SocialIQA_pt: A Translation of a Common-Sense Reasoning Dataset about Social Interactions

Fabio Grassiotto

30 June 2024

Abstract

The objective of this work is to create a Portuguese language translation of the English dataset Social IQa, a benchmark of 38,000 multiple choice questions for the evaluation of emotional and social intelligence in a number of everyday situations. The approach taken here was to perform machine translation using popular models available at Hugging Face, rank translations using a GPT-driven evaluation (GEMBA) and select the best for our dataset.

1 Introduction

Common-Sense intelligence, usually defined as the human ability of applying practical knowledge for decisions in everyday life, is still considered a challenging task for AI systems. As it can be readily perceived when interacting with chatbots, these systems lack the ability to, through intuition, reason about common situations and events. Such reasoning requires background knowledge about how the world works, including the rich nuanced interaction between people in the social sphere [1, 7].

Therefore, in artificial intelligence, the availability of datasets capable of benchmarking this task is of utmost importance. Datasets such as Social IQa are readily available in the English language, but we are not aware of any in Portuguese [10].

The Portuguese language cannot be truly considered a low-resource language as it is the case with some African and Asian languages, as there are datasets already available in Portuguese that have millions of tokens. There are, however, certainly gaps that should be addressed [4].

One of these gaps lies in the availability of datasets that deal with common sense-reasoning and in special datasets that address social interactions.

This work is structured as follows:

- Section 1 **(this section)**: Here we introduce the work and its motivation.
- Section 2 **Dataset**: We describe the SocialIQA original English language dataset.

- Section 3 **Methodology**: We describe the translation process and the evaluation system we employed.
- Section 4 **Results**: We describe the translated sets results and compare them in detail.
- Section 5 **Conclusion and Future Works**: We analyze the results we achieved and describe next steps to be taken.

2 Dataset

The Social Intelligence QA dataset was the first available resource upon its publication for the measurement of social and emotional intelligence for AI systems [3]. The dataset, collected using a crowd-sourced framework, is composed of around 38k multiple choice questions.

The Social IQa dataset is divided into 3 separate bases:

- Development, with 1954 questions/answers.
- Training, with 33410 questions/answers.
- Test, with 2224 questions/answers.

Each row of the dataset consists of the following sequences of characters:

- **Context** An observed context for a common situation.
- **Question** A question requiring reasoning about the social causes and effects of the observed situation.
- **Answer A, Answer B, Answer C** Multiple choices for answers. Besides one correct answer, the other two are collected from four negative options that were handwritten to be similar to the correct answers in terms of topic, length, and style but are subtly incorrect.
- **Correct** The expected correct answer to the question as determined by human peers.

An example of the types of questions and choices can be seen below on Table 1. These are the first few rows from the development section of the development section of the dataset.

Context	Question	Answer A	Answer B	Answer C	Correct
Tracy didn't go home that evening and resisted Riley's attacks.	What does Tracy need to do before this?	make a new plan	Go home and see Riley	Find somewhere to go	C
Sydney walked past a homeless woman asking for change but did not have any money they could give to her. Sydney felt bad afterwards.	How would you describe Sydney?	sympathetic	like a person who was unable to help	incredulous	A
Sasha protected the patients' rights by making new laws regarding cancer drug trials.	What will patients want to do next?	write new laws	get petitions signed	live longer	B
Jordan was in charge of taking the food on the camping trip and left all the food at home.	How would Jordan feel afterwards?	horrible that he let his friends down on the camping trip	happy that he doesn't need to do the cooking on the trip	very proud and accomplished about the camping trip	A

Table 1: Example rows from the SocialIQa English dataset.

3 Methodology

Our methodology consisted of two macro-phases, (1) machine translation using neural network models and (2) comparative evaluation of the translations we obtained using a large language model. After these macro-phases the best ranked translation for each row of the dataset was selected for our final dataset.

3.1 Machine Translation

Three machine translation models were selected based on their popularity (amount of downloads) from the Hugging Face website for this phase. The descriptions

below are taken verbatim from the model cards on the site.

3.1.1 Helsinki-NLP/opus-mt-tc-big-en-pt

> This model is part of the OPUS-MT project [11], an effort to make neural machine translation models widely available and accessible for many languages in the world. All models are originally trained using the amazing framework of Marian NMT, an efficient NMT implementation written in pure C++. The models have been converted to PyTorch using the transformers' library by hugging face. Training data is taken from OPUS and training pipelines use the procedures of OPUS-MT-train.

3.1.2 unicamp-dl/translation-en-pt-t5

> This repository brings an implementation of T5 for translation in PT-EN and EN-PT tasks using a modest hardware setup. We propose some changes in tokenizator and post-processing that improves the result and used a Portuguese pretrained model for the translation [9].

