# SYSTRAN @ WAT 2019: Russian↔Japanese News Commentary task

**Jitao Xu**[†], **MinhQuang Pham**[†‡], **TuAnh Nguyen**[†], **Josep Crego**[†], **Jean Senellart**[†]
[†]SYSTRAN / 5 rue Feydeau, 75002 Paris, France
firstname.lastname@systrangroup.com
[‡]LIMSI, CNRS, Université Paris-Saclay 91405 Orsay, France
firstname.lastname@limsi.fr

## Abstract

This paper describes SYSTRAN's submissions to WAT 2019 Russian↔Japanese News Commentary task. A challenging translation task due to the extremely low resources available and the distance of the language pair. We have used the neural Transformer architecture learned over the provided resources and we carried out synthetic data generation experiments which aim at alleviating the data scarcity problem. Results indicate the suitability of the data augmentation experiments, enabling our systems to rank first according to automatic evaluations.

## 1 Introduction

This paper describes the SYSTRAN neural MT systems employed for the $6^{th}$ Workshop on Asian Translation (WAT) (Nakazawa et al., 2019), an open evaluation campaign focusing on Asian languages. This is our first participation in the workshop and the first year the workshop includes the Russian↔Japanese News Commentary task, with the objective of studying machine translation under extremely low resource conditions and for distant language pairs.

The lack of sufficient data together with the distance and richness of the language pair constitute very challenging conditions. A rather common situation in the translation industry, that motivated us to explore techniques that can help in the construction from scratch of efficient NMT engines. We present systems built using only the data provided by the organisers for both translation directions (Russian↔Japanese) and using the Transformer network introduced by (Vaswani et al., 2017). We enhance the baseline systems with several experiments that aim at alleviating the data scarcity problem. More precisely we run experiments following the back-translation method proposed by (Sennrich et al., 2016b) in which target monolingual corpora are translated back into the source language. Thus, creating synthetic parallel data. In addition, we present an updated version of back-translation where synthetic data is created with higher diversity by means of side constraints.

The remaining of this paper is structured as follows: We first describe statistics of the datasets provided in Section 2. Section 3 outlines our neural MT system. In Section 4 we detail the data augmentation methods employed to alleviate data scarcity. Experiments are reported in Section 5. We analyse results in Section 6 and conclude in Section 7.

## 2 Resources

Datasets used for the evaluation can be found listed in the shared task web site[1]. WAT organisers kindly provide a manually aligned, cleaned and filtered Japanese↔Russian, Japanese↔English and English↔Russian train, development and test corpora (JaRuNC)[2] as well as a news domain Russian↔English corpus (NC)[3]. In addition, use of the next out-of-domain data is encouraged:

- Japanese↔English Wikipedia articles related to Kyoto (KFTT)[4].
- Japanese↔English Subtitles (JESC)[5],
- Japanese↔English asian scientific paper abstracts (ASPEC)[6],

[1] lotus.kuee.kyoto-u.ac.jp/WAT/WAT2019
[2] github.com/aizhanti/JaRuNC
[3] lotus.kuee.kyoto-u.ac.jp/WAT/News-Commentary/news-commentary-v14.en-ru.filtered.tar.gz
[4] www.phontron.com/kftt/
[5] datarepository.wolframcloud.com/resources/Japanese-English-Subtitle-Corpus
[6] lotus.kuee.kyoto-u.ac.jp/ASPEC/

*Proceedings of the 6th Workshop on Asian Translation*, pages 189–194
Hong Kong, China, November 4, 2019.

- Russian↔English transcriptions of TED talks (TED)[7],

- Russian↔English pair of the United Nations Parallel Corpus (UN)[8],

- The Russian↔English Yandex corpus v1.3 (Yandex)[9]

Statistics of the training bitexts are shown by Table 1, summarising for each language the total number of sentences, running words, vocabulary size and average sentence length. Note that despite listed as an official resource we do not use the ASPEC corpus as we never received the download link of the ASPEC corpus from the corpus owners. Statistics are computed after performing a light tokenisation by means of the OpenNMT tokeniser[10] (aggressive mode) which basically splits-off punctation. No additional parallel resources are used in our experiments.

