The only chance to understand: machine translation of the severely endangered low-resource languages of Eurasia

Anna Mossolova, Kamel Smaïli

The only chance to understand: machine translation of the severely endangered low-resource languages of Eurasia

Anna Mosolova
Université de Lorraine
Nancy, France
a.mosolova333@gmail.com

Kamel Smaïli
Loria, Campus Scientifique
Vandoeuvre-Lès-Nancy, France
smaili@loria.fr

Abstract

Numerous machine translation systems have been proposed since the appearance of this task. Nowadays, new large language model-based algorithms show results that sometimes overcome human ones on the rich-resource languages. Nevertheless, it is still not the case for the low-resource languages, for which all these algorithms did not show equally impressive results. In this work, we want to compare 3 generations of machine translation models on 7 low-resource languages and make a step further by proposing a new way of automatic parallel data augmentation using the state-of-the-art generative model.

1 Introduction

Being one of the oldest tasks of natural language processing (NLP), machine translation changed many different state-of-the-art approaches over the past 70 years. Starting with old dictionary-based systems, then going forward with statistical algorithms, switching to neural approaches with sequence-to-sequence methods, currently, the best MT systems use language models (LM) with Transformer architecture inside.

All these new language models rely on huge data corpora from which they are able to extract general patterns about any language including grammar, vocabulary, discourse characteristics, etc. Their results are especially remarkable on the translation tasks from one rich-resource language to another, where they achieve results sometimes indistinguishable from the human ones.

However, when it comes to the low-resource languages, not all models can perform well. Recently, new large LMs were developed especially for few-shot learning, however, they are still evaluated on the datasets containing several tens of thousands of samples.

In this work, we want to evaluate the performance of the algorithms coming from 3 different generations of models: statistical, sequence-to-sequence and transformer-based. Additionally, we want to propose a new fully automatic parallel data augmentation method based on GPT model and compare the quality of the models fine-tuned with the generated data.

For our purposes, we will take 7 extremely low-resource languages of Eurasia coming from 4 different language families. All these languages will be the source languages of translation, while Russian will be the target one. These languages are mainly spoken in Russia, so linguists have already collected small corpora of linguistic data for these languages including the sentences translations to Russian. We extracted sentences from these corpora and composed 7 datasets of parallel sentences. The minimum size of the corpus used is 586 training pairs and the maximum size is equal to 8619.

In the following, we will start by presenting the related work that has been carried out so far (Section 2), then the languages used for this study (section 3). After this, we will describe the experiments (Section 4) and analyse the obtained results (Section 5). All the contributions made in this paper will be summed up in the conclusion (Section 6) as well as the direction of the future work.

2 Related work

Throughout the history of machine translation, numerous models have been proposed. During the era of statistical machine translation, one of the first models were word-based models such as IBM ones (Brown et al., 1993). These models were then followed by the phrase-based systems (Koehn et al., 2003) which became widely used for several years.

The next decade was marked by the appearance of neural network-based machine translation algorithms starting with a sequence-to-sequence model with LSTM layers (Sutskever et al., 2014) which was then modified with CNN (Gehring et al., 2017) and different types of attention mechanism (Luong

et al., 2015).

In the recent years, the Transformer architecture (Vaswani et al., 2017) appeared and it showed groundbreaking results in many NLP tasks. For machine translation, firstly, the original Transformer paper showed new state-of-the-art scores and then the metrics were improved by the T5 model (Raffel et al., 2020), the multilingual mBART-25 (Liu et al., 2020) and mBART-50 (Tang et al., 2020) models followed by the other transformer-based architectures.

These three generations of models have also been used in the low-resource settings. We can find adaptions of all kinds of algorithms for the under-resourced conditions. For example, the phrase-based statistical models have been used for the translation of the low-resource Arabic dialects (Meftouh et al., 2015).

As for the sequence-to-sequence models, an interesting approach to data processing and further Seq2Seq model training and tuning was shown in the paper by Goel et al. (2020). The authors transliterated all low-resource languages that they had into the common alphabet shared with a rich-resource language coming from the same language group. Then they pre-trained a sequence-to-sequence model using the corpus of a rich-resource language and fine-tuned it with small corpora of the low-resource languages. Another example of the successful application of the Seq2Seq model to the low-resource machine translation is the multi-task training using the translation task from and to several dialects at the same time (Moukafih et al., 2021).

The Transformer-based models have also been tested in the low-resource conditions. For example, Garcia et al. (2021) proposed a new 3-stage training approach with no data for the low-resource languages. The authors trained the Transformer model using the corpora of the close rich-resource languages. Additionally, they used the so-called synthetic corpora which contained the translations of the sentences from all zero-source languages which they generated using the model obtained after the first stage of training.

As we can see, the main approach that is used to improve the quality of the translation is transferring some knowledge from the languages that are coming from the same language group or family. Moreover, these language families have many daughter languages that are popular in the world. However, in our work, some of the languages either are the only remaining living languages of their family or come from the families which are not widely known, so we will try to exploit some approaches that do not rely on the languages similarities.

