Data-adaptive Transfer Learning for Low-resource Translation: A Case Study in Haitian

Anonymous ACL submission

Abstract

Multilingual transfer techniques often improve low-resource machine translation (MT). Many of these techniques are applied without considering data characteristics. We show in the context of Haitian-to-English translation that transfer effectiveness is correlated with amount of training data and relationships between knowledge-sharing languages. Our experiments suggest that beyond a threshold of authentic data, back-translation augmentation methods are counterproductive, while cross-lingual transfer during training is preferred. We complement this finding by contributing a rule-based French-Haitian orthographic and syntactic engine and a novel method for phonological embedding. When used with multilingual techniques, orthographic transformation significantly improves performance over conventional methods, and phonological transfer greatly improves performance in Jamaican MT.

1 Introduction and Motivation

Machine translation (MT) for low resource languages (LRL) requires special attention due to data scarcity. Often LRL MT is aided by knowledge transfer from languages with more abundant resources (Tars et al., 2021; Neubig and Hu, 2018; Zoph et al., 2016). In this work we report a case study showing that transfer techniques based on back-translation can improve poor scores in very low-resource settings but be counterproductive once a threshold of authentic data is reached.

We show that beyond this threshold, multi-source MT methods are more effective (Zoph et al., 2016). In these settings, MT systems map from a small amount of data in a LRL and a larger amount of data in a related high resource language (HRL) to a target language (TGT), in order to improve LRL-to-TGT translation quality. (See §2.) In addition to applying these methods conventionally, we present novel techniques for harnessing syntactic, orthographic, and phonological similarities between source languages. Prior to training, we transform HRL data to resemble LRL orthography and syntax by harnessing morphological and syntactic relationships between related languages. For phonologically similar languages, we present novel phonological word embeddings via PanPhon (Mortensen et al., 2016) and use these to initilize MT models.

We conduct these experiments in a case study of Haitian-to-English MT. We also contribute a rule-based French-Haitian (FRA-HAT) orthographic and syntactic engine that transforms French to Haitian text with 59.5% character error rate (CER) and 1.60 BLEU (Papineni et al., 2002) on a single-reference set of 50 sentences. To demonstrate how these techniques can be applied to other LRL, we adapt these strategies to Jamaican and show significant improvements over baseline performance, particularly via phonological transfer.

In summary, our findings suggest that despite back-transltion's reputation for usefulness in some settings, it cannot result in usable MT in others, in which case other transfer methods are needed for further improvement. To our knowledge, this is the first work to present this finding.

Case Study: Haitian We consider Haitian as a paradigm low-resource language. This language has critical importance for the global community, particularly in the context of recent immigration and disaster relief efforts. Haitian is closely related to high-resource French, but the two have an unconventional relationship: high phonological and lexical similarity with low syntactic and orthographic similarity. This is comparable to a large number of language pairs such as Thai and Lao, Arabic and Maltese, Jamaican and English, etc.

The Haitian government did not formalize a Haitian writing system until the 20th century. Still today, Haitians often write in French rather than Haitian due to social pressures, which contributes to a lack of written and digitized materials. Despite this lack of resources, Haitian is a widely spoken language. Roughly 12 million people speak it natively, including about 1 million immigrants in the USA and over a million more in Brazil, the Bahamas, Canada, Chile, Cuba, the Dominican Republic, France, Mexico, and elsewhere. Not many other residents of these countries learn Haitian. As a result, the lives of many Haitian speakers could be greatly improved by high-quality MT technology.

2 Related Work and Approach

We are not the first researchers to explore Haitian-to-English MT. Frederking et al. (1998) developed early statistical systems for Haitian MT and automatic speech recognition. In 2010 a devastating earthquake in Haiti's capital killed roughly a quarter million people. This disaster renewed international interest in Haitian MT systems for disaster relief efforts, the deployment of which was a "widely heralded success story" (Neubig and Hu, 2018).

Back-translation Augmentation Many researchers have employed back-translation to augment LRL data (Sennrich et al., 2016). This technique requires a small LRL-TGT bitext and a larger monolingual TGT corpus. Rather than mapping from LRL to HRL sentences in the small bitext, Sennrich et al. (2016) proposed a new method: (1) use the small bitext to train a TGT-to-LRL system, (2) translate the large TGT corpus to LRL, creating a large *synthetic* HRL-LRL bitext, then (3) train a system that maps from the LRL to the HRL on both the small authentic bitext and large synthetic bitext. In this paradigm, the quality of the synthetic translations may be low because they were produced by a system trained on a small bitext. The idea is that a small amount of high-quality data mixed with a large amount of low-quality data is preferable to a small amount of high-quality data alone. Back-translation has shown improvements in multiple MT settings (Popel et al., 2020). Xia et al. (2019) extended variations of this idea to a multilingual framework. They investigated translating to English (ENG) from an LRL that has a closely related HRL. A large HRL-ENG bitext, and small bitexts between the LRL and the two other languages are assumed, as well as a large monolingual ENG corpus. They proposed producing synthetic LRL-ENG aligned data in three ways:

