Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 6654–6661
Marseille, 20-25 June 2022

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

Cheikh M. Bamba Dione, Alla Lo, Elhadji Mamadou Nguer, Sileye Oumar Ba
University of Bergen, Gaston-Berger University, UVS, Loreal Research & Innovation
Norway, Senegal, Senegal, France
dione.bamba@uib.no, lo.alla@ugb.edu.sn, elhadjimamadou.nguer@uvs.edu.sn,sileye.ba@dailymotion.com

Abstract

In this paper, we propose two neural machine translation (NMT) systems (French-to-Wolof and Wolof-to-French) based on sequence-to-sequence with attention and Transformer architectures. We trained our models on the parallel French-Wolof corpus (Nguer et al., 2020) of about 83k sentence pairs. Because of the low-resource setting, we experimented with advanced methods for handling data sparsity, including subword segmentation, backtranslation and the copied corpus method. We evaluate the models using BLEU score and find that the transformer outperforms the classic sequence-to-sequence model in all settings, in addition to being less sensitive to noise. In general, the best scores are achieved when training the models on subword-level based units. For such models, using backtranslation proves to be slightly beneficial in low-resource Wolof to high-resource French language translation for the transformer-based models. A slight improvement can also be observed when injecting copied monolingual text in the target language. Moreover, combining the copied method data with backtranslation leads to a slight improvement of the translation quality.

Keywords: French, neural machine translation, sequence-to-sequence, transformers, Wolof

1. Introduction

Neural Machine Translation (NMT) based on the encoder-decoder framework has achieved impressive results in high-resource data conditions (Sutskever et al., 2014; Cho et al., 2014). For low-resource conditions, recent studies have shown that NMT performance can be improved by using subword units (Sennrich et al., 2016b), backtranslation (Sennrich et al., 2015), and by adapting NMT systems to low-resource settings (Sennrich and Zhang, 2019).

In this research work, we investigate the design and implementation of NMT systems to translate between French (FR)[1] and Wolof (WO, ISO 639-3), a low-resource Niger-Congo language mainly spoken in Senegal (Gamble, 1950). We selected two NMT architectures: (i) an attentional sequence-to-sequence model (henceforth S2S) based on gated recurrent units (GRU) (Cho et al., 2014), and (ii) a Transformer (Vaswani et al., 2017; Hassan et al., 2018). According to recent studies, transformers have shown great promise as an approach to NMT for low-resource languages (Abbott and Martinus, 2018). At the same time, however, transformer models remain difficult to optimize and require careful hyper-parameter tuning to achieve good translation quality (Nguyen and Salazar, 2019).

As a main contribution of the paper, we have implemented and benchmarked state-of-the-art NMT systems, which are of high interest for French↔Wolof translators, thereby considerably facilitating their work. Our work will provide to large Wolof communities useful language resources they would not otherwise have access to. We have also conducted several experiments to deal with the data lack problem faced by low-resource languages. We hope that our work will allow the machine learning community to advance research on low-resource African languages.

The remainder of this paper is organized as follows. First, Section 2 provides background research. Section 3 outlines the methods used to address sparsity in low-resource NMT. Section 4 describes the NMT models considered in this work. Section 5 presents the various experiments conducted to assess the performance of the models. Section 6 presents the results and examines the implications of these findings. Finally, section 7 concludes the discussion.

2. Background

So far, minimal attention has been given to machine translation (MT) for African languages. A major difficulty hindering the progress of MT of such languages is that they are mostly very low-resource and the few resources that exist are often scattered and difficult to obtain. Recently, there have been a few attempts at using common NMT techniques for some South African languages (Martinus and Abbott, 2019; Nyoni and Bassett, 2021). More recent researches have benchmarked NMT between English and five African languages: Swahili, Amharic, Tigrigna, Oromo, and Somali (Lakew et al., 2020). Also, Tapo et al. (2020) investigated the case of Bambara in low-resource NMT setting. In contrast, Senegalese languages have so far not been subject to statistical or neural machine translation. In fact, none of these languages is currently supported by Google Translate. To our knowledge, Lo et al. (2019; Lo et al. (2020) and Dione (2019) are among

[1] As the official language of Senegal (the country of the most Wolof speakers), it is easier to find parallel data between French and Wolof than between e.g. English and Wolof.

the very few studies that have so far explored Wolof using neural network-based methods. Lo et al. (2020) developed an encoder-decoder NMT system based on bidirectional LSTMs and the attention mechanism. Although Lo et al. (2020) achieved interesting performance in terms of BLEU score, their results were quite biased due to a large overlap between the training, validation and test sets. Circa 60% of the sentences in the test set were also found in the training data. In that sense, the test set was not totally blind, leading to a high inflation of the BLEU scores.

