Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 910–918
Marseille, 20-25 June 2022

The Multilingual Microblog Translation Corpus: Improving and Evaluating Translation of User-Generated Text

Paul McNamee and Kevin Duh
Johns Hopkins University
Human Language Technology Center of Excellence
810 Wyman Park Drive, Baltimore MD 21211, USA
mcnamee@jhu.edu, kevinduh@cs.jhu.edu

Abstract

Translation of the noisy, informal language found in social media has been an understudied problem, with a principal factor being the limited availability of translation corpora in many languages. To address this need we have developed a new corpus containing over 200,000 translations of microblog posts that supports translation of thirteen languages into English. The languages are: Arabic, Chinese, Farsi, French, German, Hindi, Korean, Pashto, Portuguese, Russian, Spanish, Tagalog, and Urdu. We are releasing these data as the Multilingual Microblog Translation Corpus to support further research in translation of informal language. We establish baselines using this new resource, and we further demonstrate the utility of the corpus by conducting experiments with fine-tuning to improve translation quality from a high performing neural machine translation (NMT) system. Fine-tuning provided substantial gains, ranging from +3.4 to +11.1 BLEU. On average, a relative gain of 21% was observed, demonstrating the utility of the corpus.

Keywords: Machine Translation, Informal Language, Domain Adaptation, Evaluation Corpora

1. Introduction

Large collections of parallel text, or bitext, are increasingly available in many languages and in ever growing quantities. One well known project is OPUS, the Open Parallel Corpus, a web portal containing many open source parallel corpora which is maintained by Jörg Tiedemann (2012). It is hard to overstate the impact that comes from the wide availability of such data; it is now possible to build translation systems for language pairs that might have been infeasible just a few years ago. However, the largest sources of parallel data continue to be specialized domains such as governmental documents (*e.g.,* European Parliament) or religious texts. This creates an issue of domain mismatch, where the sources of training data are frequently unlike the types of data being translated by end users.

One genre where this is problematic is in user-generated texts found in online social media platforms (*e.g.,* Facebook, Reddit, and Twitter). Unlike copy-edited news, user-produced text has a much higher spelling error rate. The language used on these platforms is often vernacular vs. formal, and contains a high proportion of abbreviations and other shorthand notations. There are many peer-to-peer messages written in the first or second person point of view, unlike news or reports written in the third person. There is a higher percentage of non-linguistic or quasi-linguistic content such as user IDs, URLs, hashtags, and emoji. And the use of case and punctuation is often atypically applied for stylistic effect. These issues and many others contribute to domain mismatch, and as a result translation performance is degraded when a model trained on standard datasets is applied without modification. Common adaptation methods (such as fine-tuning) are often not applicable due to the lack of sufficient in-domain or in-genre bitext.

In this paper we describe our efforts to create a microblog translation corpus that is suitable for evaluating the quality of machine translation systems on user-generated texts, and additionally can be used to adapt out-of-domain models for use in microtext translation. Our corpus contains over 200,000 manual translations covering thirteen language-pairs, and our goal is to create a benchmark that enables research in this domain.

In addition to creating this valuable resource, we investigate the following questions:

- What are the best practices and guidelines for manual translation of microblogs?
- What improvements in translation quality can be realized using domain adaptation methods, and how much data is required to obtain tangible improvements?

In Section 2 we present the Multilingual Microblog Translation Corpus (MMTC). In Section 3 we describe development of the corpus, including some of the annotation challenges we faced. Section 4 presents the guidelines used in this translation task. We then describe experimental results using the corpus in Section 5. Section 6 highlights related work. Finally, in Section 7 we summarize our findings.

2. Corpus Description

The Multilingual Microblog Translation Corpus contains microblog posts from Twitter (or tweets) in thirteen source languages that have been translated to American English. The source languages are Arabic, Chinese, Farsi, French, German, Hindi, Korean, Pashto, Portuguese, Russian, Spanish, Tagalog, and Urdu.

Language	*Tweets*	*Src Len*	*Eng Len*	*% live*	*% RT*	*% handle*	*% hash*	*% emoji*	*% URL*
Arabic	24,634	77.4	108.4	75.5	40.7	30.1	5.8	1.5	0.23
Chinese	5,580	55.6	173.7	50.7	35.4	33.7	0.6	2.3	7.54
Farsi	6,647	75.8	95.3	64.3	27.6	43.9	0.9	1.0	1.58
French	6,485	62.4	60.5	50.7	33.1	38.8	0.1	2.9	0.22
German	2,418	81.3	81.3	88.9	26.0	66.7	0.3	3.7	0.74
Hindi	901	93.5	106.7	87.5	41.8	45.6	0.2	2.7	0.11
Korean	39,225	38.7	83.4	51.4	14.9	25.8	0.1	0.1	0.01
Pashto	1,390	80.2	87.1	79.7	20.9	11.8	2.0	4.1	4.46
Portuguese	6,085	56.0	57.6	43.9	26.4	15.7	0.0	1.6	0.05
Russian	42,734	69.5	75.4	65.1	18.5	35.9	0.6	1.8	0.46
Spanish	62,247	59.8	62.9	75.6	30.9	36.3	0.2	2.5	1.07
Tagalog	5,066	61.9	63.6	64.4	21.5	19.9	0.6	4.6	0.16
Urdu	5,118	74.0	84.2	78.4	35.4	29.4	0.7	2.0	9.22

