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

University of Cambridge, Department of Engineering,
Trumpington St, Cambridge CB2 1PZ, UK
{am969, mjfg}@eng.cam.ac.uk

Abstract

Automatic spoken language assessment systems are gaining popularity due to the rising demand for English second language learning. Current systems primarily assess fluency and pronunciation, rather than semantic content and relevance of a candidate's response to a prompt. However, to increase reliability and robustness, relevance assessment and off-topic response detection are desirable, particularly for spontaneous spoken responses to open-ended prompts. Previously proposed approaches usually require prompt-response pairs for all prompts. This limits flexibility as example responses are required whenever a new test prompt is introduced.

This paper presents a initial study of an attention based neural model which assesses the relevance of prompt-response pairs without the need to see them in training. This model uses a bidirectional Recurrent Neural Network (BiRNN) embedding of the prompt to compute attention over the hidden states of a BiRNN embedding of the response. The resulting fixed-length embedding is fed into a binary classifier to predict relevance of the response. Due to a lack of off-topic responses, negative examples for both training and evaluation are created by randomly shuffling prompts and responses. On spontaneous spoken data this system is able to assess relevance to both seen and unseen prompts.

Index Terms: Spoken Language Assessment, Relevance Assessment, Deep Learning

1. Introduction

Automatic assessment systems are becoming attractive with a growing demand for assessment of English as an additional language [1]. They allow language assessment programmes to economically scale their operations whilst decreasing throughput time and provide testing on demand. Spoken language proficiency is assessed based on a candidate's responses to a series of question prompts, such as 'describe a difficult situation at work, why was it difficult?". These assessment systems operate on features derived from recordings of the candidate's responses. Automatic speech recognition (ASR) is used to transcribe the responses to provide structured features, in addition to features derived directly from the audio. Modern systems, such as ETS' *SpeechRater* [2] and Pearson's *AZELLA* [3], typically only assess pronunciation and fluency. Although these are highly correlated with spoken language proficiency, reliable and robust high-stakes assessment requires the assessment of the semantic content, construction and relevance of the response to the question prompt. Such a system should assess whether the candidate has given an off-topic response, either due to misunderstanding the question and/or memorizing a response. This is the problem addressed in this paper.

A standard approach to assessing topic relevance and off-topic response detection, both for essays and speech, is based on measuring the similarity between a response and the test question or prompt. Commonly, this is done by measuring the similarity between vector representations of responses and prompts, such as TF-IDF, Latent Semantic Analysis (LSA) [4, 5] or Latent Dirichlet Allocation (LDA) [6, 7]. There are two major deficiencies with this approach. Firstly, it is based on bag-of-words vector representations which lose sequential information important to evaluating the semantic content of responses. Secondly, such systems require having prompt-response pairs for all prompts in the test and can only assess relevance to prompts which they have seen in the training data. The approach proposed in [8] overcomes the first limitation. It uses a topic adapted Recurrent Neural Network Language Model (RNNLM) to rank the topic-conditional probabilities of a response sentence. However, this approach still requires having prompt-response pairs for all prompts and cannot assess relevance to new and previously unseen test prompts. Furthermore, re-training the system may be computationally costly. This limits the flexibility and increases the cost of deployment of such systems, as it is necessary to collect example responses to newly introduced prompts in order to have a system which is able to detect off-topic responses to these prompts. This work aims to overcome this limitation.

Recent work in the fields of Neural Machine Translation and Question Answering [9, 10] has come up with a number of attention-based deep learning architectures. Their key advantage is their ability to use an attention mechanism to extract relevant information from a variable-length sequence model in the form of a fixed-length embedding, conditioned on another embedding. This approach was used by [9] to achieve breakthrough results in English-to-French machine translation. Such approaches have also been successfully applied to assessment of multiple-choice questions [11]. In a related piece of work, a Recurrent Neural Network was used to extract optimal sequence features from spoken assessment in [12].

This paper presents an initial investigation of a novel neural attention-based model for assessing the relevance of spontaneous spoken responses to open ended prompts without the need to see them in training. This model uses a Bidirectional Recurrent Neural Network (BiRNN) embedding of a prompt to attend over a BiRNN embedding of a response. The resulting fixed-length prompt-conditional response embedding is fed into a binary classifier to predict the relevance of the response to the prompt. The model is trained on ASR transcriptions of spoken responses. Due to a lack of off-topic responses, negative examples for both training and evaluation are created by randomly shuffling prompts and responses. The model is evaluated using the area under a binary classification receiver-operator curve

(ROC AUC) metric. The ability of this model to assess the relevance and detect off-topic responses to prompts which were both seen, and crucially, not seen in the training data is demonstrated on spoken data from the Cambridge Business Language (BULATS) exam.

