A HIERARCHICAL ATTENTION BASED MODEL FOR OFF-TOPIC SPONTANEOUS SPOKEN RESPONSE DETECTION

Andrey Malinin, Kate Knill and Mark J. F. Gales

University of Cambridge, Department of Engineering,
Trumpington St, Cambridge CB2 1PZ, UK

ABSTRACT

Automatic spoken language assessment and training systems are becoming increasingly popular to handle the growing demand to learn languages. However, current systems often assess only fluency and pronunciation, with limited content-based features being used. This paper examines one particular aspect of content-assessment, off-topic response detection. This is important for deployed systems as it ensures that candidates understood the prompt, and are able to generate an appropriate answer. Previously proposed approaches typically require a set of prompt-response training pairs, which limits flexibility as example responses are required whenever a new test prompt is introduced. This paper extends the attention based neural topic model (ATM) which can assess the relevance of prompt-response pairs regardless of whether the prompt was seen in training. This model uses a bidirectional Recurrent Neural Network (BiRNN) embedding of the prompt to attend over the hidden states of a BiRNN embedding of the response to compute a fixed-length embedding used to predict relevance. A hierarchical variant of the ATM (HATM) is also described, which computes an interpretable prompt embedding by interpolating all prompts seen in training data given a prompt of interest via a second attention mechanism. On spontaneous spoken data, taken from BULATS tests, these systems are able to assess relevance to both seen and unseen prompts.

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

1. INTRODUCTION

A key part of learning a language is learning how to speak fluently and with confidence. This is assessed through spoken language proficiency tests where candidates are prompted to respond to a series of open-ended questions, such as ”describe a difficult situation at work, why was it difficult?”. Human examiners assess the candidate’s spontaneous speech replies in terms of pronunciation, hesitations/extent, use of grammar and vocabulary, and how coherent their discourse is. The increasing demand for language learning and for practice tests available at any time make the development of automatic systems to undertake this assessment and provide feedback an attractive proposition [1]. Structured features derived from automatic speech recognition (ASR) generated transcriptions of the candidate’s responses are combined with features derived directly from the audio as input to automatic spoken language assessment systems. Current automatic assessment is primarily focused on pronunciation and fluency (both of which are highly correlated with proficiency), such as ETS’ *SpeechRater* [2] and Pearson’s *AZELLA* [3]. It is not clear to what extent content is currently assessed. Reliable, robust assessment requires the evaluation of the semantic content, construction and relevance of a response to the question prompt. Such a system should assess if a 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.

Standard approaches [4, 5] to assessing semantic content topic relevance, both for essays and speech, are based on measuring the similarity between vector representations of responses and prompts. Such systems need to have seen in training prompt-response pairs for all prompts in a test to assess the relevance of a test response. This limits the flexibility and increases the cost of deployment of such systems, as example responses have to be collected for newly introduced prompts. Re-training the system may be computationally costly. This limitation is overcome in the approach proposed in [6], called the Attention-based Topic Model (ATM),to assess the relevance of spontaneous spoken responses to open-ended prompts. The ATM allows the assessment of relevance to prompts not seen in the training data. Unfortunately, while the system achieves excellent performance on prompts with responses seen in training, performance on unseen prompts is not as good. Furthermore, the ATM is not particularly interpretable, does not explicitly exploit the similarity between different prompts and [6] used a fixed set of prompt-response matchings as negative examples during training.

This paper presents extensions to the ATM. A hierarchical variant of the ATM (HATM) is proposed in an attempt to improve performance on unseen prompts and increase interpretability. The HATM explicitly leverages similarity between prompts via a second attention mechanism which interpolates all prompts seen in the training data given a prompt of interest. This allows the construction of an prompt ontology. Furthermore, a dynamic sampling mechanism is added to generate negative examples and the use of ASR confidence scores as additional features is investigated. The ability of these models to assess the relevance and detect off-topic responses to prompts which are 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 models, section 3 describes the data and experimental setup, section 4 contains the experimental results and analysis, and section 5 is the conclusion.

2. MODEL

This section describes the ATM and HATM models for assessing the relevance of responses to prompts. The ATM (Fig. 1) consists of a prompt encoder (red), a response encoder and an attention mechanism over responses (blue) and a binary classifier (green). The

This research was funded under the ALTA Institute, Cambridge University. Thanks to Cambridge English Language Assessment for supporting this research and providing access to the BULATS data.