3 Study of several low-resource languages

3.1 Motivation

In this study, we want to evaluate the performance of 3 different types of models on a particularly difficult type of the machine translation task which is the translation of extremely low-resource languages. The target language for all our experiments is Russian, while the source sentences come from 7 low-resource languages.

Being a member of the *Indo-European language family*, Russian is considered to be a high-resource language with a common word order Subject-Verb-Object (SVO) and fusional type of inflection. Apart from many European languages, Russian uses Cyrillic alphabet which makes it difficult to transfer the knowledge of the pre-trained monolingual language models by fine-tuning them on a small Russian corpus. However, many popular language models were trained on huge Russian corpora (for example, Common Crawl (Eberius et al., 2015) or Taiga (Shavrina and Shapovalova, 2017)), such as BERT[1], T5[2] or GPT[3] and then applied on various down-stream tasks.

The low-resource languages that we use in this study are: Karelian, Ludic, Veps, Selkup, Evenki, Chukchi and Ket. They are spoken in Eurasia, mainly in Russia and adjacent countries, however, none of them belongs to the Indo-European language family. We chose these languages as they are the heritage of the nationalities that use these languages as the native one and of the countries to which these nationalities belong. Unfortunately, currently these languages are not widely spoken any more, as it becomes more and more popular to use Russian as a native language and learn English as a second one. In Russia, studying the language that represents the identity of a region is mandatory only during the first 4 years of education in school, so many students stop using their national language

[1] https://huggingface.co/DeepPavlov/rubert-base-cased

[2] https://huggingface.co/cointegrated/rut5-base-multitask

[3] https://github.com/ai-forever/ru-gpts

once they are 11 years old. With our work, we want to draw attention to these languages, as some of them are on the verge of extinction.

3.2 Languages description

In this section, we will give some linguistic facts about the studied languages such as their language family and which word order, word formation method and alphabet they use. We will also provide the examples of the translation of a Russian sentence *Ja ne ponimaj tebja* (I do not understand you), when possible.

Karelian, Ludic, Veps and Selkup[4] languages come from the *Uralic language family*. All of them have SVO word order, are agglutinative and are written with the Latin alphabet. Karelian phonetic system consists of 8 vowels and 19 consonants, Ludic has the same number of vowels and one more consonant, Selkup contains 25 vowels and 16 consonants, while Veps has 10 vowels and 34 consonants.

The difference between these languages can be seen from the examples. For instance, the sentence *I do not understand you* in Karelian is *en ymärrä teitä*, in Ludic is *en elgenda teid*, in Veps is *mina en el'genda teid* and in Selkup is *mat assa sintit tenimak* (all words are transliterated into Latin where necessary). Selkup's translation does not resemble others at all, while Ludic and Veps are almost similar except for the pronoun *minä* (En: I) in Veps.

Chukchi language[5] is a member of the *Chukotko-Kamchatkan language family* with Subject-Object-Verb (SOV) word order, agglutination and Latin alphabet which consists of 6 vowels and 14 consonants. The example of a sentence in this language is: *wanewan mesisewtek* (I do not understand you), where the first word expresses the negation and the tense and the second word expresses the verb's meaning, the subject (*me-*) and the object (*-tek*).

Evenki language[6] is a part of the *Tungusic language family*, it uses SOV word order, agglutination for word formation and inflection and Cyrillic alphabet for writing which contains 11 vowels and 18 consonants. The following phrase is an example of a sentence in the Evenki language: *bi sine ehim tylle* (I do not understand you), where *bi* is a subject, *sine* is an object, *ehim* expresses the negation (*e-*), the present tense (*-hi-*) as well as person and number (*-m*) and *tylle* is a verb which also carries the meaning of negation (*-le*).

Ket language[7] is the only living member of the *Yeniseian language family*. This language uses SVO word order and Cyrillic alphabet as well. Its phonetic system has 11 vowels and 20 consonants. It has fusional type of word formation and inflection. Here is an example of a sentence in the Ket language: *bu duoton kolet* (he sees the city), where *bu* is the subject, *kolet* is the object and *duoton* is a verb in which the grammatical information about the subject is expressed in the *du-* part and the grammatical information about the object is shown with the *-o-* part.

The summary of the sizes of the corpora available for our study is presented in the table 1.

Language	Training corpus size
Ket	586
Chukchi	806
Ludic	1100
Karelian	1571
Selkup	1932
Evenki	4524
Veps	8619

Table 1: Corpora sizes for 7 low-resource languages. The size is represented by the number of parallel sentences in an X language and Russian

4 Machine translation models for the languages of Eurasia

In this section, we will describe 4 different machine translation models that we trained on our datasets.

