The Development of a Comprehensive Data Set for Systematic Studies of Machine Translation

Jörg Tiedemann[1][0000−0003−3065−7989]

University of Helsinki, Department of Digital Humanities
P.O. Box 24, FI-00014 Helsinki, Finland
jorg.tiedemann@helsinki.fi

Abstract. This paper presents our on-going efforts to develop a comprehensive data set and benchmark for machine translation beyond high-resource languages. The current release includes 500GB of compressed parallel data for almost 3,000 language pairs covering over 500 languages and language variants. We present the structure of the data set and demonstrate its use for systematic studies based on baseline experiments with multilingual neural machine translation between Uralic languages and other language groups. Our initial results show the capabilities of training effective multilingual translation models with skewed training data but also stress the shortcomings with low-resource settings and the difficulties to obtain sufficient information through straightforward transfer from related languages.

Keywords: machine translation · low-resource languages · multilingual NLP

1 Introduction

Massively parallel data sets are valuable resources for various research fields ranging from cross-linguistic research, language typology and translation studies to neural representation learning and cross-lingual transfer of NLP tools and applications. The most obvious application is certainly machine translation (MT) that typically relies on data-driven approaches and heavily draws on aligned parallel corpora as their essential training material.

Even though parallel data sets can easily be collected from human translations that naturally appear, their availability is still a huge problem for most languages and domains in the world. This leads to a skewed focus in cross-linguistic research and MT development in particular where sufficient amounts of real-world examples of reasonable quality are only available for a few well-resourced languages. The success of modern neural machine translation (NMT) is amazing, however very much limited to a small number of language pairs due to the lack of data.

Furthermore, current research in NLP is dominated by benchmark-driven research with a heavy focus on state-of-the-art results in established tasks and

In Hämäläinen, M., Partanen, N., Alnajjar, K. (eds.) Multilingual Facilitation (2021), pages 248–262.

test frameworks. This fact reinforces the emphasis on a small number of selected languages and limits the possibilities of exploring other domains covering a larger proportion of the linguistic diversity in the world. Even more striking is the issue that research done in low-resource NLP is often carried out in artificial setups for the sake of comparison using synthetic data or sub-sampled data sets that simulate low-resource scenarios. The problem with this approach has many aspects: First of all, the research does not benefit any low-resource task with a practical solution and merely focuses on the publication of the methodology with theoretical claims of its potentials in realistic settings. Secondly, the artificial setup does not consider the peculiarities of low-resource scenarios and often builds on clean and well-prepared data that is simply reduced in size. Moreover, the data selected in machine translation research often comes from multiparallel data sets such as WIT[3] [2] or Europarl [8], which is a highly unrealistic scenario for most languages in the world. Strong claims about transfer learning and zero-shot translation based on experiments on such idealistic data sets has the unavoidable consequence that results are exaggerated and create the impression of immediate success of straightforward techniques in multilingual NLP (see, e.g., [3–5, 11]). Those results typically contradict the experience with actually endangered languages and neglect essential efforts that are necessary to build basic infrastructures for such languages [15, 14] and basic prerequisites for work on low-resource machine translation [13].

The problem has certainly been recognized but the underlying problem of benchmark-driven research is still strong. Luckily, new test sets appear all the time and in this paper, we address the issue of low-resource machine translation by establishing a new data set that attempts to increase the scope and to provide test beds for realistic experiments in MT with a much wider support of languages and language pairs. We describe our on-going effort, which is just a small but important step in the direction of a more inclusive research in this field.

Our data set is based on OPUS,[1] a growing collection and widely recognized data hub for parallel corpora. We rely on the coverage in terms of domains and languages and base the benchmarks we propose on the entire collection without removing essential parts to artificially test low-resource settings. In particular, we create a data set that covers over 500 languages naturally including language pairs with limited training data in order to test the impact of resources and noise on translation quality. More details are given below. The complete list of supported languages is available in the appendix. The benchmark itself is based on Tatoeba,[2] a crowd-sourced collection of user-provided translations in hundreds of languages. The sentences in this collection are, admittedly, rather simple in most cases but its language coverage is impressive and the fact that it is constantly growing as a community effort is appealing. We plan to follow-up with new releases to let our benchmark grow as well in order to increase the linguistic coverage of our benchmark even further.

[1] https://opus.nlpl.eu
[2] https://tatoeba.org/

In Hämäläinen, M., Partanen, N., Alnajjar, K. (eds.) Multilingual Facilitation (2021), pages 248–262.

Below, we first present the general structure of the data set and benchmarks we created and after that we present a few baselines in order to demonstrate the use of the collection. Here we emphasize the use of Uralic languages and the possibility to train multilingual models with various subsets of languages that can easily be grouped according to established language groups and families.

2 The Tatoeba MT challenge data set

Based on the selected source of the test and development data we named our data set the *Tatoeba MT Challenge*. Our initial release covers 565 languages and provides over 500GB of compressed, sentence-aligned bitexts for 2,961 language pairs. We release dedicated splits into training, development and test data in a convenient plain text format with aligned sentences on corresponding lines uniformly encoded in Unicode UTF-8. Figure 1 illustrates the basic structure of the collection using the example of the German-English language pair.

```
data/deu-eng/
data/deu-eng/train.src.gz
data/deu-eng/train.trg.gz
data/deu-eng/train.id.gz
data/deu-eng/dev.id
data/deu-eng/dev.src
data/deu-eng/dev.trg
data/deu-eng/test.src
data/deu-eng/test.trg
data/deu-eng/test.id
```

Fig. 1. Released data packages: training data, development data and test data. Language labels are stored in ID files that also contain the name of the source corpus for the training data sets.

We apply ISO-639-3 language codes and assume symmetric bitexts, i.e. we only provide parallel data in one translation direction as we assume that there is no difference for the set in the other direction. This is certainly a shortcoming, but due to the lack of annotation in the original OPUS data, there is no reasonable other way. Furthermore, many parallel data sets will not come from directly translated sources between the given languages but will be rather done via an intermediate pivot language such as English.

