Evaluating Multiway Multilingual NMT in the Turkic Languages

Jamshidbek Mirzakhalov[a,b], **Anoop Babu**[a,b], **Aigiz Kunafin**[a], **Ahsan Wahab**[a],
Behzod Moydinboyev[a,b], **Sardana Ivanova**[a,c], **Mokhiyakhon Uzokova**[a,d],
Shaxnoza Pulatova[a,e], **Duygu Ataman**[a,f], **Julia Kreutzer**[a,g],
Francis Tyers[a,h], **Orhan Firat**[a,g], **John Licato**[a,b], **Sriram Chellappan**[a,b]

[a]Turkic Interlingua, [b]University of South Florida,
[c]University of Helsinki, [d]Tashkent State University of Uzbek Language and Literature,
[e]Namangan State University, [f]NYU, [g]Google Research, [h]Indiana University

Abstract

Despite the increasing number of large and comprehensive machine translation (MT) systems, evaluation of these methods in various languages has been restrained by the lack of high-quality parallel corpora as well as engagement with the people that speak these languages. In this study, we present an evaluation of state-of-the-art approaches to training and evaluating MT systems in 22 languages from the Turkic language family, most of which being extremely under-explored (Joshi et al., 2019). First, we adopt the TIL Corpus (Mirzakhalov et al., 2021) with a few key improvements to the training and the evaluation sets. Then, we train 26 bilingual baselines as well as a multi-way neural MT (MNMT) model using the corpus and perform an extensive analysis using automatic metrics as well as human evaluations. We find that the MNMT model outperforms almost all bilingual baselines in the out-of-domain test sets and finetuning the model on a downstream task of a single pair also results in a huge performance boost in both low- and high-resource scenarios. Our attentive analysis of evaluation criteria for MT models in Turkic languages also points to the necessity for further research in this direction. We release the corpus splits, test sets as well as models to the public[1].

1 Introduction

The last few years have seen encouraging advances in low-resource MT development with the increasing availability of public multilingual corpora (Agić and Vulić, 2019; Ortiz Suárez et al., 2019; Schwenk et al., 2019; El-Kishky et al., 2020; Tiedemann, 2020; Goyal et al., 2021; ∀ et al., 2020) and more inclusive multilingual MT models (Arivazhagan et al., 2019; Tiedemann and Thottingal, 2020; Fan et al., 2020). In this study, we take the Turkic language family into focus, which has not been studied at large in MT research (detailed review in Section 2). Most recently, in a wide evaluation of translation between hundreds of languages with a multilingual model (M2M-124) trained on large web-mined parallel data, translation into, from, and between Turkic languages was shown to be very challenging compared to other language families (Goyal et al., 2021). With the promise of strong transfer capabilities of multilingual models especially for related languages, we hope that the inclusion of a wider set of Turkic languages into a joint model can unlock automatic translation even for the very low-resourced Turkic languages where no prior translation models exist (Koehn, 2005; Choudhary and Jha, 2011; Post et al., 2012; Nomoto et al., 2018; Esplà-Gomis et al., 2019; ∀ et al., 2020).

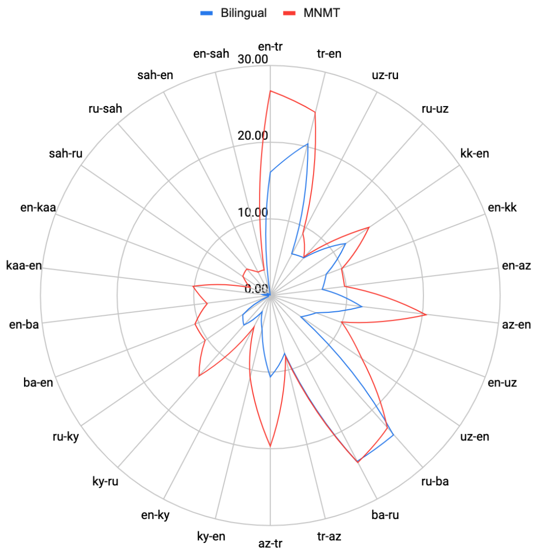

Figure 1: Performance comparison between bilingual baselines and the MNMT model on X-WMT test set.

To this aim, we adopt the TIL Corpus (Mirzakhalov et al., 2021) compiled by the Turkic Inter-

[1] https://github.com/turkic-interlingua/til-mt

Proceedings of the Sixth Conference on Machine Translation (WMT), pages 518–530
November 10–11, 2021. ©2021 Association for Computational Linguistics

Name	Codes	Speakers	Data	MT?
English	en, eng	400.0M	38.6M	✓
Russian	ru, rus	258.0M	23.3M	✓
Turkish	tr, tur	85.0M	52.6M	✓
Kazakh	kk, kaz	13.2M	5.3M	✓
Uzbek	uz, uzb	27.0M	2.9M	✓
Azerbaijani	az, aze	23.0M	2.2M	✓
Tatar	tt, tat	5.2M	1.8M	✓
Kyrgyz	ky, kir	4.3M	1.7M	✓
Chuvash	cv, chv	1.0M	1.5M	✓
Turkmen	tk, tuk	6.7M	910.4K	✓
Bashkir	ba, bak	1.4M	880.5K	✓
Uyghur	ug, uig	10.0M	334.8K	✓
Karakalpak	kaa	583.0K	253.8K	✗
Khakas	kjh	43.0K	219.0K	✗
Altai	alt	56.0K	192.6K	✗
Crimean Tatar	crh	540.0K	185.3K	✗
Karachay-Balkar	krc	310.0K	162.8K	✗
Gagauz	gag	148.0K	157.4K	✗
Sakha	sah	450.0K	157.1K	✓
Kumyk	kum	450.0K	156.8K	✗
Tuvinian	tyv	280.0K	100.3K	✗
Shor	cjs	3.0K	2.3K	✗
Salar	slr	70.0K	766	✗
Urum	uum	190.0K	491	✗

Table 1: (The table indicates the language codes used for the Turkic languages along with the number of L1 speakers, amount of available data (in sentences) in our corpus. The column MT? indicates if there are currently available online machine translation systems for the language. K: thousand, M: million.)

lingua[2] community (Mirzakhalov, 2021) including X-WMT test sets with a few key improvements (Section 3). We train a multi-way NMT (MNMT) model on the entire parallel corpus, which constitutes the first large-scale multilingual translation model specifically for Turkic languages (Section 4). We perform an extensive analysis of the strengths and weaknesses of this model, comparing it to the bilingual baselines and evaluating it under a domain shift. We find that the MNMT model outperforms almost all bilingual baselines in the out-of-domain tests while it performs comparably or underperforms in the in-domain tests. We further analyze its capacity for transfer learning by fine-tuning the model on several language pairs all of which experience gains, both in- and out-of-domain scenarios. In addition, we complement the automatic evaluation with a human evaluation study for multiple languages (Section 5), gaining insights into types of common mistakes that the model makes and the suitability of different automatic metrics for Turkic languages. We plan on releasing the improved corpus, evaluation sets, and all the models to the public.

