Adam Mickiewicz University's English-Hausa Submissions to the WMT 2021 News Translation Task

Artur Nowakowski and **Tomasz Dwojak**
Faculty of Mathematics and Computer Science, Adam Mickiewicz University, Poznań, Poland
{artur.nowakowski,t.dwojak}@amu.edu.pl

Abstract

This paper presents the Adam Mickiewicz University's (AMU) submissions to the WMT 2021 News Translation Task. The submissions focus on the English↔Hausa translation directions, which is a low-resource translation scenario between distant languages. Our approach involves thorough data cleaning, transfer learning using a high-resource language pair, iterative training, and utilization of monolingual data via back-translation. We experiment with NMT and PB-SMT approaches alike, using the base Transformer architecture for all of the NMT models while utilizing PB-SMT systems as comparable baseline solutions.

1 Introduction

We describe the Adam Mickiewicz University's submissions to the WMT 2021 News Translation Task. We focused on translation between Hausa and English – a low-resource translation scenario between distant languages. Our methods combine data cleaning with OpusFilter (Aulamo et al., 2020) and fastText (Joulin et al., 2016), transfer learning (Aji et al., 2020; Zoph et al., 2016), iterative training, and back-translation (Sennrich et al., 2016a).

All NMT models were trained with FAIRSEQ (Ott et al., 2019), while the first iteration of the back-translation was generated with Moses (Koehn et al., 2007).

The results presented in the paper are based on the first released development set ("Dev-1"), which consists of 1000 sentences, the final development set ("Dev-full"), which adds additional 1000 sentences to the first development set, and the released test set without additional test suites ("Test"). The test set consists of 1000 sentences in English→Hausa direction and 997 sentences in Hausa→English direction.

The final submissions significantly outperform the vanilla NMT baselines in terms of BLEU (Papineni et al., 2002) metric results, as implemented in SACREBLEU (Post, 2018) with default settings.

All systems were trained in a constrained scenario i.e., using the data provided by the organizers of WMT 2021 only.

2 Data preparation

The quality of the training data has a great impact on the final performance of the NMT models (Rikters, 2018). The data preparation consisted of data cleaning and filtering performed by using OpusFilter (Aulamo et al., 2020) pipelines. We specified separate pipelines for monolingual and parallel data. Data cleaning phase consisted of normalizing punctuation, removing non-printable characters, and decoding HTML entities by using Moses (Koehn et al., 2007) pre-processing scripts.

We applied subword segmentation on filtered data by using SentencePiece (Kudo and Richardson, 2018) tool with byte-pair-encoding (BPE) (Sennrich et al., 2016b) algorithm. The corpora we used for model training, along with the number of sentences before filtering, are specified in Table 1. Number of sentences after filtering is presented in Table 2.

Monolingual data filtering For the monolingual data filtering, we defined an OpusFilter pipeline that consists of the following filters:

- deduplication filter,
- sentence length filter,
- word length filter,
- Latin character score filter,
- language identification filter.

The sentence length filter requires that the sentence contain a minimum of 3 and a maximum

Proceedings of the Sixth Conference on Machine Translation (WMT), pages 167–171
November 10–11, 2021.

Data type	Sentences	Corpora
Parallel en-ha	751,560	Khamenei, Opus, ParaCrawl
Monolingual en	41,428,626	News crawl (only 2020)
Monolingual ha	2,311,959	News crawl, CommonCrawl
Parallel de-en	8,600,361	Tilde Rapid, CommonCrawl, Europarl, News commentary, ParaCrawl

Table 1: Corpora statistics before filtering.

of 100 words. A maximum of 40 characters is required for the word length. The required Latin character score for a sentence is set to 100%. Language identification filter is based on a fastText (Joulin et al., 2016) language identifier. The open-source fastText language identification models do not identify Hausa, so we used the JW300 corpus from the English-Hausa Opus collection to train our custom language identifier. A sentence must pass all filters to be included in the training data.

Data type	Sentences
Monolingual en	39,812,834
Monolingual ha	1,227,921
Parallel ha-en	494,246

Table 2: Monolingual corpora statistics after filtering.

