic, and Shah 2020).

2.1 Utilitarian multilingual model selection

When we focus only on the performance, the choice space of models can be identified with their performance vectors, i.e., we identify a model A with its performance vector $\mathbf{a} = (a_{11}, \ldots, a_{mn})$, whose coordinates are its accuracies on different languages for different tasks. When building state of the art (SOTA) MultiLMs, the ideal is to create a model whose performance surpasses others in all languages and tasks. This is formalized by the well-known *Pareto-efficiency principle* stated below.

Definition 1. *(Pareto-efficiency and Pareto-optimality)*

1. *The* Pareto-efficiency principle *suggests that a model A is better than a model B, if $\mathbf{a}$ Pareto-dominates $\mathbf{b}$, i.e., $a_{ij} \geq b_{ij}$, for all i, j.*
2. *A model A is called Pareto-optimal, if there is no other model that performs better than A in every coordinate, i.e., no other model Pareto-dominates it.*

Note that Pareto-optimality does not necessarily mean Pareto-dominance over all other models but rather a guarantee that no other model would Pareto-dominate it. To understand Pareto-optimality, we need to understand the choice space of all available models. A common assumption about the choice space is called *non-polarization*, as defined below.

Definition 2. *(Non-polarization) The choice space satisfies* non-polarization*, if the set of all performance vectors $\mathbf{a}$, as A varies over all available model choices, is a convex set $\mathcal{C}$.*

Non-polarization is a reasonable assumption because we can always do a thought experiment that picks the model A with probability p and the model B with probability $1 - p$ to achieve performance vector given by the convex combination $p\mathbf{a} + (1 - p)\mathbf{b}$. The following is a folklore result, commonly manifested as von Neumann-Morgenstern utility theorem in decision theory (von Neumann, Morgenstern, and Rubinstein 1944) and Harsanyi's utilitarian theorem in economics (Harsanyi 1955).

Proposition 1. *Assuming non-polarization, a model A is Pareto-optimal if and only if A maximizes the weighted accuracy $\sum_{i=1}^{m}\sum_{j=1}^{n} w_{ij}a_{ij}$, for some non-negative weights w_{ij}, over all models in the choice space.*

By rescaling, we can assume that the weights give a convex combination, i.e., the weight vector $\mathbf{w}$ lies in $\Delta = \{\mathbf{w} : \sum_{i=1}^{m}\sum_{j=1}^{n} w_{ij} = 1 \text{ and } 0 \leq w_{ij} \leq 1, \forall i, j\}$.

It is important to note that Proposition 1 does not prescribe any specific weights. The natural choice being equal weights leads to the average accuracy objective. The utilitarian approach of maximizing the weighted accuracy is justified only when the weights are chosen appropriately after calibrating the accuracies across languages and tasks against each other, based on considerations, such as, the size of the training and test data across languages, and difficulty level and usefulness of the tasks. Thus, for MultiLMs, the choice of weights can be argued further based on domain expertise and the economics of model deployment.

However, the utilitarian promise of Pareto-optimality falls short for several reasons. Firstly, the choice of weights that makes a model Pareto-optimal may be debatable, leading to two or more SOTA models such that no single model Pareto-dominates the others. Secondly, if we think of the choice of weights as reflecting the balance of test data across languages and tasks, or usefulness of a task, then this balance may change over time after deployment, and so would our choice for the best MultiLM. It is a priori unclear if there is a choice of weights that is both utilitarian and robust to the above considerations. In case of unknown weights, a naturally robust and conservative choice of weights is given by

$$\mathbf{w}_{\text{robust}} = \operatorname*{argmin}_{\mathbf{w}\in\Delta} \max_{\mathbf{a}\in\mathcal{C}} \sum_{i=1}^{m}\sum_{j=1}^{n} w_{ij}a_{ij}. \tag{1}$$

In practice, Δ can be a carefully chosen set based on domain knowledge and robustness requirement. We shall revisit Equation (1) in a different light in the next subsection.

2.2 Fair multilingual model selection

Fairness across different languages has been an implicit goal in MultiLMs, as better models must have high accuracy on both high-resource and low-resource languages. A commonly studied notion of fairness in machine learning is egalitarian, where an ideal fair classifier must have predictive parity (e.g., equal accuracy, equal false positive rates) across certain demographics (Barocas, Hardt, and Narayanan 2019). Similarly, equal accuracy across different languages is a desirable long-term goal for MultiLMs. However, equal or near-equal accuracy is a constraint and not an objective in itself. In group-fairness literature, there is a long line of work in training a fair model to maximize average accuracy subject to egalitarian constraints across different groups; see (Celis et al.; Zafar et al. 2019) and the references therein. However, the training guarantees do not easily carry over to the test data. Moreover, insisting on equal or near-equal accuracy across different languages may lead us to pick a sub-optimal model of diminished overall accuracy, especially, when the availability of training data is inherently unequal across languages and the difficulty of each task is inherently different.