Table 1: Data size and the relative frequency of certain phenomena in the released corpus, by language. Columns are: the number of tweets that are manually translated, the average length of the source tweet and English translation, in characters; percent of tweets still accessible on 12/29/21; percent of retweets; percent with at least one user handle; percent with at least one hashtag; percent containing emoji; and, the percent with a recognizable URL. Retweet marks and user handles are quite prevalent, occurring in about a third of messages. Hashtags, emoji, and URLs are only found in a few percent of the released corpus.

The data size and frequency of various phenomena in the corpus (*e.g.,* retweets, emoji) are presented in Table 1. Most language-pairs have over 5000 manual translations, which is sufficient to support research in fine-tuning and domain adaptation. Our largest language pairs ({Arabic, Korean, Russian, Spanish}→English) each have tens of thousands of manual translations. The distribution of lengths of the source language tweets is shown in Figure 1.

Partitions for held-out evaluation ("test") are available in each language. In most languages there is data for training or adaptation, labeled "train", and data for parameter tuning labeled "valid". See Section 3.5.

Translations are released along with associated tweet IDs; original tweets can be obtained using the Twitter API.[1]

3. Corpus Development

We describe the development of the corpus below. Statistics pertaining to each stage of the pipeline are shown in Table 2.

3.1. Selection

Messages were sampled from dates in certain intervals between April 2016 and October 2021. The temporal span of the corpus is graphically depicted in Figure 2. No attempt was made to select tweets with specific content, for example, containing specific hashtags or named entities, or mentioning particular events. Selection was random, but subject to the filtering described below.

3.2. Language Identification & Filtering

After messages were downloaded, automated language identification (LID) was performed using a tool based on prediction by partial matching (McNamee, 2016), a compression-inspired use of language modeling. Then several additional steps were taken to control the properties of the selected tweets. Post-LID checks were undertaken to ensure the character sets matched expectations, messages with URLs were undersampled[2] so that they remained a minority, and short messages (*e.g.,* less than 20 characters) were avoided. Whitespace was normalized to a single space character. Additionally a shingling-based method of near-duplicate detection was used to avoid exact duplicates and nearly identical content (*i.e.,* as happens with popular retweets). Messages that passed the previous checks were then shuffled randomly and formed into batches for translation.

3.3. Translation

Single reference translations of the tweets, into English, were created by professional translators. Five messages at a time were presented in a web-based annotation tool (see Figure 3). A link to the original post was available in case viewing the original or surrounding messages could supply helpful context. There was also a button to bring up the translation guidelines, described in Section 4. Translators were asked to confirm the source language, and to indicate a degree of confidence in the translation.

Translators were told that they could skip translation of a tweet for any reason. As these are randomly sampled social media content, there are not-safe-for-work (NSFW) posts that contain vulgar or offensive content. If a translation was performed, profanity should be in-

[1] `https://developer.twitter.com/en/docs/labs/tweets-and-users/quick-start/get-tweets`

[2] 15% of messages containing a URL were retained.

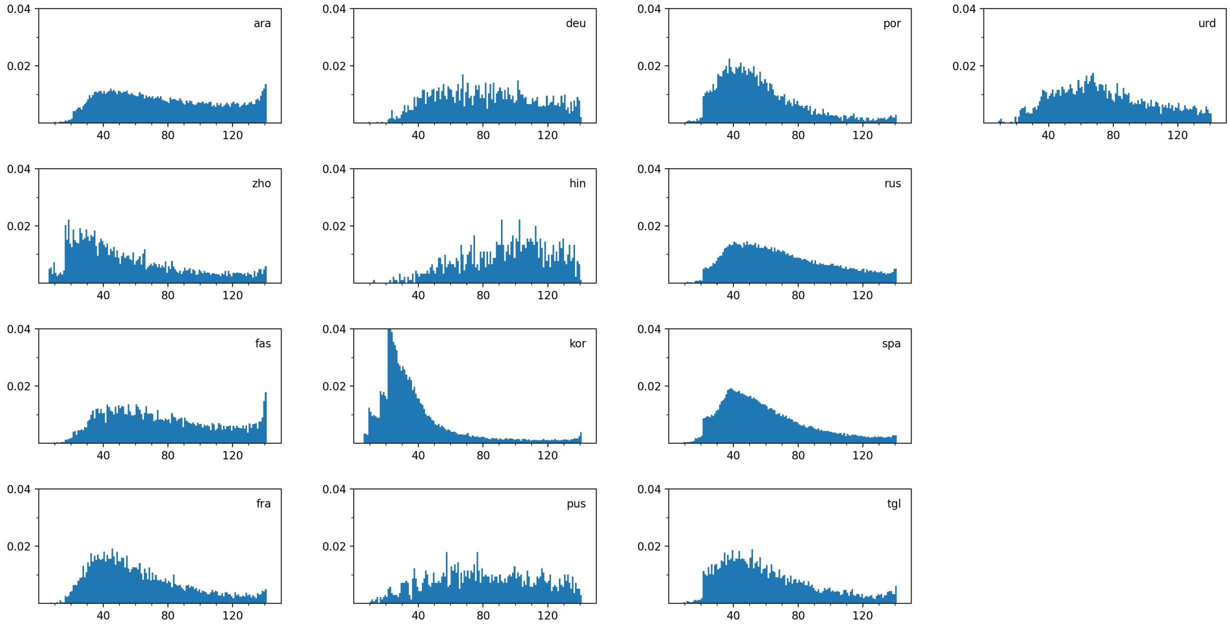


Figure 1: Probability distribution of tweet lengths (characters), by language.

