Prompt-based Content Scoring for Automated Spoken Language Assessment

Keelan Evanini
Educational Testing Service
Princeton, NJ 08541, USA
kevanini@ets.org

Shasha Xie
Microsoft
Sunnyvale, CA 94089
shxie@microsoft.com

Klaus Zechner
Educational Testing Service
Princeton, NJ 08541, USA
kzechner@ets.org

Abstract

This paper investigates the use of prompt-based content features for the automated assessment of spontaneous speech in a spoken language proficiency assessment. The results show that single highest performing prompt-based content feature measures the number of unique lexical types that overlap with the listening materials and are not contained in either the reading materials or a sample response, with a correlation of $r = 0.450$ with holistic proficiency scores provided by humans. Furthermore, linear regression scoring models that combine the proposed prompt-based content features with additional spoken language proficiency features are shown to achieve competitive performance with scoring models using content features based on pre-scored responses.

1 Introduction

A spoken language proficiency assessment should provide information about how well the non-native speaker will be able to perform a wide range of tasks in the target language. Therefore, in order to provide a full evaluation of the non-native speaker's speaking proficiency, the assessment should include some tasks eliciting unscripted, spontaneous speech. This goal, however, is hard to achieve in the context of a spoken language assessment which employs automated scoring, due to the difficulties in developing accurate automatic speech recognition (ASR) technology for non-native speech and in extracting valid and reliable features. Because of this, most spoken language proficiency assessments which use automated scoring have focused on restricted speech, and have included tasks such as reading a word / sentence / paragraph out loud, answering single-word factual questions, etc. (Chandel et al., 2007; Bernstein et al., 2010).

In order to address this need, some automated spoken language assessment systems have also included tasks which elicit spontaneous speech. However, these systems have focused primarily on a non-native speaker's pronunciation, prosody, and fluency in their scoring models (Zechner et al., 2009), since these types of features are relatively robust to ASR errors. Some recent studies have investigated the use of features related to a spoken response's content, such as (Xie et al., 2012). However, the approach to content scoring taken in that study requires a large amount of responses for each prompt to be provided with human scores in order to train the content models. This approach is not practical for a large-scale, high-stakes assessment which regularly introduces many new prompts into the assessment–obtaining the required number of scored training responses for each prompt would be quite expensive and could lead to potential security concerns for the assessment. Therefore, it would be desirable to develop an approach to content scoring which does not require a large amount of actual responses to train the models. In this paper, we propose such a method which uses the stimulus materials for each prompt contained in the assessment to evaluate the content in a spoken response.

Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications, pages 157–162,
Atlanta, Georgia, June 13 2013. ©2013 Association for Computational Linguistics

2 Related Work

There has been little prior work concerning automated content scoring for spontaneous spoken responses (a few recent studies include (Xie et al., 2012) and (Chen and Zechner, 2012)); however, several approaches have been investigated for written responses. A standard approach for extended written responses (e.g., essays) is to compare the content in a given essay to the content in essays that have been provided with scores by human raters using similarity methods such as Content Vector Analysis (Attali and Burstein, 2006) and Latent Semantic Analysis (Foltz et al., 1999). This method thus requires a relatively large set of pre-scored responses for each test question in order to train the content models. For shorter written responses (e.g., short answer questions targeting factual content) approaches have been developed that compare the similarity between the content in a given response and a model correct answer, and thus do not necessarily require the collection of pre-scored responses. These approaches range from fully unsupervised text-to-text similarity measures (Mohler and Mihalcea, 2009) to systems that incorporate hand-crafted patterns identifying specific key concepts (Sukkarieh et al., 2004; Mitchell et al., 2002).

For extended written responses, it is less practical to make comparisons with model responses, due to the greater length and variability of the responses. However, another approach that does not require pre-scored responses is possible for test questions that have prompts with substantial amounts of information that should be included in the answer. In these cases, the similarity between the response and the prompt materials can be calculated, with the hypothesis that higher scoring responses will incorporate certain prompt materials more than lower scoring responses. This approach was taken by (Gurevich and Deane, 2007) which demonstrated that lower proficiency non-native essay writers tend to use more content from the reading passage, which is visually accessible and thus easier to comprehend, than the listening passage. The current study investigates a similar approach for spoken responses.