1. Train an ENG-to-LRL system on the small LRL-ENG bitext, and translate the large monolingual English corpus to LRL (i.e. back-translation)

2. Train an HRL-to-LRL system on the small LRL-HRL bitext, and translate the large ENG-aligned HRL data to LRL

3. Train an ENG-to-HRL system on the HRL-ENG bitext, and using the system from the previous step, translate the large ENG monolingual corpus to HRL and then to LRL

In the current work, we apply these augmentation methods for Haitian-to-English translation with HRL French. We refer to the synthetic bitext produced by step 1 as `synth_mono`, by step 2 as `synth_mix1`, and by step 3 as `synth_mix2`.

Multi-source MT Multi-source MT incorporating one or more HRL-TGT bitexts into training has been shown to improve LRL-TGT translation. (Freitag and Firat, 2020; Zoph et al., 2016). Neubig and Hu (2018) trained systems that map from an LRL and one related HRL to English. This improved LRL-ENG BLEU score significantly. In our work we show that this method is more effective than back-translation when more authentic data is available, and we expand it through syntactic, orthographic, and phonological data representations to exploit relations between source languages.

3 Methodology and Experiments

Our experiments use a HAT-ENG bitext with 189,182 aligned sentence pairs (LRL-ENG) and a FRA-ENG bitext with 315,577 (HRL-TGT). These data come from broadcasts and literature produced by the Church of Jesus Christ of Latter-day Saints, with small additions from OPUS[1]. Because of overlap between the English portions of these two bitexts, we have an implicit FRA-HAT bitext of length 77,121. We have a large monolingual ENG corpus of text from Wikipedia, the Toronto book corpus (Zhu et al., 2015), and text scraped from Reddit.

All our models are attention-based (Vaswani et al., 2017), adapted from The Annotated Transformer (Klein et al., 2017), and trained using the Adam optimizer (Kingma and Ba, 2017). Hyperparameters are detailed in Appendix A.1 Because

[1] https://opus.nlpl.eu

Original French:	elle ne pensait pas descendre de sa maison pour lui rendre le livre, comme elle a fait ce matin
Orthograph transform:	lwi panse pa dèsann son kay pou lwi rann la liv, konm lwi gen fè sa maten
Syntax transform:	il pas tape penser descendre maison il pour rendre li livre le comme il té faire matin ce
Both transforms:	li pa tap panse dèsann kay li pou rann li liv la konm li te fè maten sa
Actual Haitian translation:	li pa tap panse desann sòti kay li pou rann li liv la, jan li te fè maten sa
English:	she did not want to descend from her house to give him the book, like she did this morning

Table 1: Outputs of the Haitian-approximating orthographic and syntactic engines applied to transform French text

we are comparing data sets produced with different transfer methods, we used this same model configuration for all experiments.

Haitian Back-translation We employed the same back-translation data augmentation strategies outlined in the numbered items of §2. To observe effects of this augmentation on varying amounts of authentic data, we augmented gradually. Starting with 5K, 25K and 189K lines of authentic aligned data, we added 5K, then, 25K, then 200K lines of synth_mono data. Then to the 200K of synth_mono we added 5K, 25K, then 200K of synth_mix1 data, and we followed suit with synth_mix2 data. Results from training on these 30 different sets are discussed in §4.

Multi-source Training We also trained multi-source MT models with HAT and LRL, FRA as HRL, and ENG as TGT. We conducted the same experiment with Spanish (SPA) as the HRL and with all three source languages together. We selected French and Spanish because of their proximity to Haitian. However, the nature of this proximity introduces interesting challenges. Roughly 90% of Haitian lexemes are of French origin, and the two languages are phonologically close. However they have few shared word forms because of their distinct orthography systems. And they are syntactically different. Because traditional MT transformers do not access phonological information, this similarity does not provide any benefit in using French as co-source with Haitian.

Orthographic, Syntactic, and Phonological Transfer To experiment with different methods of multi-source training, we developed a pipeline that orthographically transforms French to Haitian. The first engine changes word orthography via transformation rules based on French and Haitian grammar. The process resembles other automatic orthography transliterators like Epitran (Mortensen et al., 2018). The second engine uses the Berkeley Neural constituency parser (Kitaev et al., 2019) to change word order in French sentences, approximating Haitian syntax. This 922-line script tuned on zero data produces HAT reference translations from a single set with BLEU 1.60 and CER 59.5%[2].

