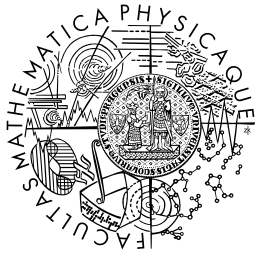

FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

MASTER THESIS

Bohdan Ihnatchenko

Multi-Target Machine Translation

Institute of Formal and Applied Linguistics

Supervisor of the master thesis: doc. RNDr. Bojar Ondřej, Ph.D.
Study programme: Computer Science
Study branch: Artificial Intelligence

Prague 2020

I declare that I carried out this master thesis independently, and only with the cited sources, literature and other professional sources. It has not been used to obtain another or the same degree.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Sb., the Copyright Act, as amended, in particular the fact that the Charles University has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 subsection 1 of the Copyright Act.

In ..... date ....

Author's signature

Dedication.

Title: Multi-Target Machine Translation

Author: Bohdan Ihnatchenko

Institute: Institute of Formal and Applied Linguistics

Supervisor: doc. RNDr. Bojar Ondřej, Ph.D., Institute of Formal and Applied Linguistics

Abstract: Abstract.

Keywords: Machine translation words

Contents

Introduction

With increasing availability of computational resources and enormous amount of publicly available corpora it is now possible to obtain a MT system which produces translations of acceptable quality. But in the use cases similar to conferences, where one speech is translated into multiple target languages, the same amount of models needs to be deployed. Another option is to use multilingual MT system for all needed languages together, which may lead to a decreased quality of translations.

1. Background

1.1 History of machine translation

1.2 Transformer model

1.3 Translation evaluation

2. Experiment setup

In this chapter we describe the data used for experiments, training setup and In this chapter we describe the data, training setup and experiments that were run to answer the questions asked in this thesis.

2.1 Dataset(s)

2.1.1 English to 36 languages

To observe effects of linguistic similarity of target languages is is important to examine enough possible variations of those. The OPUS dataset (Tiedemann [2012]) is an open and free collection of texts covers more than 90 languages with data from several domains.[1]

For our experiments the source language is English only. Sampling and splitting of the data is the one used for ELITR project.[2] For each of language pairs XXX TODO: link to the table and table itself and each sub-dataset the data was splitted to training, validation and testing sets. For each of the two latter sets 2000 random sentences were selected and the rest of the data remained for the training set. In cases where the sub-dataset contained less than 16000 sentence pairs no data went to the validation set. Later for each language pair there were 1000000 sentence pairs sampled from all training sub-sets. Firstly, if available, the sentences were taken from Europarl, then EUbooks, OpenSubtitles, and then all remaining sub-datasets. XXX should I write about this more/less? The same procedure was used to sample x000 of validation set sentences per each language pair. The test sets were left separate, so that the result on each domain would be observable.

Later an overlap in the source side of different language pairs was found. Although this would not directly lead to unfair increase of the test score, such sentence pairs were removed from the training sets. This filtering decreased the amound of sentence pairs to 0.85-0.95 millions per language pair.

2.2 Experiments

Given that the main question is in possible effects of target language groupings, the variable object in experiments is the data itself. Due to that, the setup similar to Johnson et al. [2017] was chosen. At the beginning of a source sentence there is a tag that signals into which target language the sentence needs to be translated.

XXX descripe task en2l1l2 is subsampled dataset with en2l1 and en2l2 sentences XXX Experiments: monolingual baseline, n-lingual baselines (random), group by language group, group by linguistic similarity.

Model tuning. First of all, the hyperparameters of MT model are tuned on couple of language pairs from one dataset. The parameters leading to the same

[1] Available at `http://opus.nlpl.eu/`

[2] `https://elitr.eu/wp-content/uploads/2019/07/D11.FINAL_.pdf`

result in shorter time were preferred. Then the selected parameters were used on all experimends with the dataset.

2.3 Training

2.3.1 Marian-NMT

Considering that varying element of described experiments is the data but not the model itself and, no less importantly, the amount of models needed to be trained, *MARIAN-NMT* (Junczys-Dowmunt et al. [2018]) was chosen for its speed and ease to use. The framework has efficient *sequence to sequence* and *Transformer* models implemented. Considering significantly better BLEU scores and training cost[3] the *Transformer* model was chosen to conduct the experiments.

