PROMT Systems for WMT 2018 Shared Translation Task

Alexander Molchanov
PROMT LLC
17E Uralskaya str. building 3, 199155,
St. Petersburg, Russia
Alexander.Molchanov@promt.ru

Abstract

This paper describes the PROMT submissions for the WMT 2018 Shared News Translation Task. This year we participated only in the English-Russian language pair. We built two primary neural networks-based systems: 1) a pure Marian-based neural system and 2) a hybrid system which incorporates OpenNMT-based neural post-editing component into our RBMT engine. We also submitted pure rule-based translation (RBMT) for contrast. We show competitive results with both primary submissions which significantly outperform the RBMT baseline.

1 Introduction

This paper provides an overview of the PROMT submissions for the WMT 2018 Shared News Translation Task. This year we participate with neural MT systems for the first time. We participate only in the English-Russian language pair, but with three different systems.

The paper is organized as follows: Section 2 is a brief overview of the submitted systems. Section 3 describes the data preparation, preprocessing and statistics in detail. Section 4 provides a description of the systems. In Section 5 we present and discuss the results. Section 6 concludes the paper.

2 Systems overview

We submitted three systems for the WMT 2018 Shared News Translation Task:

- A (almost) pure NMT system based on the Marian (Junczys-Dowmunt et al., 2018) toolkit. The system features a rule-based names processing module and backoff to RBMT baseline in a few cases.
- A hybrid NMT system based on the PROMT RBMT engine with OpenNMT-based (Klein et al., 2017) neural post-editing module.
- pure RBMT system.

3 Data

We use the data provided by the WMT organizers, some private in-house news parallel data (approximately 600k parallel sentences crawled from various news web-sources and dated between 2015 and 2017) and the TED Talks corpus from the OPUS website (Tiedemann, 2012). The NewsCommentary, TED and in-house corpora are used as is.

We do not use any data for fine-tuning. We use the WMT newstest2017 set as our validation set. We also report results for newstest2018.

3.1 Data filtering

The CommonCrawl and (especially) ParaCrawl corpora were heavily filtered and normalized using the PROMT tools and algorithms (including language recognition, removal of meaningless sentences, in-house tools for parallel sentences classification, spellchecker etc.). We discarded roughly 50% of the CommonCrawl and 60% of the ParaCrawl data.

The MultiUN corpus was only checked for sentence length ratio using a simple rule-based algorithm. Less than 1% of the original data was discarded.

After that, we applied the bilingual data selection algorithm (Axelrod et al., 2011) to the filtered versions of ParaCrawl and MultiUN. We use the English and Russian news 2016-2017

Proceedings of the Third Conference on Machine Translation (WMT), Volume 2: Shared Task Papers, pages 460–464
Brussels, Belgium, October 31 - November 1, 2018. ©2018 Association for Computational Linguistics
https://doi.org/10.18653/v1/W18-6420

corpora from statmt.org as the in-domain corpora. After this procedure we selected 1.5M sentences from the ParaCrawl corpus and 6M sentences from the MultiUN corpus.

The final statistics for the training data are shown in Table 1.

3.2 Data preprocessing

Pure NMT system

We adopt a standard preprocessing scheme using the scripts provided by the Marian toolkit. The data is tokenized using the Moses toolkit (Koehn et al., 2007) tokenizer; after that we apply truecasing and, finally, byte pair encoding (BPE) (Sennrich et al., 2016) with 85K operations for source and target. We do not use a shared vocabulary due to the Cyrillic nature of Russian alphabet.

Hybrid NMT system

We adopt a slightly different pipeline for the OpenNMT-based system. The data is tokenized with the OpenNMT tokenizer. The tokenizer provides a nice and handy option of applying the case feature, thus there is no need for truecasing. Then, we apply BPE with the same size 85K operations for source and target using the OpenNMT BPE script. The OpenNMT BPE learning algorithm is an extended version of the original BPE script adopted in Marian and has the following additional features: 1) the BPE merge operations are learnt to distinguish subword units at the beginning, in the middle and at the end of the word and 2) the BPE merge operations are learnt in case-insensitive mode (as we use the case feature to handle that). The OpenNMT system architecture does not support shared embeddings so despite the fact that both source (RBMT translations) and target (human translations) data is encoded in Cyrillic we train separate BPE models.

3.3 Synthetic data

We use three types of additional synthetic data described in detail below. The final size of the training data for the pure NMT system is roughly 4 times the total size of the filtered data in Table 1, while the final size of the training data for the hybrid system is approximately 6 times the size of the filtered data.

Corpus	#sent	#tokens EN	#tokens RU
MultiUN	6.0	140.8	129
ParaCrawl	1.5	28.4	24.3
Yandex corpus	0.6	16.8	15.4
Private data	0.6	15.6	15
CommonCrawl	0.4	10.3	9.5
NewsCommentary	0.3	6.2	5.9
TED Talks	0.1	2.4	2.1
Total	**9.5**	**220.5**	**201.2**

Table 1: Statistics for the filtered parallel English-Russian data in millions of sentences (#sent) and tokens.