Before we started the experiments, we uniformly preprocessed the datasets. The following steps were applied: punctuation removal, lower-casing, deleting the sentences that are longer than a certain threshold. For each language, we determined the optimal maximum length of the sentences on the basis of the loss curve during the training. We noticed that loss values are abnormally big on the long sequences, so for each language we built the plots with the dependency between loss values and sentence's lengths and chose the maximum length by finding an optimal point, where we do not lose too many training samples and loss values are not

[4]The datasets are composed from the extracts of the corpus presented in Zaytseva N. G. (2017) and Brykina et al. (2018)

[5]The dataset is composed from the sentences extracted from the corpus of the Siberian Lang project

[6]The dataset is composed from the sentences extracted from the corpus of the Siberian Lang project

[7]The dataset is composed from the sentences extracted from the corpus of the Chucklang website

extremely high. In general, we deleted from 10 to 20 pairs from each dataset.

4.1 Statistical Machine Translation

Despite the existence of neural approaches to machine translation, statistical machine translation still remains a preferable solution in some cases. It is attractive due to the fact that it does not require as mush data as neural approaches and, additionally, the vocabulary used to translate the sentences is sometimes richer than the one of neural models, especially, in the low-resource settings. Another advantage of the statistical model is the speed of training. In our experiments, it took only a few minutes to fully train a model for one language.

We used the Moses system (Koehn et al., 2007) to evaluate the quality of statistical MT approaches. For our purposes, we took the phrase-based system with the trigram KenLM language model (Och and Ney, 2003) and the GIZA++ alignment model (Heafield et al., 2013). We trained the translation model on the training corpus of each language and tuned it on the validation part.

4.2 Sequence-to-Sequence

In this study, we used the sequence-to-sequence model with LSTM layers and attention (Luong et al., 2015) from the OpenNMT library (Klein et al., 2017). We used Adam as an optimizer and a batch size equal to 64 for our training. We also experimented different learning rate values and chose 1e-5 as a final one, because the model was overfitting with the bigger ones and underfitting otherwise.

4.3 mBART

A popular mBART architecture has shown SoTA results on many rich and medium-resource language, so we decided to check if it is possible to transfer some of its knowledge to the new, unseen languages. For these purposes, we took a large mBART-CC25 model from the Fairseq repository[8] and fine-tuned it using the parallel corpora of 7 low-resource languages. We preprocessed the corpora using the mBART SentencePiece model[9]. For the fine-tuning, we took the standard Adam optimizer and a learning rate equal to 3e-05 to prevent model from forgetting the knowledge about the language it extracted from the corpora during the pre-training. We also set the early stopping to 10 epochs without validation loss improvement.

4.4 GPT

The model from the GPT family are known for their generative abilities, that is why we decided to check if a decoder Transformer-based model is able to learn the translation task. To train a model for this task, we tried several prompts and ended up with a form *"<Source language name>: <Sentence in a source language>. Russian: <Translation of the sentence into Russian> <|endoftext|>"*.First, we tried using the mGPT model (Shliazhko et al., 2022) which is said to be a GPT-3 model based on GPT-2. This model was trained on 60 different languages including Russian.

However, this model kept producing the translations on other languages, so we switched to the ruGPT-3 model[10] which was trained only on the Russian corpus. For this model, we translated the prompt so it became: *"<Source language name in Russian>: <Sentence in a source language>. <Word 'translation' written in Russian>: <Translation of the sentence into Russian> <|endoftext|>"*.

4.5 Augmentation with GPT

As GPT-3 is a generative model, we tried to use it to generate new samples of the data. After training the model to translate from one of the source languages to Russian, we prompted it with a name of a source language (for example, *"Evenki: "*, but written in Cyrillic) to check if it can generate the source sentence and its translation. For these purposes, we used the Beam search with the following parameters: maximum length = 40, repetition penalty = 1.2, top-k = 50, top-p = 0.95, temperature = 0.7.

This combination produced examples that sometimes were a real translation pair. However, many pairs were wrong due to the fact that the model continued generating the Russian translation up to the maximum length, so we filtered all examples that had more then twice words in the translation than in the source sentence. Additionally, we checked if all words from the source part were present in the training dataset. Hypothetically, the model could have learnt how to conjugate verbs or decline nouns. Nevertheless, none of the authors is a native speaker of any source language from this study, so we decided to stick to the definitely

[8] https://github.com/facebookresearch/fairseq/tree/main/examples/mbart

[9] https://github.com/google/sentencepiece

[10] https://github.com/ai-forever/ru-gpts

existing words to avoid fine-tuning a model with the fake data.

We augmented all the datasets with 10% of the newly generated translation pairs and fine-tuned the mBART models using these new datasets.

5 Results

In this section, we will show and discuss the performance of all machine translation models that we implemented. The figures with the comparison of all results for every language are presented in the Appendix A.

5.1 Phrase-based statistical model

BLEU scores that we obtained with the Moses model are shown in the table 2.

