NRC Machine Translation System for WMT 2017

Chi-kiu Lo **Boxing Chen** **Colin Cherry** **George Foster***
Samuel Larkin **Darlene Stewart** **Roland Kuhn**
National Research Council Canada
1200 Montreal Road, Ottawa, ON K1A 0R6, Canada
FirstName.LastName@nrc-cnrc.gc.ca

Google Research
Montreal, Canada
fosterg@google.com

Abstract

We describe the machine translation systems developed at the National Research Council of Canada (NRC) for the Russian-English and Chinese-English news translation tasks of the Second Conference on Machine Translation (WMT 2017). We conducted several experiments to explore the best baseline settings for neural machine translation (NMT). In the Russian-English task, to our surprise, our best-performing system is one that rescores phrase-based statistical machine translation outputs using NMT rescoring features. On the other hand, in the Chinese-English task, which has far more parallel training data, NMT is able to outperform SMT significantly. The NRC MT systems is the best constrained system in Russian-English (out of nine participants) and the fourth best constrained system in Chinese-English (out of twenty participants) in WMT 2017 human evaluation.

1 Introduction

We present NRC's submission to the Russian-English and Chinese-English news translation task of WMT 2017. In contrast to last year, when we participated in the Russian-English task only, with our well-developed phrase-based statistical machine translation system (Lo et al., 2016; Larkin et al., 2010; Foster et al., 2009), this year we built large-scale state-of-the-art neural machine translation (NMT) systems for these two language pairs to facilitate further understanding and discussion of NMT.

Russian-English and Chinese-English are both challenging language pairs for machine translation. Russian is a highly inflectional and free word order language. The skewed Russian to English word type ratio introduces a data sparsity problem that cannot be solved by discarding word inflections, since they play an important role in disambiguating the meaning of sentences. Chinese does not have clear word boundaries. The number of Chinese word types created by automatic word segmentation software is high, while naive character segmentation would result in a skewed Chinese to English sentence length ratio. These characteristics make it difficult for machine translation systems to learn the correct association between words in Chinese and English.

Since this was the first time we deployed NMT models in an evaluation, we first tried to replicate the results of previous work (Sennrich et al., 2016a). Our NMT systems are based on Nematus (Sennrich et al., 2017). We used automatic back-translation (Sennrich et al., 2016b) of a subselected monolingual News corpus as additional training data, and all the training data is segmented into subword units using BPE (Sennrich et al., 2016c). We also experimented with pervasive dropout as implemented in Nematus.

For Russian-English, our WMT16 PBMT system scored higher than all the NMT systems we built this year. We therefore experimented with using the NMT systems as features for rescoring the 1000-best output from our WMT16 PBMT system. This strategy yielded almost 2 BLEU point improvement over the PBMT baseline. For Chinese-English, we exploited different domain adaptation techniques to boost the system performance on in-domain news translation. We also integrated various regularization methods to avoid the systems overfitting to the small development set.

The NRC Russian-English and Chinese-English news translation systems achieve competitive per-

*Work performed while at NRC.

Proceedings of the Conference on Machine Translation (WMT), Volume 2: Shared Task Papers, pages 330–337
Copenhagen, Denmark, September 711, 2017.

formance (third place in both language pairs) in the preliminary automatic evaluation of WMT 2017. In this paper, we discuss the lessons learned in building large-scale state-of-the-art NMT systems.

2 Russian-English news translation

We used all the Russian-English parallel corpora available for the constrained news translation task. They include the CommonCrawl corpus, the NewsCommentary v12 corpus, the Yandex corpus and the Wikipedia headlines corpus. In total, 2.6 million parallel Russian-English sentences are used to train the baseline system. We use the news translation test set of WMT 15 as development set and that of WMT 16 as test set. The Russian and English texts in the training/development/test corpora were kept in their original true case and tokenized, then the Russian and English texts were combined to train a BPE model with vocabulary size of 30k.

