Automated Scoring Across Different Modalities

Anastassia Loukina and Aoife Cahill
Educational Testing Service
660 Rosedale Rd
Princeton, NJ 08541, USA
aloukina@ets.org, acahill@ets.org

Abstract

In this paper we investigate how well the systems developed for automated evaluation of written responses perform when applied to spoken responses. We compare two state of the art systems for automated writing evaluation and a state of the art system for evaluating spoken responses. We find that the systems for writing evaluation achieve very good performance when applied to transcriptions of spoken responses but show degradation when applied to ASR output. The system based on sparse n-gram features appears to be more robust to such degradation. We further explore the role of ASR accuracy and the performance and construct coverage of the combined model which includes all three engines.

1 Introduction

In this paper we evaluate how well the systems developed for automated evaluation of *written* responses perform when applied to *spoken* responses. We use a corpus of spoken responses to an English language proficiency test and compare the performance of two state-of-the-art systems for evaluating writing and a state of the art system for evaluating spoken responses.

Automated speech scoring, until recently, primarily focused on evaluating pronunciation and prosody of highly constrained read speech (Bernstein et al., 1990; Neumeyer et al., 1996; Witt and Young, 2000). With the improvement in automatic speech recognition technology, automated scoring has lately also been applied to constructed responses where the content of the response may not be known in advance (Zechner et al., 2009; Cheng et al., 2014). Earlier scoring systems for such responses still primarily evaluated delivery aspect of the response, but there has also been a growing amount of work on automatic evaluation of grammar, vocabulary and content of spoken responses (Bernstein et al., 2010; Chen and Zechner, 2011; Xie et al., 2012; Bhat and Yoon, 2015).

While automatic evaluation of these high-level aspects of language proficiency is a relatively new field in automated speech scoring, there exists a substantial body of research on evaluating these constructs in written responses including several systems already used operationally for scoring responses to high-stakes language proficiency tests (see Shermis (2014) for a comprehensive overview).

Automated scoring systems for spoken and written responses generally share a common structure: they extract a set of features measuring different aspects of language proficiency and use a machine learning algorithm to map those features to a human score. There is also a substantial overlap in the criteria used to score grammar and vocabulary of spoken and written responses. Therefore, it is not unreasonable to expect that some of the features developed for evaluating writing will also be applicable to scoring spoken responses (cf. Crossley and McNamara (2013)).

On the other hand, the performance of such features can be affected by a number of factors. First of all, many grammatical features rely on knowledge of sentence boundaries in order to parse the response into syntactic constituents. In written responses the sentence boundaries can be established

Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications, pages 130–135,
San Diego, California, June 16, 2016. ©2016 Association for Computational Linguistics

based on punctuation. In spoken responses, however, these have to be estimated using machine learning algorithms such as the ones described in Chen and Yoon (2011). Furthermore, sentence boundaries in speech are often ambiguous. These factors may lead to a decrease in feature performance.

Second, in automated speech scoring the transcription of the spoken responses necessary to evaluate grammar and vocabulary is obtained using automated speech recognition (ASR) (Higgins et al., 2011). These systems may incorrectly recognize certain words introducing additional noise into the feature input and consequently lowering their performance.

Finally, spoken and written discourse differ in what is considered appropriate in terms of language use (Chafe and Tannen, 1987; Biber and Gray, 2013). Thus, for example, sentence fragments typically considered inappropriate for written language are generally very common in unscripted spoken responses. This may also impact how well the features developed for written responses perform on spoken responses.

We first introduce three state-of-the-art operational systems for automated scoring. We then apply the engines for evaluating writing to a corpus of spoken responses. Finally, we evaluate whether combining different engines leads to further improvement in system performance and construct coverage.