Language		*Start*	*Skip'd*	*Lang?*	*Confidence: H/R/L*			*Final*	*Rev'd*	*Edit%*
Arabic	ara	28,103	2,628	6	17,284	7,350	835	24,634	8,862	74.3%
Chinese	zho	8,068	2,142	30	3,351	2,229	316	5,580	1,584	76.7%
Farsi	fas	8,746	1,668	22	4,566	2,081	409	6,647	—	—
French	fra	7,153	519	2	5,339	1,146	147	6,485	3,652	29.5%
German	deu	2,494	7	3	1,739	679	66	2,418	—	—
Hindi	hin	911	9	1	888	13	0	901	—	—
Korean	kor	47,587	5,951	12	21,956	17,269	2,399	39,225	20,559	51.9%
Pashto	pus	6,982	5,258	104	711	679	230	1,390	—	—
Portuguese	por	6,173	44	5	5,740	345	39	6,085	—	—
Russian	rus	45,780	1,680	24	29,500	13,234	1,342	42,734	26,159	36.4%
Spanish	spa	65,785	2,059	33	49,594	12,653	1,446	62,247	38,982	34.4%
Tagalog	tgl	6,198	921	173	4,101	965	38	5,066	—	—
Urdu	urd	6,090	829	7	3,484	1,634	136	5,118	—	—
Totals		240,070	23,715	422	148,253	60,277	7,403	208,530	99,798	42.5%

Table 2: Statistics related to the creation of the Multilingual Microblog Translation Corpus. From left to right the columns are: the number of messages examined; the number skipped by translators; the number tagged as a different language than expected; the number given a confidence of High, Reasonably, or Low; the final number of messages in each language; the number that received a secondary review; and, the percentage of those additionally reviewed that were changed in any way.

cluded and spelling should be correct. Other reasons for skipping included lack of context to reliably translate (*e.g.,* missing anaphors), ambiguity caused by misspellings, or foreign text from a different language in the post. Approximately 10% of presented tweets were skipped. 96.6% of translations were marked highly or reasonably confident. Those marked as being of low confidence were not included in the corpus. Table 3 contains a sample of comments from the translators.

3.4. Secondary Review

In some languages we were able to perform an additional review for a portion of the translations. This step was undertaken by our most experienced translators, and their edits, if any, were used in the released corpus. The number of additionally reviewed translations, and how often they were edited is reported in Table 2. In languages with additional review, edits to first pass translations ranged from a low of about 30% in French to 77% in Chinese.

3.5. Partitioning

Priority was given to placing live tweets in the *test* and *valid* partitions to maximize use of the corpus for evaluation purposes. We believe the higher quality translations are those that underwent secondary review, followed by those translations that identified as "highly confident" by the translators. Therefore we also priori-