The rest of this paper is structured as follows: section 2 introduces and describes the proposed model, section 3 describes the data and experimental setup, section 4 contains the results and analysis, and section 5 is the conclusion.

2. Model

This section describes the proposed neural attention-based model for assessing the relevance of responses to prompts. The model is illustrated in Figure 1. It consists of four components; a prompt encoder, a response encoder, an attention mechanism and a binary classifier.

The proposed model assesses the relevance of responses to prompts by using the prompt to extract information from the response which is used to assign a relevance score. This is accomplished by learning to dynamically compute a representation (embeddings) of the prompt using the prompt encoder. This prompt embedding is used to attend over a representation (embedding) of the response via an attention mechanism, which should highlight the parts of the response most relevant to the prompt. Based on this information, a binary classifier assigns the probability of the response being relevant to the prompt.

The prompt (eq. 1) and response (eq. 2) encoders are Bidirectional Recurrent Neural Networks (BiRNN) [13] with LSTM recurrent units [14, 15] which process the words of the prompt and response, respectively. The prompt and response are represented by the word sequences $\boldsymbol{w}^p = \{w_1^p, \cdots, w_L^p\}$ and $\boldsymbol{w}^r = \{w_1^r, \cdots, w_T^r\}$. The prompt embedding $\tilde{\boldsymbol{h}}^p$ is computed by concatenating the final forward in time $\overrightarrow{\boldsymbol{h}}_L^p$ and backward in time $\overleftarrow{\boldsymbol{h}}_1^p$ hidden states of the prompt encoder (eq. 3). The forward in time $\overrightarrow{\boldsymbol{h}}_t^r$ and backward in time $\overleftarrow{\boldsymbol{h}}_t^r$ hidden states of the response encoder are concatenated at every time step to produce a hidden state $\tilde{\boldsymbol{h}}_t^r$ (eq. 3), which contains information about how the complete surrounding context relates to the current word.

$$\boldsymbol{h}_{1:L}^p = \texttt{LSTM}^p(\boldsymbol{w}^p;\ \boldsymbol{\theta}^p) \tag{1}$$

$$\boldsymbol{h}_{1:T}^r = \texttt{LSTM}^r(\boldsymbol{w}^p;\ \boldsymbol{\theta}^r) \tag{2}$$

$$\tilde{\boldsymbol{h}}^p = \begin{bmatrix} \overrightarrow{\boldsymbol{h}}_L^p \\ \overleftarrow{\boldsymbol{h}}_1^p \end{bmatrix} \quad \tilde{\boldsymbol{h}}_t^r = \begin{bmatrix} \overrightarrow{\boldsymbol{h}}_t^r \\ \overleftarrow{\boldsymbol{h}}_t^r \end{bmatrix} \tag{3}$$

A fixed-length prompt-conditional embedding $\boldsymbol{c}$ of the response is computed as a weighted sum of the hidden states $\tilde{\boldsymbol{h}}_t^r$ of the response encoder given a set of attention weights α_t via an attention mechanism (eq. 5). The attention weights for each hidden state are computed as a softmax (eq. 6), where the logits are given by a similarity function between the prompt embedding and the response hidden state. The similarity function (eq. 7) computes how strongly a hidden state of the response encoder relates to the embedding of the prompt. The parameters of the attention mechanism are $\boldsymbol{\theta}^a = \{\boldsymbol{v}_e, \boldsymbol{\Lambda}_1, \boldsymbol{\Lambda}_2, \boldsymbol{b}\}$. This similarity function was used in [9] for neural machine translation. Alternative attention mechanisms, with different similarity functions

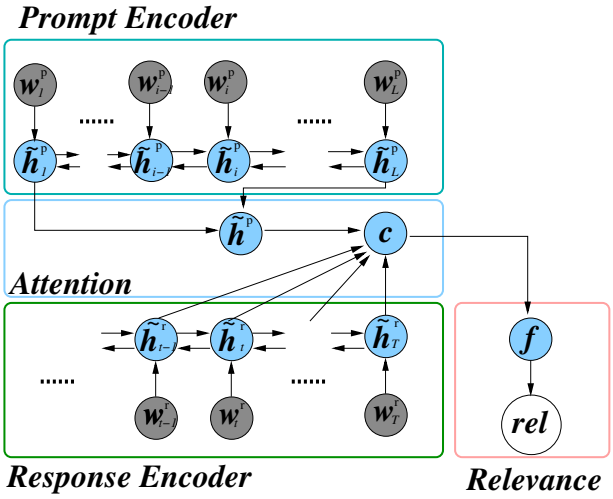

Figure 1: *Neural attention-based response-prompt relevance model.*

[16] and attention sharpening [10] could potentially be used.