In this manner we transform our French-English bitext into a pseudo-Haitian-English bitext and train jointly with that and our authentic Haitian-English data. To observe the different effects of transfer from orthographic similarity and from syntactic similarity in MT training, we also transform French to pseudo-Haitian using the two engines in isolation. See Table 1 for output examples.

Many languages are not lexically or phonologically close but share syntactic features, such as Jamaican and Haitian. We explore this more generalizable case in §4.

We employ a separate method to exploit phonological similarity between source languages. We convert Haitian and French words to IPA feature vectors using Epitran (Mortensen et al., 2018) and PanPhon (Mortensen et al., 2016). We represent each word as the sum of its phone vectors and use these to initialize transformer embeddings. In this way, the model can know that French *unité* (IPA: ynite) and its Haitian translation *inite* (IPA: inite) are closely related. This method does not involve transforming or altering either language and can be applied readily to other language pairs. For this application, we made significant improvements to Epitran for its French setting.

4 Results and Discussion

Figure 1 shows translation performance scores across a progression of back-translation-based augmentation as discussed in §3. These techniques improve performance when the amount of authentic data is very small. But once it crosses a threshold, they become counter-productive.

Results for multilingual source training experiments are in Table 2. This illustrates that bi- and trilingual source training can improve MT even when we use all 189K authentic HAT-ENG pairs. As mentioned in §3, our MT models cannot take

[2] BLEU is a poor metric for this engine since a majority of its errors are word choice differences and misspellings.

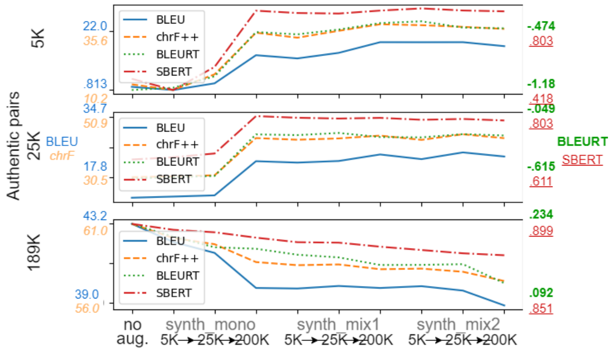

Figure 1: Scores in four performance metrics across models employing back-translation techniques. Back-translation augmentation increases to the right.

Source	BLEU	BLEURT
HAT	43.94*	.6810*
HAT+FRA	46.05*	.7026*
HAT+SPA	**46.51***	.7065
HAT+FRA+SPA	46.41*	**.7131***
HAT+JPN	30.41	-.1554

Table 2: HAT-ENG translation scores from multi-source training, best results **bolded**
*Significant improvement over next-best score, p=1e-6, details in Appendix B.1

full advantage of Haitian's similarity to French. Note that augmenting with French is still more helpful than with an unrelated language, Japanese, which degrades performance. The best configurations used Haitian and Spanish, evaluated using BLEU and BLEURT (Sellam et al., 2020).

Table 3 displays the results from different transformations of French source data to augment for HAT-ENG training. *Synt* and *Orth* refer to data transformation from our syntactic and orthographic FRA-to-HAT engines, respectively. *Phon* indicates phonological encoded similarity via PanPhon. *All* indicates all of these transfers employed at once. Overall, our best HAT-to-ENG model uses orthographically transformed FRA data, and the second-best uses both *Synt* and *Orth*.

Although these methods all score significantly higher than zero augmentation (and significantly higher than the untransformed FRA baseline in BLEU), their margin of improvement is smaller than expected. We hypothesize this could be improved by learning phonological embeddings that preserve phone order in the case of *Phon* and by tuning our FRA-HAT pipeline to a small amount of real data in the case of *Synt* and *Orth*.

Transform.	BLEU	BLEURT
No HRL	43.94	.6810
No transf. on FRA	46.05*	.7026
Synt	46.08*	.7015*
Orth	**46.88***	**.7061**
Synt+Orth	46.43*	.7057
Phon	46.21*	.7050
All	46.37*	.7053

Table 3: French co-source data transformed in three different ways to resemble Haitian, best results **bolded**
*Significant improvement over next-best score, p=1e-6

Transform.	BLEU	BLEURT
JAM→ENG (baseline)	4.868	.3873*
JAM+HAT →ENG (synt.)	10.32*	.4483*
JAM+cs-ENG →FRA (orth.)	7.807*	.1698
JAM+ENG phon. embeds. (phon.)	**81.31***	**.6861***

Table 4: Experiments for harnessing syntactic, orthographic, and phonological relatedness to higher-resourced languages for Jamaican translation
*Significant improvement over next-best score, p=1e-6

Rapid Adaptation to New Languages We show rapid adaptation of these methodologies to new languages, without language-specific transformation engines, by exploring Jamaican (JAM) MT. In this setting, phonological transfer is highly effective (see Figure 4). For this experiment we created a new Jamaican setting in Epitran via 37 mapping rules. (Note this step would be unnecessary for adaption to any of the 77 languages supported by Epitran.) This simple technique improves both BLEU and BLEURT scores markedly. We used ENG as HRL (with FRA as TGT) in this experiment and in orthographic transfer, which consisted of code-switched English data using a dictionary of 200 Jamaican words. For syntactic transfer we simply used HAT as the HRL, since Jamaican is even lower-resourced, and the two are syntactically close.