| Bitext | | sent. | words | vocab. | $L_{mean}$ |
|---|---|---|---|---|---|
| JaRuNC | ja | 47.1K | 1.3M | 48.3K | 26.9 |
| | en | | 1.0M | 51.2K | 22.1 |
| KFTT | ja | 440K | 10.9M | 118K | 24.9 |
| | en | | 12.0M | 173K | 27.2 |
| JESC | ja | 2.8M | 23.2M | 155K | 8.3 |
| | en | | 25.6M | 133K | 9.2 |
| ASPEC | ja | - | - | - | - |
| | en | | - | - | - |
| JaRuNC | ru | 82.1K | 1.7M | 140K | 20.1 |
| | en | | 1.9M | 67.0K | 23.0 |
| NC | ru | 279K | 7.1M | 204K | 25.5 |
| | en | | 7.6M | 67.5K | 27.2 |
| TED | ru | 185K | 3.3M | 165K | 17.7 |
| | en | | 3.9M | 58.5K | 21.0 |
| Yandex | ru | 1.0M | 22.9M | 704K | 23.0 |
| | en | | 25.2M | 322K | 25.2 |
| UN | ru | 11.7M | 309M | 870K | 26.5 |
| | en | | 340M | 408K | 29.2 |
| JaRuNC | ru | 12.4K | 235K | 42.0K | 19.0 |
| | ja | | 341K | 21.9K | 27.6 |

Table 1: *Statistics of training bitexts. Note that K stands for thousands and M for millions.*

Table 2 illustrates statistics of the development and test sets extracted from the corresponding JaRuNC corpora. We now include the number of out-of-vocabulary words. As it can be seen, Japanese↔Russian parallel resources are extremely scarce with only 12,4K sentence pairs.

| Side | sent. | words | vocab. | $L_{mean}$ | OOV |
|---|---|---|---|---|---|
| Development (JaRuNC) | | | | | |
| ja | 589 | 21.5K | 3.5K | 36.4 | 288 |
| en | | 16.4K | 3.7K | 27.9 | 273 |
| ru | 313 | 7.6K | 3.2K | 24.3 | 278 |
| en | | 8.3K | 2.3K | 26.4 | 83 |
| ru | 486 | 11.2K | 4.4K | 23.1 | 1297 |
| ja | | 16.0K | 2.9K | 33.0 | 470 |
| Test (JaRuNC) | | | | | |
| ja | 600 | 22.5K | 3.5K | 37.5 | 302 |
| en | | 16.9K | 3.7K | 28.2 | 316 |
| ru | 600 | 15.6K | 5.6K | 25.9 | 661 |
| en | | 16.9K | 3.7K | 28.2 | 223 |
| ru | 600 | 15.6K | 5.6K | 25.9 | 1873 |
| ja | | 22.5K | 3.5K | 37.5 | 661 |

Table 2: *Statistics of development and test sets.*

## 3 Neural MT System

We use the state-of-the-art Transformer model (Vaswani et al., 2017) implemented in `OpenNMT-tf`[11] toolkit (Klein et al., 2017). A neural network following the encoder-decoder architecture, where:

- Each word $x_j$ in the input sentence $x_1^J$ is encoded in a continuous space. Fixed positional embeddings are also added to the word vectors to represent a word embedding $\bar{x}_j$.

- The encoder is a self-attentive module that maps an input sequence of words $\bar{x}_1^J$ into a sequence of continuous representations $h_1^J$.

  $$h_1^J = H_{enc}(\bar{x}_1^J; \theta_{enc})$$

  where $\theta_{enc}$ are encoder parameters.

- The decoder is also a self-attentive module that at each time step outputs a single hidden state $s_i$, conditioned on the sequence of previously seen embedded target words $\bar{y}_{<i}$ and the encoder outputs $h_1^J$.

  $$s_i = H_{dec}(h_1^J, \bar{y}_{<i}; \theta_{dec})$$

  where $\theta_{dec}$ are decoder parameters.

[7] wit3.fbk.eu

[8] cms.unov.org/UNCorpus/

[9] translate.yandex.ru/corpus?lang=en

[10] pypi.org/project/pyonmttok/

[11] github.com/OpenNMT/OpenNMT-tf

- The hidden state $s_i$ is projected to the output vocabulary and normalised with a $softmax$ operation resulting in a probability distribution over target words.

$$p(y_i|y_{<i}, x_1^J) = softmax(W \cdot s_i + b)$$

## 4 Data Augmentation

### 4.1 Back-translation

We follow the *back-translation* method proposed by (Sennrich et al., 2016b) in which target monolingual corpora are translated back into the source language. This synthetic parallel data is then used in combination with the actual parallel data to further train the model. This approach yields state-of-the-art results even when large parallel data are available, currently common practice in academia and industry scenarios (Poncelas et al., 2018).

