# Off-topic Response Detection for Spontaneous Spoken English Assessment

**Andrey Malinin, Rogier C. Van Dalen, Yu Wang, Kate M. Knill, Mark J. F. Gales**
**University of Cambridge, Department of Engineering**
**Trumpington St, Cambridge CB2 1PZ, UK**
{am969, yw396, kate.knill, mjfg}@eng.cam.ac.uk

## Abstract

Automatic spoken language assessment systems are becoming increasingly important to meet the demand for English second language learning. This is a challenging task due to the high error rates of, even state-of-the-art, non-native speech recognition. Consequently current systems primarily assess fluency and pronunciation. However, content assessment is essential for full automation. As a first stage it is important to judge whether the speaker responds on topic to test questions designed to elicit spontaneous speech. Standard approaches to off-topic response detection assess similarity between the response and question based on bag-of-words representations. An alternative framework based on Recurrent Neural Network Language Models (RNNLM) is proposed in this paper. The RNNLM is adapted to the topic of each test question. It learns to associate example responses to questions with points in a topic space constructed using these example responses. Classification is done by ranking the topic-conditional posterior probabilities of a response. The RNNLMs associate a broad range of responses with each topic, incorporate sequence information and scale better with additional training data, unlike standard methods. On experiments conducted on data from the Business Language Testing Service (BULATS) this approach outperforms standard approaches.

## 1 Introduction

As English has become the global *lingua franca*, there is growing demand worldwide for assessment of English as a second language (Seidlhofer, 2005). To assess spoken communication, spontaneous speech is typically elicited through a series of questions such as 'describe the photo' or 'plan a meeting'. Grades are awarded based on a candidate's responses.

Automatic assessment systems are becoming attractive as they allow second language assessment programmes to economically scale their operations while decreasing throughput time and provide testing on demand. Features for automatic graders are derived from the audio and from hypotheses produced by automatic speech recognition (ASR) systems. The latter is highly errorful due to the large variability in the input speech; disfluencies common to spontaneous speech, non-native accents and pronunciations. Current systems, such as ETS' *SpeechRater* (Zechner et al., 2009) and Pearson's *AZELLA* (Metallinou and Cheng, 2014), primarily assess pronunciation and fluency. Although these are clearly indicative of spoken language ability, full assessment of spoken communication requires judgement of high-level content and communication skills, such as response construction and relevance. The first stage of this is to assess whether the responses are off-topic, that is, has the candidate misunderstood the question and/or memorised a response.

While there has been little work done on detecting off-topic responses for spoken language assessment, detection of off-topic responses and content assessment has been studied for essay assessment. One approach for essay content assessment uses features based on semantic similarity metrics between vector space representations of responses. Common vector representations include lexical Vector Space Models and Latent Semantic Analysis (LSA) (Yannakoudakis, 2013). This approach was first applied to spoken assess-



*Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1075–1084,
Berlin, Germany, August 7-12, 2016. 

ment in (Xie et al., 2012) and then in (Evanini et al., 2013). Following this, (Yoon and Xie, 2014) investigated the detection of responses for which an automatic assessment system will have difficulty in assigning a valid score, of which off-topic responses are a specific type. A decision tree classifier is used with features based on cosine similarity between a test response and *tf-idf* vectors of both aggregate example responses and questions, as well as pronunciation and fluency. In (Evanini and Wang, 2014) text reuse and plagiarism in spoken responses are detected using a decision tree classifier based on vector similarity and lexical matching features which compare a response to a set of example 'source texts' . This task is similar to off-topic response detection in that it is based on comparing a test response to example responses. Thus, a standard approach to off-topic response detection would be based on measuring the similarity between vector representations of a spoken response and the test question. A major deficiency of this approach is that it is based on bag-of-words vector representations, which loses information about the sequential nature of speech, which is important to evaluating response construction and relevance. Additionally, adapting the approach to model a range of responses for each topic causes classification time to scale poorly with training data size and the number of questions.

To address these issues a general off-topic content detection framework based on topic adapted Recurrent Neural Network language models (RNNLM) has been developed and applied to off-topic response detection for spoken language assessment. This framework uses example responses to test questions in training of the language model and construction of the topic-space. The RNNLM learns to associate the example responses with points in the topic-space. Classification is done by ranking the topic-conditional posterior probabilities of a response. The advantage of this approach is that sequence information can be taken into account and broad ranges of responses can be associated with each topic without affecting classifcation speed. Two topic vector representations are investigated: Latent Dirichlet Allocation (LDA) (Blei et al., 2003; Griffiths and Steyvers, 2004) and Latent Semantic Analysis (LSA) (Landauer et al., 1998). They are compared to standard approaches on data from the Cambridge Business English (BULATS) exam.

The rest of this paper is structured as follows: Section 2 discusses the RNNLM adaptation and topic spaces; Section 3 discusses approaches to topic detection; Section 4 presents data sets and experimental infrastructure; Section 5 analyzes experimental results; Section 6 concludes the paper.

