NICT's Unsupervised Neural and Statistical Machine Translation Systems for the WMT19 News Translation Task

Benjamin Marie[1*], **Haipeng Sun**[2,1*], **Rui Wang**[1†], **Kehai Chen**[1],
Atsushi Fujita[1], **Masao Utiyama**[1], **and Eiichiro Sumita**[1]
1 National Institute of Information and Communications Technology (NICT)
2 Harbin Institute of Technology
{bmarie, sun.haipeng, wangrui, khchen, atsushi.fujita, mutiyama, eiichiro.sumita}@nict.go.jp

Abstract

This paper presents the NICT's participation in the WMT19 unsupervised news translation task. We participated in the unsupervised translation direction: German-Czech. Our primary submission to the task is the result of a simple combination of our unsupervised neural and statistical machine translation systems. Our system is ranked first for the German-to-Czech translation task, using only the data provided by the organizers ("constraint"), according to both BLEU-cased and human evaluation. We also performed contrastive experiments with other language pairs, namely, English-Gujarati and English-Kazakh, to better assess the effectiveness of unsupervised machine translation in for distant language pairs and in truly low-resource conditions.

1 Introduction

This paper describes the unsupervised neural (NMT) and statistical machine translation (SMT) systems built for the participation of the National Institute of Information and Communications Technology (NICT) to the WMT19 shared News Translation Task. Only one translation direction was proposed in the unsupervised track of task: German-to-Czech (de-cs). Our submitted systems are constrained, in other words, we used only the provided monolingual data for training our models and the provided parallel data for development, i.e., validation and tuning. We trained unsupervised NMT (UNMT) and unsupervised SMT (USMT) systems, and combined them through training a pseudo-supervised NMT model with merged pseudo-parallel corpora and n-best list reranking using different informative features as proposed by Marie and Fujita (2018a). This simple combination method performed the best among unsupervised MT systems at WMT19 by BLEU [1] and human evaluation (Bojar et al., 2019). In addition to the official track, we also present the unsupervised systems for English-Gujariti and English-Kazakh for contrastive experiments with much more distant language pairs.

The remainder of this paper is organized as follows. In Section 2, we introduce the data preprocessing. In Section 3, we describe the details of our UNMT, USMT, and pseudo-supervised MT systems. Then, the combination of pseudo-supervised NMT and USMT is described in Section 4. Empirical results produced with our systems are shown and analyzed in Section 6 and 7, and Section 8 concludes this paper.

2 Data and Preprocessing

2.1 Data

As monolingual training data to train our de-cs UNMT and USMT systems, we randomly extracted 50 million sentences from WMT monolingual News Crawl datasets.[2] Bilingual development data (16.6K sentences) from "last years' parallel dev and test sets"[3] were also officially provided "for bootstrapping" the UNMT systems.[4] Among the large number of possible approaches for exploiting the development data, we only used it for tuning USMT, validate UNMT models, train a reranking system, and finally to fine-tune our pseudo-supervised NMT systems.

[*] Equal contribution in alphabetical order. This work was conductd when Haipeng Sun visited NICT as an internship student.

[†] Corresponding author.

[1] http://matrix.statmt.org/matrix/systems_list/1897

[2] http://data.statmt.org/news-crawl/

[3] http://data.statmt.org/wmt19/translation-task/dev.tgz

[4] http://www.statmt.org/wmt19/translation-task.html

Proceedings of the Fourth Conference on Machine Translation (WMT), Volume 2: Shared Task Papers (Day 1) pages 294–301
Florence, Italy, August 1-2, 2019.

