A Large-Scale Study of Machine Translation in Turkic Languages

Mirzakhalov, Jamshidbek

A Large-Scale Study of Machine Translation in the Turkic Languages

Jamshidbek Mirzakhalov[a,b]**, Anoop Babu**[a,b]**, Duygu Ataman**[a,c]**, Sherzod Kariev**[a,b]**,
Francis Tyers**[a,d]**, Otabek Abduraufov**[a,b]**, Mammad Hajili**[a,e]**, Sardana Ivanova**[a,f]**,
Abror Khaytbaev**[a,b]**, Antonio Laverghetta Jr.**[a,b]**, Behzodbek Moydinboyev**[a,b]**, Esra Onal**[a,d]**,
Shaxnoza Pulatova**[a,g]**, Ahsan Wahab**[a]**, Orhan Firat**[a,h]**, Sriram Chellappan**[a,b]

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

Abstract

Recent advances in neural machine translation (NMT) have pushed the quality of machine translation systems to the point where they are becoming widely adopted for building competitive systems. However, there is still a large number of languages that are yet to reap the benefits of NMT. In this paper, we provide the first large-scale case study of the practical application of MT in the Turkic language family in order to realize the gains of NMT for Turkic languages under high-resource to extremely low-resource scenarios. In addition to presenting an extensive analysis that identifies the bottlenecks towards building competitive systems to ameliorate data scarcity, our study has several key contributions, including, i) a large parallel corpus covering 22 Turkic languages consisting of common public datasets in combination with new datasets of approximately 2 million parallel sentences, ii) bilingual baselines for 26 language pairs, iii) novel high-quality test sets in three different translation domains and iv) human evaluation scores. All of our data, software and models are publicly available.[1]

1 Introduction

Having been studied widely over the last few decades, machine translation (MT) evaluation has traditionally focused on European languages, due to limitations of the available technology as well as resources. Although low-resource MT has recently started to gain more attention and new evaluation benchmarks are becoming available (Guzmán et al., 2019; Ojha et al., 2020; Fraser, 2020; Ansari et al., 2020), there are still a large amount of underrepresented languages excluded from MT evaluation. In addition to the cost of preparing such labor-intensive annotations, the lack of training resources also limits the evaluation of MT models in

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

Name	Codes	Articles	Speakers	MT?
English	en, eng	6,237,470	400M	✓
Russian	ru, rus	1,694,280	258M	✓
Turkish	tr, tur	388,641	85.0M	✓
Uzbek	uz, uzb	139,635	27.0M	✓
Azerbaijani	az, aze	177,536	23.0M	✓
Kazakh	kk, kaz	228,123	13.2M	✓
Uyghur	ug, uig	4,898	10.0M	✓
Turkmen	tk, tuk	5,876	6.70M	✓
Tatar	tt, tat	237,332	5.20M	✓
Kyrgyz	ky, kir	80,738	4.30M	✓
Bashkir	ba, bak	55,477	1.40M	✓
Chuvash	cv, chv	45,275	1.04M	✓
Karakalpak	kaa	1,882	583K	✗
Crimean Tatar	crh	8,633	540K	✗
Sakha (Yakut)	sah	13,027	450K	✓
Kumyk	kum	—	450K	✗
Karachay-Balkar	krc	2,049	310K	✗
Tuvan	tyv	3,164	280K	✗
Urum	uum	—	190K	✗
Gagauz	gag	2,737	148K	✗
Salar	slr	—	70K	✗
Altai	alt	—	56K	✗
Khakas	kjh	—	43K	✗
Shor	cjs	—	3K	✗

Table 1: Number of Wikipedia articles for Turkic languages compared to English and Russian along with number of L1 speakers and two- and three-letter language codes. The column **MT?** indicates if there are currently available online machine translation systems for the language. (K: thousand, M: million.)

terms of their applicability across a wide range of world languages. On the other hand, many studies have pointed to the limited applicability of prominent methods in MT research including models and evaluation metrics (Birch et al., 2008; Stanojević et al., 2015; Bugliarello et al., 2020) in translating languages with varying linguistic typology.

In order to extend the evaluation of the state-of-the-art methods in MT (Joshi et al., 2019) and ultimately aid in designing methods with wider range of applicability, in this paper, we present a large-scale case study of MT methods in a very challenging case of the Turkic language family. The Turkic

language family consists of around 35 languages spoken by communities across Eurasia by around 200 million people. Of this number, around 20 are official languages of a state, or sub-national entity, with the remaining being minority languages. The languages are distinct in their highly complex use of morphology, and thus create extremely sparse vocabularies, presenting a challenging case of evaluation of statistical models, in particular MT systems (Tantuğ et al., 2008) and n-gram language models (Bender, 2011; Tsarfaty et al., 2020). Table 1 presents the amount of resources and the number of speakers in Turkic languages[2] which aids our analysis on the feasibility in crowdsourcing, based on the approach of Moshagen et al. (2014).

Our study includes the preparation of novel public resources covering many languages in the Turkic family, most of which included for the first time in parallel corpora. We also present new benchmarks for MT which could be used for assessing different factors determining the limits of MT methods in various languages, such as data size, evaluation metrics, translation domain, linguistic typology, relatedness, and the writing system. We test the use of our resources in MT and present the first evaluation results for many Turkic languages. Our novel resources consist of *i)* a large-scale multicentric parallel corpus of 75M+ sentence pairs in 22 Turkic languages and their translations into English, Russian, as well as in-family languages, covering over 400 translation directions, *ii)* 3 new test sets for each translation direction curated from our corpus in 3 different translation domains, *iii)* bilingual baselines in 26 different language pairs. Our baselines are evaluated using automatic metrics as well as human assessments against commercial or open-source systems where applicable. We release our parallel corpora, test sets, and baseline systems publicly to encourage future research in Turkic languages.