This work will not only enrich the landscape of languages currently considered in MT research and spur future research on NLP for Turkic languages but will hopefully also inspire the building of new translation engines and derived technologies for populations with millions of native speakers (Table 1).

2 Related Work

This section discusses the previous work on MT of these languages including the available corpora and languages resources. The 19 Turkic languages covered in the study are: Altai, Azerbaijani, Bashkir, Crimean Tatar, Chuvash, Gagauz, Karachay-Balkar, Karakalpak, Khakas, Kazakh, Kumyk, Kyrgyz, Sakha, Turkmen, Turkish, Tatar, Tuvan, Uyghur, and Uzbek. There are several other widely spoken languages that are left out from our study such as Shor, Salar, Urum, Nogai, Khorasani Turkic, Qashqai, and Khalaj, due to the lack (or very limited amount) of any available parallel corpora. Future work will focus on extending the corpus to these languages as well.

2.1 MT of Turkic Languages

The need for more comprehensive and diverse multilingual parallel corpora has sped up the creation of such large-scale resources for many language families and linguistic regions (Koehn, 2005; Choudhary and Jha, 2011; Post et al., 2012; Nomoto et al., 2018; Esplà-Gomis et al., 2019; ∀ et al., 2020). Tiedemann (2020) released a large-scale corpus for over 500 languages covering thousands of translation directions. The corpus currently includes 14 Turkic languages and provides bilingual baselines for all translation directions present in the corpus. However, most of the 14 Turkic languages contain a few hundred or a dozen samples. In addition, the varying and limited size of the test sets does not allow for the extensive analysis and comparisons between different model artifacts, linguistic features, and translation domains. More recently, Goyal et al. (2021) extended the previous Flores benchmark by providing human translated evaluation sets for 101 languages, among which 5 of them are from the Turkic family: Azerbaijani, Kazakh, Kyrgyz, Turkish, and Uzbek. Similarly, they train a large MNMT model and evaluate its performance

[2] https://turkicinterlingua.org/

using the benchmark.

A Russian-Turkic parallel corpus was curated for 6 different Turkic languages, and their bilingual baselines have been reported for both directions using different NMT-based approaches Khusainov et al. (2020). However, the dataset, test sets, and models are not released to the public which limits its use to serve as a comparable benchmark. Additionally, a rule-based MT framework for Turkic languages has been presented with 4 language pairs Alkım and Çebi (2019). Also, several rule-based MT systems have been built for Turkic languages which are publicly available through the Apertium[3] website Washington et al. (2019).

For individual languages in our corpus, there are several proposed MT systems and linguistic resources: Azerbaijani (Hamzaoglu, 1993; Fatullayev et al., 2008), Bashkir (Tyers et al., 2012), Crimean Tatar (Gökırmak et al., 2019; Altıntaş, 2001), Karakalpak (Kadirov, 2015), Kazakh (Assylbekov and Nurkas, 2014; Sundetova et al., 2015; Littell et al., 2019; Briakou and Carpuat, 2019; Tukeyev et al., 2019), Kyrgyz (Çetin and Ismailova), Sakha (Ivanova et al., 2019), Turkmen (Tantuğ and Adalı, 2018), Turkish (Turhan, 1997; El-Kahlout and Oflazer, 2006; Bisazza and Federico, 2009; Tantuğ et al., 2011; Ataman et al., 2017), Tatar (Salimzyanov et al., 2013; Khusainov et al., 2018; Valeev et al., 2019; Gökırmak et al., 2019), Tuvan (Killackey, 2013), Uyghur (Mahsut et al., 2004; Nimaiti and Izumi, 2012; Song and Dai, 2015; Wang et al., 2020), and Uzbek (Axmedova et al., 2019). Yet to our knowledge, there has not been a study that covers Turkic languages to such a large extent as ours, both in terms of multilingual parallel corpora and multiway NMT benchmarks across these languages.

3 TIL Corpus

As we adopt the TIL Corpus as the training data, we perform a few key modifications to better the quality of the datasets.

First, we notice that the alignments for the Bible[4] and TedTalks[5] datasets were not optimal as most ”sentences” were actually comprised of multiple sentences in order to preserve the quality of the alignment with target sequence. For example, in the case of TedTalks, the original speech utterance may have been 2-3 sentences in text but the translation of that speech may end up differing by 1 or even more sentences depending on the translator. Common practice in this situation, as seen through multiple corpora across OPUS[6], is to leave the entire utterance as is to preserve the quality of the alignment even if the number of sentences do not match. Instead, we drop the examples where the total number of sentences do not match and split (and realign) the cases where they do. This naturally increased the overall number of sentence alignments in both the Bible and TedTalks corpora for all language pairs.

Second, we perform a corpus-wide length and length-ration filtering where we drop sentence pairs that are single words as well as the entries where source and target ratio is over 2.

Third, we re-curate the in-domain evaluation sets following the improvements to the corpus. Details on the evaluation sets are described further in Section 3.1.

3.1 Curation of evaluation sets

The original TIL Corpus introduced three evaluation sets with different domains (Bible, TedTalks, and X-WMT). To simplify the analysis of the models, we re-curate the in-domain evaluation sets by randomly sampling from each corpora. X-WMT is used as the out-of-domain test set since it is from the news domain with substantial amount of new words/terms that most of the language pairs lack. The curation steps for the test sets are presented below.

3.1.1 In-domain Evaluation Sets

In-domain development and test sets are randomly sampled from each language pair and can serve as evaluation sets for both bilingual and multilingual models. The size of the development and test sets depends on the amount of training data available. More specifically, development and test sizes are 5k each if the train size is over 1 million parallel sentences, 2.5k if over 100k, 1k if over 10k, and 500 if over 2.5k. All test and development samples are removed from the training corpus for that language pair. Overall, this yields development and test sets for exactly 400 language pairs.

