MRL-EuropeanPT - Report

European Portuguese PIQA-style Dataset

MRL 2025 Shared Task on Multilingual Physical Reasoning Datasets

Spreadsheet with created/annotated examples:
MRL-EuropeanPT dataset in European Portuguese

Overall Description

The MRL-EuropeanPT dataset in European Portuguese is composed of 200 examples, of which 108 are original examples and 92 are translated to European Portuguese from the English PIQA dataset.

For the 108 examples that make-up the original portion of the dataset, there were eight contributors, all native speakers of European Portuguese. The aim was to include both general common sense examples as well as some relating to Portuguese culture (which included references to the preparation of traditional dishes, festivities, holidays, etc). After the dataset was completed, it was reviewed by two of the contributors.

There were two main types of items included:

- one is a question-answer format in which the prompt is a question answered by a solution pair, with only one of the solutions being the correct one;
- and the other is a sentence format in which the prompt is a sentence that is completed by a solution pair, with only one of the solutions being the correct one.

For the most part, the solution pairs being provided differ very little, with only one or two words being different.

Here are two examples:

{"prompt": "Na plataforma de uma estação de comboio, existe uma linha amarela ao longo do chão. Onde é recomendável os passageiros ficarem à espera do comboio?",
"solution0": "Atrás da linha amarela.",
"solution1": "À frente da linha amarela.",
"label": 0 }

{"prompt": "Para cortar o pão, o utensílio mais indicado é",
"solution0": "a colher, visto que a sua aresta serrilhada e afiada permite cortar o pão mais facilmente.",
"solution1": "a faca, visto que a sua aresta serrilhada e afiada permite cortar o pão mais facilmente.",
"label": 1 }

English translation of the examples:

{“prompt”: “On a train station platform, there is a yellow line along the floor. Where is it recommended that passengers wait for the train?”,
‘solution0’: “Behind the yellow line.”,
“solution1”: “In front of the yellow line.”,
“label”: 0 }

{“prompt”: “To cut bread, the most suitable tool is”,
“solution0”: “a spoon, since its serrated and sharp edge allows you to cut bread more easily.”,
“solution1”: “a knife, since its serrated and sharp edge allows you to cut bread more easily.”,
“label”: 1 }

The translated portion of the dataset, which includes 92 examples, was obtained through post-edited machine translation and revised by a translator.

Quality-control

Manual revision. All samples were subject to careful revision, with respect to punctuation, grammar, and overall structure, as well as checking each question and solution pair, making sure that the solution labelled as correct was accurate and made sense within the context of the dataset. This revision was carried out by two of the collaborators, both of whom are language specialists and European-Portuguese native speakers.

Human-Study. After the dataset was completed, it was stripped of its labels and delivered to two collaborators who did not participate in the development of the dataset. They then had to fill out the labels, basing their answers solely on their own common sense and reasoning. The results are as follows:

- Collaborator A was able to obtain a total score of 89.0%, with a higher score in the original examples (90.7%) than in the translated examples (87.0%).
- Collaborator B was able to obtain a total score of 89.0%, with a higher score in the original examples (95.4%) than in the translated examples (81.5%).

In total, the collaborators’ answers were both right in 82.5% of the answers and both wrong in 4.5% of the answers, with no match in 13.0% of the answers. For the original examples, the collaborators were both right in 88.9% of the answers and both wrong in 2.8% of the answers, with no match in 8.3% of the answers. For the translated examples, the collaborators were both right in 75.0% of the answers and both wrong in 6.5% of the answers, with no match in 18.5% of the answers.

As such, it is possible to conclude that for the original portion of the examples the collaborators obtained higher scores and were both right more often, and that for the translated portion of the examples they were both wrong or in disagreement more often, furthering the idea that PIQA-style datasets should be constructed in the intended language from inception.

Collaborators: Diogo Tavares <dc.tavares@campus.fct.unl.pt>, Inês Vieira <im.vieira@campus.fct.unl.pt>, Inês Calvo <inesmcalvo@gmail.com>, David Semedo <df.semedo@fct.unl.pt>, Afonso Simplício <am.simplicio@campus.fct.unl.pt>, Diogo Glória-Silva <dmgc.silva@campus.fct.unl.pt>, Rui Guerra <rp.guerra@fct.unl.pt>, Ana Condez <a.condez@campus.fct.unl.pt>, Pedro Valente <pm.valente@campus.fct.unl.pt>