The initial selection is made with respect to Popel and Bojar [2018].

2.3.2 Computational cluster

Many computations - cluster used.

Resourses are used by other people, disc quota is limited – parallel launching of experiments, switching to the next each 2 hours, saving only best models and the last one, removing subsampled datasets

[3]Vaswani et al. [2017]

3. Random choise of target languages

3.1 Overview

In this chapter we explore the effect of increasing number of target languages on the model performance in general. Multiple possible outcomes can be expected at this experiment: either performance drop due to the increased amount of languages to be processed by the model of the same size, or the opposite - performance increase due to shared knowledge gained by the model from bigger and varying dataset. Also, either of these options can be true for different target languages in different scale.

First of all, performance drop is expected. Considering that the size of the model is fixed, so is its capacity. At some moment adding more target languages should lead to the decrease in translation quality for each of every target language

3.2 Performance drop on massively multilingual setup

1-to-3, 5, 7, etc. models on en-to-36 dataset (0.9 mil. sentences per target language)

When the size of the model is fixed, adding more translation directions usually causes worsening of its performance. Multiple studies have shown this to be truth for many-to-many setup.

In Aharoni et al. [2019] models with up to 103 languages were tested. English centric in-house dataset was used to train En→Any and Any→En multilingual models. The average number of examples per language pair is 940k: for 13 out of the 102 pairs there were less than one million examples available. All languages from 5-to-5 model are present in 25-to-25, same is true for all languages from 25-to-25 with respect to 50-to-50 and so forth. In all cases they trained large Transformer model with 473.7M parameters. As can be seen on Table 3.1, the quality of translation is significantly worse when model is trained to translate more languages. However, it is worth reminding that this many-to-many experiment may have different reasons due to many-to-one direction present in it.

The decrease of model's performance with adding more target langueges is clearly shown in Aharoni et al. [2019].

3.3 Performance decrease on richer data sets

1 to 3, 4, 5 on UN corpus (much more sentence pairs per target language)

	En-Ar	En-Fr	En-Ru	En-Uk
5-to-5	**12.42**	**37.3**	**24.86**	16.48
25-to-25	11.77	36.79	23.24	**17.17**
50-to-50	11.65	35.83	21.95	15.32
75-to-75	10.69	34.35	20.7	14.59
103-to-103	10.25	34.42	19.9	13.89

Table 3.1: From [Aharoni et al., 2019] (part of Table 7): higher number of target languages decreases the model's performance

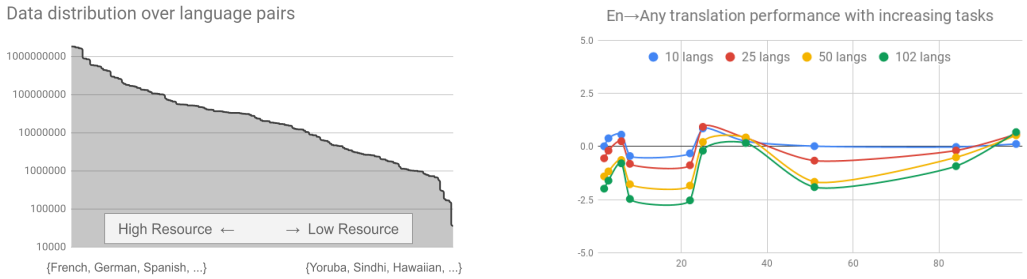

Figure 3.1: Performance decrease Arivazhagan et al. [2019]. On axis X there are languages sorted by amount of training data. The points visualized are 10 languages that are present in all setups from En $\leftrightarrow$ 10 to En $\leftrightarrow$ 102.

n_targets	mean	std	count
2	40.68	0.18	6
3	39.51	0.61	16
4	39.45	0.53	4
5	38.48	0.51	12

Table 3.2: BLEU score for Bulgarian at dataset Europarl/v7

n_targets	mean	count	std
2	18.8625	8.0	0.306769
3	17.5875	8.0	0.482368
4	17.8000	4.0	0.346410

Table 3.3: BLEU score for ru at dataset OpenSubtitles/v2016

4. Group by language groups

4.1 Language groups

1 to 2, 3, 4, 5, etc. models on en-to-36 dataset (0.9 mil. sentences per target language) compared with random runs

