Machine Translation of Folktales: small-data-driven and LLM-based approaches

Olena Burda-Lassen, Ph.D.
Independent Research Scientist, Colorado, United States
oburdalassen@gmail.com

Abstract

Can Large Language Models translate texts with rich cultural elements? How "cultured" are they? This paper provides an overview of an experiment in Machine Translation of Ukrainian folktales using Large Language Models (Open AI), Google Cloud Translation API, and Opus MT. After benchmarking their performance, we have fine-tuned an Opus MT model on a domain-specific small dataset specially created to translate folktales from Ukrainian to English. We have also tested various prompt engineering techniques on the new Open AI models to generate translations of our test dataset (folktale 'The Mitten') and have observed promising results. This research explores the importance of both small data and Large Language Models in Machine Learning, specifically in Machine Translation of literary texts, on the example of Ukrainian folktales.

1 Introduction

"ChatGPT has already become a good translator" (Jiao et al., 2023) is an increasingly popular statement. We see an exponential increase in using Open AI models for various Machine Learning tasks and wanted to further explore this new tendency.

In addition to human translation, machine translation has undeniable potential in connecting people and cultures. Therefore, improving accuracy and accessibility to high-quality machine translation tools is very important.

We chose Ukrainian folktales for this experiment due to their unique nature and rich linguistical ecosystem. Folktales are usually passed on from one generation to another, going back hundreds and sometimes thousands of years, creating immense depth of knowledge and layers of cultural relevance.

The Ukrainian language has an extensive collection of myths, legends, proverbs, songs, and folktales. Even though these texts have literary translations available, many of them are rather transcreations, meaning that stories are retold and adapted to the target language and culture.

This experiment uses a recently created corpus of domain-specific curated parallel training data: Ukrainian-To-English Folktale Corpus (Burda-Lassen, 2022).

We wanted to use this curated corpus for fine-tuning machine translation models and see the impact on the accuracy and quality of translation.

2 Machine Translation Process

2.1 Overview of resources and machine translation models

For the creation of the Ukrainian-To-English Folktale Corpus, we used familiar Ukrainian folktales that were available in English: folktales from various websites for children's literature, blogs about Ukrainian traditions, bilingual children's books, as well as English translations from the Gutenberg Project[1].

Training a reliable machine translation system requires a large number of parallel sentences in two languages, which is often widely unavailable in low-resource language pairs (Sánchez-Cartagena

[1] https://www.gutenberg.org/cache/epub/29672/pg29672.txt

Proceedings of the 2023 CLASP Conference on Learning with Small Data, pages 68–71

et al., 2021). Even though Ukrainian is not considered a low-resource language anymore, the availability of the Ukrainian-To-English Folktale Corpus was very helpful in our experiment.

2.2 Applied Methods

Since the focus of our research was comparing the performance of 3 machine translation approaches and models for the translation of Ukrainian folktales, we have selected specific cultural terms that would be more domain-specific and, therefore, more challenging to translate.

While most common phrases are already being translated accurately by available machine translation engines, rare or cultural terms are often mistranslated or generalized. Adding an extra layer of culturally significant information can significantly improve the outcome of the translation process.

We have tested the translation at the word and text levels. We have chosen a subset of the words that have a high level of cultural sensitivity and are challenging to translate (Table 1).

For text-level translation, we have chosen as our test dataset the Ukrainian folktale 'The Mitten.' We have translated this folktale into English, using carefully selected human translation techniques, to preserve culturally specific elements, their meaning, and literary style. We have then translated this text using Google Cloud Translation API, the pre-trained model 'Helsinki-NLP/opus-mt-uk-en' and 2 Open AI models ('text-davinci-002' and the more recent 'gpt-3.5-turbo-16k'). We have also fine-tuned 'Helsinki-NLP/opus-mt-uk-en' on the Ukrainian-To-English Folktale Corpus and tested the accuracy of the translation by the fine-tuned version of the model. We have used sacreBLEU and BERTScore as evaluation metrics.

