The Samsung and University of Edinburgh's submission to IWSLT17

Pawel Przybysz[1], Marcin Chochowski[1], Rico Sennrich[2], Barry Haddow[2] and Alexandra Birch[2]

[1]Samsung R&D Institute Poland
[2]School of Informatics, University of Edinburgh
{m.chochowski,p.przybysz}@samsung.com
bhaddow@inf.ed.ac.uk, {rico.sennrich,a.birch}@ed.ac.uk

Abstract

This paper describes the joint submission of Samsung Research and Development, Warsaw, Poland and the University of Edinburgh team to the IWSLT MT task for TED talks. We took part in two translation directions, en-de and de-en. We also participated in the en-de and de-en lectures SLT task. The models have been trained with an attentional encoder-decoder model using the BiDeep model in Nematus. We filtered the training data to reduce the problem of noisy data, and we use back-translated monolingual data for domain-adaptation. We demonstrate the effectiveness of the different techniques that we applied via ablation studies. Our submission system outperforms our baseline, and last year's University of Edinburgh submission to IWSLT, by more than 5 BLEU.

1. Introduction

This paper describes the system submission of Samsung R&D Institute Poland and the University of Edinburgh team. The models have been trained with a deep attentional encoder-decoder neural machine translation model using Nematus [1]. In this year's submission, we focused on the core NMT architecture, for which we selected the BiDeep model by [2], training data filtering via language identification and MT-based sentence alignment scores, and domain adaptation with back-translated, domain-filtered monolingual training data, and fine-tuning towards the in-domain training set with MAP-L2 regularization towards the baseline model [3].

Corpus	raw	aligned	filtered
Commoncrawl [4]	2.40M	2.22M	1.62M
Europarl v7 [5]	1.92M	1.90M	1.85M
GoldAligment	509	508	486
MultiUN [6]	0.16M	0.16M	0.15M
News Com. v12 [4]	0.27M	0.26M	0.26M
Opensubtitles2016 [7]	13.88M	12.08M	9.04M
QED Corpus	0.07M	0.07M	0.06M
Rapid 2016	1.33M	1.28M	1.12M
Wikipedia Corpus	2.46M	2.16M	1.18M
WIT3 (in-domain) [8]	0.22M	0.21M	0.20M
Total	22.72M	20.35M	15.47M

Table 1: Admissible parallel corpora used for training, with number of sentences before and after filtering

2. Training data and data selection

2.1. Parallel corpora

For the English-German language pair, we used the corpora listed in Table 1. IWSLT provides a large amount of permissible parallel training data. We performed filtering based on sentence alignment and language identification.

To obtain a sentence alignment score, we follow the idea that we can automatically translate the source text, and use BLEU between the automatic translation and the target side as a feature to predict probable alignments [9]. We trained a Phrase-based Statistical MT model, using significance filtering [10] to remove improbable phrases. Then we translated German sentences into English with a fast Statistical MT engine. Then, a sentence aligner BLEU-Champ[1] was applied to score each parallel training sentence. We also scored each sentence pair with a sentence-level language

[1]https://github.com/emjotde/bleu-champ

recognition tool. After these operations each sentence pair had assigned BLEU-Champ scores and language recognition scores. We selected small subset of 3k sentences from the corpora and performed manual evaluation for each sentence pairs scoring from 1 (very bad) to 5 (very good). Then we trained a regression model to predict human score based on BLEU-Champ and language recognition scores. Finally, we used the regression model to score whole parallel corpora and select potentially good sentences (predicted score above 2). We also removed lines with Wiki markup as we observed negative impact of such lines in our baseline model. Corpus sizes after these steps are shown in column **aligned** and **filtered**. The filtering method removed less than 5% of high quality corpora like News Commentary, but it removed over 50% of Wiki corpus. Additionally monolingual training data from the Commoncrawl [11] was used for creating synthetic parallel training data, see section 2.2 and 2.3 for details.

2.2. Selecting pseudo in-domain monolingual data

In order to reduce the amount of training data and possibly improve domain-adaptation effects, we decided to select data that matched the domain of TED talks based on Moore-Lewis filtering [12]. We followed the procedure described in Edinburgh's submission to IWSLT16 [13]. We used the TED talk data from WIT3 as seed data to create the in-domain language model and a matching amount of randomly chosen out-of-domain data for the contrasting language model.