Parallel data filtering The filters used in the parallel data filtering pipeline are nearly identical to those used in the monolingual data filtering pipeline. Filters are applied to both the source and target sentences in this scenario. We also included a length ratio filter with a threshold of 2, indicating that a sentence on the source side can be up to twice as long as a sentence on the target side and vice versa.

A similar pipeline was applied to the German-English data that was used for transfer learning. We downsampled 3M sentence pairs from ParaCrawl due to the imbalance in the German-English data.

3 Approach

Our models combine transfer learning from a high-resource language pair (German-English), iterative training, and back-translation. We used FAIRSEQ (Ott et al., 2019) toolkit in our experiments with NMT models, while we used Moses (Koehn et al., 2007) toolkit for our experiments with PB-SMT models.

All of our NMT models follow the base Transformer architecture (Vaswani et al., 2017), using ReLU as the activation function and Adam (Kingma and Ba, 2015) as the optimizer with the following parameters: $\beta_1 = 0.9$, $\beta_2 = 0.98$, $\epsilon = 1\text{e}{-8}$. We set the inverse square root learning rate scheduling with a peak value of $1\text{e}{-3}$. We used learning rate warmup stage for 4000 updates with initial learning rate of $1\text{e}{-7}$. Dropout probability was set to 0.2, while the attention dropout probability was set to 0.1. We also used label smoothing with a value of 0.1. In the case of baseline English-Hausa models, the joint vocabulary was based on both English and Hausa data. In all cases, the vocabulary size was set to 32,000.

The PB-SMT models were trained with default settings with Moses (Koehn et al., 2007) toolkit. In addition, we trained a 5-gram Operation Sequence Model (Durrani et al., 2013). All language models are 5-gram models and were binarized with KenLM (Heafield et al., 2013). The models were trained on tokenized, word-level, lowercased sentences. Re-casing was applied to the model outputs. After training the base models, we also applied MERT (Minimum Error Rate Training) (Och, 2003; Bertoldi et al., 2009) tuning on the development set.

3.1 Baseline systems

We decided to train baseline models of two types: vanilla Transformer (base) and PB-SMT. The experiments conducted on the first release of the development set showed that PB-SMT performs significantly better than NMT: we achieved +1.8 BLEU score on Hausa→English and +0.7 on English→Hausa. Based on these results, we decided to use PB-SMT models to generate data for the first iteration of iterative training.

When the test set was published, we computed the scores for the baselines. To our surprise, the scores obtained by NMT are much higher than PB-SMT, especially in the Hausa→English direction.

System	HA → EN		EN → HA	
	Dev-1	Test	Dev-1	Test
NMT baseline	12.21	11.44	10.28	11.05
PB-SMT baseline	14.00	6.59	11.02	9.36

Table 3: Baseline results according to the automatic evaluation with BLEU metric.

3.2 Transfer learning

According to recent studies, transfer learning (TL) enhances translation quality in low-resource scenarios (Zoph et al., 2016; Aji et al., 2020). We chose the German→English translation direction as a base. In general, we followed (Nguyen and Chiang, 2017) and trained a shared Hausa-German-English vocabulary (BPE). Then, we trained a German→English model using parallel data from the WMT 2021 Translation Task, which was filtered similarly to Hausa-English data. Finally, we used the Hausa-English data to fine-tune the pre-trained German→English model. We obtained a BLEU score of 13.31 on the "Dev-1" development set (+1.1 BLEU compared to the NMT baseline), which was lower than the PB-SMT baseline.

3.3 Iterative back-translation

Monolingual data has been widely employed in MT to enrich parallel corpora with synthetic data to improve the quality of MT systems, particularly in low-resource scenarios (Bojar and Tamchyna, 2011; Bertoldi and Federico, 2009). We applied the back-translation technique (Edunov et al., 2018) iteratively (Hoang et al., 2018) to translate Hausa and English monolingual data into the other language, using intermediate models to generate incrementally better translations.

1. First, we used the best baseline model (PB-SMT based on Moses) in English→Hausa direction to translate 5M English sentences into Hausa.