2 Turkic Languages & MT

This section gives a brief overview of Turkic languages from a linguistic perspective as well as presenting the previous work on MT of these languages. In our study, we include 22 Turkic languages: Altai, Azerbaijani, Bashkir, Crimean Tatar, Chuvash, Gagauz, Karachay-Balkar, Karakalpak, Khakas, Kazakh, Kumyk, Kyrgyz, Sakha, Salar, Shor, Turkmen, Turkish, Tatar, Tuvan, Uyghur,

Urum, and Uzbek. There are several other widely spoken languages that were left out from our study such as Nogai, Khorasani Turkic, Qashqai, and Khalaj, due to the lack of any available parallel corpora. Future work will focus on extending the corpus to these languages as well.

2.1 Linguistic Typology

The Turkic languages are spoken in a wide area that stretches from south-eastern Europe to northeastern Asia. The languages are of the agglutinative morphological type and uniformly have Subject-Object-Verb main constituent order.

Nominal morphology is highly similar between the languages, with all of them exhibiting inflection for number, possession, and case. There are a variable number of cases, but the six-core cases of nominative, genitive, accusative, dative, locative, and ablative are extant in the vast majority of languages. As part of the nominal inflectional system, the languages also have a derivational process whereby locatives and genitives can be pronominalized and constitute full noun phrases in their own right. Verbal inflection, on the other hand, is more heterogeneous between the languages with each language having a variety of strategies for encoding tense, aspect, voice, modality, and evidentiality. One common feature however is that each of the languages has an extensive system of non-finite forms: verbal adjectives, verbal nouns, and verbal adverbs. These are full clauses that can be used as either modifiers (in the case of verbal adjectives and verbal adverbs) or heads (in the case of verbal nouns). Many of the languages also have constructions consisting of a non-finite verbal form and an auxiliary verb which constitute a single predicate, with the auxiliary verb giving extra information about tense or mood (Johanson and Johanson, 2015).

The modern Turkic languages are written in a variety of scripts, with Latin, Cyrillic, and Perso-Arabic being most common. Many of the languages have been written in several writing systems over the past century, making collecting texts more problematic. For example, we can find instances where the same language have texts that are written in Perso-Arabic before the 1920s, in Latin until the 1930s, in Cyrillic until the 1990s, and then in Latin again (Róna-Tas, 2015). In addition, many languages have gone through several orthographic norms based on the same script, and

some languages are currently written in different scripts depending on which country the speakers are in. This orthographic diversity makes collecting and collating text resources difficult, as many texts may be available only in a previously-used orthography and conversion between orthographic systems is never deterministic owing to the large number of loan words in many texts.

2.2 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 includes 14 Turkic languages and provides bilingual baselines for all translation directions present in the corpus. However, 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. Khusainov et al. (2020) collected a large-scale Russian-Turkic parallel corpus for 6 language pairs and reports bilingual baselines using a number of NMT-based approaches, although the dataset, test sets, and the models are not released to the public which limits its use to serve as a comparable benchmark. Alkım and Çebi (2019) introduces a rule-based MT framework for Turkic languages and demonstrates the performance with 4 language pairs. Washington et al. (2019) demonstrates several rule-based MT systems built for Turkic languages which are available through the Apertium[3] website.

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 (Tantug et al., 2007), 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, 2014; Song and Dai, 2015; Wang et al., 2020), Uzbek (Axmedova et al., 2019). Yet, to our knowledge, there has not been a study that covers Turkic languages in such a large extent as ours, both in terms of multi-lingual parallel corpora and benchmarks including multi-way comparable test sets in all languages.

3 TIL Corpus

Our parallel corpus is collected through unifying publicly available datasets and additional parallel data we prepare by crawling public domain resources. Table 2 shows the total amount of sentences in that particular language across the corpus along with number of sentences that are newly introduced (previously unavailable). This section describes the details of our data collection process.

3.1 Public Datasets

In our corpus we include the following public data sets:

- The Tatoeba corpus (Tiedemann, 2020) provides training and test sets for over 500 languages and thousands of translation pairs. It uses the latest version of OPUS[4] (Tiedemann and Nygaard, 2004) as training sets and use parallel sentences from the Tatoeba project for testing. Tatoeba consists of 58 language pairs of interest. For the purposes of our corpus, we merge the training, development, and test sets into a single set for all available languages.

- JW300 (Agić and Vulić, 2019) is a public dataset available for download through OPUS. Although most of the parallel data in JW300 was provided through the Tatoeba corpus, we have identified several pairs that were missing in Tatoeba but present in JW300. To avoid further data loss, we have obtained the JW300 dataset directly from OPUS and deduplicated it against the Tatoeba corpus. This dataset provided data for 59 language pairs of interest and resulted in 5.2 million parallel sentences.

- GoURMET[5] is another dataset available through OPUS and provides parallel sen-

---

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

[4]http://opus.nlpl.eu/
[5]https://gourmet-project.eu/

Language	Data	Script	Category	New Data
Turkish	52.6M	Latin	The Underdogs (4)	755.9K
Kazakh	5.3M	Arabic, Cyrillic, Latin	The Rising Star (3)	201.9K
Uzbek	2.9M	Arabic, Cyrillic, Latin	The Rising Star (3)	1.7M
Azerbaijani	2.2M	Arabic, Cyrillic, Latin	The Scraping-Bys (1)	284.8K
Tatar	1.8M	Arabic, Cyrillic	The Scraping-Bys (1)	192.0K
Kyrgyz	1.8M	Arabic, Cyrillic	The Scraping-Bys (1)	188.6K
Chuvash	1.5M	Cyrillic	The Scraping-Bys (1)	191.0K
Turkmen	921.0K	Arabic, Cyrillic, Latin	The Scraping-Bys (1)	191.7K
Bashkir	893.1K	Cyrillic	The Scraping-Bys (1)	713.9K
Uyghur	343.0K	Arabic, Cyrillic, Latin	The Scraping-Bys (1)	187.0K
Karakalpak	253.8K	Cyrillic, Latin	The Scraping-Bys (1)	274.3K
Khakas	219.0K	Cyrillic	The Left-Behinds (0)	242.8K
Altai	192.6K	Cyrillic	The Left-Behinds (0)	190.0K
Crimean Tatar	185.3K	Cyrillic, Latin	The Scraping-Bys (1)	197.6K
Kumyk	165.6K	Cyrillic	The Left-Behinds (0)	192.4K
Karachay-Balkar	162.8K	Cyrillic, Latin	The Scraping-Bys (1)	182.6K
Gagauz	157.4K	Cyrillic, Latin	The Scraping-Bys (1)	177.1K
Sakha	157.1K	Cyrillic	The Scraping-Bys (1)	174.8K
Tuvinian	103.2K	Cyrillic	The Scraping-Bys (1)	148.3K
Shor	2.3K	Cyrillic	The Left-Behinds (0)	6.9K
Salar	766	Latin	The Left-Behinds (0)	1.5K
Urum	491	Greek, Cyrillic, Latin	The Left-Behinds (0)	491