Wolof has around 10 million speakers mostly located in the West African countries of Senegal and Gambia (Gamble, 1950). There are some official Wolof↔French dictionaries (Fal et al., 1990; Cissé, 1998; Diouf, 2003). The amount of written Wolof monolingual data could be estimated to ca. 2 million tokens (including encyclopedic, narrative, didactic, informative and literary works). A good part of these data (ca. 70%) can be accessed online using web crawling. Recently, efforts towards the acquisition and construction of a bilingual parallel corpus between French and Wolof have been reported by Nguer et al. (2020).

In this research work, we apply NMT techniques to a low-resource French-Wolof dataset and assess the performance of our models in terms of BLEU scores (Papineni et al., 2002). A more ambitious long-term goal is to effectively use deep learning methods to design NMT models that can be applied to local Senegambian languages. Research advancements in this area will represent monumental stepping stones towards providing support for translating several low-resource local Senegambian languages (e.g. Fula and Soninke).

3. Techniques to handle data sparsity

As Wolof is low-resource, we explored advanced techniques for handling issues of rare words and data bottlenecks. Accordingly, for each of the NMT architectures used in this work, we developed four different translation models: (i) a word-level vs. (ii) a subword-based model, (iii) a version using back-translation vs. (iv) a model without back-translated data. In addition, we used copied target-language data in combination with subword units to exploit monolingual corpora in the target language. Due to time constraints, backtranslation and the copied method were only tested on the best translation system, i.e. the Transformer.

3.1. Subword unit

Subword segmentation is a common method used to enable open-vocabulary translation by encoding rare words with sequences of subword units. This technique allows NMT models to translate or generate unseen words at inference time, while effectively decreasing the vocabulary size of the whole training. In this work, the subword units are learned by applying Byte Pair Encoding (BPE) (Sennrich et al., 2016b) on the union of the source and target corpora (i.e. joint BPE segmentation). Original vocabulary sizes of the French and Wolof baseline training data are 38,152 and 33,286 tokens, respectively. After applying BPE, the common vocabulary size was (empirically) set to 15,000 pieces.

3.2. Backtranslation

We also experimented with using automatic back-translation of monolingual data as a way of tackling issues of data bottlenecks. This method has been shown to be helpful for statistical machine translation (SMT) (Bojar and Tamchyna, 2011), supervised NMT (Sennrich et al., 2015) as well as unsupervised MT (Lample et al., 2018). Also, merging NMT/SMT back-translated data can have positive effects (Poncelas et al., 2019). For this purpose, we created synthetic parallel data by translating target-language monolingual text into the source language. 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. We applied back-translation for each direction (FR→WO and WO→FR).

3.3. Copied monolingual data

Additionally, we investigated incorporating monolingual training data into our NMT models as a way to circumvent the data sparsity problem. Following Currey et al. (2017), we generated a bitext from the monolingual data in the target language so that each source sentence is identical to the target sentence. We then added this copied material to the baseline parallel corpus and train the NMT models on the mixed corpus. In other words, we combined e.g. French→French and Wolof→French into one system for the purpose of improving Wolof→French quality. For the opposite direction, we applied the same technique. Like Currey et al. (2017), we used this method on models trained with subword units (not on word-level models). According to Currey et al. (2017), the copied data method proved to be beneficial for low-resource NMT (e.g. for English→Turkish and English→Romanian). An important advantage of that method (compared to back-translation) is that it does not require to train an additional target to source NMT system.

4. The NMT Models

This section briefly describes the attentional S2S model and the transformer used in this work. Both systems are implemented in TensorFlow Keras (Chollet and others, 2015).

