PROMT Systems for WMT 2020 Shared News 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 2020 Shared News Translation Task. This year we participated in four language pairs and six directions: English-Russian, Russian-English, English-German, German-English, Polish-English and Czech-English. All our submissions are MarianNMT-based neural systems. We use more data compared to last year and update our back-translations with better models from the previous year. We show competitive results in terms of BLEU in most directions.

1 Introduction

This paper provides an overview of the PROMT submissions for the WMT 2020 Shared News Translation Task. This year we participate with neural MT systems for the third time. We participate in four language pairs and six directions. We describe our data preparation pipelines, models training setups and present the results on the newstest sets.

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 detailed description of the systems. In Section 5 we present and discuss the results. Section 6 concludes the paper.

2 Systems overview

We submitted six systems based on the `MarianNMT` (Junczys-Dowmunt et al., 2018) toolkit: English-Russian, Russian-English, English-German, German-English, Polish-English and Czech-English. All systems are unconstrained (we use the allowed data, private data and publicly available unconstrained data like OpenSubtitles). The English-German and German-English systems have the same basic architecture. The English-Russian and Russian-English systems are slightly different as we use separate vocabularies. The Polish-English system was trained jointly in both directions. The Czech-English is a multilingual system trained to translate from Croatian, Serbian, Slovak and Czech to English.

3 Data

We use all data provided by the WMT organizers, private in-house parallel data and other publicly available data, mainly from the OPUS website (Tiedemann, 2012). The human parallel data for the German-English system is exactly the same as for the English-German system, the two systems only have different synthetic back-translated data. This also applies to the English-Russian and Russian-English systems.

We use the Tatoeba sets as our validation sets and the newstest2019 is our test set. The reason why we choose the Tatoeba corpus for validation is that we aim at building general-domain (and not just news-domain) models. Besides, the Tatoeba corpus is available for many language pairs beyond the scope of the WMT Translation Task.

We only do fine-tuning for the Czech-English system. This will be described in detail in Section 3.4 below.

3.1 Data filtering

There are several stages in our data filtering pipeline. The statistics for the final training data are shown in Table 1. Note that for the multilingual Czech|Croatian|Serbian|Slovak-English system the table provides statistics only for the Czech-English part. The size of the filtered versions of the Croatian, Serbian and Slovak parts

Proceedings of the 5th Conference on Machine Translation (WMT), pages 248–253
Online, November 19–20, 2020. ©2020 Association for Computational Linguistics

are 21.8M, 21.7M and 14M parallel sentences respectively (more than 95% of the Serbian data is OpenSubtitles).

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. In addition, we remove lines with rare words from the Bookshop and the OpenSubtitles corpora (using frequency lists built on large monolingual corpora including all monolingual data from WMT, private data and Wikipedia dumps).

Deduplication

We remove duplicate translations and keep only the most frequent translation for the source sentence if it repeats more than two times. This procedure is applied to some corpora, e.g. OpenSubtitles and MultiUN which contain a lot of various (and often incorrect) translations for common phrases. For example, the English phrase '*No.*' is encountered almost 100k times in the source side of the English-Russian OpenSubtitles corpus. It has more than 78k unique translations, second most popular among which is '*Да.*' ('*Yes.*' in Russian).

Language detection

The algorithm is a fairly simple ensemble of three tools: `pycld2`[1], `langid` (Lui and Baldwin, 2012), `langdetect`[2].

Parallel data filtering with NMT and language models

We apply this step to all data. Last year we used our own algorithm based on `Hunalign` (Varga et al., 2005) and our inhouse classifier to identify and discard unparallel sentence pairs. This year we use a different approach. We score parallel data with NMT models in both directions. We also score the source and target sides of the data with statistical language models built on large sets of what we assume to be good-quality data (basically, the newscrawl data from the statmt.org website). The scores are normalized by sentence length and summed up. We also apply weights (from 0.1 to 0.3 depending on the corpus type) to the statistical LMs scores as we mostly rely on the scores produced by the NMT models. The data is then sorted according to the final scores, and we select a subset of the data according to a certain threshold set individually for different corpora by our linguists.