Table 2: Corpus details for each Turkic language. *Data* shows the aggregated amount of sentences across the corpus. *Category* refers to the language classes based on data resource according to (Joshi et al., 2020).

tences for 7 language pairs including English-Turkish and English-Kyrgyz. They are not available in Tatoeba due to a more recent release. English-Kyrgyz consists of 14.5 thousand sentence pairs while English-Turkish contains 1.3 million.

In addition to this, with the permission from the owners, we include privately owned corpora for English-Azerbaijani[6] containing data from news articles, English-Uzbek[7] containing data from KhanAcademy website localization, and Bashkir-Russian[8] having a mix of data from news articles and literary works.

3.2 Data Crawling

We obtained additional parallel data from a few different public domain websites that contain a large amount of text translated into many different languages. One of these includes *TED Talks*,[9] which contains talks across various domains that

are translated by volunteers. Qi et al. (2018) compiled a dataset for 60 languages, however, only a few Turkic languages were available at their time of curation. We have compiled an updated version of this dataset and obtained sentence pairs for 8 Turkic languages. *Bible.is*[10] is another website that contains an extensive list of languages into which religious texts and books are translated. 19 out of 22 Turkic languages were covered in this source with an average of approximately 8,000 sentence pairs for each translation direction. Additionally, we have crawled other public websites, online dictionaries, and resources with parallel data that were identified by native speakers of these languages. The full list of online resources we used in our crawling is given in the Appendices.

3.3 Data Alignment

All crawled documents are aligned using Hunalign (Varga et al., 2005), with a threshold of either 0.2 or 0.4 depending on the availability of a native speaker for the language. When crawling pre-aligned sources such as TED Talks, we noticed

---

[6]https://github.com/derintelligence/en-az-parallel-corpus/
[7]https://uz.khanacademy.org/
[8]https://github.com/AigizK/bashkort-parallel-corpora
[9]https://www.ted.com/talks

[10]http://www.bible.is

serious alignment issues with certain Turkic languages, especially when the source and target differ greatly in size. In these cases, we split both sides into sentences using NLTK sentence tokenizer[11] and realign using the Hunalign tool. Specifically for the Bible dataset, all the data has been aligned at the verse level first, then split into sentence-level bitexts whenever possible. This results in parallel texts that are relatively longer while ensuring higher quality alignments.

3.4 Data Preprocessing

Many of the languages in our dataset are written using multiple scripts, which creates consistency problems for building MT systems. Therefore, we transliterate three of the languages in our dataset that have a high mix of multiple scripts. Namely, we transliterate Uzbek into a Latin script, while all Karakalpak text is converted into Cyrillic. Although the performance of transliteration tools (Uzbek[12] and Karakalpak[13]) were not strictly evaluated, the tools we have used were recommended and widely adopted by the native speakers of the languages. Once we combine the entire corpus data, we deduplicate the sentences in each language pair.

4 Bilingual Baselines

We train bilingual baselines for 26 language pairs in three different resource categories: high (¿5M), medium (100K-5M) and low (¡100K). The choice of pairs to train was based on multiple factors such as the availability of test sets, native speakers (for human evaluation), and other comparable MT systems.

4.1 Model Details

All models are Transformers (Vaswani et al., 2017) (*transformer-base*) whose exact configuration depends on the amount of data available for training. Models for low-resource pairs use 256-dimensional embeddings and hidden layers. Models for mid-resource pairs use 512-dimensonal embeddings and hidden layers. The models for high-resource pairs use the same 512-dimensonal embedding and hidden layer sizes for the encoder, but for the decoder both dimensions are increased to 1024. All models are trained with the Adam optimizer (Kingma and Ba, 2015) over cross-entropy loss with a maximum learning rate of $3 * 10^{-4}$ and a minimum of

$1 * 10^{-8}$, which warms up for the first 4800 training steps and then decays after reaching the maximum. We use a training batch size of 4096. We use perplexity as our early stopping metric with a patience of 5 epochs. We set a dropout (Srivastava et al., 2014) probability of 0.3 in both the encoder and the decoder. We apply a byte pair encoding (BPE) (Sennrich et al., 2015; Dong et al., 2015) with a joint vocabulary size of 4K and 32K for low- and mid/high-resource scenarios respectively.

All models use the Joey NMT (Kreutzer et al., 2019) implementation and apex[14] where possible to speed up training. Models were trained on pre-emptible GPUs freely available on Google Colab.[15]

4.2 Test Sets

High-quality and diverse test sets are essential in evaluating the strength and weaknesses of MT systems. We curate 3 test sets covering 3 translation domains: religious (Bible), conversational (TED Talks), and news (X-WMT).

Bible dataset is the main source that exists across almost all of the 24 language pairs that are included in our corpus. From this dataset, around 400 to 800 most commonly present sentences for every language pair were separated to create a test set. This allowed having a test set comparable in all language pairs, which we find essential for a controlled evaluation and believe would be a useful resource in future studies involving multilingual models.

TED Talks is another resource we use for collecting sentences across multiple languages to create a language-wise comparable test set in the conversational domain. This allows our approach to be comparable also across different domains. After deduplication, 3000-5000 sentences per language pair are picked as a part of our TED Talks test set.