We analysed the translations and noticed that the model leaves all words for which it cannot find the corresponding translation unchanged in the translation. Comparing the SMT results to the other models, we can see that this behaviour allowed it to achieve the highest BLEU scores among other models for 3 languages with the smallest training corpora (Chukchi, Ket, Ludic). Neural network-based models were not able to understand the structure of the language with such a small number of sentences, while statistical approach not only retained all possible correct translations, but also copied the words from the input to the output instead of repeating or generating random words. It was especially helpful in the case of Ket, where native speakers sometimes included Russian words in the Ket sentences.

5.2 Seq2Seq

Table 3 presents the results of the Seq2Seq model. We can see that these results are the worst ones among other models, as this model needed to learn the grammar and the vocabulary from scratch using only our small training corpora which were not sufficient for the network. In the original Seq2Seq paper(Sutskever et al., 2014), the authors showed BLEU scores of 34.81 after training on the corpus of 12M parallel sentences which can explain close to 0 results of our models which did not have that much training data. During the analysis of the results, we have noticed that Seq2Seq models tend to repeat simple words or replace some words with the <*unk*> token which also affected the final results.

An additional reason of the low scores for some languages is the fact that the models needed to learn to translate from the Latin alphabet to Cyrillic and from the languages with a completely different grammatical structure. For example, the Chukchi language has the SOV word order and tends to incorporate the information about the subject and the object into a verb (see Section 3.2 for an example).

5.3 mBART

In the table 4, the performance of the mBART models is shown.

During the evaluation of the translations produced by the model, we noticed that sometimes it replaces some words with their synonyms, so the BLEU score may show lower results, despite the fact that the translations were still understandable.

One can see that the Ket language performance is again better than for almost all other languages which is related to the sentences size, small vocabulary size and the fact that some sentences already contained words in Russian which did not need any translation.

We can also see that the mBART model achieves the highest result for the Karelian language and almost highest results for the other Uralic languages. This is related to the fact that mBART was trained on Finnish and Estonian languages, so the knowledge transfer was made not only for the target translations in Russian, but also for the source sentences in our low-resource languages.

The low results of the Veps model are caused by the tendency of the model to overfit and predict the same token instead of translations. For this reason, we stopped the training before the repeating token started occurring in the translations which led to worse results. We suppose this behavior is related to the bigger corpus size compared to the other languages.

5.4 ruGPT-3

BLEU scores we obtained with the ruGPT-3 model are shown in the table 5.

We can see that the results are comparable with the mBART model when the alphabet used by the language is Cyrillic, while for other languages the BLEU values are smaller. The only exception is Veps language which shows better results then the mBART model due to the problems with the mBART model.

When analysing the results, we have also noticed that the GPT model sometimes is not able to trans-

Language	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Evenki	33.2	23.1	16.5	12.1
Chukchi	19.9	12.8	8.4	5.6
Ket	53.2	42.3	34.5	27.4
Selkup	27.6	16.8	10.5	6.8
Ludic	30.5	17.7	11.4	7.8
Veps	43.6	28.7	19.9	14.5
Karelian	49.3	34.3	24.8	18.3

Table 2: Phrase-based model results on translation task from 7 low-resource languages to Russian

Language	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Evenki	15.47	7.81	3.4	1.59
Chukchi	10.03	3.0	0.0	0.0
Ket	22.04	10.66	4.76	0.0
Selkup	16.92	8.05	3.43	0.0
Ludic	16.38	6.81	2.86	1.62
Veps	18.23	7.66	3.5	1.85
Karelian	20.14	8.22	4.05	0.0

Table 3: Seq2Seq model results on translation task from 7 low-resource languages to Russian

Language	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Evenki	36.8	24.8	16.9	12.2
Chukchi	13.9	6.9	4.1	2.5
Ket	36.8	28.1	20.7	14.9
Selkup	23.0	13.0	7.8	5.3
Ludic	25.1	14.2	8.7	5.6
Veps	28.0	15.3	8.3	4.7
Karelian	50.2	36.9	27.1	20.1

Table 4: mBART model results on translation task from 7 low-resource languages to Russian

late Karelian sentences correctly because of their length which was up to 220 symbols.

5.5 mBART with ruGPT-3 augmentation

Table 6 represents the BLEU scores that we obtained with the mBART model after fine-tuning it with the augmented data.

The results show that Evenki and Selkup models have improved some of their BLEU scores compared to the mBART models trained with the original datasets. As for the other models, we have noticed that the change in quality of the model is proportional to the size of the dataset. This correlation is shown in the figure 1. The training corpus size of each language is presented on the x axis, the difference between two BLEU-1 scores is presented on the y axis. We can see that the quality of the ruGPT-3 generation depends severely on the size of the training corpus. This fact is proved by the results of the translation models trained on the generated data. One can see that the results are much worse for the models with less than 1000 examples and starting from 1000 examples the difference becomes less and less. It means that the GPT model is able to generate coherent examples which are helpful during the training of the translation model starting from 2000 examples.

5.6 Example analysis

In the table 7 we present the example of the translation of one sentence by each system. The source sentence for all systems was the following phrase in the Evenki language: *tar ahi albaran ilatčami togoi.*. The expected output is the first line of the table.

