Machine Translation into Low-resource Language Varieties

Sachin Kumar[♣] **Antonios Anastasopoulos**[♢] **Shuly Wintner**[♡] **Yulia Tsvetkov**[♠]
[♣]Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, USA
[♢]Department of Computer Science, George Mason University, Fairfax, VA, USA
[♡]Department of Computer Science, University of Haifa, Haifa, Israel
[♠]Paul G. Allen School of Computer Science & Engineering, University of Washington
sachink@cs.cmu.edu, antonis@gmu.edu, shuly@cs.haifa.ac.il, yuliats@cs.washington.edu

Abstract

State-of-the-art machine translation (MT) systems are typically trained to generate "standard" target language; however, many languages have multiple varieties (regional varieties, dialects, sociolects, non-native varieties) that are different from the standard language. Such varieties are often low-resource, and hence do not benefit from contemporary NLP solutions, MT included. We propose a general framework to rapidly adapt MT systems to generate language varieties that are close to, but different from, the standard target language, using no parallel (source–variety) data. This also includes adaptation of MT systems to low-resource typologically-related target languages.[1] We experiment with adapting an English–Russian MT system to generate Ukrainian and Belarusian, an English–Norwegian Bokmål system to generate Nynorsk, and an English–Arabic system to generate four Arabic dialects, obtaining significant improvements over competitive baselines.

1 Introduction

Despite tremendous progress in machine translation (Bahdanau et al., 2015; Vaswani et al., 2017) and language generation in general, current state-of-the-art systems often work under the assumption that a language is homogeneously spoken and understood by its speakers: they generate a "standard" form of the target language, typically based on the availability of parallel data. But language use varies with regions, socio-economic backgrounds, ethnicity, and fluency, and many widely spoken languages consist of dozens of varieties or dialects, with differing lexical, morphological, and syntactic patterns for which no translation data are typically available. As a result, models trained to translate from a source language (SRC) to a standard language variety (STD) lead to a sub-par experience for speakers of other varieties.

Motivated by these issues, we focus on the task of adapting a trained SRC→STD translation model to generate text in a different target variety (TGT), having access only to limited monolingual corpora in TGT and no SRC–TGT parallel data. TGT may be a dialect of, a language variety of, or a typologically-related language to STD.

We present an effective transfer-learning framework for translation into low resource language varieties. Our method reuses SRC→STD MT models and finetunes them on synthesized (pseudo-parallel) SRC–TGT texts. This allows for rapid adaptation of MT models to new varieties without having to train everything from scratch. Using word-embedding adaptation techniques, we show that MT models which predict continuous word vectors (Kumar and Tsvetkov, 2019) rather than softmax probabilities lead to superior performance since they allow additional knowledge to be injected into the models through transfer between word embeddings of high-resource (STD) and low-resource (TGT) monolingual corpora.

We evaluate our framework on three translation tasks: English to Ukrainian and Belarusian, assuming parallel data are only available for English→Russian; English to Nynorsk, with only English to Norwegian Bokmål parallel data; and English to four Arabic dialects, with only English→Modern Standard Arabic (MSA) parallel data. Our approach outperforms competitive baselines based on unsupervised MT, and methods based on finetuning softmax-based models.

2 A Transfer-learning Architecture

We first formalize the task setup. We are given a parallel SRC→STD corpus, which allows us to

[1]Code, data and trained models are available here: https://github.com/Sachin19/seq2seq-con

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Short Papers), pages 110–121
August 1–6, 2021.

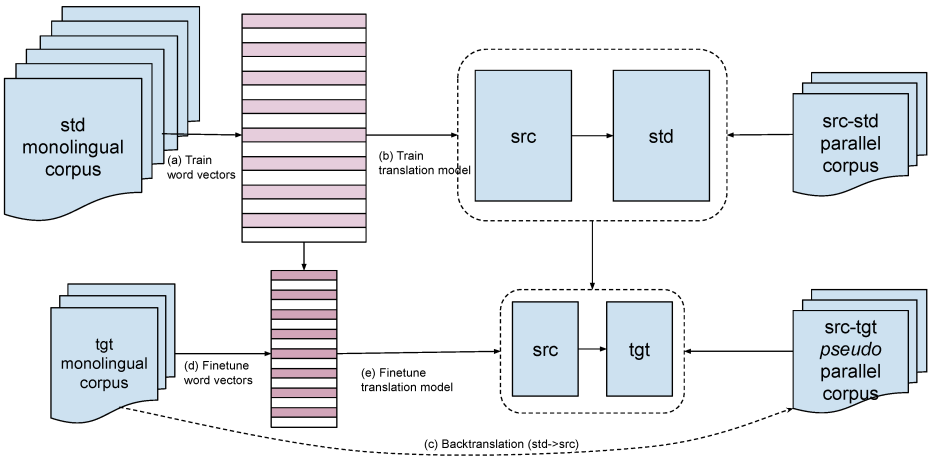

Figure 1: An overview of our approach. (a) Using the available STD monolingual corpora, we first train word vectors using `fasttext`; (b) we then train a SRC→STD translation model using the parallel corpora to predict the pretrained word vectors; (c) next, we train STD→SRC model and use it to translate TGT monolingual corpora to SRC; (d) now, we finetune STD subword embeddings to learn TGT word embeddings; and finally (e) we finetune a SRC→STD model to generate TGT pretrained embeddings using the back-translated SRC→TGT data.

train a translation model $f(\cdot;\theta)$ that takes an input sentence x in SRC and generates its translation in the standard veriety STD, $\hat{y}_{\text{STD}} = f(x;\theta)$. Here, θ are the learnable parameters of the model. We are also given monolingual corpora in both the standard STD and target variety TGT. Our goal now is to modify f to generate translations $\hat{y}_{\text{TGT}}$ in the target variety TGT. At training time, we assume no SRC–TGT or STD–TGT parallel data are available.