2.1 NMT baseline system

Our NMT baseline system is developed using Nematus (Sennrich et al., 2017). The dimension of word embeddings is set to 512 and that of the hidden layers is set to 1024. We train the models with rmsprop (Tieleman and Hinton, 2012), reshuffling the training corpus between epochs. We use minibatches of size 100 and validate the model every 8000 minibatches against BLEU on the WMT 15 news translation test set. We perform early stopping on the baseline system. We use AmuNMT C++ decoder (Junczys-Dowmunt et al., 2016a) with a beam size of 4.

2.2 Synthetic training data

In statistical machine translation, large monolingual corpora in the output language have traditionally been used for training language models to make the system output more fluent. However, it is difficult to integrate language models in current NMT architectures. Instead of ignoring such large monolingual corpora, Sennrich et al. (2016b) exploited large corpora in the output language by translating a subset of them into the input language and then using the resulting synthetic sentence pairs as additional training data. We translated monolingual English text into Russian using an English to Russian NMT system mirroring the one described in Section 2.1,[1] and then employed the machine-translated Russian and perfect English sentence pairs as additional data to train the Russian-English MT system.

To select sentences for back-translation, we used a semi-supervised convolutional neural network classifier (Chen and Huang, 2016). We sampled two million sentences from the English monolingual News Crawl 2015 & 2016 corpora according to their classifier scores, which reflect their similarity to the the English half of our development set.

2.3 Pervasive dropout

Pervasive dropout prevents the NMT system from overfitting. We apply the variant of Gal and Ghahramani (2016) pervasive dropout that is implemented in Nematus to all layers in the network. This variant has the characteristic that the random dropout is applied at the token level, instead of at the word-type level. We set the dropout probability for the source words, target words and embedding layers to 0.15. For the hidden layers, we set the dropout probability to 0.3.

2.4 Minimum risk training

Minimum risk training (MRT) (Shen et al., 2016) allows model optimization to arbitrary loss functions, which do not need to be differentiable, thus enabling direct model tuning against automatic MT evaluation metrics. It uses the MT evaluation metric as the loss function and minimizes the expected loss on the training data at sentence-level. We experimented with further model optimization using MRT on the whole training corpus against sentence BLEU at the final stage.

2.5 Greedy model averaging

A common practice for avoiding overfitting to the training data is ensembling the last few models saved as checkpoints. Recently, Junczys-Dowmunt et al. (2016b) showed that one can see nearly the same benefits by performing a component-wise average of all parameters across checkpoints. We extended this technique by using a greedy strategy to average a wider range of models. Instead of considering only the last few saved models, we considered 30 saved models having the best BLEU performance on the validation set one-by-one. For each checkpoint, in descending

[1] This scores 21.05 BLEU on the WMT 15 test set.

order of BLEU score, we add the checkpoint to our running average to create a model candidate. We then use the candidate to decode our development set. If this results in improved BLEU, we accept the candidate, and it becomes our new running average. We find that this process generally selects between 5 and 8 checkpoints to include in the average.

2.6 Portage - NRC WMT16 PBMT system

The core of the NRC WMT16 MT system (Lo et al., 2016) is *Portage* (Larkin et al., 2010). Portage is a conventional log-linear phrase-based SMT system.

The system was trained on all the Russian-English parallel training corpora and WMT 12 and WMT 13 Russian-English news translation test set and tuned on the WMT 14 test set. Both the Russian and English text in the parallel and monolingual corpora in the training/development/test corpora were tokenized and lowercased.

The system employed Russian lemmatization extensively in building word alignments for translation models, a hierarchical distortion model, a sparse feature model and neural network joint models or NNJMs (Devlin et al., 2014). The system also made extensive use of monolingual English corpora in building language models. Last but not least, it had comprehensive Russian OOV handling, which included a fallback Russian lemma-based phrase table and a Russian transliteration model.

2.7 Rescoring and truecasing

We rescored 1000-best lists output from the phrase-based decoder using a rescoring model (Och et al., 2004; Foster et al., 2009) consisting of 13 features: 3 NMT models, 2 language models, 5 NNJMs and 3 n-best features. The rescoring model was tuned using n-best MIRA (Cherry and Foster, 2012).

