# Improving NMT via Filtered Back Translation

**Nikhil Jaiswal, Mayur Patidar, Surabhi Kumari, Manasi Patwardhan**
**Shirish Karande, Puneet Agarwal, Lovekesh Vig**
TCS Research, New Delhi, India
{nikhil.jais, patidar.mayur, surabhi.kumari6,
shirish.karande, puneet.a, lovekesh.vig}@tcs.com

## Abstract

Document-Level Machine Translation (MT) has become an active research area among the NLP community in recent years. Unlike sentence-level MT, which translates the sentences independently, document-level MT aims to utilize contextual information while translating a given source sentence. This paper demonstrates our submission **(Team ID - DEEPNLP)** to the Document-Level Translation task organized by WAT 2020[1]. This task focuses on translating texts from a business dialog corpus while optionally utilizing the context present in the dialog. In our proposed approach, we utilize publicly available parallel corpus from different domains to train an open domain base NMT model. We then use monolingual target data to create filtered pseudo parallel data and employ Back-Translation to fine-tune the base model. This is further followed by fine-tuning on the domain-specific corpus. We also ensemble various models to improvise the translation performance. Our best models achieve a BLEU score of 26.59 and 22.83 in an unconstrained setting and 15.10 and 10.91 in the constrained settings for En → Ja & Ja → En direction, respectively.

## 1 Introduction

Neural Machine Translation (Bahdanau et al., 2015; Vaswani et al., 2017a) has performed impressively in recent years, especially for high resource language pairs. However, one of the shortcomings while translating texts in the form of a paragraph or a document is that the inter-relations among sentences are ignored and the sentences are translated independently. Document Level MT (Maruf et al., 2020; Zhang and Zong, 2020; Kim et al., 2019b) aims to utilize these inter-sentential context information to deal with context-dependent phenomena such as coreference, lexical cohesion, and consistency, lexical disambiguation, etc. (Voita et al., 2019; Lopes et al., 2020) The meaning of a translated sentence can deviate from its originality when treated independently. WAT 2020's (Nakazawa et al., 2020) Document-level Business Scene Dialogue (BSD) Translation sub-task aims to foster research in the area of document-level MT. To tackle this task, we perform the following steps. Firstly, we gather several publicly available English Japanese corpus and combine them to train an open domain base model. Then, we utilize the monolingual corpus in the target language to create the pseudo parallel corpus. Since the generated pseudo parallel corpus might consist of noisy translated sentences, we use a sentence-level similarity-based filtration technique to filter out such pairs. We then fine-tune the base model on the filtered data followed by fine-tuning on in-domain parallel BSD[2] data. We also utilize checkpoint ensembles to further improve the translation performance.

## 2 Problem Description

This task aims to translate all the sentences in the BSD test file from Ja → En and vice-versa. Participants could participate either in the constrained setting in which only the official BSD corpus needs to be used or in an unconstrained setting where other resources such as parallel corpora, monolingual corpora, and parallel dictionaries in addition to the official corpora could be utilized. We participate in both settings. BLEU (Papineni et al., 2002), RIBES [3] and AMFM (Banchs et al., 2015) are used as the official automatic evaluation metrics and are calculated on the tokenized version of translated and reference sentences using different tokenizers such as Juman, KyTea, MeCab, & Moses.