As we can see, the statistical model did not manage to find the translation for the word *togoi* and left it unchanged in the text. As for the Seq2Seq model,

Language	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Evenki	34.8	23.7	16.7	12.2
Chukchi	14.4	8.1	4.9	3.1
Ket	37.9	30.6	24.0	19.5
Selkup	20.6	11.5	6.6	4.4
Ludic	17.2	8.4	5.4	3.7
Veps	36.0	22.6	15.2	10.3
Karelian	27.0	16.1	9.9	6.2

Table 5: ruGPT-3 model on translation task from 7 low-resource languages to Russian

Language	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Evenki	37.4	25.6	17.8	13.1
Chukchi	8.2	3.7	1.9	0.0
Ket	22.1	16.2	11.7	8.5
Selkup	23.4	13.3	7.6	4.9
Ludic	24.0	12.3	7.5	4.4
Veps	23.3	12.5	6.9	4.0
Karelian	49.2	35.2	26.0	19.9

Table 6: mBART model results after augmentation of the datasets by 10% on translation task from 7 low-resource languages to Russian

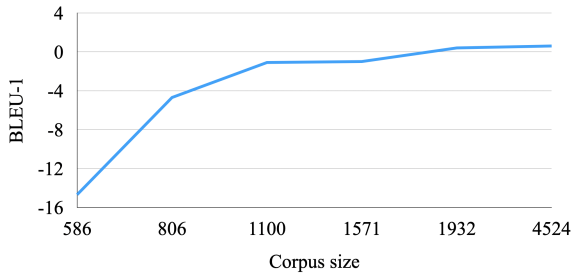

Figure 1: Correlation between the training corpus size and BLEU-1 difference between mBART and augmented mBART model. We did not include the Veps model results do to the problem explained in the Section 5.3

one can notice that it suffers from the lexical repetition problem and additionally it missed the main verb of a sentence. Both mBART models translated the sentence almost correctly, the only missing point is the possessive pronoun *svoj* which is expressed by the last letter *i* in the source word *togoi*. The ruGPT model translated the words correctly, but made 2 grammatical errors: in the subject by declining it to the instrumental case (*zenŝinoj* instead of *zenŝina*) and in the auxiliary verb by using the masculine ending instead of the feminine one (*smog* instead of *smogla*).

Overall, the translation quality is pretty high and it is possible to understand the source meaning of the sentence from all the generated translations.

6 Conclusion

In this study, we have presented our work on the machine translation for 7 low-resource languages of Eurasia. We have compared the phrase-based statistical model, the Seq2seq model, the mBART model and the ruGPT-3 model. We have shown that the statistical model achieves the highest quality for the majority of the languages and mBART model shows the best quality for the remaining ones.

We have also proposed the new way of augmenting the dataset with parallel sentences generated by the GPT-model fine-tuned for the translation task. The study has shown that this method allows to increase the quality of the model starting from a certain size of the training dataset, otherwise the quality decreases as the GPT model is not able to generate coherent examples.

Our future directions of research include training other Transformer-based architectures like M2M100 and using multi-task learning during the fine-tuning stage.

By this work, we would like to bring attention to the low-resource languages of Eurasia and encourage other researchers to continue our work. Every language is the part of the world's treasure and it is important to do our best trying to preserve them.

Model	Translation	English translation
Target	eta ženŝina ne smogla razžeč svoj ogon	this woman did not manage to start her fire
Moses	ta ženŝina ne smogla razžeč togoi	this woman did not manage to start *her fire*
Seq2Seq	eta ženŝina ne mogla i ogon ogon	this woman was not able and fire fire
mBART	ta ženŝina ne smogla razžeč ogon	that woman did not manage to start the fire
ruGPT-3	ženŝinoj i ne smog razžeč ogon	*woman* and *did not manage* to start the fire
mBART+	ta ženŝina ne smogla razžeč ogon	that woman did not manage to start the fire

Table 7: An example of the generated translations. The first line is the target translation from the corpus, other lines represent different models. mBART+ refers to the mBART model trained with the augmented dataset. Words in italics represent errors that are not obvious from the English translation and are explained in the Section 5.6

Acknowledgements

Experiments presented in this paper were carried out using the Grid'5000 experimental testbed, being developed under the INRIA ALADDIN development action with support from CNRS, RENATER and several Universities as well as other funding bodies (see https://www.grid5000.fr).

References

Peter F. Brown, Stephen A. Della-Pietra, Vincent J. Della-Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation. *Computational Linguistics*, 19(2):263–313.

Maria Brykina, Svetlana Orlova, and Beáta Wagner-Nagy. 2018. Inel selkup corpus. version 0.1. *The INEL corpora of indigenous Northern Eurasian languages., Hamburg, December. Hamburger Zentrum für Sprachkorpora*.

Julian Eberius, Maik Thiele, Katrin Braunschweig, and Wolfgang Lehner. 2015. Top-k entity augmentation using consistent set covering. SSDBM '15.