[3] https://www.apertium.org/

[4] https://bible.is/

[5] https://www.ted.com/participate/translate

[6] https://opus.nlpl.eu/

	en	ru	ba	tr	uz	ky	kk	az	sah	kaa
en	—									
ru	**1000**	—								
ba	1000	**1000**	—							
tr	**800**	800	800	—						
uz	**900**	900	900	600	—					
ky	500	**500**	500	400	500	—				
kk	700	700	700	500	**700**	500	—			
az	**600**	600	600	500	600	500	500	—		
sah	300	**300**	300	300	300	300	300	300	—	
kaa	300	300	300	**300**	300	300	300	300	300	—

Table 2: X-WMT test sets. Bolded entries indicate the original translation direction.

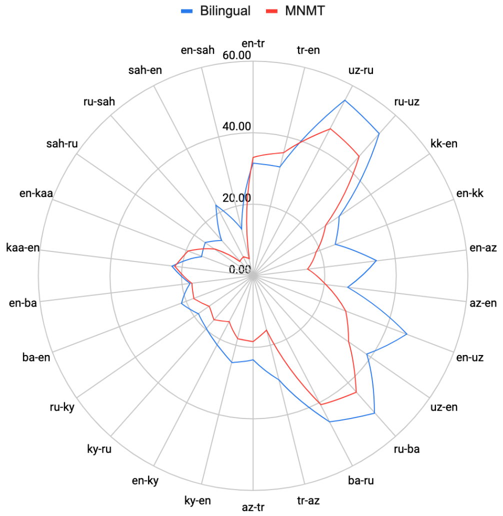

Figure 2: Performance comparison between bilingual baselines and the MNMT model on the in-domain test set.

3.1.2 X-WMT Test Set

X-WMT is a challenging and human-translated test set in the news domain based on the professionally translated test sets in English-Russian from the WMT 2020 Shared Task (Mathur et al., 2020). It was originally introduced in the TIL Corpus and we adopt the test sets as they are. Currently, the test set extends into 8 Turkic languages (Bashkir, Uzbek, Turkish, Kazakh, Kyrgyz, Azerbaijani, Karakalpak, and Sakha) paired with English and Russian. Table 2 highlights the currently available test set directions. Bolded entries in the table indicate the original direction of the translation.

4 Experimental Setup

4.1 Bilingual Experiments

To serve as initial baselines, we train 26 bilingual baselines using the corpus and report the performance on the in-domain test set as well as the X-WMT set (out-of-domain) as described in Section 3.1.2. The selection of the language pairs was constricted by the availability of both in-domain and out-of-domain test sets to enable more meaningful insights from the experiments.

4.1.1 Model details

All models are Transformers (*transformer-base*) (Vaswani et al., 2017b) and are trained using the JoeyNMT framework (Kreutzer et al., 2019). In the preprocessing stage, we use Sacremoses for tokenization and apply byte pair encoding (BPE) (Sennrich et al., 2015; Dong et al., 2015) with a joint vocabulary size of 4k and 32k. Models use 512-dimensional word embeddings and hidden layers and are trained with the Adam optimizer (Kingma and Ba, 2015). A learning rate of $3*10^{-4}$ is applied along with a dropout rate of 0.3. We use a batch size of 4096 BPE tokens with 8 accumulations to simulate training on 8 GPU machines. All models, except English-Turkish and Turkish-English, are trained on Google Colab's freely availably pre-emptible GPUs.

4.2 Multilingual Experiments

To examine the extent of transfer learning and generalization within our corpus, we train a multiway multilingual NMT model on the entire dataset covering almost 400 language directions. We then compare the performance of the model on the in-domain and out-of-domain test sets across a range of language pairs.

4.2.1 Data Preprocessing

Similar to the bilingual data preprocessing, the entire corpus has been tokenized using Sacremoses[7] and samples longer than 300 words have been filtered out. In addition, we perform cross-filtering of test and dev sets of all language pairs from the training corpus, as it is very necessary to do so in any MNMT model using a multiway corpus. Since the corpus is relatively unbalanced, we perform a temperature-based sampling with a value of 1.25. Although a higher temperature value between 2 and 3 would further balance our corpus, it would increase the dataset size by 8x with t=2 and 25x with t=3. This increase would limit our ability to train the model due to the restrained compute resources. Originally, the overall training set size is at around 133 million samples and this increases to 244 million after the sampling procedure. We ap-

[7] https://github.com/alvations/sacremoses

		In-Domain Test					**X-WMT Test**				
Pairs	**Train size**	**Bilingual**		**MNMT**			**Bilingual**		**MNMT**		
		BLEU	**Chrf**	**BLEU**	**Chrf**	**PPL**	**BLEU**	**Chrf**	**BLEU**	**Chrf**	**PPL**
en-tr	35.8M	31.45	0.51	**33.09**	**0.51**	8.18	16.04	0.55	**26.74**	**0.56**	12.76
tr-en	35.8M	31.37	0.50	**35.48**	**0.52**	7.19	20.39	0.51	**24.66**	**0.55**	10.88
ru-uz	1.3M	**53.12**	**0.76**	44.73	0.71	3.02	6.58	0.41	**6.70**	**0.42**	82.20
uz-ru	1.3M	**55.39**	**0.76**	46.42	0.71	3.27	6.08	0.36	**9.16**	**0.39**	16.70
en-kk	564.8K	**24.53**	**0.54**	18.92	0.49	10.45	7.82	0.40	**9.92**	**0.43**	10.02
kk-en	564.8K	**29.17**	**0.51**	24.67	0.48	7.47	12.00	0.42	**15.71**	**0.44**	26.02
az-en	548.9K	**26.65**	**0.48**	20.47	0.42	7.70	12.01	0.41	**20.41**	**0.49**	14.46
en-az	548.9K	**34.73**	**0.56**	15.27	0.42	8.74	6.79	0.38	**9.71**	**0.43**	10.59
en-uz	529.6K	**45.95**	**0.66**	27.80	0.51	6.04	6.34	0.40	**9.89**	**0.42**	47.45
uz-en	529.6K	**38.72**	**0.58**	32.44	0.50	6.15	4.81	0.24	**14.45**	**0.45**	19.08
ba-ru	523.7K	**46.02**	**0.69**	40.59	0.64	3.75	24.39	**0.58**	**24.57**	0.57	5.49
ru-ba	523.7K	**51.26**	**0.74**	43.44	0.67	3.24	**24.31**	**0.59**	23.13	0.56	6.29
az-tr	410.1K	**23.47**	**0.48**	18.40	0.43	8.87	10.61	0.43	**19.63**	**0.48**	23.42
tr-az	410.1K	**29.97**	**0.53**	15.71	0.42	8.37	7.78	0.39	**8.21**	**0.42**	14.51
en-ky	312.6K	**21.66**	**0.44**	14.54	0.38	10.77	2.33	0.27	**4.64**	**0.34**	19.57
ky-en	312.6K	**24.96**	**0.42**	18.01	0.38	11.02	4.65	0.29	**10.87**	**0.39**	35.64
ky-ru	293.7K	**19.63**	**0.40**	16.30	0.38	10.04	5.23	0.30	**14.08**	**0.44**	9.43
ru-ky	293.7K	**18.57**	**0.43**	14.82	0.40	9.58	4.42	0.35	**10.35**	**0.45**	11.52
ba-en	34.3K	**21.51**	0.36	17.79	**0.37**	10.81	0.32	0.19	**10.55**	**0.40**	37.89
en-ba	34.3K	**17.78**	0.33	17.29	**0.35**	10.52	0.16	0.14	**8.35**	**0.34**	21.43
en-kaa	17.1K	15.34	0.40	**19.42**	**0.46**	8.83	0.31	0.19	**2.82**	**0.27**	77.93
kaa-en	17.1K	**22.82**	0.43	21.95	**0.48**	8.56	1.04	0.21	**10.21**	**0.38**	38.17
ru-sah	9.2K	**13.26**	**0.35**	5.46	0.19	30.82	0.12	0.16	**4.64**	**0.17**	58.01
sah-ru	9.2K	**16.35**	**0.36**	13.11	0.26	23.00	0.42	0.18	**4.41**	**0.25**	40.68
en-sah	8.1K	**13.45**	**0.36**	4.98	0.18	34.31	0.04	**0.14**	**3.46**	0.12	75.38
sah-en	8.1K	**22.19**	**0.40**	5.90	0.23	24.58	0.16	0.21	**3.38**	**0.24**	110.50