4.1.1 Germanic group

4.1.2 Slavic with cyrillic script

n_targets	mean		count		std	
	cyrillic	random	cyrillic	random	cyrillic	random
2	14.950000	15.2625	6.0	8.0	1.209545	0.757699
3	15.383333	14.5625	6.0	8.0	0.725029	0.616297
4	15.200000	14.2750	2.0	4.0	0.282843	0.206155

(a) MultiUN/v1

n_targets	mean		count		std	
	cyrillic	random	cyrillic	random	cyrillic	random
2	22.883333	22.875	6.0	8.0	0.507609	0.514782
3	22.333333	21.050	6.0	8.0	0.273252	0.705084
4	21.050000	20.950	2.0	4.0	0.494975	0.465475

(b) NewsCommentary/v11

n_targets	mean		count		std	
	cyrillic	random	cyrillic	random	cyrillic	random
2	19.266667	18.7375	6.0	8.0	0.273252	0.396187
3	19.350000	17.7875	6.0	8.0	0.653452	0.418970
4	18.900000	17.5750	2.0	4.0	0.282843	0.613052

(c) OpenSubtitles/v2018

Table 4.1: Mean BLEU score, its standard deviation and number of trained models (count) for Russian at various datasets

4.2 Selecting target languages by linguistic similarity

n_targets	mean		len		std	
	cyrillic	random	cyrillic	random	cyrillic	random
2	40.433333	40.683333	6.0	6.0	0.273252	0.183485
3	39.216667	39.506250	6.0	16.0	0.318852	0.611521
4	37.750000	39.450000	2.0	4.0	0.494975	0.525991
5	NaN	38.483333	NaN	12.0	NaN	0.511386

(a) Europarl/v7

n_targets	mean		len		std	
	cyrillic	random	cyrillic	random	cyrillic	random
2	23.683333	23.250000	6.0	6.0	0.598052	0.225832
3	23.216667	22.406250	6.0	16.0	0.470815	0.619106
4	22.300000	22.600000	2.0	4.0	0.424264	0.081650
5	-	21.741667	-	12.0	-	0.494439

(b) OpenSubtitles/v2018

Table 4.2: BLEU score for Bulgarian at various datasets

Conclusion

Bibliography

Roee Aharoni, Melvin Johnson, and Orhan Firat. Massively Multilingual Neural Machine Translation. In *Proceedings of the 2019 Conference of the North {A}merican Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3874–3884, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1388. URL https://www.aclweb.org/anthology/N19-1388.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, Wolfgang Macherey, Zhifeng Chen, and Yonghui Wu. Massively Multilingual Neural Machine Translation in the Wild: Findings and Challenges. jul 2019. URL http://arxiv.org/abs/1907.05019.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. *Transactions of the Association for Computational Linguistics*, 5:339–351, 2017. doi: 10.1162/tacl_a_00065. URL https://www.aclweb.org/anthology/Q17-1024.

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. Marian: Fast Neural Machine Translation in {C++}. In *Proceedings of ACL 2018, System Demonstrations*, Melbourne, Australia, 2018. URL https://arxiv.org/abs/1804.00344.

Martin Popel and Ondřej Bojar. Training tips for the transformer model. *The Prague Bulletin of Mathematical Linguistics*, 110, 03 2018. doi: 10.2478/pralin-2018-0002.

Jörg Tiedemann. Parallel data, tools and interfaces in opus. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may 2012. European Language Resources Association (ELRA). ISBN 978-2-9517408-7-7.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In I Guyon, U V Luxburg, S Bengio, H Wallach, R Fergus, S Vishwanathan, and R Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf.

