Off-topic spoken response detection for language assessment

Andrey Malinin, Rogier van Dalen, Yu Wang, Kate Knill and Mark Gales

(amm969,yw396,kate.knill,mjfg)@eng.cam.ac.uk

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 \( t_i \) from concatenated example responses using:
   - Latent Dirichlet Allocation (LDA)
   - Latent Semantic Analysis (LSA)
   - Each topic \( T_i \) is associated with a point \( t_i \) in the topic space

Correct answers to “How do you feel about working on Sundays?”
Correct answers to “Describe this chart about Pepsico.”

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 \( P(w) \) to a word sequence:
  \[
P(pepsico \text{ profit} \text{ rose by}) = 0.6
  \]
- Want the language model to be topic conditional \( P(w|T_w) \)
- Sentences which match the topic should have a higher probability:
  \[
P(pepsico \text{ profit} \text{ rose by} | T_1) = 0.9
  \]
  \[
P(pepsico \text{ profit} \text{ rose by} | T_2) = 0.1
  \]
- Use topic-adapted Recurrent Neural Network Language Model (RNNLM):
  - A topic vector \( \vec{t}_i \) is an extra input to the network

Training - train RNNLM on:
- Individual example responses \( W_i = \{w_{i1}, \ldots, w_{iN} \} \) for all topics \( T_i \)
- Concatenated example response topic vectors \( \vec{t}_i \forall i \in \{1, \ldots, N\} \)
- Inference - for each test response \( w \) assign \( \hat{T}_w = \arg \max P(w|T_w) \)
- 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

<table>
<thead>
<tr>
<th>System</th>
<th>Trn.Data</th>
<th>% Equal Error Rate</th>
<th>Directed</th>
<th>Naive</th>
</tr>
</thead>
<tbody>
<tr>
<td>KNN1</td>
<td>490</td>
<td>12.5</td>
<td>9.0</td>
<td></td>
</tr>
<tr>
<td>RNN1</td>
<td>490</td>
<td>8.0</td>
<td>6.0</td>
<td></td>
</tr>
<tr>
<td>RNN2</td>
<td>10004</td>
<td>5.0</td>
<td>4.5</td>
<td></td>
</tr>
</tbody>
</table>

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