The Impact of Multilinguality and Tokenization on Statistical Machine Translation

Alidar Asvarov, Andrey Grabovoy
Antiplagiat
Moscow, Russia
asvarov,grabovoy@ap-team.ru

***Abstract*—Multilingual neural machine translation systems has achieved state-of-the-art results on translation quality, especially for low-resource languages, yet statistical machine translations systems has not been trained and examined in similar multilingual setup. This work defines a multilingual statistical machine translation system as a many-to-one system capable of translating from any of the predefined languages to the one target language. We study how the multilingual setting affects translations quality compared to a regular one-to-one language machine translation system. And we examine how this setting affects related languages with different amount of training data. The research is conducted in multiple languages of different language families. The impact of different tokenizers and preprocessing methods is researched as well. Specifically, we compare the default Moses tokenizer with the SentencePiece tokenizer, as well as dedicated Chinese and Japanese word splitters. We also investigate the impact of lowercasing and conduct our experiments on data of different sizes. We find out that multilinguality gives a small gain across all of the metrics. Languages with sufficient amount of good quality training data do not affect the quality of related languages with lesser quality data. The SentencePiece tokenizer shows lower BLEU scores on average, but outperforms other tokenizers on chrF++ and METEOR metrics. Lowercasing increases scores of all metrics in all of the scenarios.**

I. Introduction

For the last decade Statistical Machine Translation (SMT) was superseded by Neural Machine Translation (NMT), which achieves much higher translation quality and can generalize much better for out-of-domain data [1]–[3]. Apart from that, neural based systems proved to be capable of translating from and to hundreds of languages [4]. This multilingual capabilities hugely reduces the deployment cost at training and inference, as one system to serve translation for multiple directions. However, training and inferencing an NMT system requires a GPU and large amount of parallel corpora, which could be difficult to obtain, especially for low-resource languages [5], [6]. In addition, renting and scaling CPUs in cloud environments is much cheaper than GPUs. While distilled and light-weight versions of massively multilingual models exists [7], they still require a GPU, and scaling this type of processing power as the number of translation requests grows is expensive. And in some production circumstances computational resources are limited and the best translation quality is not required. But what is more important is translation speed and good (or not bad) quality [8]. In this scenarios statistical machine translation is still in use and state-of-the-art SMT toolkit is Moses [9].

An essential text pre-processing step for training any machine translation system is tokenization. The tokenization process breaks text into chunks which can be considered as discrete elements. There are two main type of tokenizers: word tokenizers and subword tokenizers. The former usually use whitespaces and punctuations as delimiters of words. The latter should be trained and is based on subword frequencies, i.e. characters that often appear together will be merged into a subword. Languages, like Chinese or Japanese, are difficult to tokenize, as they doesn't have separator between words. Thus, developing efficient word segmentation algorithm is essential for machine translation on these languages. This makes subword tokenizers exceptionally convenient, as they are language-agnostic, meaning that it is possible to preprocess an entire multilingual corpus with just one tokenizer, skipping tokenizer selection for each language. Moreover, subword tokenization eliminates the problem of out-of-vocabulary words in machine translataion. However, a subword tokenizer should be trained on high quality and diverse corpora.

Historically, statistical translation systems were developed only with word tokenizers and subword tokenizers are widely used in neural machine translation.

In this work we hypothesize that utilizing sub-word tokenization and training an SMT system to translate from multiple languages may improve overall translation quality and related language may help each other, especially in situations when one of them is rich in available parallel resources, while others aren't. In addition, replacing multiple machine translation systems with one may prove to be effective and save computational and disk resources in production environment.

II. Related Work

Many studies on comparing statistical and neural machine translation systems, different tokenizers and pre-processing techniques have been conducted. However, to the best of our knowledge, this paper is the first study of multilingual statistical machine translation model training across multiple languages of different language families.

Nevertheless, some simpler forms of multilingualism for SMT have been studied.

1) *pivot-based*: utilizing *pivot* or *bridge* language to circumvent the data bottleneck [10],

2) *multi-source*: method of input combination to generate lattices for multi-source translation within a single translation model [11],
3) *SMT involving related languages*: with the mose common approaches involved script unification by mapping to a common script such as Devanagari [12] or transliteration [13]. One of the studies on multilingual SMT is [14], where translation system from standard and dialect versions of Arabic to English is implemented. The authors also propose classifier based multilingual system (a classifier chooses). However, the study concludes that the input text classification and subsequent selection of the monolingual system outperforms multilingual and monolingual systems.

The paper [15] assess the impact of the tokenization on the quality of the final translation on neural machine translation (NMT). The authors experiment on five tokenizers over ten language pairs and come to the conclusion that the tokenization significantly affects the resulting translation quality. The best tokenizer should be carefully selected for each language pair and can achieve gains of up to 12 points of BLEU.

In *SentencePiece Experiments* [16] multiple experiments on neural machine translation are provided comparing different segmentation algorithms (subword and word based) and various pre-tokenization methods. The study concludes, that subword methods, such as SentencePiece [17], [18], outperform word-based methods for Japanese-English language pair. Different pre-tokenization methods can improve translation quality as well.

The work [19] studies the impact of word segmentation for Chinese-English machine translation. The study shows that the translation directory matters and word segmentation is a necessity for Chinese-to-English translation, but not for English-to-Chinese one. The authors examine different segmentation strategies (both statistical and dictionary-based) and come to the conclusion that the key to better machine translation is not the segmentation strategy choice, but the linguistic resources for supporting segmenters.

The paper [20] systematically compares SMT and NMT models for Arabic-English translation on data preprecossed by various tokenization algorithms. Experiment results show that applying sub-word tokenization gives a slight improvement for statistical machine translation system.

Our contributions are as follows:

- research the impact of training an SMT system in multilingual manner, i.e. translating from many languages to one,
- study the impact of sub-word and word based tokenizers on quality of statistical machine translation,
- experimentation with different language pairs of different language families, different training data size and text case,
- analyze and explain the dynamics of evaluation metrics across experiments.