Our solution (Figure 1) is based on a transformer-based encoder-decoder architecture (Vaswani et al., 2017) which we modify to predict word vectors. Following Kumar and Tsvetkov (2019), instead of treating each token in the vocabulary as a discrete unit, we represent it using a unit-normalized d-dimensional pre-trained vector. These vectors are learned from a STD monolingual corpus using `fasttext` (Bojanowski et al., 2017). A word's representation is computed as the average of the vectors of its character n-grams, allowing surface-level linguistic information to be shared among words. At each step in the decoder, we feed this pretrained vector at the input and instead of predicting a probability distribution over the vocabulary using a softmax layer, we predict a d-dimensional continuous-valued vector. We train this model by minimizing the von Mises-Fisher (vMF) loss—a probabilistic variant of cosine distance—between the predicted vector and the pre-trained vector. The pre-trained vectors (at both input and output of the decoder) are not trained with the model. To decode from this model, at each step, the output word is generated by finding the closest neighbor (in terms of cosine similarity) of the predicted output vector in the pre-trained embedding table.

We train f in this fashion using SRC–STD parallel data. As shown below, training a softmax-based SRC→STD model to later finetune with TGT suffers from vocabulary mismatch between STD and TGT and thus is detrimental to downstream performance. By replacing the decoder input and output with pre-trained vectors, we separate the vocabulary from the MT model, making adaptation easier.

Now, to finetune this model to generate TGT, we need TGT embeddings. Since the TGT monolingual corpus is small, training `fasttext` vectors on this corpus from scratch will lead (as we show) to low-quality embeddings. Leveraging the relatedness of STD and TGT and their vocabulary overlap, we use STD embeddings to transfer knowledge to TGT embeddings: for each character n-gram in the TGT corpus, we initialize its embedding with the corresponding STD embedding, if available. We then continue training `fasttext` on the TGT monolingual corpus (Chaudhary et al., 2018). Last, we use a supervised embedding alignment method (Lample et al., 2018a) to project the learned TGT embeddings in the same space as STD. STD and TGT are expected to have a large lexical overlap, so we use identical tokens in both varieties as supervision for this alignment. The obtained embeddings, due to transfer learning from STD, inject additional knowledge in the model.

Finally, to obtain a SRC→TGT model, we finetune f on psuedo-parallel SRC–TGT data. Using a STD→SRC MT model (a back-translation model

trained using large STD–SRC parallel data with standard settings) we (back)-translate TGT data to SRC. Naturally, these synthetic parallel data will be noisy despite the similarity between STD and TGT, but we show that they improve the overall performance. We discuss the implications of this noise in §4.

3 Experimental Setup

Datasets We experiment with two setups. In the first (synthetic) setup, we use English (EN) as SRC, Russian (RU) as STD, and Ukrainian (UK) and Belarusian (BE) as TGTs. We sample 10M EN-RU sentences from the WMT'19 shared task (Ma et al., 2019), and 80M RU sentences from the CoNLL'17 shared task to train embeddings. To simulate low-resource scenarios, we sample 10K, 100K and 1M UK sentences from the CoNLL'17 shared task and BE sentences from the OSCAR corpus (Ortiz Suárez et al., 2020). We use TED dev/test sets for both languages pairs (Cettolo et al., 2012).

The second (real world) setup has two language sets: the first one defines English as SRC, with Modern Standard Arabic (MSA) as STD and four Arabic varieties spoken in Doha, Beirut, Rabat and Tunis as TGTs. We sample 10M EN-MSA sentences from the UNPC corpus (Ziemski et al., 2016), and 80M MSA sentences from the CoNLL'17 shared task. For Arabic varieties, we use the MADAR corpus (Bouamor et al., 2018) which consists of 12K 6-way parallel sentences between English, MSA and the 4 considered varieties. We ignore the English sentences, sample dev/test sets of 1K sentences each, and consider 10K monolingual sentences for each TGT variety. The second set also has English as SRC with Norwegian Bokmål (NO) as STD and its written variety Nynorsk (NN) as TGT. We use 630K EN-NO sentences from WikiMatrix (Schwenk et al., 2021), and 26M NO sentences from ParaCrawl (Esplà et al., 2019) combined with the WikiMatrix NO sentences to train embeddings. We use 310K NN sentences from WikiMatrix, and TED dev/test sets for both varieties (Reimers and Gurevych, 2020).

Preprocessing We preprocess raw text using Byte Pair Encoding (BPE, Sennrich et al., 2016) with 24K merge operations on each SRC–STD corpus trained separately on SRC and STD. We use the same BPE model to tokenize the monolingual STD data and learn `fasttext` embeddings (we consider character n-grams of length 3 to 6).[2] Splitting the TGT words with the same STD BPE model will result in heavy segmentation, especially when TGT contains characters not present in STD.[3] To counter this, we train a joint BPE model with 24K operations on the concatenation of STD and TGT corpora to tokenize TGT corpus following Chronopoulou et al. (2020). This technique increases the number of shared tokens between STD and TGT, thus enabling better cross-variety transfer while learning embeddings *and* while finetuning. We follow Chaudhary et al. (2018) to train embeddings on the generated TGT vocabulary where we initialize the character n-gram representations for TGT words with STD's `fasttext` model wherever available and finetune them on the TGT corpus.

Implementation and Evaluation We modify the standard OpenNMT-py seq2seq models of PyTorch (Klein et al., 2017) to train our model with vMF loss (Kumar and Tsvetkov, 2019). Additional hyperparameter details are outlined in Appendix B. We evaluate our methods using BLEU score (Papineni et al., 2002) based on the SacreBLEU implementation (Post, 2018).[4] For the Arabic varieties, we also report a macro-average. In addition, to measure the expected impact on actual systems' users, we follow Faisal et al. (2021) in computing a population-weighted macro-average ($\textbf{avg}_{\textbf{pop}}$) based on language community populations provided by Ethnologue (Eberhard et al., 2019).