Source language files are marked with the file extension *.src* whereas target language files will have the extension *.trg*. The additional ID files (with extension *.id*) provide language codes and domain labels in the case of training data in a simple TAB-separated format. Language codes are important to mark sub-languages that are subsumed under the same macro-language label or

to indicate different scripts. More information about language labels is given in Section 2.1 below. Domain labels refer to the original source in the OPUS corpus and make it possible to divide the data acording to its source. In the current release, we include data from many different collections and the compilation for specific language pairs depends on the availability of the data in each of them. Sources include, among others, publications of the European Union, translated movie subtitles, localisation data of open source software, translated news items, aligned subtitles from TED talk, crawled bitexts and bilingual data extracted from Wikipedia. For most languages, only a few of those domains are available and, especially in low-resource scenarios, the data variety is naturally pretty limited. For more information about domains and sources, we refer to the OPUS collection and the data we provide in the Tatoeba MT Challenge set.

In practice, we compile the data from the pre-aligned bitexts released in OPUS but also further clean them in various ways. For example, we remove non-printable characters and strings that violate Unicode encoding principles using a set of regular expressions and *recode* (v3.7), a popular character conversion tool.[3] We de-escape XML characters that are still in the data applying Moses tools for that purpose [9] and, finally, we also use automatic language identification to further reduce noise from the data taking the compact language detect library (CLD2) through its Python bindings.[4] Another standard Python library is used for converting between different ISO-639 standards.[5]

As we are working with a large number of languages, we opt for the extended set of supported languages in CLD2 that covers 172 languages. We apply the "best effort" option and provide a language hint coming from the annotation of the original data. There are still a lot of unsupported languages that we need to handle and for those we reverse the language identification filter and remove all examples that are detected to be English instead of another or an unknown language. The inclusion of English is a common problem in some corpora and using this strategy we are able exclude a large portion of untranslated text in, e.g. localization corpora and other noisy data sets. In general, we rely on detected languages only if they are flagged as reliable by the softwar.

Furthermore, we decided to use macro-languages from the ISO639-3 standard if available and merge sub-languages from all corpora accordingly. The specific languages are marked using the labels in the given ID files as explained above. We then deduplicate and shuffle the entire data set per language pair using GNU/Unix tools such as `sort`, `uniq` and `terashuf`[6] that is capable to efficiently shuffle large data sets. We also release the procedure of compiling the corpus to make the compilation as transparent as possible.[7]

The entire collection is naturally skewed due to the varying availability of data per language and language pair. English-French is unsuprisingly the largest

[3] https://github.com/pinard/Recode
[4] https://pypi.org/project/pycld2/
[5] https://pypi.org/project/iso-639/
[6] https://github.com/alexandres/terashuf
[7] https://github.com/Helsinki-NLP/Tatoeba-Challenge

In Hämäläinen, M., Partanen, N., Alnajjar, K. (eds.) Multilingual Facilitation (2021), pages 248–262.

data set with over 180 million aligned sentence pairs. It is important to stress that our release provides realistic settings for many language pairs and another 173 of them contain over 10 million sentence pairs. We also plan regular updates of the data to keep the collection on par with the increasing amounts of training data in modern machine translation. With this we want to avoid to artifically keep the sources down that would lead to wrong conclusions in comparison to the state of the art.

2.1 Language labels and scripts

One of the main efforts done when compiling the data set is related to harmonizing the language codes. We converted the original language IDs given by OPUS into a standardized form and provide consistent three-letter codes from ISO-639-3. Furthermore, we added information about the writing system in cases where various scripts are in use. For the latter, we implemented a simple script detection procedure based on Unicode character classes and regular expressions counting letters from specific script properties. The writing scripts are specified using standardized four-letter codes ISO-15924 and we attach them to the language code. Mixed content is marked by the most frequent script that is found in the string. A special code (Zyyy) is specified to refer to characters that cannot be used to distinguish scripts. We omit the script tag if it is the only one used in that language and we also omit the tag of the default script of a language (if that kind of information is known about the language in question). All of the labels are assigned automatically and we cannot effort human validation of the annotation. Mistakes are hard to avoid but we tried to reduce noise as much as possible.

It is also important to note that the use of macro-languages leads to data sets that incorporate various sub-languages. Together with the variation of writing systems this may cause quite a number of variants in one bitext. Below is the example of Serbo-Croatian (hbs), Japanese (jpn) and Chinese (zho):

Serbo-Croatian: bos_Latn, hrv, srp_Cyrl, srp_Latn
Japanese: jpn, jpn_Hani, jpn_Hira, jpn_Kana, jpn_Latn
Chinese: cjy_Hans, cjy_Hant, cmn, cmn_Bopo, cmn_Hans, cmn_Hant, cmn_Latn, gan, lzh, lzh_Bopo, lzh_Hang, lzh_Hani, lzh_Hans, lzh_Hira, lzh_Kana, lzh_Yiii, nan_Hani, nan_Latn, wuu, wuu_Bopo, wuu_Hang, wuu_Hani, wuu_Hira, yue_Hans, yue_Hant, yue_Latn

Using macro-languages addresses the scarcity of some languages and puts them together with naturally closely related languages. Nevertheless, with the provided labels it is still possible to divide them into separate language pairs and those labels can also be used to remove additional noise, for example, the unusual use of Latin script in Japanese data, which is most likely an indication for erroneous data. Still, it is important to remember that script detection can also fail and the result needs to be taken with a grain of salt. For example, distinguishing between traditional (Hant) and simplified Chinese (Hans) can be ambiguous and noise in the data can have an effect on the detection algorithm.

In Hämäläinen, M., Partanen, N., Alnajjar, K. (eds.) Multilingual Facilitation (2021), pages 248–262.