## 2 Topic Adapted RNNLMs

### 2.1 RNNLM Architecture

A statistical language model is used to model the semantic and syntactic information in text in the form of a probability distribution over word sequences. It assigns a probability to a word sequence $\mathbf{w} = \{w_0, w_1, \cdots, w_L\}$ as follows:

$$\mathtt{P}(w_i|w_{i-1}, \cdots, w_0) = \mathtt{P}(w_i|\mathbf{h}_0^{i-1}) \quad (1)$$

$$\mathtt{P}(\mathbf{w}) = \prod_{i=1}^{L} \mathtt{P}(w_i|\mathbf{h}_0^{i-1}) \quad (2)$$

where $w_0$ is the start of sentence symbol <s>. In this work a language model is trained to model example responses to questions on a spoken language assessment test. $\mathtt{P}(w_i|\mathbf{h}_0^{i-1})$ can be estimated by a number of approaches, most notably N-grams and Recurrent Neural Networks (Mikolov et al., 2010).

Recurrent Neural Network language models (RNNLMs) (Figure 1) (Mikolov, 2012) are a variable context length language model, capable of representing the entire context efficiently, unlike N-grams. RNNLMs represent the full untruncated history $\mathbf{h}_0^{i-1} = \{w_{i-1}, \cdots, w_0\}$ for word $w_i$ as the hidden layer $\mathbf{s}_{i-1}$, a form of short-term memory, whose representation is learned from the data. Words and phrases are represented in a continuous space, which gives RNNLMs greater generalization abilities than other language models, such as N-grams.

RNNLMs can be adapted by adding a feature vector $\mathbf{f}$ which represents information absent from the RNN (Mikolov and Zweig, 2012). In this work, the vector representation of a spoken language test question topic $\mathbf{f}_q$ is used for the context vector $\mathbf{f}$. Architecturally, a context adapted RNNLM is described by equations 3-5. $\mathtt{e}(x)$ and $\mathtt{g}(x)$ are element-wise sigmoid and softmax acti-

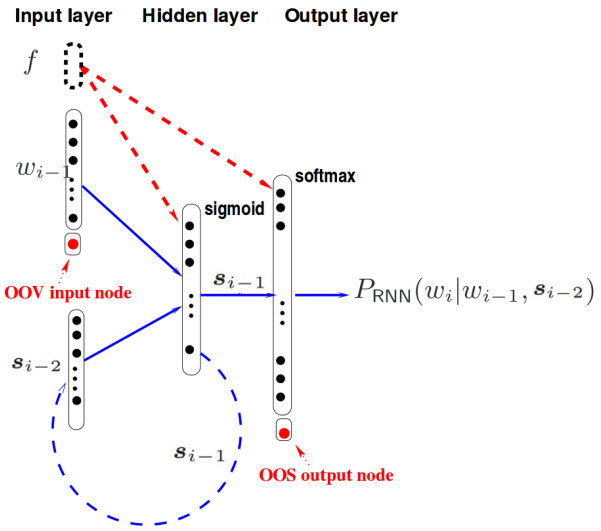


Figure 1: Context adapted RNN language model

vation functions.

$$\mathrm{P}(w_i|\mathbf{h}_0^{i-1}, \mathbf{f}) = \mathrm{P}_{\text{RNN}}(w_i|w_{i-1}, \mathbf{s}_{i-2}, \mathbf{f}) \quad (3)$$

$$\mathrm{P}_{\text{RNN}}(w_i|w_{i-1}, \mathbf{s}_{i-2}, \mathbf{f}) = \mathrm{g}(\mathbf{V}\mathbf{s}_{i-1} + \mathbf{H}\mathbf{f}) \quad (4)$$

$$\mathbf{s}_{i-1} = \mathrm{e}(\mathbf{U}w_{i-1} + \mathbf{W}\mathbf{s}_{i-2} + \mathbf{G}\mathbf{f}) \quad (5)$$

Through the process of adaptation the RNNLM learns to associate particular types of responses with particular topics, thereby becoming more discriminative. Thus, a sentence's topic-conditional probability $\mathrm{P}_{\text{RNN}}(\mathbf{w}|\mathbf{f}_q)$ will be higher if it corresponds to the topic $q$ than if it does not.

### 2.2 Example Response Based Topic Space

In order for the topic vectors $\mathbf{f}_q$ to be informative they must span the space of all question topics in the test. Thus a topic space needs to be defined. Example responses, which are necessary to train the RNNLM, are used to define a topic space because typical responses to a question will be definitive of the question's topic. Multiple example responses to a particular question are merged into one aggregate response to capture a broad range of response variations and increase the robustness of the vector representation estimation.

By default a topic $t$ is defined for each question $q$. However, multi-part questions are common, where candidates are given a scenario such as providing tourist information in which individual questions ask about food, hotels or sights. Since the underlying topic is related this can confuse a classifier. The responses for all these related questions could be merged to form a single aggregate vector, but the statistics of the responses to each question can be sufficiently different that less distinct topics are formed. Instead the aggregrate example responses for each question are assigned the same topic label. Thus, a mapping between questions and topics and its inverse is introduced:

$$\mathcal{M} : q \rightarrow t \quad (6)$$

$$\mathcal{M}_t^{-1} : \{q \in Q | \mathcal{M}(q) = t\} \quad (7)$$

A vector representation of a question topic is computed using the aggregate example responses. As mentioned in Section 1, two common representations are LDA and LSA; both are investigated in this work.