3.1 Experiments

Our proposed framework, **LANGVARMT**, consists of three main components: (1) A supervised SRC→STD model is trained to predict continuous STD word embeddings rather than discrete softmax probabilities. (2) Output STD embeddings are replaced with TGT embeddings. The TGT embeddings are trained by finetuning STD embeddings on monolingual TGT data and aligning the two embedding spaces. (3) The resulting model is finetuned with pseudo-parallel SRC→TGT data.

We compare LANGVARMT with the following competitive baselines. **SUP(SRC→STD)**: train a standard (softmax-based) supervised SRC→STD model, and consider the output of this model as

[2] We slightly modify `fasttext` to not consider BPE token markers "@@" in the character n-grams.

[3] For example, both RU and UK alphabets consist of 33 letters; RU has the letters Ёё, ъ, ы and Ээ, which are not used in UK. Instead, UK has Ґґ, Єє, Ii and Її.

[4] While we recognize the limitations of BLEU (Mathur et al., 2020), more sophisticated embedding-based metrics for MT evaluation (Zhang et al., 2020; Sellam et al., 2020) are unfortunately not available for low-resource language varieties.

Size of TGT corpus	UK 10K	UK 100K	UK 1M	BE 10K	BE 100K	BE 1M	NN 300K	Arabic Varieties (10K) Doha	Beirut	Rabat	Tunis
SUP(SRC→STD)	1.7	1.7	1.7	1.5	1.5	1.5	11.3	3.7	1.8	2.0	1.3
UNSUP(SRC→TGT)	0.3	0.6	0.9	0.4	0.6	1.4	2.7	0.2	0.1	0.1	0.1
PIVOT	1.5	8.6	14.9	1.15	3.9	8.0	11.9	1.8	2.1	1.7	1.1
SOFTMAX	1.9	12.7	15.4	1.5	4.5	7.9	14.4	14.5	7.4	4.9	3.9
LANGVARMT	**6.1**	**13.5**	**15.3**	**2.3**	**8.8**	**9.8**	**16.6**	**20.1**	**8.1**	**7.4**	**4.6**

Table 1: BLEU scores on translation from English to Ukrainian, Belarusian, Nynorsk, and Arabic dialects with varying amounts of monolingual target data (TGT sentences) available for finetuning. Our approach (LANGVARMT) outperforms all baselines.

TGT under the assumption that STD and TGT may be very similar. **UNSUP(SRC→TGT)**: train an unsupervised MT model (Lample et al., 2018a) in which the encoder and decoder are initialized with cross-lingual masked language models (MLM, Conneau and Lample, 2019). These MLMs are pre-trained on SRC monolingual data, and then finetuned on TGT monolingual data with an expanded vocabulary as described above. This baseline is taken from Chronopoulou et al. (2020), where it showed state-of-the-art performance for low-monolingual-resource scenarios. **Pivot**: train a UNSUP(STD→TGT) model as described above using STD and TGT monolingual corpora. During inference, translate the SRC sentence to STD with the SUP(SRC→STD) model and then to TGT with the UNSUP(STD→TGT) model. We also perform several ablation experiments, showing that every component of LANGVARMT is necessary for good downstream performance. Specifically, we report results with LANGVARMT but using a standard softmax layer (**SOFTMAX**) to predict tokens instead of continuous vectors.[5]

4 Results and Analysis

Table 1 compares the performance of LANGVARMT with the baselines for Ukrainian, Belarusian, Nynorsk, and the four Arabic varieties. For reference, note that the EN→RU, EN→MSA, and EN→NO models are relatively strong, yielding BLEU scores of 24.3, 21.2, and 24.9, respectively.

Synthetic Setup Considering STD and TGT as the same language is sub-optimal, as is evident from the poor performance of the non-adapted SUP(SRC→STD) model. Clearly, special attention ought to be paid to language varieties. Direct unsupervised translation from SRC to TGT performs poorly as well, confirming previously reported results of the ineffectiveness of such methods on unrelated languages (Guzmán et al., 2019).

[5]Additional ablation results are listed in Appendix C.

Translating SRC to TGT by pivoting through STD achieves much better performance owing to strong UNSUP(STD→TGT) models that leverage the similarities between STD and TGT. However, when resources are scarse (e.g., with 10K monolingual sentences as opposed to 1M), this performance gain considerably diminishes. We attribute this drop to overfitting during the pre-training phase on the small TGT monolingual data. Ablation results (Appendix C) also show that in such low-resource settings the learned embeddings are of low quality.

Finally, LANGVARMT consistently outperforms all baselines. Using 1M UK sentences, it achieves similar performance (for EN→UK) to the softmax ablation of our method, SOFTMAX, and small gains over unsupervised methods. However, in lower resource settings our approach is clearly better than the strongest baselines by over 4 BLEU points for UK (10K) and 3.9 points for BE (100K).

To identify potential sources of error in our proposed method, we lemmatize the generated translations and test sets and evaluate BLEU (Qi et al., 2020). Across all data sizes, both UK and BE achieve a substantial increase in BLEU (up to +6 BLEU; see Appendix D for details) compared to that obtained on raw text, indicating morphological errors in the translations. In future work, we will investigate whether we can alleviate this issue by considering TGT embeddings based on morphological features of tokens (Chaudhary et al., 2018).