Back-translated data

Using the filtered data presented in Table 1 we train two initial auxiliary target-to-source NMT systems using the filtered data:

- A Russian-English NMT system using Marian (s2s with default parameters);
- A Russian-to-RBMT NMT system using OpenNMT (dbrnn, 2 layers, RNN size 1024 units).

The trained systems are then used to back-translate the 2017 news corpus from statmt.org (in case with the Marian system, we translate from Russian into English; the OpenNMT systems translates from Russian into the "Rule-based Russian", mimicking the rule-based machine translation accent and structure). The size of the synthetic corpus is approximately equivalent to the size of human training data.

Replicated data with unknown words

Similar to (Pinnis et al., 2017), we again roughly double our parallel data by creating a synthetic parallel corpus using the following steps: first, we perform word-alignment of our initial parallel training corpus using the MGIZA tool (Gao and Vogel, 2008). Then, we randomly replace from one to three unambiguously (one-to-one) aligned subword units in both source and target parallel sentences with the special <UNK> placeholder. The same pipeline is applied to both pure NMT system (for which we augment the English-Russian corpus) and the hybrid NMT system (for which we augment the RBMT-human Russian corpus) and to both the initial and back-translated data.

Monolingual data

To benefit from the fact that we have data in Cyrillic in both source (RBMT) and target (human Russian) when dealing with the hybrid system, we add the 2017 Russian news corpus from statmt.org to the source side of the training data of the hybrid NMT system and replicate it on the target side. Currey et al. (2017) claim that this technique can yield improvements for translation of named entities. The BPE models learnt on the initial training data are applied.

4 Systems architecture

This section describes the trained systems in detail.

4.1 RBMT system

The PROMT RBMT System is a mature machine translation system with huge linguistic structured databases containing morphological, lexical and syntactic features for most European and Russian languages. We did not do any specific tuning for our submission.

4.2 Pure NMT system

For the pure NMT system we train a transformer (Vaswani et al., 2017) model. We use the recipe available at the Marian website [1]. The system configuration, hyperparameters and training steps follow those in the recipe. There are two minor differences: 1) we check the validation translation less frequently and set a higher early-stopping threshold to allow the model iterate over the training data for several epochs; 2) we do not use shared vocabulary because of the different alphabets in English and Russian. First, we trained the baseline system on the initial parallel data and back-translated data. After that, we trained 4 other models with different seeds using the whole data augmented with unknown words (see section 3.3).

Model configuration

We use an ensemble of all 5 transformer models as our baseline translation system; in addition, we use RBMT as our back-off system (this will be described in detail in the next section). We use the beam of size 12 and the "normalize" parameter is set to 1.

[1] https://github.com/marian-nmt/marian-examples/tree/master/wmt2017-transformer

Back-off to RBMT

At first we had in mind training a classifier to choose when to fall back to the RBMT model. However, linguistic analysis of the neural translation of the validation set showed us that the NMT output is of good quality. We only encountered two minor problems: 1) the model sometimes outputs English text (less than 1% of the validation set sentences) and 2) from time to time the decoder outputs multiple recurring words or n-grams (this is a well-known problem of NMT systems). We deal with both problems using simple rules. First, the model output is checked using language recognition tool. If the language is other than the Russian, we fall back to the RBMT translation. Additionally, we check the neural translation for recurring words or n-grams: if a word recurs more than twice or an n-gram recurs more than once, we also fall back to the RBMT system.

Handling proper names

We noticed that our transformer models have a problem translating proper names, especially rare ones or the ones not seen in the training data. Linguistic analysis led us to the conclusion that problems occur most often with the proper names which either 1) appear less than 5 times in the training corpus or 2) are split by the BPE model. To deal with this issue, we developed the following pipeline. We use the Stanford NER tool (Finkel et al., 2005) to identify proper names in the source text (person names, organizations and locations). We check the name frequency in the training data and whether it is split by the BPE model. If the frequency of any part of the name is low or it is split, we replace the whole name with the <UNK> placeholder in the source sentence. Then we translate the sentence by an ensemble of 4 models trained to reproduce unknown words allowing the decoder to reproduce unknown words in the output. Finally, we substitute the <UNK> placeholders in the output with the translations of the names produced by the RBMT system. If for some reason we can't match the names to their RBMT translations or the number of the <UNK> placeholders in the NMT system output is not equal to the number of the placeholders in the source sentence, we fall back to the baseline NMT system described in Subsection 4.2 above.

Source sentence	NMT	NMT+names	Reference
The Russians represented in qualifying were Anton Chupkov, Evgeny Koptelov, Alexander Sukhorukov, and Grigory Tarasevich.	В квалификации россияне представляли Антона Чупакова, Евгения Коптева, Александра Сухокова и Григория Тараскевича.	В квалификации были представлены россияне Антон Чупков, Евгений Коптелов, Александр Сухоруков и Григорий Тарасевич.	Россиян в квалификации представили Антон Чупков, Евгений Коптелов, Александр Сухоруков и Григорий Тарасевич.
They all lived in the small town of Greenfield, Massachusetts.	Все они жили в небольшом городе Гринсфилл, штат Техас.	Все они жили в небольшом городе Гринфилд, Массачусетс.	Все они жили в небольшом городке Гринфилд в штате Массачусетс.