List of Figures

List of Tables

List of Abbreviations

A. Attachments

A.1 First Attachment

A.1.1 Slavic with cyrillic script

n_targets	mean		len		std	
	cyrillic	random	cyrillic	random	cyrillic	random
2	13.516667	12.5125	6.0	8.0	2.117939	0.390741
3	14.266667	11.8875	6.0	8.0	1.148332	0.464258
4	14.100000	12.3000	2.0	6.0	0.565685	1.203329

Table A.1: BLEU score for mk at dataset GlobalVoices/v2015

n_targets	mean		len		std	
	cyrillic	random	cyrillic	random	cyrillic	random
2	13.483333	12.575000	6.0	8.0	1.880869	0.395511
3	14.433333	12.062500	6.0	8.0	0.989276	0.465794
4	14.150000	12.233333	2.0	6.0	0.494975	1.155278

Table A.2: BLEU score for mk at dataset GlobalVoices/v2017q3

n_targets	mean		len		std	
	cyrillic	random	cyrillic	random	cyrillic	random
2	5.700000	6.0875	6.0	8.0	0.950789	0.464258
3	6.333333	5.9750	6.0	8.0	0.791623	0.547070
4	6.250000	6.6000	2.0	6.0	0.212132	0.244949

Table A.3: BLEU score for mk at dataset KDE4/v2

n_targets	mean		len		std	
	cyrillic	random	cyrillic	random	cyrillic	random
2	22.866667	23.425	6.0	8.0	0.970910	0.528475
3	23.116667	22.325	6.0	8.0	0.397073	0.443203
4	21.950000	22.350	2.0	6.0	0.494975	0.372827

Table A.4: BLEU score for mk at dataset OpenSubtitles/v2016

n_targets	mean		len		std	
	cyrillic	random	cyrillic	random	cyrillic	random
2	24.050000	24.0500	6.0	8.0	0.561249	0.526444
3	23.533333	23.2125	6.0	8.0	0.417931	0.313676
4	22.700000	23.1000	2.0	6.0	0.565685	0.404969

Table A.5: BLEU score for mk at dataset OpenSubtitles/v2018

n_targets	mean		len		std	
	cyrillic	random	cyrillic	random	cyrillic	random
2	10.483333	10.337500	6.0	8.0	2.851958	1.151319
3	12.166667	10.125000	6.0	8.0	1.531883	0.662786
4	12.800000	10.683333	2.0	6.0	0.141421	1.231936

Table A.6: BLEU score for mk at dataset SETIMES/v1

n_targets	mean		len		std	
	cyrillic	random	cyrillic	random	cyrillic	random
2	14.516667	13.187500	6.0	8.0	2.961362	0.598659
3	15.333333	12.312500	6.0	8.0	1.823915	0.488255
4	15.400000	12.833333	2.0	6.0	0.424264	1.233964

Table A.7: BLEU score for mk at dataset SETIMES/v2

n_targets	mean		len		std	
	cyrillic	random	cyrillic	random	cyrillic	random
2	32.400000	32.816667	6.0	6.0	0.328634	0.285774
3	31.283333	32.350000	6.0	16.0	0.402078	1.002663
4	29.900000	32.650000	2.0	4.0	0.424264	0.264575
5	NaN	31.500000	NaN	12.0	NaN	0.463191

Table A.8: BLEU score for bg at dataset DGT/v4

n_targets	mean		len		std	
	cyrillic	random	cyrillic	random	cyrillic	random
2	15.150000	15.383333	6.0	6.0	0.327109	0.865833
3	14.616667	15.593750	6.0	16.0	0.331160	0.899977
4	14.050000	15.950000	2.0	4.0	0.636396	0.310913
5	NaN	15.550000	NaN	12.0	NaN	0.545227

Table A.9: BLEU score for bg at dataset EMEA/v3

n_targets	mean		len		std	
	cyrillic	random	cyrillic	random	cyrillic	random
2	36.550000	36.9500	6.0	6.0	0.367423	0.333167
3	34.833333	35.3625	6.0	16.0	0.422690	0.732917
4	32.500000	35.7500	2.0	4.0	0.707107	0.412311
5	NaN	34.2500	NaN	12.0	NaN	0.633174

Table A.10: BLEU score for bg at dataset EUbookshop/v2

n_targets	mean		len		std	
	cyrillic	random	cyrillic	random	cyrillic	random
2	40.433333	40.683333	6.0	6.0	0.273252	0.183485
3	39.216667	39.506250	6.0	16.0	0.318852	0.611521
4	37.750000	39.450000	2.0	4.0	0.494975	0.525991
5	NaN	38.483333	NaN	12.0	NaN	0.511386

