Complementary Systems for Off-Topic Spoken Response Detection

Vatsal Raina, Mark J.F. Gales, Kate Knill
Dept. of Engineering, Cambridge University
Cambridge, UK
{vr311, mjfg, kate.knill}@eng.cam.ac.uk

Abstract

Increased demand to learn English for business and education has led to growing interest in automatic spoken language assessment and teaching systems. With this shift to automated approaches it is important that systems reliably assess all aspects of a candidate's responses. This paper examines one form of spoken language assessment; whether the response from the candidate is relevant to the prompt provided. This will be referred to as off-topic spoken response detection. Two forms of previously proposed approaches are examined in this work: the hierarchical attention-based topic model (HATM); and the similarity grid model (SGM). The work focuses on the scenario when the prompt, and associated responses, have not been seen in the training data, enabling the system to be applied to new test scripts without the need to collect data or retrain the model. To improve the performance of the systems for unseen prompts, data augmentation based on easy data augmentation (EDA) and translation based approaches are applied. Additionally for the HATM, a form of prompt dropout is described. The systems were evaluated on both seen and unseen prompts from Linguaskill Business and General English tests. For unseen data the performance of the HATM was improved using data augmentation, in contrast to the SGM where no gains were obtained. The two approaches were found to be complementary to one another, yielding a combined $F_{0.5}$ score of 0.814 for off-topic response detection where the prompts have not been seen in training.

1 Introduction

Spoken language assessment of English is on the rise as English is the chosen language of discourse for many situations. Businesses and academic institutes demand rigorous assessment methods to ensure prospective employees and students exceed a baseline standard for English proficiency so they can succeed and contribute in their new environment. Standardised assessments such as IELTS (Cullen et al., 2014), Pearson Test of English Academic (Longman, 2010) and TOEFL (ETS, 2012) include "*free speaking*" tasks where the candidate speaks spontaneously in response to a prompted question to ensure their speaking skills are fully assessed. A candidate might attempt to achieve a higher grade by speaking a pre-prepared response, irrelevant to the prompt. For scoring validity it is important that measures are taken to detect any off-topic responses so they do not influence the final grade. This is particularly true for automatic assessment systems which are increasingly being deployed to cope with the growing demand for examinations and may see increased cheating if candidates are aware that a computerised system is responsible for grading them (e.g. (Mellar et al., 2018)). These systems are more susceptible to inaccurate scoring due to emphasis given to criteria such as fluency, pronunciation and language use, over topic relevance (Lochbaum et al., 2013; Higgins and Heilman, 2014).

Automatic off-topic spoken response detection systems based on attention (Malinin et al., 2017b,a) and similarity grid (Wang et al., 2019) models have shown good performance for prompts seen in training. For operational reasons it would be cost and time effective to be able to use the same systems on responses to new prompts, unseen in training i.e. removing the need to collect new data and retrain models prior to deployment. Yoon et al. (2017) has had some success with handling unseen prompts. This is still a challenging research problem, however, with significant degradation observed on even the best performing hierarchical attention-based topic model (HATM) (Malinin et al., 2017a), and with no assessment to date of Wang et al. (2019)'s similarity grid model (SGM) approach. This paper

Proceedings of the 15th Workshop on Innovative Use of NLP for Building Educational Applications, pages 41–51
July 10, 2020.

therefore focuses on investigating how to improve performance on unseen prompts for these models. It presents extensions to the HATM and SGM with the goal of learning robust representations of seen prompts for effective generalisation to unseen prompts. The resulting systems are shown to have complementary detection characteristics, yielding improved off-topic response detection when combined.

The remainder of the paper is structured as follows: Section 2 presents related work; Section 3 details the components of the HATM and SGM, proposing modifications to each (universal regularisation and multi-channel cosine-similarity, respectively) to make them more robust; data augmentation is proposed in Section 4 to overcome limited training data; the experimental set-up and structure of the data is described in Section 5; Section 6 presents the experimental results and analysis; conclusions are given in Section 7.

2 Related Work

Initial off-topic spoken response detection systems were based on vector space models, measuring the relevance between the spoken response and test prompts inspired by systems for written essays (Higgins et al., 2006; Louis and Higgins, 2010). Cheng and Shen (2011)'s approach using speech confidence derived features is unsuited to general free speaking tasks, whereas Yoon and Xie (2014) required hundreds of example responses for each prompt from highly proficient speakers.

Using word embeddings and deep neural networks to measure sentence similarity has since become dominant. Rei and Cummins (2016) generated sentence-level relevance scores for written essays by using various similarity metrics based on word embeddings. For spoken responses, Malinin et al. (2016) proposed a topic-adapted recurrent neural network language model (RNNLM) to rank prompt-response pairs. This handles sequential information but cannot handle prompts unseen in training so Malinin et al. (2017b) introduced an attention-based topic model (ATM) which can. The deep learning ATM architecture uses an attention mechanism to attend over response word embeddings with the prompt sentence embedding as the key. A hierarchical variant of the ATM (HATM) is proposed in Malinin et al. (2017a) where an additional (prompt) attention mechanism is incorporated to attend over sentence embeddings of a set of seen prompts with the test prompt embedding acting as the key. Hence, an unseen prompt is able to lock onto the vector representation of a combination of the seen prompts. The HATM assumes the set of seen prompts are sufficiently diverse to capture aspects of all possible unseen prompts. Malinin et al. (2017a) observed lower performance on unseen prompts due to a lack of diversity.

