The University of Edinburgh's systems submission to the MT task at IWSLT

Marcin Junczys-Dowmunt[1,2], Alexandra Birch[1]

[1]School of Informatics, University of Edinburgh
[2]Faculty of Mathematics and Computer Science, Adam Mickiewicz University in Poznań
junczys@amu.edu.pl a.birch@inf.ed.ac.uk

Abstract

This paper describes the submission of the University of Edinburgh team to the IWSLT MT task for TED talks. We took part in four translation directions, en-de, de-en, en-fr, and fr-en. The models have been trained with an attentional encoder-decoder model using Nematus, training data filtering and back-translation have been applied for domain-adaptation purposes.

1. Introduction

This paper describes the submission of the University of Edinburgh team to the IWSLT MT task for TED talks. We submitted translation results for four directions: en-de, de-en, en-fr, and fr-en. The models have been trained with an attentional encoder-decoder neural machine translation model using Nematus [1]. Due to the large amount of admissible parallel and monolingual training data and in order to benefit from domain-adaptation effects, we filtered all data sets towards the task domain. Furthermore back-translation has been applied for to utilize domain-filtered monolingual data.

2. Training data and data selection

2.1. Used corpora

Due to the addition of the Opensubtitles 2016 corpus and the United Nations Parallel Corpus v1.0, the amount of available parallel training data was a lot larger than in previous years. For the English-German and English-French language pair, we used the corpora listed in Table 1.

Additionally monolingual training data from the Commoncrawl [7] was used for back-translations, see section 2.2 and 2.3 for details.

Corpus	de-en	fr-en
WIT3 (in-domain) [2]	0.2M	0.2M
Commoncrawl [3]	2.3M	3.2M
Europarl v7 [4]	1.9M	2.0M
Giga Fr-En [3]	–	22.5M
News Commentary v11 [3]	0.2M	0.2M
Opensubtitles 2016 [5]	13.4M	33.5M
United Nations v1.0 [6]	–	25.8M

Table 1: Admissible parallel corpora used for training, with number of segments per language pair

2.2. Selecting pseudo in-domain training data

In order to reduce the amount of training data and possibly improve domain-adaptation effects, we decided to select data that seems to match the domain of TED talks based on Moore-Lewis filtering [8]. For the German-English pair, no parallel data filtering was performed, as the total number of training sentences was smaller than 20M.

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. For parallel data we filtered based on the English half only, for monolingual data we filtered using the respective language.

Prior to filtering, the data was tokenized and true-cased; to avoid issues with out-of-domain words, subword units [9] were applied for filtering. Subword units were computed on a small subset of the to-be-filtered data. Preprocessing was later reversed, the produced true-casing model and subwords symbols were not reused in further steps; we trained these models from scratch from the final data used for training as described in the next section.

As seen in Tables 2 and 3, the average cross-entropy

Lang.	Total	Selected	Avg. score	Sel. score
de-en	18M	18M	0.3753	0.3753
fr-en	87M	20M	0.5979	0.0800

Table 2: Selected parallel data. The average score is the average cross-entropy score before selection across the total number of sentences, the selected score is the average cross-entropy over the selected 20M segments. The lower the more in-domain.

Lang.	Total	Selected	Avg. score	Sel. score
de	2.9G	20M	0.4639	-0.0935
en	3.0G	20M	0.3797	-0.0394
fr	3.0G	20M	0.4403	-0.0185

Table 3: Selected monolingual data. Interpretation of figures is the same as for parallel data.

scores with regard to the TED seed language model decreases significantly after filtering, indicating higher similarity to actual in-domain data.

2.3. Preprocessing and subword units

Training data has been tokenized with the Moses tokenizer and true-cased. To avoid the large-vocabulary problem in NMT models [10], we used byte-pair-encoding (BPE) to achieve open-vocabulary translation with a fixed vocabulary of subword symbols [9]. For all languages we set the number of subword units to 50,000. Segmentation into subword units was applied after any other preprocessing step. During evaluation, subwords were reassembled, tokenization and true-casing were reversed.

2.4. Back-translation

The positive impact of adding back-translated monolingual in-domain data to the actual parallel data has been demonstrated in [11]. We back-translate the selected monolingual data due to time constraints with a phrase-based system. The parallel in-domain data as well as the parallel selected data have been used to train Moses baseline models. Monolingual files were split into pieces with $50,000$ lines each. The translation process has been accelerated using GNU parallel [12].

Translation direction	Progress set 2015 BLEU	Progress set 2015 TER	Test set 2016 BLEU	Test set 2016 TER
de-en	0.3383	0.4605	0.3256	0.4615
en-de	0.3042	0.5202	0.2734	0.5526
en-fr	0.3914	0.4445	0.3688	0.4602
fr-en	0.3969	0.4038	0.3756	0.4095

Table 4: Results for the IWSLT TED translation task

3. Neural translation systems

The neural machine translation system is an attentional encoder-decoder [13], which has been trained with Nematus [1]. We used mini-batches of size 40, a maximum sentence length of 50, word embeddings of size 500, and hidden layers of size 1024. We clipped the gradient norm to 1.0 [14]. Models were trained with Adam [15], reshuffling the training corpus between epochs. All models have been trained with scaling dropout over all GRU steps and with dropout over input embeddings [1], both with dropout probabilities of 0.1. The models were trained until convergence, saving every $30,000$ iterations (mini-batches).

For training the general models we used the $20,000,000$ most in-domain parallel sentences from the out-of-domain parallel training data (or all of it if less data was available) and $20,000,000$ back-translated sentences from the domain-selected monolingual data. The in-domain TED data was oversampled 20 times and also added to the full training data.

3.1. Finetuning and ensembling

We performed the same finetuning and ensembling method for each language pair: starting with the best model trained on the complete data, we fine-tuned each model for three epochs on the TED in-domain data only and repeated this process five times. For each fine-tuning step we saved the best model according to the dev set. The resulting five models were ensembled to produce the final results. We observed slight improvements on the dev set with each additional fine-tuned model added to the ensemble.

4. Results

The organizers of the shared task supplied the results listed in Table 4 for our submitted systems. We are

waiting for the complete rankings to assess the results in comparison to other submissions.

5. Acknowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement 688139 (SUMMA).

6. References

[1] R. Sennrich, B. Haddow, and A. Birch, "Edinburgh neural machine translation systems for WMT 16," in *WMT*, 2016, pp. 371–376.

[2] 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.

[3] 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

[4] 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

[5] 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.

[6] M. Ziemski, M. Junczys-Dowmunt, and B. Pouliquen, "The United Nations Parallel Corpus v1.0," in *LREC 2016*. ELRA, 2016.

[7] 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.

[8] 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

[9] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," *CoRR*, vol. abs/1508.07909, 2015.

[10] T. Luong, I. Sutskever, Q. V. Le, O. Vinyals, and W. Zaremba, "Addressing the rare word problem in neural machine translation," in *ACL*, 2015.

[11] R. Sennrich, B. Haddow, and A. Birch, "Improving neural machine translation models with monolingual data," *CoRR*, vol. abs/1511.06709, 2015. [Online]. Available: http://arxiv.org/abs/1511.06709

[12] O. Tange, "Gnu parallel - the command-line power tool," *;login: The USENIX Magazine*, vol. 36, no. 1, pp. 42–47, Feb. 2011. [Online]. Available: http://www.gnu.org/s/parallel

[13] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *CoRR*, vol. abs/1409.0473, 2014.

[14] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *ICML*, 2013, pp. 1310–1318.

[15] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014. [Online]. Available: http://arxiv.org/abs/1412.6980