$$\boldsymbol{c} = \sum_{t=1}^{T} \alpha_t \tilde{\boldsymbol{h}}_t^r \tag{4}$$

$$\alpha_t = \frac{\exp(s_t(\tilde{\boldsymbol{h}}^p, \tilde{\boldsymbol{h}}_t^r))}{\sum_{\tau=1}^{T} \exp(s_\tau t(\tilde{\boldsymbol{h}}^p, \tilde{\boldsymbol{h}}_\tau^r))} \tag{5}$$

$$s_t(\tilde{\boldsymbol{h}}^p, \tilde{\boldsymbol{h}}_t^r) = \boldsymbol{v}_{\boldsymbol{e}}^{\mathrm{T}} \tanh(\boldsymbol{\Lambda}_1 \tilde{\boldsymbol{h}}^p + \boldsymbol{\Lambda}_2 \tilde{\boldsymbol{h}}_t^r + \boldsymbol{b}) \tag{6}$$

The fixed-length response embedding $\boldsymbol{c}$ is the fed into a binary classifier f (eq. 7) which outputs the probability $\texttt{P}(\texttt{rel}|\boldsymbol{w}^r, \boldsymbol{w}^p)$ of the response relating to the question. In this work f is a deep neural network (DNN) with parameters $\boldsymbol{\theta}^f$.

$$\texttt{P}(\texttt{rel}|\boldsymbol{w}^r, \boldsymbol{w}^p) = f(\boldsymbol{c};\ \boldsymbol{\theta}^f) \tag{7}$$

This model is trained using minibatch stochastic gradient descent with a logistic loss error function (eq. 8) over all parameters $\boldsymbol{\theta} = \{\boldsymbol{\theta}^p, \boldsymbol{\theta}^r, \boldsymbol{\theta}^a, \boldsymbol{\theta}^f\}$. The model is trained on a balanced data set of prompt-response pairs containing both positive and negative examples of relevance.

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{N} \sum_{i=1}^{N} & t_i \log(\texttt{P}(\texttt{rel}|\boldsymbol{w}_i^r, \boldsymbol{w}_i^p)) \\ & + (1 - t_i) \log(1 - \texttt{P}(\texttt{rel}|\boldsymbol{w}_i^r, \boldsymbol{w}_i^p))\end{aligned} \tag{8}$$

2.1. Relation to Previous Work

Previously proposed methods, such as [4, 17, 18, 19, 20, 8] require an active set of question or prompt representations to be maintained. Typically, these are vector representations of topic based on TF-IDF, LSA or LDA [6, 7]. These are commonly constructed from example responses to the questions or prompts. Thus if a new prompt is introduced, there is a need to collect example responses and to re-train the model, both of which could be expensive and time-consuming, limiting the flexibility of deployment of such models.

The primary advantage of the proposed attention-based relevance model is that, unlike previous methods, it does not need to maintain an active set of topic (prompt) embeddings, but can automatically embed any prompt into the appropriate space

via the prompt encoder. This also eliminates the need to precompute a set of topic representations from examples responses. All components of the model are trained jointly, which allows them to learn the necessary representations and transformations which make this possible. This allows the model to assess the relevance of responses to newly introduced prompts without the need to collect example responses to the new prompt or for the model to be re-trained. However, the model needs to be trained to generalize well in order to effectively handle unseen prompts, especially if they are quite different to the prompts seen in the training data.