3 Data

The data used in this study was drawn from TOEFL iBT, an international assessment of academic English proficiency for non-native speakers. For this study, we focus on a task from the assessment which elicits a 60 second spoken response from the test takers. In their response, the test takers are asked to use information provided in reading and listening stimulus materials to answer a question concerning specific details in the materials. The responses are then scored by expert human raters on a 4-point scale using a scoring rubric that takes into account the following three aspects of spoken English proficiency: delivery (e.g., pronunciation, prosody, fluency), language use (e.g., grammar, lexical choice), and topic development (e.g., content, discourse coherence). For this study, we used a total of 1189 responses provided by 299 unique speakers to four different prompts[1] (794 responses from 199 speakers were used for training and 395 responses from 100 speakers were used for evaluation).

4 Methodology

We investigated several variations of simple features that compare the lexical content of a spoken response to following three types of prompt materials: 1) *listening passage*: a recorded lecture or dialogue containing information relevant to the test question (the number of words contained in each of the four listening passages used in this study were 213, 223, 234, and 318), 2) *reading passage*: an article or essay containing additional information relevant to the test question (the number of words contained in the two reading passages were 94 and 111), and 3) *sample response*: a sample response provided by the test designers containing the main ideas expected in a model answer (the number of words contained in the four sample responses were 41, 74, 102, and 133).

The following types of features were investigated for each of the materials: 1) *stimulus_cosine*: the cosine similarity between the spoken response and the various materials, 2) *tokens/response*, *types/response*: the number of word tokens / types that occur in both the spoken response and each of

[1] Two out of the four tasks in this study had only listening materials; responses to these tasks are not included in the results for the features which require reading materials.

the materials, divided by the number of word tokens / types in the response,[2] and 3) *unique tokens*, *unique types*: the number of word tokens / types that occur in both the spoken response and one or two of the materials, but do not occur in the remaining material(s).

As a baseline, we also compare the proposed content features based on the prompt materials to content features based on collections of scored responses to the same prompts. This type of feature has been shown to be effective for content scoring both in non-native essays (Attali and Burstein, 2006) and spoken responses (Xie et al., 2012), and is computed by comparing the content in a test response to content models trained using responses from each of the score points. It is defined as follows:

- Sim_i: the similarity score between the words in the spoken response and a content model trained from responses receiving score i ($i \in 1, 2, 3, 4$ in this study)

The Sim_i features were trained on a corpus of 7820 scored responses (1955 for each of the four prompts), and we investigated two different methods for computing the similarity between the test responses and the content models: Content Vector Analysis using the cosine similarity metric (CVA) and Pointwise Mutual Information (PMI).

The spoken responses were processed using an HMM-based triphone ASR system trained on 800 hours of non-native speech (approximately 15% of the training data consisted of responses to the four test questions in this study), and the ASR hypotheses were used to compute the content features.[3]

5 Results

We first examine the performance of each of the individual features by calculating their correlations with the holistic English speaking proficiency scores provided by expert human raters. These results for the training partition are presented in Table 1.[4]

Feature Set	Feature	*r*
stimulus_cosine	listening	0.384
	reading	0.176
	sample	0.384
tokens/response	listening	0.022
	reading	0.096
	sample	0.121
types/response	listening	0.426
	reading	0.142
	sample	0.128
unique tokens	L'RS	0.116
	L'RS'	0.162
	LR'S	0.219
	LR'S'	0.337
unique types	L'RS	0.140
	L'RS'	0.166
	LR'S	0.259
	LR'S'	0.450
CVA	Sim_1	0.091
	Sim_2	0.186
	Sim_3	0.261
	Sim_4	0.311
PMI	Sim_1	0.191
	Sim_2	0.261
	Sim_3	0.320
	Sim_4	0.361

Table 1: Correlations of individual content features with holistic human scores on the training partition

As Table 1 shows, some of the individual content features based on the prompt materials obtain higher correlations with human scores than the baseline CVA and PMI features based on scored responses. Next, we investigated the overall contribution of the content features to a scoring model that takes into account features from various aspects of speaking proficiency. To show this, we built a baseline linear regression model to predict the human scores using 9 features from 4 different aspects of speaking

[2] Dividing the number of matching word tokens / types by the number of word tokens in the response factors out the overall length of the response from the calculation of the feature.

