

# Automatic evaluation of spoken summaries: the case of language assessment

Anastassia Loukina, Klaus Zechner, Lei Chen

Educational Testing Service (ETS)

Princeton, NJ 08541, USA

aloukina@ets.org, kzechner@ets.org, lchen@ets.org

## Abstract

This paper investigates whether ROUGE, a popular metric for the evaluation of automated written summaries, can be applied to the assessment of spoken summaries produced by non-native speakers of English. We demonstrate that ROUGE, with its emphasis on the recall of information, is particularly suited to the assessment of the summarization quality of non-native speakers' responses. A standard baseline implementation of ROUGE-1 computed over the output of the automated speech recognizer has a Spearman correlation of  $\rho = 0.55$  with experts' scores of speakers' proficiency ( $\rho = 0.51$  for a content-vector baseline). Further increases in agreement with experts' scores can be achieved by using types instead of tokens for the computation of word frequencies for both candidate and reference summaries, as well as by using multiple reference summaries instead of a single one. These modifications increase the correlation with experts' scores to a Spearman correlation of  $\rho = 0.65$ . Furthermore, we found that the choice of reference summaries does not have any impact on performance, and that the adjusted metric is also robust to errors introduced by automated speech recognition ( $\rho = 0.67$  for human transcriptions vs.  $\rho = 0.65$  for speech recognition output).

## 1 Introduction

In this paper we explore whether metrics commonly used for the automated evaluation of written summaries can be used to evaluate spoken summaries in the context of language assessment.

The performance of automatic summarization systems is routinely evaluated using content met-

rics such as ROUGE (Lin and Rey, 2004), which measures the  $n$ -gram overlap between the candidate summary and a set of reference summaries (see also Rankel et al. (2013) for historical background). ROUGE is a recall-oriented metric inspired by its precision-oriented counterpart BLEU, developed to evaluate machine translations (Papineni et al., 2002). Recent research in this area has been focused on identifying the most reliable variants of ROUGE and best practices in the application of the metric (Owczarzak et al., 2012; Rankel et al., 2013). These studies (reviewed in more detail in Section 2.1) showed that less commonly used variants of ROUGE may in fact be more consistent with human judgments, at least in the context of automatic summary evaluation.

Beyond the research in automatic summarization systems, ROUGE has also been used to evaluate written summaries in the context of educational assessment. Madnani et al. (2013) showed that one of the variants of ROUGE, in combination with other metrics, performed consistently well for the automated scoring of written responses to summary tasks produced by middle- and high-school students. They did not investigate the effect of using other variants of ROUGE.

In this paper, we explore whether ROUGE can be used to automatically evaluate the content coverage of spoken summaries produced by non-native speakers in the context of language assessment. As in case of automatic text summaries, the human raters who score these responses are asked to assess whether the summary accurately conveys the information contained in the stimulus. While the length of the spoken responses is more loosely constrained than in case of automatic text summaries, human raters do not penalize for extraneously irrelevant language. Therefore recall-oriented ROUGE is an attractive evaluation metric for this task.

At the same time, unlike automatic text sum-

maries, spoken summaries are abstractive and often contain ungrammatical sequences, repetitions, repairs, and other disfluencies. Further ‘noise’ is introduced by transcription errors generated by the automated speech recognition system. In this study, we assess whether (a) ROUGE is robust against this type of noise; (b) how many reference summaries are necessary to obtain reliable evaluation; and (c) how the choice of specific reference summaries affects the performance of the metric (Section 4.1). We also assess which variants of ROUGE have the most agreement with human judgments on this type of summary and what adjustments can be made to mitigate the effects of disfluencies and errors introduced by automated speech recognition (Section 4.2). Finally, we test how well our adjusted variant of ROUGE can predict the human scores on unseen data (Section 4.3).

