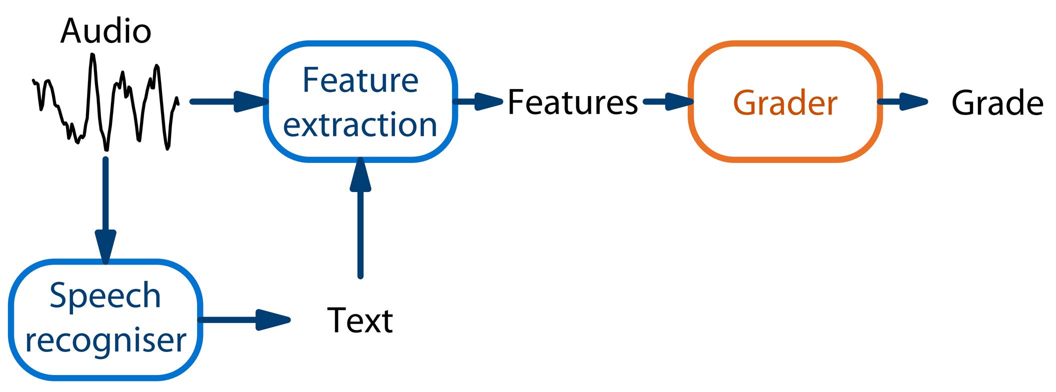
Off-topic spoken response detection for language assessment

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

1. Introduction

Assessment of spoken English for language learners:

- Many people are learning English → want official qualifications
- To help meet this demand:
  - Automatic assessment of spontaneous spoken English



- An automatic grader:
  - is more consistent than human graders
  - has significantly higher throughput
- Currently uses fluency features → cannot detect if candidate:
  - Failed to construct valid response to question
  - Misunderstood question
  - Gave a memorized response
- Need to detect off-topic responses before grading
  - Increases validity of automatic assessment
- Solution: Construct response topic classifier →
  - Reject and pass off-topic responses to human graders

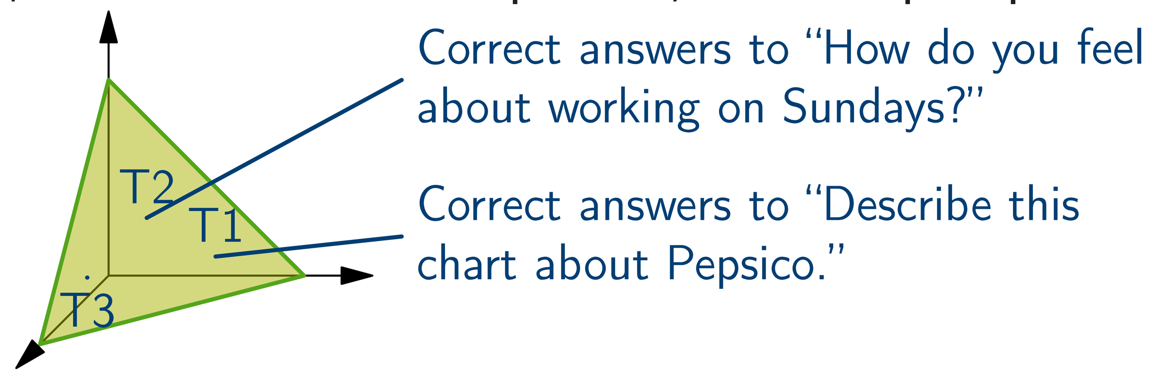
2. Data

- Experiments run on BULATS test data - 5 test sections, 21 question in total
  - Sections A and B - Simple questions and read-aloud
  - Sections C and E - Constructed response to open-ended questions
  - Section D - Describe and analyse a chart or graph

3. Topic Space Construction

1. Construct topic space using example responses
2. Assign a topic T_i to each question in the test
3. Concatenate all example responses belonging to the same topic T_i
4. Compute topic vectors $\mathbf{t}_i$ from concatenated example responses using:
   - Latent Dirichlet Allocation (LDA)
   - Latent Semantic Analysis (LSA)

- Each topic T_i is associated with a point $\mathbf{t}_i$ in the topic space



4. Standard Approaches to Topic Classification

- Training
  1. Construct topic space
  2. Project individual example responses into the existing topic space
     - Each topic T_i is associated with a cloud of points in topic space
     - Captures variability in responses to same question
- Inference - for each test response:
  1. Project the test response into the topic space
  2. Compute pairwise cosine distance with each topic vector
  3. Classify using a (K) Nearest Neighbour classifier
- Limitations:
  1. Inference time scales with training data size
  2. Sequence information is not modelled