Ukrainian	English: human translation	Open AI: text-davinci-002	Google Cloud Translation API	Opus MT	Open AI: gpt-3.5-turbo-16k
мишка-шкряботушка	Scratchy-Mouse	Mouse-squeak	Mouse-scratcher	Roaster mouse	Bear-scratchy
жабка-скрекотушка	Croaky-Frog	toad-croaker	frog-scratch	snot frog	frog-squeaky
ведмідь-набрідь	Bear, the Wanderer	badger-beard	bear-bear	bear-brick	bear-wanderer
кабан-іклан	Boar-With-Tusks	wild boar-tusks	wild boar	boar-aklan	boar-tusk
коза-дереза	Bully Goat	goat-bristles	goat-skunk	goat-tree	goat-stick
солом'яний бичок	a little straw bull	straw-haired hog	straw bull	Somus-Treaby	Straw bull
бичок-третячок	Three-year-old bull calf	third-class hog	bull-tretyachok	monkey	bull-third
Мавка	Mavka, the forest spirit	fly-agaric	magpie	monkey	Mavka wood nymph
мед-вино	honey-wine	honey-wine	honey-wine	honey wine	honey-wine

Table 1: Translation examples of selected culture-specific terms.

2.3 Key Findings

After reviewing machine translation predictions, we can identify a few specific translation techniques and tendencies: calque (loan translation), generalization, and transcription.

As we can see in Table 1, many examples were mistranslated (especially the word to describe one of the female spirits in Ukrainian mythology *Mavka, the forest spirit*). If a term consisted of commonly known words, it was translated more precisely. Therefore, all tested models heavily rely on general language corpora and do not predict the values of specialized terminology, which formalizes the need for fine-tuning these models on small datasets that are very specific and rich in such valuable domain information.

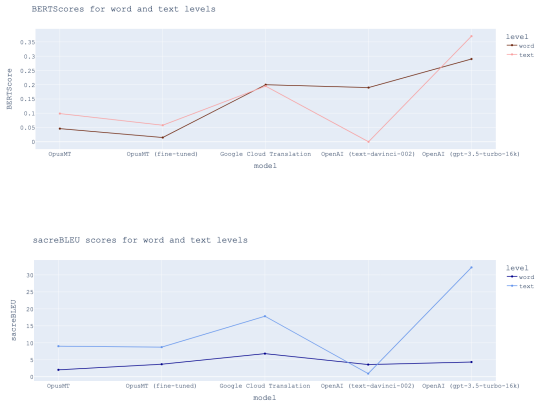

Table 2: BERTScore and sacreBLEU scores for models at text and word levels.

Word level measurements refer to the accuracy of translation of the specific subset of cultural terms, while text level accuracy evaluates the translation of the folktale 'The Mitten.'

We have observed an increase in sacreBLEU score on the word level after fine-tuning the Opus MT model (from *2.05* to *3.70*), and not on the text level. The highest accuracy in translating 'The Mitten' was observed by running predictions using the Open AI model 'gpt-3.5-turbo-16k'. We noticed that it outperformed all other models on the text level, and not on the word level (where Google Cloud Translation API was in the lead).

However, BERTScore showed the highest accuracy values for word and text levels for Open AI 'gpt-3.5-turbo-16k', with Google Cloud Translation API consistently showing second-best accuracy scores.

Additionally, we have noticed a rather significant difference between the two tested Open

AI models. While no notable accuracy improvement was observed from using 'text-davinci-002' at the word level translation, the translation of 'The Mitten' was unsuccessful: the story was simply summarized into eight sentences with most of the plot missing.

Open AI model 'gpt-3.5-turbo-16k', on the other hand, performed really well. We have applied several prompt engineering techniques, and the model was able to translate text according to the requirements of applying different styles. For example, when prompted with *'This text is for a children's book and needs to be fun and interactive.\ Create interesting compound words.'*, it created new compound words, very similar to the ones from the human translation. While prompted with '*This text is informational for English-speakers to learn more about Ukrainian culture*', the model translated all terms as a combination of two very appropriate human translation techniques: transliteration with an explanation.

