The University of Edinburgh's English-German and English-Hausa Submissions to the WMT21 News Translation Task

Pinzhen Chen **Jindřich Helcl** **Ulrich Germann** **Laurie Burchell** **Nikolay Bogoychev**
Antonio Valerio Miceli Barone **Jonas Waldendorf** **Alexandra Birch** **Kenneth Heafield**
School of Informatics, University of Edinburgh
{pinzhen.chen, jhelcl, ulrich.germann, laurie.burchell, n.bogoych,
amiceli, jwaldend, a.birch, kenneth.heafield}@ed.ac.uk

Abstract

This paper presents the University of Edinburgh's constrained submissions of English-German and English-Hausa systems to the WMT 2021 shared task on news translation. We build En-De systems in three stages: corpus filtering, back-translation, and fine-tuning. For En-Ha we use an iterative back-translation approach on top of pre-trained En-De models and investigate vocabulary embedding mapping.

1 Introduction

We describe the University of Edinburgh's participation in English↔German (En↔De) and English↔Hausa (En↔Ha) at the WMT 2021 news translation task. We apply distinct sets of techniques to the two language pairs separately, as the two pairs are very different in terms of language proximity and the availability of resources. We follow the constrained condition where we only use the provided data available to all participants.

For En↔De we first employ rule-based and dual conditional cross-entropy filtering to clean the datasets. Then we add to training back-translations generated in a few ways: tagged, greedy, beam search and sampling. We fine-tune our models on past years' test sets, and finally tune a few configurations: length normalization, test sentence splitting, and German post-processing.

For En↔Ha we adopt iterative back-translation, where at each iteration we initialize the model parameters from an En-De model in the corresponding direction (En→De for En→Ha and De→En for Ha→En). These En-De models are trained in the same way as those submitted to the En-De task, except that their vocabulary includes subwords from the Hausa language. Besides, we experiment with vocabulary mapping at the embedding level.

Some configurations are kept consistent across language pairs and systems. Sentences are tokenized using SentencePiece (Kudo and Richardson, 2018) with a 32K shared vocabulary, except that we added a few extra tokens for tagged back-translation. All models are trained following Marian's Transformer-Big task preset (Vaswani et al., 2017; Junczys-Dowmunt et al., 2018) unless otherwise specified: 6 encoder and decoder layers, 16 heads, 1024 hidden embedding size, tied embeddings (Press and Wolf, 2017), etc.[1]

Section 2 and Section 3 describe the detailed model building process for En↔De and En↔Ha respectively. While awaiting human evaluation results, we summarize our automatic metric scores on the WMT 2021 test sets computed by the task organizers in Table 1.

Direction	BLEU	ChrF
En→De	29.90	0.59
De→En	51.78	0.66
En→Ha	14.81	0.45
Ha→En	14.89	0.42

Table 1: Automatic metric scores on WMT21 test computed by the task organizers.

2 English↔German

2.1 Data and cleaning

English-German is considered to be a high-resource language pair, with over 90 million parallel and hundreds of millions monolingual sentences provided in the shared task. Following our last year's submission (Germann, 2020), we divide the data into three categories, and we use all the parallel data, as well as monolingual news from 2018 to 2020:

- High-quality parallel: News Commentary, Europarl and Rapid.

[1] https://github.com/marian-nmt/marian/blob/master/src/common/aliases.cpp

- Crawled parallel: ParaCrawl, WikiMatrix, CommonCrawl, and WikiTitles.
- Monolingual news: News Crawl

The majority of parallel data are mined and aligned sentences from the web (Bañón et al., 2020; Schwenk et al., 2021), so our first step is corpus filtering to remove noisy sentences which could harm neural machine translation (Khayrallah and Koehn, 2018). We run rule-based filtering using FastText language identification (Joulin et al., 2016), and various handcrafted features such as sentence length, character ratio and length ratio. Similar rules are applied on the monolingual data, omitting the features designed for parallel data. More details can be found in our cleaning script which is made public.[2]