The three NMT systems used as rescoring features were: 1) baseline further trained with synthetic data, 2) dropout baseline further trained with synthetic data and with dropout turned off, and 3) the previous model optimized to the development set using minimum risk training.

The five NNJM rescoring features include two Russian-word NNJMs and three Russian-lemma ones. Following Devlin et al. (2014), we take advantage of the rescoring framework to have our NNJMs view each candidate translation from perspectives not available during decoding. The Russian-lemma NNJMs are rescored using normal, target-to-source, and right-to-left perspectives. The Russian-word NNJMs are rescored using normal and right-to-left perspectives. The choice of which perspectives to include was made based on empirical devtest (WMT 16) performance.

The two language models were: a left-to-right 6-gram LM and a right-to-left 6-gram LM. Both were trained on the WMT 16 monolingual English training corpus,

The final output was truecased and detokenized in the same way as described in Lo et al. (2016).

	dev		test
System	single best	ave.	
a: baseline	23.6	24.8	23.8
b: (a)+synthetic	25.6	26.3	25.3
c: dropout baseline	26.3	26.3	25.6
d: (c)+synthetic	27.7	27.8	26.6
e: (d)+mrt	27.8	27.8	26.1
f: WMT16 Portage	28.2	–	28.6
g: (f) rescored by (d)	29.9	–	29.6
h: (f) final rescoring	–	–	30.4

Table 1: Selected results from our Russian-English development experiments. The ave. column shows the result of greedy model averaging, where applicable.

2.8 Results

Table 1 shows the results of selected models from our development experiments. It can be seen that synthetic training data generated by back-translation of large output language monolingual corpora consistently improves the baseline by 1.4 to 2 BLEU. However, this result is rather disappointing by comparison with the exciting improvement reported in Sennrich et al. (2016a), i.e. 3-4 BLEU.

Another disappointing result is that model averaging does not work well with the dropout models. We can see model averaging yields around 1 BLEU gain on non-dropout systems. However, the improvement achieved by model averaging drops to 0-0.1 BLEU on dropout systems. In other experiments not shown here, we also saw no improvement from ensembling the checkpoints of our dropout systems.