Real-world Setup The effectiveness of LANGVARMT is pronounced in this setup with a dramatic improvement of more than 18 BLEU points over unsupervised baselines when translating into Doha Arabic. We hypothesize that during the pretraining phase of unsupervised methods, the extreme difference between the size of the MSA monolingual corpus (10M) and the varieties' corpora (10K) leads to overfitting. Additionally, compared to the synthetic setup, the Arabic varieties we consider are quite close to MSA, allowing for easy and effective adaptation of both word embeddings and

EN→MSA models. LANGVARMT also improves in all other Arabic varieties, although naturally some varieties remain challenging. For example, the Rabat and particularly the Tunis varieties are more likely to include French loanwords (Bouamor et al., 2018) which are not adequately handled as they are not part of our vocabulary. In future work, we will investigate whether we can alleviate this issue by potentially including French corpora (transliterated into Arabic) to our TGT language corpora. On average, our approach improves by 2.3 BLEU points over the softmax-based baseline (cf. 7.7 and 10.0 in Table 2 under $\mathbf{avg}_{\mathcal{L}}$) across the four Arabic dialects. For a population-weighted average ($\mathbf{avg}_{\mathbf{pop}}$), we associate the Doha variety with Gulf Arabic (ISO code: `afb`), the Beirut one with North Levantine Arabic (`apc`), Rabat with Moroccan (`ary`), and the Tunis variety with Tunisian Arabic (`aeb`). As before, LANGVARMT outperforms the baselines. The absolute BLEU scores in this highly challenging setup are admittedly low, but as we discuss in Appendix D, the translations generated by LANGVARMT are often fluent and input preserving, especially compared to the baselines.

Finally, due to high similarity between NO and NN, the SUP(EN→NO) model also performs well on NN with 11.3 BLEU, but our method yields further gains of over 4 points over the baselines.

5 Discussion

Fairness The goal of this work is to develop more equitable technologies, usable by speakers of diverse language varieties. Here, we evaluate the systems along the principles of *fairness*. We evaluate the fairness of our Arabic multi-dialect system's utility proportionally to the populations speaking those dialects. In particular, we seek to measure how much average benefit will the people of different dialects receive if their respective translation performance is improved. A simple proxy for fairness is the standard deviation (or, even simpler, a $\max - \min$ performance) of the BLEU scores across dialects (A higher value implies more unfairness across the dialects) Beyond that, we measure a system's *unfairness* with respect to the different dialect subgroups, using the adaptation of generalized entropy index (Speicher et al., 2018), which considers equities within and between subgroups in evaluating the overall unfairness of an algorithm on a population Faisal et al. (2021) (See Appendix F for details and additional discussion).

Table 2 shows that our proposed method is fairer across all dialects, compared to baselines where only MSA translation produces comprehensible outputs.

Model	$\mathbf{avg}_{\mathcal{L}}\uparrow$	$\mathbf{avg}_{\mathbf{pop}}\uparrow$	max−min↓	unfair↓
SUP(SRC→STD)	2.2	1.8	19.9	0.037
UNSUP(SRC→TGT)	0.1	0.1	21.1	0.046
PIVOT	1.7	1.8	20.1	0.037
SOFTMAX	7.7	5.7	17.3	0.020
LANGVARMT	**10.0**	**7.3**	**16.6**	**0.016**

Table 2: Average performance and fairness metrics across the four Arabic varieties. This evaluation includes MSA (with a BLEU score of 21.2 on the SUP(EN→MSA) model).

Negative Results Our proposed method relies on two components: (1) quality of TGT word embeddings which is dependent on STD and TGT shared (subword) vocabulary, and (2) the psuedo-parallel SRC–TGT obtained by back-translating TGT data through a STD→SRC model. If STD and TGT are not sufficiently closely related, the quality of both of these components can degrade, leading to a drop in the performance of our proposed method. We present results of two additional experiments to elucidate this phenomenon in Appendix E.

Related Work We provide an extensive discussion of related work in Appendix A.

6 Conclusion

We presented a transfer-learning framework for rapid and effective adaptation of MT models to different varieties of the target language without access to any source-to-variety parallel data. We demonstrated significant gains in BLEU scores across several language pairs, especially in highly resource-scarce scenarios. The improvements are mainly due to the benefits of continuous-output models over softmax-based generation. Our analysis highlights the importance of addressing morphological differences between language varieties, which will be in the focus of our future work.

Acknowledgements

This research was supported by Grants No. 2017699 and 2019785 from the United States-Israel Binational Science Foundation (BSF), by the National Science Foundation (NSF) under Grants No. 2040926 and 2007960, and by a Google faculty research award. We thank Safaa Shehadi for evaluating our model outputs, Xinyi Wang and Aditi Choudhary for helpful discussions, and the anonymous reviewers for much appreciated feedback.

References

Željko Agić and Ivan Vulić. 2019. JW300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3204–3210, Florence, Italy. Association for Computational Linguistics.

Kemal Altintas and Ilyas Cicekli. 2002. A machine translation system between a pair of closely related languages. In *In Seventeenth International Symposium On Computer and Information Sciences*.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised neural machine translation. In *International Conference on Learning Representations*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, and Kemal Oflazer. 2018. The MADAR Arabic dialect corpus and lexicon. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. WIT3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Annual conference of the European Association for Machine Translation*, pages 261–268, Trento, Italy. European Association for Machine Translation.

Aditi Chaudhary, Chunting Zhou, Lori Levin, Graham Neubig, David R. Mortensen, and Jaime Carbonell. 2018. Adapting word embeddings to new languages with morphological and phonological subword representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3285–3295, Brussels, Belgium. Association for Computational Linguistics.

Monojit Choudhury and Amit Deshpande. 2021. How linguistically fair are multilingual pre-trained language models? In *Proceedings of the AAAI Conference on Artificial Intelligence*, Online. AAAI.