X-WMT is our 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). This set contains approximately 1,000 sentences curated both from English and Russian-centric news sources. Through the engagement of native speakers and professional translators[16], we partially translate this test set into 8 Turkic languages (Bashkir, Uzbek, Turkish, Kazakh, Kyrgyz, Azerbaijani, Karakalpak, and Sakha).

---

[11]http://www.nltk.org/api/nltk.tokenize.html
[12]https://github.com/kodchi/uzbek-transliterator
[13]http://www.transliteration.kpr.eu/kaa/

[14]https://github.com/NVIDIA/apex
[15]https://colab.research.google.com/
[16]A total of 22 volunteer translators

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

Table 3 highlights the currently available test set directions. Bolded entries in the table indicate the original direction of the translation. While Bashkir and Sakha have been translated by professional translators, other languages have been translated and validated (by another person) by proficient bilingual speakers of both the source and target language. The curation of this test set is an ongoing and growing effort currently covering 88 language directions.

5 Evaluation

Automatic evaluation metrics are very commonplace in MT research, and there has been a recent line of work exploring better metrics that capture translation quality beyond the syntactic and lexical features (Zhang et al., 2019; Sellam et al., 2020; Rei et al., 2020). Methods relying on contextual embeddings to capture the semantic similarity between the hypothesis and references fall short in terms of their language coverage. This is largely due to the pretraining of these evaluation models that require a significant of monolingual data which most of the low-resource languages lack. In this study, we evaluate our systems using both automatic metrics and human evaluation of translations.

5.1 Automatic Metrics for MT

We employ two widely adopted metrics: BLEU (Papineni et al., 2002) and ChrF (Popović, 2015). BLEU utilizes modified n-gram precision where the consecutive n-grams of the system translation are compared with the consecutive n-grams of the reference translation. We use the standard Sacre-BLEU implementation (Post, 2018). ChrF applies the same method at the level of character n-gram and we use the original implementation from the paper as provided through NLTK library.[17]

[17] https://github.com/m-popovic/chrF

5.2 Human Evaluation

To perform a more holistic analysis of MT systems, it is critical to involve native speakers in the evaluation process. We conducted a human evaluation campaign using a randomly sampled subset of 250 sentences from X-WMT or Bible (whenever X-WMT was not available) for evaluating the outputs of 14 bilingual baseline models. Our assessment is based on Direct Assessment (DA) test (Nießen et al., 2000; Papineni et al., 2002; Doddington, 2002), where annotators were asked to rate a translation according to *adequacy* and *fluency* on a 5 point Likert scale. All participants of the study were bilingual speakers of the source and target language. To better understand the importance of directionality (e.g. English-X vs X-English) and avoid variance in scores, we ensure that both directions of the same pair are evaluated by the same annotator (whenever possible). While reporting, we average the scores for each pair but report *adequacy* and *fluency* separately. Adequacy is defined as how much information is preserved in the translation. A score of 1 would mean that the translation is meaningless and has no correlation with the target sentence. A score of 5 would mean the translation retains all of the information. Fluency is defined as how grammatically, syntactically, and stylistically correct the translation is. A score of 1 would mean the sentence makes no sense grammatically or syntactically. A score of 5 would mean the sentence is perfectly correct.

6 Results & Discussion

The upper section of Table 4 highlights the bilingual baselines for high-resource pairs and their evaluation scores in the three domains. Despite the large training size, both models perform relatively modestly on the Bible and TED Talks with the *en-tr* model slightly better than *ru-tr*. Our hypothesis is that the domain of the Bible test set is far from the rest of the training set for both pairs, as most of the training data for Turkish comes from OpenSubtitles.[18] Another likely bottleneck is the suboptimal model size and hyperparameters, which were not tuned due to limited computational resources.

Baseline results for the mid- and low-resource pairs are in the lower part of Table 4. While there are a lot of fluctuations in the results, it is important to note the large disparities in BLEU scores

[18] http://www.opensubtitles.org/

Pair	Train size	Test size	Bible		Test size	Ted Talks		Test size	X-WMT	
			BLEU	ChrF		BLEU	ChrF		BLEU	ChrF
en-tr	39.9m	416	7.15	0.30	5.2k	12.32	0.43	800	19.87	0.51
ru-tr	16.8m	455	7.44	0.33	5.1k	8.64	0.38	800	8.81	0.41
ru-uz	1.22M	684	6.01	0.41	2.7K	4.51	0.76	800	5.95	0.39
uz-ru	1.22M	684	9.84	0.51	2.7K	7.57	0.73	800	7.45	0.37
en-az	784K	455	10.56	0.24	3.3K	10.58	0.29	600	8.88	0.41
az-en	784K	455	21.17	0.45	3.3K	17.01	0.17	600	12.14	0.42
en-ky	733K	451	6.47	0.32	-	-	-	500	3.18	0.19
ky-en	733K	451	13.08	0.43	-	-	-	500	4.30	0.40
tr-az	634K	606	13.78	0.65	3.6K	20.50	0.40	500	9.68	0.33
az-tr	634K	606	11.66	0.71	3.6K	24.20	0.95	500	11.53	0.49
en-kk	601K	453	3.62	0.61	3.6K	6.31	0.29	700	6.99	0.38
kk-en	601K	453	11.22	0.27	3.6K	9.78	0.30	700	9.75	0.46
en-uz	555K	465	5.23	0.40	3.2K	5.89	0.20	800	6.60	0.42
uz-en	555K	465	16.20	0.63	3.2K	11.61	0.18	800	12.32	0.48
tr-uz	161K	486	6.50	0.14	2.9K	4.28	0.20	700	1.58	0.23
uz-tr	161K	486	7.40	0.32	2.9K	3.92	0.26	700	1.73	0.22
kk-ky	6.4K	696	2.39	0.33	-	-	-	500	0.14	0.09
ky-kk	6.4K	696	2.53	0.24	-	-	-	500	0.11	0.13
en-krc	6.5K	374	5.57	0.25	-	-	-	-	-	-
krc-en	6.5K	374	11.57	0.22	-	-	-	-	-	-
kk-tt	7.7K	678	4.13	0.22	-	-	-	-	-	-
tt-kk	7.7K	678	3.75	0.17	-	-	-	-	-	-
ru-sah	8K	759	2.48	0.27	-	-	-	300	0.08	0.20
sah-ru	8K	759	2.44	0.23	-	-	-	300	0.31	0.16
uz-kaa	8.9K	772	9.90	0.71	-	-	-	300	5.39	0.41
kaa-uz	8.9K	772	9.58	0.60	-	-	-	300	5.24	0.44

