SLEDE: A classification-based evaluation model of second language English conversational dialogue

Anonymous ACL submission

Abstract

The majority of English conversation datasets are collected and evaluated from a native English perspective. English is considered the world's leading second language, with estimates ranging from 750 million to over a billion people speaking it as a non-native language. The requirements for second-language English data are increasing facing the demands of communicating with second-language English speakers. However, no datasets are available for demonstrating the quality of communicative features among English second language speakers in dialogue. Yet, some second-language English datasets are available for training purposes, such as International Corpus of Learners English (ICLE Ver. 2) with a certain focus on second-English written data; Spoken Learner Corpus was collected from second-language speakers in a monologic way. Still, none of the above datasets has focused on interactive dialogue in an open domain.

For dialogue quality evaluation, automatic metrics mainly focus on turn or utterance level quality, such as BLUE, METEOR, or ADEM, while lacking more comprehensive evaluation metrics on diversified features in dialogue level (e.g., engagement; feedback; topic development) of conversational datasets. Current metrics usually employ datasets containing human annotations that measure the quality of the responses ignoring the ability of generative responses from the whole dialogue level.

Up to date, massive research has been done to investigate label classification and prediction in dialogues and conversations. However, most spoken datasets used in dialogue evaluation studies focused on label classification instead of using labels to quantify the dialogue quality. The gap exists in the lack of a baseline classifier trained to predict labels further for dialogue quality evaluation. In spoken datasets, dialogue acts indicate the engagement and participation of interlocutors involved in the conversations. Thus, the requirement of labels related to dialogue acts from three levels is needed for further evaluation of dialogue quality.

The major contribution of this study is that we aim to propose a novel ESL dataset with annotations in three levels, including token levels, utterance levels, and dialogue levels. The dataset used in this study is a spoken dataset from stay-abroad ESL speakers in Australian universities. The data is suitable to address research questions in this study for analyzing the dialogue quality. The approximate amount of the recording is about 60 dialogues (around 60 hours in total), together with transcriptions in text formats and the full annotation.