A radically different approach was proposed by Wang et al. (2019) based on initial work in Lee et al. (2017) and Yoon et al. (2017). Very deep CNNs are employed for automatic detection of off-topic spoken responses in which a prompt-response pair is represented as a similarity grid that can be interpreted as an image (this model will be referred to here as *SGM*). Similarity measurements are made based on word embeddings (or other distance metrics) of the prompt/response content words. The training data used in Wang et al. (2019) was from exams with generally short prompts which resulted in a limited number of content words to use. It was not assessed on unseen prompts.

3 Off-Topic Spoken Response Detection

This paper builds upon the hierarchical attention-based topic model (HATM) (Malinin et al., 2017a) (section 3.1) and the similarity grid in CNNs model (SGM) (Wang et al., 2019) (section 3.2) for detection of off-topic spoken responses to prompted questions. The candidate can answer freely so automatic speech recognition (ASR) is needed to determine the words in their response in each case. Both systems assign a probability that the ASR obtained response, $\hat{\mathbf{z}}$, is relevant to the prompt, $\hat{\mathbf{x}}$.

The performance of the HATM and SGM models on responses to prompts *seen* in training is high, with $F_{0.5}$ scores above 0.9. A key issue, however, in the practical deployment of off-topic response detection systems is handling responses to prompts that were *unseen* in training so that new examination questions can be asked without requiring example responses to be collected and the detection system retrained. Although Malinin et al. (2017a) improved off-topic detection on *unseen* prompts compared to earlier work, the performance is still quite far below that of *seen*, and the SGM has not been evaluated in the unseen prompt scenario. This section presents two approaches to potentially improve performance on unseen prompts: universal regularisation in an attention mechanism as a structural modification to the HATM in order to en-

courage generalisation; multi-channel SGM based on cosine distance (MSGM). Data augmentation strategies to increase the number of prompts available for training are presented in Section 4.

3.1 Attention-Based Model

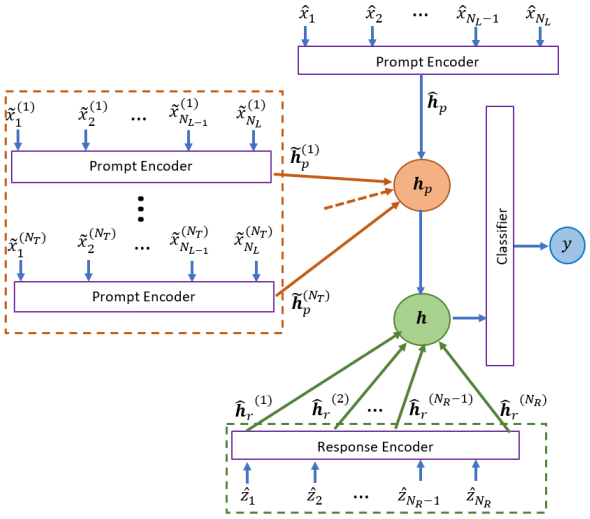

Figure 1: Hierarchical Attention-based Topic Model (HATM).

The Hierarchical Attention-based Topic Model (HATM) (Malinin et al., 2017a) is depicted in Figure 1. The system uses an encoding of the prompt word sequence, $\mathbf{h}_p$, as the key for an attention mechanism over an embedding of the response to yield a fixed length vector, $\mathbf{h}$, that is used to predict the probability, y, that the response was relevant to the prompt. To improve the robustness of the estimate of the prompt embedding an additional attention mechanism is run over all N_T embeddings of the prompts seen in training, $\tilde{\mathbf{h}}_p^{(1)}, \ldots, \tilde{\mathbf{h}}_p^{(N_T)}$. This attention mechanism uses an embedding of the test prompt, $\hat{\mathbf{h}}_p$, as the key to yield $\mathbf{h}_p$. This additional attention mechanism over the prompts was found to improve the performance of the system when the prompt had not been seen in the training data (Malinin et al., 2017a). The same network configuration as that used in Malinin et al. (2017a) was implemented in this work:

- bi-directional (Schuster and Paliwal, 1997) LSTMs (Hochreiter and Schmidhuber, 1997) were used as the encoders for both the prompts and the responses. Separate models were used for the prompt and response encoders[1];
- additive attention mechanisms were used for both the attention mechanism over the training prompts and that over the responses;
- the classifier used ReLU activation functions.

The parameters of the network were optimised using cross-entropy training. For the training of the prompt attention mechanism, the actual prompt was excluded from the attention mechanism, otherwise the attention mechanism simply focuses on the matched prompt embedding.

One of the issues observed with the HATM is that the performance of the system on unseen prompts is significantly poorer than the performance on seen prompts that have been seen, along with relevant responses, in the training data. This motivates the need for the model to improve the generalisation of the system to unseen prompts. Here, the prompt attention mechanism (see Figure 1) is targeted. A specific form of dropout, where training prompt embeddings are excluded from the prompt attention mechanism, referred to as *prompt-dropout*, is proposed. Denoting $\alpha_k = \texttt{Softmax}\left[\hat{\mathbf{h}}_p, \tilde{\mathbf{h}}_p^{(k)}, \theta_{pa}\right]$ as the attention weight for the k^{th} training prompt embedding, $\tilde{\mathbf{h}}_p^{(k)}$, with the test prompt embedding, $\hat{\mathbf{h}}_p$, as the key, random attention weights are set to zero during training in prompt-dropout such that

$$\alpha_k = \begin{cases} 0 & \text{w.p. } 1-\kappa \\ \texttt{Softmax}\left[\hat{\mathbf{h}}_p, \tilde{\mathbf{h}}_p^{(k)}, \theta_{pa}\right] & \text{w.p. } \kappa \end{cases} \quad (1)$$

where κ represents the keep-probability. The attention weights must be re-normalised after prompt-dropout. Sampling κ from a probabilistic distribution for each attention weight is motivated by Garnelo et al. (2018). In initial experiments this distribution of dropout rate was found to outperform selecting a fixed dropout rate.