III. Approach

Given a set

$$\mathbf{X}_{l_{src}} \in \{x_1^{l_{src}}..x_n^{l_{src}}\}$$

of n sentences $x_i^{l_{src}}$ from a source language l_{src} and a set $\mathbf{Y}_{l_{tgt}} \in \{y_1^{l_{tgt}}..y_n^{l_{tgt}}\}$ of n translations $y_i^{l_{tgt}}$ to a target language l_{tgt}, we define a monolingual dataset as a pair $\{\mathbf{X}_{l_{src}}; \mathbf{Y}_{l_{tgt}}\}$. Given a set of source languages

$$\mathcal{L} \in \{l_{src_1}, .., l_{src_j}\}$$

, their sentences set $\mathcal{X} = \{\mathbf{X}_{l_{src_1}}, .., \mathbf{X}_{l_{src_j}}\}$ and their translations set $\mathcal{Y} = \{\mathbf{Y}_1, .., \mathbf{Y}_j\}$, we define a multilingual dataset as a set $\{\mathcal{X}; \mathcal{Y}\}$ of source sentences on different languages and their translations to the target language l_{tgt}. In all the experiments we set l_{tgt} as English language.

In this paper we refer to a a monolingual SMT as a system trained on a pair $\{\mathbf{X}_{l_{src}}; \mathbf{Y}_{l_{tgt}}\}$, i.e. a system for translation from one language to another. And we refer to a multilingual SMT as a system trained on a pair $\{\mathcal{X}; \mathcal{Y}\}$, i.e. a system capable of translating from any of the predefined languages to the one target language.

The goal of experiments is to study capabilities of multilingual and monolingual versions of Moses and compare their quality across all of the selected languages. We hypothesize that training Moses in multilingual setting may improve performance for related languages. We also compare different types of tokenizers, specifically the default Moses tokenizer and SentencePiece tokenizer. The impact of lowercasing is also studied.

In this work we report BLEU [21], [22], chrF++ [23], [24] and METEOR [25] scores. We utilize `sacrebleu` library [26] for BLEU and chrF++ calculations and Huggingface's `evaluate` library [27] for calculating METEOR score. We use devtest split of the FLORES-200 Evaluation Benchmark [4], [28], [29] to measure the quality of trained models in our experiments.

IV. Experiment 1: Multilingual Moses

A. Experimental settings

In this experiment we train multiple Moses SMT systems to translate from the following languages to English.

1) *The Romance language family*: Italian, French, Spanish, Portuguese.
2) *The Germanic language family*: German and Dutch.
3) *The Slavic language family*: Russian, Belarussian, Polish and Czech.
4) *The Turkic language family*: Kazakh, Kyrgyz, Turkish, Uzbek.
5) *Other language families*: Hindi, Simplified Chinese and Japanese.

Some of these languages are high-resource languages, some are not. We hypothesize that by training in the multilingual setting the quality on low-resource languages may be improved by high-resource related languages.

B. Data preparation

To avoid domain shift one needs to employ data sources which were collected in the same way for each language. The No Language Left Behind (NLLB) dataset [4] fits these criteria. It was built by large-scale bitexts mining from the web using LASER3 sentence encoder [30]. The dataset was downloaded from the OPUS project [31] website [32].

Sentence pairs in the NLLB dataset are sorted by their LASER3 score. Often sentences of the same domain have similar LASER3 scores and are located together. For example, religious texts, Bible and Quran translations usually have a big score and are often located in the beginning of the dataset for many languages. Thus, to avoid domain shift we employ the following sentence pair selection algorithm. We consider sentence pairs whose LASER3 score is greater than 1.07. Then we limit the number of pairs to 20 millions and take uniform random samples of 10000 and 250000 sentence pairs, which creates two dataset versions for each language. `MosesPunktNormalizer` from `sacremoses` library [33] was used to normalize sentences, then sentences were tokenized. Additional version of lowercased sentences is also created, lowercasing is applied before tokenization. We utilize `MosesTokenizer` from `sacremoses` library. As a sentencepiece tokenizer we employ pre-trained NLLB tokenizer [34] and after the tokenization process we remove `<unk>` tokens. We filter out too short and too long sentences with the default `clean-corpus-n.perl` Moses cleaninig utility using parameters 1 and 80 respectively.

We apply this procedure to every language pair, so each language pair data has eight versions with varying size (10k and 250k), casing (normalcased and lowercased), tokenization (Moses and SentencePiece). Multilingual version of the dataset is created by concatenating sentence pairs of the same size, case and tokenization for all languages, which gives 8 additional dataset For short, we refer to monolingual datasets as mono and to concatenated multilingual datasets as multi.

It turned out that we have to give up with Chinese and Japanese languages. The reason is that with the data tokenized by the Moses tokenizer the training or prunning processes constantly fail and produce errors. Even after preprocessing, a lot of problematic tokens remain in training dataset and those tokens make Moses produce errors. Quick look at Chinese and Japanese subsets on NLLB made us realize that sentences in that subsets are extremely contaminated and still contain a lot of artefacts, regardless of performed preprocessing. The experiment was conducted without these languages.

C. Results

In total we trained 128 SMT systems with different data configurations: monolingual and multilingual, Moses tokeinzation and SentencePiece tokenization, lowercase and normalcase, 10k and 250k size, mono and multi versions.

The default Moses tokenizer achieves higher BLEU score on average, which we report in Table I for 250k normalcased dataset version. Better results for the same dataset version on METEOR and chrF++ metrics are achieved by SentencePiece tokenizer, which are reported in Tables II and III respectively. However, BLEU score is higher for SentencePiece across all of the Turkic language family, which can be explained by their agglutinative nature. METEOR and chrF++ scores are higher for the Romance language family, which can also be explained as English comes from the same language family, has a lot of similar and shared vocabulary.

Scores for other dataset versions, as well as full phrase table and lexical reordering table sizes comparisons can be found in appendix A.

Multilingual settings achieve a little bit higher scores across all of the metrics, however this may simply be caused by training on more data. Related languages don't help each other in any significant way. For example, according to metrics, Russian has much more quality data, than Belarussian. However, training them together gives increase of the same order of magnitude for both languages.

In general, everything depends on the quality of the data. One can clearly see big difference in minimum and maximum scores across all the languages. This could be explained by the quality of parallel sentences for those languages. Moreover, most of the European languages have a significant vocabulary intersection with English and among themselves, have parallel data of sufficient quality and are written in the Latin script. So even on smaller data there are high scores for these languages, especially when comparing to other languages.