Xavier Garcia, Aditya Siddhant, Orhan Firat, and Ankur Parikh. 2021. Harnessing multilinguality in unsupervised machine translation for rare languages. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1126–1137, Online. Association for Computational Linguistics.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional sequence to sequence learning. In *International conference on machine learning*, pages 1243–1252. PMLR.

Pranav Goel, Suhan Prabhu, Alok Debnath, Priyank Modi, and Manish Shrivastava. 2020. Hindi TimeBank: An ISO-TimeML annotated reference corpus. In *16th Joint ACL - ISO Workshop on Interoperable Semantic Annotation PROCEEDINGS*, pages 13–21, Marseille. European Language Resources Association.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable modified Kneser-Ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 690–696, Sofia, Bulgaria. Association for Computational Linguistics.

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.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions*, pages 177–180.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase based translation. In *Proceedings of the Joint Conference on Human Language Technologies and the Annual Meeting of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL)*.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation.

Karima Meftouh, Salima Harrat, Salma Jamoussi, Mourad Abbas, and Kamel Smaili. 2015. Machine translation experiments on padic: A parallel arabic dialect corpus. In *Proceedings of the 29th Pacific Asia conference on language, information and computation*, pages 26–34.

Youness Moukafih, Nada Sbihi, Mounir Ghogho, and Kamel Smaïli. 2021. Improving Machine Translation of Arabic Dialects through Multi-Task Learning. In *20th International Conference Italian Association for Artificial Intelligence:AIxIA 2021*, MILAN/Virtual, Italy.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Tatiana Shavrina and Olga Shapovalova. 2017. To the methodology of corpus construction for machine learning:«taiga» syntax tree corpus and parser. *Proceedings of the "Corpora*, pages 78–84.

Oleh Shliazhko, Alena Fenogenova, Maria Tikhonova, Vladislav Mikhailov, Anastasia Kozlova, and Tatiana Shavrina. 2022. mgpt: Few-shot learners go multilingual.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008.00401*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Kryzhanovskaya A.A. Pellinen N.A. Rodionova A.P. Zaytseva N. G., Kryzhanovskii A.A. 2017. Otkryty korpus vepsskogo i karelskogo yazykov (vepkar): predvaritelny otbop materialov i slovarnaya chast sistemi. *Trudi mezhdunarodnoi konferrencii «Korpusnaya lingvistika – 2017»*., pages 172–177.

A The comparison of all models

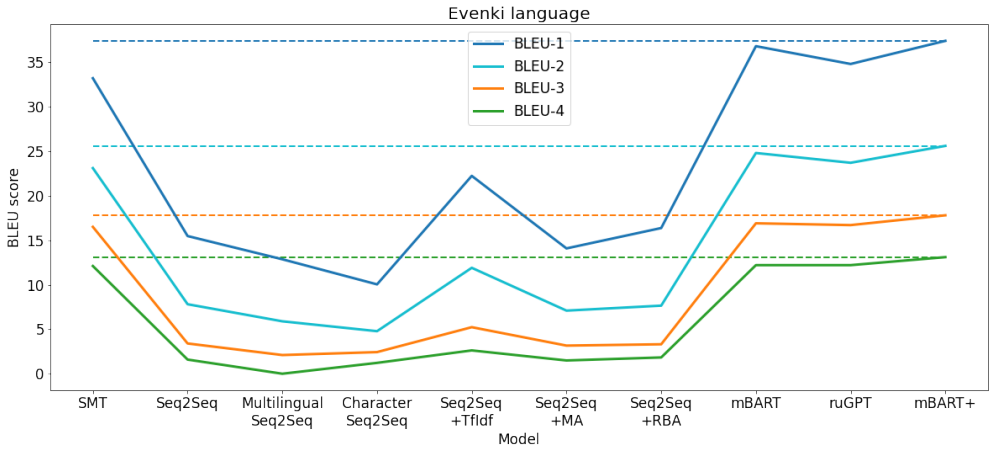

Figure 2: The comparison of the results of all the models trained from Evenki to Russian. Dashed lines represent the best result for the corresponding BLEU score. RBA = rule-based augmentation, MA = manual augmentation, mBART+ = mBART trained on the augmented with ruGPT dataset

