Guidelines for Evaluating the Quality of Data and Labelling for FauxPas-EAI

Description of the annotation task

In this task you are asked to evaluate the correctness of data items and automatic labelling efficiency, from the FauxPas-EAI benchmark. FauxPas-EAI is a dataset based on the faux pas test, used in clinical psychology to test theory-of-mind and social intelligence in children. The objective of the annotation task is to evaluate the correctness of the data items and the efficiency of the automatic evaluation, based on the Levenshtein distance metric. The task is composed of two sub-tasks.

You are given some FauxPas-EAI items, composed of story, question, and ground truth answer, the model prediction, and the automatic evaluation score (1, if the model prediction matches the ground truth, and 0, otherwise).

Sub-task 1: Evaluating Data and Automatic Labelling

In the first sub-task, you are asked to evaluate each data item, with relative model prediction and automatic label, using three possible categories: *data issue, evaluation issue, and no issue*.

A) ***Data issues*** include issues related to the given story, question or ground truth answer.

B) ***Evaluation issues*** represent issues related to the automatic label, assigned using the Levenshtein distance to compare the ground truth answer and the model prediction.

C) ***No issues*** represent items for which there are no problems discussed above. For these the model can answer good or bad, but we are sure that it is not penalized in any way.

These issues are sorted in relevance order, since when an item contains data issues, it will be removed from the dataset.

Copyright (©) 2022 University of Trento, Signal and Interactive Systems Laboratory

Sub-task 2: Providing an explanation for the selected choice

Following the evaluation, you are asked to specify a note about your choice. The note should be between 3 to 30 words long. For example:

"Steve is the wrong agent executing the action"

or

"The answer 'to the supermarket' is valid compared to 'he droved to the supermarket'"

You will be asked for the note only for the first two types of issues, when there are no issues the comment is not mandatory.

Examples

Examples of Data Issues

FauxPas-EAI Item		
Context: Simon bought Robert a toy car for his birthday. A few months later, they were playing with it, and Simon dropped it. 'Don't worry,' said Robert, 'it was only an accident'. **Question:** Did Simon know Robert had given him the toy car for his birthday? Answer with 'Yes' or 'No' only, without explanations. In case of doubt, answer according to the most probable answer. **Ground truth answer:** Yes **Model prediction:** No		
Automatic Score	**Evaluation**	**Explanation**
0	Data Issue	Simon and Robert are in swapped order in the question compared to the context

Examples of Evaluation Issues

FauxPas-EAI Item
Context: Robert had just started at a new school. He said to his new friend, Andrew, 'My Mum is a dinner lady at this school.' Then Claire came over and said, 'I hate dinner ladies. They're horrible.' 'Do you want to come and play rounders?' Andrew asked Claire. 'No' she replied 'I'm not feeling very well. **Question:** What job does Robert's Mum do? Answer the question only, without explanations.

Copyright (©) 2022 University of Trento, Signal and Interactive Systems Laboratory

Ground truth answer: Robert's Mum is a dinner lady. **Model prediction:** She is a dinner lady at the school.

Automatic Score	Evaluation	Explanation
0	Evaluation Issue	The model answers correctly, but it's still penalized by the automatic evaluation

Examples of No Issues

FauxPas-EAI Item
Context: Kim helped her Mum make an apple pie for her uncle when he came to visit. She carried it out of the kitchen. "I made it just for you," said Kim. "Mmm", replied Uncle Tom, "That looks lovely. I love pies, except for apple, of course!" **Question:** What kind of pie had Kim made ? Answer the question only, without explanations. **Ground truth answer:** Kim made an apple pie. **Model prediction:** Apple pie.

Automatic Score	Evaluation	Explanation
1	No Issues	

Copyright (©) 2022 University of Trento, Signal and Interactive Systems Laboratory