Previous Deep Learning based approaches to off-topic response detection [8], which were also evaluated on the BULATS data used in this work, constructed a topic-adapted RNNLM which is conditioned on an active, fixed set of LSA topic embeddings trained separately on example responses. This is a discriminative sentence model conditioned on the topic $\mathtt{P}(\boldsymbol{w}^r|\boldsymbol{w}^p)$. Topic relevance is assessed via the approximation in eq. 9. By using a uniform prior $\mathtt{P}(\boldsymbol{w}^p)$ over topics, it is possible to induce an implicit generative model over topics via Bayes rule. Since a single response may be related to multiple topics to different degrees, relevance is assessed by taking the top-N highest ranking probabilities.

$$\mathtt{P}(\mathtt{rel}|\boldsymbol{w}^r, \boldsymbol{w}^p) \approx \mathtt{P}(\boldsymbol{w}^p|\boldsymbol{w}^r) \approx \frac{\mathtt{P}(\boldsymbol{w}^r|\boldsymbol{w}^p)}{\sum_{\forall p} \mathtt{P}(\boldsymbol{w}^r|\boldsymbol{w}^p)} \quad (9)$$

In contrast, the proposed model calculates $\mathtt{P}(\mathtt{rel}|\boldsymbol{w}^r, \boldsymbol{w}^p)$ directly, and there is no need to use ranking to assess topic relevance. Furthermore, the proposed model does not have an explicit model of topic $\mathtt{P}(\boldsymbol{w}^p|\boldsymbol{w}^r)$. Since a response can be relevant to varying degrees to several different topics, a potential disadvantage of the proposed model is that a certain amount of confusion can be introduced by having negative examples which are very similar to the positive example.

3. Data and Experimental Setup

A series of experiments were run to assess the ability of the proposed automatic systems to rate the relevance of responses to prompts. Data from the Business Language Testing Service (BULATS) English tests was used for training and test. The text for each response was generated using an ASR system. The 1-best recognition hypothesis was then passed to a relevance assessment system, which decided whether the candidate had spoken off topic by assigning a probability of whether the response was relevant to the prompt. To avoid a data mismatch, the recognition hypothesis was used both in training and test.

3.1. BULATS Test Format and Data

The BULATS Online Speaking Test has five sections [21]. This work focuses on the 3 sections where open ended prompts (which appear on screen) elicit spontaneously constructed responses:

- C Candidates talk about a work-related topic (e.g. the perfect office).
- D Candidates must describe a graph such as a pie or a bar chart related to a business situation (e.g. company sales).
- E Candidates are asked to respond to 5 open-ended prompts related to a single context prompt (e.g. a set of 5 questions about organizing a stall at a trade fair).

The 3 sections consist of 7 prompts in total.

Data	#Topics	#Resp.	#Words	#Resp./ Topic	Avg.Resp. Length
TRN	379	292.9K	13.4M	772.8	45.7
EVAL1	92	1319	64.7K	14.3	49.1
EVAL2	179	1445	59.6K	8.1	41.3
EVAL3	180	1496	63.8K	8.3	42.6
ALL	222	4260	188.1K	19.2	44.2

Table 1: *Topic, response and word statistics of the prompt-response data sets based on 1-best recognition hypotheses.*

The training (*TRN*) data set is used as the source of prompt-response pairs for training the model. It contains 13.4M words in 292.9K responses covering 42K candidates. There are a total of 379 unique prompts in *TRN*. Each prompt relates to one topic, making the terms interchangeable. For multi-part prompts, each part is considered a distinct topic. For each of the topics (prompts) there are an average of 772.8 example responses, with an average response length of 45.7 words. *TRN* consists of candidates from a wide range of L1 (native) languages, with the largest proportion being Gujarati L1 candidates.

The evaluation data sets, described in table 1, are designed to have an (approximately) even distribution over CEFR grades levels [22] as well as over the different topics (prompts). *EVAL1* is composed of only Gujarati L1 speakers, *EVAL2* of only Spanish L1 candidates and *EVAL3* is composed of Arabic, Dutch, French, Polish, Thai and Vietnamese L1 candidates. The evaluation data set *ALL* is the combination of *EVAL1-3*.

3.2. Training Data Construction

As the data is taken from tests run with human examiners the responses are virtually all on topic. To produce negative, off topic training examples, the responses and prompts for both training and evaluation were shuffled. As was shown in [8], responses to prompts from the same section tend to be more similar so are more confusable. Thus, two topic shuffling strategies are considered: *Naive*, where prompts are shuffled across all sections; and *Directed*, where prompts are shuffled only within the same section [8]. *Naive* topic shuffling represents a more likely scenario, as real off-topic responses are unlikely to come from the same section. The data sets were balanced, so that for every response there are as many matched positive examples as there are mismatched negative examples in the training data. Thus, if more than one negative example is shown for a particular response, the positive example would be over-sampled the corresponding number of times. For multi-part prompts, which contain a main prompt that describes the overall question, and several (5 in this case) sub-prompts, all sub-prompts were pre-appended with the main prompt. These sub-prompts are considered distinct topics. During training, sub-prompts to the same overall prompt are considered competing negative examples to each other during shuffling.