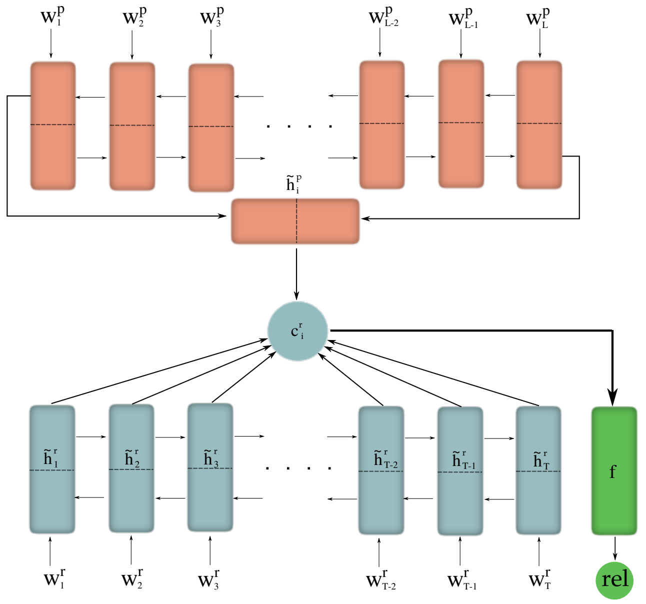


Fig. 1. Attention-based Topic Model

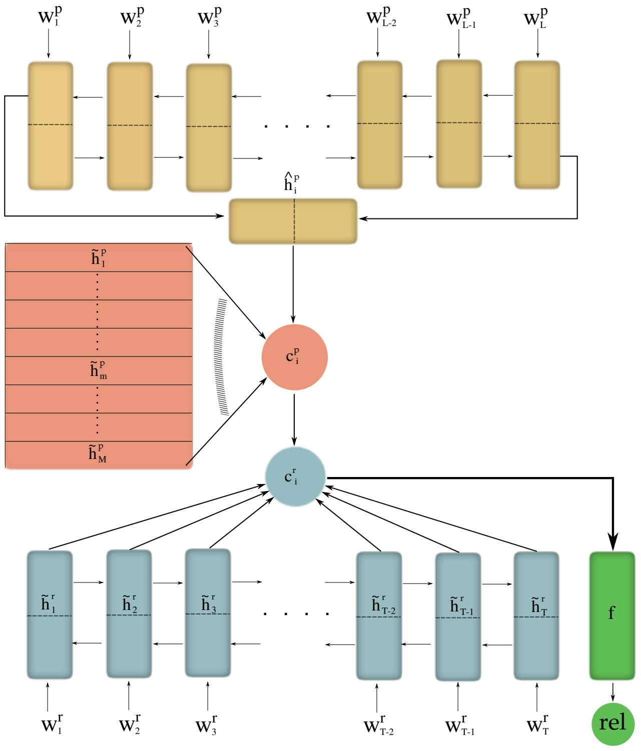


Fig. 2. Hierarchical Attention-based Topic Model

HATM (Fig. 2) additionally has a prompt attention mechanism and prompt-search embedding (yellow).

The ATM assesses the relevance of responses to prompts by using the prompt to extract information from the response which is then used to assign a relevance score. This is accomplished by learning to dynamically compute a representation (embedding) 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 a response is relevant to the prompt.

The prompt (eq. 1) and response (eq. 2) encoders are Bidirectional Recurrent Neural Networks (BiRNN) [7] with LSTM recurrent units [8, 9] which process the word sequences $\boldsymbol{w}^p = \{w^p_1, \cdots, w^p_L\}$ and $\boldsymbol{w}^r = \{w^r_1, \cdots, w^r_T\}$ of the prompt and response, respectively. The prompt embedding $\tilde{\boldsymbol{h}}^p$ is computed by concatenating the final forward in time $\overrightarrow{\boldsymbol{h}}^p_L$ and backward in time $\overleftarrow{\boldsymbol{h}}^p_1$ hidden states of the prompt encoder (eq. 3). The forward in time $\overrightarrow{\boldsymbol{h}}^r_t$ and backward in time $\overleftarrow{\boldsymbol{h}}^r_t$ hidden states of the response encoder are concatenated at every time step to produce a hidden state $\tilde{\boldsymbol{h}}^r_t$ (eq. 3), which contains information about how the complete surrounding context relates to the current word.