Table 2. Examples of translation with names processing. The NMT+names stands for the system with proper names processing as described in Section 4.2.

4.3 Hybrid NMT system

We mentioned earlier that OpenNMT does not support the transformer model architecture. Due to this fact we train a model with a deep bidirectional encoder and a decoder with attention (Luong et al., 2015). Both encoder and decoder consist of two layers each with 1024 hidden units. The word embeddings size is 500 and the case feature embeddings size is 4. As with the pure NMT system, we first trained a baseline model on the initial parallel data and back-translated (Russian-to-RBMT) data. After that, we retrained the baseline model on the whole data augmented with unknown words and monolingual data (see Section 3.3 for details). We train the baseline model for 8 epochs and then retrain the model on all data for two more epochs. We use the beam of size 8 for translation.

Linguistic analysis of the translation of the validation set didn't show any problems regarding the NMT post-editing component. Thus, we made a decision not to make any special processing of names or fall back to RBMT and submit the hybrid post-edited translation as is.

5 Results and discussion

In this section we present the BLEU (Papineni et al., 2002) scores for our systems on two test sets and the linguistic analysis of the results.

The scores are presented in Table 3. Calculation is done using the `multi-bleu-detok.perl` script from the Moses toolkit.

We also studied the impact of the proper names processing applied to the NMT translation. Our pipeline affected 815 (27%) out of 3000 sentences in the test set. As we can see, unfortunately the BLEU is a bit lower than for the default translation. We see two reasons for that: first, we lose precision because frequently a name, even translated correctly, appears in the wrong case in the output. Russian is a highly inflective language and this is a problem. Marian does not support factored translation yet, so we couldn't teach the system to output the case feature for our placeholders. Secondly, the system was trained to reproduce placeholders for subword units and not the whole words, as we generated the synthetic data from the already BPE-segmented parallel bitexts. We chose, however, the translation with names processing to be our final submission as we decided that a system which is a little less fluent but more accurate at translating names would be better. Examples of translations with and without the names processing can be found in Table 2.

6 Conclusions and Future work

In this paper we have described our English-Russian submissions for the WMT 2018 Shared News Translation Task. Overall we have made three submissions: 1) a pure NMT system developed with the Marian toolkit, 2) a hybrid system with a NMT post-editing component

System	newstest2017	newstest2018
RBMT	22.9	18.1
NMT	**31.0**	**27.4**
NMT+names	**30.9**	**27.3**
Hybrid	29.5	25.3

Table 3: Results for the submitted systems. The NMT+names stands for the system with proper names processing as described in Section 4.2.

trained with the OpenNMT toolkit, and 3) pure RBMT system.

The pure NMT system with the state-of-the-art transformer architecture proved to be the best among our submissions in terms of BLEU.

We also present a names processing and translation pipeline which can be improved by teaching the system to output the translations in the correct case.

References

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain adaptation via pseudo in-domain data selection. In *Proceedings of the ACL Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 355–362, Edinburgh, Scotland, UK.

Anna Currey, Antonio Valerio Miceli Barone, and Kenneth Heafield. 2017. Copied monolingual data improves low-resource neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, pages 148–156, Copenhagen, Denmark.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, pages 363–370, Ann Arbor, MI, USA.

Qin Gao and Stephan Vogel. 2008. Parallel implementations of word alignment tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing, SETQA-NLP 08*, pages 49–57, Stroudsburg, PA, USA.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast Neural Machine Translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, Alexander M. Rush. 2017. OpenNMT: Open-Source Toolkit for Neural Machine Translation. *Computing Research Repository*, arXiv:1701.02810. Version 2.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, ACL 07*, pages 177–180, Stroudsburg, PA, USA.

Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (2015)* pages 1412–1421, Lisbon, Portugal.

Kishore Papineni, Salim Roukos, Todd Ward, and WeiJing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 02)*, pages 311–318, Philadelphia, PA, USA.

Marcis Pinnis, Rihards Krišlauks, Daiga Deksne, and Toms Miks. 2017. Neural Machine Translation for Morphologically Rich Languages with Improved Sub-word Units and Synthetic Data. In *Proceedings of the 20th International Conference of Text, Speech and Dialogue (TSD2017)*, pages 237–245, Prague, Czechia.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. 2016. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (2016)*, pages 35–40, San Diego, CA, USA.

Michel Simard, Cyril Goutte, and Pierre Isabelle. 2007. Statistical phrase-based post-editing. In *Proceedings of The North American Chapter of the 342 Association for Computational Linguistics Conference (NAACL-07)*, pages 508–515, Rochester, NY, USA.

Jorg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, USA.