LDA is a generative model which allows documents to be modelled as distributions over latent topics $z \in Z$. Each latent topic $z$ is described by a multinomial distribution over words $\mathrm{P}(w_i|z)$, and each word in a document is attributed to a particular latent topic (Blei et al., 2003). Thus, the adaptation vector $\mathbf{f}_\mathbf{w}$ represents a vector of posterior probabilities over latent topics for word sequence $\mathbf{w}$:

$$\mathbf{f}_\mathbf{w} = [\mathrm{P}(z = 1|\mathbf{w}), \cdots, \mathrm{P}(z = K|\mathbf{w})]^T \quad (8)$$

$$\mathrm{P}(z = k|\mathbf{w}) = \frac{\sum_{i=1}^{N} \delta(z_{w_i} = k)}{N} \quad (9)$$

LDA was found to perform better for RNNLM adaptation than other representations in (Mikolov and Zweig, 2012; Chen et al., 2015).

LSA (Landauer et al., 1998) is a popular representation for information retrieval tasks. A word-document matrix $\mathbf{F}$ is constructed using example responses and each word is weighted by its term frequency-inverse document frequency (TF-IDF). Then a low-rank approximation $\mathbf{F}_k$ is computed using Singular Value Decomposition (SVD):

$$\mathbf{F}_k = \mathbf{U}_k \mathbf{\Sigma}_k \mathbf{V}_k^T \quad (10)$$

$$\mathbf{F}_k = [\mathbf{f}_1, \cdots, \mathbf{f}_Q]^T \quad (11)$$

$$\mathbf{f}_\mathbf{w} = \Sigma_k^{-1} \mathbf{U}_k^T \mathbf{f}_{tfidf} \quad (12)$$

Only the k largest singular values of the singular value matrix $\Sigma$ are kept. $\mathbf{F}_k$ is a representation of the data which retains only the k most significant factors of variation. Each row is a topic vector $\mathbf{f}_q$. New vectors $\mathbf{f}_\mathbf{w}$ for test responses can be projected into the LSA space via equation 12, where $\mathbf{f}_{tfidf}$ is the TF-IDF weighted bag-of-word representation of a test response.

## 3 Topic Detection and Features

This section discusses the standard, vector similarity feature-based, and the proposed topic adapted RNNLM approaches for topic detection.

### 3.1 Vector Similarity Features

The essence of the standard approach to topic detection is to assess the semantic distance $\mathtt{D}_{\mathtt{sem}}$ between the test response $\mathbf{w}$ and the aggregate example response $\mathbf{w}_q$ by approximating it using a vector distance metric $\mathtt{D}_{\mathtt{vec}}$ between vector representations of the response $\mathbf{f}_{\mathbf{w}}$ and the topic $\mathbf{f}_q$. Classification is done by selecting the topic closest to the test response:

$$\hat{t}_{\mathbf{w}} = \mathcal{M}\big( \arg\min_q \{\mathtt{D}_{\mathtt{sem}}(\mathbf{w}, \mathbf{w}_q)\}\big) \quad (13)$$

$$\mathtt{D}_{\mathtt{sem}}(\mathbf{w}, \mathbf{w}_q) \approx \mathtt{D}_{\mathtt{vec}}(\mathbf{f}_{\mathbf{w}}, \mathbf{f}_q) \quad (14)$$

The selection of an appropriate distance metric $\mathtt{D}_{\mathtt{vec}}(\mathbf{f}_{\mathbf{w}}, \mathbf{f}_q)$ can have a large effect on the classification outcome. A common metric used in topic classification and information retrieval is cosine similarity, which measures the cosine of the angle between two vectors. A distance metric based on this, *cosine distance*, can be defined as:

$$\mathtt{D}_{\mathtt{cos}}(\mathbf{f}_{\mathbf{w}}, \mathbf{f}_q) = 1 - \frac{\mathbf{f}_{\mathbf{w}}{}^{\mathsf{T}} \mathbf{f}_q}{|\mathbf{f}_{\mathbf{w}}||\mathbf{f}_q|} \quad (15)$$

While topics are robustly defined, this approach fails to capture the range of responses which can be given to a question. A different approach would be to maintain a separate point in this topic space for every example response. This retains the robust topic definition while allowing each topic to be represented by a cloud of points in topic space, thereby capturing a range of responses which can be given. A K-nearest neighbour (KNN) classifier can be used to detect the response topic by computing distances of the test response to each of the training points in topic space. However, classification may become impractical for large data sets, as the number of response points scales with the size of the training data. Low-cost distance measures, such as cosine distance, allow this approach to be used on large data sets before it becomes computationally infeasible. This approach is used as the baseline for comparison with the proposed RNNLM based method. For multi-part questions, topic vectors relating to the same overall topic are simply given the same topic label.