Table 3: Experiments results from bilingual baselines and MNMT model evaluated on the in-domain and out-of-domain test sets. *BLEU* and *Chrf* uses the SacreBLEU implementation and *PPL* refers to the internal perplexity of the MNMT model.

ply the sentencepiece[8] implementation of the byte pair encoding (BPE) (Sennrich et al., 2016) with a joint vocabulary size of 64k. Following the method from Ha et al. (2016) and Johnson et al. (2017), we prepend a target language token to the source sentences to enable many-to-many translation.

4.2.2 Model details

We train the model using the Transformer architecture in the *transformer-base* configuration. More specifically, we use the *transformer_wmt_en_de* version from Fairseq (Ott et al., 2019) implementation[9] with 6 layers both in the encoder and decoder. Configuration of the model closely follows the original implementation of the Transformer (Vaswani et al., 2017a) with the model dimension set at 512 and hidden dimension size at 2048. We apply a dropout rate of 0.3, the learning rate of $5 * 10^{-4}$, and warm-up updates of 40k. The effective batch size is 16,384 BPE tokens. The model is trained using 4 NVIDIA V100 GPU machines for a little over 1 million steps which takes about 36-48 hours.

4.3 Evaluation of Models

Automatic evaluation metrics used to compare the performance of bilingual baselines and MNMT are token-based corpus BLEU (Papineni et al., 2002) and character-based Chrf (Popović, 2015). While corpus BLEU is the de-facto standard in MT (Marie et al., 2021), Chrf might work better for morphologically rich languages because it can reward partially correct words. We also report the MNMT model's internal perplexity to better highlight the language pairs in which the model struggles most. We evaluate the models on the in-domain and X-WMT evaluation sets. The gap between scores on in-domain

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

[9] `https://github.com/pytorch/fairseq/tree/master/examples/translation`

	Bilingual		MNMT		Gain	
	BLEU	Chrf	BLEU	Chrf	BLEU	Chrf
XX-En	6.92	0.31	13.78	0.42	+6.86	+0.11
En-XX	4.57	0.30	9.37	0.36	+4.80	+0.06
XX-Ru	9.03	0.36	13.06	0.41	+4.03	+0.06
Ru-XX	8.86	0.38	11.21	0.40	+2.35	+0.02
XX-XX	8.99	0.38	12.49	0.41	+3.49	+0.04

Table 4: Performance comparison with different language groups and their overall gains in the MNMT setup. XX refers to the Turkic languages in the corpus.

versus out-of-domain translations is particularly interesting since it gives us an estimate of domain robustness and generalization, as well as mimics a realistic shift from the training domain to the domain of interest for potential users or downstream applications.

4.4 Bilingual baselines vs MNMT

Table 3 shows all the results for the bilingual baselines and MNMT as evaluated on two test tests. The first obvious trend in the table is the dominance of the bilingual baselines on the in-domain test sets as they overperform the MNMT model in most of the high- to mid-resource language pairs. As the train size decreases, the results become more comparable in terms of BLEU and even better for MNMT when evaluated in Chrf. When tested under a domain shift with the X-WMT set, MNMT results in gains across almost all pairs. However, it is important to note that there is a noticeable performance drop that follows the domain shift as can be seen in Figures 1 and 2. This highlights a realistic phenomenon of generalization and sets an expectation of the model's capabilities in real-world use cases.

Another observation in Table 3 is that all of the language pairs having fewer than 100k training samples (8 total) in our bilingual baselines barely pass the mark of 1 BLEU score or 0.2 Chrf in the out-of-domain test. However, in the MNMT setup, the average BLEU and Chrf score for those 8 low-resource pairs are 5.98 and 0.27 respectively. While these scores indicate that these pairs are still extremely low in quality and potentially unusable in practice, gains are promising given the amount of resources and a moderately-sized MNMT model.

To examine the generalization of the MNMT model into different language groups, we calculate the average gains for all pairs translating into English (XX-En), from English (En-XX), into Russian (XX-Ru), from Russian (Ru-XX), and direct pairs (XX-XX). Table 4 shows the average gains per category in terms of BLEU and Chrf. As it looks, translating from and into English sees the most gains, which is very consistent with the findings from the community (Arivazhagan et al., 2019; Goyal et al., 2021). A positive trend is the increasing quality of direct pairs which are very comparable to the non-Turkic pairs. We hypothesize that one of the main reasons for this is that the TIL Corpus is a multi-centric dataset with training data between almost all language pairs which allows us to train a complete Multilingual Neural Machine Translation (cMNMT) (Freitag and Firat, 2020). As shown in (Freitag and Firat, 2020; Fan et al., 2021), MNMT models trained on multi-centric parallel corpora tend to result in performance gains between non-English pairs.