For transparency, we release all our code that we use for language code normalization and writing script detection and the software can be downloaded from github.[8]

2.2 Monolingual data

We also provide monolingual text in addition to the parallel data sets. Those collections come from public data provided by the Wikimedia foundation in popular resources such as Wikipedia, Wikibooks, Wikinews, Wikiquote and Wikisource. We extract sentences from data dumps provided in JSON format[9] and process them with jq,[10] a lightweight JSON processing tool. Cleaning and language identification is done in the same way as explained above and all data sets are marked with ISO-639-3 language codes. We apply the Moses tools [9] and UDPipe [16] for sentence boundary detection and provide shuffled and de-duplicated versions of the data as well as collections that preserve document boundaries to enable experiments with document-level models future work.

3 The translation benchmark

The main purpose of the release is to provide a broad set of benchmarks for machine translation with a focus on low-resource languages and lesser common language combinations. For convenience, we divide the data into different categories depending on the size of available training data in our collection. For this, we define some (rather arbitrary) thresholds to create the following categories:

zero-shot tasks: 40 language pairs with no training data
low-resource tasks: 87 language pairs with less than 100,000 training examples; 24 of them have less then 10,000 training examples (which we keep as a distinct sub-category in this class)
medium-sized resource tasks: 97 language pairs with less than 1 million training examples but more than 100,000 sentence pairs
high-resource tasks: 298 language pairs with at least one million training examples; 173 of them contain over 10 million sentence pairs

In total, we have 522 benchmarks using the data above and test sets containing at least 200 sentences. Only 101 of them involve English, which gives a good amount of less common non-English test cases. 288 of the benchmarks include over 1,000 sentence pairs making them reliable enough for empirical comparisons. The upper size limit for our test sets is set to 10,000, which we achieve for 76 language pairs. The remaining sentences in those cases are reserved for validation purposes in disjoint development data. Test sets are reduced to 5,000 sentences (19 in the current release) if there are less than 20,000 examples in the

[8] https://github.com/Helsinki-NLP/LanguageCodes
[9] https://dumps.wikimedia.org/other/cirrussearch/current
[10] https://stedolan.github.io/jq/

urj-sla	BLEU	chrF2
chm-rus (4K)	1.7	0.159
est-rus (7M)	46.2	0.667
fin-pol (29M)	41.5	0.622
fin-rus (12M)	41.0	0.621
hun-bul (32M)	36.2	0.591
hun-ces (40M)	39.2	0.598
hun-pol (40M)	37.8	0.603
hun-rus (19M)	38.2	0.590
hun-ukr (1.5M)	38.3	0.586

sla-urj	BLEU	chrF2
rus-chm	1.4	0.176
rus-est	51.1	0.703
pol-fin	38.0	0.605
rus-fin	38.3	0.624
bul-hun	41.3	0.634
ces-hun	44.1	0.652
pol-hun	39.3	0.624
rus-hun	36.2	0.590
ukr-hun	40.2	0.647

urj-sit	BLEU	chrF2
fin-zho (9M)	29.8	0.236
zho-fin (9M)	23.3	0.393

urj-bat	BLEU	chrF2
fin-lit (9M)	34.5	0.589
lit-fin (9M)	39.0	0.621

urj-sem	BLEU	chrF2
fin-heb (18M)	29.8	0.534
heb-fin (18M)	33.8	0.589

Table 1. Results from selected languages coming out of the multilingual translation model between Uralic languages (urj) and Slavic languages (sla), Sino-Tibetan languages (sit), Baltic languages (bat) and Semitic languages (sem) measured on the Tatoeba test set using BLEU and chrF2. Numbers in brackets refer to the original size of the data from which the training sets were sampled (M for millions of sentences and K for thousands).

Tatoeba data collection for that language pair. We further reduce to 2,500 test sentences (48 language pairs) with less than 10,000 sentences in Tatoeba and 1,000 test sentences for data sets below 5,000 examples in total (currently 78 language pairs). For language pairs below 2,000 translated sentences in Tatoeba, we keep everything as test and skip validation data.

Naturally, test and validation data are strictly disjoint and none of the examples from Tatoeba are explicitly included in the training data. Accidental overlaps, however, do appear but the average proportion is rather low – around 5.5% across all data sets with a median percentage of 2.3% and 2.9% for test and validation data, respectively.

4 Baseline models for Uralic languages

The intention of publishing the data set is to motivate machine translation researchers to compare their systems based on the given benchmarks and to use the data to push MT development for a wide variety of language pairs. We also release our models and below we discuss a few baselines that we trained to study the capabilities of multilingual NMT to capture essential information for the translation from and to Uralic languages in our collection.

For this purpose, we train state-of-the-art transformer models [18] using Marian-NMT,[11] a stable production-ready NMT toolbox with efficient training and decoding capabilities [7]. Our setup applies a common architecture with 6 self-attentive layers in both, the encoder and decoder network using 8 attention heads in each layer. The hyper-parameters follow the general recommendations given in the documentation of the software.[12] The training procedures follow the

[11] https://marian-nmt.github.io

[12] https://github.com/marian-nmt/marian-examples/tree/master/transformer

strategy implemented in OPUS-MT [17] and detailed instructions are available from github.[13]

Training is performed on v100 GPUs with early-stopping after 10 iterations of dropping validation perplexities. We use SentencePiece [10] for the segmentation into subword units and apply a shared vocabulary of a maximum of 65,000 items. Language label tokens in the spirit of [6] are used in case of multiple language variants or scripts in the target language. All our models also apply a simple sampling strategy to balance the amount of data available for each language pair that is included in the model. In particular, we restrict the training data to a maximum of one million training examples per language pair taken from the top of the shuffled data set. Furthermore, we over-sample language pairs that come with less than one million sentence pairs by repeating the data to fill up the quota. However, we set a threshold of a maximum of 50 copies to avoid an over-representation of repeated (and possibly highly noisy) content in very-low-resource languages.

Table 1 lists a few results from our experiments covering selected languages in the multilingual data sets of Uralic languages and Slavic languages, Sino-Tibetan languages, Baltic languages and Semitic languages. Note that only a few language pairs in each language group can properly be tested with the benchmarks available and hopefully, a larger coverage will be achieved in the near future. In order to help this proces, we encourage the reader to directly submit translations to the Tatoeba initiative, which will immediately lead to an increase of our test data.