3.2 Similarity Grid Model

The Similarity Grid Model (SGM) (Wang et al., 2019) represents a prompt-response pair as a "similarity image". This image is transformed by an Inception network into a measure of the degree of relevance between the test prompt and test response.

Initially all stop words are removed such that a given prompt and response pair only consists of

[1]More complex sentence and word embeddings, such as BERT (Devlin et al., 2019) were examined in initial experiments but were not found to yield performance gains.

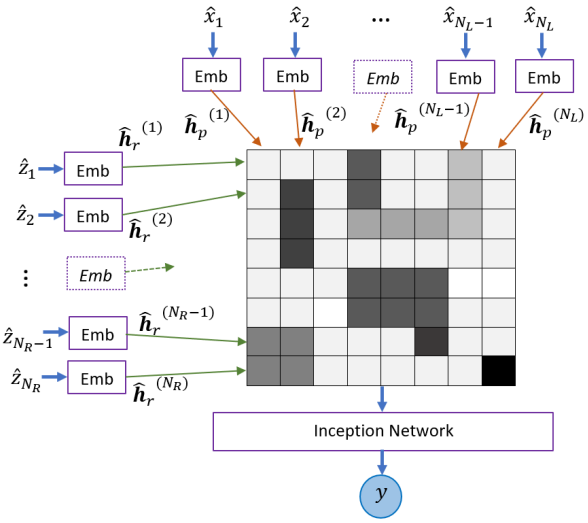

Figure 2: Similarity Grid Model (SGM) where the grid "pixel" colours indicate the level of similarity.

content words[2]. After this pre-processing the similarity grid model for a prompt-response pair is as shown in Figure 2. For all content words in the test prompt, $\{\hat{x}_i\}_{i=1}^{N_L}$, and test response, $\{\hat{z}_i\}_{i=1}^{N_R}$, word embeddings, $\left\{\hat{\mathbf{h}}_p^{(i)}\right\}_{i=1}^{N_L}$ and $\left\{\hat{\mathbf{h}}_r^{(i)}\right\}_{i=1}^{N_R}$, are computed. These word embeddings are used to construct a similarity grid, $\mathbf{S}$. This two-dimensional grid has N_L columns and N_R rows and the cell position (i, j) holds an inverse similarity metric between the i^{th} prompt word embedding, $\hat{\mathbf{h}}_p^{(i)}$, and the j^{th} response word embedding, $\hat{\mathbf{h}}_r^{(j)}$. $\mathbf{S}$ is then resized to 180×180 in order to make the similarity grid of standard size regardless of the number of content words in the test prompt and response. Perceiving the similarity grid as an image, an Inception network transforms the resized image to a value $0 \leq y \leq 1$, indicating the degree of relevance between the test prompt and test response.

The network configuration used in this work is closely related to that used in Wang et al. (2019):

- context independent word embeddings are computed for each word in the prompt and the response. The embeddings for both the prompt and response are tied;
- a cosine distance to compute the distance between each prompt response word embedding pair, followed by a bilinear transform was used to resize the similarity grid;
- the Resnet-152 (He et al., 2016b) Inception network was used.

The similarity grid has one channel i.e. one single measurement value per cell. This can be extended to multiple channels (MSGM), with different forms of embeddings or distance functions used to compute the grid in each channel. Wang et al. (2019) used cosine distances in the first channel and inverse document frequency (IDF) values of the prompt and response words for the second and third channel, respectively. For this paper a MSGM with three channels where each channel represents the cosine distance between prompt and response word embeddings with a different set of word embeddings learnt for each channel is used[3]. The variety in the embeddings, and resulting channel, is achieved by using different initialisation seeds with the same network configuration. As the Inception network filters over the channels are simultaneously trained the resulting filters will be complementary.

4 Data Augmentation

In general, the performance of off-topic response systems is limited by insufficient unique prompts being available for training. Data augmentation, where the training data is modified in some way to create new examples, is regularly applied on low resource tasks in areas such as speech recognition (e.g. (Cui et al., 2015)) and computer vision (e.g. (Shorten and Khoshgoftaar, 2019)) and has had some success in NLP (e.g. (Zhang et al., 2015; Kafle et al., 2017; Wei and Zou, 2019)). This motivates investigating if augmentation of the training prompts such that multiple versions of each prompt are generated can help improve robustness[4]. Prompt augmentation will permit the model to explore the region around each unique prompt rather than being restricted to a discrete point in the high-dimensional prompt-space.

Both structured and unstructured data augmentation techniques are considered here. Augmentation of prompts is performed *on-the-fly* during training. Note, the hierarchy of seen prompts of the HATM

[2]A comprehensive list of stop words is provided by `nltk.corpus` https://www.nltk.org/api/nltk.corpus.html.

