Similarity-Based Non-Scorable Response Detection for Automated Speech Scoring

Su-Youn Yoon
Educational Testing Service
Princeton, NJ, USA
syoon@ets.org

Shasha Xie
Microsoft
Sunnyvale, CA, USA
shxie@microsoft.com

Abstract

This study provides a method that identifies problematic responses which make automated speech scoring difficult. When automated scoring is used in the context of a high stakes language proficiency assessment, for which the scores are used to make consequential decisions, some test takers may have an incentive to try to game the system in order to artificially inflate their scores. Since many automated proficiency scoring systems use fluency features such as speaking rate as one of the important features, students may engage in strategies designed to manipulate their speaking rate as measured by the system.

In order to address this issue, we developed a method which filters out non-scorable responses based on text similarity measures. Given a test response, the method generated a set of features which calculated the topic similarity with the prompt question or the sample responses including relevant content. Next, an automated filter which identified these problematic responses was implemented using the similarity features. This filter improved the performance of the baseline filter in identifying responses with topic problems.

1 Introduction

In spoken language proficiency assessment, some responses may include sub-optimal characteristics which make it difficult for the automated scoring system to provide a valid score. For instance, some test takers may try to game the system by speaking in their native languages or by citing memorized responses for unrelated topics. Others may repeat questions or part of questions with modifications instead of generating his/her own response. Hereafter, we call these problematic responses non-scorable (NS) responses. By using these strategies, test takers can generate fluent speech, and the automated proficiency scoring system, which utilizes fluency as one of the important factors, may assign a high score. In order to address this issue, the automated proficiency scoring system in this study used a two-step approach: these problematic responses were filtered out by a "filtering model," and only the remaining responses were scored using the automated scoring model. By filtering out these responses, the robustness of the automated scoring system can be improved.

The proportion of NS responses, in the assessment of which the responses are scored by human raters, are likely to be low. For instance, the proportion of NS responses in the international English language assessment used in this study was 2%. Despite this low proportion, it is a serious problem which has a strong impact on the validity of the test. In addition, the likelihood of students engaging in gaming strategies may increase with the use of automated scoring. Therefore, an automated filtering model with a high accuracy is a necessary step to use the automated scoring system as a sole rater.

Both off-topic and copy responses have topic-related problems, although they are at the two extremes in the degree of similarity. Focusing on the intermediate levels of similarity, Metzler et al. (2005) presented a hierarchy of five similarity levels: unrelated, on the general topic, on the specific topic, same facts, and copied. In the automated scoring of spontaneous speech, responses that fell into *unrelated* can be considered as off-topic, while the ones that fell into *copied* can be considered as repetition or plagiarism. Following this approach, we developed a non-scorable response identification method utilizing similar-

Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications , pages 116–123,
Baltimore, Maryland USA, June 26, 2014.

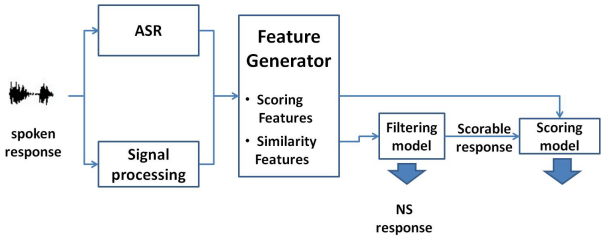

Figure 1: A diagram of the overall architecture of our method.

ity measures. We will show that this similarity based method is highly efficient in identifying off-topic or repetition responses. Furthermore, we will show that the method can effectively detect NS responses that are not directly related to the topicality issue (e.g, non-English responses).