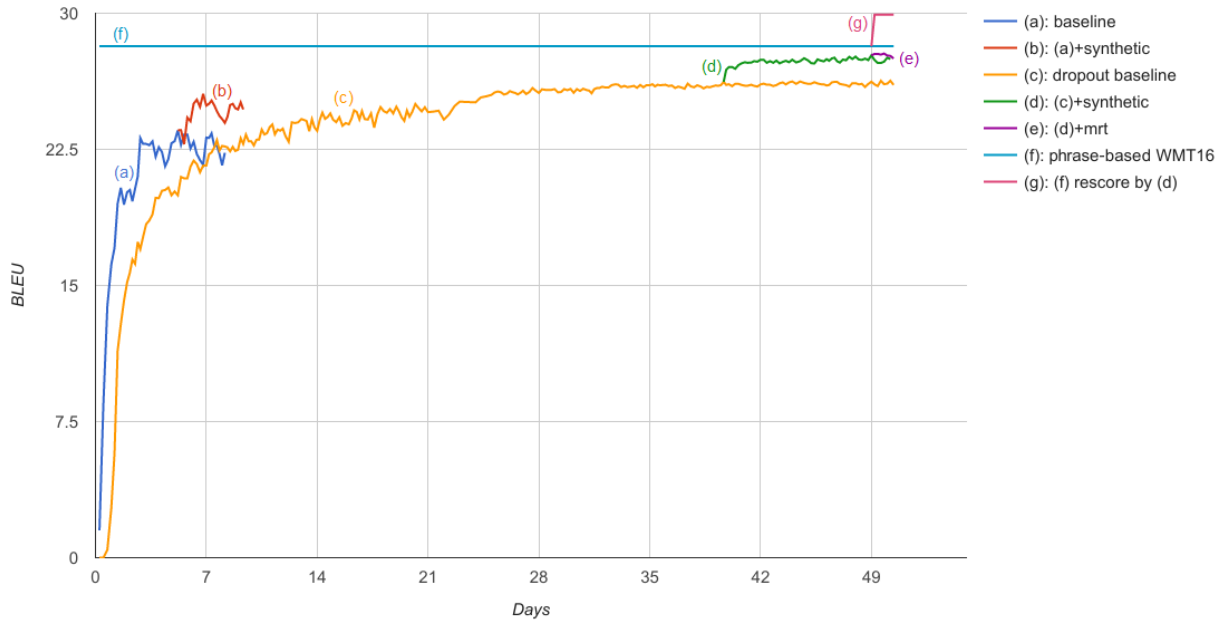

Figure 1: Russian-English learning curve on development set in cased BLEU of selected models: a) NMT baseline, b) NMT baseline further trained with synthetic data, c) NMT dropout baseline, d) NMT dropout baseline further trained with synthetic data while dropout is turned off, e) NMT dropout baseline with synthetic data optimized to sentence-level BLEU on the training data using MRT, f) our WMT16 PBMT submission and g) the PBMT rescored by one of the rescoring features.

The most interesting observation in our experiments is that the dropout baseline continues to improve over the course of many weeks. Figure 1 shows the learning curve of the selected models for all 7 weeks of development. Line (c) of this figure shows that the dropout baseline continues training and improving until the end of the evaluation campaign, achieving a development BLEU score that is 2.7 BLEU points beyond our best single NMT system that does not use dropout. This system can be further improved by adding synthetic data, as in line (d), however, we found that we needed to switch dropout off after adding the synthetic data.

Although in figure 1 we see that none of the NMT systems manage to beat our WMT16 PBMT submission, the more interesting result is that there is more than 1.8 BLEU gain on the development set and 1.1 BLEU gain on the test set by rescoring the PBMT 1000-best list using just one of our NMT systems and no other features, as in line (g). The final rescoring with weighted collections of NMT systems, language model features, NNJM features and n-best features shows 1.8 BLEU improvement over the WMT 16 submission on the test set.

3 Chinese-English news translation

We used all the Chinese-English parallel corpora available for the constrained news translation task. They include the UN corpus, the NewsCommentary v12 corpus and the CWMT corpus. In total, 25 million parallel Chinese-English sentences were used to train the baseline system. We used half of the WMT 17 news translation development set as our development set and the other half as internal test set. The English texts in the training/development/test corpora were tokenized and lowercased while the Chinese texts in the training/development/test corpora were segmented using the ICTCLAS segmenter (Zhang et al., 2003). Then the Chinese and English text were combined to train a BPE model with vocabulary size of 90k.

3.1 NMT baseline system

Our Chinese-English NMT baseline system is similar to the Russian-English baseline as described in section 2.1: Nematus-based, word embeddings with 512 dimensions, hidden layers with 1024 dimensions, etc.