2 Automated scoring systems

2.1 *e-rater*®

e-rater® (*E*) is an engine that can automatically provide feedback on students' writing, as well as automatically assign a score to that writing. Using statistical and rule-based NLP methods, *E* identifies and extracts several feature classes for model building and essay scoring (Attali and Burstein, 2006; Burstein et al., 2013)). Individual feature classes typically represent an aggregate of a larger feature set and are designed to capture a specific aspect of the construct being measured. The feature classes used in this paper include the following: (a) grammatical errors (e.g., subject-verb agreement errors), (b) word usage errors (e.g., their versus there), (c) presence of essay-based discourse elements (e.g., thesis statement, main points, supporting details, and conclusions), (d) development of essay-based discourse elements, (e) a feature that considers correct usage of prepositions and collocations (Futagi et al., 2008), and (f) sentence variety. To train a new scoring model, features are extracted from a training data set, and a linear model (roughly equivalent to non-negative least squares regression) is learned.

2.2 c-rater-ML

c-rater-ML (*C*) is an automated scoring engine originally designed to evaluate the content of a student response. It is typically applied to short responses ranging from a few words to a short paragraph. Therefore, in contrast to *E* and *S* many of the features used in the *C* engine are sparse lexicalized features similar to the ones described in Heilman and Madnani (2013). In addition to word and character n-gram features, the models also include syntactic dependency features. As a result of the large number of sparse features, the modeling technique for this kind of feature set needs to be different from a straightforward linear model. *C* employs a Support Vector Regressor with a radial basis function kernel. We use this alternative approach to scoring (with many sparse lexical features and a non-linear learning function) to contrast with the typical scoring models used for evaluating speech or writing quality.

2.3 SpeechRater

SpeechRater$^{\text{sm}}$ (*S*) (Zechner et al., 2009) is an automated scoring engine that is designed to evaluate the quality of spontaneous spoken responses. Using signal processing as well as NLP techniques, *S* extracts features for evaluating both the delivery characteristics (e.g. fluency, pronunciation) and the language use characteristics (e.g. grammar, vocabulary) of each response.

The features are extracted from the sound recording of the response using the two stage method described in Higgins et al. (2011) where the transcription of the responses is obtained using automated speech recognition technology. The ASR engine incorporated into *S* was trained on over 800 hours of non-native speech from the same assessment used in this study with no speaker overlap. The ASR system uses a GMM-based crossword triphone acoustic model and a 4-gram language model with a vocabulary size of 65,000 words.

The *S* model used in this study contained 18 features. Of these, 15 features covered various aspects of delivery such as fluency, pronunciation and rhythm. Three features measured language use. These feature were: (1) average log of the frequency of all content words (Yoon et al., 2012), (2) CVA-based comparison between the lexical content of each response and the reference corpus (based on Xie et al. (2012)) and (3) a CVA-based comparison computed based on part-of-speech tags (Bhat and Yoon, 2015). As in case of *E*, the final score is computed as a linear combination of these features.

3 Data and methodology

3.1 Corpus of spoken responses

The study is based on a corpus of 5,884 spoken responses to an English language proficiency test obtained from 996 speakers. The corpus contains up to six responses from each speaker. Each response was unscripted and around 1 minute long.

The corpus was equally split into training and evaluation sets (2,941 responses each). There was no overlap of speakers or prompts in the two sets. All responses were assigned a holistic proficiency score by expert raters. The scores ranged from 1 (low proficiency) to 4 (high proficiency). The raters evaluated the overall intelligibility of responses, grammar, the use of vocabulary, and topic development. To obtain human benchmarks, 136 responses from the evaluation set were scored by two raters. The agreement between the two raters was Pearson's $r = 0.56$.

All responses were transcribed by a professional transcription agency. The average length of transcribed responses was 104 words ($\sigma = 30$) with the length of 50% of the responses falling between 84 and 124 words. The responses from more proficient speakers were generally longer: the number of words in the response was moderately correlated with proficiency score with Pearson's $r = 0.51$ ($p < 0.0001$). Table 1 shows the number of responses in the evaluation set assigned to each score category as well as the mean and standard deviation of the number of words in the transcriptions of these responses.