The *least difference principle* proposed by Rawls (1999) in distributive justice offers another perspective at the egalitarian approach, and proposes to narrow the gap between unequal accuracies instead of insisting on equalizing them.

Definition 3. *(Least difference principle)*
Consider two models A and B with performance vectors $\mathbf{a} = (a_{11}, \ldots, a_{mn})$ and $\mathbf{b} = (b_{11}, \ldots, b_{mn})$, respectively. Suppose there exist two indices (i, j) and (k, l) such that $a_{ij} < b_{ij} < b_{kl} < a_{kl}$, and $a_{pq} = b_{pq}$ on all other indices $(p, q) \notin \{(i, j), (k, l)\}$. Then the least different principle prefers model B over A because it reduces the gap between two unequal coordinates while keeping the rest unchanged.

Decision theory and social choice theory literature contains additional axioms that a *well-behaved* choice process

must satisfy, e.g., the von Neumann-Morgenstern (VNM) rationality axioms (von Neumann, Morgenstern, and Rubinstein 1944) and the *Independence of Irrelevant Alternatives (IIA)* (Ray 1973). In our context, the most important properties for a choice process to be *well-behaved* are as follows.

1. For any permutation σ of the coordinates, if we prefer $\mathbf{a}$ over $\mathbf{b}$ then we must prefer $\sigma(\mathbf{a})$ over $\sigma(\mathbf{b})$.
2. For any monotone function $f : [0,1] \to [0,1]$, if we prefer $\mathbf{a}$ over $\mathbf{b}$ then we must prefer $\tilde{\mathbf{a}} = (f(a_{11}), \ldots, f(a_{mn}))$ over $\tilde{\mathbf{b}} = (f(b_{11}), \ldots, f(b_{mn}))$.

Keeping the mathematical description aside, a high-level takeaway from the independence of irrelevant alternatives for MultiLMs is that if an ideal ordering or preference among models is well-behaved then any modification to training or testing methodology that affects all the accuracies in similar fashion should not affect the ideal ordering or preference among models.

Previous work in economics has shown that the above principles and axioms guide us to a well-defined max-min objective for finding an optimal model from a given set of feasible models. Although, the max-min objective looks prioritarian, where the benefits of the worst-off matter more than the benefits of the better-off, the following proposition shows that it can be derived from utilitarian considerations of Pareto-efficiency and the Rawlsian fairness through least different principle. It is a direct consequence of a theorem due to Hammond (Hammond 1976) and Strasnick (Strasnick 1976), which we restate from (Yaari 1981).

Proposition 2. *Assuming non-polarization, Pareto-efficiency and least difference principles, any well-behaved optimal choice A^* must be a solution to* $\max_{\mathbf{a}\in\mathcal{C}} \min_{i,j} a_{ij}$.

Proposition 2 implies that any optimal MultiLM model that satisfies Pareto-efficiency and Rawlsian fairness must maximize the minimum accuracy. This is in contrast with the egalitarian solutions that must equalize accuracies. Moreover, we can rewrite $\max_{\mathbf{a}\in\mathcal{C}} \min_{i,j} a_{ij}$ as the minimum over convex combinations and use Sion's minimax theorem (Sion 1958) and Equation (1) to get

$$\max_{\mathbf{a}\in\mathcal{C}} \min_{\mathbf{w}\in\Delta} \sum_{i=1}^{m}\sum_{j=1}^{n} w_{ij}a_{ij} = \min_{\mathbf{w}\in\Delta} \max_{\mathbf{a}\in\mathcal{C}} \sum_{i=1}^{m}\sum_{j=1}^{n} w_{ij}a_{ij}$$
$$= \sum_{i=1}^{m}\sum_{j=1}^{n} (\mathbf{w}_{\text{robust}})_{ij}a_{ij}.$$

Thus, it is interesting to note that the optimal choice for Rawlsian fairness coincides with the optimal robust choice made from a purely utilitarian viewpoint.

2.3 Deontological multilingual model selection

The approaches discussed above try to find the right objective so that the outcome (namely, the optimal choice of MultiLM) would meet certain efficiency and fairness principles. In ethics, the outcome-oriented approaches are known as consequentialist. On the other hand, deontological approaches divide the benefits proportional to their rights, rewarding intended beneficiaries or compensating the victims. For example, an egalitarian view of dividing the rights equally leads to weighted accuracy with equal weights. Deontological approach has also seen renewed interest in fair machine learning (Leben 2020; Saleiro et al. 2018).