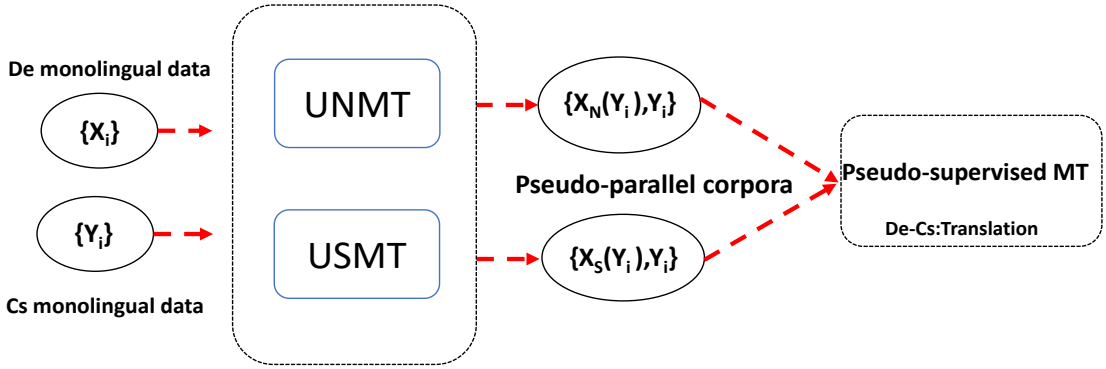

Figure 1: Our training framework. UNMT can generate the pseudo-parallel corpora $\{X_N(Y_i), Y_i\}$; USMT can generate the pseudo-parallel corpora $\{X_S(Y_i), Y_i\}$. These pseudo-parallel corpora were merged to train the pseudo-supervised MT system.

2.2 Tokenization, Truecasing, and Cleaning

We used `Moses` tokenizer (Koehn et al., 2007)[5] and truecaser for both languages. The truecaser was trained on one million tokenized lines extracted randomly from the monolingual data. Truecasing was then performed on all the tokenized data. For cleaning, we only applied the `Moses` script `clean-corpus-n.perl` to remove lines in the monolingual data containing more than 50 tokens, and replaced characters forbidden by `Moses`. Note that we did not perform any punctuation normalization.

3 Systems

Our entire system is illustrated in Figure 1.

3.1 Unsupervised NMT

To build competitive UNMT systems, we chose to rely on the Transformer-based UNMT initialized by a pre-trained cross-lingual language model (Lample and Conneau, 2019) since it had been shown to outperform UNMT initialized with word embeddings, in quality and efficiency. In order to limit the size of the vocabulary of the UNMT model, we segmented tokens in the training data into sub-word units via byte pair encoding (BPE) (Sennrich et al., 2016b). We determined 60k BPE operations jointly on the training monolingual data for German and Czech, and used a shared vocabulary for both languages with 60k tokens based on BPE.

We used 50M monolingual corpora to train a cross-lingual language model using `XLM`[6] in order to pre-train the UNMT model. We used the accumulate gradient method to train the language model on 1 GPU[7] to solve the out-of-memory problem caused by big batch. The accumulate size was set to 8. The other parameters for training the language model were set as listed in Table 1. Then we trained a Transformer-based UNMT model with the pre-trained cross-lingual language model using `XLM` toolkit.

```
--lgs 'cs-de' --mlm_steps
'cs,de' --emb_dim
1024 --n_layers 6
--n_heads 8 --dropout 0.1
--attention_dropout 0.1
--gelu_activation true
--batch_size 32 --bptt 256
--optimizer adam,lr=0.0001
```

Table 1: Parameters for training `XLM`.

The auto-encoder of UNMT architecture cannot learn useful knowledge without some constraints; it would merely become a copying task that learns to copy the input words one by one (Lample et al., 2018). To alleviate this issue, we utilized a denoising auto-encoder (Vincent et al., 2010), and added noise in the form of random token swapping in input sentences to improve the model learning ability (Hill et al., 2016; He et al., 2016).