The classification rate can be improved by taking the top $N$ $\hat{\mathbf{t}}_N = \{\hat{t}_1, \cdots, \hat{t}_N\}$ results into account. The KNN classifier can be modified to yield the N-best classification by removing all training points from the 1-best class from the KNN classifier and re-running the classification to get the 2-best results, and so on.

One of the main deficiencies of methods based on computing distances between vector representations is that commonly used representations, such as LSA and LDA, ignore word-order in documents, thereby throwing away information which is potentially useful for topic classification. Additionally, if any of the test or example response utterances are short, then their topic vector representations may not be robustly estimated.

### 3.2 Topic Adapted RNNLM Framework

The RNNLM based approach to topic detection is based on different principles. By combining equations 2 and 3 the log-probability $\mathtt{L}(q)$ of a response sentence given a particular topic vector $\mathtt{P}_{\mathtt{RNN}}(\mathbf{w}|\mathbf{f}_q)$ is computed. For each response $\mathbf{w}$ in the test set $\mathtt{L}(q)$ is computed (equation 16) for all topic vectors $\mathbf{f}_q$. $\mathtt{L}(q)$ is calculated using equation 17 for multi-part questions with responses $\mathbf{w}_p$ where $p \in t$. Classification is done by ranking log-probability $\mathtt{L}(q)$ for an utterance $\mathbf{w}$ and $\mathtt{L}(q)$ for all $q \in M_t^{-1}$ are averaged (equation 18).

$$\mathtt{L}(q) = \begin{cases} \log[\mathtt{P}_{\mathtt{RNN}}(\mathbf{w}|\mathbf{f}_q)] & (16) \\ \sum_p \frac{1}{N_p} \log[\mathtt{P}_{\mathtt{RNN}}(\mathbf{w}_p|\mathbf{f}_q)] & (17) \end{cases}$$

$$\hat{t}_{\mathbf{w}} = \arg\max_t \{ \frac{1}{|\mathcal{M}_t^{-1}|} \sum_{q \in M_t^{-1}} \mathtt{L}(q)\} \quad (18)$$

It is trivial to extend this approach to yield the N-best solutions by simply taking the top $N$ outputs of equation 18.

The RNNLM approach has several benefits over standard approaches. Firstly, this approach explicitly takes account of word-order in determining the topical similarity of the response. Secondly, there is no need to explicitly select a distance metric. Thirdly, the problems of robustly estimating a vector representation $\mathbf{f}_{\mathbf{w}}$ of the test response are sidestepped. Furthermore, the RNNLM accounts for a broad range of responses because it is trained on individual response utterances which it associates with a question topic vector. This makes

it more scalable than the KNN approach because the number of comparisons which need to be made scales only with the number of questions, not the size of the training data. Thus, arbitrarily large data sets can be used to train the model without affecting classification time.

The RNNLM could be used in a KNN-style approach, where it associates each example response with its *individual* topic vector, using $\mathtt{L}(q)$ as a distance metric. However, this is computationally infeasible since computing $\mathtt{L}(q)$ is significantly more expensive than cosine distance and the previously mentioned scalability would be lost.

## 4 Data and Experimental Set-up

Data from the Business Language Testing Service (BULATS) English tests is used for training and testing. At test time, each response is recognised using an ASR system and the 1-best hypothesis is passed to the topic classifier. The topic detection system decides whether the candidate has spoken off topic by comparing the classifier output to the topic of the question being answered.

### 4.1 BULATS Test Format and Data

The BULATS Online Speaking Test has five sections (Chambers and Ingham, 2011):

A Candidates respond to eight simple questions about themselves and their work (e.g. what is your name, where do you come from?).

B Candidates read aloud six short texts appropriate to a work environment.

C Candidates talk about a work-related topic (e.g. the perfect office) with the help of prompts which appear on the screen.

D Candidates must describe a graph or chart such as a pie or a bar chart related to a business situation (e.g. company exports).

E Candidates are asked to respond to five open-ended questions related to a single context prompt. For example a set of five questions about organizing a stall at a trade fair.

Candidates are given a test consisting of 21 questions, however, only the last three sections, consisting of 7 questions, are spontaneously constructed responses to open ended question, and therefore of relevance to this work. Each unique set of 7 questions is a *question script*.

Training, development and evaluation data sets composed of predominantly Gujarati L1 candidates are used in these experiments. The data sets are designed to have an (approximately) even distribution over grades as well as over the different *question scripts*.

During operation the system will detect off-topic responses based on ASR transcriptions, so for the system to be matched it needs to be trained on ASR transcriptions as well. Thus, two training sets are made by using the ASR architecture described in section 4.2 to transcribe candidate responses. Each training set covers the same set of 282 unique topics. The first training set consists of data from 490 candidates, containing 9.9K responses, with an average of 35.1 responses per topic. The second, much larger, training set consists of data from 10004 candidates, containing 202K responses, with an average of 715.5 responses per topic.

| Characteristic | Section | | | | |
|---|---|---|---|---|---|
| | A | B | C | D | E |
| # Unique Topics | 18 | 144 | 17 | 18 | 85 |
| # Questions/Section | 6 | 8 | 1 | 1 | 5 |
| Av. # Words/Resp. | 10 | 10 | 61 | 77 | 20 |

Table 1: Data Characteristics.