Figure 1 shows the overall architecture of our method including the automated speech proficiency scoring system. For a given spoken response, the system performs speech processing including speech recognition and generates a word hypotheses and time stamps. In addition, the system computes pitch and power; the system calculates descriptive statistics such as the mean and standard deviation of pitch and power at both the word level and response level. Given the word hypotheses and descriptive features of pitch/power, it derives features for automated proficiency scoring. In addition, the similarity features are generated based on the word hypotheses and topic models. Finally, given both sets of features, the filtering model filters out non-scorable responses, and the remainder of the responses are scored using a scoring model. A detailed description of the system is available from Zechner et al. (2009). In this study, we will only focus on the filtering model.

This paper will proceed as follows: we first review previous studies in section 2, then describe the data in section 3, and present the method and experiment set-up in sections 4 and 5. The results and discussion are presented in section 6, and the conclusions are presented in section 7.

2 Related Work

Filtering of NS responses for automated speech scoring has been rarely recognized. Only a few pieces of research have focused on this task, and most studies have targeted highly restricted speech. van Doremalen et al. (2009) and Lo et al. (2010) used normalized confidence scores of a speech recognizer in recasting speech. They identified non-scorable responses with promising performances (equal error rates ranged from 10 to 20%). Cheng and Shen (2011) extended these studies and combined an acoustic model score, a language model score, and a garbage model score with confidence scores. They applied this new filter to less constrained items (e.g., picture description) and identified off-topic responses with an accuracy rate of 90% with a false positive rate of 5%.

Although normalized confidence scores achieved promising performances in restricted speech, they may not be appropriate for the items that elicit unconstrained spontaneous speech. Low confidence scores signal the use of words or phrases not covered by the language model (LM) and this is strongly associated with off-topic responses in restricted speech in which the target sentence is given. However, in spontaneous speech, this is not trivial; it may be associated with not only off-topic speech but also mismatch between the LM and speech input due to the low coverage of the LM. Due to the latter case, the decision based on the confidence score may not be effective in measuring topic similarity.

The topic similarity between two documents has been frequently modeled by relative-frequency measures (Hoad and Zobel, 2003; Shivakumar and Garcia-Molina, 1995), document fingerprinting (Brin et al., 1995; Shivakumar and Garcia-Molina, 1995; Shivakumar and Garcia-Molina, 1996)), and query based information retrieval methods using vector space models or language model (Sanderson, 1997; Hoad and Zobel, 2003).

Document similarity measures have been applied in automated scoring. Foltz et al. (1999) evaluated the content of written essays using latent semantic analysis (LSA) by comparing the test essays with essays of known quality in regard to their degree of conceptual relevance and the amount of relevant content. In another approach, the lexical content of an essay was evaluated by comparing the words contained in each essay to the words found in a sample of essays from each score category (Attali and Burstein, 2006). More recently, Xie et al. (2012) used a similar approach in automated speech scoring; they measured the similarity using three similarity measures, including a lexical matching method (Vector Space Model) and two semantic similarity measures (Latent Semantic Analysis and Pointwise Mutual Information). They showed moderately high correlations

between the similarity features and human proficiency scores on even the output of an automatic speech recognition system. Similarity measures have also been used in off-topic detection for non-native speakers' essays. Higgins et al. (2006) calculated overlaps between the question and content words from the essay and obtained an error rate of 10%.

Given the promising performance in both automated scoring and off-topic essay detection, we will expand these similarity measures in NS response detection for speech scoring.

3 Data

In this study, we used a collection of responses from an international English language assessment. The assessment was composed of items in which speakers were prompted to provide spontaneous speech.

Approximately 48,000 responses from 8,000 non-native speakers were collected and used for training the automated speech recognizer (ASR set). Among 24 items in the ASR set, four items were randomly selected. For these items, a total of 11,560 responses were collected and used for the training and evaluation of filtering model (FM set). Due to the extremely skewed distribution of NS responses (2% in the ASR set), it was not easy to train and evaluate the filtering model. In order to address this issue, we modified the distribution of NS responses in the FM set. Initially, we collected 90,000 responses including 1,560 NS responses. While maintaining all NS responses, we downsampled the scorable responses in the FM set to include 10,000 responses. Finally, the proportion of NS responses was 6 times higher in FM set (13%) than ASR set. This artificial increase of the NS responses reduces the current problem of the skewed NS distribution and may make the task easier. However, the likelihood of students engaging in gaming strategies may increase with the use of automated scoring, and this increased NS distribution may be close to this situation.