The denoising auto-encoder acts as a language model that has been trained in one language and

[5] https://github.com/moses-smt/mosesdecoder

[6] https://github.com/facebookresearch/XLM

[7] NVIDIA @ Tesla @ P100 16Gb.

```
--lgs 'cs-de' --ae_steps
'cs,de' --bt_steps
'cs-de-cs,de-cs-de'
--word_shuffle 3
--word_dropout 0.1
--word_blank 0.1 --lambda_ae
'0:1,100000:0.1,300000:0'
--encoder_only false
--emb_dim 1024 --n_layers
6 --n_heads 8 --dropout
0.1 --attention_dropout
0.1 --gelu_activation
true --tokens_per_batch
2000 --batch_size 32
--bptt 256 --optimizer
adam_inverse_sqrt,beta1=0.9,
beta2=0.98,lr=0.0001
--eval_bleu true
```

Table 2: Parameters for training UNMT.

does not consider the final goal of translating across different languages. Therefore, back-translation (Sennrich et al., 2016a) was adapted to train a translation system in a true translation setting based on monolingual corpora. The pseudo-parallel sentence pairs generated by the model at the previous iteration is used to train the new translation model.

We used 50M monolingual corpora to train the UNMT model for 50000 iterations. The de-cs UNMT system was trained on 4 GPUs, with the parameters listed in Table 2.

3.2 Unsupervised SMT

Previous work has shown that USMT performs similarly or better than UNMT (Artetxe et al., 2018c). Marie and Fujita (2018b) has also shown that USMT can be used to train a standard NMT system to obtain significant improvements in translation quality while the whole training framework remains unsupervised.

We built USMT systems using a framework similar to the one proposed in Marie and Fujita (2018b). The first step of USMT consists in inducing a phrase table from the monolingual corpora. We first collected phrases of up to six tokens from the monolingual News Crawl corpora using `word2phrase`.[8] As phrases, we also considered all the token types in the corpora. Then, we selected the 300k most frequent phrases in the monolingual corpora to be used for inducing a phrase table. All possible phrase pairs are scored, as in Marie and Fujita (2018b), using bilingual word embeddings (BWE), and the 300 target phrases with the highest scores were kept in the phrase table for each source phrase. In total, the induced phrase table contains 90M phrase pairs. BWE of 512 dimensions were obtained using word embeddings trained with `fastText`[9] and aligned in the same space using unsupervised `Vecmap` (Artetxe et al., 2018b)[10] for this induction. In total four scores, to be used as features in the phrase table, for each of these phrase pairs were computed to mimic phrase-based SMT: forward and backward phrase and lexical translation probabilities. Then, the phrase table was plugged into a `Moses` system that was tuned on the development data using `KB-MIRA`. We performed four refinement steps to improve the system using at each step 3M synthetic parallel sentences generated by the forward and backward translation systems, instead of using only either forward (Marie and Fujita, 2018b) or backward translations (Artetxe et al., 2018c). We report on the performance of the systems obtained after the fourth refinement step.

3.3 Pseudo-supervised MT

As shown in Marie and Fujita (2018b), pseudo-parallel data generated by unsupervised MT can be directly used as training data to train a standard NMT system with a significantly better translation quality. We adopted the same strategy for our unsupervised systems. We generated pseudo-parallel corpora with our USMT and UNMT systems. Then we trained a Transformer-based NMT model (Vaswani et al., 2017) on these pseudo-parallel corpora. Since the pseudo-parallel corpora generated by USMT and UNMT are of very different nature, and that USMT and UNMT perform similarly in translation quality, we can expect that the complementarity of both data will be useful to train a better NMT system in contrast to using only data generated either by USMT or UNMT. Our synthetic parallel corpora for training this system was composed of 6M sentence pairs generated by USMT and 20M

[8] https://code.google.com/archive/p/word2vec/

[9] https://github.com/facebookresearch/fastText

[10] https://github.com/artetxem/vecmap

```
 --type transformer
--max-length  100
--transformer-dim-ffn  4096
--dim-vocabs  50000  50000
-w  12000  --mini-batch-fit
--valid-freq 5000 --save-freq
5000 --disp-freq 500
--valid-metrics  ce-mean-words
perplexity  translation
--quiet-translation
--sync-sgd --beam-size
12 --normalize=1
--valid-mini-batch 16
--keep-best  --early-stopping
20 --cost-type=ce-mean-words
--enc-depth 6 --dec-depth
6  --tied-embeddings
--transformer-dropout
0.1 --label-smoothing
0.1  --learn-rate  0.0003
--lr-warmup  16000
--lr-decay-inv-sqrt
16000 --lr-report
--optimizer-params  0.9
0.98  1e-09 --clip-norm  5
--exponential-smoothing
```

