Off-topic spoken response detection for language assessment

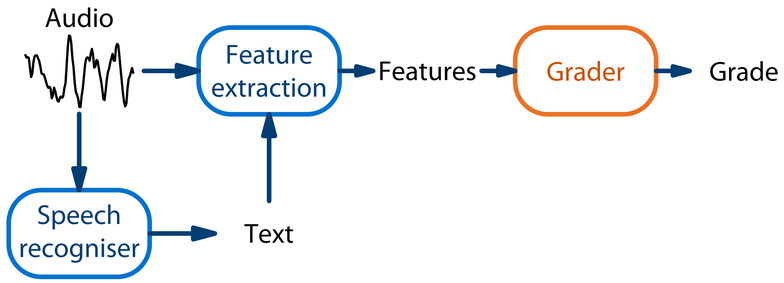
Andrey Malinin, Rogier van Dalen, Yu Wang, Kate Knill and Mark Gales
{am969,yw396,kate.knill,mjfg}@eng.cam.ac.uk

UNIVERSITY OF CAMBRIDGE

ALTA Institute / Department of Engineering, University of Cambridge

Introduction

Assessment of spoken English for language learners:
- Many people are learning English → need official qualifications.
- To help meet this demand:
  Automatic assessment of spontaneous spoken English.

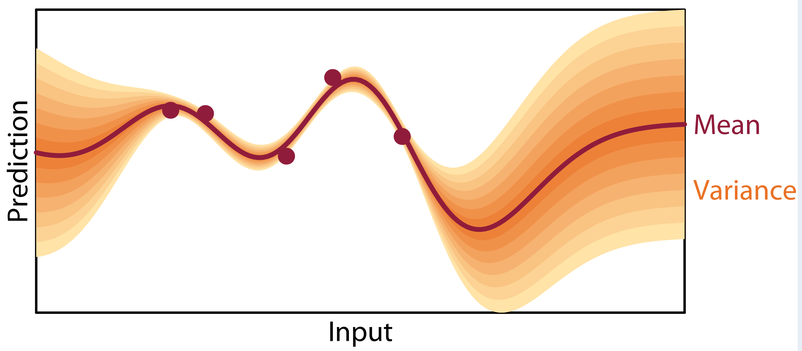


- An automatic grader is more consistent than humans graders.
- However, necessary to back-off to human graders when:
  - Grader has low confidence in the grade.
  - Topic of response is inappropriate for question.

Gaussian Process Grader

A Gaussian process is a non-parametric model
- Mean = prediction
- Variance = uncertainty about prediction → detect outliers

Reject low confidence grades:
- Pass to human graders

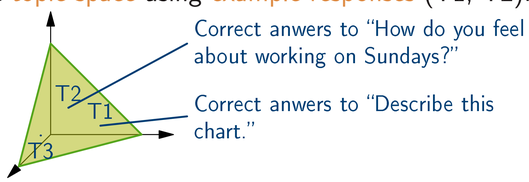


Uses fluency features → cannot detect if candidate:
- Failed to construct response
- Misunderstood question
- Gave memorized response.

Need to detect off-topic responses for assessment validity.

Standard Approaches to Topic Classification

1. Construct topic space using example responses (T1, T2).

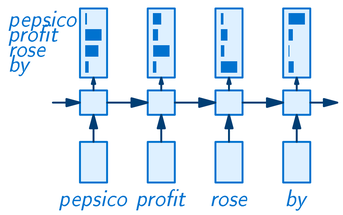


   Text → vector space representation
   - Latent Dirichlet Allocation (LDA)
   - Latent Semantic Analysis (LSA)
2. Classify using a (K) Nearest Neighbour classifier

Limitations:
- Complexity scales with training data size.
- Sequence information is not modelled.

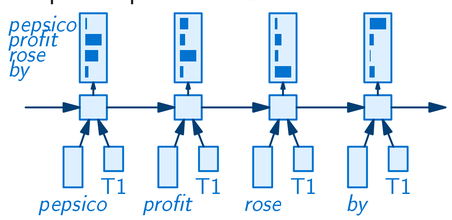
Statistical Language Model Topic Classification

A language model assigns a probability $\mathrm{P}(\mathbf{w})$ to a word sequence:
- $\mathrm{P}(\textit{pepsico profit rose by}) = 0.8$
- Implement as Recurrent Neural Network Language Model (RNNLM)



We want the language model to be topic conditional → $\mathrm{P}(\mathbf{w}|T_{\mathbf{w}})$:
- $\mathrm{P}(\textit{pepsico profit rose by} \mid T_1) = 0.9$
- $\mathrm{P}(\textit{pepsico profit rose by} \mid T_2) = 0.1$

Implemented as topic-adapted RNNLM:



- Use topic vector T_i as extra input
- Sentences which match topic vector have higher probability.
- Assign topic $\hat{T}_{\mathbf{w}}$ which maximizes $\mathrm{P}(\mathbf{w}|T_{\mathbf{w}})$
- Complexity scales with number of topics.

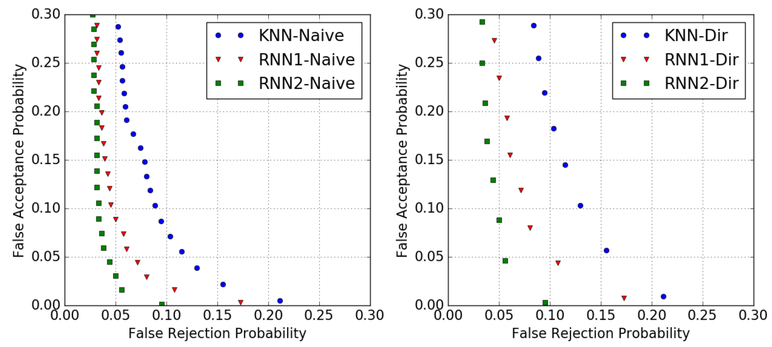
Data

Systems
- Train and test on 30% WER ASR transcriptions of responses
- Two training sets: 490 candidates, 10004 candidates
- Evaluate on 1560 candidates.

Experiments

Two part topic detection experiment:
- Detect topics of responses
- Detect off-topic responses
  - Data only contains valid responses. →
  - Randomly select responses to other questions
- How to select off-topic responses to a question? → Two strategies
  - Naive: responses from any test section
  - Directed: responses only from the same test section

System	Trn.Data # Cands.	% Equal Error Rate Directed	Naive
KNN	490	12.5	9.0
RNN1	490	8.0	6.0
RNN2	10K	5.0	4.5



Conclusions

Grading tests of spontaneous spoken English:
- Use a Gaussian process
- Use variance to detect outliers

Detect off-topic responses
- Create topic space using example responses
- Use topic adapted RNNLM to rank responses based on topics
- Work accepted in ACL 2016, Berlin.