[3] Transcriptions were not available for the spoken responses used in this study, so the exact WER of the ASR system is unknown. However, the WER of the ASR system on a comparable set of spoken responses is 28%.

[4] For the *unique tokens* and *unique types* features, each row lists how the prompt materials were used in the similarity comparison as follows: R = reading, L = listening, S = sample, and ' indicates no lexical overlap between the spoken response and the material. For example, L'RS indicates content from the test response that overlapped with both the reading passage and sample response but was not contained in the listening material.

proficiency (fluency, pronunciation, prosody, and grammar) produced by SpeechRater, an automated speech scoring system (Zechner et al., 2009), as shown in Table 2.

Category	Features
Fluency	normalized number of silences > 0.15 sec, normalized number of silences > 0.495 sec, average chunk length, speaking rate, normalized number of disfluencies
Pronunciation	normalzied Acoustic Model score from forced alignment using a native speaker AM, average normalized phone duration differnce compared to a reference corpus
Prosody	mean deviation of distance between stressed syllables
Grammar	Language Model score

Table 2: Baseline speaking proficiency features used in the scoring model

In order to investigate the contribution of the various types of content features to the scoring model, linear regression models were built by adding the features from each of the feature sets in Table 1 to the baseline features. The models were trained using the 794 responses in the training set and evaluated on the 395 responses in the evaluation set. Table 3 presents the resulting correlations both for the individual responses (N=395) as well as the sum of all four responses from each speaker (N=97).[5]

As Table 3 shows, all of the scoring models using feature sets with the proposed content features based on the prompt materials outperform the baseline model. While none of the models incorporating features from a single feature set outperforms the baseline CVA model using features based on scored responses, a model incorporating all of the proposed prompt-based content features, *all prompt-based*, does outperform this baseline. Furthermore, a model incorporating all of the content features (both the proposed features and the baseline CVA / PMI features), *all content*, outperforms a model using only the baseline CVA and PMI features.[6]

Feature Set	response r	speaker r
Baseline	0.607	0.687
+ *types/response*	0.612	0.701
+ *tokens/response*	0.615	0.700
+ *unique tokens*	0.616	0.695
+ *stimulus_cosine*	0.630	0.716
+ *unique types*	0.658	0.761
+ CVA	0.665	0.762
+ all prompt-based	0.677	0.779
+ PMI	0.723	0.818
+ CVA and PMI	0.723	0.818
+ all content	0.742	0.838

Table 3: Performance of scoring models with the addition of content features

6 Discussion and Conclusion

This paper has demonstrated that the use of content scoring features based solely on the prompt stimulus materials and a sample response is a viable alternative to using features based on content models trained on large sets of pre-scored responses for the automated assessment of spoken language proficiency. Under this approach, automated scoring systems for large-scale spoken language assessments involving spontaneous speech can begin to address an area of spoken language proficiency (content appropriateness) which has mostly been neglected in systems that have been developed to date. Compared to an approach using pre-scored responses for training the content models, the proposed approach is much more cost effective and reduces the risk that test materials will be seen by test takers prior to the assessment; both of these attributes are crucial benefits for large-scale, high-stakes language assessments. Furthermore, the proposed prompt-based content features, when combined in a linear regression model with other speaking proficiency features, outperform a baseline set of CVA content features which use models trained on pre-scored responses,

[5] Three speakers were removed from the evaluation set for this analysis since they provided fewer than four responses.