[3]In this work, the use of IDF values in the second and third MSGM channels did not improve performance over the SGM so was not used.

[4]Data augmentation of training responses is also possible. This was found to degrade performance in initial experiments, possibly due to there being a large number of diverse responses available for training without augmentation and issues with generating sensible back-translations on ASR output.

in the prompt attention mechanism (Figure 1) does not include the additional augmented prompts because the expectation is that augmented prompts will not dramatically differ from the original unique prompts.

Easy Data Augmentation (EDA) techniques were trialled by Wei and Zou (2019). They proposed that different variants of any textual data can be generated using a combination of synonym replacement, random insertion, random swap and random deletion of words. A single hyper-parameter, α, controls the fraction of words that are modified in the original text. Using the default value of $\alpha = 0.1$, prompts are augmented using the above techniques to replace, insert, swap or delete 10% of the words randomly in the original prompt. This structured augmentation approach should enable the model to learn a more robust representation of each unique prompt.

Back-translation is employed as an unstructured method to augment the amount of available training data. A machine translation model is employed to translate a given training prompt into a foreign language by taking the maximum likelihood output. Then a reverse machine translation model takes the prompt in the foreign language and translates it back into English. The expectation is that the original and final pair of English prompts will be very similar in meaning but will have a different ordering and choice of specific words. Therefore, the back-translated prompt can be treated as a new prompt which can be paired with the original prompt's response to generate a new prompt-response pair. The use of several different languages permits the creation of several variants of the same prompt. Translation can be achieved using standard machine translation packages.

5 Data and Experimental Set-Up

The HATM and SGM models and the proposed extensions were assessed on their ability to detect off-topic responses to prompts in free speaking tests where the candidates can talk for up to one minute in answering the question.

5.1 Training and evaluation data

Data from the Cambridge Assessment English Linguaskill Business and Linguaskill General English tests[5] are used in the training and evaluation of the systems. The two tests are similar in format but with different foci and therefore vary in the topics discussed by candidates and their associated vocabularies. They are multi-level, global, tests - i.e. taken by candidates from across the CEFR levels, A1-C2, with a wide range of first languages (L1s) and variation in response proficiency.

Both Linguaskill speaking tests are comprised of five parts. For this paper only prompts and corresponding responses from the three long free speaking parts are used. The candidate has 60 seconds in parts 3 and 4 to talk on a topic such as *advice for a colleague/friend* and discuss a picture or graph, respectively. Part 5 consists of 20 second responses to a set of five contextualised prompts, such as *starting a retail business*, or *talk about a hobby*. The diversity of these prompts is discussed by Malinin et al. (2017a).

Data	TRN	SEEN	UNS
#Prompts	379	219	56
#Responses	257.2K	40.8K	85.0K
Avg. prompt length	51	51	55
content words	28	28	29
Avg. resp. length	48	43	42
content words	22	20	19

Table 1: Prompt/response statistics for training (TRN) and *seen* (SEEN) and *unseen* (UNS) evaluation data sets.

Table 1 outlines the statistics for the training (TRN) and two evaluation data sets (SEEN and UNS). TRN and SEEN are taken from the Linguaskill Business test and UNS from the Linguaskill General English test. There is no overlap in speakers between any of the data sets. The response texts are generated automatically from the 1-best hypotheses from an ASR system with a word error rate (WER) of 25.7% on Business English data. TRN consists of a total of 257.2K responses to 379 unique prompts, an average of 679 responses per prompt compared with 186 for SEEN and 1518 for UNS. The average number of words are similar across the 3 data sets, with prompts (51-55) being slightly longer than responses (42-48) on average. This reduces by about half when content words only are included. The HATM is trained and evaluated using all the words in every textual prompt-response pair while the SGM is trained and evaluated using only the content words in every prompt-response pair.

[5] https://www.cambridgeenglish.org/exams-and-tests/linguaskill/

5.1.1 Training data construction

All responses are taken from tests assessed by human examiners, which permits the assumption that all responses in the data are on-topic. Therefore synthetic off-topic responses have to be created to train the systems. The off-topic data is generated using a dynamic sampling mechanism; this matches responses from one prompt with a different prompt. Balance is maintained such that the empirical distribution of topics in the on-topic examples is mimicked in the generation of synthetic off-topic examples. Off-topic examples for training data are generated *on-the-fly* (Malinin et al., 2017a) instead of producing a fixed set of negative examples prior to training as in Malinin et al. (2016) because dynamic sampling allows the diversity of negative examples to be efficiently increased. For each on-topic example, one off-topic example is generated.

For the data augmentation experiments the number of prompts was increased by a factor of 10 using EDA (Wei and Zou, 2019) or machine translation, and 20 when both were applied. The default value of $\alpha = 0.1$ was used in EDA to change 10% of words in the original prompt by replacing, inserting, swapping and/or deleting. Machine translation was performed offline using the Babylon MT system [6]. Back-translations were generated using 9 different languages[7].

5.1.2 Evaluation data construction

Due to the scarcity of real off-topic examples, negative off-topic examples are generated by permuting on-topic examples for SEEN and UNS. Each on-topic example has ten off-topic examples generated and duplicated ten times to maintain balance. Data set SEEN is formed from prompts that have been seen during training and negative responses that correspond to a different set of prompts seen during training. Data set UNS consists of prompts that are unseen during training and negative responses that correspond to prompts that are unseen during training too. Forming negative examples by permuting on-topic examples is reasonable because real off-topic examples by candidates are anticipated to consist of responses to a different prompt to that being answered.