As Table 1 shows, the average response length varies across sections due to the nature of the sections. Shorter responses to questions are observed for sections A, B and E, with longer responses to C and D. Estimating topic representations for sections A, B and E questions based on individual responses would be problematic due to the short response lengths. However, by aggregating example responses across candidates, as described in section 2.2, the average length of responses in all sections is significantly longer, allowing the example-response topic space to be robustly defined.

Section E topics correspond to topics of sub-questions relating to an overall question, thus there are only 15 unique questions in section E. However, the sub-questions are sufficiently distinct to merit their own topic vectors. At classification time confusions between sub-questions of an overall section E question are not considered mistakes.

Held-out test sets are used for development, *DEV*, and evaluation, *EVAL*, composed of 84 and 223 candidates, respectively. ASR transcriptions are used for these test sets, as per the operating

scenario. A version of the DEV set with professionally produced transcriptions, *DEV REF*, is also used in training and development.

The publicly available Gibbs-LDA toolkit (Phan and Nguyen, 2007) is used to estimate LDA posterior topic vectors and the scikit-learn 17.0 toolkit (Pedregosa et al., 2011) to estimate LSA topic representations. The topic adapted RNNLM uses a 100-dimensional hidden layer. DEV REF is used as a validation set for early stopping to prevent over-fitting. The CUED RNNLM toolkit v0.1 (Chen et al., 2016) is used for RNNLM training, details of which can be found in (Chen et al., 2014; Mikolov et al., 2010)

### 4.2 ASR System

A hybrid deep neural network DNN-HMM system is used for ASR (Wang et al., 2015). The acoustic models are trained on 108.6 hours of BULATS test data (Gujarati L1 candidates) using an extended version of the HTK v3.4.1 toolkit (Young et al., 2009; Zhang and Woodland, 2015). A Kneser-Ney trigram LM is trained on 186K words of BULATs test data and interpolated with a general English LM trained on a large broadcast news corpus, using the SRILM toolkit (Stolcke, 2002). Lattices are re-scored with an interpolated trigram+RNN LM (Mikolov et al., 2010) by applying the 4-gram history approximation described in (Liu et al., 2014), where the RNNLM is trained using the CUED RNNLM toolkit (Chen et al., 2016). Interpolation weights are optimized on the DEV REF data set. Table 2 shows the word error rate (WER) on the DEV test set relative to the DEV REF references for each section and the combined spontaneous speech sections (C-E).

| % WER | | | | | |
|---|---|---|---|---|---|
| A | B | C | D | E | C-E |
| 30.6 | 23.2 | 32.0 | 29.9 | 32.3 | 31.5 |

Table 2: ASR performance on DEV.

## 5 Experiments

Two forms of experiment are conducted in order to assess the performance of the topic-adapted RNNLM. First, a topic classification experiment is run where the ability of the system to accurately recognize the topic of a response is evaluated. Second, a closed-set off-topic response detection experiment is done.

In the experimental configuration used here a response is classified into a topic and the accuracy is measured. The topic of the question being answered is known and all responses are actually on-topic. A label (on-topic/off-topic) is given for each response based on the output of the classifier relative to the question topic. Thus, results presented are in terms of false rejection (FR) and false acceptance (FA) rates rather than precision and recall.

Initial topic detection experiments were run using the DEV test set with both the reference transcriptions (REF) and recognition hypotheses (ASR) to compare different KNN and RNN systems. After this, performance is evaluated on the EVAL test set. The systems were trained using data sets of 490 and 10004 candidates, as described in section 4.1.

### 5.1 Topic Classification

Performance of the topic-adapted RNNLM is compared to the KNN classifier in Table 3. The RNN1 system outperforms the KNN system by 20-35 % using the LDA topic representation. Furthermore, the KNN system performs worse on section E than it does on section C, while RNN1 performance is better on section E by 7-10% than on section C. The LSA topic representation consistently yields much better performance than LDA by 25-50% for both systems. Thus, the LDA representation is not further investigated in any experiments.

When using the LSA representation the RNN1 system outperforms the KNN system only marginally, due to better performance on section E. Additionally, unlike the KNN-LDA system, the KNN-LSA system does not have a performance degradation on section E relative to section C. Notably, the RNN1 system performs better on section E by 5-13% than on section C. Clearly, section C questions are hardest to assess. Combining both representations through concatenation does not effect performance of the KNN system and slightly degrades RNN1 performance on section C. KNN and RNN1 systems with either topic representation perform comparably on REF and ASR. This suggests that the systems are quite robust to WER rates of 31.5% and the differences are mostly noise.