In deontological approach to fairness, the philosophical concepts of moral luck and desert play a key role. Considering the resources from different languages that went into training a MultiLM, the benefits must be divided in some way. What we consider as resources are often split further into luck and work parts. In the context of languages and MultiLMs, an example of luck is typological similarities (Bender 2011; Joshi et al. 2020), e.g., a model trained on English would do much better on Dutch than Japanese because of typological similarities (Lin et al. 2019). One interpretation of the work part can be the amount of collective effort or money spent by the NLP community to create labeled data for different languages. If we take an approach that the inequality between high-resource and low-resource languages is pure luck, then a libertarian objective would be a weighted accuracy, where the weights are proportional to the resources that went into training a MultiLM. However, if we follow the approach of moral desert, an equally weighted accuracy or considering only the gains above a common baseline, would be well-justified. If we make a finer distinction between luck and work parts, we could say that a MultiLM is deontologically fair as long as it does not penalize any language due to typological difference but it is okay to have unequal accuracy across languages based on the availability of training data.

3 Reinterpreting the Choices Made in MultiLM Literature

In this section, we will review the current trends in resolution of MMSP, and reinterpret those decisions in light of the fairness and efficiency principles discussed in the previous section. Our aim here is not to conduct a thorough technical survey of MultiLMs; neither it is to quantify the statistics of normative principles that are followed. Rather, we shall discuss a few representative work to illustrate two broad trends that we observe in the literature - Pareto-efficiency based resolution, and the average performance across languages, both of which are based on the utilitarian principle. Then we discuss how some of these resolutions would have changed if we were to follow the Rawlsian fairness principles. We also present some of the recent critiques of MMSP, which, as we shall argue, are mostly deontological in nature.

A Note on Tasks and Datasets: MultiLMs are typically evaluated on test-benches such as XNLI (Conneau et al. 2018), XGLUE (Liang et al. 2020) [a cross-lingual evaluation benchmark consisting of NER (Sang 2002), POS (Zeman et al. 2019), News Classification (NC), MLQA (Lewis et al. 2019), XNLI, PAWS-X (Yang et al. 2019), Query-Ad Matching (QADSM), Web Page Ranking (WPR), QA Matching (QAM), Question Generation (QG) and News Title Generation (NTG) tasks], XTREME (Hu et al. 2020) [also a collection of cross-lingual test benches consisting of XNLI, PAWS-X, POS, NER (Pan et al. 2017),

XQUAD (Artetxe, Ruder, and Yogatama 2020), MLQA, TyDiQA (Clark et al. 2020), BUCC (Zweigenbaum, Sharoff, and Rapp 2018) and Tatoeba (Artetxe and Schwenk 2019)], GLUECoS (Khanuja et al. 2020), and LinCE (Aguilar, Kar, and Solorio 2020). While comparing MultiLMs, researchers report and compare the performance of the models on the datasets, rather than the tasks. For instance, MLQA and XQuAD involve the same NLP task, namely Question Answering through span extraction from Wikipedia, but the accuracies on these are reported independently. Therefore, here we shall use the term *task* to refer to the dataset, rather than the task itself.

3.1 Pareto-Efficiency Principle

Arguably, the aim of most model building exercises is to come up with MultiLM architectures, training processes or simply larger models that are Pareto-efficient with respect to tasks and languages that it has been tested on, in comparison to the existing SOTA MultiLMs. Whenever possible, this indeed would be the ideal resolution of MMSP under any of the consequentialist approaches - Utilitarian as well as Prioritarian (or Rawlsian). Some of the popular MultiLMs had achieved Pareto-efficiency over the existing SOTA, often by a considerable margin for all languages. For example, XLM (Conneau and Lample 2019) Pareto-dominates mBERT (Devlin et al. 2019) as well as (Artetxe and Schwenk 2019) for XNLI; Unicoder (Huang et al. 2019) Pareto-dominates XLM, and XLM-R (Conneau et al. 2020) Pareto-dominates Unicoder, again for XNLI. On the GLUECoS dataset, Khanuja et al. (2020) show that their proposed method is Pareto-efficient across all *tasks* for English-Spanish code-mixing, though it is not the case for English-Hindi, where mBERT and the proposed method are Pareto-optimal over the tasks.

3.2 Utilitarian Principle

Whenever Pareto-efficiency for a model is not observed, usually the *average* performance over all languages is used as a metric to determine the "better" model. An illustrative case-study is that of Liang et al. (2020), where along with a new test-bench, XGLUE, authors also propose a new MultiLM - the *Unicoder*. Unicoder is compared to mBERT and XLM-R on 11 tasks across 19 languages (Table 4 in the paper), where average across languages, and then average across tasks are used as the indicators to declare that Unicoder is the best performing model. This is clearly an application of the *utilitarian principle*.

In Table 1, we summarize the performance of the models across tasks (except for the two text generation tasks, QG and NTG, for which a different metric of performance is used) by indicating the average and minimum values, as well as the Pareto-optimality of the models. As we can see, Unicoder has a higher average for all but two tasks - POS and QADSM. Nevertheless, it is Pareto-efficient only for three tasks - XNLI, PAWS-X and QAM.