3.3. ASR System

A speaker independent hybrid deep neural network - hidden Markov model (DNN-HMM) system is used for ASR [23]. The acoustic models are trained on 108.6 hours of BULATS test data (Gujarati L1 speakers) using the HTK v3.5 toolkit [24, 25]. A Kneser-Ney trigram language model is trained on this data and is then interpolated with a general English language model

trained on a large broadcast news corpus, using the SRILM toolkit [26]. This ASR system has a word error rate of 32% on a Gujarati L1 ASR development set taken from the BULATs data. Performance on other L1s varies from 42-53%.

3.4. Model and Training Hyper-parameters

The proposed relevance assessment model was implemented in Tensorflow [27]. It consists of 2 BiRNN encoders with 400 LSTM recurrent units with hyperbolic tangent (TanH) nonlinearities, 200 for the forward states and 200 for the backward states. The model was trained for 5 epochs with the Adam optimizer [28], with an initial learning rate of 1e-3, and an exponentially decaying learning rate with decay factor 0.96 per epoch. Dropout regularization [29] was used with a keep probability of 0.8, dropout was applied to all layers except for the LSTM hidden-to-hidden connections and the word embeddings. The binary classifier was a DNN with 2 hidden layers of 200 rectified linear (ReLU) units and with a 1-dimensional logistic output. The word embeddings, shared by both the response and prompt BiRNNs were initialized from an RNNLM language model trained on the TRN responses and were kept fixed during training. Four main models are examined in this work: models N1 and D1, with *Naive* and *Directed* topic shuffling of training data, respectively, and 1 negative example per response, and models N5 and D5, with *Naive* and *Directed* topic shuffling of training data, respectively, and 5 negative examples per response. N1 and D1 take roughly 2.5 hours to train while N5 and D5 take 12 hours to train in an nVidia GTX 980M graphics card.

3.5. Assessment Criteria

The models are evaluated using the area under a Receiver-Operator Characteristic (AUC), which plots the True Positive vs. the False positive rate at different decision thresholds. In order to be able to do this, negative examples (true negatives) need to be introduced into the evaluation data sets. This is done using the same method as for training, with one positive and one negative example for every response, both with *Naive* and *Directed* shuffling. It must be noted, that results are based on a particular instance of shuffling the prompts for evaluation.

4. Experiments

This section presents the results of investigations into the properties of the proposed model. Subsection 4.1 investigates several key properties of the model when all the prompts are seen. First, the baseline performance of a model trained on data with *Naive* topic shuffling and 1 negative example per response. Secondly, the effect of CEFR grade level [22] on relevance assessment performance is investigated. Thirdly, the effect of using training data with 5 negative examples per response is assessed. Finally, the effect of using training data with *Directed* topic shuffling is investigated. Subsection 4.2 investigates the performance of the model on unseen topics (prompts).

4.1. Baseline Performance

Table 2 and Figure 2 show the AUC scores for the baseline N1 model for all evaluation data sets. There are several notable trends in the data. Firstly, overall, the model achieves a high AUC of 0.95 on *ALL* evaluation data with *Naive* topic shuffling, and a lower AUC of 0.90 with *Directed* topic shuffling. This supports the findings in [8] which state that it is more difficult to distinguish prompts from the same section than from across sections. However, the AUC score of 0.95 reflects the more likely operating scenario, as *Naive* topic shuffling is more representative of real off-topic responses. This trend holds for all evaluation subsets. The performance on subset *EVAL1* was highest, which reflects both the dominance of Gujarati L1 candidates in the training data as well as the better quality of the ASR transcriptions of responses of Gujarati candidates.