Here, we only provide automatic metrics (BLEU with a 4-gram maximum and character-F2-scores) computed using *sacrebleu* [12] and leave more detailed analyses of the results to future work. The scores suggest that the models are capable of decent translation performances for the rather well-equipped language pairs. For the low-resource language in our example, Mari (chm), we can see a dramatic drop, which stresses the issue with realistically under-resourced languages. In the Tatoeba MT data, we have less than four thousand training examples for Mari-Russian (with roughly 14,000 words in each language) and this is certainly not enough for our setup and no significant transfer learning across languages is happening in neither of the translation directions.

Table 2 lists additional results connecting Uralic languages to Germanic languages. We can see a similar effect that high-resource languages are well covered and show a reasonable translaton performance. The translation into Uralic language still seems to be harder compared to the translation into Germanic languages, something that has been observed in previous work as well. An especially interesting case is Karelian, which appears to translate rather well into high resource languages such as English and Dutch whereas the other translation direction reveals the shortcomings of the model when generating the low-resource language instead of translating it. It has to be noted that there is no Karelian-English nor Karelian-Dutch in the training data and, hence, this can be seen as a typical example for a zero-shot translation task. Here, we can see that a sufficient

[13] https://github.com/Helsinki-NLP/OPUS-MT-train/blob/master/doc/TatoebaChallenge.md

urj-gmw	BLEU	chrF2
chm-deu (4K)	1.9	0.183
est-deu (18M)	45.9	0.652
est-eng (25M)	54.4	0.698
fin-deu (28M)	41.4	0.610
fin-eng (45M)	46.6	0.644
fin-nld (9M)	56.2	0.721
hun-deu (26M)	36.5	0.579
hun-eng (56M)	45.3	0.626
hun-nld (35M)	45.2	0.636
krl-eng (0)	30.0	0.474
krl-nld (0)	30.9	0.446

gmw-urj	BLEU	chrF2
deu-chm	0.8	0.131
deu-est	51.6	0.716
eng-est	48.7	0.679
deu-fin	35.0	0.591
eng-fin	33.0	0.580
nld-fin	51.4	0.716
deu-hun	28.5	0.539
eng-hun	33.9	0.574
nld-hun	38.8	0.612
eng-krl	9.1	0.206
nld-krl	11.1	0.236

urj-gmq	BLEU	chrF2
fin-dan (24M)	52.4	0.676
fin-nor (9M)	45.4	0.668
fin-swe (29M)	50.4	0.666
hun-nor (8M)	54.8	0.730
hun-swe (21M)	48.2	0.647

gmq-urj	chrF2	BLEU
dan-fin	35.8	0.595
nor-fin	34.4	0.596
swe-fin	40.6	0.638
nor-hun	45.6	0.665
swe-hun	38.8	0.617

Table 2. Results from the multilingual translation models between Uralic languages and various Germanic languages from the Northern and Western subdivision.

transfer can be achieved for closely related languages especially in the encoder but the decoder still struggles a lot when insufficient or no direct training data is available.

Note that the results above are meant to provide an illustration of the experiments that can be carried out with the data we provide in the Tatoeba MT data set. The purpose is to provide a useful test bed for the development of low-resource machine translation, lesser studied language pairs and for systematic investigations on the effect of transfer learning in multilingual translation models as well as the general impact of data size on translation quality. Below, we give yet another example of the possibilities offered by the Tatoeba MT Challenge: An investigation of data augmentation techniques and how that approach can be studied using our data collection.

5 Data augmentation via back-translation

The additional monolingual data we provide as described in Section 2.2 makes it possible to augment the data via automatic back-translation. This approach refers to the common trick of tanslating monolingual target language data into the source language using a model that has been trained in the reverse direction. The Tatoeba challenge data set provides a perfect test bed for the study of this technique in various scenarios and here we provide some initial results in a lesser-resource setting using the example of the translation from English into Breton.

The original training data available for this language pair is limited with respect to common requirements for building NMT models. The training data includes roughly 380,000 mostly short and noisy training examples with less than 1.5 million tokens per language. The majority of the text (363,000 sentence pairs) comes from repetitive software localization sources such as GNOME, KDE4 and Ubuntu and about 17,000 aligned sentences refer to movie subtitles. The perfor-

model	data size	BLEU	chrF2
English-Breton (eng-bre)	375,648	4.2	0.233
+ back-translation	1,063,251	15.5	0.386
+ French-Breton (fra-bre)	1,234,516	17.2	0.402

Table 3. English to Breton translation quality when trained with and without back-translation data and when training multi-source models.

mance on the Tatoeba benchmark is very low as expected and achieves a BLEU score of 4.2, which deems the model to be completely useless.

On the other hand, we know that we can successfully train multilingual models that enable efficient transfer learning across languages especially when used in the encoder. Hence, we trained a multilingual model that translates Indo-European languages (ine) to English using samples from the Tatoeba MT Challenge data set. This model achieves a modest BLEU score of 24.4 on the Tatoebe benchmark for Breton to English, potentially good enough to boost NMT in the opposite translation direction.

Using that model, we now are able to translate the monolingual data sets taken from Wikipedia, Wikiquote and Wikisource from the original Breton to English in order to produce an artificial parallel corpus of English-Breton training examples. This procedure produces a new data set of 687,603 sentence pairs with broken English but original Breton as the target language to be used in the augmented data set for training an improved model for the translation between the two languages.

The approach is simple but very effective as we can see from our initial results. Without further adjustments, the performance of the model is boosted to a score of 15.5 BLEU as measured on the Tatoeba benchmark. Even though the result is still quite modest, the absolute improvement of over 11 BLEU points is quite astonishing considering the limited quality of the back-translation model applied in this case.

As a final test we also created a multilingual model that incorporates French as an additional input language to the training procedures. This approach provides additional training examples from a more reasonable source (the OfisPublik corpus) but for a different language pair. The final model that incorporates all existing training data for both language pairs in the Tatoeba Challenge data set as well as the monolingual Breton data back-translated to English achieves the best score so far with another significant improvement over the bilingually augmented translation model. Table 3 summarizes the results.