5.2 Hyper-parameters and models

The HATM consists of two 400 dimensional BiLSTM encoders with 200 forward and backward states each and TanH non-linearities. 200 dimensional parameters are used for the prompt attention mechanism. The binary classifier is a DNN with 2 hidden layers of 200 rectified linear (ReLU) units and a 1-dimensional logistic output. Dropout regularisation (Srivastava et al., 2014) with a keep probability of 0.8 is applied to all layers except for the LSTM recurrent connections and word embeddings. The universal regularisation samples its keep probability using $\kappa \sim \mathcal{U}(0.05, 0.95)$. The HATM is initialised from an attention-topic model (ATM) as described in Malinin et al. (2017a). It is trained for 3 epochs using an Adam optimizer, with an exponentially decaying learning rate initialised at 1e-3 and decay factor of 0.85 per epoch. The first two epochs train only the prompt attention mechanism and the final epoch is used to train the whole network apart from the DNN binary classifier. This configuration was optimised using seen development data, similarly for the SGM. The ATM takes approximately 3 hours to train and an additional 1 hour for the HATM on an nVidia GTX 980M graphics card.

The SGM learns 200 dimensional word embeddings for each word in the prompt and response. ResNet-152 (He et al., 2016a)[8] with 152 residual layers is used as the Inception network with a 1-dimensional logistic output. The SGM is trained for 1 epoch using an Adam optimizer with a learning rate of 1e-3. It takes about 2 hours to train on an nVidia GTX 980M graphics card. The extended HATM and the SGM were built in Tensorflow[9].

The HATM and SGM results reported are computed on an ensemble of 15 models unless noted otherwise. Each model has an identical architecture and training parameters but each has a different initial seed value, creating modeling diversity. For this work a large ensemble is reported to minimise variance on the ensemble performance results. No analysis of efficiency is given. Approaches such as ensemble distillation (Hinton et al., 2015) can be directly applied to reduce computational cost.

[6] https://translation.babylon-software.com/english/Offline/

[7] Machine translation languages: Arabic, French, German, Greek, Hebrew, Hindi, Japanese, Korean, Russian.

[8] Code available at https://github.com/KaimingHe/resnet-1k-layers.

[9] Code available at https://github.com/VatsalRaina/HATM.

5.3 Performance criteria

Following Wang et al. (2019), precision and recall are used to assess performance except $F_{0.5}$ is preferred over F_1 as there is a greater interest in achieving a higher precision compared to recall: a candidate's response should not be mistakenly classified as off-topic as such responses are to be assigned a score of 0. This is a more standard metric than the area under the curve (AUC) used in Malinin et al. (2017a) and more intuitive in terms of test evaluation. Note, the results are given for a particular instance of permuting the off-topic examples for evaluation.

6 Experimental Results

This section presents the results of experiments performed on SGM, MSGM and extended HATM systems. Section 6.1 compares the performance of the baseline HATM with the MSGM on the *unseen* (UNS) and *seen* (SEEN) evaluation data sets. Section 6.2 explores the improvement in performance due to extending the baseline HATM using universal regularisation and prompt augmentation strategies. Finally, 6.3 investigates the complementary nature of the MSGM and the extended HATM. The prompt-specific performance of the combined system is considered in Section 6.4.

6.1 Baseline systems

	Model	P	R	$F_{0.5}$
SEEN	HATM	—	—	0.918 ±0.010
	-ensemble	0.963	0.841	0.936
	MSGM	—	—	0.905 ±0.009
	-ensemble	0.943	0.838	0.920
UNS	HATM	—	—	0.612 ±0.032
	-ensemble	0.815	0.370	0.657
	MSGM	—	—	0.767 ±0.019
	-ensemble	0.833	0.691	0.800

Table 2: Comparison of baseline HATM [$\mathcal{B}$] and MSGM, for *seen* (SEEN) and *unseen* (UNS).

Table 2 shows the baseline performance for the HATM and MSGM models. There is a relatively low variance between individual system results but combining the outputs in an ensemble improves the $F_{0.5}$ score in each case, with a larger gain (0.045/0.033 vs 0.018/0.015) observed on the unseen data. The HATM performs slightly better on the seen data than the MSGM, with a higher $F_{0.5}$ and a similar but always slightly higher precision-recall curve (Figure 3). For unseen data, however, the reverse is true with MSGM having a higher $F_{0.5}$ score of 0.800 compared to 0.657 for the baseline HATM. From Figure 3 it can be seen that the precision-recall curves for the HATM and SGM/MSGM systems on unseen data are quite different in shape. The HATM has a higher precision at the lowest recall but this drops quickly as the threshold increases. The degradation in the MSGM precision is much more gradual.

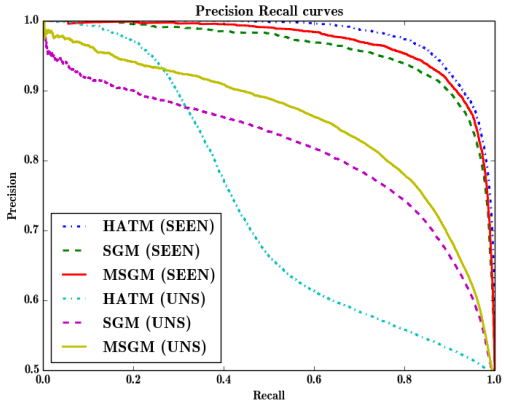

Figure 3: Comparison of precision-recall curves for baseline ensemble systems for HATM, SGM and MSGM for *seen* (SEEN) and *unseen* (UNS).