Here are a few examples:

"Mishka-shkryabotushka" - A term used to describe a small bear that scratches itself.
"Zhabka-skrekotushka" - A term used to describe a small frog that makes a screeching sound.
"Vedmid-nabrid" - A term used to describe a bear that is annoying or bothersome.
"Kaban-iklan" - A term used to describe a boar with long tusks.
"Koza-dereza" - A term used to describe a goat with sharp horns.
"Solom'yanyi bychok" - A term used to describe a scarecrow made of straw.
"Bychok-tretyachok" - A term used to describe a small bull or young bull.
"Mavka" - A term used to describe a mythical creature from Ukrainian folklore, often depicted as a forest nymph or spirit.
"Med-vyno" - A term used to describe mead, an alcoholic beverage made from fermented honey.

Even though there is an error in translating the term '*Mishka-shkryabotushka'* (it is a small mouse, not a bear), with the correct transliteration being '*Myshka-shkryabotushka',* this definitely was an interesting machine translation output, which calls for further study and research.

While small datasets with domain-specific information can help train the traditional neural machine translation models and increase accuracy, especially if examples are carefully curated and hand-picked, Large Language Models have the potential to increase translation accuracy and create style-specific translations.

3 Conclusion

More research is necessary to increase the size of the Ukrainian-To-English Folktale Corpus to include a broader range of cultural terms, which will help further explore the preferable size of small data to make a more noticeable impact on accuracy score.

Since we have noticed an increase in translation accuracy at the word level after fine-tuning an Opus MT model, it would be valuable to explore the depth and volume of cultural terms needed to increase the accuracy score even further.

Another area of research could be prompt engineering and fine-tuning LLMs, while exploring their added benefit of creating machine translation tailored to specific literary styles.

Contrary to human translation of folklore, machine translation techniques must be more literal and descriptive. Therefore, a significant difference exists between human and machine translation techniques for folktales. That's where using a more informational translation style could be very valuable.

We believe that this type of research would be important for other language pairs as well. The domain of literary translation, specifically the translation of folklore and other culturally specific texts, is a vibrant environment full of fascinating challenges and great potential.

References

Eleftherios Avramidis, Marta R. Costa-jussà, Christian Federmann, Josef van Genabith, Maite Melero, and Pavel Pecina. 2012. A Richly Annotated, Multilingual Parallel Corpus for Hybrid Machine Translation. In Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), pages 2189–2193, Istanbul, Turkey. European Language Resources Association (ELRA).

Olena Burda-Lassen. 2022. Ukrainian-To-English Folktale Corpus: Parallel Corpus Creation and Augmentation for Machine Translation in Low-Resource Languages. In Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Workshop 2: Corpus Generation and Corpus Augmentation for Machine Translation), pages 28–31, None. Association for Machine Translation in the Americas.

Dmitriy Genzel, Jakob Uszkoreit, and Franz Och. 2010. "Poetic" Statistical Machine Translation: Rhyme and Meter. In Proceedings of the 2010 Conference on Empirical Methods in Natural

Language Processing, pages 158–166, Cambridge, MA. Association for Computational Linguistics.

Natalia Grabar, Kanishcheva Olga, Hamon Thierry. Multilingual aligned corpus with Ukrainian as the target language. SLAVICORP, Sep 2018, Prague, Czech Republic. ffhalshs-01968343f

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, and Zhaopeng Tu. 2023. Is chatgpt a good translator? a preliminary study. arXiv preprint arXiv:2301.08745.

Víctor M. Sánchez-Cartagena, Miquel Esplà-Gomis, Juan Antonio Pérez-Ortiz, and Felipe Sánchez-Martínez. 2021. Rethinking Data Augmentation for Low-Resource Neural Machine Translation: A Multi-Task Learning Approach. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 8502–8516, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021. WikiMatrix: Mining 135M Parallel Sentences in 1620 Language Pairs from Wikipedia. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pages 1351–1361, Online. Association for Computational Linguistics.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 2021. CCMatrix: Mining Billions of High-Quality Parallel Sentences on the Web. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 6490–6500, Online. Association for Computational Linguistics.

Yao, B., Jiang, M., Yang, D., & Hu, J. (2023). Empowering LLM-based Machine Translation with Cultural Awareness. arXiv preprint arXiv:2305.14328.

Yıldız, Eray & Tantuğ, Ahmet & Diri, Banu. (2014). The Effect of Parallel Corpus Quality vs Size in English - Toturkish SMT. Computer Science & Information Technology. 4. 21-30. 10.5121/csit.2014.4710.