4.1. The GRU-based S2S model

Our S2S system follows the common sequence-to-sequence framework (Sutskever et al., 2014; Cho et al., 2014). It is implemented as an encoder-decoder network with gated recurrent units (GRU) cells. During

training, the encoder learns the representations (embeddings) of the input sentence and generates a context vector that comprises the "meaning" of the sentence. Initialized with this vector, the decoder uses the start of sentence <SOS> symbol as input to generate the target sentence for the input sentence. At every time step t, we feed the predicted output word from the previous time step (y_{t-1}), and the previous hidden state (d_{t-1}), as an input to the decoder (d_t) at the current time step, and predict the current output (y_t). The predictions are used to calculate the loss. Teacher forcing is applied to decide the next input to the decoder. The gradients are calculated and applied to the optimizer before back-propagation. The inference process occurs in a similar manner, except that it does not use teacher forcing. The input to the decoder at each time step is its previous predictions along with the hidden state and the encoder output. This process stops when the model predicts the end of sentence <EOS> token.

The GRU-based model is extended with attention (Bahdanau et al., 2014), a module that assigns weights to each of the words in the source sentence when modeling target words. Thus, instead of taking the last hidden state as a context vector and using it for the decoder, we take the sum of all hidden states from the encoder and use it as a context vector. This mechanism allows to rank these hidden states in terms of their importance to generate the target word at time t. The attention weights are derived by first computing the similarity score between each of the encoder's hidden states with the decoder's hidden state using an alignment function. Then, this similarity score is converted into probabilities using the softmax function.

4.2. Transformer

Our transformer follows the architecture proposed by Vaswani et al. (2017). It has a block of 4 encoders and a block of 4 decoders which deal with source sequences and target sequences, respectively (see Figure 1).

During encoding, each input word is turned into a vector using an embedding layer. Then, positional encoding is employed to inject positional information into the input embeddings. Each encoder uses two layers to convert input to a continuous representation with attention information: a multi-head attention and a position-wise feed-forward neural network (FFN). The multi-head attention uses the self-attention mechanism, which allows the models to associate each word in the input to other words. Self-attention is achieved by first creating the query (**Q**), key (**K**), and value (**V**) vectors from the input. Then, we compute a score matrix by multiplying the query with the key vector and dividing by the square root of the dimension of the key vectors (noted as d_k). We apply softmax on the scaled score matrix to obtain the attention weights that are used to get an output vector (see eq. 1). Adding the latter vector to the original positional input embedding creates residual connection. The residuals go through layer

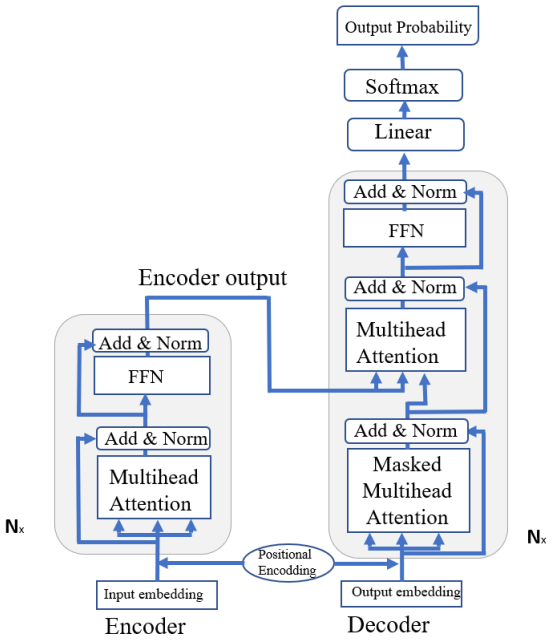

Figure 1: Our baseline transformer model

normalization and get projected through the pointwise FFN, i.e. a couple of linear layers with a ReLU activation in between. The output of that is then again added to the input of the pointwise FFN and further normalized. The pointwise feedforward layer is used to project the attention outputs potentially giving it a richer representation.