Also interesting are the experiments on pivots for XNLI, where the authors show that the performance of the model across languages are substantially influenced by the choice of the pivot language (Table 5 in the paper). Here too, since no pivot language provides a Pareto-efficient model, the authors resolve the MMSP by looking at the average performance across languages, concluding that "*the best pivot languages are Spanish (es), Greek (el) and Turkish (tr), rather than English (en).*" Authors do not offer any principled approach for breaking this tie.

Tasks	mBERT	XLM-R	Unicoder
NER	**78.2, 69.2**	**79.0, 70.4**	**79.7, 71.8**
POS	74.7, 43.3	**79.8, 55.2**	**79.6, 56.3**
NC	82.7, 78.0	**83.4, 78.2**	**83.5, 78.5**,
MLQA	**60.7, 47.9**	**65.1, 60.5**	**66.0, 62.1**,
XNLI	66.3, 50.4	74.2, 64.7	**75.3, 66.3**,
PAWS-X	87.2, 82.9	89.5, 86.9	**90.1, 87.4**,
QADSM	64.2, 60.3	**68.6, 65.8**	**68.4, 64.6**
WPR	**73.5, 64.5**	**73.8, 63.9**	**73.9, 64.4**,
QAM	66.1, 64.7	68.4, 67.8	**68.9, 68.4**,

Table 1: Performance of mBERT, XLM-R and Unicoder taken from (Liang et al. 2020). The first and second numbers in each cell denote the average and minimum performances respectively. The numbers in bold indicate that the model is Pareto-optimal on that task; if there is only one bold cell in a row, it indicates that the model is Pareto-efficient. Gray cells denote the Rawlsian choice for the task.

Some other studies that use average across languages for reporting and/or model selection purposes include (Aharoni, Johnson, and Firat 2019) for machine translation, (Ahmad et al. 2019) for parsing, and (Rijhwani et al. 2019) for entity linking. We have not come across any work that applies weighted average, presumably because weights are difficult to determine and justify. Also, it is uncommon to average over tasks; usually, the performance range across tasks, and the nature and complexity of the tasks are too diverse for average to be a meaningful quantity.

3.3 Hypothetical Resolutions under Rawlsian Fairness

What if instead of Pareto-efficiency or the utilitarian notion of averaging, we were to resolve the MMSP through Rawlsian fairness? Recall that Rawlsian fairness recommends the selection of the model that maximizes the minimum performance over the languages. While this looks like a prioritian objective, it is actually also the robust utilitarian choice under the very realistic assumption that the language-resources and utility of tasks change over time.

Let us focus again on Table 1, where the cells highlighted in gray show the Rawlsian choice of the model for a task; or in other words, the cells that have the maximum value for the minimum (the second number in the tuples) in a row. Unicoder emerges as the Rawlsian choice for 7 out of the 9 tasks (actually 9 of the 11 tasks), except for QADSM and WPR. Further, if we were to look at the max-min across all tasks and languages to make our final choice, it would still be the Unicoder. Thus, instead of averaging across languages and tasks, Unicoder's superiority over the other models could as well be established through this Rawlsian resolution of MMSP.

Tasks	mBERT	XLM	XLM-R	MMTE
XNLI	65.4, 49.7	69.1, 58.7	**79.2, 71.2**	67.5, 61.9
PAWS-X	81.9, 69.6	80.9, 64.8	**86.4, 79.0**	81.3, 69.2
POS	**70.3, 41.7**	70.1, 20.5	**72.6, 15.9**	**72.3, 43.1**
NER	**62.2, 3.6**	**61.2, 0.3**	**65.4, 1.3**	**58.3, 3.9**
XQuAD	64.5, 42.7	59.8, 35.4	**76.6, 59.3**	64.4, 48.4
MLQA	61.4, 50.2	48.5, 34.4	**71.6, 62.1**	60.3, 46.2
TyDiQA	**59.7, 49.3**	43.6, 14.2	**65.1, 31.9**	58.1, 49.9
BUCC	56.7, 50.0	56.8, 46.6	**66.0, 56.7**	**59.8, 53.3**
Tatoeba	**38.7, 11.5**	**32.6, 12.4**,	**57.3, 14.1**	37.9, -

Table 2: A summary of results on the XTREME benchmark as reported in (Hu et al. 2020). We follow the same convention as for Table 1.

Interestingly, under the Rawlsian fairness assumption, the utilitarian tie among the three pivot languages - Greek, Turkish and Spanish (Table 5, Liang et al. 2020) breaks in favor of Turkish. More importantly and surprisingly, Swahili, for which the accuracies are the lowest for most pivots, emerges as the overall Rawlsian choice for the pivot language.