Table 4: Bilingual baselines separated by high-res., mid-res., and low-res. pairs (K: thousand, M: million).

between models when translated in and out of non-Turkic languages. However, these differences are not as prominent when evaluated using ChrF, which is a character-level metric. This can partially be attributed to the complex morphology of Turkic languages which penalizes lexical mispredictions at a much higher rate than in English for example (Tantuğ et al., 2008). This in return would lead to lower BLEU scores. To examine this phenomena in more detail, we compare the results of X-WMT against human evaluations for the translations these models produced in Section 6.1.

Another notable aspect is the importance of scripts in the performance of the models. Language pairs with more than one script consistently underperform (both in automatic and human evaluations) the ones where both the source and target language use the same script. In fact, the best 6 models on the X-WMT test sets all have Latin scripts in both the source and target language. A suboptimal performance in the face of a script disparity is a known phenomenon (Anastasopoulos and Neubig, 2019; Murikinati et al., 2020; Aji et al., 2020; Amrhein

and Sennrich, 2020), where techniques such as transliteration show to improve performance. This is mostly attributable to model's inability to represent both languages in a shared space effectively when they do not share the same script, which can be damaging for the downstream performance.

6.1 Comparing Human Evaluations to BLEU

Using the Direct Assessment (DA) surveys described in Section 5.2, we obtain average scores of *adequacy* and *fluency* for almost all baseline models. Figures 1 show the scores for BLEU/ChrF and *adequacy*/*fluency* respectively. Comparing the scores from native speakers of these languages, it is quite evident that the disparities in BLEU scores between two translation directions are exaggerated and, even misleading (e.g. *en-az* vs *az-en*). Results in the human evaluations for mid-resource pairs seem a lot more closely clustered than in the BLEU/ChrF figure. These results further emphasize the pitfalls of automatic metrics of MT evaluation and emphasize the role of native speakers in the MT process.

Pair	Test size	Baseline		Google Translate		Yandex Translate		Apertium	
		BLEU	ChrF	BLEU	ChrF	BLEU	ChrF	BLEU	ChrF
en–tr	800	19.87	0.51	**69.24**	0.83	40.03	0.69	–	–
ru–tr	800	8.81	0.41	**24.79**	0.54	16.64	0.44	–	–
tr–uz	700	1.58	0.23	**27.25**	0.60	6.58	0.42	–	–
uz–tr	700	1.73	0.22	**28.03**	0.58	5.58	0.38	4.31	0.33
en–uz	800	6.60	0.42	**48.50**	0.72	15.66	0.51	–	–
uz–en	800	12.32	0.48	**32.35**	0.39	6.93	0.41	–	–
en–kk	700	6.99	0.38	**26.60**	0.55	5.51	0.39	–	–
kk–en	700	9.75	0.46	22.50	0.47	**23.2**	0.50	–	–
tr–az	500	9.68	0.33	**36.78**	0.65	5.53	0.38	–	–
az–tr	500	11.53	0.49	**32.67**	0.62	11.75	0.44	–	–
en–ky	500	3.18	0.19	**26.97**	0.56	5.21	0.36	–	–
ky–en	500	4.30	0.40	**21.66**	0.50	3.89	0.20	–	–
en–az	600	8.88	0.41	**78.54**	0.89	6.59	0.40	–	–
az–en	600	12.14	0.42	**39.42**	0.65	12.54	0.46	–	–
ru–uz	800	5.95	0.39	**22.26**	0.56	13.19	0.50	–	–
uz–ru	800	7.45	0.37	**19.00**	0.48	10.87	0.43	–	–
kk–tt*	678	4.13	0.22	**5.45**	0.35	1.58	0.24	2.77	0.28
tt–kk*	678	3.75	0.17	**5.44**	0.35	1.41	0.22	–	–
ru–sah	300	0.08	0.20	–	–	**8.27**	0.40	–	–
sah–ru	300	0.31	0.16	–	–	**24.93**	0.54	–	–
uz–kaa	300	5.39	0.41	–	–	–	–	**11.71**	0.42
kaa-uz	300	**5.24**	0.44	–	–	–	–	5.22	0.30
kk-ky	500	0.14	0.09	**20.56**	0.51	4.78	0.35	9.12	0.35
ky-kk	500	0.11	0.13	**20.57**	0.52	3.52	0.34	6.55	0.34

Table 5: Bilingual Baseline Compared to online MT Systems on X-WMT (Pair with * uses Bible Data).

	Adequacy		Fluency	
	BLEU	ChrF	BLEU	ChrF
Turkic	0.62	**0.71**	0.75	0.67
Non-Turkic	**0.75**	0.68	**0.83**	**0.86**

Table 6: Correlation between scores from human evaluation and automatic metrics for translating into Turkic and non-Turkic. Correlation is measured using Pearson's r.

6.2 Turkic Languages on the target side

Even though BLEU scores do not offer a holistic way to compare two MT systems, they are effective in telling which system performs better. As seen clearly from the results in Table 4, the performance of the baseline system as measured by the BLEU metric when translating into a Turkic language from English is substantially worse than when translating into English from a Turkic language. Translating into the Turkic language is typically twice as bad in terms of BLEU as translating from the Turkic language. The reliability of the BLEU score also decreases especially in the case of translating into morphologically-rich languages,

which has indeed been shown to correlate poorly with human judgments in Turkic languages (Ma et al., 2018, 2019). Table 6 shows the correlation between BLEU/Chrf and *adequacy/fluency* scores. BLEU seems to correlate with *adequacy/fluency* a lot better when the target side is a non-Turkic language, which emphasizes our earlier points regarding the language morphology. ChrF's correlation to *adequacy* scores is about the same regardless of the target language.