3.1.3 facebook/nllb-200-distilled-1.3B

> NLLB-200 is a machine translation model primarily intended for research in machine translation, especially for low-resource languages. It allows for single sentence translation among 200 languages [2].

3.2 Translation Evaluation

There are a number of alternatives for evaluating translations [8]. Evaluations using large language models have been producing results closer to human evaluation, and therefore we selected this method for this work.

The three translations were thus compared using the evaluation metric GEMBA - GPT Estimation Metric Based Assessment [6] using GPT3.5-turbo. To use the metric, the prompt was modified to not use human reference in the evaluation and to evaluate the translations of the three models of this work simultaneously.

The modified GEMBA prompt used to query GPT3.5 turbo was:

```
You are a helpful evaluator of the quality of translations.
Score the following translations from English to Portuguese on a
continuous scale from 0 to 100, where a score of zero means
"no meaning preserved" and score of one hundred means "perfect
meaning and grammar".

En: source_seg
Pt 1: target_seg1
Pt 2: target_seg2
```

```
Pt 3: target_seg3
Reply with only the number scores of your evaluation, in a
python list:
```

Where *source_seg* is the original language group of sentences, and *target_seg** are the proposed translations from each model. GPT-3.5-turbo replies to this prompt with a Python-style list with the rankings e.g. [80, 85, 100].

3.3 Workflow

Our workflow consisted of the five steps described in Figure 1 below. All steps are executed using Python notebook files available at.

- **Step I: English dataset conversion** In this step, the original English dataset was converted from JSON format into a temporary comma-delimited file format for easier processing. The conversion was performed on file read_dataset_en.ipynb.
- **Step II: Machine Translation** In this step, the original dataset, available in three separate files for the development, training and testing sets, was translated to the Portuguese language using the three distinct models described in 3.1. The implementations are on files translator_marian.ipynb, translator_t5.ipynb and translator_nllb.ipynb.
- **Step III: Translation Evaluation** In this step, translations were evaluated using the evaluation metric GEMBA as described in 3.2
- **Step IV: Translation selection and ranking** The highest ranked translation set was selected based on the metric described above and used in the dataset. The implementation for steps III and IV is executed on files evaluator_gemba_dev.ipynb, evaluator_gemba_train.ipynb and evaluator_gemba_tsyt.ipynb respectively for the development, training and test sets.
- **Step V: Portuguese Dataset publishing** In this step, the comma-delimited file with the final translation contents was converted to JSONL format for publishing. The conversion was performed on file publish_dataset_pt.ipynb.

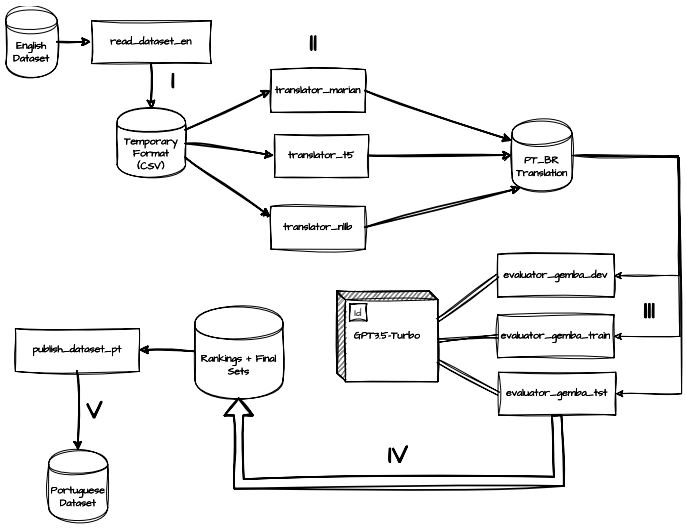

Figure 1: Translation process including machine translation and GPT-based evaluation to select the best possible translation.

4 Results

We have compared the results of the translations using the GEMBA method as can be seen below.

4.1 Selected Translations per Model

As can be seen on Figure 2, the model we selected most of our translations was the NLLB-1.3b from Meta. We believe this is due to the fact that the strengths of this model lie on the translations of short sentences. It is worthy noting that this is the largest model of the group.

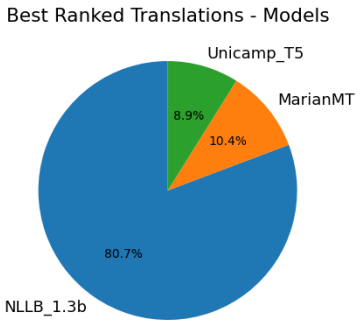

Figure 2: Selected Translations per Model.

4.2 Kernel Density Estimates

On figure 3 we plot the Kernel Density Estimates for the distribution of the ratings for each model. We note that NLLB-1.3b has the most evaluations towards the largest evaluations, while the T5 model has a more flat distribution.