Table A.11: BLEU score for bg at dataset Europarl/v7

n_targets	mean		len		std	
	cyrillic	random	cyrillic	random	cyrillic	random
2	30.466667	31.00000	6.0	6.0	0.355903	0.464758
3	29.750000	30.16875	6.0	16.0	0.388587	0.662036
4	29.200000	30.75000	2.0	4.0	0.282843	0.310913
5	NaN	29.45000	NaN	12.0	NaN	0.239317

Table A.12: BLEU score for bg at dataset JRC-Acquis/v3.0

n_targets	mean		len		std	
	cyrillic	random	cyrillic	random	cyrillic	random
2	7.100000	7.183333	6.0	6.0	0.328634	0.098319
3	7.066667	7.356250	6.0	16.0	0.196638	0.576158
4	7.100000	7.075000	2.0	4.0	0.141421	0.492443
5	NaN	6.966667	NaN	12.0	NaN	0.379793

Table A.13: BLEU score for bg at dataset KDE4/v2

n_targets	mean		len		std	
	cyrillic	random	cyrillic	random	cyrillic	random
2	19.550000	18.833333	6.0	6.0	0.463681	0.216025
3	18.616667	18.025000	6.0	16.0	0.318852	0.637181
4	18.000000	17.975000	2.0	4.0	0.282843	0.206155
5	NaN	17.591667	NaN	12.0	NaN	0.501739

Table A.14: BLEU score for bg at dataset OpenSubtitles/v1

n_targets	mean		len		std	
	cyrillic	random	cyrillic	random	cyrillic	random
2	23.016667	22.60000	6.0	6.0	0.530723	0.219089
3	22.316667	21.75625	6.0	16.0	0.348807	0.542794
4	21.150000	21.32500	2.0	4.0	0.212132	0.427200
5	NaN	21.02500	NaN	12.0	NaN	0.587947

Table A.15: BLEU score for bg at dataset OpenSubtitles/v2016

n_targets	mean		len		std	
	cyrillic	random	cyrillic	random	cyrillic	random
2	23.683333	23.250000	6.0	6.0	0.598052	0.225832
3	23.216667	22.406250	6.0	16.0	0.470815	0.619106
4	22.300000	22.600000	2.0	4.0	0.424264	0.081650
5	NaN	21.741667	NaN	12.0	NaN	0.494439

Table A.16: BLEU score for bg at dataset OpenSubtitles/v2018

n_targets	mean		len		std	
	cyrillic	random	cyrillic	random	cyrillic	random
2	23.016667	22.966667	6.0	6.0	0.910860	1.050079
3	22.866667	22.506250	6.0	16.0	0.585377	0.868308
4	22.250000	23.025000	2.0	4.0	0.070711	0.386221
5	NaN	22.350000	NaN	12.0	NaN	0.414510

Table A.17: BLEU score for bg at dataset SETIMES/v1

n_targets	mean		len		std	
	cyrillic	random	cyrillic	random	cyrillic	random
2	27.45	27.15000	6.0	6.0	0.320936	0.504975
3	26.90	26.21875	6.0	16.0	0.572713	0.693031
4	25.65	26.45000	2.0	4.0	0.636396	0.387298
5	NaN	25.75000	NaN	12.0	NaN	0.556776

Table A.18: BLEU score for bg at dataset SETIMES/v2

n_targets	mean		len		std	
	cyrillic	random	cyrillic	random	cyrillic	random
2	5.966667	5.833333	6.0	6.0	0.273252	0.216025
3	5.983333	5.675000	6.0	16.0	0.222860	0.191485
4	5.900000	5.650000	2.0	4.0	0.141421	0.057735
5	NaN	5.725000	NaN	12.0	NaN	0.195982

Table A.19: BLEU score for bg at dataset Tanzil/v1

n_targets	mean		len		std	
	cyrillic	random	cyrillic	random	cyrillic	random
2	11.766667	12.333333	6.0	6.0	0.366970	0.403320
3	11.716667	12.481250	6.0	16.0	0.331160	1.119654
4	10.950000	12.775000	2.0	4.0	0.070711	0.262996
5	NaN	12.566667	NaN	12.0	NaN	0.600505

