Evaluating Machine Translation models using both Direct Assessment and automatic metrics

1st Semester of 2021-2022

Mihnea Mihai
mihnea.mihai7@s.unibuc.ro

Abstract

This report presents the evaluation of three pretrained machine translation models (**LibreTranslate**, **OPUS** and **MBart50**) for the language pairs **German** to **English** and **English** to **Romanian** with a test suite comprising several domains (**legal**, **medical**, **news** and **literary**) based both on self-provided Direct Assessment adequacy scores and automatic scores (**BLEU**).

1 Introduction

1.1 Models

This report compares several pretrained models in order to provide useful insights pertaining to their relative performance in various domains and help compare and align the hierarchy generated by direct assessment with the one generated by automatic evaluation methods.

1.1.1 LibreTranslate

LibreTranslate[1] is an open-source Machine Translation model providing a free public API. It does not rely on proprietary providers, instead it is powered by the open source **Argos Translate**[2] library, which in turn uses **OpenNMT**[3] (Klein et al., 2017) for translation.

1.1.2 OPUS

OPUS[4] (Tiedemann and Thottingal, 2020) is an open translation model based on **Marian-NMT**[5] (Junczys-Dowmunt et al., 2018) framework and trained on the **OPUS**[6] corpus.

1.1.3 MBart-50

MBart-50[7] (Liu et al., 2020) is a denoising autoencoder pretrained on large monolingual corpora using the BART (Lewis et al., 2020) objective.

1.2 Test suites

The evaluation is done on several test sets from different domains, in order to provide an accurate depiction of the models' performance. All sentences have been hand-picked in order to ensure variate distribution of long, short, context-dependant or context-independent sentences.

1.2.1 News

The news test set (**German>English**) is extracted from the news translation task[8] of the Sixth Conference on Machine Translation (WMT21). It contains news article excerpts originally in German with an English translation as reference.

1.2.2 Legal

The legal test set (**German>English**) is manually built based on German law and its translation into English provided by the German Federal Ministry of Justice[9], which makes it valuable as it comprises actual law jargon in current use in Germany and translations provided by law experts.

1.2.3 Medical

The medical test set (**English>Romanian**) is extracted from a multilingual corpus made of documents from the European Medicine Agency (EMA).

1.2.4 Literary

The literary test set (**English>Romanian**) is extracted from the MULTEX-East "1984" (Erjavec

[1] https://libretranslate.com/
[2] https://www.argosopentech.com/
[3] https://opennmt.net/
[4] https://github.com/Helsinki-NLP/OPUS-MT
[5] https://marian-nmt.github.io/
[6] https://opus.nlpl.eu/
[7] https://github.com/pytorch/fairseq/tree/main/examples/mbart
[8] https://github.com/wmt-conference/wmt21-news-systems/
[9] https://www.gesetze-im-internet.de/Teilliste_translations.html

et al., 2010) annotated corpus of the novel **1984** by George Orwell comprising the original in English and an official Romanian translation.

2 Approach

The code and processed data is available **here**.

The test sentences were extracted from the sources mentioned above and were processed into JSON format for ease of use and access.

In order to obtain the translations to be evaluated we used either the model's own API endpoint (in the case of LibreTranslate) or loaded the pretrained model from the **Huggingface**[10] framework.

The direct assessment scores for adequacy were provided manually in a simplified CLI interface.

In order to ensure reproducibility, the automation scores were computed using the Python library `sacrebleu` (Post, 2018).

2.1 Automatic metrics

As the manual assessment through annotators requires significant resources, automatic metrics are a possible alternative for

2.1.1 BLEU score

The BLEU score (Papineni et al., 2002) is the standard metric of automatically assessing machine translation output quality.

It counts the number of n-grams of the output also occurring in the reference translation(s) while dividing by the total number of n-grams in the reference(s).

The precision scores are computed for several values of n (usually 4) and their logarithms are averaged with uniform weights, which equivalent to the geometric mean.

Additional adjustments are required in order to prevent repeated n-grams to be counted more than their actual number of occurrences in the reference(s).

Lastly, brevity penalties are also applied for the case where most n-grams in the output do match the reference, but some n-grams from the reference are missing completely.

2.1.2 chrF score

The character n-gram F-score (Popović, 2015) extends the BLEU score by taking into consideration not only precision (n-grams in the output with counterparts in the reference), but also recall (n-grams in the reference with counterparts in the output).

[10] https://huggingface.co/

	BLEU	chrF	TER	DA
OPUS	**46.67**	**71.09**	**40.24**	**92.6**
MBart-50	25.84	57.00	62.02	70.5

Table 1: Medical domain (EN>RO)

	BLEU	chrF	TER	DA
OPUS	**19.64**	**43.68**	**77.39**	**71.52**
MBart-50	15.45	40.74	82.04	64.4

Table 2: Literary domain (EN>RO)

2.1.3 TER score

The Translation Edit Rate (Snover et al., 2006) is another automatic metric computing the edits needed to make the output identical to the reference, relative to the whole length of the reference.

3 Conclusions and Future Work

3.1 Correlation

Although automatic evaluation may not be ideal, we can observe a strong correlation between the direct assessment results and the metrics obtained automatically, which again proves their extreme usefulness.

3.2 Risks

In the MT evaluation task the risk of overfitting must be seriously taken into account, as some corpora are not split in test sets and train test and there is the possibility that a pretrained model already had knowledge of some test sentences.

3.3 Hierarchy

On average, the opensource **LibreTranslate** model performed below the other two models.

In some domains, **OPUS** clearly surpassed **MBart-50** in performance, whereas in others they are very similar (given also the relatively limited sample size).

3.4 Architecture difference

The fact that **MBart-50** has a noising step in its processing pipeline can be noticed by the several 100% fluent outputs which had however missing, conflicting or even random words.

References

Tomaž Erjavec, Ana-Maria Barbu, Ivan Derzhanski, Ludmila Dimitrova, Radovan Garabík, Nancy

	BLEU	chrF	TER	DA
LibreTranslate	22.85	52.25	58.99	64.61
OPUS	**32.80**	**61.34**	**52.27**	84.61
MBart-50	28.05	58.95	54.19	**86.34**

Table 3: News domain (DE>EN)

	BLEU	chrF	TER	DA
LibreTranslate	15.35	41.49	63.28	61.6
OPUS	21.24	**47.84**	59.56	**85.4**
MBart-50	**24.38**	51.37	**58.03**	83.5

Table 4: Legal domain (DE>EN)

Ide, Heiki-Jaan Kaalep, Natalia Kotsyba, Cvetana Krstev, Csaba Oravecz, Vladimír Petkevič, Greg Priest-Dorman, Behrang QasemiZadeh, Adam Radziszewski, Kiril Simov, Dan Tufiş, and Katerina Zdravkova. 2010. MULTEXT-east "1984" annotated corpus 4.0. Slovenian language resource repository CLARIN.SI.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation.

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

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

Matt Post. 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.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.