In order to facilitate the process of systematic data augmentation, we have recently released over 550 million sentences of monolingual data translated from 188 non-English monolingual data sets to English using models that we have trained on our Tatoeba MT Challenge collection. The synthetic bitexts are all

available for download[14] and can be used in further experiments and MT development.

6 Conclusions and outlook into future work

In this paper we present a new comprehensive data set and benchmark for machine translation that covers roughly 3,000 language pairs and over 500 languages and language variants. The collection includes training and test data that can be used to explore realistic low-resource scenarios and zero-shot machine translation. The data set is carefully annotated with standardized language labels including variations in writing scripts and with information about the original source. We have presented some baseline results to illustrate the use of the collection. Our initial results show the benefits of the collection enabling systematic studies across a variety of languages and language families. We are looking forward to intensive work with the benchmarks and further contributions to the development of the resource. In particular, we would like to encourage the community to improve the data sets that are available to us. The benchmarks require further support to increase language and domain coverage and we would need to develop a strategy sufficient quality control in the long run. Needless to say that such a wide collection is difficult to maintain with appropriate accuracy for all language pairs included. The only way forward is to involve language experts and native speakers and our goal is to provide an open platform where anyone interested can contribute and help to improve the resource.

We currently work on automatic filters [1] that provide a basic framework for removing noise but additional human effort is certainly necessary to remove errors and control language-specific properties. The training data itself can be of mixed quality as one of the main tasks is to develop methods that can learn from noisy examples and manages to handle noisy input. However, the test data we provide require quality control as we want to avoid additional misleading results based on inappropriate benchmarks. We are aware of the shortcomings of the current data set but certainly hope that it will lead to a more realistic view on data-driven machine translation in low-resource settings.

[14] https://github.com/Helsinki-NLP/Tatoeba-Challenge/blob/master/Backtranslations.md

In Hämäläinen, M., Partanen, N., Alnajjar, K. (eds.) Multilingual Facilitation (2021), pages 248–262.

References

1. Aulamo, M., Virpioja, S., Tiedemann, J.: OpusFilter: A configurable parallel corpus filtering toolbox. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. pp. 150–156. Association for Computational Linguistics, Online (Jul 2020). https://doi.org/10.18653/v1/2020.acl-demos.20, https://www.aclweb.org/anthology/2020.acl-demos.20
2. Cettolo, M., Girardi, C., Federico, M.: Wit^3: Web inventory of transcribed and translated talks. In: Proceedings of the 16^{th} Conference of the European Association for Machine Translation (EAMT). pp. 261–268. Trento, Italy (May 2012)
3. Firat, O., Cho, K., Bengio, Y.: Multi-way, multilingual neural machine translation with a shared attention mechanism. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 866–875. Association for Computational Linguistics, San Diego, California (Jun 2016). https://doi.org/10.18653/v1/N16-1101, https://www.aclweb.org/anthology/N16-1101
4. Firat, O., Sankaran, B., Al-onaizan, Y., Yarman Vural, F.T., Cho, K.: Zero-resource translation with multi-lingual neural machine translation. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. pp. 268–277. Association for Computational Linguistics, Austin, Texas (Nov 2016). https://doi.org/10.18653/v1/D16-1026, https://www.aclweb.org/anthology/D16-1026
5. Ha, T.L., Niehues, J., Waibel, A.: Toward multilingual neural machine translation with universal encoder and decoder (2016)
6. Johnson, M., Schuster, M., Le, Q.V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., et al.: Google's multilingual neural machine translation system: Enabling zero-shot translation. Transactions of the Association of Computational Linguistics **5**(1), 339–351 (2017)
7. Junczys-Dowmunt, M., Grundkiewicz, R., Dwojak, T., Hoang, H., Heafield, K., Neckermann, T., Seide, F., Germann, U., Aji, A.F., Bogoychev, N., Martins, A.F.T., Birch, A.: Marian: Fast neural machine translation in C++. In: Proceedings of ACL 2018, System Demonstrations. pp. 116–121 (2018). https://doi.org/10.18653/v1/P18-4020, https://www.aclweb.org/anthology/P18-4020
8. Koehn, P.: Europarl: A parallel corpus for statistical machine translation. In: MT summit. vol. 5, pp. 79–86. Citeseer (2005)
9. Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., Herbst, E.: Moses: Open source toolkit for statistical machine translation. In: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions. pp. 177–180. Association for Computational Linguistics, Prague, Czech Republic (Jun 2007), https://www.aclweb.org/anthology/P07-2045
10. Kudo, T., Richardson, J.: SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 66–71. Association for Computational Linguistics, Brussels, Belgium (Nov 2018). https://doi.org/10.18653/v1/D18-2012, https://www.aclweb.org/anthology/D18-2012

11. Lakew, S.M., Federico, M., Negri, M., Turchi, M.: Multilingual neural machine translation for low-resource languages. IJCoL - Italian Journal of Computational Linguistics **4**(1) (2018), emerging Topics at the Fourth Italian Conference on Computational Linguistics
12. Post, M.: A call for clarity in reporting BLEU scores. In: Proceedings of the Third Conference on Machine Translation: Research Papers. pp. 186–191. Association for Computational Linguistics, Brussels, Belgium (Oct 2018). https://doi.org/10.18653/v1/W18-6319, https://www.aclweb.org/anthology/W18-6319
13. Rueter, J., Hämäläinen, M., et al.: Prerequisites for shallow-transfer machine translation of Mordvin languages: Language documentation with a purpose. Материалы Международного образовательного салона (2020)
14. Rueter, J., Partanen, N., Ponomareva, L.: On the questions in developing computational infrastructure for Komi-permyak. In: Proceedings of the Sixth International Workshop on Computational Linguistics of Uralic Languages. pp. 15–25. Association for Computational Linguistics, Wien, Austria (10–11 Jan 2020), https://www.aclweb.org/anthology/2020.iwclul-1.3
15. Rueter, J., Tyers, F.: Towards an open-source universal-dependency treebank for Erzya. In: Proceedings of the Fourth International Workshop on Computational Linguistics of Uralic Languages. pp. 106–118. Association for Computational Linguistics, Helsinki, Finland (Jan 2018). https://doi.org/10.18653/v1/W18-0210, https://www.aclweb.org/anthology/W18-0210
16. Straka, M., Hajič, J., Straková, J.: UDPipe: Trainable pipeline for processing CoNLL-u files performing tokenization, morphological analysis, POS tagging and parsing. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). pp. 4290–4297. European Language Resources Association (ELRA), Portorož, Slovenia (May 2016), https://www.aclweb.org/anthology/L16-1680
17. Tiedemann, J., Thottingal, S.: OPUS-MT — Building open translation services for the World. In: Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT). Lisbon, Portugal (2020)
18. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. CoRR **abs/1706.03762** (2017), http://arxiv.org/abs/1706.03762