6.3 Comparison to Existing Systems

Table 5 compares our baselines to three commercial/open-source MT systems: Google Translate,[19] Yandex Translate,[20] and Apertium (Forcada et al., 2011). Google Translate results are significantly higher than our baselines and other MT systems. There are quite a few reasons for the score disparities. First, commercial systems have access to more data for training and possibly also include the public data we exclude from our test sets. Moreover, several test-set translators used Google Translate to do the translations and performed post-

[19] https://translate.google.com/
[20] https://translate.yandex.com/
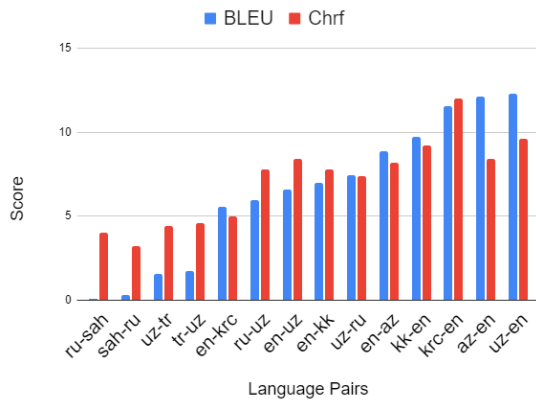
(a) BLEU and ChrF scores for select pairs. Note: ChrF scores were multiplied by 20 for better visibility.
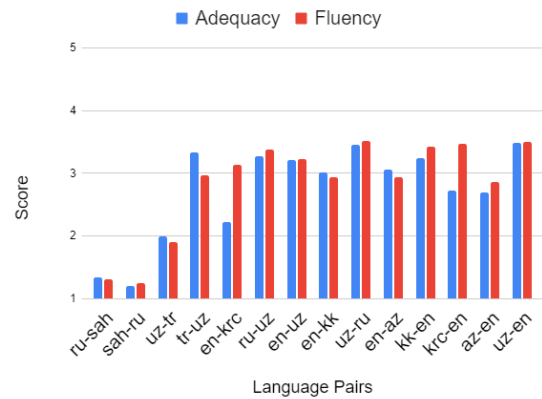
(b) Adequacy and Fluency scores (1-5) obtained from human evaluations.

Figure 1: Comparison between BLEU/ChrF scores and Adequacy/Fluency scores. Best viewed in color.

edits afterwards (e.g. *en-uz*) which creates a bias favoring sentences generated by Google's service. A safer comparison of the baselines is achieved with Yandex Translate, which despite the lower performance also supports more Turkic languages (8 in Google and 9 in Yandex). However, it is important to note that their API yielded worse results than their web interface. Apertium is a rule-based MT framework that supports several Turkic-Turkic pairs and we include the results whenever one is available. For those pairs, the results are comparable with our baselines and Yandex Translate.

7 Conclusion & Future Work

In this paper, we introduce a large parallel corpus covering 22 Turkic languages along with in-domain and out-of-domain evaluation sets. We also train the first baseline models for several language pairs and take the initial steps to address the challenges associated with machine translation in the Turkic languages. This study was carried out as in a participatory research setting by a diverse community of researchers, engineers, language specialists, and native speakers of Turkic languages. Future work will focus on studies of methods for effective cross-lingual transfer, extending of the coverage of the corpus to more languages and domains, and increasing the size of the test sets to provide more comprehensive benchmarks.

Acknowledgements

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.

Alham Fikri Aji, Nikolay Bogoychev, Kenneth Heafield, and Rico Sennrich. 2020. In neural machine translation, what does transfer learning transfer? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7701–7710.

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.

Chantal Amrhein and Rico Sennrich. 2020. On romanization for model transfer between scripts in neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 2461–2469.

Antonios Anastasopoulos and Graham Neubig. 2019. Pushing the limits of low-resource morphological inflection. *arXiv preprint arXiv:1908.05838.*

Ebrahim Ansari, Nguyen Bach, Ondřej Bojar, Roldano Cattoni, Fahim Dalvi, Nadir Durrani, Marcello Federico, Christian Federmann, Jiatao Gu, Fei Huang, et al. 2020. Findings of the iwslt 2020 evaluation campaign. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 1–34.

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.

Emily M Bender. 2011. On achieving and evaluating language-independence in nlp. *Linguistic Issues in Language Technology*, 6(3):1–26.

Alexandra Birch, Miles Osborne, and Philipp Koehn. 2008. Predicting success in machine translation. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 745–754, Honolulu, Hawaii. Association for Computational Linguistics.

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

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.

Emanuele Bugliarello, Sabrina J. Mielke, Antonios Anastasopoulos, Ryan Cotterell, and Naoaki Okazaki. 2020. It's easier to translate out of English than into it: Measuring neural translation difficulty by cross-mutual information. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1640–1649, Online. Association for Computational Linguistics.

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.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, pages 138–145.

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.

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.

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*.

Mikel L Forcada, Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, and Francis M Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine translation*, 25(2):127–144.

Alexander Fraser. 2020. Findings of the WMT 2020 shared tasks in unsupervised MT and very low resource supervised MT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 765–771, Online. Association for Computational Linguistics.

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.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. Two new evaluation datasets for low-resource machine translation: Nepali-english and sinhala-english. *CoRR*, abs/1902.01382.

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.

Lars Johanson and Éva Ágnes Csató Johanson. 2015. *The Turkic Languages*. Routledge.

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*.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the nlp world. *arXiv preprint arXiv:2004.09095*.

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.

Qingsong Ma, Ondřej Bojar, and Yvette Graham. 2018. Results of the wmt18 metrics shared task: Both characters and embeddings achieve good performance. In *Proceedings of the third conference on machine translation: shared task papers*, pages 671–688.