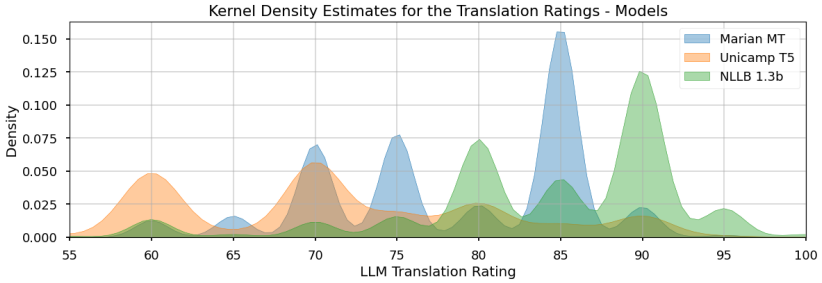

Figure 3: KDE distributions per Model.

4.3 Ratings per Sequence Length

We proceed then to evaluate the impact of sequence length on the translation quality in Figure 4. We did not notice appreciable difference among the models we used.

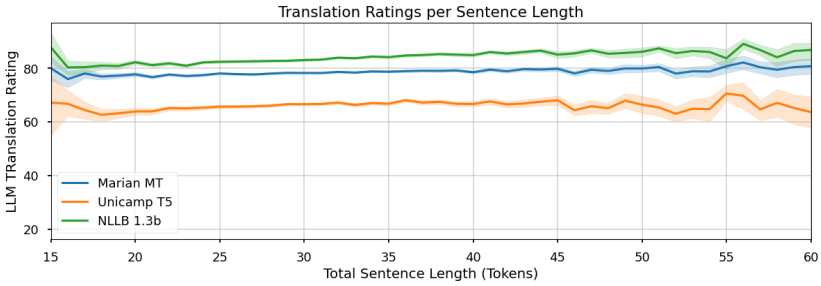

Figure 4: Rating per Sequence Length per Model.

5 Conclusion and Future Work

The Portuguese language dataset was published at Hugging face at [5].

We should note here some advantages and difficulties we found during the conduction of this work. First, machine translation models with low hardware requirements were necessary enablers for this work. All the translations were performed locally with a consumer-grade graphics card with 16Gb of video RAM. The usage of the GPT-3.5-turbo API in 2024 for a comparatively large dataset (38k lines) was only feasible due the low cost. For the translation evaluation, over 50 thousand requests and 16 million tokens were sent using the OpenAI API at an estimated total cost of less than US$ 10. It is worth noting, however, that OpenAI API speed was quite slow requiring over 50 hours for the whole process.

As future work, a thorough evaluation of the resulting dataset will be required to remove mistakes from the translation process. We note also that due to the fact that common-sense reasoning being very sensitive to cultural differences, a review of the dataset by Portuguese language speakers would be extremely valid for improving the quality of the translation results.

References

[1] Yejin Choi. The curious case of commonsense intelligence. *Daedalus*, 151(2):139–155, 2022.

[2] Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*, 2022.

[3] Allen Institute for AI. Social iqa. `https://allenai.org/data/socialiqa`, 2019. Accessed: 06/30/2024.

[4] Abdul Ghafoor, Ali Shariq Imran, Sher Muhammad Daudpota, Zenun Kastrati, Rakhi Batra, Mudasir Ahmad Wani, et al. The impact of translating

resource-rich datasets to low-resource languages through multi-lingual text processing. *IEEE Access*, 9:124478–124490, 2021.

[5] Fabio Grassiotto. Socialiqa dataset v1.4 (pt) at hugging face. `https://huggingface.co/datasets/fabiogr/social_i_qa_pt`, 2024. Accessed: 06/30/2024.

[6] Tom Kocmi and Christian Federmann. Large language models are state-of-the-art evaluators of translation quality. *arXiv preprint arXiv:2302.14520*, 2023.

[7] Stefanie Krause and Frieder Stolzenburg. Commonsense reasoning and explainable artificial intelligence using large language models. In *European Conference on Artificial Intelligence*, pages 302–319. Springer, 2023.

[8] Seungjun Lee, Jungseob Lee, Hyeonseok Moon, Chanjun Park, Jaehyung Seo, Sugyeong Eo, Seonmin Koo, and Heuiseok Lim. A survey on evaluation metrics for machine translation. *Mathematics*, 11(4):1006, 2023.

[9] Alexandre Lopes, Rodrigo Nogueira, Roberto Lotufo, and Helio Pedrini. Lite training strategies for portuguese-english and english-portuguese translation. *arXiv preprint arXiv:2008.08769*, 2020.

[10] Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. Socialiqa: Commonsense reasoning about social interactions. *arXiv preprint arXiv:1904.09728*, 2019.

[11] Jörg Tiedemann and Santhosh Thottingal. 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, 2020.