2. We used this additional data to train the Hausa→English model by applying transfer learning from the German→English model. We upsampled the original parallel data 10 times to match the size of the back-translated data. We used the resulting NMT model to translate all Hausa monolingual data into English via sampling.

3. We combined the obtained back-translated data with the original parallel corpora to train the English→Hausa model in a manner similar to step 2, with the exception that we did not upsample the parallel data in this scenario due to the fact that back-translated data was generated through sampling.

4. This technique was applied iteratively, resulting in the systems shown in Table 4. In all Hausa→English systems except the last, we utilized 5M English monolingual sentences in the model training; in the last system, we used 25M sentences. We used all accessible Hausa monolingual data in all English→Hausa systems.

System	HA → EN	EN → HA
1	16.22	-
2	-	13.04
3	20.05	-
4	-	14.38
5	22.85	-
6	-	14.77

Table 4: Iterative back-translation results of the NMT systems on the "Dev-1" development set according to the automatic evaluation with BLEU metric.

4 Final results

Table 5 presents the final results for both the English→Hausa and Hausa→English translation directions for both the development and test sets. These results were produced by the final models from the iterative back-translation step described in section 3.3.

Direction	Dev-1	Dev-full	Test
EN → HA	14.77	21.21	16.15
HA → EN	22.85	25.23	14.13

Table 5: Final results according to the automatic evaluation with BLEU metric.

We notice a severe decrease in BLEU metric results on the test set as compared to the development set, particularly in the Hausa→English direction. This could suggest a domain shift between the two sets. Because our models are heavily based on the back-translated data, some vocabulary, especially proper names, may be missing from the training data.

5 Post-submission work

Due to a lack of computing power and time, our experiments and submissions were based on single model training. After the submission deadline, we retrained the final models three times with different seeds. Table 6 presents the results for the ensemble of four models in both directions. We obtained slight improvements on both test sets, but the differences are insignificant. On the other hand, the ensemble performed worse on the development set, especially on the first version.

Direction	Dev-1	Dev-full	Test
EN → HA	14.68	21.00	16.34
HA → EN	21.24	26.25	14.87

Table 6: Post-submission models ensemble results according to the automatic evaluation with BLEU metric.

References

Alham Fikri Aji, Nikolay Bogoychev, Kenneth Heafield, and Rico Sennrich. 2020. In neural machine translation, what does transfer learning transfer? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7701–7710, Online. Association for Computational Linguistics.

Mikko Aulamo, Sami Virpioja, and Jörg Tiedemann. 2020. OpusFilter: A configurable parallel corpus filtering toolbox. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 150–156. Association for Computational Linguistics.

Nicola Bertoldi and Marcello Federico. 2009. Domain adaptation for statistical machine translation with monolingual resources. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 182–189, Athens, Greece. Association for Computational Linguistics.

Nicola Bertoldi, Barry Haddow, and Jean-Baptiste Fouet. 2009. Improved minimum error rate training in moses. *The Prague Bulletin of Mathematical Linguistics*, 91:7–16.

Ondřej Bojar and Aleš Tamchyna. 2011. Improving translation model by monolingual data. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 330–336, Edinburgh, Scotland. Association for Computational Linguistics.

Nadir Durrani, Alexander Fraser, Helmut Schmid, Hieu Hoang, and Philipp Koehn. 2013. Can Markov models over minimal translation units help phrase-based SMT? In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 399–405, Sofia, Bulgaria. Association for Computational Linguistics.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable modified Kneser-Ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 690–696, Sofia, Bulgaria. Association for Computational Linguistics.

Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. 2018. Iterative back-translation for neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 18–24, Melbourne, Australia. Association for Computational Linguistics.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Toan Q. Nguyen and David Chiang. 2017. Transfer learning across low-resource, related languages for neural machine translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 296–301, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167, Sapporo, Japan. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

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 on Association for Computational Linguistics*, ACL '02, page 311–318, 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, Belgium, Brussels. Association for Computational Linguistics.

Matīss Rikters. 2018. Impact of Corpora Quality on Neural Machine Translation. In *In Proceedings of the 8th Conference Human Language Technologies - The Baltic Perspective (Baltic HLT 2018)*, Tartu, Estonia.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. 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.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. 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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

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.