[1] http://lotus.kuee.kyoto-u.ac.jp/WAT/WAT2020/index.html

[2] https://github.com/tsuruoka-lab/BSD

[3] http://www.kecl.ntt.co.jp/icl/lirg/ribes/index.html



*Proceedings of the 7th Workshop on Asian Translation*, pages 154–159

## 3 Related Work

There are two major directions in MT, which has attracted a lot of attention from the research community in recent years - Document Level MT & MT on low resource language pairs. Numerous works have been proposed to tackle document-level MT (Maruf et al., 2019; Miculicich Werlen et al., 2018). This area's work involves utilizing contexts on source, target, or both side and designing architectures using either the single (Ma et al., 2020) or additional encoder (Zhang et al., 2018; Miculicich et al., 2018) to handle contextual information. Some work in this area also tries to analyze the contextual errors (Kim et al., 2019a). The work related to low resource language pairs involves making use of monolingual data to create pseudo parallel corpus using back translation (Sennrich et al., 2016), iterative back translation (Hoang et al., 2018) & filtered back translation techniques (Junczys-Dowmunt, 2018; Dou et al., 2020), etc. For filtering noisy pairs, Imankulova et al. (2017) uses the Round Trip BLEU score between true and synthetic sentences. Wang et al. (2019) propose dynamic domain-data selection along with dynamic clean-data selection.

## 4 System Description

We describe our proposed approach in this section. For the unconstrained setting, we first create *BASE* models by training $NMT_{s \to t}$ and $NMT_{t \to s}$ on the open domain dataset. Then, we use these trained $NMT_{s \to t}$ & $NMT_{t \to s}$ models to translate the monolingual data $M_s$ & $M_t$ to $M'_t$ & $M'_s$ respectively. We then utilize the pseudo parallel data $M'_t$ and $M_s$ along with equal amount of true parallel data to fine-tune $NMT_{t \to s}$ model. Similarly, we use the pseudo parallel data $M'_s$ and $M_t$ along with equal amount of true parallel data to fine-tune $NMT_{s \to t}$ model. This results in the creation of the back-translated (*BT*) models. In other settings, instead of utilizing the entire pseudo parallel data, we apply the filtering technique described below on these data to filter out noisy pairs. Then we use these filtered pairs along with an equal amount of true parallel data to fine-tune the $NMT_{t \to s}$ & $NMT_{s \to t}$ models. This results in the creation of the filtered back-translated *(FBT)* models. We further fine-tune *BT* as well as *FBT* models on the BSD corpus. For the constrained setting, we train $NMT_{s \to t}$ and $NMT_{t \to s}$ models directly on the BSD corpus. We also experiment with fine-tuning mBART (Liu et al., 2020) model on the BSD corpus. We finally build several ensembles by averaging checkpoints of different trained models.

| Dataset | Sentences |
|---|---|
| WikiMatrix | 3,895,992 |
| JESC | 2,801,388 |
| Wiki Titles v2 | 705,962 |
| KFTT | 442,614 |
| Japanese-English Legal Corpus | 262,449 |
| TED talks | 226,834 |
| MTNT | 12,281 |
| News Commentary | 1,869 |

Table 1: Open Domain Corpus Statistics

| | Training | Development | Test |
|---|---|---|---|
| Sentences | 20,000 | 2,051 | 2,120 |

Table 2: BSD Corpus Statistics

**Filtering Technique:** We apply a naive filtering model based on sentence similarity to filter out noisy pseudo parallel data. Given a monolingual source sentence S, we obtain the corresponding translated sentence T using the trained $NMT_{s \to t}$ model. We then apply MUSE (Multilingual Universal Sentence Encoder) (Yang et al., 2019) to obtain the sentence embeddings of S and T. Then cosine similarity is calculated on the obtained embeddings of S and T, and if the cosine score is below a certain threshold, we treat this pair as noisy. The threshold value is decided based on the cosine score on the entire monolingual data and its corresponding generated translations. We also utilize this filtering strategy to sample sentence pairs from the true parallel data. For this, we sort the entire true parallel data in decreasing order of similarity scores. Then we remove pairs that contain the text in the same language in the source and target side using Langid (Lui and Baldwin, 2012) library. We also remove pairs where the same text is present on the source and target side. Finally, we return the top n sentence pairs from the above data where n is the number of samples required from the true parallel data.

