An Attention-based model for off-topic spontaneous spoken response detection: An initial study

Andrey Malinin, Kate Knill, Anton Ragni, Yu Wang and Mark J. F. Gales

30 August 2017
Automating learning/assessment increasingly popular
  - cheap, and easily scalable
  - state-of-the-art auto-marking “good” performance

Many challenges still remain: for example
  - exploiting knowledge that an auto-marker is used

This work examines “off-topic” response detection:
  - candidate unable to formulate valid response
  - candidate does not understand the prompt/question
  - candidate deliberately “cheats”
Automatic Assessment Pipeline

Audio

Feature extraction

Features

Grader

Grade

Speech recogniser

Text

Off-Topic Response Detection
Off-topic response (relevance) takes:

- \( w^p \): prompt (question) from script
  \[ w^p = \{ \text{Discuss a company that you admire} \} \]

- \( w^r \): response from candidate derived from speech recognition
  \[ w^r = \{ \text{Cambridge Assessment is wonderful, it ...} \} \]

and derives probability of relevance

\[ P(\text{rel}|w^r, w^p) \]

Two standard options for model:

- Generative Model of Responses
- Discriminative Model of Relevance
Generative Model of Responses

- Prob. response given prompt: $P(w^r|w^p) \approx P(w^r|t_p)$
- Then probability of relevance derived from:

$$P(\text{rel}|w^r, w^p) \approx P(w^p|w^r) \approx P(t_p|w^r) = \frac{P(w^r|t_p)P(t_p)}{\sum_i P(w^r|t_p)P(t_p)}$$
Generative Model of Responses

- Simple intuitive model
  - leverage state-of-the-art language model technology
  - but two stage process ...
- Requires a “prompt” representation $w^p \rightarrow t_p$
  - standard approaches: LSA/LDA
  - projecting prompt “training” responses into space

$$w^p \rightarrow \{w^r_1, \ldots, w^r_N\} \rightarrow t_p$$

- requires example responses

- Model does not directly give probability of relevance
  - a response may be relevant to multiple prompts ...
Discriminative Response Model

- Directly model the probability of relevance

\[ P(\text{rel}| \mathbf{w}^r, \mathbf{w}^p) \]

- Split the process into sequence of steps:
  1. \( \mathbf{w}^p \rightarrow \tilde{\mathbf{h}}^p \): prompt embedding
  2. \( \mathbf{w}^r|\tilde{\mathbf{h}}^p \rightarrow \mathbf{c}^r \): response encoding (given prompt encoding)
  3. \( P(\text{rel}|\mathbf{c}^r) \): probability of relevance

- Each of these stages will be discussed below
Prompt Embedding

- Embedding of prompt $\tilde{h}^p$ using Bidirectional LSTM

$$w^p = w_1^p, \cdots, w_L^p \rightarrow \begin{bmatrix} h_{L}^p \\ h_{1}^p \end{bmatrix} = \tilde{h}^p$$

- Fixed-Length embedding (first and last embeddings)
Response Encoding: Embedding

- Response sequence into embedding sequence using Bi-LSTM

\[ w^r = w_1^r, \ldots, w_T^r \rightarrow \begin{bmatrix} \overrightarrow{h_1^r} \\ \overleftarrow{h_1^r} \end{bmatrix} \cdots \begin{bmatrix} \overrightarrow{h_T^r} \\ \overleftarrow{h_T^r} \end{bmatrix} = \tilde{h}_1^r: \tilde{h}_T^r \]

- resulting sequence same length as the response (in words)
- Map embedded response sequence to a fixed length vector
- Vary “importance” of words

**Attention mechanism**

- Distribution over embeddings ($\alpha_t$)

$$c^r = \sum_{t=1}^{T} \alpha_t \tilde{h}_t^r; \quad \alpha_t = f(\tilde{h}^p, \tilde{h}_t^r); \quad \sum_{t=1}^{T} \alpha_t = 1$$
Probability of Relevance

- Need to map from the fixed length vector to relevance

\[ P(\text{rel}|\mathbf{w}', \mathbf{w}^P) = P(\text{rel}|\mathbf{c}') = f(\mathbf{c}') \]

- standard mapping process
- use deep neural network to perform mapping

- Parameters optimised for relevance: requires
  - relevant (positive) training responses
  - not relevant (negative) training responses

- The prompt embedding can be applied to any prompt
  - naturally handles unseen (in training data) prompts
BULATS Spoken Tests

- BULATS - low-stakes test of communication skills:
  A Introductory Questions: your name where you are from
  B Read Aloud: read specific sentences
  C Topic Discussion: discuss a company that you admire