A List of supported languages

Below is a complete list of supported ISO-639 languages including the number of released data sets that include each language. Note that the collection contains artificial languages such as Interlingua and that the table lists macro-languages if available in ISO-639 instead of individual languages subsumed under that category. The list contains languages for which no test sets are available in Tatoeba. For those, we provide training data aligned to English if available in OPUS.

In Hämäläinen, M., Partanen, N., Alnajjar, K. (eds.) Multilingual Facilitation (2021), pages 248–262.

sets	ISO	language
610	eng	English
181	fra	French
150	spa	Spanish
137	deu	German
123	jpn	Japanese
116	rus	Russian
115	epo	Esperanto
106	tur	Turkish
94	zho	Chinese
93	ita	Italian
79	por	Portuguese
78	pol	Polish
78	nor	Norwegian
76	nld	Dutch
72	yid	Yiddish
71	heb	Hebrew
71	ara	Arabic
67	hun	Hungarian
67	ces	Czech
66	lat	Latin
64	ido	Ido
63	ukr	Ukrainian
62	vol	Volapük
62	dan	Danish
61	ina	Interlingua
58	lfn	Lingua Franca Nova
58	lad	Ladino
58	bzt	Brithenig
57	enm	Middle English
56	fin	Finnish
54	swe	Swedish
53	tzl	Talossan
51	tlh	Klingon
50	kur	Kurdish
49	toki	Toki Pona
49	ron	Romanian
48	ile	Interlingue
46	msa	Malay
44	bul	Bulgarian
42	nov	Novial
42	kab	Kabyle
42	jbo	Lojban
42	fas	Persian
41	bel	Belarusian
41	ang	Old English
40	tmr	Babylonian Aramaic
40	lit	Lithuanian
39	fry	Western Frisian
39	avk	Kotava
38	vie	Vietnamese
38	tat	Tatar
37	sjn	Sindarin
37	gos	Gronings
36	kor	Korean
36	afr	Afrikaans
35	nds	Low German
35	ldn	Láadan
35	est	Estonian
35	ber	Berber languages
34	lav	Latvian
33	osp	Old Spanish
33	ltz	Luxembourgish
32	slv	Slovenian
31	ell	Modern Greek
30	qya	Quenya
28	isl	Icelandic
27	hbs	Serbo-Croatian
26	mkd	Macedonian
23	swa	Swahili
23	kaz	Kazakh
23	gle	Irish
22	cat	Catalan
20	kat	Georgian
20	cym	Welsh
19	uzb	Uzbek
18	fao	Faroese
17	tgl	Tagalog
17	cor	Cornish
17	aze	Azerbaijani
16	uig	Uighur
16	tuk	Turkmen
16	ota	Ottoman Turkish
16	oci	Occitan
16	eus	Basque
16	dtp	Kadazan Dusun
15	swg	Swabian
15	bak	Bashkir
14	tha	Thai
14	sux	Sumerian
14	stq	Saterfriesisch
14	hin	Hindi
14	glg	Galician
14	bar	Bavarian
13	urd	Urdu
12	mlt	Maltese
12	khm	Khmer
12	dws	Dutton WS
12	ast	Asturian
12	asm	Assamese
11	pms	Piemontese
11	orv	Old Russian
11	gla	Scottish Gaelic
11	bre	Breton
10	sco	Scots
10	phn	Phoenician
10	mar	Marathi
10	grn	Guarani
10	gcf	Guad. Creole French
10	crh	Crimean Tatar
10	cos	Corsican
10	chm	Mari (Russia)
10	che	Chechen
10	ceb	Cebuano
9	hye	Armenian
9	grc	Ancient Greek
9	frr	Northern Frisian
9	fkv	Kven Finnish
9	dsb	Lower Sorbian
9	cycl	Cyc Language
9	chv	Chuvash
8	ksh	Kölsch
8	kir	Kirghiz
8	ben	Bengali
7	vep	Veps
7	sme	Northern Sami
7	shs	Shuswap
7	run	Rundi
7	mon	Mongolian
7	liv	Liv
6	xal	Kalmyk
6	shy	Tachawit
6	sah	Yakut
6	prg	Prussian
6	oss	Ossetian
6	moh	Mohawk
6	krl	Karelian
6	jpa	Palestinian Aramaic
6	got	Gothic
6	cha	Chamorro
5	tgk	Tajik
5	sqi	Albanian
5	rom	Romany
5	roh	Romansh
5	pap	Papiamento
5	nys	Nyungar
5	mya	Burmese
5	lld	Ladin
5	lij	Ligurian
5	kha	Khasi
5	izh	Ingrian
5	ilo	Iloko
5	hsb	Upper Sorbian
5	hrx	Hunsrik
5	hat	Haitian
5	ful	Fulah
5	egl	Emilian
5	cbk	Chavacano
4	wln	Walloon
4	tpi	Tok Pisin
4	scn	Sicilian
4	pag	Pangasinan
4	oar	Old Aramaic
4	mwl	Mirandese
4	lmo	Lombard
4	lin	Lingala
4	lao	Lao
4	kum	Kumyk
4	kal	Kalaallisut
4	dng	Dungan
4	cay	Cayuga
3	yor	Yoruba
3	wol	Wolof
3	war	Waray (Philippines)
3	vec	Venetian
3	tpw	Tupí
3	ton	Tonga (Tonga Islands)
3	tah	Tahitian
3	som	Somali
3	san	Sanskrit
3	rif	Tarifit
3	que	Quechua
3	pmn	Pam
3	nog	Nogai
3	nch	Central Huasteca Nahuatl
3	nah	Nahuatl languages
3	mvv	Tagal Murut
3	mal	Malayalam
3	jav	Javanese
3	iba	Iban
3	hil	Hiligaynon
3	guj	Gujarati
3	gsw	Swiss German
3	glv	Manx
3	fvr	Fur
3	csb	Kashubian
3	bub	Bua
3	bho	Bhojpuri
3	arg	Aragonese
3	afh	Afrihili
2	zza	Zaza
2	udm	Udmurt
2	tyv	Tuvinian
2	tts	Northeastern Thai
2	toi	Tonga (Zambia)
2	tmh	Tamashek
2	tly	Talysh
2	tel	Telugu
2	tam	Tamil
2	syc	Classical Syriac
2	srd	Sardinian
2	sna	Shona
2	smo	Samoan
2	sma	Southern Sami
2	sgs	Samogitian
2	rue	Rusyn
2	pus	Pushto
2	ppl	Pipil
2	pdc	Pennsylvania German
2	pcd	Picard
2	otk	Old Turkish
2	ori	Oriya
2	non	Old Norse
2	nep	Nepali
2	nau	Nauru
2	myv	Erzya
2	mai	Maithili
2	kkt	Koi
2	kjh	Khakas
2	kek	Kekchí
2	jam	Jamaican Creole English
2	hoc	Ho
2	hmn	Hmong
2	frm	Middle French
2	fij	Fijian
2	ewe	Ewe
2	emx	Erromintxela
2	cpi	Chinese Pidgin English
2	chr	Cherokee
2	cho	Choctaw
2	awa	Awadhi
2	ava	Avaric
2	atj	Atikamekw
2	asf	Auslan
2	arn	Mapudungun
2	arc	Official Aramaic
2	aoz	Uab Meto
2	amy	Ami
2	amu	Guerrero Amuzgo
2	alz	Alur
2	alt	Southern Altai
2	aka	Akan
2	ain	Ainu (Japan)
2	aik	Ake
2	agr	Aguaruna
2	ada	Adangme
2	acu	Achuar-Shiwiar
2	ach	Acoli
2	ace	Achinese
2	aar	Afar
1	zul	Zulu
1	zne	Zande
1	zha	Zhuang
1	zea	Zeeuws
1	zdj	Ngazidja Comorian
1	zap	Zapotec
1	yua	Yucateco
1	ybb	Yemba
1	yaq	Yaqui
1	yap	Yapese
1	yao	Yao

