NTREX-128 – News Test References for MT Evaluation of 128 Languages

Christian Federmann and **Tom Kocmi** and **Ying Xin**
Microsoft
One Microsoft Way
Redmond, WA-98052, USA
{chrife,tomkocmi,yinxin}@microsoft.com

Abstract

We release NTREX-128, a data set for machine translation (MT) evaluation from English into a total of 128 target languages. The paper describes the data creation process and proposes a quality filtering method based on human evaluation. We show experimental results which confirm that the directionality of test sets translation indeed plays an important role wrt. the usefulness of the corresponding metrics' scores. Thus, we recommend that the NTREX-128 data set should be used for evaluation of English-sourced translation models but not in reverse direction. The test set release introduces another benchmark for the evaluation of massively multilingual machine translation research.

1 Introduction

Research on massively multilingual neural machine translation models requires test data to evaluate the models' quality. The creation of such resources is expensive—especially when one considers test sets for 100+ languages—so the amount of available test data is limited. This hinders progress.

While there already exist a few multilingual benchmark test sets more data will be needed to boost research efforts. Thus, we follow recent "open data" approaches undertaken in the field with this release.

As our research shifted its focus to massively multilingual models we started collecting test data for this scenario. We now release this data to the community as an additional benchmark for the evaluation of massively multilingual machine translation models.

NTREX-128, a data set containing "**N**ews **T**ext **R**eferences of **E**nglish into **X** Languages", expands multilingual testing for translation from English into 128 target languages. Our test data is based on WMT19 (Barrault et al., 2019) test data and compatible with SacreBLEU (Post, 2018).

We release NTREX-128 in the hope that it may be useful for the scientific community.

Data set	# of Languages
TICO-19	37
FLORES-101	101
FLORES-200	200

Table 1: Number of supported languages for three multilingual test data sets. Language sets do not fully overlap and text domains differ across the data sets.

2 Literature Review

Recently, the Conference on Machine Translation (WMT) has added a shared task on large-scale, multilingual machine translation. Such tasks require benchmark data sets for their evaluation. Three examples of such data are:

- TICO-19 (Anastasopoulos et al., 2020);
- FLORES-101 (Goyal et al., 2021; Guzmán et al., 2019); and
- FLORES-200 (#NLLB Team, 2022).

Table 1 shows the total number of languages supported by each of the aforementioned data sets. We will provide brief descriptions of all three data sets below.

TICO-19 is a data set released by the "Translation Initiative for Covid-19". It was a joint effort from several partners from academia and industry. The benchmark includes 30 documents (3,071 sentences, 69.7k words) translated from English into 37 target languages.

FLORES-101 is a data set released by Meta AI researchers. It includes 842 documents (3,001 sentences) translated from English into 101 target languages.

FLORES-200 extends the above data set to a total of 200 target languages. It is based on the same English source data as FLORES-101.

Proceedings of the First Workshop on Scaling Up Multilingual Evaluation, pages 21–24

3 Data Set

3.1 Creation Process

To produce this data set we sent out the original English WMT19 (Barrault et al., 2019) test set ('newstest2019') to professional human translators. This work started after the release of the WMT19 test data and continued in parallel to our work on new translation models since then. Translators did have the full document context available but we do not know if (or to which degree) they have used this information.

3.2 Quality Assurance

Test data has to be of a high-enough quality level to be useful. We specified two main requirements: 1) we require translations which are performed by native speakers of the respective target language who are bilingual in English; and 2) reference translations should not be created based on post-editing MT output.

Our translation provider, as part of their translation process, performed quality assurance before delivery of the test set files. Upon receipt of the files we then sent them out to human evaluation via source-based direct assessment (src-DA), as implemented in the Appraise framework (Federmann, 2018). To avoid potential bias, annotation work was performed by an independent vendor.

As the result of the human evaluation process, we obtain segment-level quality scores based on the assessment of bilingual annotators who are native speakers of the respective target language. Scores range from $0 - 100$ and express the 'quality of the semantic transfer' between source and target language. This focuses more on adequacy than on fluency but, based on previous research findings, we consider this an acceptable trade-off.

Segments with scores < 25 are deemed defective, while any score in the $[25, 50)$ range is considered suspect. We return any segments with a score < 50 to the translation vendor for repairs. We have found that this method allows us to check quality for all translated segments; it scales well to thousands of segments with acceptable cost. As a side effect we have observed an increased level of quality control on the translation provider's side as they have understood that we will routinely verify their translation output for the full data sets, instead of random samples.

3.3 Avoiding post-edited reference output

Reference-based evaluation metrics, by design, have an inherent problem with reference bias. Even when dealing with professional translators there is a chance that reference translations may have been created by post-editing machine translation output. This is a problem for two reasons: First, it gives the respective MT system an unfair advantage in competitive evaluations. Second, it means that the reference translations are not independently produced anymore and, thus, may be of inferior quality compared to human translation from scratch.

4 Statistics

The NTREX-128 benchmark includes 123 documents (1,997 sentences, 42k words) translated from English into 128 target languages. More details are available in Appendix C.