D Interpret and Discuss Chart/Slide: example above
E Answer Topic Questions: 5 questions about organising a meeting

- Only sections C-E of interest for this experiment
Experimental setup: Data

<table>
<thead>
<tr>
<th>Data</th>
<th>#Prompt</th>
<th>#Resp.</th>
<th>#Resp. Prompt</th>
</tr>
</thead>
<tbody>
<tr>
<td>TRN</td>
<td>379</td>
<td>292.9K</td>
<td>772.8</td>
</tr>
<tr>
<td>EVAL</td>
<td>222</td>
<td>4.3K</td>
<td>18.6</td>
</tr>
</tbody>
</table>

- Data partitioned:
  - **TRN** covers a wide range of L1 languages
  - **EVAL** covers 8 L1 languages
- EVAL approximately uniform over CEFR grade (merge C1/C2)
- Training data responses assumed to be valid
  - randomly select negative responses from other prompts
  - all EVAL prompts seen in TRN
  - some responses later found not to be valid
Experimental setup: ASR

- First stage in process - Speech Recognition
  - non-native speakers, wide-range of levels

<table>
<thead>
<tr>
<th></th>
<th>A1</th>
<th>A2</th>
<th>B1</th>
<th>B2</th>
<th>C</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>60.3</td>
<td>54.0</td>
<td>44.9</td>
<td>41.8</td>
<td>41.4</td>
<td>45.7</td>
</tr>
</tbody>
</table>

- For these experiments only baseline system used
  - DNN acoustic models trained on \(\approx 100\) hours (Gujarati L1)
  - N-gram language model (interpolated with general LM)
- High word error rates
  - but system trained on ASR output
**Experimental setup: Performance Metric**

- **Performance metric** - Area Under ROC Curve (AUC)
  - ROC curve → true positive vs. false positive
Results: Prompts all Seen

<table>
<thead>
<tr>
<th></th>
<th>A1</th>
<th>A2</th>
<th>B1</th>
<th>B2</th>
<th>C</th>
<th>ALL</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.88</td>
<td>0.94</td>
<td>0.94</td>
<td>0.97</td>
<td>0.97</td>
<td>0.95</td>
</tr>
</tbody>
</table>

- Performance overall high (0.95)
  - all evaluation prompts seen in training
  - examples of positive responses seen in training
  - all negative responses taken from seen prompts
- Performance as expected with CEFR level
  - easier to detect relevance with higher grade level
- Not realistic for many scenarios
Experimental Setup: Unseen Prompts

- No unseen topics ➔
  - Hold out subset of topics from data

<table>
<thead>
<tr>
<th>Data</th>
<th>#Topics</th>
<th>#Resp.</th>
</tr>
</thead>
<tbody>
<tr>
<td>TRN-fixed</td>
<td>178</td>
<td>142.8K</td>
</tr>
<tr>
<td>TRN-xVal</td>
<td>201</td>
<td>150.1K</td>
</tr>
<tr>
<td>EVAL-sub</td>
<td>201</td>
<td>2955</td>
</tr>
</tbody>
</table>

- TRN-fixed always used
- TRN-xVal used in 10-fold cross “training”

- Average performance on EVAL-SUB
Results: Seen & Unseen Prompts

- Partition results in 4 ways - aspects of generalisation
  - prompts can be **Seen** or **Unseen**
  - negative responses can relate to **Seen** to **Unseen** prompts

<table>
<thead>
<tr>
<th>Prompts</th>
<th>Neg. Resp. relating to Seen Prompts</th>
<th>Unseen Prompts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seen</td>
<td>0.92</td>
<td>0.92</td>
</tr>
<tr>
<td>Unseen</td>
<td>0.78</td>
<td>0.72</td>
</tr>
</tbody>
</table>

- **Seen** prompts not sensitive to nature of response
  - **Unseen** prompts model performance drop - more sensitive
Results: Seen & Unseen ROC Curves

- ROC curve for performance with **Seen** and **Unseen** prompts
  - against balanced set of seen/unseen prompt responses
Conclusions

- Presented deep-learning-based off-topic response detection
  - uses:
    - bidirectional LSTM embeddings
    - attention based response encoding

- Initial evaluation on “spontaneous” speech
  - good performance when prompts seen in training data
  - “reasonable” performance when prompts unseen in training

- Issues with training/test configuration currently used
  - assumption that all responses are relevant
  - artificially generated “off-topic”
Thanks!
Questions
Logistic Loss - needs
- Positive/Negative Examples

Trained with SGD
- Adam Optimizer
- Initial Learning Rate 1e-3
- Decay Learning Rate
- 5 epochs of training

Dropout Regularization
- Feed-forward only