Figure 3 confirms that the focus should be on improving the performance on the unseen evaluation data set. The use of multi-channels benefits the similarity grid model as can be seen in Figure 3, with a SGM $F_{0.5}$ score of 0.908 on the seen and 0.768 on the unseen data sets, respectively. These gains are similar to those observed in Wang et al. (2019). Therefore, the results in the following sections will only be presented for the unseen evaluation data set and MSGM systems.

6.2 Regularisation and data augmentation

Universal regularisation and data augmentation were applied to the HATM to see if they improved detection performance. From Table 3 and Figure 4, it is evident that the universal regularisation on the prompt attention mechanism yields an increase in the $F_{0.5}$ score. Both the structured techniques and the machine translation (MT) prompt data augmentation strategies produce a boost in performance on the baseline HATM with universal regularisation. MT yields a much larger gain but the structured technique is shown to be complementary by

Model	P	R	$F_{0.5}$
$\mathcal{B}$	0.815	0.370	0.657
$\mathcal{B} \oplus \mathcal{P}_\mathcal{D}$	0.846	0.386	0.683
$\mathcal{B} \oplus \mathcal{P}_\mathcal{D} \oplus \mathcal{A}_\mathcal{E}$	0.790	0.464	0.693
$\mathcal{B} \oplus \mathcal{P}_\mathcal{D} \oplus \mathcal{A}_\mathcal{M}$	0.877	0.529	0.775
$\mathcal{B} \oplus \mathcal{P}_\mathcal{D} \oplus \mathcal{A}_\mathcal{E} \oplus \mathcal{A}_\mathcal{M}$	0.891	0.524	0.782

Table 3: Impact of universal regularisation, $\mathcal{P}_\mathcal{D}$, and data augmentation ($\mathcal{A}_\mathcal{E}$ = structured techniques and $\mathcal{A}_\mathcal{M}$ = machine translation) on baseline HATM, $\mathcal{B}$, for *unseen* (UNS).

	Model	P	R	$F_{0.5}$
SEEN	HATM	0.956	0.802	0.921
	MSGM	0.943	0.838	0.920
	Comb	0.962	0.839	0.935
UNS	HATM	0.891	0.524	0.782
	MSGM	0.833	0.691	0.800
	Comb	0.875	0.635	0.814

Table 4: Impact of combining models SGM and extended HATM [$\mathcal{B} \oplus \mathcal{P}_\mathcal{D} \oplus \mathcal{A}_\mathcal{E} \oplus \mathcal{A}_\mathcal{M}$] on *seen* (SEEN) and *unseen* (UNS). Comb = HATM & MSGM.

a further improvement when prompts are generated by both methods which was larger than observed when simply doubling the MT augmented prompts. Hence, the extended HATM is defined as $\mathcal{B} \oplus \mathcal{P}_\mathcal{D} \oplus \mathcal{A}_\mathcal{E} \oplus \mathcal{A}_\mathcal{M}$.

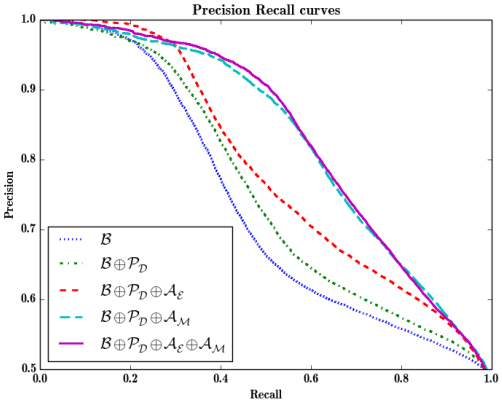

Figure 4: Impact of universal regularisation, $\mathcal{P}_\mathcal{D}$, and data augmentation ($\mathcal{A}_\mathcal{E}$ = structured techniques and $\mathcal{A}_\mathcal{M}$ = machine translation) on baseline HATM, $\mathcal{B}$, on precision-recall curves for *unseen* (UNS).

Experiments were also run on applying data augmentation to SGM. This led to significant drops in $F_{0.5}$, probably as a result of the SGM over-fitting to the training data.

6.3 Combining MSGM and extended HATM

As for the baseline HATM, the precision-recall curve for the extended HATM displays different characteristics to MSGM on the unseen data set as shown in Figure 5. These systems are complementary; combining the systems by averaging their outputs yields precision-recall curves which boost the precision at each recall level, giving a small gain over the best individual system at each threshold. The individual $F_{0.5}$ scores are boosted on the unseen data set from 0.782 and 0.800 to 0.814 and from 0.921 and 0.920 to 0.935 on the seen data set for the HATM and MSGM systems, respectively (Table 4). For comparison with Wang et al. (2019), the combined system here has an F_1 score of 0.922 and 0.807 on the seen and unseen data sets respectively.