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

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

$$\tilde{\boldsymbol{h}}^p = \begin{bmatrix} \overrightarrow{\boldsymbol{h}}^p_L \\ \overleftarrow{\boldsymbol{h}}^p_1 \end{bmatrix} \quad \tilde{\boldsymbol{h}}^r_t = \begin{bmatrix} \overrightarrow{\boldsymbol{h}}^r_t \\ \overleftarrow{\boldsymbol{h}}^r_t \end{bmatrix} \tag{3}$$

A fixed-length prompt-conditional embedding $\boldsymbol{c}^r_i$ of the response is computed as a weighted sum of the hidden states $\tilde{\boldsymbol{h}}^r_t$ of the response encoder given a set of attention weights α_t via an attention mechanism (eq. 4). The attention weights for each hidden state are computed as a softmax (eq. 5), where the logits are given by a similarity function (eq. 6) which computes how strongly a hidden state of the response encoder relates to the embedding of the prompt. The parameters of the attention mechanism are $\boldsymbol{\theta}^a = \{\boldsymbol{v}_r, \boldsymbol{\Lambda}_1, \boldsymbol{\Lambda}_2, \boldsymbol{b}\}$.

$$\boldsymbol{c}^r_i = \sum_{t=1}^{T} \alpha_{i,t} \tilde{\boldsymbol{h}}^r_t \tag{4}$$

$$\alpha_{i,t} = \frac{\exp(s(\tilde{\boldsymbol{h}}^p_i, \tilde{\boldsymbol{h}}^r_t))}{\sum_{\tau=1}^{T} \exp(s(\tilde{\boldsymbol{h}}^p_i, \tilde{\boldsymbol{h}}^r_\tau))} \tag{5}$$

$$s(\tilde{\boldsymbol{h}}^p_i, \tilde{\boldsymbol{h}}^r_t) = \boldsymbol{v}_{\boldsymbol{r}}^{\mathrm{T}} \tanh(\boldsymbol{\Lambda}_1 \tilde{\boldsymbol{h}}^p_i + \boldsymbol{\Lambda}_2 \tilde{\boldsymbol{h}}^r_t + \boldsymbol{b}) \tag{6}$$

The embedding $\boldsymbol{c}^r_i$ is then fed into a binary classifier f (eq. 7) which outputs the relevance 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 $\boldsymbol{\theta}^f$.

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

2.1. Hierarchical Attention-based Topic Model

The Hierarchical Attention-based Topic Model (HATM) (figure) extends the ATM to explicitly make use of the similarity between prompts. This assumes that there is an implicit ontology of prompts in the data, and the HATM learns it in an unsupervised fashion. This is done by expressing prompts seen in the training data $\tilde{\boldsymbol{h}}^p$ as points on a simplex and interpolating over them using a prompt attention mechanism (eq. 9-11). A separate 'search' embedding (eq 8.) of the prompt $\hat{\boldsymbol{h}}^p$ is used compute attention over all prompts $\tilde{\boldsymbol{h}}^p$. This yields a new prompt embedding $\boldsymbol{c}^p$ (eq. 9) which is used to attend over the responses. The prompts seen in the training data never directly attend over themselves - the attention mechanism is trained in a 'leave-one-out' fashion to teach it to reconstruct each prompt in the training data from all other seen prompt embeddings. Theoretically, given a rich, robust and diverse set of prompt embeddings new and unseen prompts may be expressed as an interpolation of seen prompts. This potentially allows the HATM to estimate prompt embeddings for unseen prompts more robustly. Furthermore, the learned ontology may be useful for determining which prompts are more and which are less confusable. The parameters of the prompt attention mechanism are $\boldsymbol{\theta}^{pa} = \{\boldsymbol{v}_p, \boldsymbol{\Lambda}_1^p, \boldsymbol{\Lambda}_2^p, \boldsymbol{b}^p\}$, thus two new sets of parameters are added to the system: $\{\boldsymbol{\theta}^{pa}, \boldsymbol{\theta}^s\}$.

$$\hat{\boldsymbol{h}}_{1:L}^p = \texttt{LSTM}^p(\boldsymbol{w}_i^p;\ \boldsymbol{\theta}^s)\ ;\ \hat{\boldsymbol{h}}_i^p = \begin{bmatrix} \overrightarrow{\boldsymbol{h}}_L^p \\ \overleftarrow{\boldsymbol{h}}_1^p \end{bmatrix} \tag{8}$$

$$\boldsymbol{c}_i^p = \sum_{m=1}^{M} \alpha_{i,m}^p \tilde{\boldsymbol{h}}_m^p \tag{9}$$

$$\alpha_{i,m} = \begin{cases} \dfrac{\exp(s(\hat{\boldsymbol{h}}_i^p, \tilde{\boldsymbol{h}}_m^p))}{\sum_{m=1,\neq i}^{M} \exp(s(\hat{\boldsymbol{h}}_i^p, \tilde{\boldsymbol{h}}_m^p))}, & \text{if } i \neq m \\ 0, & \text{if } i = m \end{cases} \tag{10}$$

$$s(\hat{\boldsymbol{h}}_i^p, \tilde{\boldsymbol{h}}_m^p) = \boldsymbol{v}_{\boldsymbol{p}}^{\mathrm{T}} \tanh(\boldsymbol{\Lambda}_1^p \hat{\boldsymbol{h}}_i^p + \boldsymbol{\Lambda}_2^p \tilde{\boldsymbol{h}}_m^p + \boldsymbol{b}^p) \tag{11}$$