We then train seed Transformer-Base models on the filtered high-quality data, as well as the crawled data separately, to (self-)score translation cross-entropy of the crawled parallel sentences. This enables us to rank and filter out sentences by their dual conditional cross-entropy (Junczys-Dowmunt, 2018). The method prefers the sentences in a pair to have low and similar translation cross-entropy given each other. After empirical trials, we find it is always better to score using models trained on the high-quality data, and we choose to keep the best 75% of the crawled data. The filtering efforts are reported in Table 2. Next, we train Transformer-Big models on the combination of filtered high-quality and crawled data. These models serve as baselines and are used for back-translation later.

Amount of crawled	**Scoring model**	**De→En**	**En→De**
top 25%	high-qual	41.47	-
	crawled	39.35	-
top 50%	high-qual	41.64	**43.68**
	crawled	41.51	-
top 75%	high-qual	**42.15**	**43.40**
	crawled	41.90	-
all	-	**42.02**	42.70

Table 2: BLEU of filtering experiments on WMT19 test used as dev.

2.2 Back-translation

Since its introduction, back-translation (Sennrich et al., 2016) has been widely used to boost NMT. We use ensembles of our best seed and baseline models trained on the filtered data, to generate back-translations from the monolingual news data from 2018 to 2020, hoping that the domains are similar to that of the test. For En→De we mix back-translations generated using greedy search, beam search, and sampling; for De→En, we adopt tagged back-translation (Caswell et al., 2019).

After merging the original and back-translated data, for each direction we train 4 standard Transformer-Big models, as well as a model with 8 encoder layers and 4 decoder layers. Specifically for De→En, we have an extra pre-layer normalized variant.

As we observed last year, validation BLEU does not improve after we add back-translated data to training. As a result, after the models converge, we continue training them on filtered parallel data only. The models' validation BLEU scores[3] on WMT19 test are displayed in Table 3.

Configuration	**De→En**	**En→De**
Baseline	42.2	43.4
+ BT	41.8	43.0
+ cont. training	**42.5**	**43.6**

Table 3: Average BLEU scores of BT experiments on WMT19 test used as dev.

2.3 Fine-tuning and submission

We grid search on length normalization during decoding, and find 1.2 to be ideal for En→De and 0.8 for De→En. Particularly for En→De, we have two more steps to make German text read more natural: 1) continued training on 25% title-cased parallel data to improve headline translation and 2) post-processing on German quotes to make them consistent.

Previous submissions show that fine-tuning on past years' test data helps model performance (Schamper et al., 2018; Koehn et al., 2018). In the early years of WMT news translation tasks, the test sentence pairs can originate in either source or target language, and are translated and merged into one set. However, the current evaluation is on translating sentences originally in the source language only. Therefore, we experiment with fine-tuning on the combined sets, as well as on sentence pairs originated from the source language. We fine-tune

[2] https://github.com/browsermt/students/tree/master/train-student/clean

[3] sacreBLEU (Post, 2018) with signature BLEU+case.mixed+numrefs.1+smooth.exp+tok.13a+version.1.5.1

all our models on WMT 2008-2019 test sets and validate on WMT 2020 test set.

While the training data contain mainly one sentence per line, the test set can have multiple sentences in the same segment. As a result, we split each test instance into single sentences, translate, and rejoin them. We experiment with fine-tuning and sentence splitting on the 8-encoder-4-decoder variant for both languages. Table 4 indicates that the model achieves the best BLEU (and a significant improvement over BT baseline) if we fine-tune it on previous test sentences originating in the source language only, and split long sentences in both validation and test sets.

FT on	**Dev split**	**Test split**	**De→En**	**En→De**
BT baseline			30.8	31.9
none		✓	41.7	35.2
all			34.7	-
all		✓	41.1	-
all	✓		31.2	-
all	✓	✓	41.9	36.7
orig.	✓	✓	**42.5**	**36.9**

Table 4: BLEU of fine-tuning and sentence-splitting experiments on WMT20 test

For each translation direction, we apply the best configuration to each model and ensemble them by averaging their predictions post-softmax. Overall, we have a 5-model ensemble for En→De, and a 6-model ensemble De→En.