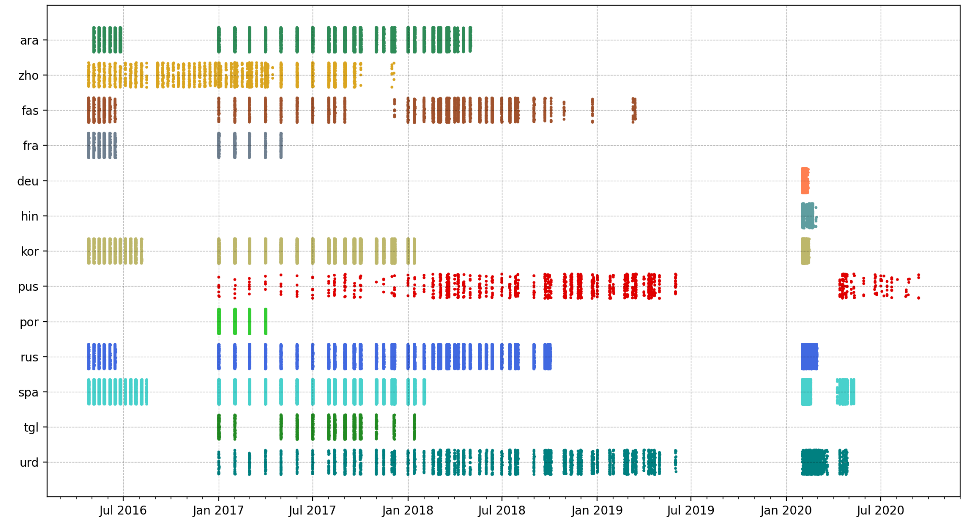
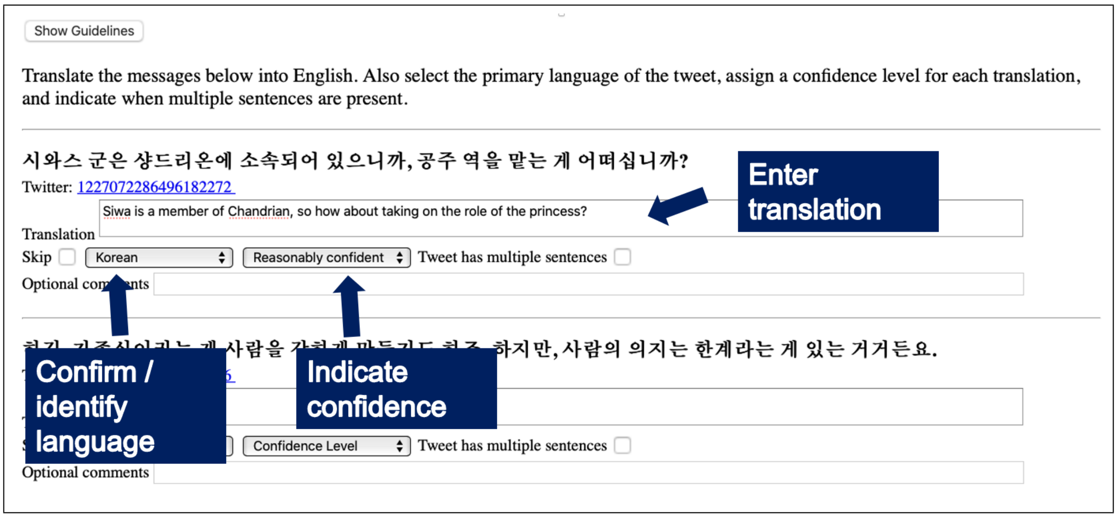


Figure 2: Temporal distribution of tweets, by language.



Translate the messages below into English. Also select the primary language of the tweet, assign a confidence level for each translation, and indicate when multiple sentences are present.

시와스 군은 샹드리온에 소속되어 있으니까, 공주 역을 맡는 게 어떠십니까?

Twitter: 1227072286496182272

Translation: Siwa is a member of Chandrian, so how about taking on the role of the princess?

Skip □ Korean Reasonably confident Tweet has multiple sentences □

Optional comments

하지만, 사람의 의지는 한계라는 게 있는 거거든요.

Confidence Level Tweet has multiple sentences □

Optional comments

Figure 3: Annotation interface for microblog translation. Translators worked with sets of five messages at a time.

1	The Arabic word for friend has extra vowels in it which could be stylistic or a mistake.
2	Because we don't have the subject it could be he, it, she or they are working in English.
3	Mateo Kovačić is a Croatian professional soccer player; Zinedine Yazid Zidane (aka Zizou), is a French former professional soccer player
4	Many grammatical errors; translation is approximate
5	The person is talking about watching a tv show called "the black pearl"
6	The word referee is misspelled.
7	I'm not sure about the cultural connotation of what is meant by 'white' days. Reading the twitter thread from which this was extracted, I can guess that it possibly means 'bleak'. This is in Egyptian dialect.
8	A misyar marriage is a form of marriage that exists in some middle-eastern countries. It can't really be translated.
9	长坂坡七进七出 references an event from Romance of the Three Kingdoms.
10	"Pepper" is referring to "Rebel Pepper" (变态辣椒), a Chinese political cartoonist in exile in the United States.
11	The word haharot is a slang word and could mean a number of different things depending on the age of the speaker and the context, which isn't provided here.
12	одногр is a shortened slang term for odnogrupnik, meaning classmate
13	bac is short for Baccalauréat and is a test that French students complete in order to graduate high school and establish eligibility for university programs. Also, QCM is "questionnaire à choix multiples" or "multiple choice quiz"
14	죽 쒀서 개 준다 is an idiom equivalent to working hard only to get no credit for it and it's given to someone else.
15	시클로로, reference to hydrochloroquine?

Table 3: A selection of comments provided by our translators.

	Test		*Valid*		*Train*	
Language	*N*	*live*	*N*	*live*	*N*	*live*
Arabic	3,000	100	3,000	100	18,634	68
Chinese	2,000	100	2,000	41	1,580	0
Farsi	2,000	100	2,000	100	2,647	10
French	2,000	100	2,000	36	2,485	0
German	2,000	100	418	36	—	—
Hindi	901	88	—	—	—	—
Korean	3,000	100	3,000	100	32,225	43
Pashto	1,390	80	—	—	—	—
Portuguese	2,000	100	2,000	34	2,085	0
Russian	3,000	100	3,000	100	36,734	59
Spanish	3,000	100	3,000	100	56,247	73
Tagalog	2,000	100	2,000	63	1,066	0
Urdu	2,000	100	2,000	100	1,118	0

Table 4: Size of partitions, including percentages of tweets live on 12/29/21.

tized placement of these presumed higher quality translation in the *test* partition, and only if required included those translations marked “reasonably” confident.