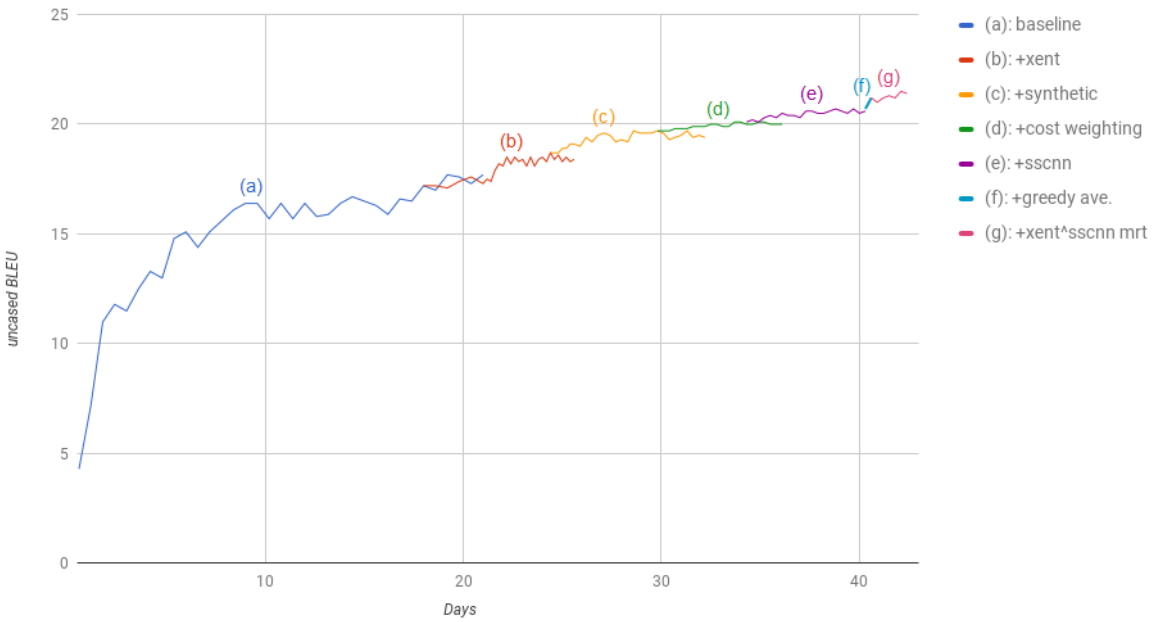

Figure 2: Chinese-English learning curves on the internal test set in uncased BLEU for selected models: a) NMT baseline, b) further trained with in-domain data selected by bilingual LM cross-entropy difference (xent), c) further trained with synthetic data, d) further trained with cost weighting, e) further trained with in-domain data selected by semi-supervised convolutional neural network classifier (sscnn), f) greedy model averaging and g) optimized against sentence-level BLEU on the intersection of the subsets of data selected by xent and sscnn using MRT.

3.2 Data selection and domain adaptation

Since the majority of the 25 million sentence pairs in the training corpus are general domain, we experimented with different data selection and domain adaptation techniques to further train the NMT system with data that are similar to the development set so as to perform better in the news domain.

Axelrod et al. (2011) introduced the bilingual language model cross-entropy difference as a similarity function for identifying sentence pairs from general-domain training corpora that are close to the target domain. We built four language models using the input and output sides of the training corpora and the development set respectively to select 3 million sentence pairs from the training corpora that are close to the news domain.

However, the development set, which consists of only 1k sentence pairs, is too tiny to be a suitable corpus for building the in-domain language models that will enable the bilingual LM cross-entropy difference data selection method to work effectively. Therefore, we also experimented with the semi-supervised convolutional neural network method in Chen and Huang (2016) to select 1 million sentence pairs from the training corpora that are close to the news domain.

Finally, we experimented with a cost weighting domain adaptation technique (Chen et al., 2017). This technique trains a domain classifier concurrently with the NMT system, and uses the classifier probabilities to weight training instances according to their similarity to the development set.

3.3 Synthetic training data

We generated synthetic Chinese and perfect English sentence pairs in a process similar to that described in section 2.2. We first used a semi-supervised convolutional neural network classifier (Chen and Huang, 2016) to sample 20 million sentences from the English monolingual News Crawl 2015 & 2016 corpora according to the development set. We then translated the selected sentences using a English-Chinese NMT baseline trained out-of-the-box using only the parallel corpora.

3.4 Greedy model averaging

Greedy model averaging is performed as described in section 2.5.

System	test
baseline	17.2
+biLM cross-entropy (xent) DS	18.7
+synthetic	19.7
+cost weighting DA	20.1
+sscnn DS	20.7
+greedy model averaging	21.2
+xent∩sscnn mrt	21.4
ensemble	24.2
rescoring	25.6

Table 2: Selected results in uncased BLEU from our Chinese-English development experiments.

