Experiments on Non-native Speech Assessment and its Consistency

Ziwei Zhou[1], Sowmya Vajjala[2] and Seyed Vahed Mirnezami[1]

[1]Iowa State University, USA.
[2]National Research Council, Canada

8th NLP4CALL workshop
30th September 2019

What did we do?

- We did initial experiments on automated non-native speech assessment using a publicly available corpus.
- We looked into the consistency of the built models and the most predictive features in them.

What did we do?

- We did initial experiments on automated non-native speech assessment using a publicly available corpus.
- We looked into the consistency of the built models and the most predictive features in them.
- On one hand, it has all been done before with different resources.
- On the other hand, there is still something new to learn from these experiments.

Research Questions

1. RQ1: Which classifier performs the best in terms of agreement with human scorers when compared using multiple performance measures?
2. RQ2: How consistent are the machine scores rendered by the best performing model?
3. RQ3: What features are influential in predicting human scores?

Experimental Setup - Corpus

- International Corpus of Network of Asian Learners (ICNALE-Spoken)
- consists of oral responses provided by college students to two opinion-based prompts
- Proficiency is indicated on the CEFR: A2_0 (N=100), B1_1 (N=211), B1_2 (N=469), and B2_0 (N=160), by converting from other existing exam scores.
- In order to protect the participants' identity, speech samples were morphed using a speech morphing system

Experimental Setup - Features

- Fluency measures (e.g., number of pauses, speech rate, articulation rate etc.) using Praat
- Audio signal features (e.g., energy, spectral flux etc) using PyAudioAnalysis
- Lexical Richness features (Lu, 2012)
- Syntactic complexity features (Lu, 2014)

Experimental Setup - Approach

- classification models trained and tested separately for each prompt, which we call intrinsic evaluation (with 10-fold cross validation)
- classification models trained on one prompt, but tested on the other, which we call extrinsic evaluation.
- Synthetic Minority Oversampling Technique (SMOTE) to address data imbalance.
- Multiple evaluation measures: accuracy, precision, recall, F1-score, Cohen's Kappa, Quadratically Weighted Kappa, and Spearman correlation.
- 95% confidence intervals to check model consistency.

Results Summary

- The best-performing model with accuracy of about 73% for both prompts is achieved by using oversampling and random forests.
- The accuracies drop substantially for the oversampled data sets, but the accuracies for the non-oversampled versions remain consistent.
- Various feature selection schemes consistently pointed to the dominance of vocabulary related features for this classification task.

Limitations and Outlook

- We relied on manual transcriptions.. should look for automatic transcriptions in future work.
- Difference between over-sampled and non-oversampled models needs further exploration to understand whether it is experimental artefact or there is something else to it.
- Dataset limitations:
  1. Morphed speech samples
  2. Labeling of the dataset is done in an indirect way by converting scores from other existing tests, not by scoring these prompt responses.

  - we don't have a way to address these, but we hope easily accessible datasets of the future address such concerns.

Thank you.

contact: ziweizh@iastate.edu, sowmya.vajjala@nrc-cnrc.gc.ca