$$Attention(Q, K, V) = softmax(\frac{Q \times K^T}{\sqrt{d_k}}) \quad (1)$$

The decoder has two multi-headed attention blocks in one layer, one for the target sequences and one for the encoder's output. The former multi-head attention is masked to prevent computing attention scores for future words. The decoder also has a pointwise feed-forward layer, residual connections, and layer normalization after each sub-layer. As with encoding, during decoding the input goes through an embedding layer and positional encoding layer to get positional embeddings. Then, these embeddings get fed into the first multi-head attention layer which computes the attention scores for the decoder's input. The second multi-headed attention layer uses the encoder's outputs as the queries and the keys, and the first multi-headed attention layer outputs are the values. This process matches the encoder's input to the decoder's input, allowing the decoder to decide which encoder input is relevant to put a focus on. The output of the second multi-headed attention goes through a pointwise FFN layer for further processing. The output of the final pointwise feed-forward layer goes through a final linear layer, that acts as a classifier.

5. Experimental setup

5.1. Datasets

The data used for the baseline training come from the French-Wolof parallel corpus (Nguer et al., 2020), which contains around 83k sentences drawn from six main domains (education, general, law, legend, religion and society). Religious texts such as the Bible and the Quran represent ca. 50% of the corpus. In addition, ca. 30% of the corpus are French sentences that were extracted from the Tatoeba project[2] and translated into Wolof by professional translators. The translation work concerns only the Wolof side (i.e. the 30% of texts that we just mentioned; the rest are parallel sentences that were obtained as such and which did not need further translation). Figure 2 is taken from Nguer et al. (2020) and shows the distribution of the domain data.

The French-Wolof parallel corpus was sharded into 78.5k sentences as training data (i.e. the baseline training corpus), 3k as validation data, and 1.5k sentences set aside as test data for evaluation. The data split was done randomly. Then, we automatically removed all sentences that occurred in the training set from both the validation and the test sets. In addition, we applied the same process between the validation and the test set. This was to make sure that the three sets are disjoint and that there is no overlap between them. Furthermore, to ensure high quality translation in the targeted direction, we carefully conducted manual inspection of the data contained in the validation and test sets. The former is used to learn the model parameters and the latter to assess the performance of our final models. We evaluate our models on both directions French↔Wolof. Table 1 provides some statistics on the parallel corpus used as baseline for training (i.e. without copied and/or back-translated data).

Language	Tokens	Vocabulary	Sentences
French	860,032	38,152	78,569
Wolof	814,619	33,286	

Table 1: Statistic summary of the French-Wolof parallel corpus used as baseline training corpus.

In addition, for each language, we collected monolingual data to be used for back-translation. For the purpose of improving French→Wolof translation quality, we randomly sampled 35k Wolof sentences from web crawled monolingual data. The data were harvested mainly from online newspapers[3] and literary works (Diop, 2003; Ba, 2007).

Conversely, for the WO→FR NMT, we back-translated 39k French sentences gathered from the Tatoeba project. This choice was motivated by the fact that ca. 30% of the baseline parallel corpus comes from the same source. So, we wanted to use data from the same domain as those found in our baseline corpus. The monolingual French data are disjoint from those sentences that were already present in the parallel corpus.[4] The monolingual data used on each side do not overlap with the data included in the baseline training. Also, we should mention that the monolingual data differ significantly in quality, size and domain. First, the Wolof monolingual data come from domains like literature and news, which are very underrepresented in the parallel corpus. In contrast, the French monolingual data come from the same source as a good part of texts in the parallel corpus. Second, while the French monolingual data are quite clean, the Wolof data are very noisy (e.g. contains misspelling, wrong word segmentation, use of non-standard orthography). Finally, the French monolingual texts have a moderate sentence length. In contrast, the number of word tokens contained in the Wolof monolingual data is significantly higher, as can be seen in Table 2.

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

Table 2: Statistic summary of the monolingual data used for back-translation.

When creating the synthetic data for back-translation, we first trained an initial target to source NMT system on the baseline parallel data, and then use this model to translate the monolingual corpus from the target language to the source language. In our experiments, this translation was done by each NMT system independently. In other words, we used the best S2S model and the best transformer model, respectively, to translate the target monolingual data into the source language. The motivation behind using the best model is to minimize the number of additional systems to train (since back-translation requires the training of an additional target → source MT system).

