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.

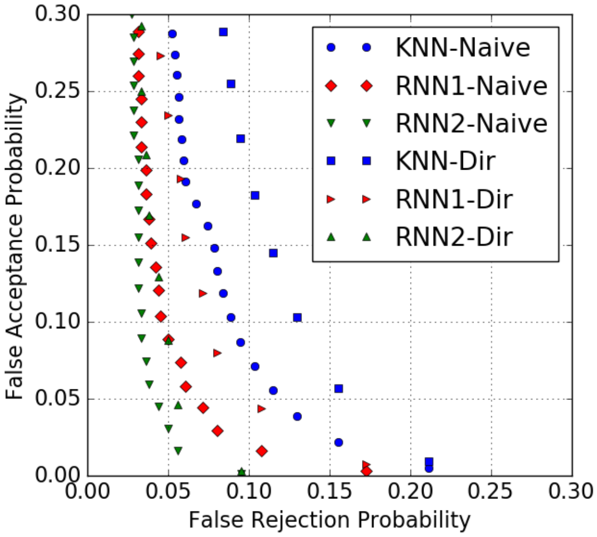

Figure 3: ROC curves of LSA topic space systems on the EVAL test set.

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