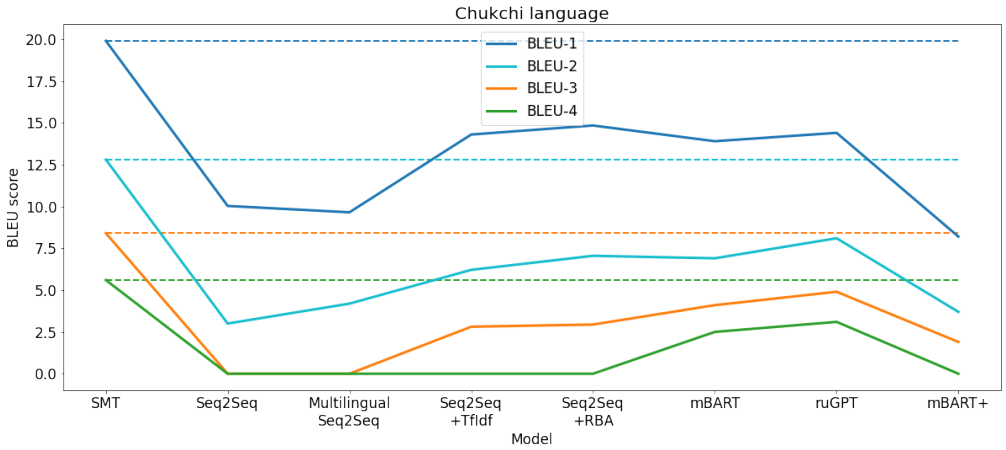

Figure 3: The comparison of the results of all the models trained from Chukchi to Russian. Dashed lines represent the best result for the corresponding BLEU score. RBA = rule-based augmentation, MA = manual augmentation, mBART+ = mBART trained on the augmented with ruGPT dataset

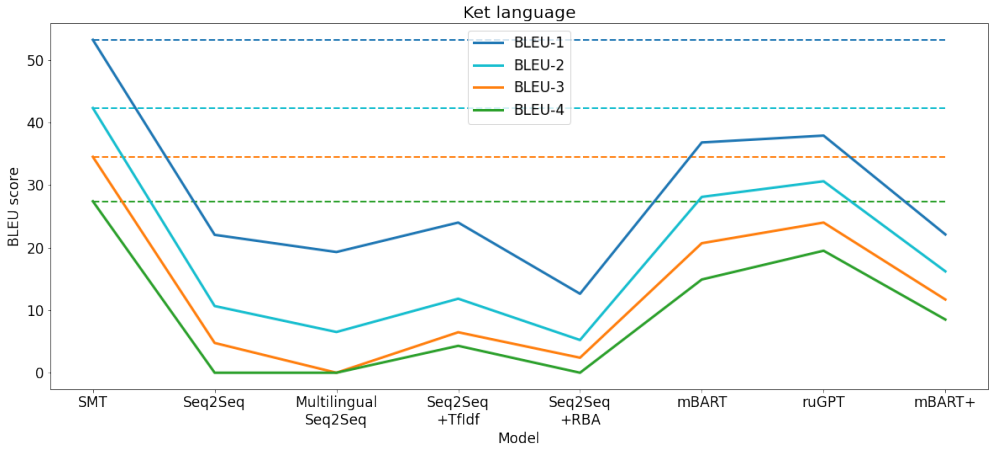

Figure 4: The comparison of the results of all the models trained from Ket to Russian. Dashed lines represent the best result for the corresponding BLEU score. RBA = rule-based augmentation, MA = manual augmentation, mBART+ = mBART trained on the augmented with ruGPT dataset

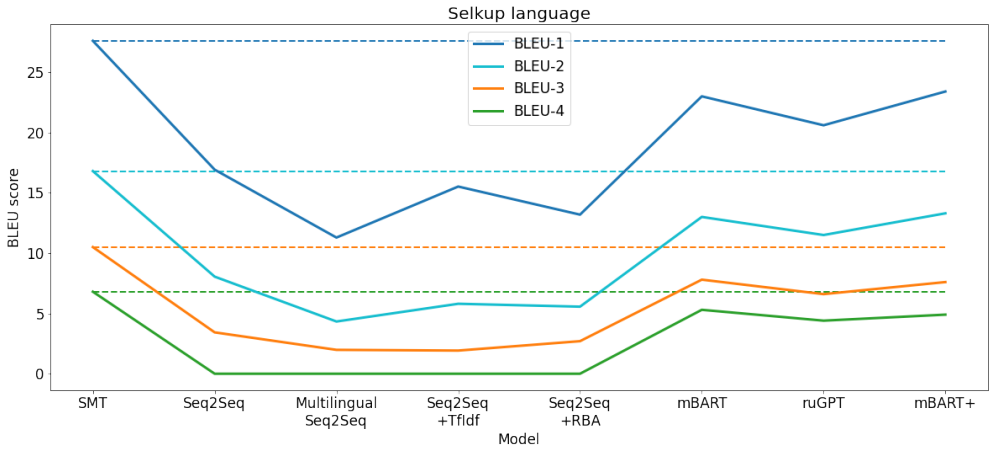

Figure 5: The comparison of the results of all the models trained from Selkup to Russian. Dashed lines represent the best result for the corresponding BLEU score. RBA = rule-based augmentation, MA = manual augmentation, mBART+ = mBART trained on the augmented with ruGPT dataset