	German-English		**Russian-English**		**Polish-English**		**Czech-English**	
	#sent	#tokens EN	#sent	#tokens EN	#sent	#tokens EN	#sent	#tokens EN
WMT	26.6	580.1	27.3	690.9	10.3	183.2	11.4	147.8
OPUS	23.8	475.9	8.3	74.9	26.8	283.7	29.1	263.8
Private	7.5	100.4	25.5	428.2	0.3	3.7	0.4	5.1
Total	57.9	1156.4	61.1	1194.0	37.4	470.6	40.9	416.7

Table 1: Statistics for the filtered parallel data in millions of sentences (#sent) and tokens (#tokens) for four language pairs. WMT stands for the data available for the News Task on the statmt.org/wmt20 website; OPUS is the data from the OPUS website apart from the data available for the News Task; Private stands for private company data.

3.2 Data preprocessing

BPE

Same as last year, we use the `OpenNMT` toolkit (Klein et al., 2017) version of byte pair encoding (BPE) (Sennrich et al., 2016b) to encode our data to subword units. The BPE merge operations are learnt in case-insensitive mode. Case-insensitive BPE model is very useful when dealing with noisy data (like, for example, OpenSubtitles where uppercase is often used to communicate emphasis) or legal and financial data where specific terms are written in title case or uppercase. News headlines are also often written in title case or uppercase.

The OpenNMT preprocessor handles case as a feature assigned to each token. As MarianNMT

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

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

does not support features yet, we perform a ‘trick’ similar to the one described in (Tamchyna et al., 2017): instead of using a feature we insert special tokens <C> and <U> after sequences in title case or uppercase. For example, a source sentence

World Championships 2017: Neil Black praises Scottish members of Team GB

is converted to

world <C> championships <C> 2017 : neil <C> black <C> pra@@ ises scottish <C> members of team <C> gb <U>

We do not use truecaser in our pipeline as it is redundant. All data is tokenized using the `Moses` toolkit (Koehn et al., 2007) tokenizer with aggressive tokenization, then the OpenNMT BPE-splitter is applied, after that we convert the case feature to separate tokens.

General tendency for our models this year is to build smaller BPE models.

English-Russian and Russian-English

Same as last year (Molchanov, 2019), we train the models with separate vocabularies due to the Cyrillic nature of Russian alphabet. Therefore we build separate BPE models for source and target, but with less merge operations (16k for English and 32k for Russian) compared to last year (35k and 45k respectively).

English-German and German-English

We train a joint BPE model for the English-German pair with 16k merge operations. We use a shared vocabulary and tie all embeddings for all translation models with joint BPE.

Polish-English

We train a joint BPE model for the Polish-English pair with 12k merge operations.

Czech-English

As was mentioned earlier, our Czech-English model is a multilingual model trained to translate from Czech, Croatian, Serbian and Slovak into English. Therefore we train a joint BPE model for all five languages with 24k merge operations. As part of Serbian data is in Cyrillic alphabet, we transliterate it into Latin using an inhouse transliteration tool.

3.3 Synthetic data

There are two types of additional synthetic training data described in detail below. The final size of the training data for the submitted systems is roughly 4 times the total size of the filtered data in Table 1 Table 1 for each language pair.

Both types of synthetic data are used for training all submitted systems. We also tag all synthetic data following (Caswell et al., 2019), i.e., insert a special token <*bt*> at the beginning of each source line for back-translations etc.

Back-translated data

Back-translations (Sennrich et al., 2016a) are a common way to improve NMT models quality. As we aim at building general-domain models, we use data from Wikipedia dumps and news from statmt.org. We shuffle the Wikipedia data and randomly select a subset of appropriate size. The selected Wikipedia subset and the news subset are roughly equal in size. The size of the whole corpus used for back-translation is approximately equivalent to the size of human training data.

For the English-Russian pair we use our last year’s English-Russian model to obtain back-translations for the Russian-English model. Then we train the Russian-English model and use it to obtain back-translations for the final English-Russian model.

We also obtain back-translations for the German-English pair using our last year’s models.

For the Polish-English and Czech-English pairs we build intermediate models using all available data excluding OpenSubtitles and Paracrawl.

We score our back-translations with the opposite-direction NMT models to discard obviously bad translations.