## 2 Related work

### 2.1 The application of ROUGE to evaluation of automatic text summarization

There exist various versions of ROUGE which differ in terms of the length of their  $n$ -grams, the use of skip-bigrams, the application of stemming, and the exclusion of stop-words. Several studies have compared these variants to identify those most consistent with human judgments. In earlier work, Lin (2004) reported that variants based on unigrams and skip-bigrams (ROUGE-SU4) or bigrams alone (ROUGE-2) performed best. ROUGE-2 was also identified as the best variant more recently by Owczarzak et al. (2012). Rankel et al. (2013) found that linear combinations of these metrics with ROUGE based on longer  $n$ -grams are more accurate in finding significantly different systems.

Previous work also explored various methods of text pre-processing prior to the computation of ROUGE, including stemming and the removal of stop-words, neither of which had any substantial effect on the performance of ROUGE (Lin and Rey, 2004; Owczarzak et al., 2012). Owczarzak et al. (2012) reported that the agreement with human judgments was, in fact, higher if the stop-words were retained.

All applications discussed so far used ROUGE to evaluate the textual summarization of written texts. There have also been attempts to apply this metric to text summaries of speech data with mixed results (see Nenkova and McKeown (2011)

for a review). ROUGE performed reasonably well for the evaluation of text summaries of spoken presentations (Hirohata et al., 2005), but was not correlated with the summary accuracy of summaries of meetings or conversations (although see (Penn and Zhu, 2008)).

Most of this work was performed on extractive summaries produced by summarization systems that used multiple summaries to evaluate each system. In this study, we explore the application of ROUGE to the evaluation of abstractive summaries produced by students in a language assessment context with an aim of producing a separate evaluation for each summary. Furthermore, the fact that these are spoken responses adds an extra layer of complexity to the analysis, therefore the results of previous studies cannot directly be applied to this new context.

### 2.2 Previous approaches to the content evaluation of spoken summaries for assessment purposes

The research on the automated scoring of content accuracy in a language assessment has primarily focused on the evaluation of written essays. Most previous approaches in this area have used so-called “bag-of-words”-based models, gleaned from the discipline of information retrieval. The basic idea is that an essay is considered to be highly content relevant to a given topic when it contains words that are similar to those seen in previously collected essays with high human-rater scores. For instance, Attali and Burstein (2006) used a vector-space model to compute the cosine similarities between word vectors found in an essay to be automatically scored and word vectors comprising previously scored essays with the same human-rater score. In a similar vein, Foltz et al. (1999) computed a compressed vector space based on singular value decomposition for a set of document-word vectors, called latent semantic analysis, and then computed similarity scores for essays based on this more compact representation.

It should be noted, though, that since all of these models do not take word sequences into account, they must be considered knowledge-poor in that they cannot distinguish between syntactic roles or a list of random words versus a well-formed sentence. In operational systems, such bag-of-words similarity features are combined with features which evaluate grammar and other aspects

of language use; therefore a random list of content words is unlikely to lead to a high overall score. However, finer-grained distinctions such as negations or subject-object relationships between words are often lost.

Applications of these methods to spontaneous speech in spoken-language assessments have been conducted much more recently as this domain of language assessment relies on the output of Automatic Speech Recognition systems (ASR) that typically have a fairly high word-error rate. These errors can negatively affect the accuracy of the methods developed for written responses. Furthermore, spoken responses differ in many properties from written ones (Biber et al., 2004) and the validity of existing methods for assessing speech needs to be established before they can be used for operational scoring.

Xie et al. (2012) presented experiments using content features on spontaneous-speech data based on vector-space models, latent semantic analysis, as well as point-wise mutual information. Some of these content features showed higher correlations with human scores than features measuring other aspects of speaking proficiency, such as fluency or pronunciation. Chen and Zechner (2012) also used a vector space model for the scoring of spontaneous speech, but extended it by using the ontological information contained in WordNet. Finally, Xiong et al. (2013) used a variety of approaches to capture the content of spontaneous responses from the same corpus that we are investigating in this paper. Approaches varied from computing the overlap between key words in the stimuli and responses to a more traditional vector space model based on content vector analysis.

