Low-resource Neural Machine Translation: Benchmarking State-of-the-art Transformer for Wolof↔French

Cheikh M. Bamba Dione [1], Alla LO[2], El Hadji M. NGUER [3], Sileye BA [4]
[1]UiB Bergen Norway, [2]LANI UGB Senegal, [3]UVS Senegal, [4]Loreal Research & Innovation
dione.bamba@uib.no, lo.alla@ugb.edu.sn, elhadjimamadou.nguer@uvs.edu.sn, sileye.ba@dailymotion.com

Introduction

- This research work proposes two neural machine translation (NMT) systems (French-to-Wolof and Wolof-to-French) based on sequence-to-sequence (S2S) with attention and Transformer architectures.
- The first NMT system (**S2S**) is implemented as an encoder-decoder network with **gated recurrent units** (GRU). The implementation follows the common sequence-to-sequence framework (Sutskever et al., 2014).
- The second NMT system is implemented as a **transformer** (Vaswani et al., 2017) with a block of 4 encoders and a block of 4 decoders which deal with source sequences and target sequences, respectively.

Handling data sparsity

Because of the low-resource setting, we used advanced methods for handling data sparsity:

- **Subword unit**: rare words are encoded with sequences of subword units which are learned by applying Byte Pair Encoding (BPE) (Sennrich et al., 2016) on the union of the source and target corpora. Common vocabulary size: 15,000 pieces.
- **Backtranslation**: We first trained an initial target to source NMT system on the available parallel data, and then used that model to translate the monolingual corpus from the target language to the source language. The resulting back-translated data was combined with the original parallel data and used to train the final source to target NMT system.
- **Copied monolingual data**: Following Currey et al. (2017), we copied each target sentence into the source language, and added this copied material to the baseline parallel corpus. Then, we train the NMT models on the mixed corpus. We combined e.g. FR→FR and WO→FR into one system for the purpose of improving WO→FR quality (and vice versa for the opposite direction).

Model configuration

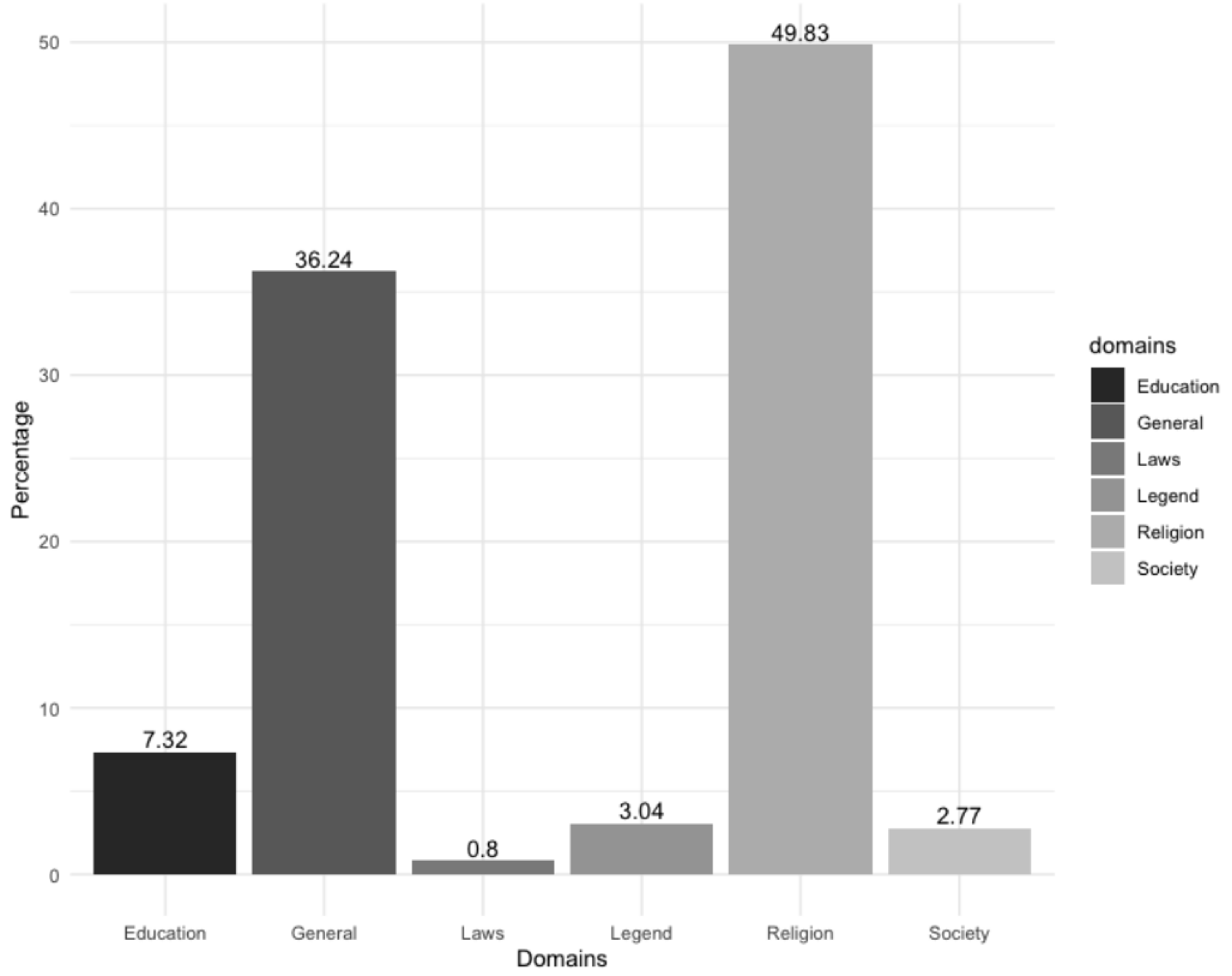
- For all models, the following hyper-parameters were kept constant across the different experiments: embedding size (256), optimizer (Adam), decay (noam), dropout rate (0.1), batch size (64 training sentences), and number of epochs (500).
- The S2S used encoder and decoder GRUs with 256 units. The transformer has a block of 4 encoders and a block of 4 decoders. Each encoder and decoder contains a self-attention layer and a feed-forward layer. In addition, each decoder has an encoder-to-decoder attention layer.