The XTREME test-bench (Hu et al. 2020) is yet another interesting case-study. Table 2 presents a summary of the results from the paper on the XTREME tasks. XLM-R performs better than the other three models - XLM, mBERT and MMTE in terms of the utilitarian objective of average. It is also Pareto-optimal for all tasks and Pareto-efficient for 5 out of 9 tasks. However, by the Rawlsian criterion (the cells highlighted in gray in the Table) one would choose MMTE for 3 tasks, namely POS, NER and TyDiQA. The authors, on the other hand, summarize their observations as: "*XLMR is the best-performing zero-shot transfer model and generally improves upon mBERT significantly... MMTE achieves performance competitive with mBERT on most tasks, with stronger results on XNLI, POS, and BUCC.*"

3.4 Deontological Critiques of MultiLMs

Critiques of the current approaches to MultiLMs have followed three main lines of arguments. The first line of argument is concerned with their limited evaluation on a few tasks and languages, even though the explicit or implicit claim has been that zero-shot transfer works for all languages that the MultiLMs have been pre-trained on (Hu et al. 2020; Wu and Dredze 2020). This, the critiques argue, is misleading because for most languages the models' performance is unknown. The second line of criticisms challenges the fundamental assumption that MultiLMs provide a solution to low-resource languages by questioning the baselines they are compared against. For instance, according to Wu and Dredze (2020), "*the 30% languages with least pretraining resources perform worse than using no pretrained language model at all. Therefore, we caution against using mBERT alone for low resource languages ... On the other end of the spectrum, the highest resource languages (top 10%) are hurt by massively multilingual joint training.*" Similarly, Joshi et al. (2020) argue that MultiLMs do not help 88% of the world's languages for which even unlabeled data is unavailable; the only classes of languages it helps, if at all, are those which have a large amount of unlabeled data but not sufficient labeled data, namely the "Rising Stars" and "The Hopefuls" which together have 47 languages.

The third set of critiques challenge the zero-shot assumption for model building and evaluation of MultiLMs (Lauscher et al. 2020; Artetxe et al. 2020). In practice, they argue, one can always build some labeled resources for a language if one intends to improve the performance of the system in that language. In general, several studies have shown that building shared MultiLMs with a smaller set of related languages (Lin et al. 2019; Ahmad et al. 2019), and fine-tuning in a few-shot rather than zero-shot mode (Lauscher et al. 2020) is more effective than massively multilingual universal models. Conneau et al. (2020) show that for a given model size, the average zero-shot performance increases only up to addition of certain number of languages during the pre-training phase, after which it starts declining; this would make truly universal MultiLMs impractically large. They refer to this phenomenon as "the curse of multilinguality".

We observe that these critiques are based on certain deontological principles of fairness, where all languages are assumed to have an *a priori right to technology* irrespective of the number of speakers or the availability of resources. The first line of arguments are more egalitarian in nature, where the ideal objective is to have equal accuracies across all languages. For instance, Hu et al. (2020) suggests a metric called the "transfer gap" which measures the average difference between the performance for the pivot and other languages. An ideal MultiLM should minimize the transfer gap, and thus, fulfill the egalitarian objective of equal accuracy across all languages. The second line of argument is more *libertarian* in nature, where the implicit objective is to do "better" for all languages with respect to the SOTA; however, the the gap in performances across languages is acceptable as long as all languages are benefiting in some way. These arguments are agnostic to the reasons that might have caused the gap at the first place, and appear to suggest that one should use the most appropriate technology, given the amount of resources available. This is precisely where the third line of criticisms tend to disagree. These arguments suggest that instead of depending on MultiLMs, or the current state of technology and resources, one should try to take the necessary steps, that is to say, do the necessary *work*, to attain a more egalitarian state. Thus, these arguments conform to the "desert principle" of fairness, where only those performance gaps between languages are acceptable that are explainable in terms of the difference in the amount of *work* put in for the language, say, in building resources. Performance differences due to inherent features of a language, such as typological factors, are akin to *luck* and cannot be allowed (for instance, by choosing a typologically dissimilar language as a pivot). Hence, the proponents recommend creation of different MultiLMs for language groups (Lin et al. 2019) and creation of labeled data for fine-tuning (Lauscher et al. 2020), whenever possible.

It is important to note that the recommendations or critiques in a paper often span multiple ideas that implicitly conform to different normative principles of efficiency and fairness. Even when these normative principles are explicitly

stated, their application in practice is not always straightforward (Binns 2018; Leben 2020; Young 1994).

4 Discussion

Language as an Entity: The notion of *distributive justice* in ethics and economics starts with the assumption that the recipient of the resource or utility is an individual or a group of individuals. The discourse on ethics in machine learning has also followed this idea of group and individual fairness (Binns 2018). In contrast, the present discussion on MMSP considers a *language* as the recipient of distributive justice. Since a language can be equated to a group of individuals, it might be argued that the principles of group fairness are applicable here as well. However, there are important philosophical and practical distinctions between the notions of "language as an entity" and "language as a group of individuals", which we believe are crucial to appropriately contextualize the present analysis. One might choose to resolve the MMSP by imagining language as a group of individuals, let us say, the users of the technologies in that language, which would be a subset of the speakers of the language. Under this assumption, the consequentialist should not look at the accuracy of the MultiLMs on test-benches. Rather, the true utility of a MultiLM, and more broadly, an NLP system can be measured only in terms of its usefulness for the end-users (Soria, Quochi, and Russo 2018). For certain languages, a predictive keyboard or speech recognition system might be of much greater value than, say, an NLI system or POS tagger. Therefore, even though "language as a group of end-users" is a practically important and useful construct, it cannot be operationalized in this context because of the way MultiLMs are evaluated currently.