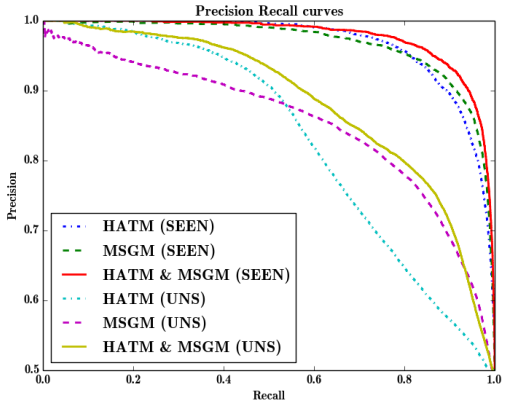

Figure 5: Impact of combining models MSGM and extended HATM [$\mathcal{B} \oplus \mathcal{P}_\mathcal{D} \oplus \mathcal{A}_\mathcal{E} \oplus \mathcal{A}_\mathcal{M}$] on *seen* (SEEN) and *unseen* (UNS).

6.4 Prompt-specific performance analysis

The performance of any off-topic response detection system is expected to depend on both the nature of the prompts, and how "close" a test prompt is to one seen in the training data. Yoon et al. (2017) found that performance varied substantially across different prompts. In this work the 10 most common, in the sense of having a large number of responses, unseen prompts in UNS were used to analyse the prompt specific performance. These common prompts should give robust per-prompt $F_{0.5}$ scores. The average performance of the combined MSGM and extended HATM system on this subset of prompts was 0.832, with a standard deviation of 0.048. This standard deviation across prompts is approximately half of the value presented in Yoon et al. (2017). As the prompts for that data are not

available, however, it is unclear whether this reduction is due to the nature of the prompts or improved generalisation of the combined model.

From Table 4 there is a large $F_{0.5}$ performance difference between prompts seen during training, 0.935, and those not seen, 0.814. Given this variation in performance, it is interesting to see whether the performance of an individual test prompt can be predicted given the set of training prompts. For prompts that are expected to perform poorly it would then be possible to collect training data.

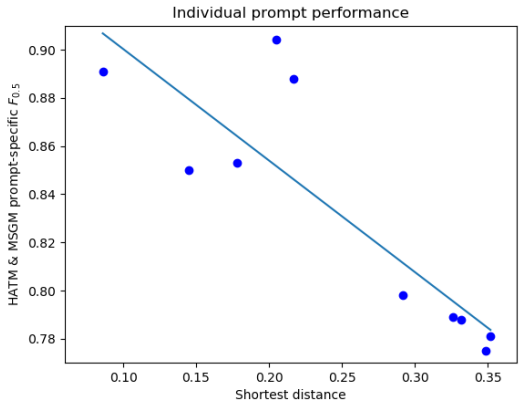

Figure 6: Relationship between unseen prompt distance to closest seen prompt and $F_{0.5}$ performance of MSGM & extended HATM $[\mathcal{B} \oplus \mathcal{P}_{\mathcal{D}} \oplus \mathcal{A}_{\mathcal{E}} \oplus \mathcal{A}_{\mathcal{M}}]$ on the unseen prompts subsets.

In Malinin et al. (2017a) the entropy of the prompt attention mechanism was used to rank performance of the prompts based on area under the curve metrics. From initial experiments this was not found to be a good predictor of $F_{0.5}$ score on these unseen test prompts. In this work, the cosine distance from the test prompt embedding, $\hat{\mathbf{h}}_p$, and each of the training prompt embeddings, $\tilde{\mathbf{h}}_p^{(i)}$, was computed. The closest distance was then used as the measure of similarity of the individual test prompt to the training data prompts. Figure 6 shows the individual $F_{0.5}$ score against this distance again using the 10 most common unseen prompts. There is a strong negative correlation, an R^2 statistic of 0.739, between the individual prompt performance and its distance to the closest seen prompt, showing the cosine distance between the prompt embeddings is a good indicator of unseen prompt performance.

From Figure 6 the cosine distance allows the unseen prompts to be partitioned into two distinct groups, *close* and *far* prompts with respect to the training prompts. The performance of all the unseen prompts was then evaluated using a distance threshold of 0.24 at the same operating point as Table 4. This yielded $F_{0.5}$ of 0.833 for *close*, and 0.777 for *far* prompts. Note for all distance thresholds examined, that resulted in a split of the unseen prompts, *close* always outperformed *far*.

7 Conclusion

This paper addresses the issue of off-topic detection in the context of unconstrained spoken responses to prompts. In particular, the problem of robustness to prompts unseen in training is considered. The Hierarchical Attention-based Topic Model (HATM) (Malinin et al., 2017a) and Similarity Grid Model (Wang et al., 2019) are compared and extended. Universal regularisation and data augmentation, from structured techniques and machine translation, increased the HATM $F_{0.5}$ by 19% relative to 0.782 on the unseen evaluation set. This contrasts with a three channel SGM (MSGM) based on cosine distances between prompt and response embeddings which yielded $F_{0.5}$ of 0.800.

The extended HATM and MSGM are shown to have very different precision-recall characteristics on unseen prompts, with the HATM having a very high precision at low recall but with a fairly sharp drop-off whilst the SGM's precision does not reach quite the same level but degrades at a much more gradual rate. The best individual systems are found to be complementary, with system combination boosting off-topic response detection on both unseen and seen prompts, achieving the best performance of $F_{0.5}$ of 0.814 on unseen and 0.935 on seen prompts. This combined system closely follows, and slightly enhances, the envelope of the best precision-recall path across the two individual systems. Finally the distance between a test prompt and the closest training is shown to predict the system performance, indicating which prompts may require additional training data to be collected.