Alexandra Chronopoulou, Dario Stojanovski, and Alexander Fraser. 2020. Reusing a Pretrained Language Model on Languages with Limited Corpora for Unsupervised NMT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2703–2711, Online. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, volume 32, pages 7059–7069. Curran Associates, Inc.

Anna Currey, Antonio Valerio Miceli Barone, and Kenneth Heafield. 2017. Copied monolingual data improves low-resource neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, pages 148–156, Copenhagen, Denmark. Association for Computational Linguistics.

David M Eberhard, Gary F Simons, and Charles D. (eds.) Fennig. 2019. Ethnologue: Languages of the world. 2019. *online. Dallas, Texas: SIL International.*

Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzmán, and Philipp Koehn. 2020. CCAligned: A massive collection of cross-lingual web-document pairs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020)*, pages 5960–5969, Online. Association for Computational Linguistics.

Miquel Esplà, Mikel Forcada, Gema Ramírez-Sánchez, and Hieu Hoang. 2019. ParaCrawl: Web-scale parallel corpora for the languages of the EU. In *Proceedings of Machine Translation Summit XVII Volume 2: Translator, Project and User Tracks*, pages 118–119, Dublin, Ireland. European Association for Machine Translation.

Fahim Faisal, Sharlina Keshava, Md Mahfuz ibn Alam, and Antonios Anastasopoulos. 2021. SD-QA: Spoken Dialectal Question Answering for the Real World. Preprint.

Xavier Garcia, Pierre Foret, Thibault Sellam, and Ankur Parikh. 2020. A multilingual view of unsupervised machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3160–3170, Online. Association for Computational Linguistics.

Xavier Garcia, Aditya Siddhant, Orhan Firat, and Ankur Parikh. 2021. Harnessing multilinguality in unsupervised machine translation for rare languages. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1126–1137, Online. Association for Computational Linguistics.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*

and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 6098–6111, Hong Kong, China. Association for Computational Linguistics.

Junxian He, Xinyi Wang, Graham Neubig, and Taylor Berg-Kirkpatrick. 2020. A probabilistic formulation of unsupervised text style transfer. In *International Conference on Learning Representations*.

Hieu Hoang and Philipp Koehn. 2008. Design of the Moses decoder for statistical machine translation. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pages 58–65, Columbus, Ohio. Association for Computational Linguistics.

Junjie Hu, Mengzhou Xia, Graham Neubig, and Jaime Carbonell. 2019. Domain adaptation of neural machine translation by lexicon induction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2989–3001, Florence, Italy. Association for Computational Linguistics.

Lidia Kidane, Sachin Kumar, and Yulia Tsvetkov. 2021. An exploration of data augmentation techniques for improving English to Tigrinya translation. In *Proceedings of the Second AfricaNLP Workshop*.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

Sachin Kumar and Yulia Tsvetkov. 2019. Von mises-fisher loss for training sequence to sequence models with continuous outputs. In *International Conference on Learning Representations*.

Surafel Melaku Lakew, Aliia Erofeeva, and Marcello Federico. 2018. Neural machine translation into language varieties. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 156–164, Brussels, Belgium. Association for Computational Linguistics.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018a. Unsupervised machine translation using monolingual corpora only. In *International Conference on Learning Representations*.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018b. Word translation without parallel data. In *International Conference on Learning Representations*.

Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. 2020. On the variance of the adaptive learning rate and beyond. In *International Conference on Learning Representations*.

Qingsong Ma, Johnny Wei, Ondřej Bojar, and Yvette Graham. 2019. Results of the WMT19 metrics shared task: Segment-level and strong MT systems pose big challenges. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 62–90, Florence, Italy. Association for Computational Linguistics.

Kelly Marchisio, Kevin Duh, and Philipp Koehn. 2020. When does unsupervised machine translation work? In *Proceedings of the Fifth Conference on Machine Translation*, pages 571–583, Online. Association for Computational Linguistics.

Luis Marujo, Nuno Grazina, Tiago Luis, Wang Ling, Luisa Coheur, and Isabel Trancoso. 2011. BP2EP - adaptation of Brazilian Portuguese texts to European Portuguese. In *Proceedings of the 15th Annual conference of the European Association for Machine Translation*, Leuven, Belgium. European Association for Machine Translation.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997, Online. Association for Computational Linguistics.

Preslav Nakov and Jörg Tiedemann. 2012. Combining word-level and character-level models for machine translation between closely-related languages. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 301–305, Jeju Island, Korea. Association for Computational Linguistics.

Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. 2020. A monolingual approach to contextualized word embeddings for mid-resource languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1703–1714, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Nima Pourdamghani and Kevin Knight. 2017. Deciphering related languages. In *Proceedings of the*

2017 Conference on Empirical Methods in Natural Language Processing, pages 2513–2518, Copenhagen, Denmark. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

John Rawls. 1999. *A Theory of Justice*. Harvard University Press.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021. WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361, Online. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

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.

Till Speicher, Hoda Heidari, Nina Grgic-Hlaca, Krishna P Gummadi, Adish Singla, Adrian Weller, and Muhammad Bilal Zafar. 2018. A unified approach to quantifying algorithmic unfairness: Measuring individual & group unfairness via inequality indices. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2239–2248.

T. Tan, S. Goh, and Y. Khaw. 2012. A malay dialect translation and synthesis system: Proposal and preliminary system. In *2012 International Conference on Asian Language Processing*, pages 109–112.

Jörg Tiedemann. 2009. Character-based PSMT for closely related languages. In *Proceedings of the 13th Annual conference of the European Association for Machine Translation*, Barcelona, Spain. European Association for Machine Translation.

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, pages 5998–6008. Curran Associates, Inc.

David Vilar, Jan-T. Peter, and Hermann Ney. 2007. Can we translate letters? In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, page 33–39, USA. Association for Computational Linguistics.