Lang.	Total	Selected	Avg. score	Sel. score
de	2.9G	20M	0.4639	-0.0935
en	3.0G	20M	0.3797	-0.0394

Table 2: Selected monolingual data. Interpretation of figures is the same as for parallel data.

As seen in Table 2 we selected 20M sentences for back-translation from much larger original corpora of 2.9G and 3.0G sentences.

2.3. Preprocessing and subword units

To avoid the large-vocabulary problem in NMT models [14], we used byte-pair-encoding (BPE) to achieve open-vocabulary translation with a fixed vocabulary of subword symbols [15]. For all languages we set the number of merge operations to 90k. Segmentation into subword units was applied after any other preprocessing step for joint source and target vocabulary. We set vocabulary threshold to 50.

2.4. Back-translation

Corpus	size	oversapling
WIT3 (in-domain) [8]	0.20M	4.17M
Other parallel	15.27M	15.27M
Synthetic	19.57M	19.57M
Total	-	39.01M

Table 3: Final corpora used for training including admissible, filtered parallel corpora, oversampled in-domain corpus and synthetic, backtranslated data.

Back-translated monolingual in-domain data has been shown to be very beneficial when added to the parallel training data [16]. We back-translated the selected monolingual data with shallow, single layer NMT model trained on raw, permissible parallel data. We call it a baseline model hereafter. The model was trained with Nematus and translation was done with Marian [17]. We present the size of the final training corpora in table 3.

3. Neural translation systems

The neural machine translation system is an attentional encoder-decoder [18], which has been trained with Nematus [1]. There have been a number of papers showing that deeper models in machine translation lead to higher quality output. We apply the BiDeep model [2], which is a combination of stack RNNs and deep recurrent RNNs. Each cell in the stack RNN consists of multiple GRU cells, as illustrated in Figure 1. We use 4 stacks of RNNs with deep recurrent GRUs with a transition depth of 2.

In these experiments we followed the implementation details described in Edinburgh's WMT 2017 submission [19]. Important features which we used were: layer normalisation, BPE Version 2 with filtering of rare subwords, dynamic batching and using tied embeddings.

Additionally, to reduce training time we experimented with data parallelism on multi-GPU. Most of the approaches ([20],[21]) use SGD optimizer with centralized parameters which all workers read and

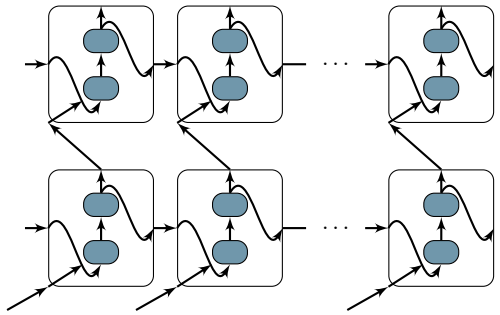

Figure 1: Illustration of BiDeep RNN architecture [2]. The architecture consists of a stack of layers of recurrent cells; each cell is composed of multiple GRU transitions.

update. We decided to use Adam optimizer instead as it was shown to converge faster. Unfortunately, the additional parameters in Adam make centralized parameter approach ineffective due to large copy overhead (one needs to store also means and variances of all parameters). In our implementation there is no centralized copy, instead each worker holds its own copy of all parameters. Each worker computes gradients on its own batch and only the gradients are summed over all workers and shared among them synchronously (we use nccl library[2]). Next, each worker independently updates its model parameters using the shared gradient. The data copied between workers is thus minimal. Since all workers are initialized equally, after the update they all still hold the same parameter values. Using N workers in this implementation can be seen as single worker case with N-times larger batch size. Thus to compare results one needs to set the training parameters like validation frequency accordingly. Table 4 compares trainings results for different number of GPUs. The results refer to de2en, BiDeep model training on filtered corpus using single node server with 8 GeForce 1080Ti. We did not present the 8-GPU case due to hardware problems with one card. The results show that our approach scales well. With increasing number of cards the throughput measured in words per second scales nearly linearly while the training time significantly reduces and the achieved BLEU is in close range.

To train the models for IWSLT 2017 submission we used 3 servers with 8 GPUs each (1 with GeForces 1080Ti, and 2 others with Teslas K80). The number of GPU used in particular training varied between 1 and 8 depending on resources availability.