We evaluate disk usage of phrase tables (pruned and binarized) and lexical reordering tables for every trained SMT system. Sizes of these tables in Kilobytes for multilingual and sum of monolingual systems are shown in Tables IV and V. Full tables with sizes for every language pair are placed in Appendix A.

Multilingual translation system in most of the scenarios requires less disk space than sum of monolingual systems. For SentencePiece versions phrase tables weigh significantly more than Moses tokenizer versions, but lexical reordering tables weight a little bit less.

V. Experiment 2: Exploring dedicated word splitters for Japanese and Chinese

A. Experimental settings

Because training SMT system for Chinese and Japanese languages has failed in the previous experiment, we conduct a second experiment, where we employ dedicated word-splitters and more quality data. For this experiment we select smaller subset of languages: Spanish, Italian, Russian, Ukrainian, Arabic, Hindi, Simplified Chinese and Japanese.

B. Data preparation

We choose HPLTv1.1 [35] parallel corpora for Hindi, Arabic and Simplified Chinese, ParaCrawl v9 [36] for Spanish, Italian and Russian, JParaCrawl v3.0 [37] for Japanese and MaCoCu v2 for Ukrainian [38]. All of the data was downloaded from the OPUS project as well. In this experiment we use again the default Moses tokenizer, NLLB SentencePiece tokenizer, as well as Jieba tokenizer [39] for Chinese language

TABLE I. SACREBLEU SCORES FOR THE MOSES SMT TRAINED ON 250K NORMALCASED SUBSET

Pair	Moses Tokenizer Multi	SentencePiece Multi	Moses Tokenizer Mono	SentencePiece Mono
it-en	**18.90**	17.98	17.96	17.03
fr-en	**27.27**	25.94	25.20	24.34
es-en	**16.37**	15.67	15.40	14.86
pt-en	**30.67**	29.73	29.02	28.31
de-en	**23.47**	22.36	22.29	21.78
nl-en	19.01	**19.09**	18.74	18.78
ru-en	**18.31**	17.32	17.09	16.33
be-en	**10.31**	9.67	9.66	9.01
cs-en	21.70	**21.97**	21.02	20.89
pl-en	**14.96**	14.73	14.37	14.12
ky-en	4.94	**5.88**	3.97	4.72
kk-en	8.50	**9.06**	7.23	8.08
uz-en	8.69	**9.55**	8.25	8.78
tr-en	10.74	**11.69**	10.35	11.04
hi-en	**15.26**	14.58	14.03	13.51
average	**16.61**	16.35	15.64	15.44

TABLE II. CHRF++ SCORES FOR THE MOSES SMT TRAINED ON 250K NORMALCASED SUBSET

Pair	Moses Tokenizer Multi	SentencePiece Multi	Moses Tokenizer Mono	SentencePiece Mono
it-en	**50.02**	49.61	49.53	49.07
fr-en	**56.17**	55.47	54.98	54.63
es-en	**47.23**	47.06	46.64	46.33
pt-en	**58.71**	58.47	57.61	57.52
de-en	50.40	**52.27**	49.84	51.99
nl-en	47.58	**48.82**	47.56	48.72
ru-en	43.13	**46.90**	42.52	45.98
be-en	36.93	**40.12**	36.55	39.38
cs-en	49.10	**50.99**	48.94	50.70
pl-en	42.77	**44.41**	42.65	44.08
ky-en	24.66	**34.05**	22.46	31.18
kk-en	30.58	**38.80**	28.53	37.13
uz-en	36.48	**40.34**	36.54	39.73
tr-en	39.32	43.41	39.62	**43.43**
hi-en	44.04	**45.95**	43.49	45.38
average	43.81	**46.44**	43.16	45.68

and MeCab tokenizer [40] for Japanese language. We repeat all the data preprocessing steps as in the first experiment. We limit ourselves with 250k normalcased training pairs for each language.

C. Results

As we have chosen different data sources, we were able to successfully train Moses SMT systems on Chinese and Japanese languages with the default Moses tokenization. In total we trained 15 machine translation systems: 2 multilingual with the Moses and the SentencePiece tokenizers, 8 for each language with the Moses tokenizer, 8 for each language with the SentencePiece tokenizer, 1 for Chinese with Jieba and 1 for Japanese with MeCab.

This experiment confirms the results from the first experiment. Multilingual Moses achieves an improvement over monolingual one across all of the metrics. The Moses tokenizer completely loses for Chinese and Japanese languages. BLEU scores are reported in Table VI and one can see that the SentencePiece tokenizer performs the worst on this metric, but has comparable results with Jieba and MeCab word splitters. SentencePiece has the best chrF++ scores, which are reported in Table VII. As in the first experiment, the Moses tokenizer achieves higher METEOR scores on Romance languages, which is presented in Table VIII. SentencePiece has comparable METEOR scores with Jieba and MeCab. Chinese and Japanese languages show significantly worse scores than other languages. This could be explained by poorer data quality or that these languages have much bigger vocabulary, much less shared vocabulary and require more data to train. Full phrase table and lexical reordering table sizes comparisons can be found in appendix B.

SentencePiece tokenizer is more versatile and if trained on large and diverse enough corpora it can be used for any language, which is an advantage.

TABLE III. METEOR SCORES FOR THE MOSES SMT TRAINED ON 250K NORMALCASED SUBSET

Pair	Moses Tokenizer Multi	SentencePiece Multi	Moses Tokenizer Mono	SentencePiece Mono
it-en	**54.45**	53.70	54.04	53.27
fr-en	**63.23**	62.52	62.43	61.75
es-en	**51.38**	50.94	50.93	50.52
pt-en	**65.46**	65.25	64.59	64.49
de-en	55.78	**57.92**	55.15	57.63
nl-en	51.11	52.63	51.20	**52.72**
ru-en	49.24	**51.21**	48.29	50.08
be-en	40.10	**41.80**	39.80	40.52
cs-en	53.23	55.42	53.48	**55.56**
pl-en	45.48	**46.93**	45.21	46.53
ky-en	25.77	**33.09**	23.67	30.47
kk-en	33.48	**40.16**	31.66	38.42
uz-en	35.73	**41.23**	35.85	41.01
tr-en	39.90	44.91	40.28	**45.46**
hi-en	50.39	**50.79**	49.97	50.25
average	47.65	**49.90**	47.10	49.25