sets	ISO	language
1	xho	Xhosa
1	wlv	Wichí Lhamtés Vejoz
1	wls	Wallisian
1	wes	Cameroon Pidgin
1	wba	Warao
1	wal	Wolaytta
1	wae	Walser
1	vmw	Makhuwa
1	vls	Vlaams
1	ven	Venda
1	usp	Uspanteco
1	urw	Sop
1	umb	Umbundu
1	tzo	Tzotzil
1	tzh	Tzeltal
1	tyj	Tai Do
1	tvl	Tuvalu
1	tum	Tumbuka
1	ttj	Tooro
1	tsz	Purepecha
1	tso	Tsonga
1	tsc	Tswa
1	trv	Taroko
1	top	Papantla Totonac
1	toj	Tojolabal
1	toh	Gitonga
1	tog	Tonga (Nyasa)
1	tob	Toba
1	tll	Tetela
1	tiv	Tiv
1	tir	Tigrinya
1	tet	Tetum
1	tdt	Tetun Dili
1	tcy	Tulu
1	tcf	Malinaltepec Me'phaa
1	szl	Silesian
1	syr	Syriac
1	sxn	Sangir
1	sun	Sundanese
1	srn	Sranan Tongo
1	srm	Saramaccan
1	son	Songhai languages
1	soh	Aka
1	snd	Sindhi
1	sml	Central Sama
1	smc	Som
1	slk	Slovak
1	sin	Sinhala
1	sid	Sidamo
1	shr	Shi
1	shn	Shan
1	sgn	sign languages
1	sfw	Sehwi
1	seh	Sena
1	sbs	Subiya
1	sat	Santali
1	sag	Sango
1	rup	Macedo-Romanian
1	rnd	Ruund
1	rcf	Réunion Creole French
1	rar	Rarotongan
1	rap	Rapanui
1	quc	K'iche'
1	ppk	Uma
1	pot	Potawatomi
1	pon	Pohnpeian
1	pnt	Pontic
1	pmy	Papuan Malay
1	pli	Pali
1	pis	Pijin
1	pih	Pitcairn-Norfolk
1	pid	Piaroa
1	pdt	Plautdietsch
1	pcm	Nigerian Pidgin
1	pck	Paite Chin
1	pbb	Páez
1	pau	Palauan
1	pan	Panjabi
1	ote	Mezquital Otomi
1	orm	Oromo
1	ong	Olo
1	oke	Okpe (Southwestern Edo)
1	nzi	Nzima
1	nyu	Nyungwe
1	nyn	Nyankole
1	nyk	Nyaneka
1	nya	Nyanja
1	nst	Tase Naga
1	nso	Pedi
1	nrm	Narom
1	nqo	N'Ko
1	nnh	Ngiemboon
1	nlv	Orizaba Nahuatl
1	niu	Niuean
1	nij	Ngaju
1	nia	Nias
1	nhn	Central Nahuatl
1	nhk	Isthmus-Cosoleacaque Nahuatl
1	nhg	Tetelcingo Nahuatl