Table 4 describes the partitions. The *test* partition contains a high percentage of live tweets, which we hope will mean that these data will be useful for some time to come. The four languages with the greatest number of translations are: Arabic, Korean, Russian, and Spanish. In these languages we used 3,000 translations for the test and validation partitions; in other languages 2,000 were used.

4. Translation Guidelines

We consulted materials produced by the LDC on the DARPA BOLT Program (Linguistic Data Consortium, 2012), which created translations of chat and SMS messages in Chinese and Egyptian Arabic. While in a similar genre, we needed to address phenomena common in online microtexts, so we adapted and augmented the BOLT guidelines for our purposes. Due to our work in many languages we endeavored to keep the guidance as language-agnostic as possible. We summarize the guidelines below. Section 4.1 lays out the most important principles and the remainder of Section 4 addresses specific issues. If atypical situations arose the translators were instructed to seek clarification.

4.1. General Principles

- Translate the source tweet into natural-sounding English; when possible prefer literal translations.

- The English translation should be faithful to the original text in terms of meaning and style.

- Do not put parenthetical comments in the text of the translation. A box is available for comments.

- Please do not use machine translation tools (*e.g.,* Google/Bing) and then edit the translation. However, it is okay to use dictionaries, or translation sites to lookup unknown terminology.

- The translation should not add or delete information, and no amplifying comments should be included. For example, if the German Chancellor is referred to simply as “Merkel”, do not render the translation as “Chancellor Merkel” or “Angela Merkel”.

- If necessary, it is okay to skip a message due to difficulty, vulgar content, garbled input, etc...

- Preserve punctuation, stylized casing, @userids, RT marks, emoji/emoticons, URLs, and hashtags. Copy/paste may help avoid error.

- Review translations before submitting to avoid incorrect spellings or other minor errors.

4.2. Proper Names

If there is an existing conventional translation of a name in English, it should be used. Also, if a foreign proper name comes from a different source language, use a direct translation to English. For example the Russian word Венеция should be translated as Venice and not transliterated as say Venetsiya, which is closer to the Italian pronunciation. When encountering an unfamiliar name, please consult a standard reference or search the Web to identify the standard or most commonly used form.

When a name appears with a title that may seem a little awkward in English (*e.g.,* Comrade Zhang or Brother Zhang), it should not rendered as Mr. Zhang. The literal translation is preferred.

If a Roman-script name has accented characters (*e.g.,* México) use the dominant English form (*e.g.,* Mexico). But it is okay to retain accents in names that are commonly written with them (*e.g.,* Bogotá, Simón Bolívar).

4.3. Capitalization

If capitalization rules are intentionally ignored follow the author’s style. Otherwise use standard rules for written English unless there is strong evidence (or accepted usage) that suggests a different treatment (*e.g.,* iPhone).

4.4. Punctuation

For grammatical and formal text, use standard rules for punctuation. But in informal texts (*e.g.,* tweets) atypical punctuation may be common, and when present it should be preserved. However, non-English punctuation marks should be avoided (*e.g.,* ¿ or 。).

4.5. Foreign Text

English text in the source sentence should generally be preserved and left untranslated, without making any change or correction. This may occasionally result in a strange looking translation where a name or concept appears twice. For example, “亚西尔·阿拉法特 (Yasser Arafat) 1929年8月出生于耶路撒冷” should be translated as “Yasser Arafat (Yasser Arafat) was born in Jerusalem in August 1929”

4.6. Errors in the Source Text

Misspellings or improper use of homophones (*e.g.,* their/there) should be corrected in the translation.

4.7. Dropped Components

Some languages allow components to be dropped due to context (*e.g.,* Chinese and Korean allow for dropped subjects and pronouns). This could result in the loss of meaning or an ungrammatical English sentence. In such cases provide the missing subject or object.

4.8. Ambiguity

Due to insufficient context aspects about the source sentence may be unclear, such as, the tense of a verb or the gender of a subject or object. In such cases, it is fine to make a best guess to resolve the issue. It is okay to choose a translation that could be correct, even if you are unsure about the author's intent. And if uncomfortable with the translation you can also choose to mark it "Not Confident" or skip it.

4.9. Special Cases for Informal Texts

Translations should capture and express the same degree of emotion and formality as the source text. Sometimes informal abbreviations will be found in social media language (*e.g.,* IDK, LOL, BRB, etc...). Try to select an equivalent in English. Emoticons and Emoji should be copied over to the English translation with cut-and-paste. URLs, user IDs, and retweet marks, should also be copied over. Hashtags, if written in non-Roman script, should be translated matching the author's style to the extend possible (*e.g.,* CamelCase, UPCASE, split_by_underscores, *etc.*). Profanity should be translated and spelled correctly as opposed to being censored; however, it is always permissible to skip a disturbing message.

4.10. Confidence Levels

Each translation should be assigned a confidence:

- *Highly Confident*: Good understanding of the source and confident in the English translation.
- *Reasonably Confident*: Good understanding of the source, but you might have uncertainty about a word or some other minor concern.
- *Not Confident*: The source text may be unclear. Important context may be missing. The vocabulary may be too challenging. You're not sure the translation is reliable.

