The Mininglamp Machine Translation System for WMT21

Shiyu Zhao, Xiaopu Li, Minghui Wu, Jie Hao
Mininglamp Technology, Beijing, China
{zhaoshiyu, lixiaopu, wuminghui, haojie}@mininglamp.com

Abstract

This paper describes Mininglamp neural machine translation systems of the WMT2021 news translation tasks. We have participated in eight directions translation tasks for news text including Chinese↔English, Hausa↔English, German↔English and French↔German. Our fundamental system was based on Transformer architecture, with wider or smaller construction for different news translation tasks. We mainly utilized the method of back-translation, knowledge distillation and fine-tuning to boost single model, while the ensemble was used to combine single models. Our final submission has ranked first for the English→Hausa task.

1 Introduction

This paper describes the Mininglamp submissions to the WMT2021 news translation tasks for eight directions including four high-resource Chinese↔English, German↔English, two medium-resource French↔German and two low-resource Hausa↔English. Furthermore, all of our systems were built with constrained data sets.

For this participation, we experimented with some smaller or wider Transformer (Vaswani et al., 2017) architectures to reach a reliable baseline based on different resource scales, sampling or beam search in back-translation to generate more suitable pseudo bilingual sentences. Particularly in the low-resource tasks, Hausa↔English, the Transformer-Small neural machine translation was built for the baseline, we presented iterative between back-translation and fine-tuning pattern which significantly improve the BLEU score on the validation set, and it worked well on English→Hausa task. Due to time constraints, we did not experiment on Hausa→English task. This path could be an experiment in the future work.

As for the data augmentation aspect, we experimented with several back-translation methods (Sennrich et al., 2016a), including the beam search, unrestricted sampling and sampling-topK (Edunov et al., 2018), to leverage the target-side monolingual data. We also applied knowledge distillation (Freitag et al., 2017) to leverage the source-side monolingual data.

Our systems followed four main steps:1) data filtering and preprocessing, 2) back-translation to generate pseudo bilingual data, 3) knowledge distillation by monolingual data, 4) fine-tuning with in-domain.

It should be emphasized that we used Marian[1] (Junczys-Dowmunt et al., 2018) to implement only for Hausa↔English baseline systems, and Fairseq[2] (Ott et al., 2019) for the rest, include Hausa↔English back-translation and knowledge distillation models.

2 System Overview

2.1 Data Filtering and Preprocessing

In this section, we discuss the preprocessing, normalization and filter techniques carried out in an attempt, in order to reduce spurious uncertainty in the modeling problem.

2.1.1 Text Preprocessing

Generally, we carried out the following text preprocessing steps prior to use in every model:

- Normalization: Unicode canonicalization, replacement of common multiple encoding errors present in training data, standardization of quotation marks into directional variants, conversion of any traditional Chinese characters into simplified forms, conversion of any Chinese full-width characters and segmental Chinese full-width punctuation into half-width forms. Normalize punctuation in all data by using Moses[3] (Koehn et al., 2007)

[1] https://github.com/marian-nmt/marian
[2] https://github.com/pytorch/fairseq
[3] https://github.com/moses-smt/mosesdecoder

Proceedings of the Sixth Conference on Machine Translation (WMT), pages 260–264
November 10–11, 2021. ©2021 Association for Computational Linguistics

(normalize-punctuation.perl) script except for every language pair.

- Segmentation: Chinese was segmented using the Jieba[4] segmentation tool, and tokenizer using Moses (tokenizer.perl) script for English, German, French and Hausa. For the Hausa tokenizer, we used English tokenizer instead.

- True-case: The word, at the start of a sentence, containing only an initial capital letter was replaced with the capitalized variant. That occurred most frequently in other positions of the English monolingual training data. Thus, in the previous sentence, the initial token would be "words" rather than "Words". We used Moses' script for true-case.

- Subword: The neural machine translation system is capable of open-vocabulary translation by representing rare and unseen words as a sequence of subword units. The model was trained based on subword-nmt[5] on the parallel training corpus.

2.1.2 Data Filtering

For all language pairs, the data filtering process for the training bilingual corpus stayed to the principle with the following rules:

- Filter out the sentence pairs that contain blank lines either from the source side or the target side.

- Filter out the sentence pairs that the source side and the target side at the same.

- Filter out the sentences with the length ratio falling outside from 0.4 to 2.5.

- Filter out the sentences whose punctuation and foreign words taking more than 40 percent.

- Remove the sentences which are longer than 200 words, or exceed a single word with 30 characters.

- Filter out the sentences which contain HTML tags or duplicated translations.

- Filter out the sentences which its word ratio between the source and the target exceeds 1:2.5 or 2.5:1.

- Identify language and delete foreign languages. Filter parallel and monolingual data by language detection using cld2[6].

The rules described above were also employed when cleaning monolingual and back-translation data. In the monolingual data particularly there were some lines that include two or more sentences, we cut them into several sentences by writing a script.

