Speaker Sensitive Response Evaluation Model

JinYeong Bak (jy.bak@kaist.ac.kr), Alice Oh (alice.oh@kaist.edu)

How can we train an AI model to automatically evaluate the appropriateness of responses for a given conversational context? A straightforward approach is to train the model from a sufficiently large human-labeled dataset, but that would be very expensive and time-consuming. Another approach is to train with a dataset of pairs of conversational context and appropriate response for that context, but it would be impossible to make all possible pairs because there is an unlimited number of appropriate responses for a given context. For example, in response to a suggestion to go to a movie, people can respond positively, negatively, or make other suggestions such as to go walking or shopping. With a dataset of limited pairs, the model would not be general enough to evaluate the appropriateness of unseen responses.

In this paper, we propose an automatic evaluation model that is able to evaluate the appropriateness of conversational responses by training from both appropriate and inappropriate responses. We create a response set that contains an appropriate response along with randomly selected inappropriate responses for a given conversational context. And we ask the model to identify the appropriate response in the set for training. This allows the model to learn the appropriateness by comparing the appropriate response with inappropriate responses.

To increase the accuracy and robustness of our model, we further augment the set by selecting inappropriate responses that are closer to the appropriate response than random. So we consider the speakers in defining the different levels of similarity. Utterances from the speaker who says the appropriate response would have similar topics and linguistic styles to the appropriate response. And utterances in the same conversation from the same speaker would be very similar to the appropriate response.

We test our model’s evaluation performance on two different casual conversation corpora - Twitter conversations and movie scripts. Our model outperforms the existing evaluation models in terms of correlation with human annotation scores. With this model, we can evaluate conversation models more like humans than before.