Table 3: Parameters for training `Marian`.

sentence pairs generated by UNMT. To train this pseudo-supervised NMT (PNMT) system, we chose `Marian` (Junczys-Dowmunt et al., 2018)[11] since it supports state-of-the-art features and is one of the fastest NMT frameworks publicly available. Specifically, the pseudo-supervised NMT system for de-cs was trained on 4 GPUs for 300,000 iterations, with the parameters listed by Table 3.

4 Combination of PNMT and USMT

Our primary submission for the task was the result of a simple combination of PNMT and USMT similarly to what we did last year in our participation to the supervised News Translation Task of WMT18 (Marie et al., 2018). As demonstrated by Marie and Fujita (2018a), and despite the simplicity of the method used, combining NMT and SMT makes MT more robust and can significantly improve translation quality, even though SMT greatly underperforms NMT. Following Marie and Fujita (2018a), our combination of PNMT and USMT works as follows.

4.1 Generation of n-best Lists

We first independently generated the 100-best and 12-best translation hypotheses[12] with N PNMT models, independently trained, and also with the ensemble of these N PNMT models. We also generated 100-best translation hypotheses with our USMT system. We then merged all these lists generated by different systems, without removing duplicated hypotheses, which resulted in a list of $(N+2)*100+(N+1)*12$ translation hypotheses for each source sentence. Finally, we rescored all the hypotheses in the list with a reranking framework using features to better model the fluency and the adequacy of each hypothesis. This method can find a better hypothesis in these merged n-best lists than the one-best hypothesis originated by the individual systems.

4.2 Reranking Framework and Features

We chose `KB-MIRA` (Cherry and Foster, 2012) as a rescoring framework and used a subset of the features proposed in Marie and Fujita (2018a). All the following features we used are described in details by Marie and Fujita (2018a). It includes the scores given by N PNMT models independently trained. We computed sentence-level translation probabilities using the lexical translation probabilities learned by `mgiza` during the training of our USMT system. We also used two 4-gram language models to compute two features for each hypothesis. One is the same language model used by our USMT system while the other is a small model trained on all the development data from which we removed the data used to train the reranking framework. To account for hypotheses length, we added the difference, and its absolute value, between the number of tokens in the translation hypothesis and the source sentence.

The reranking framework was trained on n-best lists generated by decoding the first 3k sentence pairs of the development data that we also used to validate the training of UNMT and PNMT systems and to tune the weights of USMT models.

[11] https://marian-nmt.github.io/

[12] We generated n-best with different beam size for decoding since translation quality can decrease with larger beam size (Koehn and Knowles, 2017).

#	Methods	de-cs
1	Single UNMT system	15.5
2	Single USMT system	11.1
3	Single NMT system pseudo-supervised by UNMT	15.9
4	Single NMT system pseudo-supervised by USMT	15.3
5	Single Pseudo-supervised MT system	16.2
6	Ensemble Pseudo-supervised MT system	16.5
7	Re-ranking Pseudo-supervised MT system	17.0
8	Fine-tuning Pseudo-supervised MT system	18.7
9	Fine-tuning Pseudo-supervised MT system + fixed quotes	19.6
10	Fine-tuning + re-ranking Pseudo-supervised MT system + fixed quotes	20.1

Table 4: BLEU scores of UMT. #10 is our primary system submitted to the organizers.

5 Fine-tuning and Post-processing

Fine-tuning (Luong and Manning, 2015; Sennrich et al., 2016a) is a conventional method for NMT on low-resource language pairs and domain-specific tasks (Chu et al., 2017; Chu and Wang, 2018; Wang et al., 2017a,b). The PNMT model only relying on monolingual corpora was further trained on the parallel development data to improve translation performance. Finally, fixed quotes method was applied to the final Czech translation.