The models are trained using minibatch stochastic gradient descent with a logistic loss error function (eq. 12) over all parameters $\boldsymbol{\theta}^{\boldsymbol{ATM}} = \{\boldsymbol{\theta}^p, \boldsymbol{\theta}^r, \boldsymbol{\theta}^a, \boldsymbol{\theta}^f\}$ or $\boldsymbol{\theta}^{\boldsymbol{HATM}} = \{\boldsymbol{\theta}^p, \boldsymbol{\theta}^r, \boldsymbol{\theta}^a, \boldsymbol{\theta}^f\ \boldsymbol{\theta}^{pa}, \boldsymbol{\theta}^s\}$

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta}) = \frac{1}{N} \sum_{i=1}^{N} & t_i \log(\texttt{P}(\texttt{rel}|\boldsymbol{w}_i^r, \boldsymbol{w}_i^p)) \\ & + (1 - t_i) \log(1 - \texttt{P}(\texttt{rel}|\boldsymbol{w}_i^r, \boldsymbol{w}_i^p)) \end{aligned} \tag{12}$$

3. DATA AND EXPERIMENTAL SETUP

A series of experiments were run to evaluate the ability of the ATM and HATM to assess 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 (ATM/HATM), 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 hypotheses were used both in training and test.

3.1. BULATS Test Format and Data

The BULATS Online Speaking Test has five sections [10]. This work focuses on the 3 sections where open ended prompts (which appear on screen) elicit spontaneously constructed responses. In Section C, candidates talk about a work related topic (e.g. the perfect office). Candidates must describe a graph such as a pie or bar chart related to a business situation (e.g. company sales) in Section D. In Section E candidates are asked to respond to 5 prompts related to a single context prompt (e.g. a set of 5 questions about organizing a stall at a trade fair). There are 7 prompts in total.

Table 1 gives the statistics of the prompt-response data sets used in this paper. Each prompt corresponds to one topic, making the terms interchangeable. The training data set *TRN* contains 13.4M words in 293.0K responses from 42K candidates. 379 unique prompts are seen in *TRN*, with an approximately Zipfian distribution (Fig. 3). There are an average of 773 example responses per topic (prompt), with an average response length of 45.8 words. *TRN* has a wide range of candidate L1s, with the largest proportion being Gujarati L1. The evaluation data sets, *EVAL1-3, ALL*, are designed to have an (approximately) even distribution over CEFR grades levels [11]. Fig. 3 shows that the topic distribution of the evaluation data is less skewed than the training data but still roughly zipfian. *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*. A subset of very bad responses which had very poor ASR was removed from the evaluation data, which causes some discrepancies with previous work [6].

Data	#Topics	#Resp.	#Words	#Resp./ Topic	Avg.Resp. Length
TRN	379	293.0K	13.4M	773.2	45.8
EVAL1	92	1297	64.4K	14.1	49.7
EVAL2	177	1335	58.5K	7.5	43.8
EVAL3	179	1445	63.1K	8.1	43.7
ALL	219	4077	186.0K	18.6	45.6