TABLE IV. SIZES OF PRUNED AND BINARIZED PHRASE TABLES IN KILOBYTES FOR MULTILINGUAL AND MONOLINGUAL (TOTAL SUM OF MULTIPLE SMT'S) TRANSLATION MODELS TRAINED ON 250K SENTENCE PAIRS PER LANGUAGE (FIRST EXPERIMENT)

Pair	Moses Tokenizer	SentencePiece	Moses Tokenizer Lowercased	SentencePiece Lowercased
sum	**192680**	248236	194424	249876
multi	**170848**	214252	171940	215528

TABLE V. SIZES OF LEXICAL REORDERING TABLES IN KILOBYTES FOR MULTILINGUAL AND MONOLINGUAL (TOTAL SUM OF MULTIPLE SMT'S) TRANSLATION MODELS TRAINED ON 250K SENTENCE PAIRS PER LANGUAGE (FIRST EXPERIMENT)

Pair	Moses Tokenizer	SentencePiece	Moses Tokenizer Lowercased	SentencePiece Lowercased
sum	973828	968288	943164	**934316**
multi	987444	958132	950092	**926396**

TABLE VI. SACREBLEU SCORES FOR THE SECOND EXPERIMENT

Pair	Moses Tokenizer Multi	SentencePiece Multi	Moses Tokenizer Mono	SentencePiece Mono	Dedicated Word Splitter
es-en	**16.93**	16.45	16.25	15.88	—
it-en	**19.16**	18.45	18.24	17.36	—
ru-en	**16.67**	16.23	15.89	15.17	—
uk-en	**19.27**	18.89	18.15	17.62	—
ar-en	**13.28**	12.94	12.45	12.31	—
hi-en	**12.85**	12.11	11.88	11.11	—
zh-en	0.69	**4.39**	0.71	4.17	4.15
ja-en	0.4	**6.55**	0.33	5.99	6.4
average	12.41	**13.25**	11.7	12.45	—

TABLE VII. CHRF++ SCORES FOR THE SECOND EXPERIMENT

Pair	Moses Tokenizer Multi	SentencePiece Multi	Moses Tokenizer Mono	SentencePiece Mono	Dedicated Word Splitter
es-en	47.33	**47.42**	46.81	46.93	—
it-en	49.73	**49.77**	49.14	49.06	—
ru-en	40.60	**45.73**	40.19	44.70	—
uk-en	43.54	**48.14**	42.96	47.33	—
ar-en	36.68	**41.53**	36.21	40.79	—
hi-en	41.43	**43.54**	40.80	42.69	—
zh-en	4.20	**27.45**	4.22	26.06	21.83
ja-en	3.18	**36.17**	3.23	36.14	35.75
average	33.34	**42.47**	32.95	41.71	—

TABLE VIII. METEOR SCORES FOR THE SECOND EXPERIMENT

Pair	Moses Tokenizer Multi	SentencePiece Multi	Moses Tokenizer Mono	SentencePiece Mono	Dedicated Word Splitter
es-en	**43.02**	42.92	42.15	42.49	—
it-en	**45.76**	45.47	44.95	44.51	—
ru-en	38.21	**41.11**	37.77	40.14	—
uk-en	41.62	**44.12**	41.17	43.17	—
ar-en	34.03	**37.75**	33.71	36.81	—
hi-en	37.16	**37.51**	36.70	36.55	—
zh-en	1.21	**16.47**	1.30	14.77	16.31
ja-en	0.75	25.59	0.84	26.03	**26.52**
average	30.22	**36.37**	29.82	35.56	—

TABLE IX. SIZES OF PRUNED AND BINARIZED PHRASE TABLES IN KILOBYTES FOR MULTILINGUAL AND MONOLINGUAL (TOTAL SUM OF MULTIPLE SMT'S) TRANSLATION MODELS TRAINED ON 250K SENTENCE PAIRS PER LANGUAGE (SECOND EXPERIMENT)

Pair	Moses Tokenizer	SentencePiece
sum	**113432**	148612
multi	**93140**	149636

TABLE X. SIZES OF LEXICAL REORDERING TABLES IN KILOBYTES FOR MULTILINGUAL AND MONOLINGUAL (TOTAL SUM OF MULTIPLE SMT'S) TRANSLATION MODELS TRAINED ON 250K SENTENCE PAIRS PER LANGUAGE (SECOND EXPERIMENT)

Pair	Moses Tokenizer	SentencePiece
sum	**589136**	738088
multi	**610668**	727688

VI. CONCLUSION

In this paper, we have studied how the multilingual setting affects translations quality of statistical machine translation systems. Multiple experiments were conducted on different configuration of Moses SMT system, covering multilingual and monolingual versions, different tokenizers, normal and lower cases and different training corpora sizes.

We found out that training SMT system in multilingual manner gives increase in translation quality and the gains could be attributed to simply training on bigger corpora. For most languages the default Moses tokenizer achieves higher BLEU scores on average than SentencePiece tokenizer. However, SentencePiece tokenizer gets higher scores for Turkic languages, which could be explained by agglutinative nature of this language family. Moses tokenizer has difficulties with tokenization of Japanese and Chinese languages, whereas SentencePiece tokenizer can be applied to any language. Dedi-

cated Japanese and Chinese word splitters show comparable to SentencePiece scores on all metrics. The SentencePiece tokenizer shows significantly better METEOR and chrF++ scores on average. Lowercasing achieves small increase in all metrics. The sizes of phrase and lexical reordering tables produced by training with SentencePiece tokenization are bigger than those produced by training with the default Moses tokenizer. Multilingual SMT with the default Moses tokenizer produces smaller phrase-table, comparing to monolingual versions.

Installing a multilingual SMT system makes maintaining and upgrading production environment harder, as phrase and lexical tables grow significantly and improving quality on one language direction requires retraining the whole system. However, building a multilingual SMT for languages from the same family could be considered as a trade-off between increasing translation quality and reducing deployment costs. As a future work, we would like to examine this scenario.

REFERENCES

[1] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, Eds., vol. 27. Curran Associates, Inc., 2014. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2014/file/a14ac55a4f27472c5d894ec1c3c743d2-Paper.pdf

[2] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2016.