Score	1	2	3	4
N responses	104	1005	1485	347
Average N words	51.5	91.0	111.9	126.5
Std. N words	25.4	24.7	25.5	28.9
Average WER	48%	36%	31.5%	30%

Table 1: The total number of responses in each score category, the average number of words/standard deviation of transcribed responses and the average ASR word error rate for responses in each category.

Finally, we computed the word error rate (WER) for the ASR output for each response in our corpus. The average WER for the whole corpus was 34% ($\sigma = 13.7$). The ASR was somewhat more accurate for more proficient speakers with the correlation between response WER and response score Pearson's $r = -0.24$ ($p < 0.0001$). As can be seen from Table 1, the relationship between WER and proficiency scores was non-linear: the WER was substantially greater for speakers with the lowest proficiency level (score 1), the difference between the rest of the speakers was smaller.

3.2 Method

3.2.1 Feature extraction

We used each of the three engines to extract the corresponding features for each response. Features for *S* were extracted following the operational pipeline from the sound recording of the response with all features including those related to language use computed on the ASR output. For *E* and *C* the features were extracted using three different inputs: (1) the ASR output (the same as used in *S*), (2) the expert human transcription which included punctuation, and (3) the human transcription after removing all punctuation. All transcriptions were processed to remove fillers such as 'uhm' and 'uh', word fragments and repeated words.

3.2.2 Model building

We then used various combinations of features to compare the performance of the engines. For *S* and *E* we used the actual feature values returned by each system (18 *S* features and 9 *E* features). Since *C* is based on numerous sparse features we used a stacking approach (Wolpert, 1992) to combine it with the other two engines: we used 10-fold cross-validation to generate predicted scores for all responses in the training set and used these predicted scores as a single "*C*" feature. For the evaluation set this feature corresponded to scores predicted for the evaluation set using the *C* model trained on the training set.

We then trained a series of models based on different combinations of features from the three engines. The coefficients for all models were estimated on the training set using non-negative least squares regression. The models were then used to generate predictions on the evaluation set. Finally, in all cases the predictions were re-scaled using a normal transformation to match the distribution of human scores on the training set.

4 Results

Table 2 shows the performance of all 19 models in terms of correlation (Pearson's r) between predicted and observed scores for the evaluation set. We used Steiger's method for comparing dependent correlations (Steiger, 1980)[1] to evaluate whether the differences between models are statistically significant. Unless stated otherwise, all reported differences are significant at $\alpha = 0.01$ after applying Bonferroni correction for multiple comparisons.

4.1 Performance of *E* and *C* on transcriptions

When used with human transcription, both *E* and *C* performed close to the *S* baseline (r = 0.61 for *E*, 0.61 for *C* and 0.63 for *S*, the differences are not significant). There was a small improvement from combining the two automated writing evaluation engines (*EC*) with r increasing to 0.64.

4.2 Performance of *E* and *C* on ASR output

There was a decrease in performance of both writing engines when the features were computed on ASR output. The degradation was larger for *E* (from 0.61 to 0.52). *C* appeared to be more robust to the noise introduced by ASR with the performance decreasing from 0.61 to 0.58. The model based on the combination of both engines computed on ASR output achieved r = 0.60.

We further explored the reason for degradation between the features computed on transcription and ASR output. As discussed in the introduction, ASR output does not contain sentence boundaries necessary for the computation of some of the features. To evaluate the impact of this factor we computed a new set of features using human transcriptions with punctuation removed. We found that this had no significant effect on model performance.

We next looked at the effect of response WER on the accuracy of the scoring engine for this response. We first hypothesized that the scoring error may be greater for responses with higher WER. To test this hypothesis we computed the correlation between the scoring error (the absolute difference between predicted and observed score) and the WER for each response. As expected, there was no significant correlation when automated scores were computed on transcriptions with or without punctuation. Surprisingly, the correlations between WER and scoring error for scores computed on ASR output were very low: r = 0.09 ($p < 0.00001$) for *E* and r = 0.07, p = 0.0001 for *C* . In other words, there was no linear relationship between the response WER and the scoring error.