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

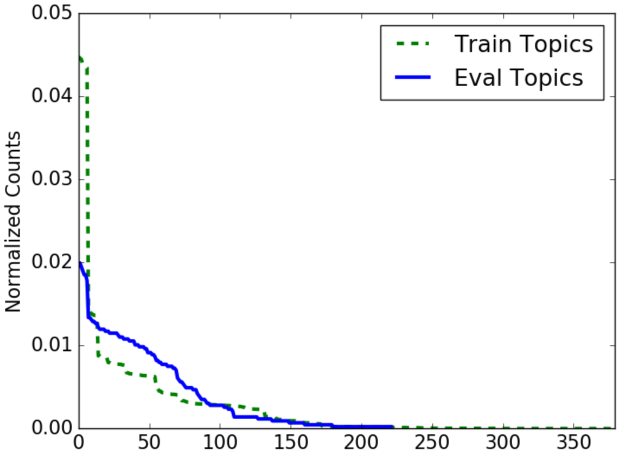


Fig. 3. Topic Distributions

3.2. Training Data Construction

The data is taken from tests run with human examiners so the responses are virtually all on topic. To produce negative, off-topic training examples the responses and prompts were shuffled during training via a dynamic sampling mechanism which samples mismatched prompts to a given response. The sampling mechanism can draw topics from the empirical topic distribution (Fig. 3) by parameterizing a distribution using empirical topic counts, or from a uniform distribution. Positive examples naturally come from the empirical topic distribution. If more than one negative example is shown for a particular response, the positive example is over-sampled the corresponding number of times to maintain a balanced training. As was shown in [12, 6], prompts from the same section tend to be more similar, and therefore more confusable. Two topic shuffling strategies were considered in [12, 6]: *Naive*, where prompts are shuffled across all sections; and *Directed*, where prompts are shuffled only within the same section. This work only considers *Naive* topic shuffling, as it represents the more likely scenario - real off-topic responses are unlikely to come predominantly from the same section. For multi-part prompts, which contain a main prompt that describes the overall question, and several (5 here) sub-prompts, all sub-prompts were pre-appended with the main prompt. These sub-prompts are considered distinct topics and thus competing negative examples to each other during shuffling.

3.3. ASR System

In this work a speaker independent hybrid DNN-HMM ASR system [13] is used. The acoustic model is trained on 108.6 hours of BULATS test data (Gujarati L1 speakers) using the HTK v3.5 toolkit [14, 15]. A Kneser-Ney trigram language model is trained on this data and interpolated with a general English language model trained on a large broadcast news corpus, using the SRILM toolkit [16]. The performance on this ASR system is described in Tables 2 and 3 relative to crowd-sourced transcriptions [17].

EVAL1	EVAL2	EVAL3	ALL
37.3	52.5	48.6	45.7

Table 2. ASR %WER on evaluation data sets

A1	A2	B1	B2	C
60.3	54.0	44.9	41.8	41.4

Table 3. ASR %WER per CEFR grade level on *ALL*

3.4. Model and Training Hyper-parameters

Both models were implemented in Tensorflow [18] and contain two 400 dimensional BiLSTM encoders with TanH non-linearities, 200 for the forward states and 200 for the backward states. The HATM also contains an additional 200-dimensional BiLSTM prompt-search encoder. The ATM was trained for 5 epochs with the Adam optimizer [19], an exponentially decaying learning rate with an initial value of 1e-3 and decay factor 0.85 per epoch. Dropout regularization [20] was applied to all layers except for the LSTM recurrent connections and word embeddings, with a keep probability of 0.8. The binary classifier was a DNN with 2 hidden layers of 200 rectified linear (ReLU) units and a 1-dimensional logistic output. The word embeddings, shared by all BiLSTMs, were initialized from an RNNLM language model trained on the *TRN* responses and kept fixed during training. The HATM was initialized from a trained ATM. For the first 3 epochs only the newly-initialized prompt-attention mechanism was trained. Further training for 1 more epoch is done with an unlocked response attention mechanism and a learning rate of 1e-4. The prompt and response encoders, as well as the DNN classifier remain locked. The ATM takes about 3.3 hours on an nVidia GTX 980M graphics card. Further training the HATM takes an extra hour.

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. To yield this, negative examples (true negatives) need to be introduced into the evaluation data sets via shuffling. The negative examples are drawn from the empirical topic distribution of the evaluation data. 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 ATM and HATM. Subsection 4.1 investigates several key properties of the models when all the prompts are seen. Firstly, the effect of sampling negative examples from the empirical and uniform distributions is assessed. Secondly, the effect of CEFR grade level [11] on relevance assessment performance is investigated. Finally, the nature of the prompt attention mechanism in the HATM is investigated. Subsection 4.2 investigates the performance of the model on unseen topics (prompts), analyses errors which the models make and compares results to previous work. Finally, subsection refsec:asr investigates the effect of using ASR confidence scores as extra features.

4.1. Baseline Performance

Table 4 shows the effect of different dynamic sampling of negative examples. For the ATM with empirical distribution samples (ATM-E) no benefit is seen increasing the number of negative examples, unlike [6]. In [6] the positive and negative prompt-response pairs were fixed, and thus using 5 negative samples increased the diversity. This is not necessary when using a sampling mechanism in training, as different negative prompt-response pairs are generated at every epoch. Using a uniform topic distribution degrades performance (ATM-U). There are likely two effects occurring - firstly there is a mismatch to the topic distributions in the evaluation data occurs. Secondly, the mismatch between in the topic distributions of the positive and negative examples likely skews the model towards treating rare topics as non-relevant. Models which use a uniform topic distribution for negative examples were not further investigated. Finally, the performance of the HATM-E and ATM-E models is comparable.

