The ADAPT Centre's Neural MT Systems for the WAT 2020 Document-Level Translation Task

Wandri Jooste, Rejwanul Haque and Andy Way
The ADAPT Centre, School of Computing
Dublin City University, Dublin, Ireland
firstname.lastname@adaptcentre.ie

Abstract

In this paper we describe the ADAPT Centre's (Team ID: adapt-dcu) submissions to the WAT 2020 document-level Business Scene Dialogue (BSD) translation task. We only considered translating from Japanese to English for this task and secured the third position in the competition as per the rankings of the MT systems based on the human evaluation scores. The machine translation (MT) systems that we built for this task are state-of-the-art Transformer models. In order to improve the translation quality of our neural MT (NMT) systems, we made use of both in-domain and out-of-domain data for training. We applied various data augmentation techniques for fine-tuning the model parameters. This paper outlines the experiments we carried out for this task and reports the MT systems' performance on the evaluation test set.

1 Introduction

We participated in the WAT 2020[1] (Nakazawa et al., 2020) document-level BSD translation task and only submitted systems that translate from Japanese-to-English (Ja-to-En). Our MT systems are Transformer models (Vaswani et al., 2017) which were trained using the Marian-NMT toolkit.[2] In this work, we applied different domain adaptation techniques, such as using synthetic data from source- and target-side monolingual data through the use of forward- and back-translation (Sennrich et al., 2016; Chinea-Ríos et al., 2017; Poncelas et al., 2018) and out-of-domain parallel data to train our models. As far as fine-tuning the model parameters is concerned, we experimented with conventional fine-tuning which consists of fine-tuning on in-domain data only, mixed fine-tuning and lastly document-level fine-tuning.

This paper is organised as follows. Firstly, the data used and processing steps are described in Section 2. Next, the methods for training the baseline MT systems are described in Section 3, and the results obtained from our MT systems are presented in Section 4. Finally, Section 5 concludes our work with the avenues for the future work.

2 Data and Preprocessing

This section outlines the corpora used and the steps that were taken to preprocess the data for training.

2.1 Corpora Used

To train the baseline models a mixture of three corpora was used, where one corpus contains in-domain sentences and the other two corpora contain out-of-domain sentences. The in-domain BSD corpus (Rikters et al., 2019) consists of a training set of 20,000 sentences, a development set of 2,051 sentences and a test set of 2,120 sentences. The out-of-domain data that was added to the BSD training set includes the JESC[3] (Pryzant et al., 2018) corpus consisting of 2.8 million sentences and the OpenSubtitles[4] (Lison and Tiedemann, 2016) corpus consisting of 2.2 million sentences. Furthermore, monolingual data from the target-side (En) of the JW300[5] corpus (Agić and Vulić, 2019) was used to create synthetic data with the use of back-translation.

Finally, source-language monolingual data with *n*-grams similar to that of the documents in the test set was mined from the Common Crawl Corpus[6] to be used as a source-side original synthetic corpus (SOSC) for fine-tuning the NMT model parameters.

[1] http://lotus.kuee.kyoto-u.ac.jp/WAT/WAT2020/index.html
[2] https://github.com/marian-nmt/marian
[3] https://nlp.stanford.edu/projects/jesc/
[4] http://www.opensubtitles.org/
[5] http://opus.nlpl.eu/JW300.php
[6] https://commoncrawl.org/the-data/

Proceedings of the 7th Workshop on Asian Translation, pages 142–146

2.2 Preprocessing

The source-side sentences (Ja) were segmented using *MeCab*[7] and the target-side sentences (En) were tokenized and lower-cased using the standard scripts from the Moses toolkit (Koehn et al., 2007). For both languages Subword-NMT[8] was used to apply byte pair encoding (BPE). We experimented with two different vocabulary sizes: 6,000 and 32,000. Details around these vocabulary sizes will be discussed in more detail in Section 3.

3 The Baseline MT Systems

We started off by training three baseline models ($B1$, $B2$ and $B3$) and the best baseline model was used to experiment with different data augmentation methods when fine-tuning model parameters. These methods will be described in detail in Section 4.

The first baseline model ($B1$) was trained on the BSD, JESC and OpenSubtitles corpora combined. As mentioned earlier, we used the MarianNMT toolkit to train our Transformer models. The setup described in Sennrich et al. (2017) was used as is for $B1$, more specifically the mini-batch size for validation was 64 and the learning rate 0.0003. The changes to this setup for the other two baseline setups ($B2$ and $B3$) are discussed below. The setup differs since BPE with a vocabulary of 32,000 was applied to the training set of $B1$ and a vocabulary size of 6,000 was used for the other two baseline models ($B2$ and $B3$). We obtained the BLEU (Papineni et al., 2002) scores to evaluate these baseline MT systems on the evaluation test set and the scores are presented in Table 1.