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[4] N. Team, M. R. Costa-jussà, J. Cross, O. Çelebi, M. Elbayad, K. Heafield, K. Heffernan, E. Kalbassi, J. Lam, D. Licht, J. Maillard, A. Sun, S. Wang, G. Wenzek, A. Youngblood, B. Akula, L. Barrault, G. M. Gonzalez, P. Hansanti, J. Hoffman, S. Jarrett, K. R. Sadagopan, D. Rowe, S. Spruit, C. Tran, P. Andrews, N. F. Ayan, S. Bhosale, S. Edunov, A. Fan, C. Gao, V. Goswami, F. Guzmán, P. Koehn, A. Mourachko, C. Ropers, S. Saleem, H. Schwenk, and J. Wang, "No language left behind: Scaling human-centered machine translation," 2022.

[5] P. Koehn and R. Knowles, "Six challenges for neural machine translation," in *Proceedings of the First Workshop on Neural Machine Translation*, T. Luong, A. Birch, G. Neubig, and A. Finch, Eds. Vancouver: Association for Computational Linguistics, Aug. 2017, pp. 28–39. [Online]. Available: https://aclanthology.org/W17-3204

[6] M. Junczys-Dowmunt, T. Dwojak, and H. Hoang, "Is neural machine translation ready for deployment? a case study on 30 translation directions," 2016.

[7] A. Mohammadshahi, V. Nikoulina, A. Berard, C. Brun, J. Henderson, and L. Besacier, "Small-100: Introducing shallow multilingual machine translation model for low-resource languages," 2022.

[8] O. Bakhteev, A. Ogaltsov, A. Khazov, K. Safin, and R. Kuznetsova, "Crosslang: the system of cross-lingual plagiarism detection," in *Workshop on Document Intelligence at NeurIPS 2019*, 2019. [Online]. Available: https://openreview.net/forum?id=BkxiG6qqIr

[9] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, "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*, S. Ananiadou, Ed. Prague, Czech Republic: Association for Computational Linguistics, Jun. 2007, pp. 177–180. [Online]. Available: https://aclanthology.org/P07-2045

[10] M. Utiyama and H. Isahara, "A comparison of pivot methods for phrase-based statistical machine translation," in *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, C. Sidner, T. Schultz, M. Stone, and C. Zhai, Eds. Rochester, New York: Association for Computational Linguistics, Apr. 2007, pp. 484–491. [Online]. Available: https://aclanthology.org/N07-1061

[11] J. Schroeder, T. Cohn, and P. Koehn, "Word lattices for multi-source translation," in *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, A. Lascarides, C. Gardent, and J. Nivre, Eds. Athens, Greece: Association for Computational Linguistics, Mar. 2009, pp. 719–727. [Online]. Available: https://aclanthology.org/E09-1082

[12] T. Banerjee, A. Kunchukuttan, and P. Bhattacharya, "Multilingual Indian language translation system at WAT 2018: Many-to-one phrase-based SMT," in *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation: 5th Workshop on Asian Translation: 5th Workshop on Asian Translation*, S. Politzer-Ahles, Y.-Y. Hsu, C.-R. Huang, and Y. Yao, Eds. Hong Kong: Association for Computational Linguistics, 1–3 Dec. 2018. [Online]. Available: https://aclanthology.org/Y18-3013

[13] P. Nakov and H. T. Ng, "Improved statistical machine translation for resource-poor languages using related resource-rich languages," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, P. Koehn and R. Mihalcea, Eds. Singapore: Association for Computational Linguistics, Aug. 2009, pp. 1358–1367. [Online]. Available: https://aclanthology.org/D09-1141

[14] A. Bastawisy and M. Elmahdy, "Multi-lingual phrase-based statistical machine translation for Arabic-English," in *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, R. Mitkov and G. Angelova, Eds. Varna, Bulgaria: INCOMA Ltd., Sep. 2017, pp. 86–89. [Online]. Available: https://doi.org/10.26615/978-954-452-049-6_013

[15] M. Domingo, M. Garcıa-Martınez, A. Helle, F. Casacuberta, and M. Herranz, "How much does tokenization affect neural machine translation?" 2019.

[16] "Sentencepiece experiments," https://github.com/google/sentencepiece/blob/master/doc/experiments.md.

[17] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, K. Erk and N. A. Smith, Eds. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 1715–1725. [Online]. Available: https://aclanthology.org/P16-1162

[18] T. Kudo and J. Richardson, "SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, E. Blanco and W. Lu, Eds. Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 66–71. [Online]. Available: https://aclanthology.org/D18-2012

[19] H. Zhao, M. Utiyama, E. Sumita, and B.-L. Lu, "An empirical study on word segmentation for chinese machine translation," in *Computational Linguistics and Intelligent Text Processing*, A. Gelbukh, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 248–263.

[20] M. Oudah, A. Almahairi, and N. Habash, "The impact of preprocessing on Arabic-English statistical and neural machine translation," in *Proceedings of Machine Translation Summit XVII: Research Track*, M. Forcada, A. Way, B. Haddow, and R. Sennrich, Eds. Dublin, Ireland: European Association for Machine Translation, Aug. 2019, pp. 214–221. [Online]. Available: https://aclanthology.org/W19-6621

[21] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, P. Isabelle, E. Charniak, and D. Lin, Eds. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, Jul. 2002, pp. 311–318. [Online]. Available: https://aclanthology.org/P02-1040

[22] M. Post, "A call for clarity in reporting BLEU scores," in *Proceedings of the Third Conference on Machine Translation: Research Papers*. Belgium, Brussels: Association for Computational Linguistics, Oct. 2018, pp. 186–191. [Online]. Available: https://www.aclweb.org/anthology/W18-6319

[23] M. Popović, "chrF: character n-gram F-score for automatic MT evaluation," in *Proceedings of the Tenth Workshop on Statistical*

Machine Translation, O. Bojar, R. Chatterjee, C. Federmann, B. Haddow, C. Hokamp, M. Huck, V. Logacheva, and P. Pecina, Eds. Lisbon, Portugal: Association for Computational Linguistics, Sep. 2015, pp. 392–395. [Online]. Available: https://aclanthology.org/W15-3049

[24] ——, "chrF++: words helping character n-grams," in *Proceedings of the Second Conference on Machine Translation*, O. Bojar, C. Buck, R. Chatterjee, C. Federmann, Y. Graham, B. Haddow, M. Huck, A. J. Yepes, P. Koehn, and J. Kreutzer, Eds. Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 612–618. [Online]. Available: https://aclanthology.org/W17-4770

[25] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. Ann Arbor, Michigan: Association for Computational Linguistics, Jun. 2005, pp. 65–72. [Online]. Available: https://www.aclweb.org/anthology/W05-0909

[26] "sacrebleu," https://github.com/mjpost/sacrebleu.