Each response was rated by trained human raters using a 4-point scoring scale, where 1 indicated a low speaking proficiency and 4 indicated a high speaking proficiency. The raters also labeled responses as NS, when appropriate. NS responses are defined as responses that cannot be given a score according to the rubrics of the four-point scale. NS responses were responses with technical difficulties (TDs) that obscured the content of the responses or responses that would receive a score of 0 due to participants' inappropriate behaviors. The speakers, item information, and distribution of proficiency scores are presented in Table 1. There was no overlap in the sets of speakers in the ASR and FM sets.

In addition, 1,560 NS responses from the FM set were further classified into six types by two raters with backgrounds in linguistics using the rubrics presented in Table 2. This annotation was used for the purpose of analysis: to identify the frequent types of NS responses and prioritize the research effort.

Type	Proportion in total NSs	Description
NR	73%	No response. Test taker doesn't speak.
OR	16%	Off-topic responses. The response is not related to the prompt.
TR	5%	Generic responses. The response only include filler words or generic responses such as, "I don't know, it is too difficult to answer, well", etc.
RE	4%	Question copy. Full or partial repetition of question.
NE	1%	Non-English. Responses is in a language other than English.
OT	1%	Others

Table 2: Types of zero responses and proportions

Some responses belonged to more than one type, and this increased complexity of the annotation task. For instance, one response was comprised of a question copy and generic sentences, while another response was comprised of a question copy and off-topic sentences. An example of this type was presented in Table 3. This was a response for the question "Talk about an interesting book that you read recently. Explain why it was interesting[1]."

For these responses, annotators first segmented them into sentences and assigned the type that was most dominant.

Each rater annotated approximately 1,000 responses, and 586 responses were rated by both

[1]In order to not reveal the real test question administered in the operational test, we invented this question. Based on the question, we also modified a sample response; the question copy part was changed to avoid disclosure of the test question, but the other part remained the same as the original response.

Data set	Num. responses	Num. speakers	Num. items	Average score	Score distribution				
					NS	1	2	3	4
ASR	48,000	8,000	24	2.63	773 2%	1953 4%	16834 35%	23106 48%	5334 11%
FM	11,560	11,390	4	2.15	1560 13%	734 6%	4328 37%	4263 37%	675 6%

Table 1: Data size and score distribution

Sentence	Type
Well in my opinion are the interesting books that I read recently is.	RE
Talking about a interesting book.	RE
One interesting book oh God interesting book that had read recently.	RE
Oh my God.	TR
I really don't know how to answer this question.	TR
Well I don't know.	TR
Sorry.	TR

Table 3: Manual transcription of complex-type response

raters. The Cohen's kappa between two raters was 0.76. Among five different NS responses, non-response was the most frequent type (73%), followed by off-topic (16%). The combination of the two types was approximately 90% of the entire NS responses.

4 Method

In this study, we generated two different types of features. First, we developed similarity features (both chunk-based and response-based) to identify the responses with problems in topicality. Secondly, we generated acoustic, fluency, and ASR-confidence features using a state-of-art automated speech scoring system. Finally, using both feature sets, classifiers were trained to make a binary distinction of NS response vs. scorable response.

4.1 Chunk-based similarity features

Some responses in this study included more than two different types of the topicality problems. For instance, the first three sentences in Table 3 belonged to the "copied" category, while the other sentences fell into "unrelated". If the similarity features were calculated based on the entire response, the feature values may fall into neither the "copied" nor "unrelated" range because of the trade-off between the two types at two extremes. In order to address this issue, we calculated chunk-based similarity features similar to Metzler et al. (2005)'s sentence-based features.