Zichao Yang, Zhiting Hu, Chris Dyer, Eric P Xing, and Taylor Berg-Kirkpatrick. 2018. Unsupervised text style transfer using language models as discriminators. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTscore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The United Nations parallel corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3530–3534, Portorož, Slovenia. European Language Resources Association (ELRA).

A Related Work

Early work addressing translation involving language varieties includes rule-based transformations (Altintas and Cicekli, 2002; Marujo et al., 2011; Tan et al., 2012) which rely on language specific information and expert knowledge which can be expensive and difficult to scale. Recent work to address this issue only focuses on cases where parallel data do exist. They include a combination of word-level and character-level MT (Vilar et al., 2007; Tiedemann, 2009; Nakov and Tiedemann, 2012) between related languages or training multilingual models to translate to/from English to different varieties of a language (e.g., Lakew et al. (2018) work on Brazilian–European Portuguese and European–Canadian French). Such parallel data, however, are typically unavailable for most language varieties.

Unsupervised translation models, which require only monolingual data, can address this limitation (Artetxe et al., 2018; Lample et al., 2018a; Garcia et al., 2020, 2021). However, when even *mono*lingual corpora are limited, unsupervised models are challenging to train and are quite ineffective for translating between unrelated languages (Marchisio et al., 2020). Considering varieties of a language as writing styles, unsupervised style transfer (Yang et al., 2018; He et al., 2020) or deciphering methods (Pourdamghani and Knight, 2017) to translate between different varieties have also been been explored but have not been shown to perform well, often only reporting BLEU-1 scores since they obtain BLEU-4 scores which are closer to 0. Additionally, all of these approaches require simultaneous access to data in all varieties during training and must be trained from scratch when a new variety is added. In contrast, our presented method allows for easy adaptation of SRC→STD models to any new variety as it arrives.

Considering a new target variety as a new domain of STD, unsupervised domain adaptation methods can be employed, such as finetuning SRC→STD models using pseudo-parallel corpora generated from monolingual corpora in target varieties (Hu et al., 2019; Currey et al., 2017). Our proposed method is most related to this approach; but while these methods have the potential to adapt the decoder language model, for effective transfer, STD and TGT must have a shared vocabulary which is not true for most language varieties due to lexical, morphological, and at times orthographic differences. In contrast, our proposed method makes use of cross-variety word embeddings. While our examples only involve same-script varieties, augmenting our approach to work across scripts through a transliteration component is straightforward.

B Implementation Details

We modify the standard OpenNMT-py seq2seq models of PyTorch (Klein et al., 2017) to train our model with vMF loss (Kumar and Tsvetkov, 2019). We use the transformer-BASE model (Vaswani et al., 2017), with 6 layers in both encoder and decoder and with 8 attention heads, as our underlying architecture. We modify this model to predict pretrained `fasttext` vectors. We also initialize the decoder input embedding table with the pretrained vectors and do not update them during model training. All models are optimized using Rectified Adam (Liu et al., 2020) with a batch size of 4K tokens and dropout of 0.1. We train SRC→STD models for 350K steps with an initial learning rate of 0.0007 with linear decay. For finetuning, we reduce the learning rate to 0.0001 and train for up to 100K steps. We use early stopping in all models based on validation loss computed every 2K steps. We decode all the softmax-based models with a beam size of 5 and all the vMF-based models greedily.

We evaluate our methods using BLEU score (Papineni et al., 2002) based on the SacreBLEU implementation (Post, 2018). While we recognize the limitations of BLEU (Mathur et al., 2020), more sophisticated embedding-based metrics for MT evaluation (Zhang et al., 2020; Sellam et al., 2020) are simply not available for language varieties.

C Additional English-Ukrainian Experiments

On our resource-richest setup of EN→UK translation using 1M UK sentences and RU as STD, we compare our method with the following additional baselines. Table 3 presents these results.

LAMPLE-UNSUP(SRC→TGT): This is another unsupervised model, based on Lample et al. (2018a) which initializes the input and output embedding tables of both encoder and decoder using cross-lingual word embeddings trained on SRC and TGT monolingual corpora. The model is trained in a similar manner to Chronopoulou et al. (2020) (UNSUP(SRC→TGT)) with iterative backtranslation and autoencoding.

PIVOT:LAMPLE(STD→TGT): This baseline is

Method	BLEU (uk)
SUP(SRC-STD)	1.7
UNSUP(SRC→TGT)	0.9
PIVOT:	14.9
LAMPLE-UNSUP(SRC→TGT)	0.4
PIVOT:LAMPLE-UNSUP(STD→TGT)	9.0
PIVOT:DICTREPLACE(STD→TGT)	2.9
LANGVARMT	15.3
LANGVARMT w/ poor embeddings	4.6
LANGVARMT-RANDOM	13.1
SOFTMAX	15.4
LANGVARMT-RANDOM-SOFTMAX	14.1

Table 3: BLEU scores on EN-UK test corpus with 1M UK monolingual corpus.

similar to the PIVOT baseline, where we replace the unsupervised model with that of Lample et al. (2018a).

PIVOT:DICTREPLACE(STD→TGT): Here we first translate SRC to STD using SUP(SRC→STD), and then modify the STD output to get a TGT sentence as follows: We create a STD–TGT dictionary using the embedding map suggested by Lample et al. (2018b). This dictionary is created on words tokenized with Moses tokenizer (Hoang and Koehn, 2008) rather than BPE tokens. We replace each token in the generated STD sentence which is not in the TGT vocabulary using the dictionary (if available). We consider this baseline to measure lexical vs. syntactic/phrase level differences between Russian and Ukrainian.

In addition to baseline comparison, we report the following ablation experiments.