### 4.2 Side Constraints

We propose a method to generate synthetic parallel data that uses a set of side constraints. Side constraints are used to guide the NMT model to produce distinct word translation alternatives based on their frequency in the training corpora. Furthermore, we employ a set of grammatical constraints (tense, voice and person) which introduce syntactic/semantic variations in translations. Thus, our method aims at enhancing translation diversity, a major drawback highlighted in back-translated data (Edunov et al., 2018). Similar to our work, side constraints have already been used on neural models in a number of different scenarios. To the best of our knowledge, side constraints were first employed to control politeness in a NMT by (Sennrich et al., 2016a). Domain-adapted translations using a unique network enhanced with side constraints is presented in (Kobus et al., 2017).

We consider 4 constraints regarding POS classes: *noun*, *verb*, *adjective* and *adverb*. For each constraint we build 3 clusters containing the set of words with `H` (high), `M` (medium) and `L` (low) frequency as computed over the training data. This is, the set of nouns occurring with highest frequency are arranged in the `NH` class, verbs with lower frequencies in `VL`, *etc.* We set the frequency thresholds to satisfy that the three clusters of a POS class have approximately the same number of occurrences in the training corpus.

Training source sentences are then tagged with the values seen on the corresponding target sentences of each POS class. Note that when a target sentence contains different values of a POS class, i.e.: two *nouns* one with high (`H`) frequency and another with low (`L`) frequency, or when no word is found belonging to one class we then use the value `N` (None). For instance, given the Russian sentence: Президент приезжает завтра (*the president arrives tomorrow*) we use as side constraints: `VH`, `NH`, `AN`, `RH`, since приезжает is a verb, президент is a noun and завтра is an adverb of high frequency, while adjectives do not appear in the sentence. Thus, the Japanese-Russian parallel sentence with corresponding side constraints illustrated in Table 3 is used in training to feed the model.

| |
|---|
| VH NH AN RH 明日大統領が到着します |
| ⇝ Президент приезжает завтра |

Table 3: French-German sentence pair with frequency constraints.

Note that when creating synthetic corpora, side constraint values are randomly generated to allow larger diversity of the generated language.

## 5 Experiments

### 5.1 Data Preprocessing

Before learning the translation network, data corresponding to each language is preprocessed following a similar workflow: word tokenisation + subword tokenisation. Tokenisation for English and Russian is performed using the OpenNMT tokeniser (aggressive mode). Japanese tokenisation is carried out by the MeCab[12] tokeniser. For subword tokenisation we trained a 30K byte-pair encoding (Sennrich et al., 2016c) (BPE) of each language, using separately English, Russian and Japanese training data.

### 5.2 Baseline Transformer

In order to alleviate the data scarcity problem, we introduce English as a third language in our baseline system to built a multi-lingual translation system following the work in (Firat et al., 2016). We concatenate both directions of all available Japanese-Russian, Japanese-English and Russian-English corpora to train our `base` model. We in-

[12] `github.com/taku910/mecab`

clude an additional token to the beginning of each source sentence to indicate the related target language (i.e. *@ru@* for Russian). In inference, the corresponding token (*@ru@* or *@ja@*) is used to request Russian or Japanese translation. Similarly, we consider an additional token to indicate whether the training sentence pair is in-domain or out-of-domain (i.e. *@in@* for in-domain data). All JaRuNC corpora and NC English-Russian corpus are considered in-domain data, the rest are deemed out-of-domain. In inference, translations are performed appending the *@in@* token.

Since BPE vocabularies were separately built for each language with 30K tokens, we then use a vocabulary of size 90K tokens for both source and target sides. Thus, covering all English, Russian and Japanese training data.

We train our model using the standard Transformer base model. We use Lazy Adam optimiser with the same learning rate decay schedule as (Vaswani et al., 2017). Learning rate is updated every 8 steps. We build our baseline model using a batch size of 3,072 over 400k steps on one GPU. The final models result of averaging the last 10 saved checkpoints in training.

### Fine-tuning

We build a second network after fine-tuning the previous baseline network. For fine-tuning we use all in-domain data. More precisely Japanese-Russian (JaRuNC), Japanese-English (JaRuNC) and Russian-English (JaRuNC and NC) datasets in both translation directions (`+FT(JaRuNC,NC)`). Fine-tuning is performed during 80K additional steps for Japanese→Russian and 50K steps for Russian→Japanese.

### Back-translation

We use the previously fine-tuned model to back-translate in-domain Russian and Japanese sentences of our datasets (aligned to English). This is, we back-translate the Japanese side of the Japanese-English (JaRuNC) corpus to extend the data available for the Russian→Japanese translation direction. Equivalently, we backtranslate the Russian side of the Russian-English (JaRuNC and NC) corpora to increase the amount of data available for the Japanese→Russian direction. Thus, building new synthetic Japanese*-Russian and Russian*-Japanese corpora. [13] Since our model is multi-lingual, we don't need additional networks to back-translate both Russian and Japanese sentences.