Qingsong Ma, Johnny Wei, Ondřej Bojar, and Yvette Graham. 2019. Results of the wmt19 metrics shared task: Segment-level and strong mt systems pose big challenges. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 62–90.

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.

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.

Sjur Moshagen, Trond Trosterud, Jack Rueter, Francis M. Tyers, and Tommi A. Pirinen. 2014. Open-source infrastructures for collaborative work on under-resourced languages. In *Proceedings of CCURL workshop 2014*.

Nikitha Murikinati, Antonios Anastasopoulos, and Graham Neubig. 2020. Transliteration for cross-lingual morphological inflection. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 189–197.

Sonja Nießen, Franz Josef Och, Gregor Leusch, Hermann Ney, et al. 2000. An Evaluation Tool for Machine Translation: Fast Evaluation for MT Research. In *LREC*.

Maimitili Nimaiti and Yamamoto Izumi. 2014. A Rule Based Approach for Japanese-Uyghur Machine Translation System. *International Journal of Software Science and Computational Intelligence (IJSSCI)*, 6(1):56–69.

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.

Atul Kr. Ojha, Valentin Malykh, Alina Karakanta, and Chao-Hong Liu. 2020. Findings of the LoResMT 2020 shared task on zero-shot for low-resource languages. In *Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages*, pages 33–37, Suzhou, China. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

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. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. 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.

Ye Qi, Devendra Singh Sachan, Matthieu Felix, Sarguna Janani Padmanabhan, and Graham Neubig. 2018. When and Why are Pre-trained Word Embeddings Useful for Neural Machine Translation?

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. *arXiv preprint arXiv:2009.09025*.

András Róna-Tas. 2015. Turkic Writing Systems. In Lars Johanson and Éva Ágnes Csató Johanson, editors, *The Turkic Languages*, chapter 6, pages 126–137. Routledge.

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

Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. BLEURT: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696*.

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

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.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958.

Miloš Stanojević, Amir Kamran, Philipp Koehn, and Ondřej Bojar. 2015. Results of the WMT15 metrics shared task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 256–273, Lisbon, Portugal. Association for Computational Linguistics.

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.

A. C. Tantug, E. Adali, and Kemal Oflazer. 2007. Machine translation between turkic languages. In *ACL*.

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

A Cüneyd Tantuğ, Kemal Oflazer, and Ilknur Durgar El-Kahlout. 2008. BLEU+: a Tool for Fine-Grained BLEU Computation. In *LREC*.

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

Jörg Tiedemann and Lars Nygaard. 2004. The OPUS Corpus-Parallel and Free: http://logos. uio. no/opus. In *LREC*. Citeseer.

Reut Tsarfaty, Dan Bareket, Stav Klein, and Amit Seker. 2020. From SPMRL to NMRL: What did we learn (and unlearn) in a decade of parsing morphologically-rich languages (MRLs)? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7396–7408, Online. Association for Computational Linguistics.

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.

5887

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*.

D. Varga, L. Németh, P. Halácsy, A. Kornai, V. Trón, and V. Nagy. 2005. Parallel corpora for medium density languages. In *Proceedings of the RANLP 2005*, pages 590–596.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All You Need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, volume 30, 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)*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. BERTscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

A Overall corpus statistics

Table 7 lists the training size (in sentences) for each language direction. It is important to note that the corpus is continuously growing and improving, so this version of the corpus was used for the bilingual baselines and human evaluations in this paper.

B Dataset Sources

Our parallel corpus is a combination of public resources and individual/group contributions. We list the sources for all the resources and websites used in curating our corpus in Table 8. More recent information on the licences and reuse of the corpus can be found in the Github repository[21].

---

[21]https://github.com/
turkic-interlingua/til-mt/tree/master/
replicate_results/corpus_paper