3 English↔Hausa

3.1 Data

The main sources of English-Hausa parallel data are OPUS (Tiedemann, 2012) and ParaCrawl. We also include data from WikiTitles[4] and the Khamenei[5] corpora, which are however much smaller. In total, we gather 759,061 parallel sentences. For back-translation, we use 9.5 million monolingual Hausa sentences from Common Crawl, Extended Common Crawl, and News Crawl provided by the task organizers. We randomly select 50 million English monolingual sentences from the News Crawl collections from 2018, 2019, and 2020.

For training, we use a mix of back-translated monolingual data and parallel data. Since the dataset sizes differ substantially, we over-sample the parallel data to achieve a balanced mix: 10× for English→Hausa, and 50× for Hausa→English. Similar to our En-De models, we used tagged back-translation to distinguish synthetic and authentic sentences in the data.

3.2 Iterative back-translation and fine-tuning

In our experiments, we combine a transfer learning approach (Zoph et al., 2016; Kocmi and Bojar, 2018) with 3 iterations of back-translation (Hoang et al., 2018; Edunov et al., 2018). In each iteration, we initialize the En→Ha model with a pre-trained En→De Transformer-Big model (and vice versa for the other direction). Then, we fine-tune the model on the English-Hausa data created by the model from the previous back-translation iteration (the initial model for the first iteration is fine-tuned on parallel data only).

We notice that the model generates a large number of empty translations. We suppress this issue by taking the second-best candidate translation from the n-best list if the first one is empty. Another problem is heavy overfitting in the models. In many translations, the sentences begin with the prefix "Never miss an important update!", followed by the actual translation. Unfortunately, we only noticed this issue after the submission.

3.3 Vocabulary embedding mapping

An additional approach we investigate is mapping the Hausa vocabulary to the German embeddings of the En→De model, when initializing the En→Ha model. We train the models with a 32K SentencePiece vocabulary obtained from datasets in all three languages. Using the frequency-based metric introduced by (Wang et al., 2020) we assign each SentencePiece token to an English, German, Hausa or joint vocabulary. This results in 9192 German tokens, 6485 Hausa tokens and a joint vocabulary of approximately 11k. Having established a separate Hausa and German vocabulary it is then possible to map between the embeddings of the two.

In order to map the vocabularies, we independently train BWEs (bilingual word embeddings) using an implementation of Bivec (Luong et al., 2015) combined with FastText (Bojanowski et al., 2017). This implementation uses a joint learning objective as described by Liu et al. (2020) utilising alignments combined with sub-word information.

[4] http://data.statmt.org/wikititles/v3/

[5] http://data.statmt.org/wmt21/translation-task/ha-en/khamenei.v1.ha-en.tsv

In lieu of a parallel De-Ha dataset an En→De NMT model is used to translate the English side of the En-Ha dataset. We constrain SentencePiece encoding using the previously extracted vocabularies for example the Huasa data is encoded using only the Hausa tokens and the joint tokens. Once both sides are encoded FastAlign is used to extract automatic alignments and the BWEs are trained.

We first map the Hausa tokens to their nearest neighbour using the Cross-Domain Similarity Local Scaling (Lample et al., 2018) distance metric in the order of Hausa tokens' frequency, and only permit a German token to be mapped to exactly one Hausa token. For tokens that do not have a one-to-one mapping, we adapt Gu et al. (2018)'s approach, whereby the embedding of a Hausa token is initialized to the weighted sum of all German embeddings. The weights are given by a probability distribution derived from the distance of the Hausa token to each German token in the bilingual embedding space. It is worth noting that we only map between the tokens in the Hausa and German vocabularies not any of the joint tokens. Finally, we initialize the embedding table using the new embeddings and remove all tokens identified as German. After initialization, we fine-tune the model using the parallel and back-translated data as described previously.