While these approaches have good correlations with human scores, they have a number of shortcomings. The best performing method suggested by Xiong et al. (2013) requires the manual annotation of the relevant key words for each prompt before the computation of the metric. Vector space models do not have this limitation, but they require a substantial number of reference summaries to achieve consistent results. Supporting this point, Chen (2013) showed that at least 50 reference responses were necessary to obtain moderate agreement between the cosine similarity measure and human judgments, with further improvement in agreement as the number of reference responses is increased to 200. These limitations pose prac-

tical difficulties when new items are added to the tests: the computation of content metrics for each new item requires either a manual annotation or a relatively large number of reference responses.

ROUGE appears promising in this context since it does not have either of these limitations. First, the computation of ROUGE does not require manual annotation. Second, research on the evaluation of written summaries suggests that relatively few reference summaries may be necessary to obtain reliable results, e.g., only four references were used for the summary evaluation at the Text Analysis Conference (Rankel et al., 2013). In addition, the recall-based nature of ROUGE is well-aligned with the evaluation criteria for these responses. Therefore in this paper, we explore whether any of the variants of ROUGE can be successfully applied to the content scoring of spoken summaries and what modifications may be necessary to achieve optimal performance.

### 3 Data and methodology

#### 3.1 Description of the corpus

The study is based on a corpus of responses collected during the pilot administration of the TOEFL®Junior™Comprehensive test, an international assessment of English proficiency targeted at middle-school students aged from 11 to 15 (see also Xiong et al. (2013) who used a subset of this corpus).

The corpus used in this study included 5,934 spoken responses produced by 1,611 speakers; all learners of English as a foreign language residing in different countries. In addition to a read-aloud task that was not relevant for this paper, the speakers were presented with four other tasks. First, the speakers were asked to describe a sequence of six pictures. For the remaining three tasks, the speakers listened to one announcement and two fragments from a lecture and were then asked to summarize the content of what they heard. The students were provided with a list of concepts that test takers were expected to cover in their responses.

For example, a student may have listened to a teacher giving an assignment in history class.<sup>1</sup> This assignment required the class to go to the library, look up information about the water supply in old and modern cities, answer the questions on their worksheet, and write a short paragraph about

<sup>1</sup><http://toefljr.caltesting.org/sampletest/s-historylesson.html>

their findings. The students were then asked to respond to the following prompt:

*Imagine that your classmate was not in class today. Tell your classmate about what the history teacher asked the students to do. Be sure to talk about the following:*

- the library
- the worksheet
- the homework

The corpus contained responses to 24 different prompts with 6 different sets of prompts. Each speaker only answered one set of prompts giving 4 responses per speaker. The recording time for each response was limited to 60 seconds. The actual number of words varied between participants with an average 72 words per response ( $\sigma = 29$ ).

From the originally recorded 6,444 responses, we excluded from further analysis 510 responses (about 8%), which contained either no speech or where the quality of the recording was too low for further analysis. All remaining 5,934 responses were scored on a scale of 1-4 by two expert human raters on a holistic scale that reflects all aspects of speaking proficiency, including pronunciation, grammar, and content coverage.<sup>2</sup> For content coverage, the raters were asked to consider whether the key information contained in the prompt was conveyed accurately or, in case of the picture description prompt, whether the story was complete. When the difference in the scores assigned by the two raters was greater than 1, the final score was assigned by an adjudicator.

The corpus was divided into non-overlapping training and testing partitions. The training partition contained 3,337 responses from 915 speakers and the test partition contained 2,597 spoken responses from 696 speakers. Both partitions included responses for the same prompts but there was no speaker overlap.

