PROMT Systems for WMT23 Shared General Translation Task

Alexander Molchanov & Vladislav Kovalenko
PROMT LLC
17E Uralskaya str. building 3, 199155,
St. Petersburg, Russia
First.Last@promt.ru

Abstract

This paper describes the PROMT submissions for the WMT23 Shared General Translation Task. This year we participated in two directions of the Shared Translation Task: English to Russian and Russian to English. Our models are trained with the MarianNMT toolkit using the transformer-big configuration. We use BPE for text encoding, both models are unconstrained. We achieve competitive results according to automatic metrics in both directions.

1 Introduction

The WMT Shared General Translation Task is an annual event where different companies and researchers build and test their systems on the test sets provided by the organizers. This year we decided to participate in two directions: English to Russian and Russian to English. We use the standard transformer-big configuration for our models. The English-Russian model is basically the same as last year, whereas the Russian-English model is a new one built for WMT23.

The rest of the paper is organized as follows: in Section 2 we describe in detail the systems we submitted to the Shared Task. In Section 3 we present and discuss the results. We conclude the paper in Section 4 with discussion for possible future work.

2 Systems overview

All of our WMT22 submissions are `MarianNMT`-trained (Junczys-Dowmunt et al., 2018) transformer-big (Vaswani et al., 2017) systems. We use the `OpenNMT` toolkit (Klein et al., 2017) version of byte pair encoding (BPE) (Sennrich et al., 2016b) for subword segmentation. Our BPE models are case-insensitive, we use special tokens in the source and target sides to process case (see Molchanov (2019) for details).

All of the systems are unconstrained, i.e. we use all data provided by the WMT organizers, all publicly available data and some private data crawled from different web-sources.

We also augment our training data with two types of synthetic data: 1) back-translations (Sennrich et al., 2016a) and 2) synthetic data with placeholders as described in Pinnis et al. (2017). The back-translations are obtained using the previous versions of our NMT models which are baseline transformers trained with less data (and without some up-to-date data like the news 2021 corpora from statmt.org). We also tag all our synthetic data with special tokens at the beginning of the source sentences as described in Caswell et al. (2019).

All models are trained with guided alignment which is used at translation time to handle named entities and document formatting. We obtain alignments using the `fast-align` (Dyer et al., 2013) tool.

The data statistics for the Russian-English language pair are presented in Table 1.

The details regarding different directions can be found in the next Section.

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 150–154
December 6–7, 2023.

	Russian-English	
	#sent	#tokens RU
WMT+OPUS	37.4	690.9
Private	30.2	542.2
Total	67.6	1233.1

Table 1: Statistics for the filtered human parallel data in millions of sentences (#sent) and tokens (#tokens) for the English-Russian language pair. WMT stands for the data available for the News Task on the statmt.org/wmt22 website; OPUS is the data from the OPUS website apart from the data available for the News Task; Private stands for private company data.

2.1 Data preparation

There are several stages in our data preparation pipeline. These are mostly common filtering techniques. The main stages of the pipeline are:

- Basic filtering
  This includes some simple length-based and source-target length ratio-based heuristics, removing tags, lines with low amount of alphabetic symbols etc. We also remove lines which appear to be emails or web-addresses and duplicates.

- Language identification
  The algorithm is a fairly simple ensemble of three tools: `pycld2`[1], `langid` (Lui and Baldwin, 2012), `langdetect`[2]. We mainly use pycld2 as it is by far the fastest tool of the three. If pycld2's output differs from the hint language, we perform additional checks using the other two libraries, and the final language is determined by majority vote. For large monolingual corpora we use only pycld2.

- Bicleaner filtering
  We use the bicleaner (Ramírez-Sánchez et al., 2020) tool to filter parallel data. We discard all sentence pairs with the score threshold <= 0.3.

- Scoring with NMT models
  We finally score all parallel data and back-translations with our intermediate models. We use a score threshold to discard a portion of the data. The exact threshold is determined by human evaluation. The discarded data includes non-parallel sentences (i.e. pairs of sentences where the source does not correspond to the target in part or fully) and low-quality synthetic translations.

- Dual conditional cross-entropy filtering
  This year we use this algorithm again for both directions as described in Junczys-Dowmunt (2018).

2.2 English-Russian

The English-Russian system is basically the same as last year (Molchanov et al., 2022). It was trained in two steps. First, we build the baseline model on all available data. Second, we fine-tune the model on data of high quality. Specifically, we remove the ParaCrawl, UN and OpenSubtitles corpora. The training corpus then consists of the remains of the human data mixed with the back-translations of the news corpora (2020, 2021) from statmt.org. This approach shows good results according to automatic metrics and general translation quality. The reason for doing this is that we aim for our models to be used mostly for translation of news and formal texts like various types of documents.
The system was trained with separate vocabularies, the sizes of the BPE models are 24k for the source side and 48k for the target side.

[1] https://pypi.org/project/pycld2/
[2] https://pypi.org/project/langdetect/

source	Model2022	Model2023	Model2023 fixed
Перенасадка башмаков и колец для колпаков, замена вентилей должны	Overpressure of shoes and rings for caps, replacement of valves shall	\\\\\\\{ 1\\\\\\\> ·\\\\\\{ 1\\\\\\}\\\\\\{ 2\\\\\\}\\\\\\< 3\\\\\\\}	Re-fitting of shoes and rings for caps, replacement of valves should
Прогноз компонентов ВВП по использованию на 2019 г. несколько изменился, что связано прежде всего с выходом фактических данных за II квартал 2019 года.	The forecast of GDP components for use for 2019 has changed somewhat, which is primarily due to the release of actual data for the second quarter of 2019.	2019 GDP Components Forecast by Usage The Bank of Russia's monetary policy is based on the following principles:	The forecast of GDP components for 2019 slightly changed, which is primarily due to the release of actual data for the second quarter of 2019.