Training the RNN2 system on 20 times as much data leads to greatly improved performance over KNN and RNN1 systems, almost halving the over-

| Topic Repn. | System | # Cands. | C | | D | | E | | ALL (C-E) | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | REF | ASR | REF | ASR | REF | ASR | REF | ASR |
| LDA | KNN | 490 | 75.0 | 81.0 | 37.0 | 42.0 | 91.8 | 91.1 | 68.0 | 71.4 |
| | RNN1 | | 61.9 | 58.3 | 28.4 | 25.9 | 48.8 | 51.2 | 46.6 | 45.4 |
| LSA | KNN | 490 | 32.1 | 28.6 | 2.5 | 3.7 | 31.3 | 33.3 | 22.0 | 21.9 |
| | RNN1 | | 29.8 | 31.0 | 4.9 | 6.2 | 23.8 | 23.8 | 19.7 | 20.5 |
| | RNN2 | 10004 | 19.0 | 19.0 | 3.7 | 3.7 | 9.5 | 10.7 | 10.8 | 11.2 |
| LDA +LSA | KNN | 490 | 30.9 | 29.8 | 2.5 | 3.7 | 31.5 | 33.3 | 21.7 | 22.3 |
| | RNN1 | | 32.1 | 35.7 | 4.9 | 4.9 | 23.8 | 22.6 | 20.5 | 21.3 |
| | RNN2 | 10004 | 25.0 | 22.6 | 4.9 | 4.9 | 10.7 | 10.7 | 13.7 | 12.9 |

Table 3: % False rejection in topic detection using KNN classifier with 6 nearest neighbour and distance weights and RNNLM classifier on the DEV test set. 280 dim. topic spaces for LDA and LSA, and 560 dim. for LDA+LSA.

all error rate. The KNN system could not be evaluated effectively in reasonable time using 20 times as many example responses and results are not shown, while RNN2 evaluation times are unaffected. Notably, RNN performance using the LSA representation scales with training data size better than with the LDA+LSA representation. Thus, we further investigate systems only with the LSA representation. Interestingly, section D performance is improved only marginally.

Performance on section D is always best, as section D questions relate to discussion of charts and graphs for different conditions for which the vocabulary is very specific. Section C and E questions are the less distinct because they have free-form answers to broad questions, leading to higher response variability. This makes the linking of topic from the training data to the test data more challenging, particularly for 1-best classification, leading to higher error rates.

Figure 2 shows the topic classification confusion matrix for the RNN1 LSA system. A similar matrix is observed for the KNN LSA system. Most confusions are with topics from the same section. This is because each section has a distinct style of questions and some questions within a section are similar. An example is shown below. Question SC-EX1 relates to personal local events in the workplace. SC-EX2, which relates to similar issues, is often confused with it. On the other hand, SC-EX3 is rarely confused with SC-EX1 as it is about non-personal events on a larger scale.

- SC-EX1: Talk about some advice from a colleague. You should say: what the advice was, how it helped you and whether you would give the same advice to another colleague.
- SC-EX2: Talk about a socially challenging day you had at work. You should say: what was the challenging situation, how you resolved it and why you found it challenging.
- SC-EX3: Talk about a company in your local town which you admire. You should say: what company it is, what they do, why you admire them, and how the company impacts life in your town.

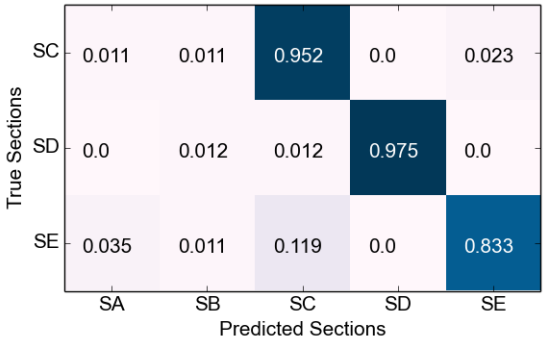


Figure 2: RNN1 LSA confusion matrix on DEV ASR.

System performance can be increased by considering $N$-best results, as described in Section 3. Results for systems trained on 490 and 10004 candidates are presented in Table 4. The error rate decreases as the value of $N$ increases for all systems. However, performance scales better with $N$ for the RNN systems than for KNN. Notably, for values of $N > 1$ performance of all systems on REF is better, which suggests that ASR errors do have a minor impact on system performance.

### 5.2 Off-topic response detection

In the second experiment off-topic response detection is investigated. Performance is measured

| N | System | # Cands. | REF | ASR |
|---|---|---|---|---|
| 1 | KNN | 490 | 22.1 | 21.9 |
| | RNN1 | | 19.7 | 20.5 |
| | RNN2 | 10004 | 10.8 | 11.2 |
| 2 | KNN | 490 | 15.9 | 16.0 |
| | RNN1 | | 13.7 | 16.1 |
| | RNN2 | 10004 | 6.8 | 7.6 |
| 3 | KNN | 490 | 13.5 | 14.3 |
| | RNN1 | | 10.4 | 11.2 |
| | RNN2 | 10004 | 6.4 | 7.2 |
| 4 | KNN | 490 | 11.1 | 12.5 |
| | RNN1 | | 8.8 | 10.0 |
| | RNN2 | 10004 | 5.2 | 6.4 |

