UNIVERSITY OF CAMBRIDGE

Cambridge ALTA
Institute for Automated Language Teaching and Assessment

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

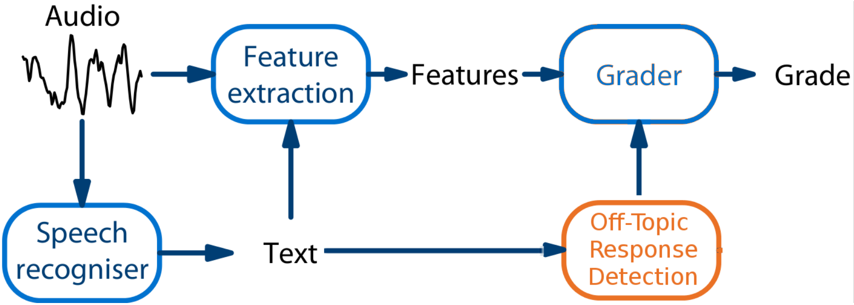
Andrey Malinin, Kate Knill, Anton Ragni, Yu Wang and Mark J. F. Gales

30 August 2017

Spoken Language Learning and Assessment

- 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



Relevance Detection

- Off-topic response (relevance) takes:
  - $\boldsymbol{w}^p$: prompt (question) from script

    $$\boldsymbol{w}^p = \{\texttt{Discuss a company that you admire}\}$$

  - $\boldsymbol{w}^r$: response from candidate derived from speech recognition

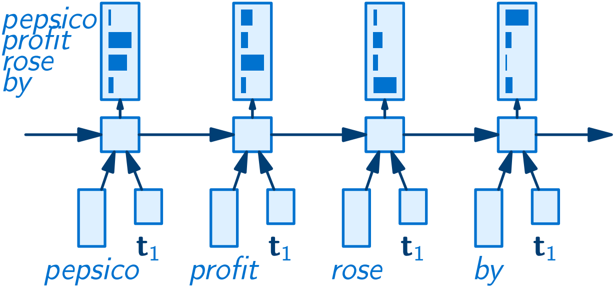
    $$\boldsymbol{w}^r = \{\texttt{Cambridge Assessment is wonderful, it ...}\}$$

  and derives probability of relevance

$$\mathtt{P}(\mathtt{rel}|\boldsymbol{w}^r, \boldsymbol{w}^p)$$

- Two standard options for model:
  - Generative Model of Responses
  - Discriminative Model of Relevance

Generative Model of Responses



- Prob. response given prompt: $\mathrm{P}(\boldsymbol{w}^r | \boldsymbol{w}^p) \approx \mathrm{P}(\boldsymbol{w}^r | \mathbf{t}_p)$
- Then probability of relevance derived from:

$$\mathrm{P}(\mathtt{rel} | \boldsymbol{w}^r, \boldsymbol{w}^p) \approx \mathrm{P}(\boldsymbol{w}^p | \boldsymbol{w}^r) \approx \mathrm{P}(\mathbf{t}_p | \boldsymbol{w}^r) = \frac{\mathrm{P}(\boldsymbol{w}^r | \mathbf{t}_p)\mathrm{P}(\mathbf{t}_p)}{\sum_i \mathrm{P}(\boldsymbol{w}^r | \mathbf{t}_p)\mathrm{P}(\mathbf{t}_p)}$$

Generative Model of Responses

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

$$\boldsymbol{w}^p \rightarrow \{\boldsymbol{w}_1^r, \ldots, \boldsymbol{w}_N^r\} \rightarrow \boldsymbol{t}_p$$

  - requires example responses
- Model does not directly give probability of relevance
  - a response may be relevant to multiple prompts ...

Discriminative Response Model

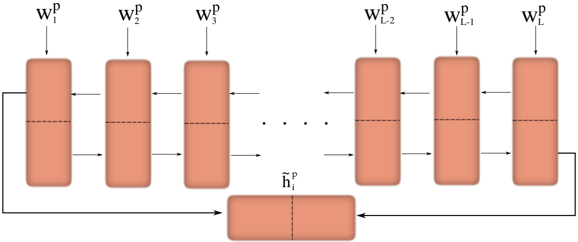
- Directly model the probability of relevance

$$\mathtt{P}(\mathtt{rel}|\boldsymbol{w}^r, \boldsymbol{w}^p)$$

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

- Each of these stages will be discussed below

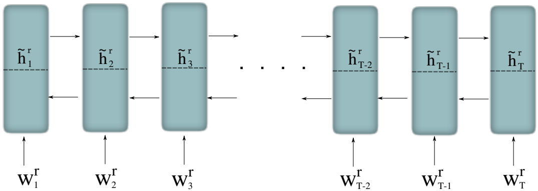
Prompt Embedding



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

$$\boldsymbol{w}^p = w_1^p, \cdots, w_L^p \rightarrow \begin{bmatrix} \overrightarrow{\boldsymbol{h}}^p_L \\ \overleftarrow{\boldsymbol{h}}^p_1 \end{bmatrix} = \tilde{\boldsymbol{h}}^p$$

