PROMT Systems for WMT22 Shared General Translation Task

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

Abstract

This paper describes the PROMT submissions for the WMT22 Shared General Translation Task. This year we participated in four directions of the Shared Translation Task: English to Russian, English to German and back, and Ukrainian to English. All our models are trained with the MarianNMT toolkit using the transformer-big configuration. We use BPE for text encoding, all of our models are unconstrained. We achieve competitive results according to automatic metrics in all 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 the Task has shifted from news to the general domain. We participate in four directions: English to Russian, English to German and back, and Ukrainian to English. We build the transformer-big models for the first time. We also explore new data filtering techniques, data preparation and model training strategies.

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.

This year we use the dual conditional cross-entropy (Junczys-Dowmunt, 2018) method for data filtering. We extend the method as proposed by the author and build neural language models for both source and target languages.

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

Proceedings of the Seventh Conference on Machine Translation (WMT), pages 342 - 345
December 7-8, 2022 ©2022 Association for Computational Linguistics

	German-English		Russian-English		Ukrainian-English	
	#sent	#tokens EN	#sent	#tokens EN	#sent	#tokens EN
WMT+OPUS	148.0	4000.1	37.4	690.9	24.8	566.7
Private	8.1	106.8	30.2	542.2	0.5	5.8
Total	156.1	4106.9	67.6	1233.1	25.3	572.5

Table 1: Statistics for the filtered human parallel data in millions of sentences (#sent) and tokens (#tokens) for three language pairs. 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.

alignments using the `fast-align` (Dyer et al., 2013) tool.

The data statistics for different language pairs are presented in Table 1.

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

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]. 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 to discard non-parallel sentence pairs and bad synthetic translations.

- Dual conditional cross-entropy filtering
  This year we use this algorithm for the first time. We apply it to the English-German language pair.

2.2 English-Russian

The English—Russian system 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 totally remove the ParaCrawl, UN and OpenSubtitles corpora and fine-tune the model using 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.

2.3 English-German and German-English

Both models were trained with the same joint vocabulary, the BPE model size is 32k. We use all available human data. We apply basic filtering for some data which we believe to be clean (e.g. private data and high-quality open-source corpora like News-Commentary). The rest of the data is filtered with the modified dual conditional cross-entropy filtering algorithm. We noticed that using only the news corpora as general for filtering as described in Junczys-Dowmunt (2018) results in the fact that the data shifts towards the news domain. For example, a perfectly fine sentence

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

pair related to the IT domain may receive low scores from News models. Therefore, we try to build a general good quality corpus comprising different domains (news, IT, technical data etc.). We do not include colloquial corpora into these general corpora because we intend for our models to be used for translating mostly formal text, be it news, formal letters or technical documents. We set the threshold for the filtering score at 0.1. Thus, we discard around 60-70% of the original data.

2.4 Ukrainian-English

We use a lot of synthetic data for this model. We decided that we could pivot the Ukrainian-English model through our Ukrainian-Russian and English-Russian data and systems. We translate the Russian side of the English-Russian data to Ukrainian and use it as synthetic data for the final model.

The Ukrainian-Russian model is a transformer-base unconstrained model. It was built jointly to translate from Ukrainian into Russian and back. We use all available parallel data and back-translations of the news and Wikipedia corpora. Although this is a transformer-base model, the Ukrainian-Russian language pair is relatively easy for the model to learn properly and achieve very good results in. Thus, we made an assumption that even the big model would benefit from this synthetic data given the fact that the Ukrainian-English is not a high-resource language pair.

To see how much we benefit specifically from using the transformer-big architecture in addition to the synthetic data from the Russian-English pair we also build a transformer-base model for this language pair.

3 Results and discussion

The results are presented in Table 2.

As we can see, we clearly 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			
Model2021	29.1	52.5	0.54
Model2022	**30.6**	**53.8**	**0.60**
English-German			
Model2021	45.3	62.8	0.49*
Model2022	**49.0**	**65.3**	**0.55***
German-English			
Model2021	47.3	62.3	0.51*
Model2022	**49.1**	**63.8**	**0.55***
Ukrainian-English			
Model2021	38.6	60.4	0.44
Model2022 base	39.7	61.3	0.46
Model2022	**41.2**	**62.6**	**0.49**

Table 2: Results for different systems and directions. The submitted systems are marked in bold. The starred scores are averaged scores over two references provided by the organizers. Model2021 stands for our previous versions of the systems which we consider the baseline. Model2022 base stands for the transformer-base configuration of the 2022 model.

We notice that our submitted models have some problems with translation of colloquial content compared to the previous versions. 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 degradations are presented in Table 3. The first example illustrates the problem when short colloquial segments are left untranslated. We think there are two major reasons for that: 1) the fine-tuned model has partially 'forgotten' how to translate colloquial speech; 2) there are many technical and IT-related texts in the fine-tuning data where large constructions (e.g. model or software program names) are left untranslated. Two other examples illustrate bad choice of meaning for specific words from the fine-tuned translation model ('screwed' is translated literally as if the kid was attached to something with a screwdriver; 'кредит' is a word from the financial domain which is inappropriate in this context).

Source text	Model2021	Model2022
You meet me	Встретишь меня	You meet me
And this kid is screwed.	И этот парень облажался.	И этот пацан прикручен.
I don't have enough credits to graduate.	У меня недостаточно баллов, чтобы закончить школу.	У меня недостаточно кредитов, чтобы получить высшее образование.

Table 3: Examples of translation degradation for colloquial content in the English-Russian direction. Model2021 stands for the previous version of the English-Russian system which we consider the baseline.

We should also note that the gain from the transformer-big configuration for the Ukrainian-English model is not that large according to the automatic scores and our human evaluation. We think this is because the synthetic translations obtained from the English-Russian data with the Russian-Ukrainian model are ultimately not of perfect quality.

4 Conclusions and future work

In this paper we presented our submissions for the WMT22 Shared General Translation Task. We show good results in all directions we participate. We clearly outperform our baselines in all directions. A detailed analysis of the translations shows us that we lose quality in translation of colloquial speech. We plan to carefully select colloquial data of very high quality and use it for the general-domain language models for dual cross-entropy data selection. We also plan to train a transformer-big Russian-Ukrainian model and rebuild the synthetic translations for the Ukrainian-English model in the future.

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.

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.