sets	ISO	language
1	ngu	Guerrero Nahuatl
1	ngt	Kriang
1	ngl	Lomwe
1	ndo	Ndonga
1	nde	North Ndebele
1	ndc	Ndau
1	ncx	Central Puebla Nahuatl
1	ncj	Northern Puebla Nahuatl
1	nci	Classical Nahuatl
1	nbl	South Ndebele
1	nba	Nyemba
1	nav	Navajo
1	nap	Neapolitan
1	mzn	Mazanderani
1	mxv	Metlatónoc Mixtec
1	mus	Creek
1	mrq	North Marquesan
1	mri	Maori
1	mos	Mossi
1	mnw	Mon
1	mni	Manipuri
1	mlg	Malagasy
1	miq	Mískito
1	mic	Mi'kmaq
1	mgr	Mambwe-Lungu
1	mgm	Mambae
1	mfe	Morisyen
1	men	Mende (Sierra Leone)
1	mdf	Moksha
1	mcp	Makaa
1	mco	Coatlán Mixe
1	maz	Central Mazahua
1	mau	Huautla Mazatec
1	map	Austronesian languages
1	mam	Mam
1	mah	Marshallese
1	mad	Madurese
1	lzz	Laz
1	luy	Luyia
1	lus	Lushai
1	luo	Luo (Kenya and Tanzania)
1	lun	Lunda
1	lug	Ganda
1	lue	Luvale
1	lub	Luba-Katanga
1	lua	Luba-Lulua
1	lrc	Northern Luri
1	loz	Lozi
1	lkt	Lakota
1	lim	Limburgan
1	lez	Lezghian
1	lbe	Lak
1	lam	Lamba
1	lah	Lahnda
1	kxi	Keningau Murut
1	kwn	Kwangali
1	kua	Kuanyama
1	ksw	S'gaw Karen
1	kss	Southern Kisi
1	kri	Krio
1	krc	Karachay-Balkar
1	kqn	Kaonde
1	kpe	Kpelle
1	koo	Konzo
1	kon	Kongo
1	kom	Komi
1	kok	Konkani (macrolanguage)
1	kmb	Kimbundu
1	kin	Kinyarwanda
1	kik	Kikuyu
1	kea	Kabuverdianu
1	kdx	Kam
1	kbp	Kabiyè
1	kbh	Camsá
1	kbd	Kabardian
1	kau	Kanuri
1	kar	Karen languages
1	kan	Kannada
1	kac	Kachin
1	kaa	Kara-Kalpak
1	jmx	Western Juxtlahuaca Mixtec
1	jiv	Shuar
1	jdt	Judeo-Tat
1	jaa	Jamamadí
1	iso	Isoko
1	ish	Esan
1	iro	Iroquoian languages
1	ipk	Inupiaq
1	inh	Ingush
1	iku	Inuktitut
1	ibo	Igbo
1	ibg	Ibanag
1	hyw	Western Armenian
1	hus	Huastec
1	hup	Hupa
1	hne	Chhattisgarhi
1	hmo	Hiri Motu
1	hif	Fiji Hindi
1	her	Herero

sets	ISO	language
1	hch	Huichol
1	haz	Hazaragi
1	haw	Hawaiian
1	hau	Hausa
1	hai	Haida
1	gym	Ngäbere
1	gxx	Wè Southern
1	guw	Gun
1	gur	Farefare
1	gum	Guambiano
1	guc	Wayuu
1	gre	Modern Greek (1453-)
1	gqr	Gor
1	glk	Gilaki
1	gil	Gilbertese
1	gcr	Guianese Creole French
1	gbm	Garhwali
1	gbi	Galela
1	gag	Gagauz
1	frp	Arpitan
1	fro	Old French
1	fon	Fon
1	fil	Filipino
1	fan	Fang (Equatorial Guinea)
1	ext	Extremaduran
1	ewo	Ewondo
1	eml	Emiliano-Romagnolo
1	efi	Efik
1	dzo	Dzongkha
1	dyu	Dyula
1	dua	Duala
1	dop	Lukpa
1	djk	Eastern Maroon Creole
1	dje	Zarma
1	din	Dinka
1	dhv	Dehu
1	dga	Southern Dagaare
1	daf	Dan
1	cuk	San Blas Kuna
1	ctu	Chol
1	cto	Emberá-Catío
1	crp	Creoles and pidgins
1	cre	Cree
1	cop	Coptic
1	cni	Asháninka
1	cnh	Hakha Chin
1	cku	Koasati
1	ckt	Chukot
1	cjp	Cabécar
1	cjk	Chokwe
1	chy	Cheyenne
1	chw	Chuwabu
1	chu	Church Slavic
1	chq	Quiotepec Chinantec
1	chk	Chuukese
1	chj	Ojitlán Chinantec
1	chg	Chagatai
1	chf	Tabasco Chontal
1	cce	Chopi
1	cak	Kaqchikel
1	cac	Chuj
1	cab	Garifuna
1	bzj	Belize Kriol English
1	byv	Medumba
1	byn	Bilin
1	bvy	Baybayanon
1	bug	Buginese
1	bua	Buriat
1	btx	Batak Karo
1	bts	Batak Simalungun
1	btg	Gagnoa Bété
1	bsn	Barasana-Eduria
1	brx	Bodo (India)
1	bpy	Bishnupriya
1	bod	Tibetan
1	bnt	Bantu languages
1	bmv	Bum
1	bin	Bini
1	bih	Bihari languages
1	bhw	Biak
1	bem	Bemba (Zambia)
1	bci	Baoulé
1	bbj	Ghomálá'
1	bbc	Batak Toba
1	bas	Basa (Cameroon)
1	bam	Bambara
1	bal	Baluchi
1	ave	Avestan
1	atk	Ati
1	arh	Arhuaco
1	aoc	Pemon
1	amh	Amharic
1	akl	Aklanon
1	ajg	Aja (Benin)
1	aha	Ahanta
1	ady	Adyghe
1	abk	Abkhazian