	Adequacy				Fluency			
	Avg	k	LL	UL	Avg	k	LL	UL
en-tr	2.97	0.33	0.23	0.43	3.20	0.12	0.04	0.21
tr-en	2.95	0.45	0.36	0.55	3.18	0.40	0.30	0.50
en-uz	2.77	0.18	0.10	0.26	2.93	0.28	0.17	0.38
uz-en	3.05	0.28	0.20	0.37	3.19	0.29	0.18	0.39
ba-ru	2.74	0.58	0.48	0.67	3.34	0.63	0.54	0.73
ru-ba	2.81	0.27	0.17	0.37	3.06	0.19	0.09	0.29

Table 5: **Avg** represents the average score for either Adequacy or Fluency given by the annotators for each language pair. **k** represents the Cohen's Kappa score. **LL** represents the Lower Limit within 95% confidence. **UL** represents the Upper Limit within 95% confidence.

4.5 BLEU vs Chrf

Figure 3 compares BLEU and Chrf for all bilingual and multilingual models on X-WMT. We distinguish between translating into and from Turkic languages since all Turkic languages feature agglutination. As hinted above, we suspect that BLEU might underestimate translation quality when translating into Turkic languages. The graph shows a clear distinction that confirms this: For translations into non-Turkic languages, the relation between Chrf and BLEU is almost linear, with a Pearson correlation of 0.98 and a rank correlation of 0.98 as well. For translation into Turkic, the trend follows a more curved line, with a largely higher Chrf-to-BLEU ratio. The Pearson correlation is much lower at 0.87, but the rank correlation is only slightly lower than for non-Turkic languages at 0.92. Consequently, we can expect the same BLEU score to correspond to a higher Chrf score when translating into Turkic languages than from them. This means that while Chrf and BLEU are likely to pro-

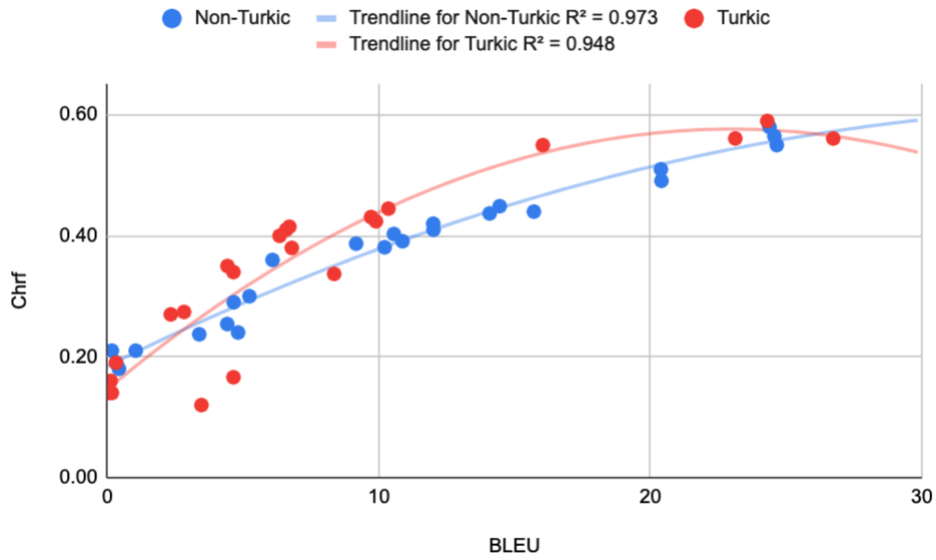

Figure 3: Correlations between BLEU and Chrf scores when the target language is Turkic and non-Turkic.

duce similar rankings of systems (at least in our scenario with standard comparable Transformer models), the Chrf score might better characterize the absolute translation quality. Our human evaluation does not cover sufficient language pairs (three from and three into Turkic languages) to yield a reliable empirical confirmation for this hypothesis. Future studies of larger scale as in the WMT metrics shared task (Mathur et al., 2020) will be needed.

5 Human Evaluation of MNMT

5.1 Human evaluation setup

To facilitate analysis on how well evaluation metrics measure the quality of the translations, we conduct human evaluations using the outputs from the MNMT model on the X-WMT set. We use Direct Assessment (DA) and follow the TAUS guidelines[10] with the only exception being the number of annotators per language pair, where we employ 2 annotators per language pair instead of 4[11]. In our DA, two hundred sentences of the MNMT model's output per language pair are evaluated based on its adequacy and fluency on respective 1-4 point scales. Annotators received an explanation of the rating scales with the task (e.g. "Adequacy: On a 4-point scale rate how much of the meaning is represented in the translation: 4: Everything 3: Most 2: Little 1: None"). To measure the inter-annotator agreement (IAA) between the two annotators of each language pair, we compute the Weighted Cohen's Kappa statistic (Cohen, 1960).

The language pairs involved in this human study are English-Turkish, Turkish-English, Bashkir-Russian, Russian-Bashkir, Uzbek-English, and English-Uzbek. These pairs were selected on the basis of language and script diversity, their performance on the X-WMT test set, and the availability of annotators.

5.2 Discussion and Results

The results of the average adequacy and fluency for each language pair are shown by Table 5. Most of the chosen language pairs received an average score of around 3 for both adequacy and fluency. This indicates that the model was largely able to convey most intended meaning in a good grammatical sense to a native speaker. Fluency is consistently rated higher than adequacy, which is a common theme in NMT evaluation (Martindale et al., 2019). The large difference in BLEU (5 BLEU points) between en-uz and uz-en is still noticeable, but much smaller according to the human evaluation. Chrf estimates a quality difference of 0.3 here, which is closer to the human estimate.