Replicated data with unknown words

We apply the technique described in (Pinnis et al., 2017) to create a synthetic parallel corpus. The procedure includes the following steps: first, we perform word-alignment of our initial parallel training corpus using the fast-align tool (Dyer et al., 2013). Then, we randomly replace from one to three unambiguously (one-to-one) aligned tokens in both source and target parallel sentences with the special <UNK> placeholder. The same pipeline is applied to both the initial and back-translated data. We train our models to reproduce the <UNK> placeholder in various contexts and use this feature for handling named entities as described in Section 4.1 below.

3.4 Data for fine-tuning

We only do fine-tuning for the Czech-English system. The model is tuned on available parallel Czech-English data mixed with back-translations of the English news 2017-2019 from statmt.org. We use the newstest2017-2019 as our devset.

4 Systems architecture

This section describes the trained systems in detail. We train transformer (Vaswani et al., 2017) models for all submitted systems. We use the recipe available at the MarianNMT website[3]. The system configuration, hyperparameters and training steps follow those in the recipe.

We use the transformer-big configuration for the English-Russian model.

We train single models for all directions.

We use the beam of size 6 and the `--normalize` parameter is set to 0.6.

4.1 Handling named entities

We preserve several types of named entities (NEs): numbers, emails, alphanumeric sequencies etc. in the following way. First, we produce the baseline NMT translation without any processing. Then we validate the translation of NEs by comparing the system's output to the source sentence. The validation is simple: we search for the corresponding strings (numbers, emails etc.) in the system's output. If some of the NEs are not translated or are translated incorrectly, we replace the entities with the <UNK> placeholder in the source sentence and translate the sentence again allowing the decoder to generate unknown words in the output. Finally, we substitute the <UNK> placeholders in the output with their initial value. If the number of the <UNK> placeholders in the NMT system's output is not equal to the number of the placeholders in the source sentence, we fall back to the baseline NMT translation without NEs processing. We do not do any specific processing for proper names.

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 analysis of the results.

The scores are presented in Table 2. Calculation is done using the `multi-bleu-detok.perl` script from the `Moses` toolkit.

We significantly outperform our last year's submissions for the News Task.

Fine-tuning for the Czech-English system does not give us significant improvements in terms of BLEU. This may be because we didn't perform any data selection this year.

System	newstest2019	newstest2020
English-Russian		
Model2019	29.5	21.7
Model2020	**32.3**	**23.3**
Russian-English		
Model2019	37.2	33.6
Model2020	**42.3**	**38.2**
Polish-English		
Model2020	**-**	**31.3**
English-German		
Model2019	38.2	29.8
Model2020	**40.7**	**31.9**
German-English		
Model2019	32.4	34.9
Model2020	**39.4**	**39.6**
Czech-English		
Model2020	**-**	25.1
Model2020 tuned	**-**	**25.6**

Table 2: Results for different systems and directions. The submitted systems are marked in bold. Model2019 stands for our last year's submitted systems which we consider the baseline.

We are among the top 10 systems in the English<>Russian directions, however, we are substantially behind the top systems in other directions in terms of BLEU. We see two reasons for that. First of all, we pay much more attention to our Russian systems, thus, our last year's Russian systems had already undergone several iterations of updated backtranslations and retraining and can be considered strong baselines. Second, we possess much more private high-quality data for the English-Russian pair compared to other language pairs.

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

6 Conclusions and Future work

In this paper we have described our submissions for the WMT 2020 Shared News Translation Task. Overall we have made six submissions in four language pairs: English-Russian, English-German, Polish-English and Czech-English.

We have documented the methodology used to prepare the training data, system training set-ups, the pipeline for handling NEs.

We show competitive results in most directions.

In future we plan to experiment once again with a shared vocabulary for the English-Russian models applying transliteration to the source side.

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.

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.

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, 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.

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.

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.

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.

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.

Aleš Tamchyna, Marion Weller-Di Marco and Alexander Fraser. 2017. Modeling Target-Side Inflection in Neural Machine Translation. In *Proceedings of the Second Conference on Machine Translation*, pages 32–42, Copenhagen, Denmark.

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.

Dániel Varga, László Németh, Péter Halácsy, András Kornai, Viktor Trón, Viktor Nagy. 2005. Parallel corpora for medium density languages. In *Proceedings of the RANLP 2005*, pages 590–596, Borovets, Bulgaria.

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.