Topic shuffling	EVAL1	EVAL2	EVAL3	ALL
Naive	0.97	0.95	0.94	0.95
Directed	0.94	0.89	0.88	0.90

Table 2: *Baseline AUC scores for model trained on data with* Naive *topic shuffling and 1 negative example per response (N1).*

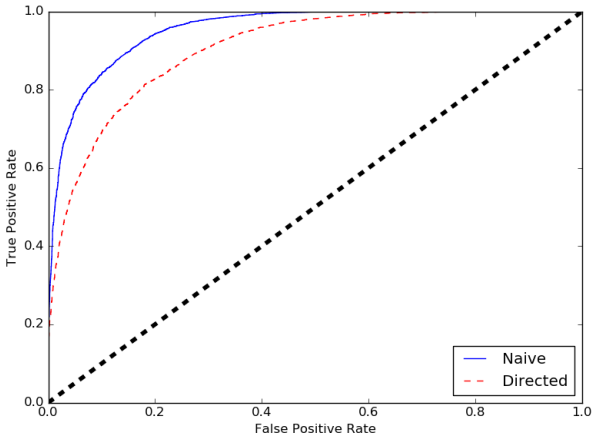

Figure 2: *ROC curve for model trained on data with* Naive *topic shuffling and 1 negative example per response (N1) evaluated on* ALL *eval data.*

Table 3 shows how the AUC performance of the baseline N1 model varies with the CEFR level of the candidates. Clearly, AUC increases with increasing proficiency level from the lowest, A1, to the highest, C. This reflects both the increasing complexity of the response, allowing it to be more easily distinguished from a response to a different prompt, and the rising quality of the transcription - it is easier to correctly transcribe the response of a good candidate using ASR. This trend holds for all subsets *EVAL1-3*.

Topic Shuffling	A1	A2	B1	B2	C
Naive	0.88	0.94	0.94	0.97	0.97
Directed	0.82	0.88	0.91	0.93	0.94

Table 3: *Per grade level breakdown of performance on* ALL *for model trained on data with* Naive *topic shuffling and 1 negative example per response (N1).*

The results of the investigation of the effect of using *Naive* vs *Directed* shuffling of training data, as well as the effect of using more negative-examples per response are presented in Table 4. Using 5 negative examples with *Naive* shuffling (N5) gives very high performance on both the *Naive*, and especially, *Directed* evaluation data. Clearly, as the model is exposed to a greater variety of negative examples it learns to generalize better. This performance boost relative to the model trained with 1

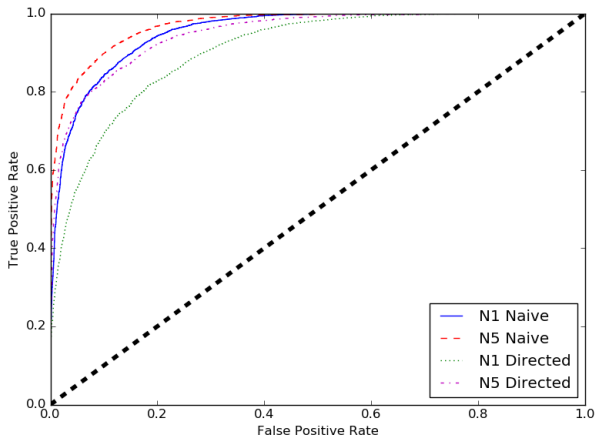

Figure 3: *Comparison of model trained on data with* Naive *topic shuffling with 1 (N1) and 5 (N5) negative examples per response evaluated on* ALL *eval data.*

negative sample is illustrated in Figure 3. The same trend can be seen for models trained with *Directed* shuffling of the training data (D1 and D5). Interestingly, model D1 has similar performance on both evaluation sets, while model N1 has clearly better performance on the *Naive* evaluation set. This distinction is blurred for the N5 and D5 models, both of which have comparable performance on all evaluation datasets.

Topic shuffling	Train topic shuffling			
	Naive		Directed	
	N1	N5	D1	D5
Naive	0.95	0.97	0.90	0.95
Directed	0.90	0.95	0.91	0.96

Table 4: *Comparison of using* Naive *and* Directed *training data and using more negative-examples on* ALL *evaluation data.*

4.2. Performance on Unseen Prompts

In the above experiments all the prompts have been seen in the training data. This section considers the scenario where some prompts are not seen in training, investigating the proposed model's ability to generalize to new prompts. Since real unseen prompt-response pairs are unavailable, 10-fold cross validation over prompts (topics) was used on the training and evaluation data. A fixed block of data, *TRN-fixed* (Table 5), is never removed from the training data, as it contains topics which dominate the training data and topics which do not appear in the evaluation set *ALL*. The *TRN-xVal* data was used in cross validation. A subset of *ALL*, called *ALL-sub*, without the dominant topics of *TRN*, was used for cross validation evaluation. All parts of related multi-part prompts are held out together.