5888

	alt	az	ba	cjs	crh	cv	en	gag	kaa	kjh	kk	krc	kum	ky	ru	sah	slr	tk	tr	tt	tyv	ug	uum	uz
alt		9.8K	11.8K	48	6.8K	9.9K	11.3K	6.6K	6.8K	6.8K	6.8K	6.8K	6.7K	11.2K	6.8K	6.6K	61	9.7K	11.3K	11.3K	4.5K	6.8K		10.5K
az	9.8K		33.0K	48	6.9K	82.3K	787.6K	6.7K	6.9K	6.9K	8.0K	6.9K	6.7K	220.1K	389.0K	6.9K	52	123.4K	636.5K	215.1K	13.6K	6.9K		227.4K
ba	11.8K	33.0K			6.8K	34.4K	64.5K	6.6K	6.8K	6.7K	6.8K	6.8K	6.7K	35.9K	36.2K	6.6K		30.2K	65.4K	67.2K	16.5K	6.7K		27.5K
cjs	48	48			49		57	48	31	47	46			47	2.4K	48	44	47	50	50		53		48
crh	6.8K	6.9K	6.8K	49		6.7K	16.4K	6.8K	6.9K	7.0K	6.9K	6.9K	6.7K	6.7K	12.3K	6.7K	52	6.7K	13.9K	8.3K		7.0K		7.3K
cv	9.9K	82.3K	34.4K		6.7K		156.4K	6.5K	6.7K	6.7K	6.8K	6.8K	6.6K	83.1K	82.9K	6.7K		78.6K	158.2K	160.0K	15.1K	6.6K		59.9K
en	11.3K	787.6K	64.5K	57	16.4K	156.4K		6.8K	7.3K	7.0K	614.5K	7.0K	6.8K	625.6K	47.1M	7.1K	837	250.4K	39.9M	572.5K	29.0K	111.2K	507	559.7K
gag	6.6K	6.7K	6.6K	48	6.8K	6.5K	6.8K		6.7K	6.8K	6.7K	6.7K	6.6K	6.6K	6.7K	6.6K	55	6.5K	6.7K	6.6K		6.8K		7.1K
kaa	6.8K	6.9K	6.8K	31	6.9K	6.7K	7.3K	6.7K		6.8K	6.9K	6.9K	6.7K	6.7K	7.3K	6.7K	32	6.6K	6.9K	6.8K		6.8K		9.8K
kjh	6.8K	6.9K	6.7K	47	7.0K	6.7K	7.0K	6.8K	6.8K		6.8K	7.7K	6.7K	6.7K	6.8K	8.1K	52	6.6K	6.8K	6.7K		7.0K		7.3K
kk	6.8K	8.0K	6.8K	46	6.9K	6.8K	614.5K	6.7K	6.9K	6.8K		6.9K	6.7K	7.0K	4.5M	7.0K	48	6.6K	65.9K	8.4K		6.8K		124.9K
krc	6.8K	6.9K	6.8K		6.9K	6.8K	7.0K	6.7K	6.9K	7.7K	6.9K		6.8K	6.7K	6.9K	7.6K		6.6K	6.9K	6.8K		6.8K		7.3K
kum	6.7K	6.7K	6.7K		6.7K	6.6K	6.8K	6.6K	6.7K	6.7K	6.7K	6.8K		6.6K	6.8K	6.6K		6.5K	6.7K	6.6K		6.7K		7.1K
ky	11.2K	220.1K	35.9K	47	6.7K	83.1K	625.6K	6.6K	6.7K	6.7K	7.0K	6.7K	6.6K		309.3K	6.7K	50	122.5K	549.0K	232.4K	15.0K	6.7K		127.0K
ru	6.8K	389.0K	36.2K	2.4K	12.3K	82.9K	47.1M	6.7K	7.3K	6.8K	4.5M	6.9K	6.8K	309.3K		8.7K		122.5K	16.8M	296.7K		54.6K		1.2M
sah	6.6K	6.9K	6.6K	48	6.7K	6.7K	7.1K	6.6K	6.7K	8.1K	7.0K	7.6K	6.6K	6.7K	8.7K		51	6.5K	6.9K	6.9K		6.6K		7.3K
slr	61	52		44	52		837	55	32	52	48			50		51		56	59	32		59		48
tk	9.7K	123.4K	30.2K	47	6.7K	78.6K	250.4K	6.5K	6.6K	6.6K	6.6K	6.6K	6.5K	122.5K	122.5K	6.5K	56		244.1K	235.4K	14.0K	6.6K		7.2K
tr	11.3K	636.5K	65.4K	50	13.9K	158.2K	39.9M	6.7K	6.9K	6.8K	65.9K	6.9K	6.7K	549.0K	16.8M	6.9K	59	244.1K		531.2K	29.2K	64.4K		254.8K
tt	11.3K	215.1K	67.2K	50	8.3K	160.0K	572.5K	6.6K	6.8K	6.7K	8.4K	6.8K	6.6K	232.4K	296.7K	6.9K	32	235.4K	531.2K		15.0K	6.7K		136.3K
tyv	4.5K	13.6K	16.5K			15.1K	29.0K							15.0K				14.0K	29.2K	15.0K				10.8K
ug	6.8K	6.9K	6.7K	53	7.0K	6.6K	111.2K	6.8K	6.8K	7.0K	6.8K	6.8K	6.7K	6.7K	54.6K	6.6K	59	6.6K	64.4K	6.7K				19.5K
uum							507																	
uz	10.5K	227.4K	27.5K	48	7.3K	59.9K	559.7K	7.1K	9.8K	7.3K	124.9K	7.3K	7.1K	127.0K	1.2M	7.3K	48	7.2K	254.8K	136.3K	10.8K	19.5K		

Table 7: Parallel corpora size for each language pair.

Source	Link	Languages	Size
Tatoeba Challenge (OPUS+Tatoeba+ Gourmet+JW300)	`https://github. com/Helsinki-NLP/ Tatoeba-Challenge`	az, ba, crh, cv, gag, kjh, kk, krc, kum, ky, sah, tk, tr, tt, tyv, ug, uz, ru, en	∼40m
UDHR	`https://www.ohchr.org/EN/ UDHR/Pages/SearchByLang. aspx`	alt, ba, az, cv, cjs, crh, gag, kaa, kjh, kk, ky, sah, slr, tk, tt, tr, ug, uz, ru, en	∼100 per direction
Bible	`https://www. faithcomesbyhearing.com/ audio-bible-resources/ recordings-database`	alt, ba, az, cjs,cv, crh, en, gag, kaa, kjh, kk, ky, sah, tk, tt, ug, uz, tr	∼9k per direction
Ted Talks	`https://www.ted.com/ participate/translate/ our-languages`	az, en, kk, ky, ru, tt, tr, tt, uz, ug	∼600k
Mozilla		az, ba, cv, en, kk, ky, sah, tk, tt, ug, uz, tr, ru	∼300 per direction
Azerbaijani News	`https://github.com/ derintelligence/ en-az-parallel-corpus`	az, en	∼68k
Uzbek/English News	`https://data.gov.uz` `https://president.uz` `https://uz.usembassy.gov` `https://www.gov.uz`	uz, en	∼60k
Uzbekistan Legislative Dataset (Law)	`https://lex.uz/`	uz, ru, en	∼1.5m
KhanAcademy Project Translations(Math/Science)	`https://uz.khanacademy. org/`	uz, en	∼200k
Karakalpak News	`https://kknews.uz` `https://www.gov.uz` `http://karakalpakstan.uz` `https://www.qrstat.uz/kk`	kaa, uz, ru, en	∼60k
Bashkir-Russian Corpus	`https://github. com/AigizK/ bashkort-parallel-corpora`	ba,ru	∼600k
Salar Language Materials	`http://www.sino-platonic. org/complete/spp043_ salar_language.pdf`	slr,en	∼700
Urum Language Materials	`https://web.archive.org/ web/20180919233848/http: //projects.turkmas.uoa. gr/urum/`	urum, en	∼500
Russian-Shor Online Dictionary	`http://tili.tadarlar.ru/ tadar/rus-shor.html`	ru,cjs	∼300

Table 8: Sources and links for resources and websites used.