When conducting the experiments, we followed the common practice which consists in tokenizing and lowercasing the parallel and monolingual training data. Similar to Bahdanau et al. (2014), we filtered out sentence pairs whose length exceeds 50 words and used padding to compensate for the empty slots in shorter sentences. We applied exactly the same parameters (including the sentence length limit) to both the training and the validation data.

When experimenting with subword units, we learned a shared byte pair encoding (BPE) model on the parallel data only (not on monolingual data). As recommended

[2]https://www.manythings.org/anki/

[3]The online sources include Wolof-online.com, https://www.defuwaxu.com, and http://saabal.com.

[4]We did not include news texts for French, because the original parallel corpus did not contain texts from that domain. Adding monolingual dataset from news texts (or other text genres) might cause some bias due to the difference between the baseline corpus and backtranslation corpus.

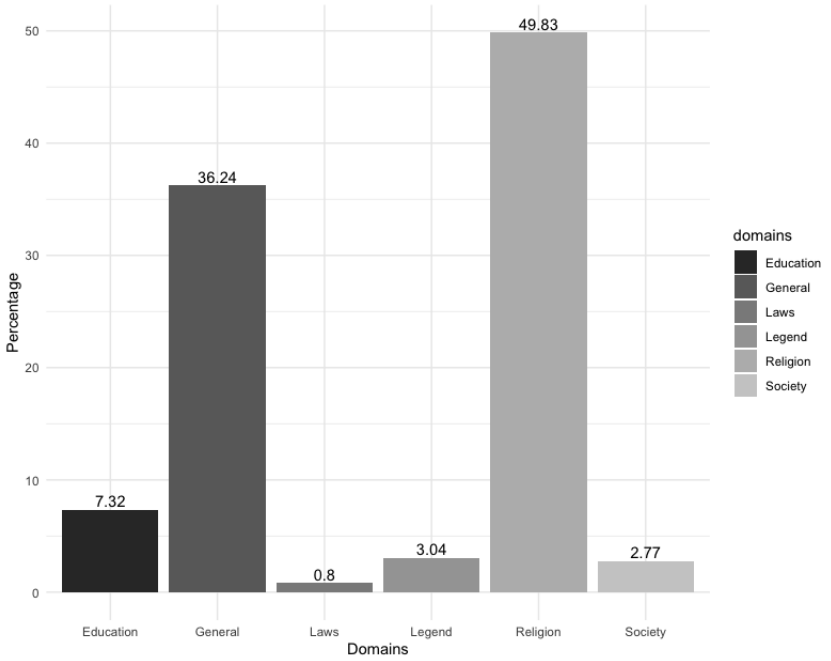

Figure 2: Distribution of domain data in the French-Wolof corpus (Nguer et al., 2020).

by Sennrich et al. (2016a), we removed diacritics from the source training data.

5.2. Model configuration

For all models, the following hyper-parameters were kept constant across the different experiment settings: embedding size, optimizer, and batch size. We used 256 dimensional word embeddings and the efficient Adam approach to stochastic gradient descent (Kingma and Ba, 2015) (adam_beta_1: 0.9; adam_beta2: 0.998), and *noam* as decay method. Our shuffled mini-batch contained 64 training sentences,[5] and each model was trained for 500 epochs. The dropout rate was set to 0.1. For the S2S system, we used encoder and decoder GRUs with 256 units. For the transformer, we chose a setting with a total of 8 blocks, where each block contains a self-attention layer, an encoder-to-decoder attention layer and a feed-forward layer. The hidden state dimension, the number of layers and the number of heads are 256, 6 and 8 respectively. As suggested by van Biljon et al. (2020), transformers with a moderate depth (i.e. 6 layers) seem to perform better than shallow (i.e. 2 transformer layers) or deep (12 transformer layers). The learning rate was set to 2.0, with warm-up of 800 steps. Decoding was performed using beam search (beam size=5). For the S2S and the transformer, the best hyperparameter settings were chosen based on results obtained during validation.