Our experiments show that although initializing the embedding table using a mapping-based approach results in faster model convergence, it does not improve the final BLEU score compared to just fine-tuning from the En-De models. This was observed for both the parallel data and the combined parallel and back-translated data. The outputs of the mapping approach to the baseline for the Ha-En system are qualitatively very similar and indicates that while the embedding mapping increases convergence there is no knowledge transfer from the German embeddings.

4 Conclusion

We describe our English-German and English-Hausa submissions to the news translation task at WMT 2021. For the En↔De task, fine-tuning and splitting test instances significantly boosts BLEU while back-translation alone does not help. In the En↔Ha task, we experiment with interesting low resource NMT techniques, but unfortunately, our submission contains translations from overfitted models.

Acknowledgements

This work was supported by funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825303 (Bergamot), 825627 (European Language Grid) and 825299 (GoURMET).

This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via contract #FA8650-17-C-9117. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

This work used the Cirrus UK National Tier-2 HPC Service at EPCC (http://www.cirrus.ac.uk) funded by the University of Edinburgh and EPSRC (EP/P020267/1).

This work was performed using resources provided by the Cambridge Service for Data Driven Discovery (CSD3) operated by the University of Cambridge Research Computing Service (www.csd3.cam.ac.uk), provided by Dell EMC and Intel using Tier-2 funding from the Engineering and Physical Sciences Research Council (capital grant EP/P020259/1), and DiRAC funding from the Science and Technology Facilities Council (www.dirac.ac.uk).

References

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. ParaCrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online. Association for Computational Linguistics.

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.

Isaac Caswell, Ciprian Chelba, and David Grangier. 2019. Tagged back-translation. In *Proceedings of* *the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 53–63, Florence, Italy. 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.

Ulrich Germann. 2020. The University of Edinburgh's submission to the German-to-English and English-to-German tracks in the WMT 2020 news translation and zero-shot translation robustness tasks. In *Proceedings of the Fifth Conference on Machine Translation*, pages 197–201, Online. Association for Computational Linguistics.

Jiatao Gu, Hany Hassan, Jacob Devlin, and Victor O.K. Li. 2018. Universal neural machine translation for extremely low resource languages. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 344–354, New Orleans, Louisiana. 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, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Marcin Junczys-Dowmunt. 2018. Dual conditional cross-entropy filtering of noisy parallel corpora. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 888–895, Belgium, Brussels. Association for Computational Linguistics.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Huda Khayrallah and Philipp Koehn. 2018. On the impact of various types of noise on neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 74–83, Melbourne, Australia. Association for Computational Linguistics.

Tom Kocmi and Ondřej Bojar. 2018. Trivial transfer learning for low-resource neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 244–252, Brussels, Belgium. Association for Computational Linguistics.

Philipp Koehn, Kevin Duh, and Brian Thompson. 2018. The JHU machine translation systems for WMT 2018. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 438–444, Belgium, Brussels. 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.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*.

Lu Liu, Yi Zhou, Jianhan Xu, Xiaoqing Zheng, Kai-Wei Chang, and Xuanjing Huang. 2020. Cross-lingual dependency parsing by POS-guided word reordering. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2938–2948, Online. Association for Computational Linguistics.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Bilingual word representations with monolingual quality in mind. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 151–159, Denver, Colorado. 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.

Ofir Press and Lior Wolf. 2017. Using the output embedding to improve language models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 157–163, Valencia, Spain. Association for Computational Linguistics.

Julian Schamper, Jan Rosendahl, Parnia Bahar, Yunsu Kim, Arne Nix, and Hermann Ney. 2018. The RWTH Aachen University supervised machine translation systems for WMT 2018. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 496–503, Belgium, Brussels. 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.

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.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

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. Curran Associates, Inc.

Zirui Wang, Jiateng Xie, Ruochen Xu, Yiming Yang, Graham Neubig, and Jaime G. Carbonell. 2020. Cross-lingual alignment vs joint training: A comparative study and a simple unified framework. In *International Conference on Learning Representations*.

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.