3.5 Minimum risk training

In contrast to the way in which we carried out MRT for the Russian-English system in section 2.4, we optimized the Chinese-English system using MRT against sentence BLEU only on the intersection of the subsets of corpora selected by the LM cross-entropy and the semi-supervised CNN in section 3.2. The size of the intersection of the two subsets of corpora is 300k sentence pairs.

3.6 Ensembling, rescoring and truecasing

Applying different combinations of the techniques described in section 3.1 to 3.5, we built 14 different NMT systems. Their uncased BLEU on the test set ranged from 19.8 to 21.4. We ensembled all the systems together using Simplex-tuned weights.

We rescored 500-best lists output from the ensembled NMT system using a rescoring model (Och et al., 2004; Foster et al., 2009) consisting of 82 features: IBM models, RNN language models (Mikolov et al., 2010), n-gram language models trained on different data subsets, neural network joint models (NNJMs) (Devlin et al., 2014) and word, n-gram, word alignment posteriors (Foster et al., 2009), etc. The rescoring model was tuned using n-best MIRA (Cherry and Foster, 2012).

The final output was truecased and detokenized using heuristic methods.

3.7 Results

Figure 2 shows that all the components we described in section 3.2 to 3.5 help improve the NMT system. The uncased BLEU on the test set in table 2 shows that all the data selection and domain adaptation methods improve the NMT systems by 0.4 to 1.5 BLEU. Similar to the results we observed in our Russian-English NMT systems, synthetic training data generated by back-translation of large output language monolingual corpora improved the NMT system score by 1 BLEU.

The most important observation in our experiments is that ensembling of NMT systems developed by different techniques achieves around 3 BLEU improvement and rescoring the n-best output from NMT systems also shows 1.4 BLEU gain on the test set.

4 Conclusion

We have presented the NRC submissions to the WMT 2017 Russian-English and Chinese-English news translation task. The Russian-English submitted system is our WMT 16 PBMT system rescored by three NMT models and other rescoring features. Our Chinese-English submitted system is an ensemble of fourteen NMT models rescored by a large set of additional features. Our system achieved the highest score for the Russian-English (among nine participants) and the fourth highest score for Chinese-English (among twenty participants) constrained news translation tasks in WMT 2017 human evaluation.

Our experiences in WMT 2017 illustrate the sharp divide between large- and medium-scale data scenarios when working with neural MT. For Russian-English, we found ourselves relying on techniques that are usually intended for low-resource scenarios, such as pervasive dropout and rescoring a phrase-based system. This is surprising, as 2.5 million sentence pairs would have been considered a large-data scenario in the not-too-distant past. Meanwhile, for Chinese-English, we were able to achieve strong individual neural systems, which were further strengthened by ensembling across various data selection and data weighting techniques. Our results also highlight the necessity to speed up convergence in the presence of dropout, so that it does not take weeks to train a single model.

References

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain Adaptation via Pseudo In-Domain Data Selection. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Edinburgh, Scotland, UK., pages 355–362. http://www.aclweb.org/anthology/D11-1033.

Boxing Chen, Colin Cherry, George Foster, and Samuel Larkin. 2017. Cost Weighting for Neural Machine Translation Domain Adaptation . In *1st Workshop of Neural Machine Translation*.

Boxing Chen and Fei Huang. 2016. Semi-supervised Convolutional Networks for Translation Adaptation with Tiny Amount of In-domain Data. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, CoNLL 2016, Berlin, Germany, August 11-12, 2016*. pages 314–323. http://aclweb.org/anthology/K/K16/K16-1031.pdf.

Colin Cherry and George Foster. 2012. Batch Tuning Strategies for Statistical Machine Translation. In *Proc. 2012 Conf. of the N. American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Montréal, Canada, pages 427–436. http://www.aclweb.org/anthology/N12-1047.

Jacob Devlin, Rabih Zbib, Zhongqiang Huang, Thomas Lamar, Richard Schwartz, and John Makhoul. 2014. Fast and Robust Neural Network Joint Models for Statistical Machine Translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. Baltimore, Maryland, pages 1370–1380. http://www.aclweb.org/anthology/P14-1129.