Training a single (word-level or subword-level) based model for 500 epochs on the baseline corpus took ca. 25 hours for S2S and 15 hours for the transformer on a single GPU (RTX 2080). Training on the combined corpus (i.e. baseline + backtranslated or copied corpus) took ca. twice longer.

[5]In the future, we plan to experiment with smaller batch sizes, as these proved to be beneficial in low-resource settings (Nguyen and Chiang, 2018; Sennrich and Zhang, 2019).

6. Results and evaluation

The quality of our translations is evaluated by comparing the predictions and ground truth using BLEU (Papineni et al., 2002). We report case-insensitive and detokenized BLEU scores computed with SacreBLEU (Post, 2018). BLEU scores of the two NMT systems trained on the baseline corpus for 500 epochs are shown in Table 3. From all models, the subword-level based Transformer achieves the best performance (36.5 BLEU points on WO→FR and 37.5 on FR→WO translation), outperforming the S2S by ca. 12 BLEU points on the WO→FR direction. For all models, BLEU score increases when words are split into subword pieces. Integrating this technique has a quite positive impact on all models and both directions.

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

Table 4 shows results of the NMT systems when adding back-translated data to the baseline corpus. For the

setting with back-translated data, we used target-side back-translated data for both directions.

Translation quality drops for the Transformer models when using backtranslation, but only for the word-based FR→WO direction. Interestingly, this technique led to a slight improvement of translation quality in WO→FR direction (by up to 2.1 BLEU points).

NMT model	Unit	Test set	
		FR→WO	WO→FR
Transformer	word	31.8	36.5
	+backtrans	26.5 (-9.0)	37.9 (+1.4)
	subword	33.6	36.5
	+backtrans	25.1 (-8.5)	37.7 (+1.2)

Table 4: FR↔WO translation performance when using **back-translated** monolingual data and training the Transformer models for 500 epochs.

The results in Table 4 also seem to indicate that back-translations in the 78.5K setup are of very poor quality and the noise is too detrimental for this low-resource setting. Interestingly, on the low-resource side (i.e. Wolof), the backtranslated output may still provide useful training signal to both the word-based as well as the subword-based transformer models.[6] On the high-resource language side, this hypothesis is difficult to verify as the signal might have been damaged by the noise from the monolingual Wolof data. In addition, the accuracy of backtranslated Wolof news is not so good, since the domain of the backtranslated data and the parallel corpus do not match. This seems to have a negative impact on the FR→WO translation.

Furthermore, we studied whether the copied corpus can help improve the translation quality. Following Currey et al. (2017), we only tested this method on the subword-based transformer, which was trained on the parallel corpus + copied data for 500 epochs. The results of this experiment are given in Table 5, showing a slight improvement of BLEU points (between 0.3 to 0.9) for both translation directions.

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

Table 5: FR↔WO translation performance of the transformer using **copied** monolingual data.

Our final experiment consisted in verifying the impact of combining the copied data method with back-translation. The results of this experiment are shown in Table 6. As can be seen, the combination of these two corpora results in a slight improvement for subword-based models.

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

Table 6: FR↔WO translation performance of the transformer using a combined corpus (copied monolingual data + backtranslated data).

To give the reader a qualitative flavor of the translations produced by our models, Tables 7 and 8 show samples that represent good and sometimes erroneous translations (marked in red) made by our models. These examples provide insights into understanding the potential of our current approach, despite the scarcity of the training data. Both NMT systems seem to perform quite well on the test set, with the transformer being more sensitive to long sentences. As can be seen, the output of the S2S model may be fluent for short sentences (ca. 7 words). However, for longer sentences (ca. 20 words), there is a substantial drop in terms of both fluency and accuracy. This issue has been observed in both translation directions. Also, the S2S model tends to output named entities which are not related at all to the input text. In contrast, the transformer seems to handle long sentences much better than the S2S model. In addition, it also often produces exact matches or paraphrases that convey a meaning very similar to the target sentence.

Translations generated by our models have been showed to professional Wolof translators (Akaademi Wolof). They found our results very promising in assisting them when translating French to Wolof. In addition, we have developed a web platform allowing these translators to concretely test and exploit our models.