## 5 Experiments and Results

### 5.1 Data Preparation & Preprocessing

We collect and merge several publicly available parallel corpus for training the *BASE* models on the open domain data. We use datasets from sev-

| LP → | En-Ja | | | Ja-En | | |
|---|---|---|---|---|---|---|
| Models → | BASE | BT | FBT | BASE | BT | FBT |
| Test Data ↓ | | | | | | |
| MTNT | 16.5 | 17.4 | **18.0** | 11.3 | 11.2 | **12.2** |
| IWSLT | 20.2 | 20.1 | **21.1** | 17.9 | 17.6 | **18.1** |
| WMT News | 24.2 | 24.4 | **25.7** | 16.0 | 16.4 | **17.5** |
| BSD | 16.2 | 15.5 | **18.9** | 15.9 | **16.1** | **16.1** |

Table 3: We use the BLEU score to compare the *BASE*, *BT*, and *FBT* models trained on the open domain corpus to evaluate on different publicly available test sets including the BSD corpus.

eral domains such as news, movie, Wikipedia articles, etc., so that the *BASE* model is domain agnostic. Table 1 presents the number of sentences in each such parallel corpora. The final corpus consists of around 8.3 million sentence pairs for training, 10K for validation and 7K for test set and is formed by combining following datasets - KFTT (Kyoto Free Translation Task) (Neubig, 2011), JESC (Japanese-English Subtitle Corpus) (Pryzant et al., 2017), Japanese-English Legal Parallel Corpus (Neubig, 2011), WikiMatrix (Schwenk et al., 2019), News Commentary[4], Wiki Titles v2[4] , TED Talks (Hochreiter and Schmidhuber, 1997) and MTNT (Machine Translation of Noisy Text) parallel corpus (Michel and Neubig, 2018). While combining the datasets, we follow the train, validation, and test set as provided in the respective corpus and use it in a similar fashion in our combined dataset. We sample 3 million monolingual data from News Crawl[4] dataset to create pseudo parallel corpus. Along with these pseudo parallel data, we randomly sample the same amount of true parallel data from the open domain dataset. We combine and shuffle both the pseudo parallel and true parallel data. Finally, we utilize the BSD corpus provided by WAT 2020. This corpus is manually created and consists of Japanese-English business conversations. We use the provided training, development, and evaluation splits, which are described in Table 2.

We use different preprocessing rules for each translation direction based on our initial experimentation results as well as the findings from the literature. For Ja-En, we train & apply sentencepiece (Kudo and Richardson, 2018) model to tokenize the raw text into subwords with the vocabulary size of 32,000 for each language. For En-Ja, we first tokenize the raw text by KyTea and the Moses tokenizer for Japanese and English, respectively. We also use Moses toolkits to truecase English words. We then further train & apply sentencepiece model to tokenize these words into subwords with the vocabulary size of 32,000 for each language.

[4] http://www.statmt.org/wmt20/translation-task.html

### 5.2 Implementation Details

Here, we describe a detailed setup of our experiments in both the constrained and unconstrained settings. For the unconstrained setting, we utilize Transformer-base (Vaswani et al., 2017b) model for training the open domain *BASE* models. The encoder and decoder consist of 6 layers, 8 attention heads, and the hidden size is kept to 512. We use Adam optimizer with an initial learning rate of 0.001 and dropout regularization, whose value is fixed at 0.3. We use Fairseq (Ott et al., 2019) to implement all our experiments. All the models are trained until the convergence with patience of five. Once the *BASE* models are trained, we use monolingual data to create pseudo parallel data and train the *BT* models. For filtering based on the sentence similarity, we use the MUSE model from the TensorFlow Hub library to obtain the sentence embeddings. For the constrained setting, we experiment with a Transformer-base model with two as well as three encoder and decoder layers for training on BSD corpus. We also experiment with fine-tuning the mBART model on the BSD corpus.