Table 4: $N$-Best % false rejection performance of KNN and RNNLM classifiers with the LSA topic space on the DEV test set

in terms of the false acceptance (FA) probability of an off-topic response and false rejection (FR) probability of an on-topic response. The experiment is run on DEV and EVAL test sets. Since neither DEV nor EVAL contain real off-topic responses, a pool $\mathbf{W}_q$ of such responses is synthetically generated for each question by using valid responses to other questions in the data set. Off-topic responses are then selected from this pool. A selection strategy defines which responses are present in $\mathbf{W}_q$. Rather than using a single selection of off-topic responses, an expected performance over all possible off-topic response selections is estimated. The overall probability of falsely accepting an off-topic response can be expressed using equation 19.

$$\mathtt{P}(\mathtt{FA}) = \sum_{q=1}^{Q} \sum_{\mathbf{w} \in \mathbf{W}_q} \mathtt{P}(\mathtt{FA}|\mathbf{w}, q)\mathtt{P}(\mathbf{w}|q)\mathtt{P}(q) \quad (19)$$

In equation 19, the question $q$ is selected with uniform probability from the set $Q$ of possible questions. The candidate randomly selects with uniform probability $\mathtt{P}(\mathbf{w}|q)$ a response $\mathbf{w}$ from the pool $\mathbf{W}_q$. The correct response to the question is not present in the pool. The conditional probability of false accept $\mathtt{P}(\mathtt{FA}|\mathbf{w}, q) = 1$ if $\mathcal{M}(q) \in \hat{\mathbf{t}}_N$, and $\mathcal{M}(q)$ is not the real topic of the response $\mathbf{w}$, otherwise $\mathtt{P}(\mathtt{FA}|\mathbf{w}, q) = 0$.

As shown in Figure 2, the main confusions will occur if the response is from the same section as the question. Two strategies for selecting off-topic responses are considered based on this: `naive`, where an incorrect response can be selected from any section; and `directed`, where an incorrect response can only be selected from the same section as the question. The `naive` strategy represents candidates who have little knowledge of the system and memorise responses unrelated to the test, while the `directed` strategy represents those who are familiar with the test system and have access to real responses from previous tests.

| Test Set | System | % Equal Error Rate | |
|---|---|---|---|
| | | Directed | Naive |
| DEV | KNN | 13.5 | 10.0 |
| | RNN1 | 10.0 | 7.5 |
| | RNN2 | 7.5 | 6.0 |
| EVAL | KNN | 12.5 | 9.0 |
| | RNN1 | 8.0 | 6.0 |
| | RNN2 | 5.0 | 4.5 |

Table 5: % Equal Error Rate for LSA topic space systems on the DEV and EVAL test sets.

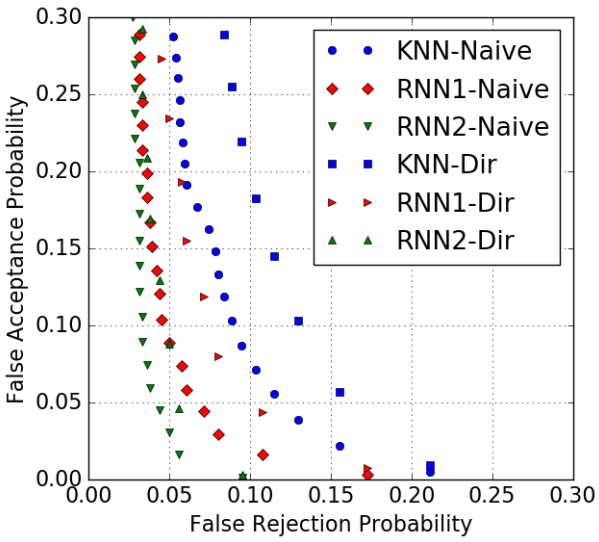


Figure 3: ROC curves of LSA topic space systems on the EVAL test set.

A Receiver Operating Characteristic (ROC) curve (Figure 3) can be constructed by plotting the FA and FR rates for a range of $N$. The RNN1 system performs better at all operating points than the KNN system for both selection strategies and evaluation test sets. Equal Error Rates (EER), where FA = FR, are given in Table 5. Results on EVAL are more representative of the difference between the KNN and RNN performance, as they are evaluated on nearly 3 times as many candidates. The

RNN2 system achieves the lowest EER. It is interesting that for better systems the difference in performance against the `naive` and `directed` strategies decreases. This indicates that the systems become increasingly better at discriminating between similar questions.

As expected, the equal error rate for the `directed` strategy is higher than for the `naive` strategy. In relation to the stated task of detecting when a test candidate is giving a memorized response, the `naive` strategy represents a lower-bound on realistic system performance, as students are not likely to respond with a valid response to a different question. Most likely they will fail to construct a valid response or will add completely unrelated phrases memorised beforehand, which, unlike responses from other sections, may not come from the same domain as the test (eg: Business for BULATs).

## 6 Conclusion and Future Work

In this work a novel off-topic content detection framework based on topic-adapted RNNLMs was developed. The system was evaluated on the task of detecting off-topic spoken responses on the BULATS test. The proposed approach achieves better topic classification and off-topic detection performance than the standard approaches.