#samples	ATM-E	ATM-U	HATM-E
1	0.97	0.95	0.96
5	0.97	0.95	0.97

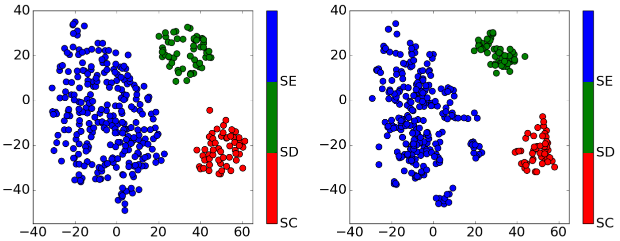
Table 4. Comparison of AUC for models with empirical (E) and uniform (U) negative sampling on *ALL*

Table 5 shows the baseline performance on *ALL* evaluation data corresponding to each CEFR [11] grade level for the ATM model with empirical distribution samples (ATM-E) evaluated on both ASR and crowd-sourced transcriptions. The latter are more accurate but mismatched to the ASR transcriptions used in the ATM training. Table 3 in section 3.3 shows that ASR error rates are lower on responses corresponding to higher grade levels, and table 5 shows that the performance of the ATM is higher on responses corresponding to higher grade levels. This trend was previously reported in [6]. However, there is very little difference between the performance on ASR and crowd-sourced transcriptions. This indicates that due to the low quality nature of the transcriptions which the system was trained on the system is unable to leverage the better quality of the crowd-sourced transcriptions. Furthermore, this is suggests the that differences in performance across grade level are not due to better transcriptions for higher grade responses, but due to the nature of the responses themselves.

SYS	A1	A2	B1	B2	C	ALL
ASR	0.91	0.96	0.98	0.98	0.98	0.97
CWD	0.92	0.96	0.98	0.98	0.99	0.97

Table 5. ATM-E Per grade level breakdown of performance on *ALL*

It is interesting to investigate what the prompt attention mechanism and the prompt encoder have learned in the HATM. Firstly, a t-SNE [21] projection of the original (ATM) prompt embeddings (Fig. 4a) is compared to the projection of the interpolated HATM embeddings (Fig. 4b). Both sets of embeddings form three distinct clusters, grouped by section. Notably, the interpolated embeddings reside in the same locations as the originals, though they appear to be more tightly grouped. The attention mechanism is also able to learn section distinctions very well and the confusion matrix (not shown) between prompt sections shows that the the system attends only over prompts of the corresponding section.



(a) Original ATM

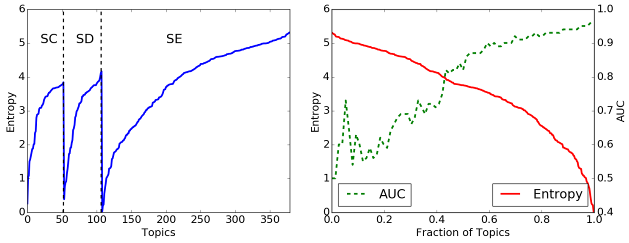
(b) Interpolated HATM

Fig. 4. Prompt embeddings

Fig. 5a shows the entropy of the attention mechanism, ordered first by section and then by increasing entropy. The plot clearly shows 3 distinct spikes, which correspond to sections C, D and E, respectively. This shows that there are topics within each section which the model is able to understand and focus on very well, and others for which it struggles to confidently find a similar topic. Furthermore, entropy is generally correlated with how common a topic is. Fig. 5b shows, on the same plot, a cumulative plot of AUC on subsets of the evaluation data *ALL* corresponding to adding topics in order of decreasing entropy, and the entropy of the added topics, in decreasing order. AUC increases as more low-entropy topics are added to the model. This suggests that entropy of the prompt attention mechanism can be used as a measure of uncertainty of the HATM's ability to accurately assess relevance. Thus, depending on the prompt, the model could reject all responses to these topics whose entropy is above a certain threshold to be assessed by humans, and process the rest automatically. This is an important advantage of the HATM over the ATM, despite their comparable performance.

4.2. Performance on Unseen Prompts

The proposed models' ability to generalize to new prompts is investigated in this section. 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 6), 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.



