# 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

## 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.

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

Reject low confidence grades:
- Pass to human graders

## Audio

Feature extraction

Text → vector space representation
- Latent Dirichlet Allocation (LDA)
- Latent Semantic Analysis (LSA)

## Text

Grader

Detect off-topic responses
- 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?

## Feature extraction

Speech recogniser

Input

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 \((T_1, T_2)\).
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 \(P(w)\) to a word sequence:
- \(P(\text{pepsi profit rose by}) = 0.8\)
- Implement as Recurrent Neural Network Language Model (RNNLM)

We want the language model to be topic conditional \(P(w | T_w)\):
- \(P(\text{pepsi profit rose by} | T_1) = 0.9\)
- \(P(\text{pepsi profit rose by} | T_2) = 0.1\)

Implemented as topic-adapted RNNLM:
- Use topic vector \(T_w\) as extra input
- Sentences which match topic vector have higher probability.
- Assign topic \(T_w\) which maximizes \(P(w | T_w)\)

## Gaussian Process Grader

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

## 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?

### 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.