The Cohen's Kappa scores for each language pair are present in Table 5. As Cohen's Kappa is a measure from 0–1 of how well the two annotators agreed with their evaluations while removing possible agreements by chance, Cohen's Kappa score serves as one metric in deciding the reliability of

[10] https://rb.gy/eqlgbm
[11] Due to limited resources.

en-tr	
Adequacy: 3.00 — **Fluency:** 4.00 — *Fluent Output with Inadequate Verbal Tense*	
Reference Toyota, Subaru'daki hissesini 'den fazla artıracağını söyledi.	Hypothesis Toyota, Subaru'daki hisseyi 'den fazla artırdığını söyledi.
Adequacy: 4.00 — **Fluency:** 3.00 — *Lexical choice preserves meaning, still not the natural construction*	
Reference Başka birisi ağır yaralandı.	Hypothesis Başka bir kişi kötü yaralandı.
tr-en	
Adequacy: 3.00 — **Fluency:** 2.00 — *Some of the translations made lost the original meaning*	
Reference The schoolgirl who died from catastrophic injuries following a **suspected** hit-and-run in Newcastle has been **pictured** for the first time.	Hypothesis After a **suspicious** hit-and-run in Newcastle's, the student who died badly was first **seen**.
Adequacy: 3.00 — **Fluency:** 4.00 — *Maintains grammatical form, but changes the meaning*	
Reference He further **dismissed** the embargo as an **attack** on the rights of citizens.	Hypothesis He also **denied** the ambargo by defending an **attack** on citizens' rights.
ba-ru	
Adequacy: 2.00 — **Fluency:** 3.00 — *"kiss" translates to "kill" and changes the meaning completely*	
Reference В ночь после выборов, пишет Ло Бьянко в своей книге, Карен Пенс отказалась поцеловать мужа.	Hypothesis В свою книгу Ло Бьянко, в ночь после выборов, Карен Пенс отказывается от смерти мужа.
Adequacy: 3.00 — **Fluency:** 2.00 — *Incorrect pronoun ("she" to "he"). Few awkward translations*	
Reference Поэтому она откликнулась на вакансию в Fast Trak Management, маленькой компании, которая называет себя "маркетинговой фирмой номер один в Северной Вирджинии".	Hypothesis Поэтому он согласился на вакансию Fast Trak Management в малой компании, которая называла себя "Первую маркетинговую фирму в Северной Вирджинии".
ru-ba	
Adequacy: 3.00 — **Fluency:** 2.00 — *When several verbs are present, some are omitted from the translation*	
Reference Видео Пирзаданың бер нисә йылан һәм аллигаторҙо тотоп торғанын күрһәтә.	Hypothesis Дәүләт еренә йәмәғәт **access** Видеоға ярашлы, Пирзада бер нисә йылан һәм алгигатор менән нисек эш итә.
Adequacy: 2.00 — **Fluency:** 3.00 — *A whole part of the original sentence is omitted from the translation*	
Reference iHandy тарафынан киң билдәле эмодзи-ҡушымталар серияһы сығарылды, әммә улар ҙа Google Play Store системаһынан шунда уҡ юйылды.	Hypothesis iHandy Google Play Store-ҙан сығарылған популяр эмодзи-приложениялар серияһы булдырылды.
uz-en	
Adequacy: 2.00 — **Fluency:** 2.00 — *Changed the order events*	
Reference Antonio Brown has indicated he's not retiring from the NFL, only a few days after announcing he was done with the league in a rant.	Hypothesis Antonio Braun said that after a few days after the NFL, he won't leave after he announced that he was engaged in league.
Adequacy: 2.00 — **Fluency:** 4.00 — *Improper changes from original nouns, and different sense of "hold"*	
Reference **Harker** says **Fed** should '**hold** firm' on interest rates	Hypothesis **Everyone** thinks that this is how to **hold** the **Federal** rate percentages.
en-uz	
Adequacy: 3.00 — **Fluency:** 3.00 — *"Gumonlanuvchi":"a suspect"."Shubhachi":"someone who suspects"*	
Reference Keyin ushbu mashinadan uch nafar **gumonlanuvchi** tushayotganini ko 'rishdi.	Hypothesis Keyinchalik uchta **shubhachi** mashinadan chiqib ketganini ko'rishdi.
Adequacy: 2.00 — **Fluency:** 2.00 — *Use of a correct but a foreign word (başarısız)*	
Reference WeWork's Neumann muvaffaqiyatsiz IPO o 'tkazilgandan so'ng o 'zini bosh direktor lavozimidan chetlatishga ovoz berdi	Hypothesis **Biz Work"s** Neumann IPO **başarısız** bo'lganidan so'ng **O'zbekiston** Bosh direktori sifatida ovoz berdi

Table 6: Qualitative Analysis of the MNMT model output for 6 language pairs. The Reference sentence shows the intended translation while the Hypothesis shows the MNMT model output.

Pairs	Train Size	In-Domain		X-WMT	
		BLEU	Chrf	BLEU	Chrf
ru-ba	523.7K	54.48 (+11.04)	0.743 (+0.07)	24.85 (+0.54)	0.569 (-0.02)
ky-en	312.6K	24.21 (+6.2)	0.42 (+0.05)	10.26 (+5.61)	0.38 (+0.09)
en-ba	34.3K	30.43 (+13.14)	0.46 (+0.11)	4.56 (+4.4)	0.22 (+0.08)
ru-sah	9.2K	49.46 (+44.00)	0.585 (0.4)	22.05 (+21.93)	0.348 (+0.19)

Table 7: Experiment results from the finetuning of the MNMT model.

the evaluations. We see that the reliability varies across language pairs and between adequacy and fluency. Translation into English or Russian has a higher agreement on average than in the opposite direction (en/tr is a tie).

5.3 Qualitative Analysis

To gain better qualitative insight into the model outputs in each of the 6 language directions, we asked the annotators to identify 2 examples that highlight the most commonly witnessed mistakes during their review. Table 6 showcases those examples along with a brief explanation for their scores. From this analysis, it seems that the severity of mistakes that the MNMT model makes in *adequacy* tends to range from certain words being translated to a slightly different meaning to the original intention of the sentence being lost. As for *fluency*, the errors seem to range from awkward wording to clear grammatical mistakes. There are a few cases where there is an off-target translation for a word or a segment of the sentence.

6 The Promise of MNMT: Cross-lingual Knowledge Transfer

One of the biggest advantages of a large MNMT model is its capacity for transfer learning as can be accomplished through fine-tuning. Since we plan on releasing the model to the public, we believe many understudied and underperforming language pairs could benefit from cross-lingual knowledge transfer. This phenomenon is well-known in the broader NLP community as well as in MT research. To test this hypothesis, we fine-tune our MNMT model on 4 language pairs ranging from high(er)-resource to extremely low-resource in training data available. Table 7 shows the results of the experiments. As it can be seen, the performance of the models improves steadily across all resource types, low-resource cases experiencing gains up to 44 BLEU points (or 0.4 Chrf) from the bilingual baselines in the in-domain evaluation. However, in out-of-domain scenarios, gains are not as significant. Mid- to high-resource pairs improve modestly in the range of 1–5 BLEU points (or 0–0.1 Chrf) while a low-resource pair, Russian-Sakha gains up to 22 BLEU points (0.19 Chrf).