All responses were converted to text using a state-of-the-art automatic speech recognizer (ASR) with constrained vocabulary (see Evanini and Wang (2013) for further details). To evaluate the effect of the errors that may have been introduced by the ASR system, all responses were

<sup>2</sup>see [http://www.ets.org/s/toefl\\_junior/pdf/toefl\\_junior\\_comprehensive\\_speaking\\_scoring\\_guides.pdf](http://www.ets.org/s/toefl_junior/pdf/toefl_junior_comprehensive_speaking_scoring_guides.pdf) for the scoring rubrics

transcribed manually by professional human transcribers. Comparison with the human transcription showed that the ASR word error rate for this corpus was 26.5% for picture narration tasks and 29.4% for the summarization tasks.

### 3.2 Computation of the metrics

**Evaluation metrics.** ROUGE was computed using equation (1) as an  $n$ -gram ( $gr_n$ ) overlap between candidate summary and each summary ( $S$ ) from the set of reference summaries ( $RS$ ).

$$ROUGE_N = \frac{\sum_{S \in RS} \sum_{gr_n \in S} Count_{overlap}(gr_n)}{\sum_{S \in RS} \sum_{gr_n \in S} Count(gr_n)} \quad (1)$$

We used  $n$ -grams whereby  $n$  was in a range from 1 to 4 (ROUGE 1-4) and a combination of unigrams with skip-bigrams with maximum step of four words (ROUGE-SU1-4). Finally, we also computed a combined measure ROUGE-ALL which is the geometrical mean of ROUGE-1–ROUGE-4, computed by using the same smoothing procedure as for BLEU (Papineni et al., 2002).

We used the cosine distance (CVA) between the response and reference summaries as a baseline metric as this metric is commonly used for evaluating document similarity in the context of language assessment. CVA was computed as the cosine distance between candidate responses and the same reference responses as used for the computation of ROUGE. All term frequencies were weighted using  $tf-idf$  where  $tf$  is the frequency of a term in a given response and  $idf$  is the inverse document frequency.  $idf$  frequencies were computed based on all of the responses in the corpus.

**Reference summaries.** The reference summaries were selected from responses with the highest human rater final score (4). This approach is similar to using system outputs as pseudo-models for the evaluation of machine-translation or automatic-summarization systems (cf. Louis and Nenkova (2013)). It has also been successfully applied to the content assessment of written answers by Madnani et al. (2013) who used one randomly selected highly scored summary as a reference summary.

Since previous work on summarization evaluation showed that multiple summaries increase the reliability of evaluations (Louis and Nenkova, 2013; Nenkova and McKeown, 2011), we tested

how many summaries were necessary to achieve consistent results. We therefore computed ROUGE for each response using up to 10 randomly selected responses with final score of 4. To investigate the effect that different choices of reference summaries may have on the metrics, we repeated the analysis for 20 randomly selected sets of reference responses.

The corpus did not contain a sufficient number of responses with the maximum score for each prompt. Therefore, this part of the analysis was based on a subset of 1,784 responses selected from the training partition. This set included only 12 prompts for which human raters assigned a score of 4 to more than 11 responses.

**Text preprocessing.** For the evaluation of written summaries, ROUGE is usually computed using the raw counts of all of the terms. In addition to using this classical approach using unstemmed terms (*'all'*), we also computed ROUGE using three other approaches: (1) excluding all stop-words (*'Non-stop'*); (2) setting the frequency of all  $n$ -grams within each summary to 1, that is, counting types instead of tokens (*'Types'*); (3) excluding all stop-words and counting types only (*'Non-stop types'*). Finally, we computed all of these ROUGE variants using raw text as well as lemmatized text. As a result, we computed 72 different variants of ROUGE for each response and each combination of reference summaries: nine different types of ROUGE (eight different  $n$ -gram lengths and ROUGE-ALL) computed using four different methods of text processing and two possible approaches to lemmatization. All of the computations were done both on ASR and manual transcriptions.