We also tested whether the relationship between the scoring error and WER was further obscured by training the models on already "noisy" ASR outputs. We retrained the *E*, *C* and *EC* models using the features computed on transcriptions and then evaluated them using features computed on ASR outputs. We found that the performance of these new models was similar (*E*) or slightly lower (*C* and *EC*) than the performance of the models trained on ASR outputs. Furthermore, the correlation between the WER and scoring error for these models were as low as the correlations observed for ASR-trained models.

4.3 Combined performance of all three engines

Finally, we evaluated the performance and construct coverage of the model based on the combination of all three engines.

We found that the performance improved if *S* features were combined with writing features computed based on transcriptions. This is not surprising considering that all *S* features were computed on ASR output and therefore the new features computed on transcriptions most likely contained the information lost due to inaccurate ASR.

For features computed on ASR output, we found that there was no further gain in performance from

[1] We used the Python implementation from https://github.com/psinger/CorrelationStats

Model	Description	Org N feats	trans	trans-no-punct	asr
S	Baseline model containing *S* features only	18			0.63
E	The model containing only *E* features	9	0.61	0.60	0.52
C	The original *C* model	(see text)	0.61	0.62	0.58
EC	The model which combined *E* features and predictions from *C* (see main text)	10	0.64	0.64	0.60
SE	The combination of *S* and *E* features	27	0.67	0.66	0.63
SC	The combination of *S* features and predictions from *C* (see main text)	19	0.66	0.66	0.64
SEC	The combination of *S* and *E* features and predictions from *C*	28	0.67	0.67	0.64

Table 2: Summary of performance (Pearson's r between the predicted and human score) of all 19 models evaluated in this study. The table shows the original number of features and the model performance for different types of input (see section 3.2.1). The final number of features in the model may be less than the original number of features since all coefficients were set to be positive.

combining *S* and *E* (*SE*), *C* (*SC*) or all three engines (*SEC*).

We also evaluated which features had the biggest contribution to the final score in different models. In the baseline *S* model, delivery features (fluency, pronunciation, and prosody) accounted for 80% of the final score. The language use features accounted for the remaining 20%. In the combined model, the relative contribution of language use features increased to 38% for *SE*, 37% for *SC* and 44% for *SEC*. Thus in the combined model the delivery and language use features are more evenly balanced.

5 Discussion

In this paper we explored how well state-of-the-art engines for evaluating written responses perform when applied to transcriptions of spoken responses. Surprisingly, we found that the writing engines achieve relatively high agreement with human scores even though they do not measure some fundamental aspects of spoken language proficiency: fluency and pronunciation. At the same time, there was no improvement between the baseline *S* system and a system that combines all three engines.

Furthermore, the drop in performance of writing engines when moving from well-formed transcribed text to ASR output is not as high as one might initially expect given the relatively high WER in this data set. We also found that the relationship between the ASR accuracy and scoring error was not straightforward and deserves further study. Finally, lack of sentence boundaries had no effect on the engine performance.

C engine showed good agreement with human scores even though there was no overlap between prompts in training and evaluation sets and therefore the model could not learn any specifics relevant to particular prompts and had to rely on general patterns of word use.

Our results highlight the complex role of construct in automated scoring. The majority of speakers who show good performance along one of the dimensions of language proficiency generally also score high along other dimensions. This is exemplified by our result that writing engines which measured only one aspect of spoken responses still showed relatively high agreement with holistic scores. Consequently, as shown in this study, the gain in performance from combining different engines is small or non-existent. However, a system which heavily relies on features measuring a single aspect of proficiency is sub-optimal both in terms of validity of the final score and the system vulnerability to various gaming behaviours.

We showed that combining the speech scoring engine with the existing features developed for scoring written responses produces a model where the contribution of different proficiency aspects to the final score is more balanced leading to a more valid system, which is potentially more robust to gaming. In future study we will investigate the individual contribution of different features in these engines.