[27] "Evaluate metric," https://huggingface.co/evaluate-metric.

[28] F. Guzmán, P.-J. Chen, M. Ott, J. Pino, G. Lample, P. Koehn, V. Chaudhary, and M. Ranzato, "The flores evaluation datasets for low-resource machine translation: Nepali-english and sinhala-english," 2019.

[29] N. Goyal, C. Gao, V. Chaudhary, P.-J. Chen, G. Wenzek, D. Ju, S. Krishnan, M. Ranzato, F. Guzman, and A. Fan, "The flores-101 evaluation benchmark for low-resource and multilingual machine translation," 2021.

[30] K. Heffernan, O. Çelebi, and H. Schwenk, "Bitext mining using distilled sentence representations for low-resource languages," 2022.

[31] J. Tiedemann, "Parallel data, tools and interfaces in opus," in *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, N. C. C. Chair), K. Choukri, T. Declerck, M. U. Dogan, B. Maegaard, J. Mariani, J. Odijk, and S. Piperidis, Eds. Istanbul, Turkey: European Language Resources Association (ELRA), may 2012.

[32] "Opus," https://opus.nlpl.eu/.

[33] "Sacremoses," https://github.com/hplt-project/sacremoses.

[34] "No language left behind," https://github.com/facebookresearch/fairseq/tree/nllb.

[35] M. Aulamo, N. Bogoychev, S. Ji, G. Nail, G. Ramírez-Sánchez, J. Tiedemann, J. van der Linde, and J. Zaragoza, "HPLT: High performance language technologies," in *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, M. Nurminen, J. Brenner, M. Koponen, S. Latomaa, M. Mikhailov, F. Schierl, T. Ranasinghe, E. Vanmassenhove, S. A. Vidal, N. Aranberri, M. Nunziatini, C. P. Escartín, M. Forcada, M. Popovic, C. Scarton, and H. Moniz, Eds. Tampere, Finland: European Association for Machine Translation, Jun. 2023, pp. 517–518. [Online]. Available: https://aclanthology.org/2023.eamt-1.61

[36] M. Bañón, P. Chen, B. Haddow, K. Heafield, H. Hoang, M. Esplà-Gomis, M. L. Forcada, A. Kamran, F. Kirefu, P. Koehn, S. Ortiz Rojas, L. Pla Sempere, G. Ramírez-Sánchez, E. Sarrías, M. Strelec, B. Thompson, W. Waites, D. Wiggins, and J. Zaragoza, "ParaCrawl: Web-scale acquisition of parallel corpora," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds. Online: Association for Computational Linguistics, Jul. 2020, pp. 4555–4567. [Online]. Available: https://aclanthology.org/2020.acl-main.417

[37] M. Morishita, K. Chousa, J. Suzuki, and M. Nagata, "JParaCrawl v3.0: A large-scale English-Japanese parallel corpus," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, and S. Piperidis, Eds. Marseille, France: European Language Resources Association, Jun. 2022, pp. 6704–6710. [Online]. Available: https://aclanthology.org/2022.lrec-1.721

[38] M. Bañón, M. Esplà-Gomis, M. L. Forcada, C. García-Romero, T. Kuzman, N. Ljubešić, R. van Noord, L. P. Sempere, G. Ramírez-Sánchez, P. Rupnik, V. Suchomel, A. Toral, T. van der Werff, and J. Zaragoza, "MaCoCu: Massive collection and curation of monolingual and bilingual data: focus on under-resourced languages," in *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*. Ghent, Belgium: European Association for Machine Translation, Jun. 2022, pp. 303–304. [Online]. Available: https://aclanthology.org/2022.eamt-1.41

[39] "jieba," https://github.com/fxsjy/jieba.

[40] "Mecab: Yet another part-of-speech and morphological analyzer," https://taku910.github.io/mecab/.

APPENDIX

A. First experiment evaluation results

TABLE XI
SACREBLEU SCORES FOR THE MOSES SMT TRAINED ON 250K LOWERCASED SUBSET

Pair	Moses Tokenizer Multi	SentencePiece Multi	Moses Tokenizer Mono	SentencePiece Mono
it-en	**19.87**	18.74	19.41	18.71
fr-en	**28.00**	26.51	26.59	25.23
es-en	**17.14**	16.42	16.76	15.94
pt-en	**31.33**	29.93	30.27	29.26
de-en	**24.35**	23.70	24.11	23.59
nl-en	20.45	20.42	20.53	**20.55**
ru-en	**19.16**	18.29	18.65	18.02
be-en	**11.06**	10.69	11.04	10.48
cs-en	22.84	**23.20**	22.58	22.47
pl-en	**16.25**	15.52	15.80	15.46
ky-en	5.74	**6.59**	4.67	5.24
kk-en	9.40	**10.02**	7.89	9.00
uz-en	9.70	**10.05**	9.52	9.79
tr-en	11.78	**12.33**	11.93	12.24
hi-en	**16.78**	16.05	15.76	14.91
average	**17.59**	17.23	17.03	16.73

TABLE XII
CHRF++ SCORES FOR THE MOSES SMT TRAINED ON 250K LOWERCASED SUBSET

Pair	Moses Tokenizer Multi	SentencePiece Multi	Moses Tokenizer Mono	SentencePiece Mono
it-en	**51.09**	50.45	51.06	50.74
fr-en	**56.99**	56.19	56.38	55.79
es-en	48.22	48.02	**48.23**	47.61
pt-en	**59.56**	58.96	59.21	58.61
de-en	51.65	**53.61**	51.93	**53.61**
nl-en	48.81	50.10	49.43	**50.20**
ru-en	44.72	**48.25**	46.29	**48.25**
be-en	38.83	41.70	40.36	**41.80**
cs-en	50.74	52.29	51.61	**52.43**
pl-en	44.41	45.38	45.39	**45.98**
ky-en	26.50	35.51	24.50	**31.91**
kk-en	32.77	**40.37**	30.83	38.21
uz-en	38.28	**41.44**	38.05	41.01
tr-en	41.10	44.65	41.28	**44.66**
hi-en	45.53	**47.86**	45.69	47.33
average	45.28	**47.65**	45.35	47.21