5. Statistical Language Model Topic Classification

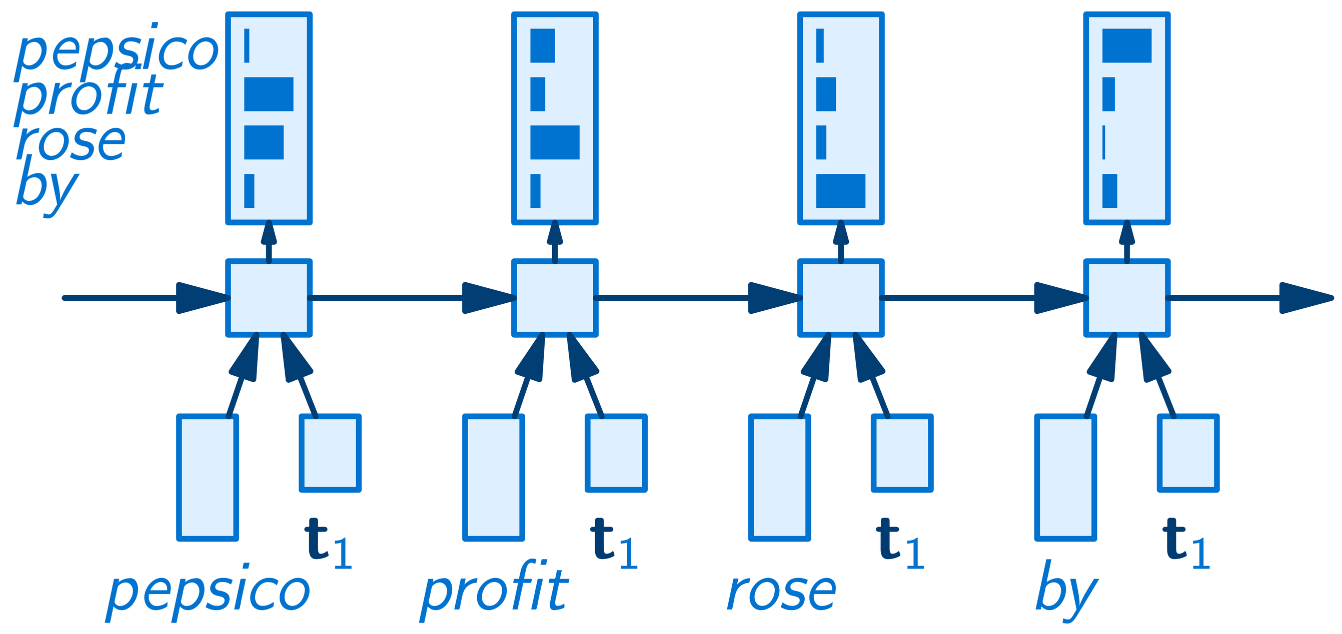
- A language model assigns a probability $\mathrm{P}(\mathbf{w})$ to a word sequence:

$$\mathrm{P}(\textit{pepsico profit rose by}) = 0.6$$

- Want the language model to be topic conditional $\mathrm{P}(\mathbf{w}|\mathrm{T}_{\mathbf{w}})$
- Sentences which match the topic should have a higher probability:

$$\mathrm{P}(\textit{pepsico profit rose by} \mid \mathrm{T}_1) = 0.9$$
$$\mathrm{P}(\textit{pepsico profit rose by} \mid \mathrm{T}_2) = 0.1$$

- Use topic-adapted Recurrent Neural Network Language Model (RNNLM):
  - Topic vector $\mathbf{t}_i$ is an extra input to the network



- Training - train RNNLM on:
  - Individual example responses $\mathbf{W}_i = \{\mathbf{w}_1^i, \cdots, \mathbf{w}_M^i\}$ for all topics T_i
  - Concatenated example response topic vectors $\mathbf{t}_i \; \forall i \in \{1, \cdots, N\}$
- Inference - for each test response $\mathbf{w}$ assign $\hat{\mathrm{T}}_{\mathbf{w}} = \arg\max_i \mathrm{P}(\mathbf{w}|\mathrm{T}_i)$
- Advantages:
  - Inference time scales with number of topics
  - Sequence information explicitly modelled