Given that we synthesised a larger number of Japanese*-Russian sentences than Russian*-Japanese we further synthesise additional data following another approach. We use English in-domain sentences (JaRuNC and NC) to produce Russian and Japanese translations. Thus, new Japanese*-Russian* synthetic bitexts become available. Translations are performed using two distinct uni-directional English→Russian and English→Japanese models. Following the same parameterisation used for the `base` model we train a new model considering all previous parallel data (`+FT(JaRuNC,NC,BT,SYN)`). Notice that following the same multi-stage strategy used in (Imankulova et al., 2019) our new model is built from `+FT(JaRuNC,NC)`.

We also used our fine-tuned multi-lingual model to translate English sentences. However, the translation quality of the multi-lingual model is much poorer than uni-directional models. Thus, hurting the performance of the final model.

### Side Constraints

As introduced in Section 4.2 we perform experiments synthesising using side constraints. Each Russian and Japanese source-side training sentence is extended with the side constraints previously described (corresponding to random frequency values of verbs, adjectives, nouns and adverbs) and are used as input sentences in order to generate the corresponding Japanese and Russian hypotheses. The synthesised parallel data is used together with the previous datasets to learn a new model (`+FT(JaRuNC,NC,BT,SYN,SC)`). Notice again that our new model is built from `+FT(JaRuNC,NC)`.

## 6 Evaluation

All our results are computed following the BLEU (Papineni et al., 2002) score. Validation sets are used to select our best performing networks, while results shown in Table 5 are computed for the official test sets.

As it can be seen, all our experiments to alleviate data scarcity boosted translation performance. A light decrease in accuracy is observed when using `SC` data for Russian→Japanese translation. The improvement is remarkable for the Japanese→Russian task for which the BLEU score is doubled from 7 to more than 14 points.

[13] We use * to denote synthetic data

| System | Ru-Ja | Ja-Ru |
|---|---|---|
| `base` | 9.76 | 6.95 |
| `+FT(JaRuNC,NC)` | 12.10 | 9.17 |
| `+FT(JaRuNC,NC,BT,SYN)` | 15.89 | 13.78 |
| `+FT(JaRuNC,NC,BT,SYN,SC)` | 15.39 | 14.36 |

Table 4: *BLEU score on JaRuNC testset.*

A final experiment is carried out considering our best performing setting so far. We repeat training work with a larger batch size of 6,144 during 250K iterations and using 3 GPUs.

| batch size | Ru-Ja | Ja-Ru |
|---|---|---|
| 3,072 | 15.89 | **14.36** |
| 6,144 | **16.41** | - |

Table 5: *BLEU score using a batch size = 6,144 with 3 GPUs.*

Given the tight schedule to submit our translations, we only run the experiment for the Russian→Japanese task. Bold figures indicate the BLEU scores of the best performing systems submitted for the evaluation.

## 7 Conclusions

We described SYSTRAN's submissions to WAT 2019 Russian↔Japanese News Commentary task. A challenging translation task due to the extremely low resources available and the distance of the language pair. Several data generation experiments were performed in order to alleviate data scarcity, one of the major difficulties of the translation task. Results showed the suitability of the experiments that boosted translation performance in both translation directions allowing our systems to rank first according to automatic evaluations.

## Acknowledgements

We would like to thank the reviewers for their insightful comments.

## References

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 866–875, San Diego, California. Association for Computational Linguistics.

Aizhan Imankulova, Raj Dabre, Atsushi Fujita, and Kenji Imamura. 2019. Exploiting out-of-domain parallel data through multilingual transfer learning for low-resource neural machine translation. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, pages 128–139, Dublin, Ireland. European Association for Machine Translation.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

Catherine Kobus, Josep Crego, and Jean Senellart. 2017. Domain control for neural machine translation. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 372–378, Varna, Bulgaria. INCOMA Ltd.

Toshiaki Nakazawa, Chenchen Ding, Raj Dabre, Hideya Mino, Isao Goto, Win Pa Pa, Nobushige Doi, Yusuke Oda, Anoop Kunchukuttan, Shantipriya Parida, Ondřej Bojar, and Sadao Kurohashi. 2019. Overview of the 6th workshop on Asian translation. In *Proceedings of the 6th Workshop on Asian Translation*, Hong Kong. Association for Computational Linguistics.

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 on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Alberto Poncelas, Dimitar Shterionov, Andy Way, Gideon Maillette de Buy Wenniger, and Peyman Passban. 2018. Investigating backtranslation in neural machine translation. *CoRR*, abs/1804.06189.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Controlling politeness in neural machine translation via side constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 35–40, San Diego, California. Association for Computational Linguistics.

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

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016c. 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)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. 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.