7 Future Work and Conclusion

In this work, we train and evaluate the first large-scale MNMT model for the Turkic language family which consists of many underexplored languages. Among many results, we find it very promising to train and finetune a MNMT model with a language family corpus as it boosts the cross-lingual knowledge transfer between the related languages and consistently improves over the strong bilingual baselines in out-of-domain scenarios. Our analysis also shows that Chrf and BLEU do not correlate in the same when the target language group if different: BLEU underestimates the translations for the Turkic languages.

In the future work, we hope to include more of the underrepresented Turkic language pairs in the study and explore the potential of transfer learning into the translation of unseen languages and language pairs (”zero-shot”).

Acknowledgements

We thanks all of the members and partners of the Turkic Interlingua (TIL) community for their contributions to the project. Namely, we would like to thank our dedicated translators and annotators: Nurlan Maharramli, Sariya Kagarmanova, Iskander Shakirov, Aydos Muxammadiyarov, Ziyodabonu Qobiljon qizi, Alperen Cantez, Doniyorbek Rafikjonov, Mukhammadbektosh Khaydarov, Madina Zokirjonova, Erkinbek Vokhabov, Petr Popov, Abilxayr Zholdybai and Akylbek Khamitov. We also acknowledge and appreciate significant dataset contributions from Rasul Karimov, Khan Academy O’zbek[12], and the Foundation for the Preservation and Development of the Bashkir Language[13].

[12] https://uz.khanacademy.org/

[13] https://bsfond.ru/

References

Željko Agić and Ivan Vulić. 2019. JW300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3204–3210, Florence, Italy. Association for Computational Linguistics.

Emel Alkım and Yalçın Çebi. 2019. Machine translation infrastructure for Turkic languages (MT-Turk). *The International Arab Journal of Information Technology*, 16(3):380–388.

Kemal Altıntaş. 2001. *Turkish to Crimean Tatar machine translation system*. Ph.D. thesis, Bilkent University.

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. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges.

Zhenisbek Assylbekov and Assulan Nurkas. 2014. Initial explorations in kazakh to english statistical machine translation. In *The First Italian Conference on Computational Linguistics CLiC-it 2014*, page 12.

Duygu Ataman, Matteo Negri, Marco Turchi, and Marcello Federico. 2017. Linguistically motivated vocabulary reduction for neural machine translation from turkish to english. *The Prague Bulletin of Mathematical Linguistics*, 108(1):331–342.

Xolisa Axmedova, Guzal Abdujalilova, and Umida Abdurahmonova. 2019. Algorithm based on linguistic models in machine translation between russian and uzbek. *ACADEMICIA: An International Multidisciplinary Research Journal*, 9(12):16–21.

Arianna Bisazza and Marcello Federico. 2009. Morphological pre-processing for turkish to english statistical machine translation. In *Proceedings of IWSLT 2009*.

Eleftheria Briakou and Marine Carpuat. 2019. The university of Maryland's Kazakh-English neural machine translation system at WMT19. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 134–140.

Mustafa Alp Çetin and Rita Ismailova. Assisting tool for essay grading for Turkish language instructors. *MANAS Journal of Engineering*, 7(2):141–146.

Narayan Choudhary and Girish Nath Jha. 2011. Creating multilingual parallel corpora in indian languages. In *Language and Technology Conference*, pages 527–537. Springer.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-Task Learning for Multiple Language Translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1723–1732, Beijing, China. Association for Computational Linguistics.

Ilknur Durgar El-Kahlout and Kemal Oflazer. 2006. Initial explorations in english to turkish statistical machine translation. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 7–14.

Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzmán, and Philipp Koehn. 2020. CCAligned: A massive collection of cross-lingual web-document pairs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5960–5969, Online. Association for Computational Linguistics.

Miquel Esplà-Gomis, Mikel L Forcada, Gema Ramírez-Sánchez, and Hieu Hoang. 2019. ParaCrawl: Web-scale parallel corpora for the languages of the EU. In *Proceedings of Machine Translation Summit XVII Volume 2: Translator, Project and User Tracks*, pages 118–119.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond english-centric multilingual machine translation. *CoRR*, abs/2010.11125.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Rauf Fatullayev, Ali Abbasov, and Abulfat Fatullayev. 2008. Dilmanc is the 1st MT system for Azerbaijani. *Proc. of SLTC-08, Stockholm, Sweden*, pages 63–64.

∀, Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Tajudeen Kolawole, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddee Hassan Muhammad, Salomon Kabongo, Salomey Osei, et al. 2020. Participatory research for low-resourced machine translation: A case study in african languages. *Findings of EMNLP*.

Markus Freitag and Orhan Firat. 2020. Complete multilingual neural machine translation. *arXiv preprint arXiv:2010.10239*.

Memduh Gökırmak, Francis Tyers, and Jonathan Washington. 2019. Machine Translation for Crimean

Tatar to Turkish. In *Proceedings of the 2nd Workshop on Technologies for MT of Low Resource Languages*, pages 24–31.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2021. The FLORES-101 evaluation benchmark for low-resource and multilingual machine translation. *CoRR*, abs/2106.03193.

Thanh-Le Ha, Jan Niehues, and Alexander Waibel. 2016. Toward multilingual neural machine translation with universal encoder and decoder. *arXiv preprint arXiv:1611.04798*.

Ilker Hamzaoglu. 1993. *Machine translation from Turkish to other Turkic languages and an implementation for the Azeri language*. Ph.D. thesis, MSc Thesis, Bogazici University, Istanbul.

Sardana Ivanova, Anisia Katinskaia, and Roman Yangarber. 2019. Tools for supporting language learning for sakha. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 155–163, Turku, Finland. Linköping University Electronic Press.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Pratik Joshi, Christain Barnes, Sebastin Santy, Simran Khanuja, Sanket Shah, Anirudh Srinivasan, Satwik Bhattamishra, Sunayana Sitaram, Monojit Choudhury, and Kalika Bali. 2019. Unsung challenges of building and deploying language technologies for low resource language communities. *arXiv preprint arXiv:1912.03457*.

Azizbek Kadirov. 2015. The algorithm of machine translation from uzbek to karakalpak. *TurkLang-2015*, page 24.

Aidar Khusainov, Dzhavdet Suleymanov, Rinat Gilmullin, and Ajrat Gatiatullin. 2018. Building the Tatar-Russian NMT system based on re-translation of multilingual data. In *International Conference on Text, Speech, and Dialogue*, pages 163–170. Springer.