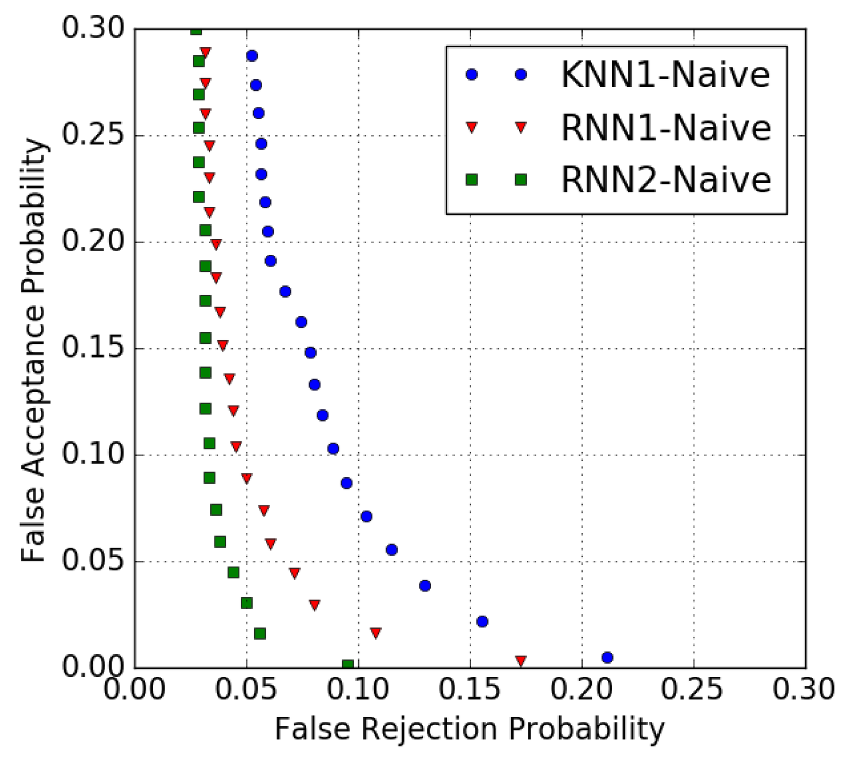
6. Experiment configuration

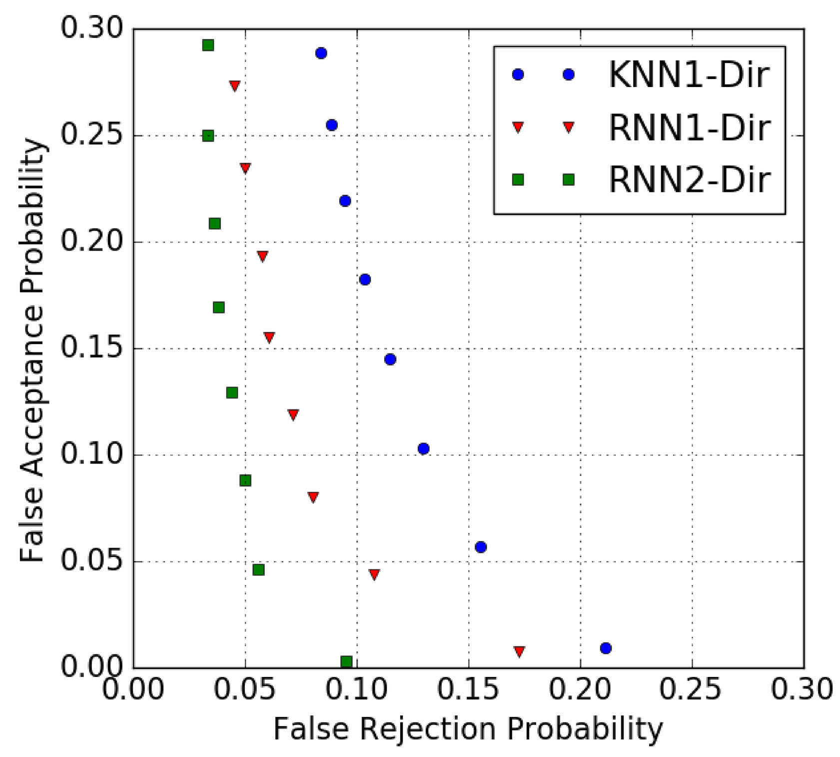
- Train and test on 30% WER ASR transcriptions of responses
- LSA topic space covering 282 topics
- Two training sets:
  1. 490 candidates → models KNN1, RNN1
  2. 10004 candidates → model RNN2
- Evaluate on 1560 candidates

7. Experiments

- Two part topic detection experiment:
  1. Detect topics of responses (ground truth known)
  2. 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
KNN1	490	12.5	9.0
RNN1	490	8.0	6.0
RNN2	10004	5.0	4.5





8. Conclusions

- Detect off-topic responses before grading
  - Increases validity of automatic assessment
- Use topic adapted RNNLM to classify response topics
  - Outperforms standard approaches
  - Scales to large datasets without affecting inference time
  - Can take advantage of progress in RNNLMs and Deep Learning