2.2 Data Augmentation

2.2.1 Back-Translation

Back-translation (Sennrich et al., 2016a) is an essential method to integrate the target side monolingual synthetic knowledge when building a state-of-the-art neural machine translation system. Especially for low-resource language tasks, it's indispensable to augment the training data by mixing the pseudo corpus with the parallel part. In that the target side, lexicon coverage was insufficient. The nucleus sampling (Holtzman et al., 2020) in back-translation to generate more suitable pseudo bilingual sentences. We attempted several data augmentation methods as follow, with different single technologies or combinations.

- Beam search: Generated target translation by beam search with beam 5.

- Sampling: Selected a word randomly from the whole distribution in each step, which increases the diversity of pseudo corpus with low precision, compared with beam search.

- Sampling Top-K: Selected a word in a restricted way that only top-K (we set K as 16) words could be chosen.

2.2.2 Forward Translation to Generate Synthetic Parallel Sentence

For Chinese↔English tasks. To generate a more diverse pseudo-parallel corpus, we use forward-translated to do generated synthetic parallel sentences on source monolingual data only by our own ensemble model.

[4] https://github.com/fxsjy/jieba

[5] https://github.com/rsennrich/subword-nmt

[6] https://github.com/CLD2Owners/cld2

2.2.3 Knowledge Distillation

We used knowledge distillation (Kim and Rush, 2016) to do distillation on the original dataset. Specifically, we translated the source-side of the bilingual data using previously trained proposal models, and generated distilled candidates. We then trained models on filtered data along with the original bilingual data and back-translation data.

2.3 Iterative Back-translation and Fine-tuning

A process which iterative twice between back-translation and fine-tuning was implemented by following steps for the low-resource Hausa↔English tasks.

2.4 Reranking

For German↔English, French↔German tasks, we followed noisy-channel (Yee et al., 2019) reranking using one neural language model and three reverse translation models.

3 Experiment

3.1 Experiment Settings

In order to demonstrate the experiments of the system, there some experiment details should be clarified. To train all of the models used in our system, we made use only of the constrained data sets provided to shared news translation task participants. On the other side, the baseline models were trained on parallel corpus only by cleaned corpus. In terms of model evaluation, the main indicator for the report was calculated according to sacreBLEU[7] (Post, 2018) based on the results which has been removed parts of post-preprocessing such as removed BPE symbols, detruecased, detokenized, etc.

The Transformer-Small was implemented based on Marian (Junczys-Dowmunt et al., 2018) as our baseline for Hausa↔English tasks. For Chinese↔English, German↔English and French↔German tasks, we implemented the Transformer-Big FFN-8192 based on Fairseq (Ott et al., 2019) as our baseline model. We used Adam optimizer (Kingma and Ba, 2014) during training, learning rate was 5e-4, β_1 = 0.9, β_2 = 0.98, weight decay was 0.0001, label smoothing was 0.1. Specifically, the learning rate warmed up over the 8,000 steps for pre-normalize architectures Transformer-Big FFN-8192 model. The system shuffled the training data before generating the training batch for each epoch, so the document context information was not considered in this case. FP16 was applied to accelerate training with few performance damage during the training process.

3.2 Chinese↔English

For Chinese↔English system, our parallel corpus included CCMT, wikititles-v3, wikimatrix-v1, para-crawl-v7.1, news-commentary-v16 corpus. While Chinese were segmented by Jieba word segmentation toolkit, English was tokenized by Moses tokenizer script. Based on the result of data Filtering, we used 17 million Chinese↔English parallel data corpus for training the baseline model. As the next step after the preprocessing, we trained BPE (Sennrich et al., 2016b) models which were learned with 32,000 merge operations for joined English and Chinese on the parallel data. We built separately vocabularies for each language, and the final vocabulary size of Chinese was 42K and English was 22K. Baseline train data we followed drop-BPE (Provilkov et al., 2020). We trained the Transformer-Big FFN-8192 model for Chinese↔English.

For back-translation, we selected 20 million News Crawl 2020 English monolingual data for Chinese→English task. All News Crawl Chinese monolingual data and selected 20 million Extended Common Crawl Chinese monolingual data were combined for English→Chinese task. Back-translation data were combined by sampling top-16 and beam search. At the same time, there was a combination between back-translation data and parallel data corpus in order to train Chinese↔English models. We selected 10 million Chinese and English sentences respectively for forward translation and knowledge distillation to generate synthetic parallel sentences.

Our final submissions consisted of three Transformer-Big FFN-8192 models with different configurations, using the beam search with a beam size of 5, and set lenpen 2.0. Table 1 shows that the translation quality was improved by using the proposed techniques.

3.3 Hausa↔English

The parallel corpus for Hausa↔English system included para-crawl-v8, wikititles-v3, Khamenei and Opus corpus, which was tokenized by Moses tokenizer script. It should be clear that Hausa used tokenizer by English mode. After the data filter-

[7] https://github.com/mjpost/sacrebleu

System	zh-en	en-zh
baseline	30.2	42.9
+ Back Translation	33.4	45.1
+ Knowledge Distillation	33.8	46.2
+ Fine-tuning	34.7	47.8
+ Ensemble	35.5	48.6

Table 1: SacreBLEU scores on newstest2020 Chinese↔English tasks.

ing, we used 550 thousand Hausa↔English parallel data corpus for training the baseline model. A joint BPE model was applied with 10,000 merge operations. Moreover, shared vocabularies were selected for Hausa↔English language pairs.