A limitation of both the standard and proposed approach is that if a new question is created by the test-makers, then it will be necessary to collect example responses before it can be widely deployed. However, since the system can be trained on ASR transcriptions, the example responses do not need to be hand-transcribed. This is an attractive deployment scenario, as only a smaller hand-transcribed data set is needed to train an ASR system with which to cost-effectively transcribe a large number of candidate recordings.

Further exploration of different topic vector representations and their combinations is necessary in future work.

## Acknowledgements

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

## References

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, March.

Lucy Chambers and Kate Ingham. 2011. The BULATS online speaking test. *Research Notes*, 43:21–25.

Xie Chen, Yongqiang Wang, Xunying Liu, Mark J.F. Gales, and P.C. Woodland. 2014. Efficient GPU-based Training of Recurrent Neural Network Language Models Using Spliced Sentence Bunch. In *Proc. INTERSPEECH*.

Xie Chen, Tian Tan, Xunying Liu, Pierre Lanchantin, Moquan Wan, Mark J.F. Gales, and Philip C. Woodland. 2015. Recurrent Neural Network Language Model Adaptation for Multi-Genre Broadcast Speech Recognition. In *Proc. INTERSPEECH*.

X. Chen, X. Liu, Y. Qian, M.J.F. Gales, and P.C. Woodland. 2016. CUED-RNNLM – an open-source toolkit for efficient training and evaluation of recurrent neural network language models. In *Proc. ICASSP*.

Keelan Evanini and Xinhao Wang. 2014. Automatic detection of plagiarized spoken responses. In *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications*.

Keelan Evanini, Shasha Xie, and Klaus Zechner. 2013. Prompt-based Content Scoring for Automated Spoken Language Assessment. In *Proc. NAACL-HLT*.

Thomas L. Griffiths and Mark Steyvers. 2004. Finding Scientific Topics. *Proceedings of the National Academy of Sciences*, 101:5228–5235.

Thomas K Landauer, Peter W. Foltz, and Darrell Laham. 1998. Introduction to Latent Semantic Analysis. *Discourse Processes*, 25:259–284.

Xunying Liu, Y. Wang, Xie Chen, Mark J.F. Gales, and Philip C. Woodland. 2014. Efficient Lattice Rescoring using Recurrent Neural Network Language Models. In *Proc. INTERSPEECH*.

Angeliki Metallinou and Jian Cheng. 2014. Using Deep Neural Networks to Improve Proficiency Assessment for Children English Language Learners. In *Proc. INTERSPEECH*.

Tomas Mikolov and Geoffrey Zweig. 2012. Context Dependent Recurrent Neural Network Language Model. In *Proc. IEEE Spoken Language Technology Workshop (SLT)*.

Tomas Mikolov, Martin Karafiát, Lukás Burget, Jan Cernocký, and Sanjeev Khudanpur. 2010. Recurrent Neural Network Based Language Model. In *Proc. INTERSPEECH*.

Tomas Mikolov. 2012. *Statistical Language Models Based on Neural Networks*. Ph.D. thesis, Brno University of Technology.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Xuan-Hieu Phan and Cam-Tu Nguyen. 2007. GibbsLDA++: A C/C++ implementation of latent Dirichlet allocation (LDA). `http://gibbslda.sourceforge.net/`.

Barbara Seidlhofer. 2005. English as a lingua franca. *ELT journal*, 59(4):339.

A Stolcke. 2002. SRILM – an extensible language modelling toolkit. In *Proc. ICSLP*.

Haipeng Wang, Anton Ragni, Mark J. F. Gales, Kate M. Knill, Philip C. Woodland, and Chao Zhang. 2015. Joint Decoding of Tandem and Hybrid Systems for Improved Keyword Spotting on Low Resource Languages. In *Proc. INTERSPEECH*.

Shasha Xie, Keelan Evanini, and Klaus Zechner. 2012. Exploring Content Features for Automated Speech Scoring. In *Proc. NAACL-HLT*.

Helen Yannakoudakis. 2013. Automated assessment of English-learner writing. Technical Report UCAM-CL-TR-842, University of Cambridge Computer Laboratory.

Su-Youn Yoon and Shasha Xie. 2014. Similarity-Based Non-Scorable Response Detection for Automated Speech Scoring. In *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications*.

Steve Young, Gunnar Evermann, Mark J. F. Gales, Thomas Hain, Dan Kershaw, Xunying (Andrew) Liu, Gareth Moore, Julian Odell, Dave Ollason, Dan Povey, Valtcho Valtchev, and Phil Woodland. 2009. *The* HTK *book (for* HTK *Version 3.4.1)*. University of Cambridge.

Klaus Zechner, Derrick Higgins, Xiaoming Xi, and David M. Williamson. 2009. Automatic scoring of non-native spontaneous speech in tests of spoken English. *Speech Communication*, 51(10):883–895. Spoken Language Technology for Education Spoken Language.

Chau Zhang and Philip C. Woodland. 2015. A General Artificial Neural Network Extension for HTK. In *Proc. INTERSPEECH*.