[2]https://github.com/NVIDIA/nccl

	BLEU	time [h.]	words/sec.	overhead
1GPU	35.01	162.2	1082	8.3%
2GPU	34.85	103.0	1890	15.6%
3GPU	35.45	80.2	2950	17.6%
4GPU	35.30	67.1	3586	18.6%
6GPU	35.16	48.2	5315	23.0%

Table 4: Multi-GPU BiDeep model traininig statistics for different number of GPU. Training performed on de2en filtered corpora. The first column reports the best BLEU, the second convergence time, the third number of processed words per second. The last one is the overhead added by using the multi-gpu mechanism (reduce-all synchronization). Note the non-zero overhead even with 1 GPU.

3.1. Training, tuning and ensembling

We perform several steps of fine-tuning of the general models, using continued training with a new selection of training data and training parameters.

For each translation direction we run several independent trainings with slightly different data and parameters to get variety of final models for most successful ensembling. In all trainings we used TED test set from 2015 as a validation set.

As an example, the training of two of the de2en models (Fig. 2) used in the final ensemble, was started on the filtered parallel training data with 20 million in-domain backtranslated sentences and TED corpus oversampled 20 times. We trained this to convergence. After convergence, we enabled dropout, with both embedding dropout and hidden layer dropout set to 0.2, and continued training until results converged again.

We repeated this procedure two times for each direction, hoping that two independent runs will give us a better ensemble model than a checkpoint ensemble.

Finally we performed a fine-tuning step where we tuned to just the TED corpus with dropout and MAP-L2 regularization towards the previous model [3]. We also performed careful validation and early stopping. For fine-tuning we selected best and second best models for each independent run from the previous step.

For the final system we choose the 4 fine-tuned models that gave the best ensembles. In de2en direction

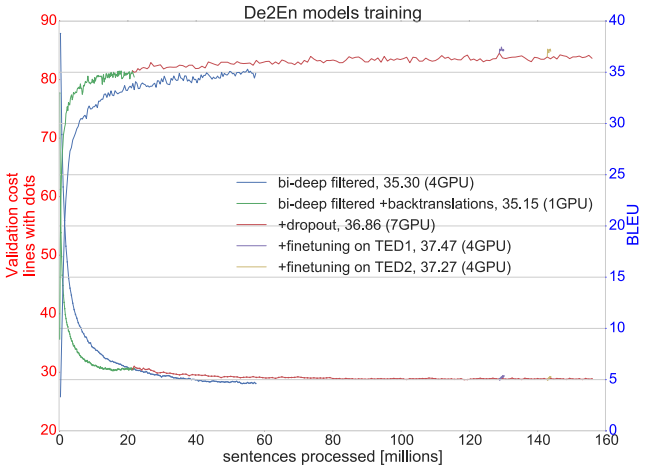

Figure 2: German-to-English models training progress. Plot shows the BLEU (straight lines) and validation error (lines with dots) on tst2015. Colors represent successive training parameters modifications. The number in plot label is the best BLEU score for particular training configuration.

the best model was a checkpoint ensemble and in en2de independent ensemble (from 2 independent models).

3.2. Spoken Language Translation

We also participated in the IWSLT Lectures spoken language translation task. This task consisted of three German lectures and two English lecutres and ten English TED talks. The test sets had no segmentation or punctuation. Our submission used an English and a German punctuation model provided by the SUMMA project. These models are essentially neural machine translation models which are trained to predict commas, full stops, question marks, exclamation marks and three dots [22, 23]. The punctuated source was then translated using the reranked ensembles described above.

4. Results

We present results in table 5. For the progress set, we also report BLEU of the University of Edinburgh submission to IWSLT16 [13], which was ranked first for en-de, and third for de-en. For results comparing our MT and SLT submissions to other systems in IWSLT please see the overview paper [24].