Acknowledgments

We thank Keelan Evanini, Su-Youn Yoon, Brian Riordan, Klaus Zechner and two anonymous BEA reviewers for their comments and suggestions.

References

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

Jared Bernstein, Michale Cohen, Hy Murveit, Dimitry Rtischev, and Mitchel Weintraub. 1990. Automatic Evaluation and Training in English Pronunciation. In *Proceedings of ICSLP 90*, pages 1185–1188.

Jared Bernstein, Jian Cheng, and Masanori Suzuki. 2010. Fluency and Structural Complexity as Predictors of L2 Oral Proficiency. *Proceedings of Interspeech 2010, Makuhari, Chiba, Japan*, pages 1241–1244.

Suma Bhat and Su-Youn Yoon. 2015. Automatic assessment of syntactic complexity for spontaneous speech scoring. *Speech Communication*, 67:42–57.

Douglas Biber and Bethany Gray. 2013. Discourse Characteristics of Writing and Speaking Task Types on the TOEFL iBT Test: A Lexico-Grammatical Analysis. *ETS Research Report*, RR-13-04.

Jill Burstein, Joel Tetreault, and Nitin Madnani. 2013. The e-rater automated essay scoring system. In Mark D. Shermis and Jill Burstein, editors, *Handbook of automated essay evaluation: Current applications and new directions*. Routledge, New York, NY.

Wallace Chafe and Deborah Tannen. 1987. The relation between written and spoken language. *Annual Review of Anthropology*, 16:383–407.

Lei Chen and Su-Youn Yoon. 2011. Detecting structural events for assessing non-native speech. *Proceedings of the 6th workshop on Innovative Use of NLP for Buidling Educational Applications*, pages 38–45.

Miao Chen and Klaus Zechner. 2011. Computing and Evaluating Syntactic Complexity Features for Automated Scoring of Spontaneous Non-Native Speech. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 722–731.

Jian Cheng, Yuan Zhao D'Antilio, Xin Chen, and Jared Bernstein. 2014. Automatic Assessment of the Speech of Young English Learners. *Proceedings of the 9th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 12–21.

Scott Crossley and Danielle McNamara. 2013. Applications of Text Analysis Tools for spoken response grading. *Language Learning & Technology*, 17(2):171–192.

Yoko Futagi, Paul Deane, Martin Chodorow, and Joel Tetreault. 2008. A computational approach to detecting collocation errors in the writing of non-native speakers of english. *Computer Assisted Language Learning*, 21:353–367.

Michael Heilman and Nitin Madnani. 2013. ETS: Domain adaptation and stacking for short answer scoring. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 275–279.

Derrick Higgins, Xiaoming Xi, Klaus Zechner, and David Williamson. 2011. A three-stage approach to the automated scoring of spontaneous spoken responses. *Computer Speech & Language*, 25(2):282–306.

Leonardo Neumeyer, Horacio Franco, Mitchell Weintraub, and Patti Price. 1996. Automatic text-independent pronunciation scoring of foreign language student speech. *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96*, 3:1457–1460.

Mark D. Shermis. 2014. State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration. *Assessing Writing*, 20:53–76.

James H Steiger. 1980. Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, 87(2):245.

Silke M. Witt and Steve J. Young. 2000. Phone-level pronunciation scoring and assessment for interactive language learning. 30(2):95–108.

David H. Wolpert. 1992. Stacked generalization. *Neural Networks*, 5:241–259.

Shasha Xie, Keelan Evanini, and Klaus Zechner. 2012. Exploring content features for automated speech scoring. In *NAACL HLT '12 Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 103–111.

Su-Youn Yoon, Suma Bhat, and Klaus Zechner. 2012. Vocabulary profile as a measure of vocabulary sophistication. *Proceedings of the Seventh Workshop on the innovative use of NLP for Building Educational Applications*, pages 180–189.

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.