(1) To measure transfer from STD to TGT embeddings, we finetune the SUP(SRC→STD) model using TGT embeddings trained from scratch (as opposed to initialized with STD embeddings).

(2) To measure the impact of initialization during model finetuning, we compare with a randomly initialized model trained in a supervised fashion on the psuedo-parallel SRC–TGT data.

Baselines On the unsupervised models based on Lample et al. (2018a), we observe a similar trend as that of Chronopoulou et al. (2020), where the LAMPLE-UNSUP(SRC→TGT) model performing poorly (0.4) with substantial gains when pivoting through Russian (9.0 BLEU).

PIVOT:DICTREPLACE(STD→TGT) gains some improvement over considering the output of SUP(SRC→STD) as TGT, probably due to syntactic similarities between Russian and Ukrainian. This result can potentially be further improved with a human-curated RU–UK dictionary, but such resources are typically not available for the low-resource settings we consider in this paper.

Ablations As shown in Table 3, training the SRC→TGT model on a randomly initialized model (LANGVAR-RANDOM) results in a performance drop, confirming that transfer learning from a SRC→STD model is beneficial. Similarly, using TGT embeddings trained from scratch (LANGVARMT w/ poor embeddings) results in a drastic performance drop, providing evidence for essential transfer from STD embeddings.

D Analysis

To better understand the performance of our models, we perform additional analyses.

Lemmatized BLEU For UK and BE, we lemmatize each word in the test sets and the translations and evaluate BLEU scores. The results, depicted in Table 4, very likely indicate that our framework often generates correct lemmas, but may fail on the correct inflectional form of the target words. This highlights the importance of considering morphological differences between language varieties. The high BLEU scores also demonstrate that the resulting translations are quite likely understandable, albeit not always grammatical.

	EN→UK			EN→BE		
	10K	100K	1M	10K	100K	1M
raw	6.1	13.5	15.3	2.3	8.8	9.8
lemma	12.8	19.5	21.3	3.5	13.7	15.8

Table 4: BLEU scores on raw vs lemmatized text with LANGVARMT.

Translation of Rare Words On the outputs of the EN→UK model, trained with 100K UK sentences, we compute the translation accuracy of words based on their frequency in the TGT monolingual corpus for LANGVARMT, our best baseline SUP(SRC→STD)+UNSUP(SRC→TGT) and the best performing ablation SOFTMAX. These results, shown in Table 5, reveal that LANGVARMT is more accurate at translating rare words (with frequency less than 10) compared to the baselines.

Examples We provide some examples of EN-UK and EN-Beirut Arabic translations generated by the three models in Tables 6 and 7. As evaluated by native speakers of the Beirut Arabic, we find that

frequency	PIVOT	SOFTMAX	LANGVARMT
1	0.0429	0.1516	0.1812
2	0.0448	0.2292	0.2556
3	0.0597	0.2246	0.2076
4	0.0692	0.2601	0.2962
[5,10)	0.0582	0.2457	0.2722
[10,100)	0.1194	0.2881	0.2827
[100,1000)	0.2712	0.4537	0.4449

Table 5: Translation accuracies of words based on their frequencies on EN→UK with 100K UK sentences.

despite a BLEU score of only 8, in a majority of cases our baseline model is able to generate fluent translations of the input, preserving most of the content, whereas the baseline model ignores many of the content words. We also observe that in some cases, despite predicting in the right semantic space of the pretrained embeddings, it fails to predict the right token, resulting in surface form errors (e.g., predicting adjectival forms of verbs). This phenomenon is known and studied in more detail in Kumar and Tsvetkov (2019).

E Negative Results

We present results for the following experiments: (a) adapting an English to Thai (EN→TH) model to Lao (LO). We use a parallel corpus of around 10M sentences for training the supervised EN→TH model from the CCAligned corpus (El-Kishky et al., 2020), around 140K LO monolingual sentences from the OSCAR corpus (Ortiz Suárez et al., 2020) and TED2020 dev/tests for both TH and LO[6] (Reimers and Gurevych, 2020). (b) adapting an English to Amharic Model (EN→AM) to Tigrinya (TI). We use training, development and test sets from the JW300 corpus (Agić and Vulić, 2019) containing 500K EN–AM parallel corpus and 100K Tigrinya monolingual sentences.

As summarized in Table 8, our method fails to perform well on these sets of languages. Although Thai and Lao are very closely related languages, we attribute this result to little subword overlap in their respective vocabularies which degrade the quality of the embeddings. This is because Lao's writing system is developed phonetically whereas Thai writing contains many silent characters. Considering shared phonetic information while learning the embeddings can alleviate this issue and is an avenue for future work. On the other hand, Amharic and Tigrinya, while sharing a decent amount of vocabulary, use different constructs and function words (Kidane et al., 2021) leading to a very noisy psuedo-parallel corpus.

[6]Although Thai and Lao scripts look very similar, they use different Unicode symbols which are one-to-one mappable to each other: https://en.wikipedia.org/wiki/Lao_(Unicode_block)

Source	And we never think about the hidden connection
Reference	Та ми ніколи не думаємо про приховані зв'язки
PIVOT	І ми ніколи не дуємо про приховану зв'язку. (And we never think about a hidden connection.)
SOFTMAX	Я ніколи не думав про прихований зв'язок. (I never thought of a hidden connection.)
LANGVARMT	І ми ніколи не думаємо про прихований зв'язок. (And we never think about a hidden connection.)
Source	And yet, looking at them, you would see a machine and a molecule.
Reference	Дивлячись на них, ви побачите машину і молекулу.
PIVOT	І бачити, дивлячись на них, ви бачите машину і молекулу молекули. (And to see, looking at them, you see a machine and a molecule of a molecule.)
SOFTMAX	І так, дивлячись на них, ви бачите машину і молекулу. (And so, looking at them, you see a machine and a molecule.)
LANGVARMT	І дивляючись на них, ви побачите машину і молекулу. (And looking at them, you will see a machine and a molecule)
Source	They have exactly the same amount of carbon.
Reference	Вони мають однакову кількість вуглецю.
PIVOT	Таким чином, їх частка вуглецю. (Thus, their share of carbon.)
SOFTMAX	Вони мають однакову кількість вуглецю. (They have the same amount of carbon.)
LANGVARMT	Вони мають точно таку ж кількість вуглецю. (they have exactly the same amount of carbon)

Table 6: Examples of EN-UK translations generated by LANGVARMT and the best performing baselines.