We used Marian trained Transformer-Small[8] model for Hausa↔English baseline, with learning rate ranging from 0.0008 to 0.001, warmup steps fixing at 48,000. Three models(3e3d, 4e4d, 6e4d) were trained under different architectures on single 2080Ti GPU.

For English↔Hausa back-translation, the standard Transformer-Big model implemented in Fairseq. We selected 4.5 million Hausa monolingual data by data filtering and language detection, and 20 million English monolingual data from the News Crawl 2020 were filtered as the back-translation dataset. Every time we handled the back-translation process, the beam search was applied. Then the back-translation and the fine-tune were executed twice. For Hausa→English, due to time constraints, it was limited to one back-translation and fine-tune.

In the fine-tuning stage, 200 sentences from the newsdev2021 were kept randomly as the validation set, and other sentences were attributed to fine-tune the model.

Table 2 shows that the translation quality was improved by using the proposed techniques. Our final submissions consisted of two Transformer-Big models.

3.4 German↔English

For German↔English task, the provided parallel sentences were completely joined together so as to get about 95 million sentence pairs. Then, sentences with lots of punctuation masks and non-alpha-number characters were removed, as well as the sentences whose length ratio was larger than 2. As a result, 52 million sentences were selected to be candidates. After that, BPE was learned jointly with 32k as the merge operations, and the size of the vocabulary was 32,168. The model's parameters for both directions were copied from the Transformer-Big in the paper "Attention is all you need" (Vaswani et al., 2017). Finally, we got three English→German models and two English↔German models for ensembling and reranking. The language model used for reranking was trained with GPT-3 using data cleaned from news 2020. All the models were trained using Fairseq. The overview of our German↔English system is listed in Table 3.

System	ha-en	en-ha
baseline	13.8	11.6
+ 1st. Back-translation	24.6	22.7
+ 1st. Fine-tuning*	29.7	25.5
+ 2nd. Back-translation*	-	26.2
+ 2nd. Fine-tuning*	-	26.9
+ Ensemble*	31.7	27.4

Table 2: SacreBLEU scores on newsdev2021 Hausa↔English tasks. Steps with extra * marks are evaluated in the tiny 200 lines new validation set.

System	de-en	en-de
baseline	44.1	40.0
+ Ensemble	45.1	41.1
+ Reranking	45.5	41.4

Table 3: SacreBLEU scores on newstest2016 German↔English tasks. Learning rate for training is 0.001 and warmup steps are 4000.

3.5 French↔German

For French↔German task, about 7 million sentences were left after removing the sentences with invalid characters or punctuations from the original parallel sentences. We trained the BPE codes with 32k as the merge operations. The final vocabulary size for German was 32,144 and for French was 32,176. We introduced forward translation in German→French direction using models trained from the original parallel dataset. In both directions, the models were based on the Transformer-Big as the basic architecture. At last, three French→German models and two German→French models, trained from forward-translation, were applied to ensembling and reranking. The language model used for reranking was

[8] The dimension of word embedding was 256, the dimension of the feed-forward network was 1024, multi-head was 4, encoder and decoder layer was 4.

trained with GPT-3 using data cleaned from news 2020. The models from this system were completely trained by Fairseq. Check the overview of our German↔French systems in Table 4.

System	de-fr	fr-de
baseline	30.9	27.6
+ Knowledge Distillation	31.3	-
+ Ensemble	32.6	28.8
+ Reranking	34.1	30.9

Table 4: SacreBLEU scores on newstest2019 French↔German tasks.

4 Conclusions

This paper described the Mininglamp submissions to the WMT2021 eight news translation tasks, and our main exploration was using more diversified architectures, back-translation, fine-tuning and ensemble. We used a similar data preprocess and filtering strategy for all the tasks, containing statistical information-based rules. And we experimented with back-translation by different decoding strategies, using the Transformer-Small model and iterative between back-translation and fine-tuning for low-resource.

References

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.

Markus Freitag, Yaser Al-Onaizan, and Baskaran Sankaran. 2017. Ensemble distillation for neural machine translation. *arXiv preprint arXiv:1702.01802*.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

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.

Yoon Kim and Alexander M. Rush. 2016. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327, Austin, Texas. Association for Computational Linguistics.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

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.

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*.

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.

Ivan Provilkov, Dmitrii Emelianenko, and Elena Voita. 2020. BPE-dropout: Simple and effective subword regularization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1882–1892, Online. Association for Computational Linguistics.

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, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Kyra Yee, Yann Dauphin, and Michael Auli. 2019. Simple and effective noisy channel modeling for neural machine translation. In *Conference on Empirical Methods in Natural Language Processing*.