### 5.3 Results and Analysis

This section discusses the results of our different experiments on both constrained and unconstrained settings. For the unconstrained setting, we first summarize the results of the *BASE* model, *BT* model, and the *FBT* model in both directions on four different publicly available test sets, including the BSD corpus in Table 3. From the table, we can observe that performing *BT* as well as *FBT* helps in improvising the BLEU score of the *BASE* model.

| LP → | En-Ja | | | Ja-En |
|---|---|---|---|---|
| **Tokenizer →** | **juman** | **kytea** | **mecab** | **moses** |
| **Models ↓** | | | | |
| **TRF_2** | 7.80 | 14.06 | 9.19 | 7.47 |
| **TRF_3** | 7.90 | 14.03 | 9.10 | 9.70 |
| **TRF_2 + TRF_3 (ours)** | 8.66 | 15.10 | 10.10 | 10.91 |
| **mBART_FT** | **14.84** | **21.29** | **16.29** | **18.01** |

Table 4: **Constrained Setting:** We use the BLEU score to compare the 2 & 3 layers Transformer (*TRF*) models trained on BSD corpus, as well as the ensemble model and the fine-tuned mBART model.

| LP → | En-Ja | | | Ja-En |
|---|---|---|---|---|
| **Tokenizer →** | **juman** | **kytea** | **mecab** | **moses** |
| **Models ↓** | | | | |
| **BASE_FT** | 18.76 | 25.86 | 20.19 | 21.74 |
| **BT_FT** | 18.99 | 25.90 | 20.54 | 21.95 |
| **FBT_FT** | 17.85 | 24.72 | 19.30 | 21.67 |
| **BASE_FT + BT_FT** | 19.20 | 26.27 | 20.77 | **22.83** |
| **BASE_FT + FBT_FT** | **19.39** | **26.59** | **20.95** | 22.75 |

Table 5: **Unconstrained Setting:** We use the BLEU score to compare the *BASE*, *BT*, *FBT* and the ensemble models which are fine-tuned on the BSD corpus. We report the results using different tokenizers in each direction.

| LP → | En-Ja | Ja-En |
|---|---|---|
| **Constrained** | 2.60 | 2.40 |
| **Unconstrained** | 4.13 | 4.10 |

Table 6: We report the Human evaluation result of the Pairwise Crowdsourcing by WAT2020. This was evaluated by 5 different workers, and the final decision is made by the voting of the judgements.

So, in the zero-shot setting where none of the BSD data is used for training, we are able to obtain a BLEU score of 18.9 and 16.1 in En → Ja & Ja → En directions, respectively. We use KyTea as the tokenizer for the Japanese sentences in the results mentioned in Table 3.

For the constrained setting, Table 4 presents the overall results. For the En → Ja translation, the BLEU score using different Japanese tokenizers such as juman, kytea, and mecab are reported. For the Ja → En direction, moses tokenizer is used for the evaluation. Although the ensemble model gave us better performance compared to the single model alone, but it is the mBART model whose fine-tuning on BSD corpus surpasses all other models by a large margin in both directions. Table 5 presents the unconstrained setting results obtained by fine-tuning the *BASE*, *BT* and *FBT* models on the BSD corpus. It also reports the results of ensembles formed by using different models. We can observe that the ensemble model comprising of fine-tuning *BASE* and *FBT* models gives us the best performance for the En → Ja direction, whereas in the case of Ja → En, ensemble model comprising of fine-tuning *BASE* and *BT* models achieves the highest BLEU score. Table 6 reports the human evaluation results in both the settings.

## 6 Conclusion

We experimented with a variety of techniques in both constrained & unconstrained settings. For the constrained setting, fine-tuning mBART on the BSD corpus gave the best translation performance in both directions. Thus, mBART can be fine-tuned for MT tasks, especially for low resource language pairs. For the unconstrained scenario, the models trained & fine-tuned using the pseudo-parallel corpus showed the best overall translation performance. We also showed that by using a simple ensemble technique of averaging different model checkpoints, the translation performance could be further improvised.