TABLE XIII
METEOR SCORES FOR THE MOSES SMT TRAINED ON 250K LOWERCASED SUBSET

Pair	Moses Tokenizer Multi	SentencePiece Multi	Moses Tokenizer Mono	SentencePiece Mono
it-en	54.52	53.64	**54.83**	54.30
fr-en	63.07	62.14	**63.11**	62.11
es-en	51.32	51.02	**52.03**	51.15
pt-en	65.45	64.92	**65.76**	64.88
de-en	55.51	57.74	56.04	**57.80**
nl-en	51.06	52.58	52.35	**53.28**
ru-en	49.87	**51.63**	51.03	51.40
be-en	40.65	42.12	42.15	**42.26**
cs-en	53.84	55.69	55.45	**56.47**
pl-en	46.23	46.80	47.44	**47.70**
ky-en	26.76	**33.44**	25.09	29.97
kk-en	34.77	**40.80**	32.94	38.60
uz-en	36.81	**41.43**	37.26	41.27
tr-en	40.94	45.41	41.78	**46.07**
hi-en	50.44	**50.86**	50.42	50.80
average	48.08	**50.01**	48.51	49.87

TABLE XIV
SACREBLEU SCORES FOR THE MOSES SMT TRAINED ON 10K NORMALCASED SUBSET

Pair	Moses Tokenizer Multi	SentencePiece Multi	Moses Tokenizer Mono	SentencePiece Mono
it-en	10.41	9.83	**10.59**	9.82
fr-en	**15.75**	14.72	15.44	14.54
es-en	**8.89**	8.58	8.62	8.40
pt-en	**17.05**	16.59	16.67	16.15
de-en	12.44	11.34	**12.48**	11.61
nl-en	11.74	10.97	**12.00**	11.70
ru-en	6.80	6.27	**7.25**	6.86
be-en	**4.24**	3.54	4.10	3.43
cs-en	8.89	8.99	**9.17**	8.54
pl-en	6.40	5.96	**6.75**	5.69
ky-en	**1.50**	1.47	1.23	0.84
kk-en	2.73	**2.87**	2.33	2.35
uz-en	3.35	**3.57**	3.02	2.80
tr-en	4.08	4.48	4.25	**4.76**
hi-en	**6.57**	5.92	5.93	5.31
average	**8.06**	7.67	7.99	7.52

TABLE XV
CHRF++ SCORES FOR THE MOSES SMT TRAINED ON 10K NORMALCASED SUBSET

Pair	Moses Tokenizer Multi	SentencePiece Multi	Moses Tokenizer Mono	SentencePiece Mono
it-en	39.68	39.97	39.95	**39.94**
fr-en	**45.06**	44.97	45.04	45.05
es-en	37.72	**38.41**	37.48	37.95
pt-en	45.23	**46.56**	45.06	45.93
de-en	38.38	40.19	38.82	**40.36**
nl-en	39.37	40.09	39.75	**40.87**
ru-en	22.13	**31.15**	22.77	29.82
be-en	19.54	**28.68**	19.98	27.50
cs-en	32.27	**36.50**	32.90	**36.50**
pl-en	28.48	**32.54**	29.06	32.36
ky-en	9.67	**21.53**	8.30	18.39
kk-en	12.72	**24.91**	12.28	22.29
uz-en	24.64	**29.26**	24.68	28.93
tr-en	25.42	30.72	26.05	**31.91**
hi-en	28.56	**31.03**	29.17	30.95
average	29.92	**34.43**	30.09	33.92

TABLE XVI
METEOR SCORES FOR THE MOSES SMT TRAINED ON 10K NORMALCASED SUBSET

Pair	Moses Tokenizer Multi	SentencePiece Multi	Moses Tokenizer Mono	SentencePiece Mono
it-en	40.09	40.12	**40.74**	40.10
fr-en	49.41	48.27	**49.44**	48.77
es-en	37.62	38.15	37.63	**38.24**
pt-en	48.65	**50.12**	48.26	49.13
de-en	41.03	42.02	41.39	**42.11**
nl-en	40.01	40.82	40.60	**41.65**
ru-en	28.16	**32.85**	28.96	31.73
be-en	23.32	**28.12**	23.72	26.99
cs-en	32.00	35.92	33.14	**36.04**
pl-en	27.02	**30.28**	27.76	30.27
ky-en	12.32	**19.04**	11.60	16.72
kk-en	16.70	**24.38**	16.37	22.17
uz-en	20.21	**26.18**	20.10	26.00
tr-en	22.44	27.74	23.42	**30.20**
hi-en	34.71	34.72	**34.99**	34.57
average	31.58	**34.58**	31.87	34.31

TABLE XVII
SACREBLEU SCORES FOR THE MOSES SMT TRAINED ON 10K LOWERCASED SUBSET

Pair	Moses Tokenizer Multi	SentencePiece Multi	Moses Tokenizer Mono	SentencePiece Mono
it-en	11.33	**11.92**	11.38	11.28
fr-en	16.39	**16.56**	16.21	16.43
es-en	9.71	**10.21**	9.23	9.46
pt-en	18.03	**18.85**	17.59	17.82
de-en	13.31	**14.20**	13.51	13.79
nl-en	12.94	12.81	13.10	**13.28**
ru-en	7.37	**8.65**	7.95	8.20
be-en	4.71	**4.79**	4.65	4.40
cs-en	9.85	**11.11**	10.21	10.27
pl-en	6.93	**7.70**	7.41	6.96
ky-en	1.61	**1.97**	1.35	1.17
kk-en	2.93	**3.80**	2.64	2.73
uz-en	3.62	**4.58**	3.72	3.66
tr-en	4.68	**5.79**	4.76	5.53
hi-en	**7.41**	6.77	6.90	6.26
average	8.72	**9.31**	8.71	8.75