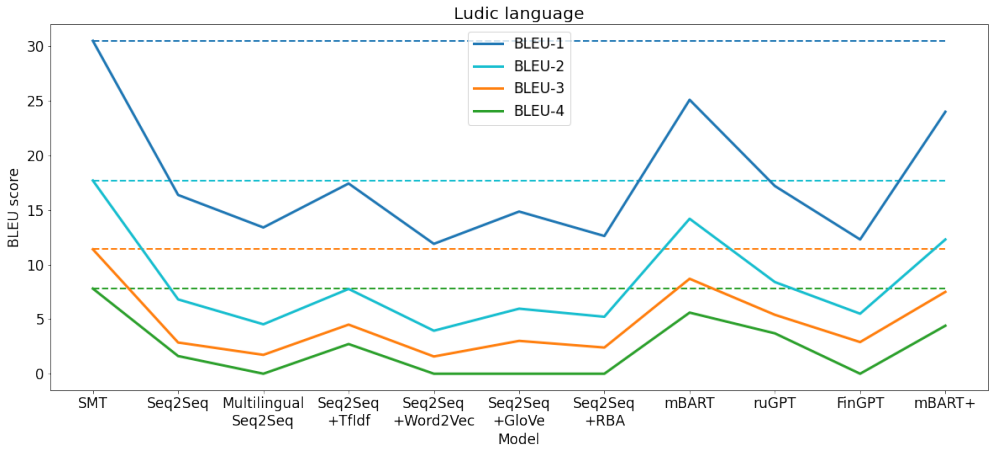

Figure 6: The comparison of the results of all the models trained from Ludic to Russian. Dashed lines represent the best result for the corresponding BLEU score. RBA = rule-based augmentation, MA = manual augmentation, mBART+ = mBART trained on the augmented with ruGPT dataset

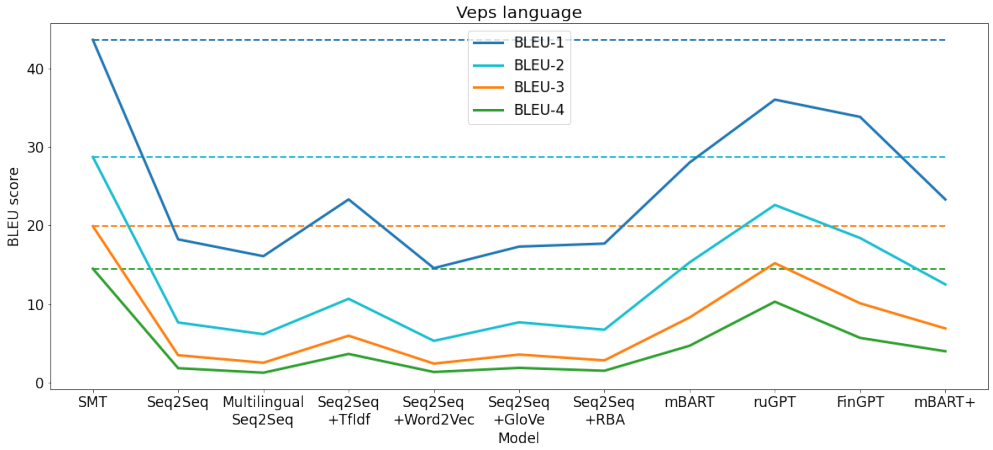

Figure 7: The comparison of the results of all the models trained from Veps to Russian. Dashed lines represent the best result for the corresponding BLEU score. RBA = rule-based augmentation, MA = manual augmentation, mBART+ = mBART trained on the augmented with ruGPT dataset

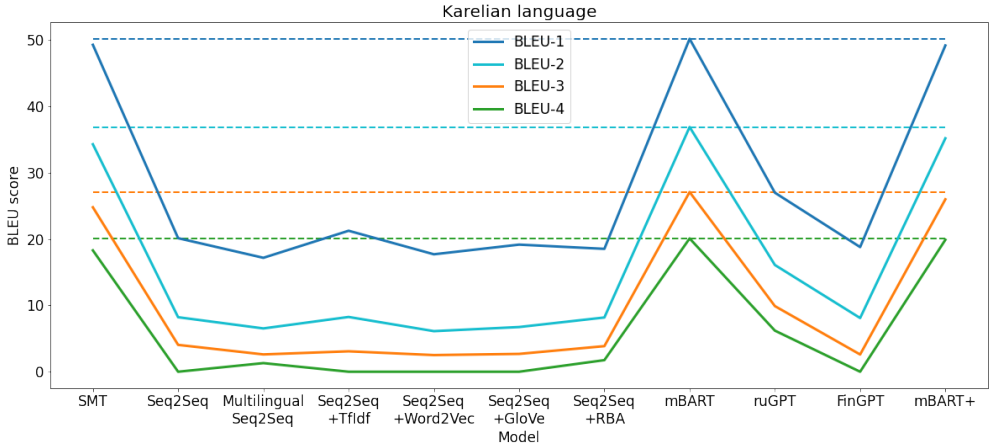

Figure 8: The comparison of the results of all the models trained from Karelian to Russian. Dashed lines represent the best result for the corresponding BLEU score. RBA = rule-based augmentation, MA = manual augmentation, mBART+ = mBART trained on the augmented with ruGPT dataset