6 Results on the German-to-Czech Task

The results of our systems computed for the Newstest2019 test set are presented in Table 4. As Table 4 shows, UNMT systems significantly outperformed our best USMT system according to BLEU. However, compared with pseudo-supervised MT model trained only on pseudo-parallel corpora generated by either UNMT (#3) or USMT (#4), merging pseudo-parallel corpora generated by UNMT and USMT (#5) can improve translation performance. Reranking `Moses` 100-best hypotheses using PNMT models (#7) significantly improved the translation quality. Another methods such as ensemble, fine-tuning, and fixed quotes also could improve translation performance.

7 Contrastive Experiments on English-Gujarati and English-Kazakh

To obtain a better picture of the feasibility of unsupervised MT, we also set up unsupervised MT for two truly low-resource and distant language pairs: English-Gujarati (en-gu) and English-Kazakh (en-kk).[13] As shown by previous work (Søgaard et al., 2018), we can expect unsupervised word embeddings to be challenging to train for distant language pairs, and subsequently to obtain unsupervised MT systems with a very poor translation quality.

Note that for these experiments, we did not train any UNMT systems. We present results only for USMT and NMT pseudo-supervised by USMT. Since training unsupervised BWE for these language pairs is particularly challenging, we also present configurations using supervised BWE trained using the approach described by Artetxe et al. (2018a) on a bilingual word lexicon extracted from the development data provided by the organizers. Our configuration of USMT and PNMT are the same as for de-cs.

As English training data, we only used all the provided News Crawl corpora as they are large in-domain corpora. For Gujarati and Kazakh, we used Common Crawl and News Crawl corpora, in addition to the provided News Commentary corpus for Kazakh. Statistics of the data preprocessed with `Moses` are presented in Table 5.

Our results are presented in Table 6. In contrast to what we observed for de-cs, unsupervised BWE are too noisy to be used in phrase table induction for USMT. For both en-gu and en-kk, we obtained unexploitable results confirming the conclusions of Søgaard et al. (2018).

Switching to supervised BWE improved significantly the translation quality of USMT but

[13] These language pairs were proposed for the supervised News Translation Task.

Corpus		en-gu		en-kk	
		en	gu	en	kk
Monolingual	#lines	187.50M	3.39M	187.50M	9.03M
	#tokens	4.39B	50.52M	4.39B	141.06M
Development	#lines	1,998	1,998	2,066	2,066
	#tokens	42,264	38,963	53,451	42,910

Table 5: Statistics of preprocessed monolingual and development data used for en-gu and en-kk.

System		en-gu		en-kk	
		→	←	→	←
Unsupervised BWE	USMT	< 1.0	< 1.0	< 1.0	< 1.0
Supervised BWE	USMT	5.7	6.2	1.4	4.7
	Pseudo-supervised NMT	8.1	8.8	2.1	5.7
Supervised NMT		10.5	17.2	6.4	26.2

Table 6: BLEU scores of our USMT and NMT pseudo-supervised by USMT systems. Note that we did not conduct experiments with pseudo-supervised NMT using USMT initialized with unsupervised BWE as the generated pseudo-parallel data were not useful to train a NMT system at all. The results of our supervised systems (last row) submitted for the News Translation Task are presented for comparison.

remains below 10 BLEU points in all our experiments. Compared with our best supervised system, the difference in translation quality appears very large especially when translating into English.

These results show that while we obtained a reasonable translation quality for de-cs, unsupervised MT is far from being useful for real world applications, i.e., truly low-resource distant language pairs. Training useful bilingual weakly-supervised/unsupervised BWE for distant language pairs remains one of the main challenges.