- Fixed-Length embedding (first and last embeddings)



- Response sequence into embedding sequence using Bi-LSTM

$$\boldsymbol{w}^r = w_1^r, \cdots, w_T^r \rightarrow \begin{bmatrix} \overrightarrow{\boldsymbol{h}}^r_1 \\ \overleftarrow{\boldsymbol{h}}^r_1 \end{bmatrix} \cdots \begin{bmatrix} \overrightarrow{\boldsymbol{h}}^r_T \\ \overleftarrow{\boldsymbol{h}}^r_T \end{bmatrix} = \tilde{\boldsymbol{h}}^r_{1:T}$$

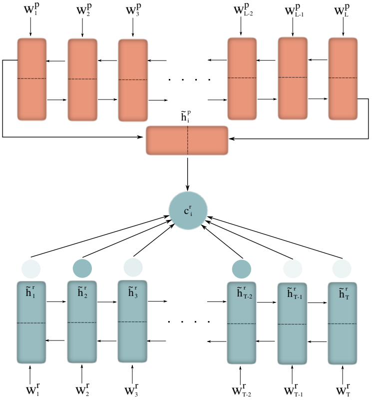
- resulting sequence same length as the response (in words)

Response Encoding: Attention Mechanism

- Map embedded response sequence to a fixed length vector
- Vary "importance" of words

Attention mechanism

- Distribution over embeddings (α_t)



$$\boldsymbol{c}^r = \sum_{t=1}^{T} \alpha_t \tilde{\boldsymbol{h}}^r_t; \qquad \alpha_t = f(\tilde{\boldsymbol{h}}^p, \tilde{\boldsymbol{h}}^r_t); \qquad \sum_{t=1}^{T} \alpha_t = 1$$

Probability of Relevance

- Need to map from the fixed length vector to relevance

$$\mathtt{P}(\mathtt{rel}|\boldsymbol{w}^r, \boldsymbol{w}^p) = \mathtt{P}(\mathtt{rel}|\boldsymbol{c}^r) = f(\boldsymbol{c}^r)$$

  - 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

Data	#Prompt	#Resp.	#Resp. Prompt
TRN	379	292.9K	772.8
EVAL	222	4.3K	18.6

- 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

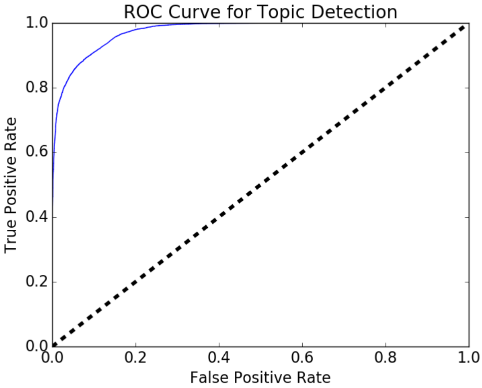
Experimental setup: ASR

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

A1	A2	B1	B2	C	All
60.3	54.0	44.9	41.8	41.4	45.7

- For these experiments only baseline system used
  - DNN acoustic models trained on ≈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

A1	A2	B1	B2	C	ALL
0.88	0.94	0.94	0.97	0.97	0.95

- 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

Data	#Topics	#Resp.
TRN-fixed	178	142.8K
TRN-xVal	201	150.1K
EVAL-sub	201	2955

  - TRN-fixed always used
  - TRN-xVal used in 10-fold cross "training"
- Average performance on EVAL-SUB

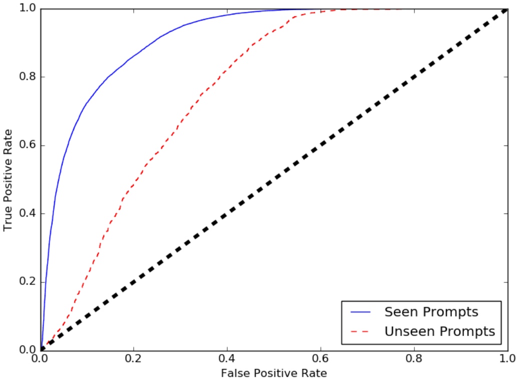
Results: Seen & Unseen Prompts

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

Prompts	Neg. Resp. relating to Seen Prompts	Neg. Resp. relating to Unseen Prompts
Seen	0.92	0.92
Unseen	0.78	0.72

- 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

Attention Based Topic Model

Logistic Loss - needs

- Postive/Negative Examples

Trained with SGD

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

Dropout Regularization

- Feed-forward only