7. Conclusion

In this paper, we proposed an attentional GRU-based S2S and a transformer for translating between French and Wolof. We experimented with methods for improving low-resource NMT, including subword segmentation, backtranslation and copied monolingual data. Our experiments showed that the best baseline results (37.5 BLEU points) are achieved when training the transformer on data segmented at subword-level. This kind of segmentation seems to have a quite positive impact on the translation quality, leading to an increase in BLEU scores for almost all models. In addition, for the transformer, backtranslation proved to be slightly beneficial when translating from the low-resource (WO) to the high-resource (FR) language. Likewise, a similar improvement can be observed when injecting copied material from target-language text. Finally, a combination of backtranslation and the copied data method proved to an effective way of improving translation quality (BLEU scores increased by up to 38.3 points).

[6]Manual inspection of the backtranslation shows that the WO→FR translation was quite good. However, the FR→WO translation quality was quite poor due to the domain mismacth and the very low quality of the target-side input.

		Translation issue(s)
Source	Je ne sais pas si j'aurai le temps de le faire. (I don't know if I will have time to do it.)	
Target	Dama xamul ndax dinaa am waxtuw def ko. I not.know if I.will have time do it.	
Transformer	xamuma ndegam dinaa am waxtuw def ko.	Translation is fine.
S2S	xamuma ndax dama koy war a def.	Fluent, but slightly inaccurate translation ("I don't know if I will have to do it.")
Source	Et maintenant je vous dis: Ne vous mêlez plus de ces hommes et laissez-les. (And now I say to you, Do nothing to these men, but let them be.)	
Target	Léegi maa ngi leen di wax génnleen ci mbirum ñooñu te bàyyi leen ñu dem.	
Transformer (word based)	Léegi maa ngi leen di wax génnleen ci mbirum ñooñu nit ñi mu ne leen ñu dem.	Almost matches exactly the reference translation, use of paraphrases
S2S (word based)	Waaye bàyyileen li ma leen wax leen ba noppi di leen yëgal yii ngeen nekk ci samay taalibe.	Less fluent, but also inaccurate translation (issues with word order and lexical choice)

Table 7: Sample **FR→WO** translations made by the S2S and transformer systems (errors are marked in red).

Source	Ba juróom ñaareelu weer wa amee fukki fan ak juróom ñaar, gaal gaa nga teereji tundi Araraat. (The ark rested in the seventh month, on the seventeenth day of the month, on Ararat's mountains.)	
Target	Et le dix-septième jour du septième mois l'arche s'arrêta sur les montagnes d'Ararat.	
Transformer	et le dix-septième mois l'arche s'arrêta sur les montagnes d'Ararat.	Missing translation + mismatch (17th instead of 7th month; 17th day not translated)
S2S	et le septième jour du côté du roi de Juda sortit de la montagne de Babylone jusqu'au septième mois.	Translation not accurate; also added unrelated named entities like Juda and Babylone

Table 8: Sample **WO→FR** translations made by the S2S and transformer systems (errors are marked in red).

Although the designed NMT systems achieved state-of-the-art performance on low-resource French↔Wolof translation, we believe there is still some room for improving the performance of our transformer models by more carefully tuning their hyper-parameters. As recent studies pointed out, low-resource NMT seems to be very sensitive to hyperparameters such as BPE vocabulary size, word dropout, and others. Future work will concentrate on exploring sophisticated techniques to refine our current models without necessarily relying on auxiliary resources such as monolingual data.

8. Bibliographical References

Abbott, J. Z. and Martinus, L. (2018). Towards neural machine translation for African languages. *arXiv preprint:1811.05467*.

Ba, M. (2007). *Bataaxal bu gudde nii*. Nouvelles Editions Africaines du Sénégal (NEAS).

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *ICLR*.

Bojar, O. and Tamchyna, A. (2011). Improving translation model by monolingual data. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 330–336. ACL.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. *EMNLP*.

Chollet, F. et al. (2015). Keras. `https://github.com/fchollet/keras`.

Cissé, M. (1998). *Dictionnaire français-wolof*.

Langues & mondes/L'Asiathèque.

Currey, A., Miceli-Barone, A. V., and Heafield, K. (2017). Copied monolingual data improves low-resource neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, pages 148–156.