The second baseline model ($B2$) was trained on the same training set as $B1$. The third baseline model's ($B3$) training set consisted of the target-side original synthetic corpus added to the training set of $B1$ (cf. Table 1). The setup for $B1$ and $B2$ was changed by setting the mini-batch size for validation to 32 and the learning rate to 0.0005.

4 Improving the Baseline MT Systems

The BLEU score of each of the baseline models described in Section 3 is shown in Table 1 and it is clear that $B2$ is the best-performing MT system out of the three baseline models. Therefore, we decided to conduct the remainder of our experiments on this baseline model ($B2$) alone.

Model	Corpus	BPE size	BLEU
$B1$	basic	32,000	16.76
$B2$	basic	6,000	**17.33**
$B3$	basic+JW300	6,000	16.69

Table 1: Comparison of baseline models and their performance. (Basic training set consist of BSD, JESC and OpenSubtitles).

In order to improve upon the scores achieved by the baseline model $B2$, we fine-tuned the parameters of the model. This was done by restarting training on model $B2$, after initial training has ended, with a newly selected corpus. We implemented four different scenarios for fine-tuning the parameters.

Scenario 1: The first scenario is the most basic, where we simply performed conventional fine-tuning of the model parameters on in-domain data only, namely the BSD training set.

Scenario 2: In the second scenario we implemented mixed fine-tuning of model parameters, where fine-tuning is conducted on the training data that consists of both in-domain data and out-of-domain data as described in Chu et al. (2017). The in-domain data was augmented by oversampling the BSD training set 50 times and the out-of-domain data is a mixture of JESC and OpenSubtitles.

Scenario 3: As for the third scenario, source-side monolingual sentences were mined that are similar in styles to the BSD test set sentences. We followed Nayak et al. (2020) and Parthasarathy et al. (2020) in order to mine those sentences from large monolingual data that could be beneficial for fine-tuning the original NMT models. We identified terms in the test set to be translated. For this, we followed the monolingual terminology extraction methods described in Haque et al. (2014, 2018), which used a large corpus that is generic in nature as a reference corpus. In our setup, we used the source-side of the authentic training bitexts on which our NMT system ($B2$) was trained as the reference corpus. The intuition is to extract those terminological expressions from the test set that do not occur or rarely occur in the training data and are more indicative of the test corpus. Given the list of extracted terms, we mined sentences from the

[7] https://github.com/SamuraiT/mecab-python3

[8] https://github.com/rsennrich/subword-nmt

Scenarios	Corpus	BLEU	RIBES	Human Evaluation Score
1	BSD	14.52	70.13	-
2	JESC+OpenSubtitles+BSD∗50	**18.70**	73.04	+3.930
3	SOSC	18.53	73.16	-
4	SOSC+BSD	18.59	**73.22**	-

Table 2: Comparison of the performance of different fine-tuning techniques. Only one of our systems was submitted for human evaluation (Scenario 2).

large monolingual corpus[9] mentioned in Section 2. The Japanese source sentences (a total of 153,402 sentences) that have been mined were translated into English using $B2$ to create synthetic data (i.e. source-side original synthetic corpus (SOSC)) to be used for fine-tuning our baseline model. Thus, in Scenario 3, the model parameters were fine-tuned on the out-of-domain data only.

Scenario 4: Finally, in Scenario 4, we once again tested the mixed fine-tuning strategy. In this case, the in-domain data consists of the BSD training set and the out-of-domain data consists of SOSC.

The results obtained from the different scenarios are shown in Table 2, where the mixed fine-tuning of Scenario 2 combined with the data augmentation technique provides us the best BLEU score on the test set. The MT system of this setup produces a 1.37 BLEU points corresponding to 7.9% relative gain over the baseline. The gain is statistically significant (Koehn, 2004).

As for Scenario 1, fine-tuning model parameters on in-domain-data only does not work well and the corresponding MT system performs poorly in comparison to the performance of other MT systems (cf. Table 2). We conjecture that this happened due to the fact that only a small-sized in-domain training corpus was used for fine-tuning.

The WAT 2020 shared task organisers reported the evaluation results in terms of the BLEU and RIBES (Isozaki et al., 2010) metrics. As for RIBES, Scenario 1 once again is found to be less effective, and the corresponding MT system produces a low score on the test set. However, the difference between the lowest and highest RIBES scores is much smaller than that of the BLEU scores. Another notable difference is that Scenario 4 seems to be most effective since the MT system of this setup provides us the highest RIBES score on the test set. This is a contrasting outcome to the one with the BLEU metric, where Scenario 2 provided the highest BLEU score on the test set.

We submitted translations of the MT system of the Scenario 2 setup for the human evaluation task conducted by the shared task organisers since it produced the highest BLEU score on the evaluation test set. As can be seen from the last column of Table 2, we received a score of 3.930 as far as the results of the human evaluation task is concerned. We secured the third position in the competition for the Japanese-to-English BSD translation task as per the rankings of the MT systems based on the human evaluation scores.