First, the response was split into the chunks which were surrounded by long silences with durations longer than 0.6 sec. For each chunk, the proportion of word overlap with the question (WOL) was calculated based on the formula (1). Next, chunks with a WOL higher than 0.5 were considered as ***question copies***.

$$WOL = \frac{|S \cap Q|}{|S|}$$
where S is a response and Q is a question,
$|S \cap Q|$ is the number of word types that appear both in S and Q,
$|S|$ is the number of word types in S
(1)

Finally, the following three features were derived for each response based on the chunk-based WOL.

- numwds: the number of word tokens after removing question copies, fillers, and typical generic sentences[2];
- copyR: the proportion of question copies in the response in terms of number of word tokens;
- meanWOL: the mean of WOLs for all chunks in the response.

4.2 Response-based similarity features

We implemented three features based on a vector space model (VSM) using cosine similarity and term frequency-inverse document frequency (tf-idf) weighting to estimate the topic relevance at the response-level.

[2]Five sentences "it is too difficult", "thank you", "I don't know", "I am sorry", and "oh my God" were stored as typical sentences and removed from responses

Since the topics of each question were different from each other, we trained a VSM for each question separately. For the four items in the FM set, we selected a total of 485 responses (125 responses per item) from the ASR set for topic model training. Assuming that the responses with the highest proficiency scores contain the most diverse and appropriate words related to the topic, we only selected responses with a score of 4. We obtained the manual transcriptions of the responses, and all responses about the same question were converted into a single vector. In this study, the term was a unigram word, and the document was the response. idf was trained from the entire set of 48,000 responses in the ASR training partition, while tf was trained from the question-specific topic model training set.

In addition to the response-based VSM, we trained a question-based VSM. Each question was composed of two sentences. Each question was converted into a single vector, and a total of four VSMs were trained. idf was trained in the same way as the response-based VSMs, while tf was trained only using the question sentences.

Using these two different types of VSMs, the following three features were generated for each response.

- $sampleCosine$: a similarity score based on the response-based VSM. Assuming that two documents with the same topic shared common words, it measured the similarity in the words used in a test response and the sample responses. The feature was implemented to identify off-topic responses (OR);

- $qCosine$: a similarity score based on the question-based VSM. It measured the similarity between a test response and its question. The feature was implemented to identify both off-topic responses (OR) and question copy responses (RE); a low score is highly likely to be an off-topic response, while a high score signals a full or partial copy;

- $meanIDF$: mean of idfs for all word tokens in the response. Generic responses (TR) tend to include many high frequency words such as articles and pronouns, and the mean idf value of these responses may be low.

4.3 Features from the automated speech scoring system

A total of 61 features (hereafter, A/S features) were generated using a state-of-the-art automated speech scoring system. A detailed description of the system is available from (Jeon and Yoon, 2012). Among these features, many features were conceptually similar but based on different normalization methods, and they showed a strong inter-correlation. For this study, 30 features were selected and classified into three groups according to their characteristics: acoustic features, fluency features, and ASR-confidence features.

The acoustic features were related to power, pitch, and MFCC. First, power, pitch and MFCC were extracted at each frame using Praat (Boersma, 2002). Next, we generated response-level features from these frame-level features by calculating mean and variation. These features captured the overall distribution of energy and voiced regions in a speaker's response. These features are relevant since NS responses may have an abnormal distribution in energy. For instance, non-responses contain very low energy. In order to detect these abnormalities in the speech signal, pitch and power related features were calculated.

The fluency features measure the length of a response in terms of duration and number of words. In addition, this group contains features related to speaking rate and silences, such as mean duration and number of silences. In particular, these features are effective in identifying non-responses which contain zero or only a few words.

The ASR-confidence group contains features predicting the performance of the speech recognizer. Low confidence scores signal low speech recognition accuracy.