5 Conclusion

Although back-translation transfer methods are effective in some MT settings, in others they are unable to improve MT performance beyond a threshold or result in usable translation. Per our explorations, methods involving multilingual transfer during training are able to make further improvements, even when more authentic data is available and baseline performance is higher. Our experiments on Haitian MT have the potential for future improvements and broad social impact. And our exploration of Jamaican demonstrates the capacity of these techniques for significant improvements in low-resource domains more generally.

References

Robert E Frederking, Ralf D Brown, and Christopher Hogan. 1998. The diplomat rapiddeployment speech mt system. *MT Summit (1997)*, pages 261–262.

Markus Freitag and Orhan Firat. 2020. Complete multilingual neural machine translation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 550–560, Online. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization.

Nikita Kitaev, Steven Cao, and Dan Klein. 2019. Multilingual constituency parsing with self-attention and pre-training.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. Opennmt: Open-source toolkit for neural machine translation. In *Proc. ACL*.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 388–395.

David R. Mortensen, Siddharth Dalmia, and Patrick Littell. 2018. Epitran: Precision G2P for many languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).

David R. Mortensen, Patrick Littell, Akash Bharadwaj, Kartik Goyal, Chris Dyer, and Lori Levin. 2016. Pan-Phon: A resource for mapping IPA segments to articulatory feature vectors. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3475–3484, Osaka, Japan. The COLING 2016 Organizing Committee.

Graham Neubig and Junjie Hu. 2018. Rapid adaptation of neural machine translation to new languages. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 875–880, Brussels, Belgium. 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.

Martin Popel, Marketa Tomkova, Jakub Tomek, Łukasz Kaiser, Jakob Uszkoreit, Ondřej Bojar, and Zdeněk Žabokrtskỳ. 2020. Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals. *Nature communications*, 11(1):1–15.

Maja Popović. 2017. chrf++: words helping character n-grams. In *Proceedings of the second conference on machine translation*, pages 612–618.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. Bleurt: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Maali Tars, Andre Tättar, and Mark Fišel. 2021. Extremely low-resource machine translation for closely related languages. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 41–52, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Mengzhou Xia, Xiang Kong, Antonios Anastasopoulos, and Graham Neubig. 2019. Generalized data augmentation for low-resource translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5786–5796, Florence, Italy. Association for Computational Linguistics.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.

A Hyperparameters, Infrastructure, and Efficiency

We will release our software publicly upon acceptation.

A.1 All Experiments

The following settings are true for all experiements reported in this paper:

architecture: Transformer (Vaswani et al., 2017)
layers: 2 encoder layers, 2 decoder layers
attention heads: 6
learning rate: 0.0005
dropout rate: 0.1
optimizer: Adam (Kingma and Ba, 2017)

Following subsections provide the settings for individual experiments.

A.2 Experiment 1: Hatian Back-Translation

parameters: 43283546
training set (sentences): 4375-690535
evaluation set (sentences): 625-98647
computing infrastructure: NVIDIA GeForce GTX 1080 Ti
average runtime: $<$ 1 hour

A.3 Experiment 2: Multi-Source Training

parameters: 43283546
training set (sentences): 165535-777440
evaluation set (sentences): 23647-111062
computing infrastructure: NVIDIA GeForce GTX 1080 Ti
average runtime: 2-3 hours

A.4 Experiment 3: Orthographic, Syntactic, and Phonological Transfer

parameters: 43283546
training set (sentences): 441665
evaluation set (sentences): 63094
computing infrastructure: NVIDIA GeForce RTX 2080 Ti
average runtime: 2 hours

A.5 Experiment 4: Jamaican MT

parameters: 43283546
training set (sentences): 6939-283069
evaluation set (sentences): 991-40438
computing infrastructure: NVIDIA GeForce RTX 2080 Ti
average runtime: 1 hour

B Evaluation Metrics

We employed four translation evaluation metrics: BLEU (Papineni et al., 2002), BLEURT (Sellam et al., 2020), chrF++ (Popović, 2017), and Sentence-BERT (SBERT) (Reimers and Gurevych, 2019)

B.1 Computing Statistical Significance

We computed statistical significance via a difference of means test over our evaluation set. We used the `stats.wilcoxon` from SciPy. For BLEURT we considered a simple difference of means, and for BLEU we bootstrapped 1000 document-level scores from our evaluation set (Koehn, 2004).