## References


Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

R. E. Banchs, L. F. D'Haro, and H. Li. 2015. Adequacy–fluency metrics: Evaluating mt in the continuous space model framework. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(3):472–482.

Zi-Yi Dou, Antonios Anastasopoulos, and Graham Neubig. 2020. Dynamic data selection and weighting for iterative back-translation.

Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. 2018. Iterative back-translation for neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 18–24, Melbourne, Australia. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, pages 1735–80.

Aizhan Imankulova, Takayuki Sato, and Mamoru Komachi. 2017. Improving low-resource neural machine translation with filtered pseudo-parallel corpus. In *Proceedings of the 4th Workshop on Asian Translation (WAT2017)*, pages 70–78, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Marcin Junczys-Dowmunt. 2018. Dual conditional cross-entropy filtering of noisy parallel corpora. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 888–895, Belgium, Brussels. Association for Computational Linguistics.

Yunsu Kim, Duc Tran, and Hermann Ney. 2019a. When and why is document-level context useful in neural machine translation? pages 24–34.

Yunsu Kim, Duc Thanh Tran, and Hermann Ney. 2019b. When and why is document-level context useful in neural machine translation?

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *CoRR*, abs/1808.06226.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation.

António Lopes, M. Amin Farajian, Rachel Bawden, Michael Zhang, and André F. T. Martins. 2020. Document-level neural MT: A systematic comparison. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 225–234, Lisboa, Portugal. European Association for Machine Translation.

Marco Lui and Timothy Baldwin. 2012. Langid.py: An off-the-shelf language identification tool. pages 25–30.

Shuming Ma, Dongdong Zhang, and Ming Zhou. 2020. A simple and effective unified encoder for document-level machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3505–3511, Online. Association for Computational Linguistics.

Sameen Maruf, André F. T. Martins, and Gholamreza Haffari. 2019. Selective attention for context-aware neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3092–3102, Minneapolis, Minnesota. Association for Computational Linguistics.

Sameen Maruf, Fahimeh Saleh, and Gholamreza Haffari. 2020. A survey on document-level neural machine translation: Methods and evaluation.

Paul Michel and Graham Neubig. 2018. MTNT: A testbed for machine translation of noisy text. *CoRR*, abs/1809.00388.

Lesly Miculicich, Dhananjay Ram, Nikolaos Pappas, and James Henderson. 2018. Document-level neural machine translation with hierarchical attention networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2947–2954, Brussels, Belgium. Association for Computational Linguistics.

Lesly Miculicich Werlen, Dhananjay Ram, Nikolaos Pappas, and James Henderson. 2018. Document-level neural machine translation with hierarchical attention networks.

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.

Graham Neubig. 2011. The Kyoto free translation task. http://www.phontron.com/kftt.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

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.

Reid Pryzant, Yongjoo Chung, Dan Jurafsky, and Denny Britz. 2017. JESC: japanese-english subtitle corpus. *CoRR*, abs/1710.10639.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2019. Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia. *CoRR*, abs/1907.05791.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017a. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, undefinedukasz Kaiser, and Illia Polosukhin. 2017b. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019. When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. *CoRR*, abs/1905.05979.

Wei Wang, Isaac Caswell, and Ciprian Chelba. 2019. Dynamically composing domain-data selection with clean-data selection by "co-curricular learning" for neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1282–1292, Florence, Italy. Association for Computational Linguistics.

Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernández Ábrego, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2019. Multilingual universal sentence encoder for semantic retrieval. *CoRR*, abs/1907.04307.

Jiacheng Zhang, Huanbo Luan, Maosong Sun, Feifei Zhai, Jingfang Xu, Min Zhang, and Yang Liu. 2018. Improving the transformer translation model with document-level context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 533–542, Brussels, Belgium. Association for Computational Linguistics.

Jiajun Zhang and Chengqing Zong. 2020. Neural machine translation: Challenges, progress and future.