Table 2: Examples of degradations for the 2023 Russian-English model.

2.3 Russian-English

The Russian-English model was built basically on the same data and in the same way as the English-Russian model. The only difference is that we use English news and Wikipedia for back-translations. The previous version of the Russian-English model was also built on the same data, but with the transformer-base configuration.

The first version of the model that we trained on this data had shown almost no improvements, both in terms of automatic and human evaluation (on average the model improved by 0.5 BLEU points on our internal test sets compared to Model2022 using the transformer-base configuration). What is more important is that we observed some serious degradations: hallucinations and critical mistakes. The examples are presented in Table 2. We investigated the problem and found out that some of our clients' data was used for training without proper filtering. This was part of our private data. We then applied the full filtering pipeline to the private data and discarded around 20k sentence pairs (roughly 0.03% of all data) with low quality. Then we retrained the model on the filtered data, and this fixed all the critical mistakes we had encountered. Surprisingly, we also gained additional 1 BLEU points on average on our internal test sets compared to the first version. All we did was just remove 0,03% of bad sentence pairs from the training data. The average BLEU score on our test sets improved from 36.66 to 38.05 points.

3 Results and discussion

The results are presented in Table 3.

As we can see, we outperform our baselines (i.e. previous versions of the models). The gains we observe, however, are not that large.

System	BLEU	chrF	COMET
English-Russian			
Model2022	30.5	55.4	82.3
Russian-English			
Model2022	32.4	58.0	-
Model2023	**32.8**	**58.4**	**80,9**

Table 3: Results for different systems in both directions. The submitted systems are marked in bold. Model2022 stands for our previous version of the Russian-English system which we consider the baseline. The English-Russian system remains the same.

However, other test sets, such as the TICO-19 evaluation set [3] (Anastasopoulos et al., 2020), show more substantial improvements. The BLEU score on that test set has grown from 33.8 to 35 points.

Poor performance on the generaltest2023 set can be due to the problems that our submitted models have with translation of colloquial content. This can be explained by our data preparation scheme. As we have already mentioned above, we want our models to translate formal text better and thus 'sacrifice' colloquial data. The examples of such mistranslations are presented in Table 4. Both examples illustrate colloquial slang which our model cannot translate properly. In the first example the word 'please' is substituted by 'pls', and thus the model 'thinks' it is a abbreviation of some kind. In the second example the author substitutes the word 'because' with a slang word 'becuz', and the model transliterates it.

We made a thorough investigation into the generaltest2023 sets. Thus, we found out that there are four major topics for the Russian-English test set: 1) movie reviews; 2) news of any

[3] https://tico-19.github.io/index.html

kind; 3) user reviews; 4) abstracts from research papers in medical domain. The English-Russian test set domains are similar. We estimate that at least half of the English-Russian test set is made up of Reddit posts and online customer reviews which often use internet slang and have spelling anomalies of some kind, e.g. "eye wud liek 2 aply 4 vilage idot", "WhY dO pPl FiNd ThE NeW SeT ExPeNsIvE." All these domains except for news were actually unexpected by our model.

source	Model2022
pls change	Изменение PLS
Becuz i have less than 3000 points.	Бекуз у меня меньше 3000 очков.

Table 4: Examples of incorrect translations for the English-Russian model considering the colloquial content.

4 Conclusions and future work

In this paper we presented our submissions for the WMT23 Shared General Translation Task. We show good results in both directions we participate. We clearly outperform our baselines in both directions. A detailed analysis of the translations shows us that we lose quality in translation of colloquial speech. We have already started to work in this direction. We have synthesized data where, e.g., 'please' is substituted with 'plz' and so on. We plan to train our model on this synthetic data so that it could deal with such colloquial examples.

References

Isaac Caswell, Ciprian Chelba, and David Grangier. 2019. Tagged Back-Translation. In *Proceedings of the Fourth Conference on Machine Translation*, pages 53–63, Florence, Italy.

Chris Dyer, Victor Chahuneau, and Noah A Smith. 2013. A Simple, Fast, and Effective Reparameterization of IBM Model 2. In *Proceedings of NAACL HLT 2013*, pages 644–648, Atlanta, USA.

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.

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.

Marco Lui and Timothy Baldwin. 2012. langid.py: An Off-the-shelf Language Identification Tool. In *Proceedings of ACL 2012, System Demonstrations*, pages 25–30, Jeju, Republic of Korea.

Alexander Molchanov. 2019. PROMT Systems for WMT 2019 Shared Translation Task. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 302–307, Florence, Italy.

Alexander Molchanov, Vladislav Kovalenko, Natalia Makhamalkina. 2022. PROMT Systems for WMT22 General Translation Task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 342–345, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

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.

Gema Ramírez-Sánchez, Jaume Zaragoza-Bernabeu, Marta Bañón and Sergio Ortiz Rojas. 2020. Bifixer and Bicleaner: two open-source tools to clean your parallel data. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 291–298, Lisboa, Portugal.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving Neural Machine Translation Models with Monolingual Data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*, Berlin, Germany.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with

subword units. 2016b. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*, pages 35–40, Berlin, Germany.

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.

Antonios Anastasopoulos, Alessandro Cattelan, Zi-Yi Dou, Marcello Federico, Christian Federmann, Dmitriy Genzel, Franscisco Guzmán, Junjie Hu, Macduff Hughes, Philipp Koehn, Rosie Lazar, Will Lewis, Graham Neubig, Mengmeng Niu, Alp Öktem, Eric Paquin, Grace Tang, and Sylwia Tur. 2020. TICO-19: the Translation Initiative for COvid-19. In Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020, Online. Association for Computational Linguistics.