[6] While the prompt-based content features do result in improvements, neither of these two differences are statistically significant at $\alpha = 0.05$ using the Hotelling-Williams Test, since both the magnitude of the increase and the size of the data set are relatively small.

and they add further improvement to a model incorporating the higher performing baseline with PMI content features.

The results in Table 1 indicate that the individual features based on overlapping lexical types (*types/response* and *unique types*) perform slightly better than the ones based on overlapping lexical tokens (*tokens/response* and *unique tokens*). This suggests that it is important for test takers to use a range of concepts that are contained in the stimulus materials in their responses. Similarly to the result from (Gurevich and Deane, 2007), Table 1 also shows that the features measuring overlap between the response and the listening materials typically perform better than the features measuring overlap between the response and the reading materials; the best individual feature, LR'S' for *unique types*, measures the amount of overlap with lexical types that are contained in the listening stimulus, but absent from the reading stimulus and sample response. This indicates that the use of content from the listening materials is a better differentiator among students of differing language proficiency levels than reading materials, likely because test takers generally have more difficulty understanding the content from listening materials.

Table 1 also shows the somewhat counterintuitive result that features based on no lexical overlap with the sample response produce higher correlations than features based on lexical overlap with the sample response, when there is lexical overlap with the listening materials and no overlap with the reading materials. That is, the LR'S' feature outperforms the LR'S feature for both the *unique types* and *unique tokens* features sets. However, as shown in Section 4, the sample responses varied widely in length (ranging from 41 to 133 words), and all were substantially shorter than the listening materials, which ranged from 213 to 318 words. Therefore, it is likely that many of the important lexical items from the sample response are also contained in the listening materials. Thus, the LR'S feature provided less information than the LR'S' feature.

The features used in this study are all based on simple lexical overlap statistics, and are thus trivial to implement. Future research will investigate more sophisticated methods of text-to-text similarity for prompt-based content scoring, such as those used in (Mohler and Mihalcea, 2009). Furthermore, future research will address the validity of the proposed features by ensuring that there are ways to filter out responses that are too similar to the stimulus materials, and thus indicate that the test taker simply repeated the source verbatim.

7 Acknowledgments

The authors would like to thank Yigal Attali for sharing his ideas about prompt-based content scoring.

References

Yigal Attali and Jill Burstein. 2006. Automated essay scoring with e-rater® V.2. *The Journal of Technology, Learning, and Assessment*, 4(3):3–30.

Jared Bernstein, Alistair Van Moere, and Jian Cheng. 2010. Validating automated speaking tests. *Language Testing*, 27(3):355–377.

Abhishek Chandel, Abhinav Parate, Maymon Madathingal, Himanshu Pant, Nitendra Rajput, Shajith Ikbal, Om Deshmuck, and Ashish Verma. 2007. Sensei: Spoken language assessment for call center agents. In *Proceedings of ASRU*.

Miao Chen and Klaus Zechner. 2012. Using an ontology for improved automated content scoring of spontaneous non-native speech. In *Proceedings of the 7th Workshop on Innovative Use of NLP for Building Educational Applications, NAACL-HLT*, Montréal, Canada. Association for Computational Linguistics.

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).

Olga Gurevich and Paul Deane. 2007. Document similarity measures to distinguish native vs. non-native essay writers. In *Proceedings of NAACL HLT*, Rochester, NY.

Tom Mitchell, Terry Russell, Peter Broomhead, and Nicola Aldridge. 2002. Towards robust computerised marking of free-text responses. In *Proceedings of the 6th International Computer Assisted Assessment (CAA) Conference*, Loughborough.

Michael Mohler and Rada Mihalcea. 2009. Text-to-text semantic similarity for automatic short answer grading. In *Proceedings of the European Chapter of the Association for Computational Linguistics (EACL 2009)*, Athens, Greece.

Jana Sukkarieh, Stephen Pulman, and Nicholas Raikes. 2004. Auto-marking 2: An update on the UCLES-Oxford University research into using computational linguistics to score short, free text responses. In

International Association of Educational Assessment, Philadelphia.

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, Montréal, Canada. 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.