George Foster, Boxing Chen, Eric Joanis, Howard Johnson, Roland Kuhn, and Samuel Larkin. 2009. PORTAGE in the NIST 2009 MT Evaluation. *Technical report, NRC-CNRC.* .

Yarin Gal and Zoubin Ghahramani. 2016. A Theoretically Grounded Application of Dropout in Recurrent Neural Networks. In *Advances in Neural Information Processing Systems 29 (NIPS)*.

Marcin Junczys-Dowmunt, Tomasz Dwojak, and Hieu Hoang. 2016a. Is Neural Machine Translation Ready for Deployment? A Case Study on 30 Translation Directions. In *Proceedings of the 9th International Workshop on Spoken Language Translation (IWSLT)*. Seattle, WA.

Marcin Junczys-Dowmunt, Tomasz Dwojak, and Rico Sennrich. 2016b. The AMU-UEDIN Submission to the WMT16 News Translation Task: Attention-based NMT Models as Feature Functions in Phrase-based SMT. In *Proceedings of the First Conference on Machine Translation*. Association for Computational Linguistics, Berlin, Germany, pages 319–325. http://www.aclweb.org/anthology/W16-2316.

Samuel Larkin, Boxing Chen, George Foster, Uli Germann, Eric Joanis, J. Howard Johnson, and Roland Kuhn. 2010. Lessons from NRC's Portage System at WMT 2010. In *5th Workshop on Statistical Machine Translation*.

Chi-kiu Lo, Colin Cherry, George Foster, Darlene Stewart, Rabib Islam, Anna Kazantseva, and Roland Kuhn. 2016. NRC Russian-English Machine Translation System for WMT 2016. In *Proceedings of the First Conference on Machine Translation*. Association for Computational Linguistics, Berlin, Germany, pages 326–332. http://www.aclweb.org/anthology/W16-2317.

Tomas Mikolov, Martin Karafit, Luks Burget, Jan Cernock, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *INTERSPEECH*. ISCA, pages 1045–1048.

Franz Josef Och, Daniel Gildea, Sanjeev Khudanpur, Anoop Sarkar, Kenji Yamada, Alexander M Fraser, Shankar Kumar, Libin Shen, David Smith, Katherine Eng, Viren Jain, Zhen Jin, and Radev Dragomir. 2004. A Smorgasbord of Features for Statistical Machine Translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*. pages 161–168.

Rico Sennrich, Orhan Firat, Kyunghyun Cho, Alexandra Birch, Barry Haddow, Julian Hitschler, Marcin Junczys-Dowmunt, Samuel Läubli, Antonio Valerio Miceli Barone, Jozef Mokry, and Maria Nadejde. 2017. Nematus: a Toolkit for Neural Machine Translation. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Valencia, Spain, pages 65–68. http://aclweb.org/anthology/E17-3017.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Edinburgh Neural Machine Translation Systems for WMT 16. In *Proceedings of the First Conference on Machine Translation*. Association for Computational Linguistics, Berlin, Germany, pages 371–376. http://www.aclweb.org/anthology/W16-2323.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. 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)*. Association for Computational Linguistics, Berlin, Germany, pages 86–96. http://www.aclweb.org/anthology/P16-1009.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016c. 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)*. Association for Computational Linguistics, Berlin, Germany, pages 1715–1725. http://www.aclweb.org/anthology/P16-1162.

Shiqi Shen, Yong Cheng, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2016. Minimum Risk Training for Neural Machine Translation. In *Proceedings of the 54th Annual Meeting of*

the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Berlin, Germany, pages 1683–1692. http://www.aclweb.org/anthology/P16-1159.

Tijmen Tieleman and Geoffrey Hinton. 2012. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. In *COURSERA: Neural networks for machine learning*.

Hua-Ping Zhang, Hong-Kui Yu, De-Yi Xiong, and Qun Liu. 2003. HHMM-based Chinese Lexical Analyzer ICTCLAS. In *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing*. Association for Computational Linguistics, Sapporo, Japan, pages 184–187.