4.4 Model training

Three filtering models were trained to investigate the impact of each feature group: a filtering model using similarity features (hereafter, the Similarity-filter), a filtering model using A/S features (hereafter, the A/S-filter), and a filtering model using a combination of the two groups of features (hereafter, the Combined-filter).

5 Experiments

An HMM-based speech recognizer was trained using the ASR set. A gender independent triphone acoustic model and a combination of bigram, tri-

gram, and four-gram language models were used. A word error rate (WER) of 27% on the held-out test dataset was observed.

For each response in the FM set, the word hypotheses was generated using this recognizer. From this ASR-based transcription, the six similarity features were generated. In addition, the 30 A/S features described in 4.3 were generated.

Using these two sets of features, filtering models were trained using the Support Vector Machine algorithm (SVM) with the RBF kernel of the WEKA machine-learning toolkit (Hall et al., 2009). A 10 fold cross-validation was conducted using the FM dataset.

6 Results and discussion

First, we will report the performance for the subset only topic-related NS responses. The similarity features were designed to detect NS responses with topicality issues, but the majority in the FM set were non-response (73%). The topic-related NS responses (off-topic responses, generic responses, and question copy responses) were only 25%. In the entire set, the advantage of the similarity features over the A/S features might not be salient due to the high proportion of non-response. In order to investigate the performance of the similarity features in the topic related NS responses, we excluded all responses other than 'OR', 'TR', and 'RE' from the FM set and conducted a 10 fold cross-validation.

Table 4 presents the average of the 10 fold cross-validation results in this subset. In this set, the total number of NS responses is 314, and the accuracy of the majority voting (to classify all responses as scorable responses) is 0.962.

	acc.	prec.	recall	fscore
Similarity-filter	0.975	0.731	0.548	0.626
A/S-filter	0.971	0.767	0.341	0.472
Combined-filter	0.977	0.780	0.566	0.656

Table 4: Performance of filters in topic-related NS detection

Not surprisingly, the Similarity-filter outperformed the A/S-filter: the F-score was approximately 0.63 which was 0.15 higher than that of the A/S-filter in absolute value. The lack of features specialized for detection of topic abnormality resulted in the low recall of the A/S-filter. The combination of the two features achieved a slight improvement: the F-score was 0.66 and it was 0.03 higher than the Similarity-filter.

In Metzler et al. (2005)'s study, the system using both sentence-based features and document-based features did not achieve further improvement over the system based on the document-based features alone. In order to explore the impact of chunk-based features, similarity features were classified into two groups (chunk-based features vs. document-based features), and two filters were trained using each group separately. Table 5 compares the performance of the two filters (Similarity-chunk and Similarity-doc) with the filter using all similarity features (Similarity).

	acc.	prec.	recall	fscore
Similarity-chunk	0.972	0.700	0.442	0.542
Similarity-doc	0.971	0.730	0.396	0.514
Similarity	0.975	0.731	0.548	0.626

Table 5: Comparison of chunk-based and document-based similarity features

In this study, the chunk-based features were comparable to the document-based features. Furthermore, combination of the two features improved F-score. The performance improvement mostly resulted from higher recall.

Finally, Table 6 presents the results using the entire FM set, including the OR, TR, and RE responses that were not included in the previous experiment. The accuracy of the majority class baseline (classifying all responses as scorable responses) is 0.865.

	acc.	prec.	recall	fscore
Similarity-filter	0.976	0.926	0.895	0.910
A/S-filter	0.974	0.953	0.849	0.898
Combined-filter	0.977	0.941	0.884	0.911

Table 6: Performance of filters in all types of NS detection

Both the Similarity-filter and the A/S-filter achieved high performance. Both accuracies and F-scores were similar and the difference

between the two filters was approximately 0.01. The Similarity-filter achieved better performance than the A/S-filter in recall: it was 0.89, which was substantially higher than the A/S-filter (0.85).