(a) Prompt Attention Entropy

(b) Prompt Attention Entropy AUC

Fig. 5. HATM entropy

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

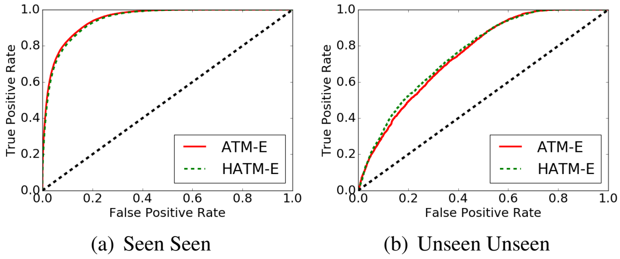
Table 6. Topic, response and word statistics of the prompt-response data sets used for 10-fold cross validation.

The prompts presented to the models in the following experiments are always either from subsets which are seen or unseen in the training data. As in section 4.1, evaluation responses are always new (not reused from the training data), but can be related to prompts either seen or unseen in training. Two strategies for shuffling evaluation responses for negative examples are considered: *seen*, *unseen*. The first uses responses to seen prompts as negative examples, the second uses responses to unseen prompts as negative examples. This produces four experiments which illustrate different aspects of how well the models understand what relates to seen prompts and how well they generalize to new, unseen prompts. 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. To decrease noise arising from particular shufflings of the evaluation data, 10 different random topic shufflings are used as negative examples and the positive examples are replicated 10 times for all 10 cross-validations folds.

Neg. Resp.	System	Seen Prompts	Unseen Prompts
Seen	ATM-E	0.949	0.855
	HATM-E	0.944	0.856
Unseen	ATM-E	0.938	0.751
	HATM-E	0.933	0.760

Table 7. Average AUC on *ALL-sub*

The results in Table 7 show that once prompts have been seen in training, the model has a clear understanding of what is relevant to them and is generally not sensitive to the nature of the negative-example responses. However, on unseen prompts there is a degradation of performance, which ranges from 0.751 to 0.855 for the ATM-E and from 0.760 to 0.856 for the HATM-E as evaluation response topic shuffling changes from *seen* to *unseen*. Clearly, the models are able to generalize well to and assess the relevance of unfamiliar responses to seen prompts, and to a lesser degree, are able to reasonably perform on new and unseen prompts, even in the extreme scenario (0.760 AUC). This is expected, as the models are exposed to a greater variety of responses than prompts. ROC curves for performance on seen and unseen prompts with corresponding response topic shuffling are shown in Fig. 6a and 6b. The experiments demonstrate a marginal advantage of the HATM over the ATM on unseen prompts.

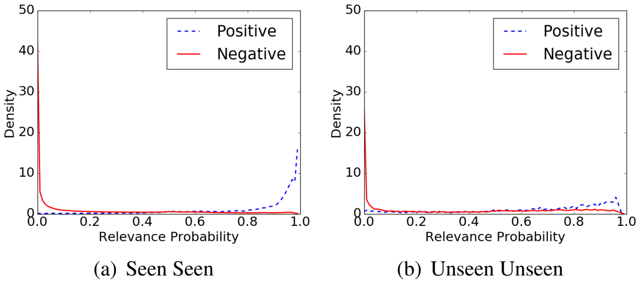


(a) Seen Seen

(b) Unseen Unseen

Fig. 6. ROC curves

It is interesting to analyze the mistakes which the system makes. To do this, the relevance probabilities for positive and negative examples are plotted as histograms for the scenarios where seen prompts are combined with seen responses (Fig. 7a) and unseen prompts with unseen responses (Fig. **??**b). The other scenarios yield similar histograms. When operating on seen data, the model is able to correctly classify most examples with very high/low relevance probabilities. However, when operating on unseen prompts it is able to confidently detect when prompts and responses are mismatched, but is unsure about matched prompt-response pairs for unseen prompts, which is the main failure case of these models.



(a) Seen Seen

(b) Unseen Unseen

Fig. 7. Relevance Probability Histograms

This suggests that the models, via the response attention mechanism, learn a 'lock and key' mechanism, where for a given response, only summation of the hidden states using weights derived from a matched prompt result in a high relevance prediction, and all other summations in a low relevance prediction. In the matched case for unseen prompts the models correctly do not yield a very low relevance score, but struggle to yield a high relevance score, which indicates a generalization issue. It should be noted that 'lock and key' behavior reflects the way the models are trained - each response in the training data is used as a positive example only once, when matched with an appropriate prompt, and many times as a negative example, when matched with any other prompt.

4.3. Use of ASR confidence Scores

As an initial experiment, word level confidence scores (mapped to remove biases [22]) from the ASR output were applied to modify the ATM input. The expectation was that these would help the ATM focus on words which the system was more confident about. The systems were evaluated on both ASR and crowd-sourced transcriptions.