5. Experiments

Our goals here are to: (a) establish benchmark results on the MMTC test set; and, (b) understand how the addition of our microblog train/valid data for domain adaptation can improve results. All neural machine translation models were trained with the transformer (Vaswani et al., 2017) using Amazon's Sockeye (v2) toolkit (Hieber et al., 2020). Key hyperparameters include: use of 6 layers in both encoder and decoder; 1,024 dimensional embeddings; 16 attention heads; 4,096 hidden units per layer; 30,000 subword byte pair encoding (BPE) units in both source and target languages; batch size of 1,024; the Adam optimizer with an initial learning rate of 2×10^{-4}.

Source	*General*	*BLEU on test*	
Language	*Data*	*flores101*	*mmtc*
Arabic	63.1	42.7	45.2
Chinese	84.8	31.5	28.0
Farsi	11.4	35.1	26.7
French	279.1	45.5	58.0
German	233.9	43.8	30.7
Hindi	8.7	35.2	23.9
Korean	15.0	29.3	23.2
Pashto	1.3	13.5	6.9
Portuguese	100.4	50.0	39.5
Russian	119.9	34.9	51.6
Spanish	65.5	29.5	51.7
Tagalog	6.3	40.3	26.3
Urdu	1.7	23.2	27.8

Table 5: General domain training data used (millions of lines of bitext), and baseline translation model performance on *flores101* benchmark, and the MMTC test partition, as BLEU scores.

5.1. Baselines

For our baseline, we trained transformers on large amounts of general-domain bitext obtained from OPUS; these do not contain microblog text in sufficient quantities. Table 5 reports the amount of training bitext used and baseline model performance on a general purpose benchmark dataset, and on the MMTC test partition. *flores101* (Goyal et al., 2021), which consists of news and travel texts from Wikipedia, is the general purpose benchmark. Translation performance is reported using case-insensitive BLEU scores (Papineni et al., 2002) calculated with *sacrebleu* (Post, 2018).[3]

5.2. Fine-Tuned Models (Adaptation)

To investigate whether using additional training data from MMTC can improve translation model performance on the held-out *test* partition, we conduct fine-tuning. We adapt the baseline models using the *valid* partition for validation, and incorporating varying amounts of data from *train*, ranging from 1,000 to 50,000 additional tweets. Figure 4 shows the improvement obtained with fine tuning (left), and the additional gains possible when greater amounts of data are available (right). The majority of the improvement attained comes from fine-tuning with only one or two thousand examples. However, in languages where more data is available, performance continues to rise, but at a much slower rate.

Relative improvements in BLEU score are reported in Table 6. Fine-tuning was effective in every lan-