TABLE XVIII
CHRF++ SCORES FOR THE MOSES SMT TRAINED ON 10K LOWERCASED SUBSET

Pair	Moses Tokenizer Multi	SentencePiece Multi	Moses Tokenizer Mono	SentencePiece Mono
it-en	41.05	**42.76**	41.29	41.98
fr-en	46.23	**47.40**	46.23	47.22
es-en	39.06	**40.80**	38.71	39.76
pt-en	46.65	**49.11**	46.41	48.00
de-en	40.55	**44.00**	40.89	43.94
nl-en	40.80	42.51	41.20	**43.01**
ru-en	23.79	**36.41**	24.64	35.25
be-en	21.24	**32.48**	21.64	31.65
cs-en	34.19	**40.30**	34.88	39.71
pl-en	30.01	**35.94**	30.68	35.33
ky-en	10.61	**24.52**	9.22	20.49
kk-en	14.23	**27.97**	13.68	25.42
uz-en	25.85	**31.13**	26.18	30.30
tr-en	27.12	34.02	27.69	**34.52**
hi-en	30.00	**34.95**	30.41	34.56
average	31.43	**37.62**	31.58	36.74

TABLE XIX
METEOR SCORES FOR THE MOSES SMT TRAINED ON 10K LOWERCASED SUBSET

Pair	Moses Tokenizer Multi	SentencePiece Multi	Moses Tokenizer Mono	SentencePiece Mono
it-en	40.88	**43.34**	41.46	42.53
fr-en	49.86	**50.74**	49.92	50.66
es-en	38.10	**41.14**	38.23	40.01
pt-en	49.59	**53.35**	49.26	51.61
de-en	41.27	**45.36**	41.58	44.74
nl-en	40.61	43.02	41.22	**43.66**
ru-en	29.49	**37.93**	30.41	36.55
be-en	24.29	**31.03**	24.75	30.07
cs-en	33.62	**40.87**	34.84	40.23
pl-en	28.20	**34.72**	29.03	34.04
ky-en	13.01	**26.47**	12.03	18.11
kk-en	17.63	**26.47**	17.04	23.99
uz-en	20.98	**28.59**	21.28	27.78
tr-en	23.98	32.63	24.85	**33.38**
hi-en	35.22	**36.52**	35.42	36.25
average	32.45	**38.15**	32.75	36.91

TABLE XX
SIZES OF PRUNED AND BINARIZED PHRASE TABLES IN KILOBYTES FOR MULTILINGUAL AND MONOLINGUAL (TOTAL SUM OF MULTIPLE SMT'S) TRANSLATION MODELS TRAINED ON 250K SENTENCE PAIRS PER LANGUAGE (FIRST EXPERIMENT)

Pair	Moses Tokenizer	SentencePiece	Moses Tokenizer Lowercased	SentencePiece Lowercased
it-en	**14904**	18304	14944	18380
fr-en	**15440**	19248	15484	19388
es-en	**14068**	16536	14088	16692
pt-en	**15012**	17940	15136	18120
de-en	**11520**	15492	11660	15880
nl-en	**13548**	17744	13592	17972
ru-en	**14340**	18316	14452	18388
be-en	**18156**	21452	18464	21552
cs-en	**14464**	18840	14648	18904
pl-en	**13416**	17932	13516	17896
ky-en	**8888**	13148	8968	12956
kk-en	**6688**	10148	6800	10092
uz-en	11068	15056	**11024**	15048
tr-en	**7580**	12292	7676	12412
hi-en	**13588**	15788	13972	16196
average	**12845**	16549	12962	16658
sum	**192680**	248236	194424	249876
multi	**170848**	214252	171940	215528

TABLE XXI
SIZES OF LEXICAL REORDERING TABLES IN KILOBYTES FOR MULTILINGUAL AND MONOLINGUAL (TOTAL SUM OF MULTIPLE SMT'S) TRANSLATION MODELS TRAINED ON 250K SENTENCE PAIRS PER LANGUAGE (FIRST EXPERIMENT)

Pair	Moses Tokenizer	SentencePiece	Moses Tokenizer Lowercased	SentencePiece Lowercased
it-en	76424	74116	74796	**72244**
fr-en	73056	69996	71532	**68708**
es-en	69548	67712	67800	**66228**
pt-en	77084	73456	75884	**72144**
de-en	62252	62944	**60592**	60760
nl-en	74004	74716	**72384**	73268
ru-en	81152	70544	79436	**68652**
be-en	75696	64368	73776	**61584**
cs-en	83268	79472	81272	**76240**
pl-en	78884	73508	76468	**69968**
ky-en	37836	50516	**35372**	47456
kk-en	41852	53212	**39084**	49748
uz-en	50140	53652	**46132**	50336
tr-en	49376	54756	**46316**	51664
hi-en	43256	45320	**42320**	45316
average	64922	64553	62877	**62288**
sum	973828	968288	943164	**934316**
multi	987444	958132	950092	**926396**

B. Second experiment evaluation results

TABLE XXII
SIZES OF PRUNED AND BINARIZED PHRASE TABLES IN KILOBYTES FOR MULTILINGUAL AND MONOLINGUAL (TOTAL SUM OF MULTIPLE SMT'S) TRANSLATION MODELS TRAINED ON 250K SENTENCE PAIRS PER LANGUAGE (SECOND EXPERIMENT)

Pair	Moses Tokenizer	SentencePiece	Dedicated WordSplitter
es-en	**14516**	19280	—
it-en	**15144**	21036	—
ru-en	**16072**	23032	—
uk-en	**14328**	21420	—
ar-en	**16040**	23452	—
hi-en	**12424**	17896	—
zh-en	22680	**6764**	14052
ja-en	**2228**	15732	15576
average	**14179**	18577	—
sum	**113432**	148612	—
multi	**93140**	149636	—

TABLE XXIII
SIZES OF LEXICAL REORDERING TABLES IN KILOBYTES FOR MULTILINGUAL AND MONOLINGUAL (TOTAL SUM OF MULTIPLE SMT'S) TRANSLATION MODELS TRAINED ON 250K SENTENCE PAIRS PER LANGUAGE (SECOND EXPERIMENT)

Pair	Moses Tokenizer	SentencePiece	Dedicated WordSplitter
es-en	**90836**	105660	—
it-en	**95960**	108848	—
ru-en	**98864**	107532	—
uk-en	**99228**	103516	—
ar-en	**78008**	95912	—
hi-en	**63144**	75248	—
zh-en	**40020**	81664	66484
ja-en	**23076**	59708	54888
average	**73642**	92261	—
sum	**589136**	738088	—
multi	**610668**	727688	—