system	BLEU
MIXER (Ranzato et al., 2016)[5]	21.8
BSO (Wiseman and Rush, 2016)	25.5
NPMT+LM (Huang et al., 2018)	30.1
MRT (Edunov et al., 2018)	32.84 ± 0.08
Pervasive Attention (Elbayad et al., 2018)	33.8
Transformer Baseline (Wu et al., 2019)	34.4
Dynamic Convolution (Wu et al., 2019)	35.2
our PBSMT (1)	28.19 ± 0.01
our NMT baseline (2)	27.16 ± 0.38
our NMT best (7)	35.27 ± 0.14

Table 3: Results on full IWSLT14 German→English data on tokenized and lowercased test set with *multi-bleu.perl*.

system	BLEU
(Gu et al., 2018b) (supervised Transformer)	5.97
phrase-based SMT	6.57 ± 0.17
NMT baseline (2)	2.93 ± 0.34
NMT optimized (8)	**10.37** ± 0.29

Table 4: Korean→English results. Mean and standard deviation of three training runs reported.

to other data conditions, and Korean→English, for simplicity.

For a comparison with PBSMT, and across different data settings, consider Figure 2, which shows the result of PBSMT, our NMT baseline, and our optimized NMT system. Our NMT baseline still performs worse than the PBSMT system for 3.2M words of training data, which is consistent with the results by Koehn and Knowles (2017). However, our optimized NMT system shows strong improvements, and outperforms the PBSMT system across all data settings. Some sample translations are shown in Appendix B.

For comparison to previous work, we report lowercased and tokenized results on the full IWSLT 14 training set in Table 3. Our results far outperform the RNN-based results reported by Wiseman and Rush (2016), and are on par with the best reported results on this dataset.

Table 4 shows results for Korean→English, using the same configurations (1, 2 and 8) as for German–English. Our results confirm that the techniques we apply are successful across datasets, and result in stronger systems than previously reported on this dataset, achieving 10.37 BLEU as compared to 5.97 BLEU reported by Gu et al. (2018b).

6 Conclusions

Our results demonstrate that NMT is in fact a suitable choice in low-data settings, and can outperform PBSMT with far less parallel training data than previously claimed. Recently, the main trend in low-resource MT research has been the better exploitation of monolingual and multilingual resources. Our results show that low-resource NMT is very sensitive to hyperparameters such as BPE vocabulary size, word dropout, and others, and by following a set of best practices, we can train competitive NMT systems without relying on auxiliary resources. This has practical relevance for languages where large amounts of monolingual data, or multilingual data involving related languages, are not available. Even though we focused on only using parallel data, our results are also relevant for work on using auxiliary data to improve low-resource MT. Supervised systems serve as an important baseline to judge the effectiveness of semisupervised or unsupervised approaches, and the quality of supervised systems trained on little data can directly impact semi-supervised workflows, for instance for the back-translation of monolingual data.

Acknowledgments

Rico Sennrich has received funding from the Swiss National Science Foundation in the project CoNTra (grant number 105212_169888). Biao Zhang acknowledges the support of the Baidu Scholarship.