8 Acknowledgements

This paper reports on research supported by Cambridge Assessment, University of Cambridge. Thanks to Cambridge English Language Assessment for support and access to the Linguaskill data. The authors would like to thank members of the ALTA Speech Team for generating the ASR transcriptions for the responses, and the initial implementation of the code from Malinin et al. (2017a).

References

J. Cheng and J. Shen. 2011. Off-topic detection in automated speech assessment applications. In *Proc. INTERSPEECH 2011*, pages 1597–1600.

X. Cui, V. Goel, and B. Kingsbury. 2015. Data augmentation for deep neural network acoustic modeling. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(9):1469–1477.

P. Cullen, A. French, and V. Jakeman. 2014. *The Official Cambridge Guide to IELTS*. Cambridge University Press.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. 2019 Conf. North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

ETS. 2012. *The Official Guide to the TOEFL® Test*, fourth edition edition. McGraw-Hill.

M. Garnelo, D. Rosenbaum, C. Maddison, T. Ramalho, D. Saxton, M. Shanahan, Y. W. Teh, D. Jimenez Rezende, and S. M. Ali Eslami. 2018. Conditional Neural Processes. In *Proc. 35th International Conference on Machine Learning, ICML*, pages 1690–1699.

K. He, X. Zhang, S. Ren, and J. Sun. 2016a. Deep residual learning for image recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016*, pages 770–778.

K. He, X. Zhang, S. Ren, and J. Sun. 2016b. Identity mappings in deep residual networks. In *Proc. European Conference Computer Vision (ECCV)*, pages 630–645.

D. Higgins, J. Burstein, and Y. Attali. 2006. Identifying off-topic student essays without topic-specific training data. *Natural Language Engineering*, 12:145 – 159.

D. Higgins and M. Heilman. 2014. Managing what we can measure: Quantifying the susceptibility of automated scoring systems to gaming behavior. *Educational Measurement: Issues and Practice*, 33(3):36–46.

G. E. Hinton, O. Vinyals, and J. Dean. 2015. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531.

S. Hochreiter and J. Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9:1735–80.

K. Kafle, M. Yousefhussien, and C. Kanan. 2017. Data augmentation for visual question answering. In *Proc. 10th International Conference on Natural Language Generation*, pages 198–202.

C. H. Lee, S.-Y. Yoon, X. Wang, M. Mulholland, I. Choi, and K. Evanini. 2017. Off-topic spoken response detection using Siamese convolutional neural networks. In *Proc. INTERSPEECH 2017*, pages 1427–1431.

K. E. Lochbaum, M. Rosenstein, P. Foltz, and M. A. Derr. 2013. Detection of gaming in automated scoring of essays with the IEA. In *Proc. 75th Annual meeting of NCME*.

P. Longman. 2010. *The Official Guide to Pearson Test of English Academic*. Pearson Education ESL.

A. Louis and D. Higgins. 2010. Off-topic essay detection using short prompt texts. In *Proc. NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 92–95.

A. Malinin, R. C. van Dalen, K. Knill, Y. Wang, and M. J. F. Gales. 2016. Off-topic response detection for spontaneous spoken english assessment. In *Proc. 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016*.

A. Malinin, K. Knill, and M. J. F. Gales. 2017a. A hierarchical attention based model for off-topic spontaneous spoken response detection. In *Proc. 2017 IEEE Automatic Speech Recognition and Understanding Workshop, ASRU*, pages 397–403.

A. Malinin, K. Knill, A. Ragni, Y. Wang, and M. J. F. Gales. 2017b. An attention based model for off-topic spontaneous spoken response detection: An initial study. In *Proc. 7th ISCA International Workshop on Speech and Language Technology in Education, SLaTE 2017*, pages 144–149.

H. Mellar, R. Peytcheva-Forsyth, S. Kocdar, A. Karadeniz, and B. Yovkova. 2018. Addressing cheating in e-assessment using student authentication and authorship checking systems: teachers' perspectives. *Int. J. Educ. Integr.*, 14(2).

Marek Rei and Ronan Cummins. 2016. Sentence similarity measures for fine-grained estimation of topical relevance in learner essays. In *Proc. 11th Workshop on Innovative Use of NLP for Building Educational Applications, BEA@NAACL-HLT 2016*, pages 283–288.

M. Schuster and K. Paliwal. 1997. Bidirectional recurrent neural networks. *Signal Processing, IEEE Transactions on*, 45:2673 – 2681.

C. Shorten and T. M. Khoshgoftaar. 2019. A survey on image data augmentation for deep learning. *J. Big Data*, 6:60.

N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958.

X. Wang, S.-Y. Yoon, K. Evanini, K. Zechner, and Y. Qian. 2019. Automatic detection of off-topic spoken responses using very deep convolutional neural networks. In *Proc. INTERSPEECH 2019*, pages 4200–4204.

J. Wei and K. Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proc. 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388.

Su-Youn Yoon, Chong Min Lee, Ikkyu Choi, Xinhao Wang, Matthew Mulholland, and Keelan Evanini. 2017. Off-topic spoken response detection with word embeddings. In *Proc. INTERSPEECH 2017*, pages 2754–2758.

Su-Youn Yoon and Shasha Xie. 2014. Similarity-based non-scorable response detection for automated speech scoring. In *Proc. Ninth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 116–123.

X. Zhang, J. Zhao, and Y. LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems 28*, page 649–657.