Table A.20: BLEU score for bg at dataset Wikipedia/v1.0

n_targets	mean		len		std	
	cyrillic	random	cyrillic	random	cyrillic	random
2	1.816667	1.9500	6.0	6.0	0.343026	0.281069
3	2.000000	1.6500	6.0	8.0	0.236643	0.507093
4	1.950000	2.2375	2.0	8.0	0.070711	0.315945

Table A.21: BLEU score for uk at dataset KDE4/v2

n_targets	mean		len		std	
	cyrillic	random	cyrillic	random	cyrillic	random
2	12.433333	11.816667	6.0	6.0	0.843010	0.116905
3	12.283333	11.350000	6.0	8.0	0.921774	0.507093
4	12.050000	11.237500	2.0	8.0	0.353553	0.459619

Table A.22: BLEU score for uk at dataset OpenSubtitles/v2016

n_targets	mean		len		std	
	cyrillic	random	cyrillic	random	cyrillic	random
2	13.150000	12.516667	6.0	6.0	0.592453	0.172240
3	12.633333	11.637500	6.0	8.0	0.366970	0.244584
4	11.950000	11.812500	2.0	8.0	0.353553	0.348210

Table A.23: BLEU score for uk at dataset OpenSubtitles/v2018

n_targets	mean		len		std	
	cyrillic	random	cyrillic	random	cyrillic	random
2	13.40	14.5500	6.0	6.0	1.744706	0.273861
3	12.25	13.6125	6.0	8.0	1.250200	0.775403
4	9.60	13.1750	2.0	8.0	0.707107	0.413176

Table A.24: BLEU score for uk at dataset Tatoeba/v2

n_targets	mean		len		std	
	cyrillic	random	cyrillic	random	cyrillic	random
2	8.300000	8.0875	6.0	8.0	0.154919	0.203101
3	8.383333	7.5625	6.0	8.0	0.213698	0.342000
4	8.200000	7.7500	2.0	4.0	0.141421	0.238048

Table A.25: BLEU score for ru at dataset Books/v1

n_targets	mean		len		std	
	cyrillic	random	cyrillic	random	cyrillic	random
2	13.516667	13.750	6.0	8.0	0.479236	0.267261
3	13.600000	12.875	6.0	8.0	0.404969	0.536523
4	13.200000	12.825	2.0	4.0	0.424264	0.206155

Table A.26: BLEU score for ru at dataset GlobalVoices/v2015

n_targets	mean		len		std	
	cyrillic	random	cyrillic	random	cyrillic	random
2	14.55	14.7875	6.0	8.0	0.557674	0.322656
3	14.80	13.7250	6.0	8.0	0.404969	0.492080
4	14.20	13.6750	2.0	4.0	0.424264	0.170783

Table A.27: BLEU score for ru at dataset GlobalVoices/v2017q3

n_targets	mean		len		std	
	cyrillic	random	cyrillic	random	cyrillic	random
2	4.733333	4.7750	6.0	8.0	0.875595	0.452769
3	5.000000	4.8625	6.0	8.0	0.544059	0.337797
4	4.950000	6.4500	2.0	4.0	0.212132	0.173205

Table A.28: BLEU score for ru at dataset KDE4/v2

n_targets	mean		len		std	
	cyrillic	random	cyrillic	random	cyrillic	random
2	14.950000	15.2625	6.0	8.0	1.209545	0.757699
3	15.383333	14.5625	6.0	8.0	0.725029	0.616297
4	15.200000	14.2750	2.0	4.0	0.282843	0.206155

Table A.29: BLEU score for ru at dataset MultiUN/v1

n_targets	mean		len		std	
	cyrillic	random	cyrillic	random	cyrillic	random
2	22.883333	22.875	6.0	8.0	0.507609	0.514782
3	22.333333	21.050	6.0	8.0	0.273252	0.705084
4	21.050000	20.950	2.0	4.0	0.494975	0.465475

Table A.30: BLEU score for ru at dataset News-Commentary/v11

n_targets	mean		len		std	
	cyrillic	random	cyrillic	random	cyrillic	random
2	16.700000	16.8625	6.0	8.0	0.536656	0.680205
3	16.016667	14.7375	6.0	8.0	0.147196	0.568048
4	14.950000	15.1000	2.0	4.0	0.212132	0.182574