5 Experiments

Based on the recent success of embedding-based, automatic evaluation metrics such as COMET (Rei et al., 2020), we run an experiment with the NTREX-128 data set in which we compare COMET-src scores for the authentic translation direction against the scores obtained in the reverse direction. As a secondary concern we investigate how COMET-src behaves for languages which it has not been trained on.

6 Results

We make the following observations:

- using COMET-src for quality estimation of test data is possible but limited as score ranges are non-comparable across language pairs;
- a sizable subset of languages sees COMET-src scores on translationese input scored higher than the corresponding authentic source data;
- while relative comparisons of COMET-src scores work for all language pairs there exists a subset of languages for which the scores appear broken. We suggest that this may be related to the fact COMET has never seen any training data examples for these languages.

See the `RESULTS` file in our repository for more details. As our main focus lies in the release of the NTREX-128 data set, we leave the further investigation of these points for future work.

7 Conclusion

We have presented our work on NTREX-128, a data set which contains 128 reference translations of the English 'newstest2019' test set originally released as part of WMT19. We intend to make it available as part of SacreBLEU. The test data will be released in the hope that it may be useful for the scientific community.

Acknowledgements

See the CONTRIBUTORS file in our repository. We would also like to thank the anonymous reviewers for their feedback.

References

Antonios Anastasopoulos, Alessandro Cattelan, Zi-Yi Dou, Marcello Federico, Christian Federmann, Dmitriy Genzel, Francisco Guzmán, Junjie Hu, Macduff Hughes, Philipp Koehn, Rosie Lazar, Will Lewis, Graham Neubig, Mengmeng Niu, Alp Öktem, Eric Paquin, Grace Tang, and Sylwia Tur. 2020. TICO-19: the Translation initiative for COvid-19. arXiv:2007.01788.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (wmt19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 conference on machine translation (wmt17). In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 169–214, Copenhagen, Denmark. Association for Computational Linguistics.

Mauro Cettolo, Marcello Federico, Luisa Bentivogli, Jan Niehues, Sebastian Stüker, Katsuhito Sudoh, Koichiro Yoshino, and Christian Federmann. 2017. Overview of the IWSLT 2017 evaluation campaign. In *Proceedings of the 14th International Conference on Spoken Language Translation*, pages 2–14, Tokyo, Japan. International Workshop on Spoken Language Translation.

Christian Federmann. 2018. Appraise evaluation framework for machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 86–88, Santa Fe, New Mexico. Association for Computational Linguistics.

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

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria. Association for Computational Linguistics.

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.

James Cross Onur Çelebi Maha Elbayad Kenneth Heafield Kevin Heffernan Elahe Kalbassi Janice Lam Daniel Licht Jean Maillard Anna Sun Skyler Wang Guillaume Wenzek Al Youngblood Bapi Akula Loic Barrault Gabriel Mejia Gonzalez Prangthip Hansanti John Hoffman Semarley Jarrett Kaushik Ram Sadagopan Dirk Rowe Shannon Spruit Chau Tran Pierre Andrews Necip Fazil Ayan Shruti Bhosale Sergey Edunov Angela Fan Cynthia Gao Vedanuj Goswami Francisco Guzmán Philipp Koehn Alexandre Mourachko Christophe Ropers Safiyyah Saleem Holger Schwenk Jeff Wang #NLLB Team, Marta R. Costa-jussà. 2022. No language left behind: Scaling human-centered machine translation.

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, Belgium, Brussels. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

A License

See the LICENSE file in our repository. In addition to these license terms we ask that you cite this paper when using NTREX-128 in your work. Thank you.

B Download

NTREX-128 data is available from our GitHub repository: https://github.com/MicrosoftTranslator/NTREX.

C List of languages

The NTREX-128 data set covers the following set of 128 languages or language variants:

Afrikaans, Albanian, Amharic, Arabic, Azerbaijani, Bangla, Bashkir, Bosnian, Bulgarian, Burmese, Cantonese, Catalan, Central Kurdish, Chinese, Chuvash, Croatian, Czech, Danish, Dari, Divehi, Dutch, English, Estonian, Faroese, Fijian, Filipino, Finnish, French, Galician, Georgian, German, Greek, Gujarati, Haitian Creole, Hebrew, Hindi, Hmong, Hungarian, Icelandic, Indonesian, Inuinnaqtun, Inuktitut, Irish, isiZulu, Italian, Japanese, Kannada, Kazakh, Khmer, Kiswahili, Korean, Kurdish, Kyrgyz, Lao, Latvian, Lithuanian, Macedonian, Malagasy, Malay, Malayalam, Maltese, Māori, Marathi, Maya, Yucatán, Mongolian, Nepali, Norwegian, Odia, Otomi, Querétaro, Pashto, Persian, Polish, Portuguese, Punjabi, Romanian, Russian, Samoan, Serbian, Slovak, Slovenian, Somali, Spanish, Swedish, Tahitian, Tajik, Tajiki, Tamil, Tatar, Telugu, Thai, Tibetan, Tigrinya, Tongan, Turkish, Turkmen, Ukrainian, Upper Sorbian, Urdu, Uyghur, Uzbek, Vietnamese, Welsh.

Note that the total count of language names is less than 128 as there are some languages for which we support multiple scripts or variants. For detailed information on language codes, see the `LANGUAGES` file in our repository, which is the most up-to-date version of this list.