On the other hand, "language as an entity" would mean that each language has a distinct identity, defined by its vocabulary, syntactic structure, its typological features, amount of available resources, and so on. Under this notion, tasks such as parsing, POS or NLI, and limited evaluation on test-benches make perfect sense. A MultiLM or any similar multilingual system should then be evaluated similarly across languages and its utility could be quantified by the performance on a fixed set of linguistic tasks. It is this notion of "language as an entity" which drives the research and discourse of MultiLMs and more broadly, multilingual NLP, including this study on MMSP.

Fairness in NLP: In NLP, the fairness discourse has been around representational biases, such as those in the word-embeddings, and performance bias against users and user groups such as gender, age or regional varieties of a language; see (Blodgett et al. 2020) and references therein. Besides a few notable exceptions (Zhao et al. 2020; Sweeney and Najafian 2019), in all the cases the object of study is a representation or system for a single language. Most related to the present work are the discussions around whether algorithms or methods are really language agnostic, as is often claimed (Bender 2011), and the issue of determining the best transfer language for a given language in a crosslingual transfer setting (Lin et al. 2019; Ahmad et al. 2019); see Ruder (2020) for an excellent compilation of reasons and critiques of English-centric NLP. While these studies raise many important ethical concerns, mostly deontological in nature, no proposals have been put forward for formal resolution of these dilemmas under different assumptions of normative principles.

On one hand, these debates are similar to the MMSP, as they treat language as an entity and as the recipient of the distributive justice. On the other hand, there is an important technical difference that unlike the case of MultiLMs, here the model parameters are not shared between languages, and are optimized for one or a pair of languages at a time. Therefore, instead of resolving MMSP, one could always choose different algorithms or different pivot/transfer languages to optimize the performance of the system for a particular language. However, we would like to reemphasize that MMSP applies even when the language models do not share any parameters, as long as we can frame it as a black-box model selection problem, where multiple agencies (could be approaches, algorithms, or even companies) offer different sets of models serving multiple languages, and one needs to choose between the agencies.

5 Practical Recommendations

Stating the Fairness and Efficiency Assumptions: We recommend that all empirical studies involving MultiLM (and/or cases where common algorithms or approaches are applied to multiple languages) should clearly articulate the fairness and efficiency principles they are following, and the assumptions they are making while resolving the MMSP, drawing conclusions and/or providing usage recommendations.

Utilitarian Resolution to MMSP: In case the researcher has a strong reason to follow the utilitarian principle, they can resolve the MMSP in favor of the model that Pareto dominates all other models. When no model Pareto dominates, the model with the highest (weighted) average performance can be chosen. If there is a tie (defined as exact or sufficiently close average performances), they could invoke the prioritarian principle to resolve the tie in favor of the model that maximizes the minimum accuracy.

Prioritarian Resolution to MMSP: In case the researcher wishes to resolve the MMSP following the prioritarian or Rawlsian principle, they can use the max-min, or more generally, the *lex-min* objective, i.e., maximize the minimum accuracy, and if two models that maximize the minimum accuracy have the same minimum accuracy, then maximize the second minimum accuracy, and so on.

Considering Outliers: Under all circumstances, the principles should be applied only after critically analyzing the outlier languages that have remarkably high or low performances. These cases should be included or excluded before applying the fairness principles – a decision that can be made based on practical considerations such as how reliable is the test-set of a particular language.

Alternative Leader-boards: Current leader-boards for MultiLMs, e.g. https://sites.research.google/xtreme, rank the models based on the average performance. We recommend alternative leader-boards or rankings based on various fairness principles, including but not limited to the prioritarian or Rawlsian principle.

Acknowledgement

We would like to thank Mr. Anirudh Srinivasan for pointers to the recent literature on MultiLMs and their critiques.

Ethics Statement

This work raises some important ethical questions regarding the selection strategies for massively multilingual pretrained models. We believe that this work will help create awareness on linguistic fairness, diversity and inclusion of all and especially low resource languages. We do not foresee any adverse ethical implications of the current study.

References

Aguilar, G.; Kar, S.; and Solorio, T. 2020. LinCE: A Centralized Benchmark for Linguistic Code-switching Evaluation. In *Proceedings of The 12th Language Resources and Evaluation Conference*, 1803–1813.

Aharoni, R.; Johnson, M.; and Firat, O. 2019. Massively Multilingual Neural Machine Translation. 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)*, 3874–3884.