F Measuring Unfairness

When evaluating multilingual and multi-dialect systems, it is crucial that the evaluation takes into account principles of fairness, as outlined in economics and social choice theory (Choudhury and Deshpande, 2021). We follow the least difference

Source	I've never heard of this address near here.
Reference	اه يف ناونعلاهب تعمس طق ام لبق نم ةقطنملا.
PIVOT	كملسيب حر. (He will hand over.)
SOFTMAX	ينه ناونعلاه نع تعمس ةرم لاو. (Not once did I hear this title here) نوه نم بيرق ناونعه نم اًدبأ تعمس ام. (I've never heard from this address near here.)
Source	What's the exchange rate today?
Reference	؟مويلا رعسلا ونش
PIVOT	؟مويلا رعس (What's the rate?)
SOFTMAX	؟مويلا فرصلا رعس ونش (What's the exchange rate today?) ؟مويلا فرصلا رعس و ش (What's the exchange rate today?)
Source	How do I get to that place?
Reference	؟حرطملاهل لصوب فيك
PIVOT	؟حصنتب فيك (How do you recommend?)
SOFTMAX	؟لحملاع لصوأ يف فيك (How can I get to the shop?) ؟لصو يف فيك (How can I get there?)
Source	Tell me when we get to the museum.
Reference	فحتملاع لصون سب يلق .
PIVOT	يناتلاع حورن حر (we will go to the other.)
SOFTMAX	فحتملاع لصون ىتميا يكحا. (Talk when we get to the museum) فحتملل انلصو ىتميا يلق. (Tell me when we got to the museum)
Source	Please take me to the morning market.
Reference	حبصلا قوس ىلع يندخ فورعم لومع.
PIVOT	ينرطن حر. (We'll wait)
SOFTMAX	حبصلا قوسلاع يندخاتنم. (You take us to the market this morning.) حبصلا قوسلاع يندخات لضفنم. (We prefer you take us to the market at the morning.)

Table 7: Examples of English to Beirut Arabic translations generated by LANGVARMT and the best performing baselines.

	EN→LO	EN→TI
SRC→STD	0.7	1.8
SOFTMAX	1.4	2.9
LANGVARMT	4.5	3.8

Table 8: BLEU scores for English to Lao and English to Tigrinya translation

principle proposed by Rawls (1999), whose egalitarian approach proposes to narrow the gap between unequal accuracies.

A simple proxy for unfairness is the standard deviation (or, even simpler, a $\max - \min$ performance) of the scores across languages. Beyond that, we measure a system's *unfairness* with respect to the different subgroups using the adaptation of generalized entropy index described by Speicher et al. (2018), which considers equities within and between subgroups in evaluating the overall unfairness of an algorithm on a population. The generalized entropy index for a population of n individuals receiving benefits $b_1, b_2, \ldots, b_n$ with mean benefit μ is

$$\mathcal{E}^{\alpha}(b_1, \ldots, b_n) = \frac{1}{n\alpha(\alpha - 1)} \sum_{i=1}^{n} \left[\left(\frac{b_i}{\mu} \right)^{\alpha} - 1 \right].$$

Using $\alpha = 2$ following Speicher et al. (2018), the generalized entropy index corresponds to half the squared coefficient of variation.[7]

If the underlying population can be split into $|G|$ disjoint subgroups across some attribute (e.g. gender, age, or language variety) we can decompose the total unfairness into individual and group-level unfairness. Each subgroup $g \in G$ will correspond to n_g individuals with corresponding benefit vector $\mathbf{b}^g = (b_1^g, b_2^g, \ldots, b_{n_g}^g)$ and mean benefit μ_g. Then, total generalized entropy can be re-written as:

$$\mathcal{E}^{\alpha}(b_1, \ldots, b_n) = \sum_{g=1}^{|G|} \frac{n_g}{n} \left(\frac{\mu_g}{\mu} \right)^{\alpha} \mathcal{E}^{\alpha}(\mathbf{b}^g)$$
$$+ \sum_{g=1}^{|G|} \frac{n_g}{n\alpha(\alpha - 1)} \left[\left(\frac{\mu_g}{\mu} \right)^{\alpha} - 1 \right]$$
$$= \mathcal{E}^{\alpha}(\mathbf{b}) + \mathcal{E}^{\alpha}_{\beta}(\mathbf{b}).$$

The first term $\mathcal{E}^{\alpha}(\mathbf{b})$ corresponds to the weighted unfairness score that is observed *within* each subgroup, while the second term $\mathcal{E}^{\alpha}_{\beta}(\mathbf{b})$ corresponds to the unfairness score *across* different subgroups.

In this measure of unfairness, we define the benefit as being directly proportional to the system's accuracy. For a Machine Translation system, each user receives an average benefit equal to the BLEU score the MT system achieves on the user's dialect. Conceptually, if the system produces a perfect translation (BLEU=1) then the user will receive the highest benefit of 1. If the system fails to produce a meaningful translation (BLEU→ 0) then the user receives no benefit ($b = 0$) from the interaction with the system.

[7] The coefficient of variation is simply the ratio of the standard deviation σ to the mean μ of a distribution.