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 inteligence, 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. It is clear that such reasoning requires background knowledge about how the world works, incluind the rich nuanced interaction between people in the social sphere. [1, 4]

Therefore, the availability of datasets capable of benchmarking AI systems on 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. [5]

The portuguese language cannot be truly considered a low-resource language as it is the case with some of the African and Asian languages, as there are datasets already available in portuguese that have millions of tokens, but there are certainly gaps that should be addressed. [2]

One of the gaps is certainly in the availability of datasets that deal with common sense-reasoning and in special datasets that adress 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 analyse 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. The dataset, collected using a crowd-sourced framework, is comprised of around 38k multiple choice questions.

Each row of the dataset consists of:

- **Context** An observed context for a common situation.
- **Question** A question requiring inferential 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 determied by human peers.

An example of the types of questions and choices can be seen below on Table 1. Some of the sentences were shortened here for brevity.

Context	Question	Answer A	Answer B	Answer C	Correct
Tracy didn't go home that evening...	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 ...	How would you describe Sydney?	sympathetic	like a person...	incredulous	A
Sasha protected the patients' rights by making ...	What will patients want to do next?	write new laws	get petitions signed	live longer	B
Jordan was in charge of taking the food...	How would Jordan feel afterwards?	horrible that he...	happy that he doesn't need...	very proud ...	A

Table 1: Example rows from the SocialIQa English dataset.

3 Methodology

Our translation process 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 three distinct models on files translator_marian.ipynb, translator_t5.ipynb and translator_nllb.ipynb.
  - Helsinki-NLP/opus-mt-tc-big-en-pt
  - unicamp-dl/translation-en-pt-t5
  - facebook/nllb-200-distilled-1.3B
- **Step III: Translation Evaluation** Translations were evaluated using the evaluation metric GEMBA - GPT Estimation Metric Based Assessment [3] 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 f
```

```
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 english lannguage text, and target_seg1..3 are the candidate portuguese language translations.

- **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.

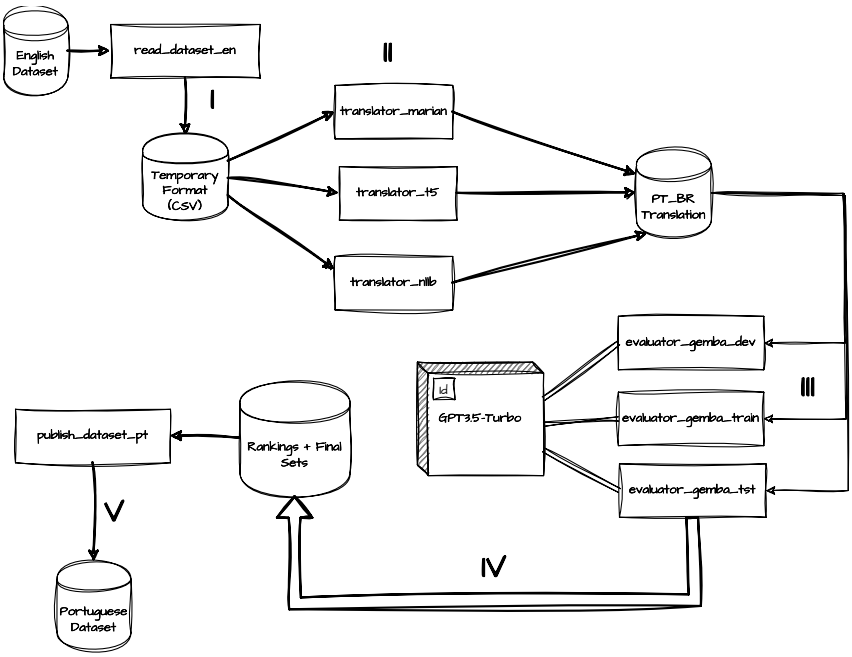

Figure 1: Translation process including machine translation and GPT-based evaluation to select the best possible translation.

4 Results

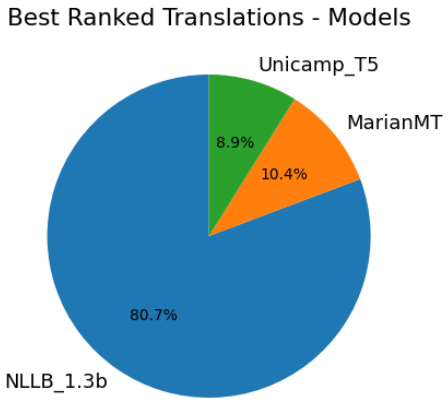

Figure 2: Figure example.

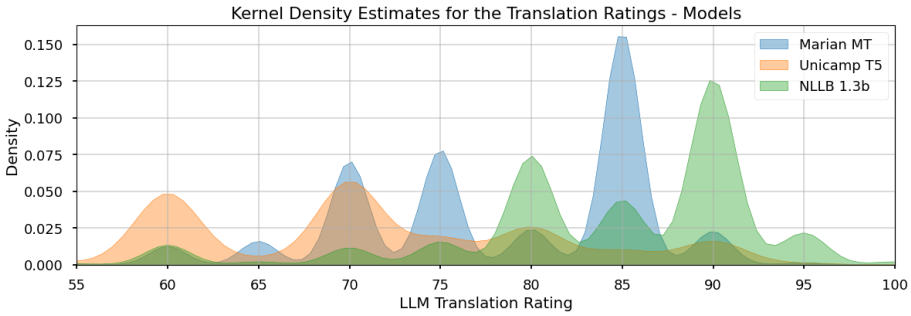

Figure 3: Figure example.

Figure 2 is an example of figure citation.

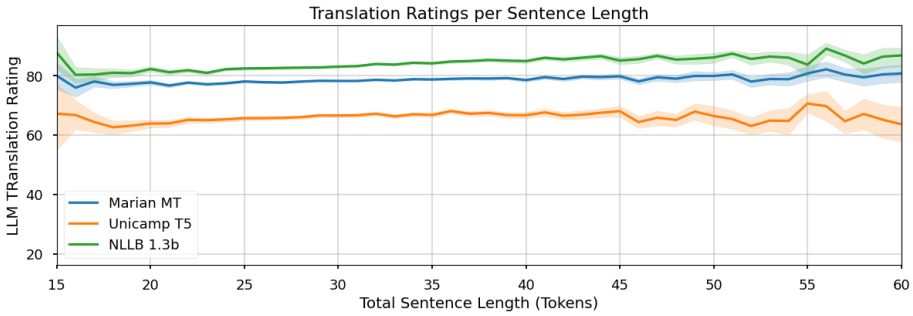

Figure 4: Figure example.

5 Conclusion and Future Work

References

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

[2] 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.

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

[4] 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.

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