Three methods of applying the mapped confidence scores were investigated: as an extra input into the response attention mechanism (ATM-E-C1); as direct multiplication of the un-normalized response attention weights by confidence scores (ATM-E-C2); weighing each response's contribution to the batch loss by the mean confidence scores that response (ATM-E-C3). For ATM-E-C1 the response similarity function was modified to use the confidence score γ_t^r as an extra scaled bias:

$$s(\tilde{\boldsymbol{h}}_i^p, \tilde{\boldsymbol{h}}_t^r, \gamma_t^r) = \boldsymbol{v}_{\boldsymbol{r}}^{\mathrm{T}} \tanh(\boldsymbol{\Lambda}_1 \tilde{\boldsymbol{h}}_i^p + \boldsymbol{\Lambda}_2 \tilde{\boldsymbol{h}}_t^r + \boldsymbol{b}_{\boldsymbol{\gamma}} \gamma_t^r + \boldsymbol{b}) \tag{13}$$

and for ATM-E-C2 the unnormalized attention weights were scaled by the confidence scores:

$$s(\tilde{\boldsymbol{h}}_i^p, \tilde{\boldsymbol{h}}_t^r, \gamma_t^r) = \gamma_t^r \cdot s(\tilde{\boldsymbol{h}}_i^p, \tilde{\boldsymbol{h}}_t^r) \tag{14}$$

For the crowd-sourced evaluation transcriptions the confidence scores were all set to 1.0. From table 8 it can be seen that there is no benefit and even a slight degradation of performance from using confidence scores, and neither do they do help the system to use the better quality of the crowd-sourced transcriptions.

Transcriptions	ATM-E	ATM-E-C1	ATM-E-C2	ATM-E-C3
ASR	0.97	0.97	0.96	0.96
CWD	0.97	0.97	0.97	0.97

Table 8. AUC performance comparison of effect of using ASR confidence scores in the ATM input.

5. CONCLUSIONS AND FUTURE WORK

This paper presented the Hierarchical Attention-based Topic Model (HATM), which extends the ATM to explicitly make use of the similarity between prompts. The HATM has comparable performance to the ATM, and on unseen prompts matched with unseen responses it performs slightly better. The primary advantage of the HATM is the prompt attention mechanism which learns a topic ontology in an unsupervised fashion. Crucially, the entropy of the prompt attention mechanism can be used as a measure of uncertainty in the HATM's ability to assess relevance.

This work analyzed the behavior and primary failure modes of the ATM and HATM, and it was determined that the models fail to classify unseen prompts with matched unseen responses as relevant with high probability. An initial study of the use of ASR confidence scores as additional features was conducted and yielded no positive results.

Clearly, the proposed models primarily suffer from a lack of topic balanced training data. Thus, data augmentation strategies should be investigated in future work to deal with the heavily skewed topic distribution of the training data. Furthermore, the training of the ATM and HATM on higher quality ASR transcriptions should also be investigated.

6. REFERENCES

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

[2] Klaus 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] Angeliki Metallinou and Jian Cheng, "Using Deep Neural Networks to Improve Proficiency Assessment for Children English Language Learners," in *Proc. INTERSPEECH*, 2014.

[4] Helen Yannakoudakis, "Automated assessment of English-learner writing," Tech. Rep. UCAM-CL-TR-842, University of Cambridge Computer Laboratory, 2013.

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

[6] Malinin A, K. Knill, A. Ragni, Y. Wang, and M.J.F. Gales, "An attention based model for off-topic spontaneous spoken respnse detection: An Initial Study," in *to be presented at ISCA Workshop on Speech and Language Technology for Education (SLaTE)*, 2017.

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

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

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

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

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

[12] Andrey Malinin, Rogier van Dalen, Kate Knill, Yu Wang, and Mark Gales, "Off-topic Response Detection for Spontaneous Spoken English Assessment," in *Proc. 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, Berlin, Germany, 2016, pp. 1075–1084.

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

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

[15] Steve Young et al., *The* HTK *book (for* HTK *version 3.5)*, University of Cambridge, 2015, `http://htk.eng.cam.ac.uk`.

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

[17] Rogier C. van Dalen, Kate M. Knill, Pirros Tsiakoulis, and Mark J. F. Gales, "Improving Multiple-Crowd-Sourced Transcriptions Using a Speech Recogniser," in *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2015.

[18] Martín Abadi et al., "TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems," 2015, Software available from tensorflow.org.

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

[20] Nitish 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.

[21] L. van der Maaten and G. Hinton, "Visualizing Data using t-SNE," *J. MLR*, vol. 1, pp. 1–49, 2008.

[22] G. Evermann and P. C. Woodland, "Large vocabulary decoding and confidence estimation using word posterior probabilities," in *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2000.