It is an encouraging result that the Similarity-filter could achieve a performance comparable to the A/S-filter, which was based on multiple resources such as signal processing, forced-alignment, and ASR. But, the combination of the two feature groups did not achieve further improvement: the increase in both accuracy and F-measure was less than 0.01.

7 Conclusions

In this study, filtering models were implemented as a supplementary module for an automated speech proficiency scoring system. In addition to A/S features, which have shown promising performance in previous studies, a set of similarity features were implemented and a filtering model was developed. The Similarity-filter was more accurate than the A/S-filter in identifying the responses with topical problems. This result is encouraging since the proportion of these responses is likely to increase when the automated speech scoring system becomes a sole rater of the assessment.

Although the Similarity-filter achieved better performance than the A/S-filter, it should be further improved. The recall of the system was low, and approximately 45% of NS responses could not be identified. In addition, the model requires substantial amount of sample responses for each item, and it will cause serious difficulty when it is used the real test situation. In future, we will explore the similarity features trained only using the prompt question or the additional prompt materials such as visual and audio materials.

References

Yigal Attali and Jill Burstein. 2006. Automated essay scoring with e–rater R v.2. *The Journal of Technology, Learning, and Assessment*, 4(3).

Paul Boersma. 2002. Praat, a system for doing phonetics by computer. *Glot International*, 5(9/10):341–345.

Sergey Brin, James Davis, and Hector Garcia-Molina. 1995. Copy detection mechanisms for digital documents. In *ACM SIGMOD Record*, volume 24, pages 398–409. ACM.

Jian Cheng and Jianqiang Shen. 2011. Off-topic detection in automated speech assessment applications. In *Proceedings of InterSpeech*, pages 1597–1600. IEEE.

Peter W. Foltz, Darrell Laham, and Thomas K. Landauer. 1999. The Intelligent Essay Assessor: Applications to educational technology. *Interactive multimedia Electronic Journal of Computer-Enhanced Learning*, 1(2).

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. 2009. The weka data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18.

Derrick Higgins, Jill Burstein, and Yigal Attali. 2006. Identifying off-topic student essays without topic-specific training data. *Natural Language Engineering*, 12(02):145–159.

Timothy C Hoad and Justin Zobel. 2003. Methods for identifying versioned and plagiarized documents. *Journal of the American society for information science and technology*, 54(3):203–215.

Je Hun Jeon and Su-Youn Yoon. 2012. Acoustic feature-based non-scorable response detection for an automated speaking proficiency assessment. In *Proceedings of the InterSpeech*, pages 1275–1278.

Wai-Kit Lo, Alissa M Harrison, and Helen Meng. 2010. Statistical phone duration modeling to filter for intact utterances in a computer-assisted pronunciation training system. In *Proceedings of Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 5238–5241. IEEE.

Donald Metzler, Yaniv Bernstein, W Bruce Croft, Alistair Moffat, and Justin Zobel. 2005. Similarity measures for tracking information flow. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 517–524. ACM.

Mark Sanderson. 1997. Duplicate detection in the reuters collection. *" Technical Report (TR-1997-5) of the Department of Computing Science at the University of Glasgow G12 8QQ, UK"*.

Narayanan Shivakumar and Hector Garcia-Molina. 1995. Scam: A copy detection mechanism for digital documents.

Narayanan Shivakumar and Hector Garcia-Molina. 1996. Building a scalable and accurate copy detection mechanism. In *Proceedings of the first ACM international conference on Digital libraries*, pages 160–168. ACM.

Joost van Doremalen, Helmet Strik, and Cartia Cucchiarini. 2009. Utterance verification in language learning applications. In *Proceedings of the SLaTE*.

Shasha Xie, Keelan Evanini, and Klaus Zechner. 2012. Exploring content features for automated speech scoring. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 103–111. Association for Computational Linguistics.

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.