The training data uses *Naive* response shuffling with 1 negative example per prompt, described in section 3.3. However, evaluation data responses are shuffled differently to the previous section, for these experiments. The prompts presented to the model are always either from the subsets which are seen or unseen in the training data. Evaluation responses are always new (not reused from the training data, same as in section 4.1), but can be related to prompts either seen or unseen in training. Three strategies for shuffling evaluation responses for negative examples are considered: *seen*, *unseen* and *balanced*. The first uses responses to seen prompts as negative examples, the second uses responses to unseen prompts as negative examples, and the last is an equal mix of the two. This produces six experiments: seen prompts with *seen*, *unseen* and *balanced* response shuffling; unseen prompts with *seen*, *unseen* and *balanced* response shuffling. The first three illustrate how well the model understands what relates to seen prompts and how well it generalizes to increasingly differing responses. The latter three experiments illustrate how well the model generalizes to new, unseen prompts. Generalization performance is increasingly stressed with *seen*, *balanced* and *unseen* evaluation response topic shuffling, since the responses become increasingly unfamiliar. Relevance probabilities are combined across all 10 folds to produce one ROC curve and AUC score for each experiment. These curves, and the associated AUC scores, represent the 'average' AUC on the data.

Data	#Topics	#Resp.	#Words
TRN-fixed	178	142.8K	6.8M
TRN-xVal	201	150.1K	6.6M
ALL-sub	201	2955	127.7K

Table 5: *Topic, response and word statistics of the prompt-response data sets used for 10-fold cross validation.*

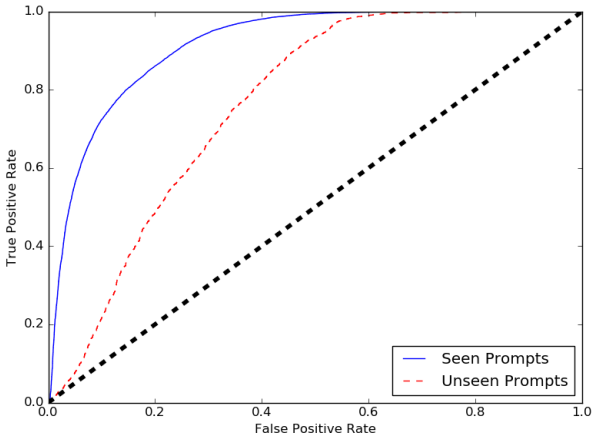

Figure 4: *Average ROC curves for seen and unseen prompts with balanced response topic shuffling across 10 folds.*

Topic shuffling	Prompts	
	Seen	Unseen
Seen	0.92	0.78
Balanced	0.92	0.76
Unseen	0.92	0.72

Table 6: *Experiments on ALL-sub.*

The results presented in Table 6 show that once prompts have been seen in training data, the model has a clear understanding of what is relevant to them and is not sensitive to the nature of the negative-example responses. However, on unseen prompts there is a degradation of performance, which ranges from 0.78 to 0.72 as evaluation response topics shuffling changes from *seen* to *unseen*. Clearly, the model is able to generalize well to unfamiliar responses, and to a lesser degree, to new prompts, even in the extreme scenario (0.72 AUC). This is expected, as the model is exposed to a greater variety of responses than prompts. ROC curves for performance on seen

and unseen prompts with balanced response topic shuffling are shown in Figure 4.

5. Conclusions and Future Work

This paper presented an initial study of a novel neural attention-based model for assessing the relevance of spontaneous spoken responses to open ended prompts. This model uses a bidirectional recurrent neural network (BiRNN) embedding of a prompt to attend over a BiRNN embedding of a response. The resulting fixed-length prompt-conditional response embedding is fed into a binary classifier to predict the relevance of the response to the prompt. Due to a lack of off-topic responses, negative examples for both training and evaluation are created by randomly shuffling prompts and responses. The primary advantage of this model is that it is able to assess the relevance and detect off-topic responses to prompts which were both seen, and crucially, not seen in the training data.

Improvements could be added to the model in future work. For example, the model could be trained with dynamic sampling of negative examples during training in order to expose the model to a greater number of competing examples at lower computation cost. Furthermore, it is interesting to investigate what the attention mechanism learns, and how its focus over particular words in a response varies across prompts. Correlation of response relevance with grade level should be investigated. Due to time constraints, it was not possible to run 10-fold cross-validation on unseen topics using a model trained on more than 1 negative example per response. A more robust method for evaluation, such as using a greater number of samples, should be considered. The proposed method should be compared to previously proposed approaches, such as [8].