Aidar Khusainov, Dzhavdet Suleymanov, Rinat Gilmullin, Alina Minsafina, Lenara Kubedinova, and Nilufar Abdurakhmonova. 2020. First Results of the "TurkLang-7" Project: Creating Russian-Turkic Parallel Corpora and MT Systems.

Rachel Killackey. 2013. Statistical Machine Translation from English to Tuvan.

Diederik P. Kingma and Jimmy Lei Ba. 2015. Adam: A Method for Stochastic Optimization. In *ICLR 2015 : International Conference on Learning Representations 2015*.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86. Citeseer.

Julia Kreutzer, Jasmijn Bastings, and Stefan Riezler. 2019. Joey NMT: A minimalist NMT toolkit for novices. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 109–114, Hong Kong, China. Association for Computational Linguistics.

Patrick Littell, Chi-kiu Lo, Samuel Larkin, and Darlene Stewart. 2019. Multi-source transformer for Kazakh-Russian-English neural machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 267–274.

Muhtar Mahsut, Yasuhiro Ogawa, Kazue Sugino, Katsuhiko Toyama, and Yasuyoshi Inagaki. 2004. An experiment on Japanese-Uighur machine translation and its evaluation. In *Conference of the Association for Machine Translation in the Americas*, pages 208–216. Springer.

Benjamin Marie, Atsushi Fujita, and Raphael Rubino. 2021. Scientific credibility of machine translation research: A meta-evaluation of 769 papers. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7297–7306, Online. Association for Computational Linguistics.

Marianna Martindale, Marine Carpuat, Kevin Duh, and Paul McNamee. 2019. Identifying fluently inadequate output in neural and statistical machine translation. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, pages 233–243, Dublin, Ireland. European Association for Machine Translation.

Nitika Mathur, Johnny Wei, Markus Freitag, Qingsong Ma, and Ondřej Bojar. 2020. Results of the WMT20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 688–725, Online. Association for Computational Linguistics.

Jamshidbek Mirzakhalov. 2021. *Turkic Interlingua: A Case Study of Machine Translation in Low-resource Languages*. Ph.D. thesis, University of South Florida.

Jamshidbek Mirzakhalov, Anoop Babu, Duygu Ataman, Sherzod Kariev, Francis Tyers, Otabek Abduraufov, Mammad Hajili, Sardana Ivanova, Abror Khaytbaev, Antonio Laverghetta Jr., Behzodbek

Moydinboyev, Esra Onal, Shaxnoza Pulatova, Ahsan Wahab, Orhan Firat, and Sriram Chellappan. 2021. A large-scale study of machine translation in the turkic languages.

Maimitili Nimaiti and Yamamoto Izumi. 2012. A rule based approach for japanese-uighur machine translation system. In *2012 IEEE 11th International Conference on Cognitive Informatics and Cognitive Computing*, pages 124–129.

Hiroki Nomoto, Kenji Okano, David Moeljadi, and Hideo Sawada. 2018. Tufs asian language parallel corpus (talpco). In *Proceedings of the Twenty-fourth Annual Meeting of the Association for Natural Language Processing*, pages 436–439.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019, pages 9 – 16, Mannheim. Leibniz-Institut für Deutsche Sprache.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Matt Post, Chris Callison-Burch, and Miles Osborne. 2012. Constructing parallel corpora for six indian languages via crowdsourcing. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 401–409.

Ilnar Salimzyanov, J Washington, and F Tyers. 2013. A free/open-source Kazakh-Tatar machine translation system. *Machine Translation Summit XIV*, pages 175–182.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, and Armand Joulin. 2019. Ccmatrix: Mining billions of high-quality parallel sentences on the WEB. *CoRR*, abs/1911.04944.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. 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)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

JL Song and L Dai. 2015. Construction of Uighur-Chinese parallel corpus. In *Multimedia, Communication and Computing Application: Proceedings of the 2014 International Conference on Multimedia, Communication and Computing Application (MCCA 2014), Xiamen, China, October 16-17, 2014*, page 353. CRC Press.

Aida Sundetova, Mikel Forcada, and Francis Tyers. 2015. A free/open-source machine translation system from english to kazakh. In *PROCEEDINGS OF THE INTERNATIONAL CONFERENCE" TURKIC LANGUAGES PROCESSING" TurkLang-2015*, pages 78–90.

Ahmet Cüneyd Tantuğ, Eşref ADALI, and Kemal OFLAZER. 2011. Türkmenceden türkçeye bilgisayarlı metin çevirisi. *İTÜDERGİSİ/d*, 7(4).

A. Cüneyd Tantuğ and Eşref Adalı. 2018. Machine translation between turkic languages. In Kemal Oflazer and Murat Saraçlar, editors, *Turkish Natural Language Processing*, pages 237–254. Springer.

Jörg Tiedemann. 2020. The Tatoeba Translation Challenge–Realistic Data Sets for Low Resource and Multilingual MT. *arXiv preprint arXiv:2010.06354*.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal.

Ualsher Tukeyev, Aidana Karibayeva, and Balzhan Abduali. 2019. Neural machine translation system for the kazakh language based on synthetic corpora. In *MATEC Web of Conferences*, volume 252, page 03006. EDP Sciences.

Cigdem Keyder Turhan. 1997. An English to Turkish machine translation system using structural mapping. In *Fifth Conference on Applied Natural Language Processing*, pages 320–323.

Francis M Tyers, Jonathan North Washington, Ilnar Salimzyanov, and Rustam Batalov. 2012. A prototype machine translation system for Tatar and Bashkir based on free/open-source components. In *First Workshop on Language Resources and Technologies for Turkic Languages*, page 11.

Aidar Valeev, Ilshat Gibadullin, Albina Khusainova, and Adil Khan. 2019. Application of Low-resource Machine Translation Techniques to Russian-Tatar Language Pair. *arXiv preprint arXiv:1910.00368*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017a. Attention is all you need. In *NeurIPS*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017b. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Dongqi Wang, Zihan Liu, Qingnan Jiang, Zewei Sun, Shujian Huang, and Jiajun Chen. 2020. NJUNLP's Machine Translation System for CCMT-2020 Uighur - Chinese Translation Task. In *China Conference on Machine Translation*, pages 76–82. Springer.

Jonathan North Washington, Ilnar Salimzianov, Francis M. Tyers, Memduh Gökırmak, Sardana Ivanova, and Oğuz Kuyrukçu. 2019. Free/open-source technologies for Turkic languages developed in the Apertium project. In *Proceedings of the International Conference on Turkic Language Processing (TURKLANG 2019)*.