[3] BLEU+case.lc+numrefs.1+smooth.exp+tok.13a+version.1.4.14

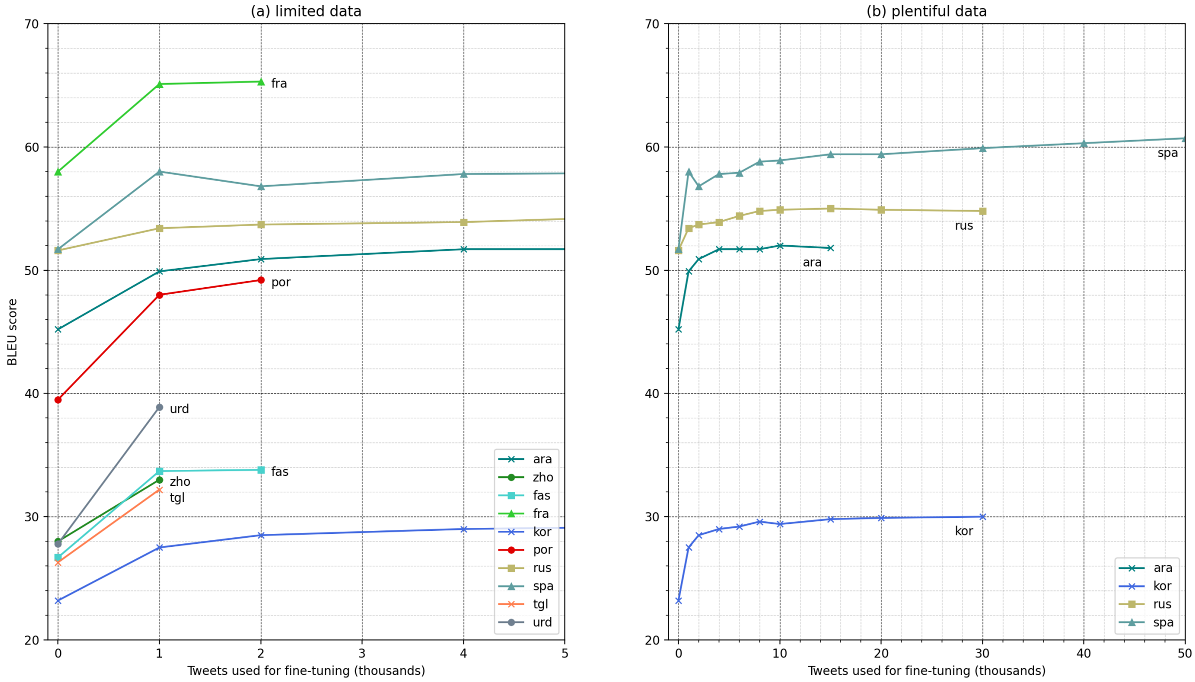


Figure 4: Gain in performance on the MMTC test partition using increasing amounts of data for fine-tuning. Baseline models (at zero) use no tweets for adaptation. Plot (a) focuses on adaptation when limited data is available; Plot (b) extends these curves for the four languages with 10,000 or more tweets available. Strong gains are observed with even small amounts of data, and when larger amounts are used performance grows further but at a slower rate.

Language	*Base*	*FT – 1,000*		*FT – Best*	
Arabic	45.2	49.9	+10.4%	52.0	+15.0%
Chinese	28.0	33.0	+17.8%	33.0	+17.8%
Farsi	26.7	33.7	+22.4%	33.8	+22.8%
French	58.0	65.1	+12.2%	65.3	+12.6%
Korean	23.2	27.5	+18.5%	30.0	+29.3%
Portuguese	39.5	48.0	+21.5%	49.2	+24.6%
Russian	51.6	53.4	+3.5%	55.0	+6.6%
Spanish	51.7	58.0	+12.2%	60.7	+17.4%
Tagalog	26.3	32.3	+22.8%	32.3	+22.8%
Urdu	27.8	38.9	+39.9%	38.9	+39.9%
$\bar{x}$			+18.1%		+20.9%

Table 6: Gains in BLEU score from fine-tuning. 1,000 examples is sufficient for substantial improvement.

guage. Gains over the baseline ranged from +3.4 BLEU (+6.6%) in Russian to +11.1 BLEU (+39.9%) in Urdu. On average a 20.9% relative gain in BLEU was observed. Even when using just 1,000 examples for fine-tuning, an average relative gain of 18.1% in BLEU score was achieved.

Overall we see that the fine-tuned models are remarkably improved on the microblog texts; this demonstrates the utility of the corpus. In Table 7 we present a couple of example translations produced by the baseline and fine-tuned models.

5.3. Domain Differences

Finally, we seek to quantify how distinct the microblog texts in MMTC are compared to general-domain data. Optimizing performance in a new domain using fine-tuning can degrade performance on general domain datasets. To assess this, we translate the *flores101* benchmark again using adapted models fine-tuned with 1,000 exemplars. We plot the results in Figure 5, where both baseline and fine-tuned models are compared on the *flores101* and *mmtc* datsets. The vertical axis of Figure 5 measures BLEU scores on *mmtc* while the horizontal axis measures BLEU for *flores101*. In every language, the shift is upwards and to the left, which is indicated with arrows. On average, the 1k fine-tuned models obtain an +18.1% gain on *mmtc* in exchange for a -5.8% decline on *flores101*. This confirms that there are substantial differences in linguistic phenomena or terminology found in the *mmtc* data compared to general-domain texts.

6. Related Work

Table 8 compares existing bilingual corpora for informal text. Automatically mined collections tend to be larger, but they favor certain kinds of content (*e.g.*, public announcements on Twitter, bilingual speakers). Corpora that are built by manual translation tend to be limited in size, with the exception of the DARPA BOLT packages in Arabic and Chinese produced by the LDC.

Src	RT @Guarromantico_: Los que usan el filtro de perrito, ¿también se asustan con los cuetes?
Ref	RT @Guarromantico_: Those that use the puppy filter, Are they scared of fireworks too?
Base	RT @Guarromantico_: Do those who use the doggy filter also get scared with the cooties?
FT	RT @Guarromantico_: Do those who use the doggie filter, also get scared with the kicks?
Src	其实吧，真心觉得考研就是一个错误的决定。不过既然已经上了这条贼船。那就走一步是一步吧！
Ref	Actually, I honestly feel like it was a mistake to go to graduate school. But of course, I've already boarded this pirate ship. Just got to take it one step at a time!
Base	In fact, I sincerely feel that the entrance examination is a wrong decision. But since I have been on this ship. Then step by step is a step!
FT	Actually, I truly believe that taking an entrance examination is a wrong decision. However, since I have already got on this pirate ship, then taking a step is a step!

Table 7: Example translations. Shown are the source tweet, the human reference, the baseline model translation, and the fine-tuned translation (w/1,000 examples).

Corpus	*Domain*	*Language*	*Size*
Manually Translated			
(Michel and Neubig, 2018)	Reddit	eng,fra,jpn	75k
(Sluyter-Gäthje et al., 2018)	Microblog	deu, eng	4k
(Munro, 2010)	SMS	hat, eng	40k
LDC2021T15 (BOLT)	SMS	arz, eng	160k
LDC2021T11 (BOLT)	SMS	zho, eng	200k
MMTC (this work)	Microblog	13 languages + eng	200k
Automatically Mined			
(Mubarak et al., 2020)	Microblog	ara, eng	170k
(Ling et al., 2013)	Microblog	zho, eng	1000k
(Vicente et al., 2016)	Microblog	spa, eus, cat, glg, por	20k

Table 8: Comparison of bilingual corpora for informal text. Size is the combined number of translations.