Dione, C. B. (2019). LSTM based Language Models for Wolof. In Zygmunt Vetulani et al., editors, *Human Language Technologies as a Challenge for Computer Science and Linguistics*, pages 190–194, Poznan, Poland. Wydawnictwo.

Diop, B. B. (2003). *Doomi Golo: Nettali*. Editions Papyrus Afrique.

Diouf, J.-L. (2003). *Dictionnaire wolof-français et français-wolof*. Editions Karthala, Paris.

Fal, A., Santos, R., and Doneux, J. L. (1990). *Dictionnaire wolof-français: suivi d'un index français-wolof*. Karthala.

Gamble, D. P. (1950). The wolof of Senegambia: Western Africa part xiv (ethnographic survey of Africa). *Routledge*.

Hassan, H., Aue, A., Chen, C., Chowdhary, V., Clark, J., Federmann, C., Huang, X., Junczys-Dowmunt, M., Lewis, W., Li, M., et al. (2018). Achieving human parity on automatic chinese to english news translation. *arXiv preprint:1803.05567*.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In Yoshua Bengio et al., editors, *3rd International Conference on Learning Representations (ICLR), San Diego, CA, USA*.

Lakew, S. M., Negri, M., and Turchi, M. (2020). Low-resource neural machine translation: A benchmark for five african languages. *arXiv preprint:2003.14402*.

Lample, G., Conneau, A., Denoyer, L., and Ranzato, M. (2018). Unsupervised machine translation using monolingual corpora only. In *International Conference on Learning Representations*.

Lo, A., Ba, S., Nguer, E. H. M., and Lo, M. (2019). Neural words embedding: Wolof language case. In *IREHI19*.

Lo, A., Dione, C. M. B., Nguer, E. M., Ba, S. O., and Lo, M. (2020). Using LSTM to translate french to senegalese local languages: Wolof as a case study. *CoRR*, abs/2004.13840.

Martinus, L. and Abbott, J. Z. (2019). A focus on neural machine translation for African languages. *arXiv preprint:1906.05685*.

Nguer, E. M., Lo, A., Dione, C. M. B., Ba, S. O., and Lo, M. (2020). Sencorpus: A French-Wolof parallel corpus. In *Proceedings of The 12th LREC*, pages 2796–2804, Marseille, France. ELRA.

Nguyen, T. Q. and Chiang, D. (2018). Improving lexical choice in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Nguyen, T. Q. and Salazar, J. (2019). Transformers without tears: Improving the normalization of self-attention. In *International Workshop on Spoken Language Translation*.

Nyoni, E. and Bassett, B. A. (2021). Low-resource neural machine translation for Southern African languages. *arXiv preprint:2104.00366*.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318.

Poncelas, A., Popovic, M., Shterionov, D. S., de Buy Wenniger, G. M., and Way, A. (2019). Combining SMT and NMT back-translated data for efficient NMT. *CoRR*, abs/1909.03750.

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

Sennrich, R. and Zhang, B. (2019). Revisiting low-resource neural machine translation: A case study. In *Proceedings of the 57th Annual Meeting of ACL*, pages 211–221, Florence, Italy. ACL.

Sennrich, R., Haddow, B., and Birch, A. (2015). Improving neural machine translation models with monolingual data. *CoRR*, abs/1511.06709.

Sennrich, R., Haddow, B., and Birch, A. (2016a). Edinburgh neural machine translation systems for WMT 16. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 371–376, Berlin, Germany. Association for Computational Linguistics.

Sennrich, R., Haddow, B., and Birch, A. (2016b). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of ACL (Volume 1: Long Papers)*, pages 1715–1725. Association for Computational Linguistics.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Tapo, A. A., Coulibaly, B., Diarra, S., Homan, C., Kreutzer, J., Luger, S., Nagashima, A., Zampieri, M., and Leventhal, M. (2020). Neural machine translation for extremely low-resource African languages: A case study on Bambara. *arXiv preprint:2011.05284*.

van Biljon, E., Pretorius, A., and Kreutzer, J. (2020). On optimal transformer depth for low-resource language translation. *arXiv preprint:2004.04418*.

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 neural information processing systems*, pages 5998–6008.