Ahmad, W. U.; Zhang, Z.; Ma, X.; Hovy, E. H.; Chang, K.-W.; and Peng, N. 2019. On Difficulties of Cross-Lingual Transfer with Order Differences: A Case Study on Dependency Parsing. In *NAACL-HLT (1)*.

Arivazhagan, N.; Bapna, A.; Firat, O.; Lepikhin, D.; Johnson, M.; Krikun, M.; Chen, M. X.; Cao, Y.; Foster, G.; Cherry, C.; et al. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges. *arXiv preprint arXiv:1907.05019* .

Artetxe, M.; Ruder, S.; and Yogatama, D. 2020. On the Crosslingual Transferability of Monolingual Representations. In *Proceedings of ACL*, 1946–1958.

Artetxe, M.; Ruder, S.; Yogatama, D.; Labaka, G.; and Agirre, E. 2020. A Call for More Rigor in Unsupervised Cross-lingual Learning. *arXiv preprint arXiv:2004.14958* .

Artetxe, M.; and Schwenk, H. 2019. Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. *Transactions of the Association for Computational Linguistics* 7(0).

Barocas, S.; Hardt, M.; and Narayanan, A. 2017. Fairness in machine learning. *NIPS Tutorial* 1.

Barocas, S.; Hardt, M.; and Narayanan, A. 2019. *Fairness and Machine Learning*. fairmlbook.org. http://www.fairmlbook.org.

Bender, E. M. 2011. On achieving and evaluating language-independence in NLP. *Linguistic Issues in Language Technology* 6(3): 1–26.

Binns, R. 2018. Fairness in Machine Learning: Lessons from Political Philosophy. 149–159. PMLR.

Blodgett, S. L.; Barocas, S.; Daumé III, H.; and Wallach, H. 2020. Language (Technology) is Power: A Critical Survey of" Bias" in NLP. *Proc of ACL* .

Celis, L. E.; Huang, L.; Keswani, V.; and Vishnoi, N. K. ???? Classification with Fairness Constraints: A Meta-Algorithm with Provable Guarantees. In *Proceedings of FAT* 19*, 319–328.

Clark, J. H.; Choi, E.; Collins, M.; Garrette, D.; Kwiatkowski, T.; Nikolaev, V.; and Palomaki, J. 2020. TyDi QA: A Benchmark for Information-Seeking Question Answering in Typologically Diverse Languages. *Transactions of the Association of Computational Linguistic* .

Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, E.; Ott, M.; Zettlemoyer, L.; and Stoyanov, V. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of ACL*.

Conneau, A.; and Lample, G. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, 7059–7069.

Conneau, A.; Rinott, R.; Lample, G.; Williams, A.; Bowman, S.; Schwenk, H.; and Stoyanov, V. 2018. XNLI: Evaluating Cross-lingual Sentence Representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2475–2485.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT (1)*.

Hammond, P. J. 1976. Equity, Arrow's Conditions, and Rawls' Difference Principle. *Econometrica* 44(4): 793–804.

Harsanyi, J. C. 1955. Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility. *Journal of Political Economy* 63(4): 309–321.

Heidari, H.; Loi, M.; Gummadi, K. P.; and Krause, A. 2019. A Moral Framework for Understanding Fair ML through Economic Models of Equality of Opportunity. In *Proceedings of FAT'19*, 181–190.

Hossain, S.; Mladenovic, A.; and Shah, N. 2020. Designing Fairly Fair Classifiers Via Economic Fairness Notions. In *Proceedings of The Web Conference 2020*, 1559–1569.

Hu, J.; Ruder, S.; Siddhant, A.; Neubig, G.; Firat, O.; and Johnson, M. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. *arXiv preprint arXiv:2003.11080* .

Huang, H.; Liang, Y.; Duan, N.; Gong, M.; Shou, L.; Jiang, D.; and Zhou, M. 2019. Unicoder: A Universal Language Encoder by Pre-training with Multiple Cross-lingual Tasks. In *Proceedings of EMNLP-IJCNLP*.

Joshi, P.; Santy, S.; Budhiraja, A.; Bali, K.; and Choudhury, M. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the ACL*.

Khanuja, S.; Dandapat, S.; Srinivasan, A.; Sitaram, S.; and Choudhury, M. 2020. GLUECoS: An Evaluation Benchmark for Code-Switched NLP. *arXiv preprint arXiv:2004.12376* .

Lauscher, A.; Ravishankar, V.; Vulić, I.; and Glavaš, G. 2020. From Zero to Hero: On the Limitations of Zero-Shot Cross-Lingual Transfer with Multilingual Transformers. *arXiv preprint arXiv:2005.00633* .

Leben, D. 2020. Normative Principles for Evaluating Fairness in Machine Learning. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, 86–92.