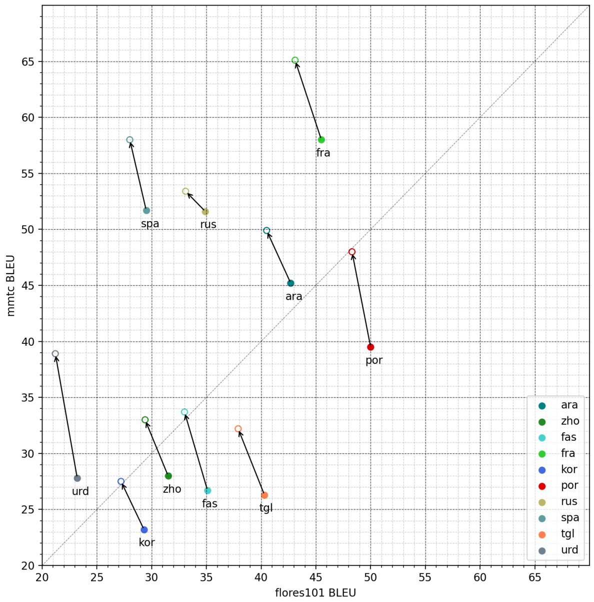


Figure 5: Baseline models are indicated with solid circles and the corresponding fine-tuned model using 1,000 examples is shown with a hollow circle. In all cases fine-tuning raises the BLEU score on the tweet corpus (vertical axis) and has a concomitant shift to the left, showing a mild reduction on the news/travel/crime *flores101* dataset (horizontal axis).

7. Conclusion

We presented a new corpus of microblog translations that we are sharing with the research community.[4] The collection contains over 200,000 translations in thirteen languages. We report first baselines for this corpus using neural machine translation models trained on open source texts. We additionally demonstrated the utility of the corpus for adapting trained models to microblog text using fine-tuning. While our intent was to build a corpus of microblog translations, the corpus may be useful for other purposes, for example, for adapting other types of online texts, or as a dataset for automated language identification of short texts.

8. Acknowledgment

We appreciate the diligent work of the many translators who supported us in developing the Multilingual Microblog Translation Corpus.

[4] The MMTC corpus can be obtained from: `https://pmcnamee.net/research/mmtc/mmtc.html`

9. Bibliographical References

Goyal, N., Gao, C., Chaudhary, V., Chen, P., Wenzek, G., Ju, D., Krishnan, S., Ranzato, M., Guzmán, F., and Fan, A. (2021). The FLORES-101 evaluation benchmark for low-resource and multilingual machine translation. *CoRR*, abs/2106.03193.

Hieber, F., Domhan, T., Denkowski, M., and Vilar, D. (2020). Sockeye 2: A toolkit for neural machine translation. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 457–458, Lisboa, Portugal, November. European Association for Machine Translation.

Ling, W., Xiang, G., Dyer, C., Black, A., and Trancoso, I. (2013). Microblogs as parallel corpora. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 176–186, Sofia, Bulgaria, August. Association for Computational Linguistics.

Linguistic Data Consortium. (2012). BOLT program annotation guidelines. Technical report, University of Pennsylvania. Accessed from: https://www.ldc.upenn.edu/collaborations/current-projects/bolt/translation.

McNamee, P. (2016). Language and dialect discrimination using compression-inspired language models. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*, pages 195–203, Osaka, Japan, December. The COLING 2016 Organizing Committee.

Michel, P. and Neubig, G. (2018). MTNT: A testbed for machine translation of noisy text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 543–553, Brussels, Belgium, October-November. Association for Computational Linguistics.

Mubarak, H., Hassan, S., and Abdelali, A. (2020). Constructing a bilingual corpus of parallel tweets. In *Proceedings of the 13th Workshop on Building and Using Comparable Corpora*, pages 14–21, Marseille, France, May. European Language Resources Association.

Munro, R. (2010). Crowdsourced translation for emergency response in Haiti: the global collaboration of local knowledge. In *Proceedings of the Workshop on Collaborative Translation: technology, crowdsourcing, and the translator perspective*, Denver, Colorado, USA, October 31. Association for Machine Translation in the Americas.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (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, July. Association for Computational Linguistics.

Post, M. (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, October. Association for Computational Linguistics.

Sluyter-Gäthje, H., Lohar, P., Afli, H., and Way, A. (2018). FooTweets: A bilingual parallel corpus of world cup tweets. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Tiedemann, J. (2012). Parallel data, tools and interfaces in opus. In *LREC*, volume 2012, pages 2214–2218.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In I. Guyon, et al., editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Vicente, I. S., Alegria, I., España-Bonet, C., Gamallo, P., Oliveira, H. G., Garcia, E. M., Toral, A., Zubiaga, A., and Aranberri, N. (2016). Tweetmt: A parallel microblog corpus. In *LREC*.