Table A.31: BLEU score for ru at dataset News-Commentary/v9.0

n_targets	mean		len		std	
	cyrillic	random	cyrillic	random	cyrillic	random
2	20.483333	20.6000	6.0	8.0	0.601387	0.453557
3	20.016667	18.7625	6.0	8.0	0.248328	0.783650
4	18.600000	18.5500	2.0	4.0	0.282843	0.369685

Table A.32: BLEU score for ru at dataset News-Commentary/v9.1

n_targets	mean		len		std	
	cyrillic	random	cyrillic	random	cyrillic	random
2	16.633333	16.5625	6.0	8.0	0.242212	0.250357
3	17.033333	15.4750	6.0	8.0	0.512510	0.517549
4	16.850000	15.9750	2.0	4.0	0.353553	0.499166

Table A.33: BLEU score for ru at dataset OpenSubtitles/v1

n_targets	mean		len		std	
	cyrillic	random	cyrillic	random	cyrillic	random
2	19.35	18.8625	6.0	8.0	0.459347	0.306769
3	19.45	17.5875	6.0	8.0	0.896103	0.482368
4	18.80	17.8000	2.0	4.0	0.282843	0.346410

Table A.34: BLEU score for ru at dataset OpenSubtitles/v2016

n_targets	mean		len		std	
	cyrillic	random	cyrillic	random	cyrillic	random
2	19.266667	18.7375	6.0	8.0	0.273252	0.396187
3	19.350000	17.7875	6.0	8.0	0.653452	0.418970
4	18.900000	17.5750	2.0	4.0	0.282843	0.613052

Table A.35: BLEU score for ru at dataset OpenSubtitles/v2018

n_targets	mean		len		std	
	cyrillic	random	cyrillic	random	cyrillic	random
2	3.466667	4.6375	6.0	8.0	0.344480	0.975320
3	3.916667	3.9375	6.0	8.0	0.194079	0.440576
4	4.150000	3.9500	2.0	4.0	0.212132	0.645497

Table A.36: BLEU score for ru at dataset PHP/v1

n_targets	mean		len		std	
	cyrillic	random	cyrillic	random	cyrillic	random
2	11.500000	12.8625	6.0	8.0	1.473771	1.712089
3	12.433333	12.2500	6.0	8.0	1.269120	0.725062
4	12.900000	13.0000	2.0	4.0	0.141421	0.588784

Table A.37: BLEU score for ru at dataset ParaCrawl/v1

n_targets	mean		len		std	
	cyrillic	random	cyrillic	random	cyrillic	random
2	15.066667	15.0625	6.0	8.0	0.287518	0.392565
3	15.383333	14.3625	6.0	8.0	0.318852	0.708998
4	15.150000	14.5750	2.0	4.0	0.494975	0.556028

Table A.38: BLEU score for ruat dataset TED2013/v1.1

n_targets	mean		len		std	
	cyrillic	random	cyrillic	random	cyrillic	random
2	2.466667	2.4875	6.0	8.0	0.051640	0.083452
3	2.583333	2.2875	6.0	8.0	0.116905	0.099103
4	2.500000	2.4750	2.0	4.0	0.000000	0.095743

Table A.39: BLEU score for ru at dataset Tanzil/v1

n_targets	mean		len		std	
	cyrillic	random	cyrillic	random	cyrillic	random
2	27.133333	26.8250	6.0	8.0	0.659293	0.720615
3	27.316667	24.4625	6.0	8.0	0.890880	0.814051
4	26.250000	25.1750	2.0	4.0	0.777817	0.613052

Table A.40: BLEU score for ru at dataset Tatoeba/v2

n_targets	mean		len		std	
	cyrillic	random	cyrillic	random	cyrillic	random
2	7.40	10.400	6.0	8.0	1.533623	2.179777
3	9.25	9.475	6.0	8.0	0.884873	0.967692
4	7.65	8.175	2.0	4.0	0.353553	0.793200

Table A.41: BLEU score for ru at dataset UN/v20090831

n_targets	mean		len		std	
	cyrillic	random	cyrillic	random	cyrillic	random
2	11.450000	12.125	6.0	8.0	1.087658	1.053904
3	11.883333	11.525	6.0	8.0	0.856543	0.567576
4	11.700000	11.700	2.0	4.0	0.141421	0.702377

Table A.42: BLEU score for ru at dataset Wikipedia/v1.0