5 Conclusion

In this paper, we described our MT systems that were submitted to the WAT 2020 document-level Business Scene Dialogue translation shared task. We presented the WAT 2020 official results that we obtained by submitting the translations of our MT systems. We showed that, in the case of limited in-domain training data, both in-domain and out-of-domain data is useful for fine-tuning model parameters, which essentially provides the best results in this translation task. Furthermore, making use of synthetic parallel data in training also greatly increased the performance of our MT systems.

In future, we aim to exploit document-level syntactic context in the fine-tuning step. We also aim to explore increasing training batch size at the fine-tuning step as this may capture wider contexts during training.

Acknowledgments

The ADAPT Centre for Digital Content Technology is funded under the Science Foundation Ireland (SFI) Research Centres Programme (Grant No. 13/RC/2106) and is co-funded under the European Regional Development Fund. This project has partially received funding from the European Union's Horizon 2020 research and innovation

[9]The monolingual corpus contains 9,923,690 sentences which are a part of the Common Crawl Corpus.

programme under the Marie Skłodowska-Curie grant agreement No. 713567, and the publication has emanated from research supported in part by a research grant from SFI under Grant Number 13/RC/2077 and 18/CRT/6224 .

References

Željko Agić and Ivan Vulić. 2019. JW300: A Wide-Coverage Parallel Corpus for Low-Resource Languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3204–3210, Florence, Italy.

Mara Chinea-Ríos, Álvaro Peris, and Francisco Casacuberta. 2017. Adapting neural machine translation with parallel synthetic data. In *Proceedings of the Second Conference on Machine Translation*, pages 138–147, Copenhagen, Denmark. 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.

Rejwanul Haque, Sergio Penkale, and Andy Way. 2014. Bilingual termbank creation via log-likelihood comparison and phrase-based statistical machine translation. In *Proceedings of the 4th International Workshop on Computational Terminology (Computerm)*, pages 42–51, Dublin, Ireland.

Rejwanul Haque, Sergio Penkale, and Andy Way. 2018. TermFinder: log-likelihood comparison and phrase-based statistical machine translation models for bilingual terminology extraction. *Language Resources and Evaluation*, 52(2):365–400.

Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. 2010. Automatic evaluation of translation quality for distant language pairs. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 944–952, Cambridge, MA. Association for Computational Linguistics.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. 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, Ondřej 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.

Pierre Lison and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

Toshiaki Nakazawa, Hideki Nakayama, Chenchen Ding, Raj Dabre, Hideya Mino, Isao Goto, Win Pa Pa, Anoop Kunchukuttan, Shantipriya Parida, Ondřej Bojar, and Sadao Kurohashi. 2020. Overview of the 7th workshop on Asian translation. In *Proceedings of the 7th Workshop on Asian Translation*, Suzhou, China. Association for Computational Linguistics.

Prashanth Nayak, Rejwanul Haque, and Andy Way. 2020. The ADAPT's submissions to the WMT20 biomedical translation task. In *Proceedings of the Fifth Conference on Machine Translation (Shared Task Papers (Biomedical)*, Punta Cana, Dominican Republic.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Venkatesh Balavadhani Parthasarathy, Akshai Ramesh, Rejwanul Haque, and Andy Way. 2020. The ADAPT system description for the WMT20 news translation task. In *Proceedings of the Fifth Conference on Machine Translation (Shared Task Papers (News))*, Punta Cana, Dominican Republic.

Alberto Poncelas, Dimitar Shterionov, Andy Way, Gideon Maillette de Buy Wenniger, and Peyman Passban. 2018. Investigating backtranslation in neural machine translation. In *Proceedings of The 21st Annual Conference of the European Association for Machine Translation (EAMT 2018)*, pages 249–258, Alicante, Spain.

Reid Pryzant, Yongjoo Chung, Dan Jurafsky, and Denny Britz. 2018. JESC: Japanese-English Subtitle Corpus. *arXiv:1710.10639 [cs]*.

Matīss Rikters, Ryokan Ri, Tong Li, and Toshiaki Nakazawa. 2019. Designing the business conversation corpus. In *Proceedings of the 6th Workshop on Asian Translation*, pages 54–61, Hong Kong, China. Association for Computational Linguistics.

Rico Sennrich, Alexandra Birch, Anna Currey, Ulrich Germann, Barry Haddow, Kenneth Heafield, Antonio Valerio Miceli Barone, and Philip Williams.

2017. The university of Edinburgh's neural MT systems for WMT17. In *Proceedings of the Second Conference on Machine Translation*, pages 389–399, Copenhagen, Denmark. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. 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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *arXiv:1706.03762 [cs]*.