8 Conclusion

We participated in the unsupervised translation direction and compared USMT and UNMT performances. We achieved the best results through the combination of both approaches thanks to an NMT framework pseudo-supervised by UNMT and USMT. We also showed that reranking of the n-best lists in this unsupervised settings can bring additional improvements in translation quality. While we achieved a reasonable translation quality for German-to-Czech, a language pair for which there exists plenty of bilingual data, our results for English-Gujarati and English-Kazakh highlighted that unsupervised machine translation is still very far from exploitable for low-resource distant language pairs.

Acknowledgments

We thank the organizers for providing the datasets and the reviewers for their valuable suggestions for improving this paper. This work was conducted under the program "Research and Development of Enhanced Multilingual and Multipurpose Speech Translation Systems" of the Ministry of Internal Affairs and Communications (MIC), Japan. Rui Wang was partially supported by JSPS grant-in-aid for early-career scientists (19K20354): "Unsupervised Neural Machine Translation in Universal Scenarios."

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018a. Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 5012–5019, New Orleans, Louisiana, USA. AAAI Press.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018b. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages

789–798, Melbourne, Australia. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018c. Unsupervised statistical machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3632–3642, Brussels, Belgium. Association for Computational Linguistics.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Christof Monz, Mathias Müller, and Matt Post. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation, Volume 2: Shared Task Papers*, Florence, Italy. Association for Computational Linguistics.

Colin Cherry and George Foster. 2012. Batch tuning strategies for statistical machine translation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 427–436, Montréal, Canada. Association for Computational Linguistics.

Chenhui Chu, Raj Dabre, and Sadao Kurohashi. 2017. An empirical comparison of domain adaptation methods for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 385–391, Vancouver, Canada. Association for Computational Linguistics.

Chenhui Chu and Rui Wang. 2018. A survey of domain adaptation for neural machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1304–1319, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. 2016. Dual learning for machine translation. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016*, pages 820–828, Barcelona, Spain. Curran Associates, Inc.

Felix Hill, Kyunghyun Cho, and Anna Korhonen. 2016. Learning distributed representations of sentences from unlabelled data. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1367–1377, San Diego California, USA. 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.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej 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.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver, Canada. Association for Computational Linguistics.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *CoRR*, abs/1901.07291.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Unsupervised machine translation using monolingual corpora only. In *Proceedings of the Sixth International Conference on Learning Representations*, Vancouver, Canada. OpenReview.net.

Minh-Thang Luong and Christopher D. Manning. 2015. Stanford neural machine translation systems for spoken language domain. In *International Workshop on Spoken Language Translation*, Da Nang, Vietnam.

Benjamin Marie and Atsushi Fujita. 2018a. A smorgasbord of features to combine phrase-based and neural machine translation. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Papers)*, pages 111–124, Boston, MA. Association for Machine Translation in the Americas.

Benjamin Marie and Atsushi Fujita. 2018b. Unsupervised neural machine translation initialized by unsupervised statistical machine translation. *CoRR*, abs/1810.12703.

Benjamin Marie, Rui Wang, Atsushi Fujita, Masao Utiyama, and Eiichiro Sumita. 2018. NICT's neural and statistical machine translation systems for the WMT18 news translation task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 449–455, Belgium, Brussels. 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.

Anders Søgaard, Sebastian Ruder, and Ivan Vulić. 2018. On the limitations of unsupervised bilingual dictionary induction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 778–788, Melbourne, Australia. 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 30: Annual Conference on Neural Information Processing Systems 2017*, pages 6000–6010, Long Beach, CA, USA. Curran Associates, Inc.

Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. 2010. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, 11:3371–3408.

Rui Wang, Andrew Finch, Masao Utiyama, and Eiichiro Sumita. 2017a. Sentence embedding for neural machine translation domain adaptation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 560–566, Vancouver, Canada. Association for Computational Linguistics.

Rui Wang, Masao Utiyama, Lemao Liu, Kehai Chen, and Eiichiro Sumita. 2017b. Instance weighting for neural machine translation domain adaptation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1482–1488, Copenhagen, Denmark. Association for Computational Linguistics.