6. Acknowledgements

This research was funded under the ALTA Institute, University of Cambridge as well as the Engineering and Physical Sciences Research Council. Thanks to Cambridge English, University of Cambridge, for support and access to the BULATS data.

7. References

[1] B. Seidlhofer, "English as a lingua franca," *ELT journal*, vol. 59, no. 4, p. 339, 2005.

[2] K. Zechner *et al.*, "Automatic scoring of non-native spontaneous speech in tests of spoken English," *Speech Communication*, vol. 51, no. 10, pp. 883–895, 2009.

[3] A. Metallinou and J. Cheng, "Using Deep Neural Networks to Improve Proficiency Assessment for Children English Language Learners," in *Proc. INTERSPEECH*, 2014.

[4] H. Yannakoudakis, "Automated assessment of English-learner writing," https://www.cl.cam.ac.uk/techreports/UCAM-CL-TR-842.pdf, University of Cambridge Computer Laboratory, Tech. Rep. UCAM-CL-TR-842, 2013.

[5] T. K. Landauer, P. W. Foltz, and D. Laham, "Introduction to Latent Semantic Analysis," *Discourse Processes*, vol. 25, pp. 259–284, 1998.

[6] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, Mar. 2003.

[7] T. L. Griffiths and M. Steyvers, "Finding Scientific Topics," *Proceedings of the National Academy of Sciences*, vol. 101, pp. 5228–5235, 2004.

[8] A. Malinin, R. van Dalen, K. Knill, Y. Wang, and M. Gales, "Off-topic Response Detection for Spontaneous Spoken English Assessment," in *Proc. 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Berlin, Germany, 2016, pp. 1075–1084.

[9] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. ICLR*, 2015.

[10] A. Graves, G. Wayne, and I. Danihelka, "Neural Turing Machines," *CoRR*, vol. abs/1410.5401, 2014. [Online]. Available: http://arxiv.org/abs/1410.5401

[11] B.-H. Tseng, S.-S. Shen, H.-Y. Lee, and L.-S. Lee, "Towards Machine Comprehension of Spoken Content: Initial TOEFL Listening Comprehension Test by Machine," in *Proc. INTERSPEECH*, 2016.

[12] Z. Yu *et al.*, "Using bidirectional LSTM recurrent neural networks to learn high-level abstractions of sequential features for automated scoring of non-native spontaneous speech," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2015, pp. 338–345.

[13] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[14] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[15] A. Graves, *Supervised Sequence Labelling with Recurrent Neural Networks*. Studies in Computational Intelligence, Springer, 2012.

[16] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in *Proc. EMNLP*, 2015.

[17] S. Xie, K. Evanini, and K. Zechner, "Exploring Content Features for Automated Speech Scoring," in *Proc. Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2012.

[18] K. Evanini, S. Xie, and K. Zechner, "Prompt-based Content Scoring for Automated Spoken Language Assessment," in *Proc. Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2013.

[19] S.-Y. Yoon and S. Xie, "Similarity-Based Non-Scorable Response Detection for Automated Speech Scoring," in *Proc. Ninth Workshop on Innovative Use of NLP for Building Educational Applications*, 2014.

[20] K. Evanini and X. Wang, "Automatic detection of plagiarized spoken responses," in *Proc. Ninth Workshop on Innovative Use of NLP for Building Educational Applications*, 2014.

[21] L. Chambers and K. Ingham, "The BULATS Online Speaking Test," *Research Notes*, vol. 43, pp. 21–25, 2011.

[22] C. of Europe, *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge, U.K: Press Syndicate of the University of Cambridge, 2001.

[23] H. Wang *et al.*, "Joint Decoding of Tandem and Hybrid Systems for Improved Keyword Spotting on Low Resource Languages," in *Proc. INTERSPEECH*, 2015.

[24] S. Young *et al.*, *The* HTK *book (for* HTK *Version 3.4.1)*. University of Cambridge, 2009.

[25] ——, *The* HTK *book (for* HTK *version 3.5)*. University of Cambridge, 2015, http://htk.eng.cam.ac.uk.

[26] A. Stolcke, "SRILM an extensible language modelling toolkit," in *Proc. International Conference on Spoken Language Processing (ICSLP)*, 2002.

[27] M. Abadi *et al.*, "TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems," 2015, software available from tensorflow.org. [Online]. Available: http://tensorflow.org/

[28] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *Proc. 3rd International Conference on Learning Representations (ICLR)*, 2015.

[29] N. Srivastava *et al.*, "Dropout: a simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.