We performed extensive ablation studies. Our results confirm the effectiveness of deep models, which yield an improvement of 0.8-1.4 BLEU. They also give evidence for the sensitivity of our models towards

Translation	Progress set (2016)		Test set (2017)	
	de-en	en-de	de-en	en-de
IWSLT16 [13]	32.56	-	27.34	-
baseline	32.52	26.05	27.84	24.33
BiDeep raw	33.92	27.27	29.28	25.14
BiDeep filtered	34.07	27.66	29.94	25.61
+backtranslations	36.27	28.81	30.93	25.24
+dropout	36.50	29.83	31.41	26.66
+finetune on TED	37.08	30.21	32.26	27.38
+checkpoint ens.	37.61	30.34	32.37	27.56
independent ens.	37.56	29.91	32.71	27.23
+right to left	**37.85**	**30.93**	**33.08**	**28.00**

Table 5: Results for the IWSLT TED translation task (BLEU). Submitted system highlighted in bold.

training data noise, and models trained on filtered data outperform those trained on the full training coropra by 0.5–0.7 BLEU. Our use of back-translated data improved performance for de-en (+1 BLEU), and on the 2016 progress set for en-de (+0.8 BLEU), but not on the 2017 test set (-0.4 BLEU). Fine-tuning towards the TED training data remains an effective strategy (+0.8 BLEU), as does ensembling and right-to-left reranking.

In total, we report improvements of over 5 BLEU over last year's IWSLT submission by the University of Edinburgh [13], which was also based on Nematus and used a similar strategy for preprocessing and training. We attribute this to technical improvements in our neural network architecture, such as layer normalisation and BiDeep networks, better regularization during fine-tuning, and the use of more out-of-domain training data, and the use of reranking with a right-to-left model. We note that even our best single model outperforms last year's ensemble of 5 models by more than 4 BLEU.

5. Conclusions

This paper describes the joint submission of Samsung R&D Institute Poland and the University of Edinburgh team to the IWSLT MT task for TED talks, for the translation directions en-de and de-en. We report strong baseline results that are on par with last year's University of Edinburgh submission to IWSLT. Our experimental results confirm the effectiveness of the BiDeep NMT architecture, and of domain adaptation

via back-translated monolingual training data, and regularized fine-tuning towards an in-domain training set. Our results also highlight the importance of clean training data for NMT training, and we obtain better translation quality with a filtered subset of the permissible parallel training data. Our submission system outperforms our baseline, and last year's University of Edinburgh submission to IWSLT, by more than 5 BLEU.

6. Acknowledgments

The research presented in this publication was conducted in cooperation with Samsung Electronics Polska sp. z o.o. - Samsung R&D Institute Poland. We thank Antonio Valerio Miceli Barone for his illustration of the BiDeep architecture. This work was supported by the H2020 project SUMMA, under grant agreement 688139.

7. References

[1] R. Sennrich, O. Firat, K. Cho, A. Birch, B. Haddow, J. Hitschler, M. Junczys-Dowmunt, S. Läubli, A. V. Miceli Barone, J. Mokry, and M. Nadejde, "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*. Valencia, Spain: Association for Computational Linguistics, April 2017, pp. 65–68. [Online]. Available: http://aclweb.org/anthology/E17-3017.pdf

[2] A. V. Miceli Barone, J. Helcl, R. Sennrich, B. Haddow, and A. Birch, "Deep Architectures for Neural Machine Translation," in *Proceedings of the Second Conference on Machine Translation, Volume 1: Research Papers*. Copenhagen, Denmark: Association for Computational Linguistics, 2017. [Online]. Available: https://arxiv.org/pdf/1707.07631

[3] A. V. Miceli Barone, B. Haddow, U. Germann, and R. Sennrich, "Regularization techniques for fine-tuning in neural machine translation," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, 2017. [Online]. Available: https://arxiv.org/pdf/1707.09920.pdf

[4] O. Bojar, R. Chatterjee, C. Federmann, B. Haddow, M. Huck, C. Hokamp, P. Koehn, V. Logacheva, C. Monz, M. Negri, M. Post, C. Scarton, L. Specia, and M. Turchi, "Findings of the 2015 Workshop on Statistical Machine Translation," in *Proceedings of the Tenth Workshop on Statistical Machine Translation*. Lisbon, Portugal: Association for Computational Linguistics, September 2015, pp. 1–46. [Online]. Available: http://aclweb.org/anthology/W15-3001

[5] P. Koehn, "Europarl: A Parallel Corpus for Statistical Machine Translation," in *Conference Proceedings: the tenth Machine Translation Summit*, AAMT. Phuket, Thailand: AAMT, 2005, pp. 79–86. [Online]. Available: http://mt-archive.info/MTS-2005-Koehn.pdf

[6] A. Eisele and Y. Chen, " MultiUN: A Multilingual Corpus from United Nation Documents," in *Proceedings of the Seventh conference on International Language Resources and Evaluation*, 2010, pp. 2868–2872.