Contribution

- Implementing and benchmarking state-of-the-art NMT systems that are relevant for FR↔WO translators, thereby considerably facilitating their work.
- Creation of useful language resources for the large Wolof communities.
- Experimenting with advanced techniques for handling low-resource African languages.

Dataset

For training, we used the French-Wolof parallel corpus (Nguer et al., 2020), which contains around 83k sentences drawn from six main domains. The corpus was sharded into 78.5k sentences as baseline training data, 3k as validation data, and 1.5k sentences set aside as test data for evaluation.



Language	Tokens	Average length	Sentences
French	381,507	9.64	39,559
Wolof	584,851	16.40	35,674

Table 1: Statistic summary of the monolingual data used for back-translation.

Results and evaluation

The quality of our translations is evaluated by comparing the predictions and ground truth using case-insensitive and detokenized BLEU (Papineni et al., 2002) scores.

NMT model	Unit	Test set	
		FR→WO	WO→FR
S2S	word	21.0	24.3
	subword	22.3 (+1.3)	26.0 (+1.7)
Transformer	word	31.8	36.5
	subword	**33.6** (+1.8)	**37.5** (+1.0)

Table 2: FR↔WO translation performance when using **word** vs. **subword** units for **baseline** training.

NMT model	Unit	Test set	
		FR→WO	WO→FR
Transformer	word	31.8	36.5
	+backtrans	26.5 (-5.3)	37.9 (+1.4)
	subword	33.6	36.5
	+backtrans	25.1 (-8.5)	37.7 (+1.2)

Table 3: FR↔WO translation performance when using **back-translated** monolingual data.

Unit	Test set	
	FR→WO	WO→FR
subword	33.6	37.5
+copied	34.5 (+0.9)	37.8 (+0.3)

Table 4: FR↔WO translation performance of the transformer using **copied** monolingual data.

Unit	Test set	
	FR→WO	WO→FR
subword	33.6	37.5
+copied + backtranslation	35.1 (+1.5)	38.3 (+0.8)

Table 5: FR↔WO translation performance when using copied monolingual data + backtranslation.

Reference

[1] Currey, A., Miceli-Barone, A. V., and Heafield, K. (2017). Copied monolingual data improves low-resource neural machine translation. Second Conference on MT, pages 148–156.

[2] Nguer, E. M., Lo, A., Dione, C. M. B., Ba, S. O., and Lo, M. (2020). Sencorpus: A French-Wolof parallel corpus. LREC, pages 2796–2804. ELRA.

[3] Sennrich, R., Haddow, B., and Birch, A. (2016). Neural machine translation of rare words with subword units. ACL.

[4] Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In Advances in NeurIPS, pages 3104–3112.

[5] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In Advances in NeurIPS, pages 5998–6008.