Lewis, P.; Oguz, B.; Rinott, R.; Riedel, S.; and Schwenk, H. 2019. MLQA: Evaluating Cross-lingual Extractive Question Answering. *ArXiv* abs/1910.07475.

Liang, Y.; Duan, N.; Gong, Y.; Wu, N.; Guo, F.; Qi, W.; et al. 2020. XGLUE: A new benchmark dataset for cross-lingual pre-training, understanding and generation. *arXiv preprint arXiv:2004.01401* .

Lin, Y.-H.; Chen, C.-Y.; Lee, J.; Li, Z.; Zhang, Y.; Xia, M.; Rijhwani, S.; He, J.; Zhang, Z.; Ma, X.; et al. 2019. Choosing transfer languages for cross-lingual learning. *arXiv preprint arXiv:1905.12688* .

Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; and Galstyan, A. 2019. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635* .

Pan, X.; Zhang, B.; May, J.; Nothman, J.; Knight, K.; and Ji, H. 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)*, 1946–1958.

Pires, T.; Schlinger, E.; and Garrette, D. 2019. How Multilingual is Multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4996–5001.

Rawls, J. 1999. *A Theory of Justice*. Harvard University Press. ISBN 9780674000773. URL http://www.jstor.org/stable/j.ctvkjb25m.

Ray, P. 1973. Independence of Irrelevant Alternatives. *Econometrica* 41(5): 987–991. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/1913820.

Rijhwani, S.; Xie, J.; Neubig, G.; and Carbonell, J. 2019. Zero-shot neural transfer for cross-lingual entity linking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 6924–6931.

Ruder, S. 2020. Why should we do NLP beyond English? URL https://ruder.io/nlp-beyond-english/.

Saleiro, P.; Kuester, B.; Stevens, A.; Anisfeld, A.; Hinkson, L.; London, J.; and Ghani, R. 2018. Aequitas: A Bias and Fairness Audit Toolkit. *arXiv preprint arXiv:1811.05577* .

Sang, E. F. T. K. 2002. Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition. *ArXiv* cs.CL/0209010.

Sen, A. 1985. Social Choice and Justice: A Review Article. *Journal of Economic Literature* 23(4): 1764–1776. ISSN 00220515. URL http://www.jstor.org/stable/2725708.

Sion, M. 1958. On general minimax theorems. *Pacific Journal of Mathematics* 8(1): 171–176.

Soria, C.; Quochi, V.; and Russo, I. 2018. The DLDP Survey on Digital Use and Usability of EU Regional and Minority Languages. In *Proceedings of LREC 2018*.

Strasnick, S. 1976. Social Choice and the Derivation of Rawls's Difference Principle. *The Journal of Philosophy* 73(4): 85–99. ISSN 0022362X. URL http://www.jstor.org/stable/2025509.

Sweeney, C.; and Najafian, M. 2019. A transparent framework for evaluating unintended demographic bias in word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1662–1667.

von Neumann, J.; Morgenstern, O.; and Rubinstein, A. 1944. *Theory of Games and Economic Behavior (60th Anniversary Commemorative Edition)*. Princeton University Press. ISBN 9780691130613. URL http://www.jstor.org/stable/j.ctt1r2gkx.

Wu, S.; and Dredze, M. 2020. Are All Languages Created Equal in Multilingual BERT? *arXiv preprint arXiv:2005.09093* .

Yaari, M. E. 1981. Rawls, Edgeworth, Shapley, Nash: Theories of Distributive Justice Re-examined. *Journal of Economic Theory* 24: 1–39.

Yang, Y.; Zhang, Y.; Tar, C.; and Baldridge, J. 2019. PAWS-X: A Cross-lingual Adversarial Dataset for Paraphrase Identification. *ArXiv* abs/1908.11828.

Young, H. P. 1994. *Equity: In Theory and Practice*. Princeton University Press. ISBN 9780691043197. URL http://www.jstor.org/stable/j.ctv10crfx7.

Zafar, M. B.; Valera, I.; Gomez-Rodriguez, M.; and Gummadi, K. P. 2019. Fairness Constraints: A Flexible Approach for Fair Classification. *Journal of Machine Learning Research* 20(75): 1–42. URL http://jmlr.org/papers/v20/18-262.html.

Zeman, D.; Nivre, J.; Abrams, M.; Aepli, N.; and al et. 2019. Universal Dependencies 2.5. URL http://hdl.handle.net/11234/1-3105. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Zhao, J.; Mukherjee, S.; Hosseini, S.; Chang, K.-W.; and Awadallah, A. H. 2020. Gender Bias in Multilingual Embeddings and Cross-Lingual Transfer. *arXiv preprint arXiv:2005.00699* .

Zweigenbaum, P.; Sharoff, S.; and Rapp, R. 2018. Overview of the Third BUCC shared task: Spotting parallel sentences in comparable corpora. In *Proceedings of 11th Workshop on Building and Using Comparable Corpora*, 39–42.