[7] P. Lison and J. Tiedemann, "OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, N. C. C. Chair), K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, Eds. Paris, France: European Language Resources Association (ELRA), may 2016.

[8] M. Cettolo, C. Girardi, and M. Federico, "WIT^3: Web Inventory of Transcribed and Translated Talks," in *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*, Trento, Italy, May 2012, pp. 261–268.

[9] R. Sennrich and M. Volk, "MT-based Sentence Alignment for OCR-generated Parallel Texts," in *The Ninth Conference of the Association for Machine Translation in the Americas (AMTA 2010)*, Denver, Colorado, USA, 2010. [Online]. Available: https://amta2010.amtaweb.org/AMTA/papers/2-14-SennrichVolk.pdf

[10] W. Ling, J. Graça, I. Trancoso, and A. Black, "Entropy-based Pruning for Phrase-based Ma-

chine Translation," in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, ser. EMNLP-CoNLL '12. Stroudsburg, PA, USA: Association for Computational Linguistics, 2012, pp. 962–971. [Online]. Available: http://dl.acm.org/citation.cfm?id=2390948.2391054

[11] C. Buck, K. Heafield, and B. van Ooyen, "N-gram Counts and Language Models from the Common Crawl," in *Proceedings of the Language Resources and Evaluation Conference*, Reykjavk, Icelandik, Iceland, May 2014.

[12] R. C. Moore and W. Lewis, "Intelligent Selection of Language Model Training Data," in *Proceedings of the ACL 2010 Conference Short Papers*, ser. ACLShort '10. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010, pp. 220–224. [Online]. Available: http://dl.acm.org/citation.cfm?id=1858842.1858883

[13] M. Junczys-Dowmunt and A. Birch, "The University of Edinburgh's systems submission to the MT task at IWSLT," in *Proceedings of IWSLT*, 2016.

[14] T. Luong, I. Sutskever, Q. V. Le, O. Vinyals, and W. Zaremba, "Addressing the Rare Word Problem in Neural Machine Translation," in *ACL*, 2015.

[15] R. Sennrich, B. Haddow, and A. Birch, "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)*. Berlin, Germany: Association for Computational Linguistics, August 2016, pp. 1715–1725. [Online]. Available: http://www.aclweb.org/anthology/P16-1162.pdf

[16] R. Sennrich, B. Haddow, and A. Birch, "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)*. Berlin, Germany: Association for Computational Linguistics, August 2016, pp. 86–96. [Online]. Available: http://www.aclweb.org/anthology/P16-1009.pdf

[17] M. Junczys-Dowmunt, T. Dwojak, and H. Hoang, "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, 2016. [Online]. Available: http://workshop2016.iwslt.org/downloads/IWSLT_2016_paper_4.pdf

[18] D. Bahdanau, K. Cho, and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," *CoRR*, vol. abs/1409.0473, 2014.

[19] R. Sennrich, A. Birch, A. Currey, U. Germann, B. Haddow, K. Heafield, A. V. Miceli Barone, and P. Williams, "The University of Edinburgh's Neural MT Systems for WMT17," in *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, Copenhagen, Denmark, 2017. [Online]. Available: https://arxiv.org/pdf/1708.00726

[20] J. Dean, G. Corrado, R. Monga, K. Chen, M. Devin, M. Mao, A. Senior, P. Tucker, K. Yang, Q. V. Le, *et al.*, "Large scale distributed deep networks," in *Advances in neural information processing systems*, 2012, pp. 1223–1231.

[21] S. Zhang, A. E. Choromanska, and Y. LeCun, "Deep learning with elastic averaging sgd," in *Advances in Neural Information Processing Systems*, 2015, pp. 685–693.

[22] O. Klejch, P. Bell, and S. Renals, "Punctuated transcription of multi-genre broadcasts using acoustic and lexical approaches," in *SLT*, 2016.

[23] O. Klejch, P. Bell, and S. Renals, "Sequence-to-sequence models for punctuated transcription combining lexical and acoustic features," in *ICASSP*, 2017.

[24] M. Cettolo, M. Federico, L. Bentivogli, J. Niehues, S. Stüker, K. Sudoh, K. Yoshino, and C. Federmann, "Overview of the IWSLT 2017 Evaluation Campaign," in *Proceedings of the 14th International Workshop on Spoken Language Translation (IWSLT)*, Tokyo, Japan, 2017.