### 3.3 Evaluation

We computed the Spearman's rank correlation between the metric and the holistic score assigned by the first rater to identify the best method of computing ROUGE and the optimal number of references. Performance of the metric may be affected by properties of the prompt (cf. (Nenkova and Louis, 2008)), therefore we first analyzed each prompt separately and then selected the variants that achieved the highest performance across all of the prompts. Since correlation coefficients are not normally distributed, we used several non-parametric methods to identify significant differences including non-parametric bootstrapping and non-parametric ANOVAs. These analyses were

done using the data from the training partition of the corpus.

We then evaluated how well the selected variants of ROUGE predicted human scores using a linear regression model trained on all of the data from the training partition using pooled data from all of the prompts. The model was tested on an unseen test partition that had not been used for any of the analyses.

Finally, we tested whether the new metrics improved the performance of the automated scoring engine for spoken responses. The current system assigns scores based on the linear combination of features with empirical weights obtained by training scoring models on scores assigned by expert raters (Zechner et al., 2009; Higgins et al., 2011). Current features measure various aspects of speaking proficiency such as fluency, pronunciation, and grammar usage. The performance of the system is evaluated with correlations and quadratic kappas between the scores assigned by the human raters and rounded predicted scores.

## 4 Results

All analyses were performed twice: each for metrics computed using ASR and manual transcriptions. We found that although the exact values of the correlation coefficients differed across these two transcriptions, the overall pattern of results remained the same. There was also a high correlation in metric values between the two types of transcription (Pearson's  $r$  for different types of ROUGE varied between 0.81 for ROUGE-4 and 0.9 for ROUGE-1). Since automated scoring relies on the output of automatic speech recognition, all numerical results reported in the main text of this section are based on ASR output. The tables report the numbers for both ASR and manual transcriptions.

### 4.1 Number and choice of reference responses

**Number of references.** To identify the optimal number of references for each prompt and metrics, we first found  $N_{best}$ , which had the highest correlation with human scores and then identified the lowest number of reference summaries for which the correlation coefficient was not significantly lower than the correlation coefficient for  $N_{best}$ .

Comparisons between different correlations

were performed using the general method suggested by Zou (2007) for comparing overlapping correlations as implemented by Baguley (2012, p.224) but we used bootstrapped confidence intervals (Wilcox, 2009). Confidence intervals for each correlation coefficient were constructed using pigeonhole bootstrapping (Owen, 2007) with 1,000 samples. For each  $N$  reference, we pooled the values computed for 20 randomly selected sets of different reference summaries. We then independently sampled responses and sets of references and selected values at each bootstrap repetition at the intersection of the two samples. The confidence intervals were constructed using the adjusted percentile method (Davison and Hinkley, 1997, p. 203-213). Since this analysis is more sensitive to Type II errors ('false negatives'), we set the significance threshold at  $\alpha = 0.15$ .

The optimal number of references varied between prompts, metrics, and methods of computation, but never exceeded 8. On average, optimal performance was achieved with 3 references. More references were required to achieve optimal performance for ROUGE based on longer  $n$ -grams (using the Kruskal-Wallis test, a non-parametric analysis of variance,  $p < 2.2 \times 10^{-16}$ ). For example, two references on average were required to achieve reliable results for ROUGE-1, but for ROUGE-4 this number was four references. The required number of references was also significantly dependent on the prompt (Kruskal-Wallis test,  $p < 2.2 \times 10^{-16}$ ) with averages varying between two and four. When the number of references was equal to or greater than the optimal number, there were no significant differences in the correlation coefficients across the different reference models.

For the analysis in the following section each of the 72 variants of ROUGE for each prompt was computed using the optimal  $N$  references identified for this variant and prompt.

#### 4.2 Types of ROUGE and different methods of computation

The correlation coefficients between the summarization metrics and human ratings depended on the length of  $n$ -grams (Kruskal-Wallis test  $p < 2.2 \times 10^{-16}$ ). While all types of ROUGE were positively correlated with human ratings, the correlation coefficients were the highest for ROUGE-1 and ROUGE-SU2-4, which performed significantly

better than ROUGE-3-4 and the combined measures ROUGE-ALL (post-hoc Tukey HSD test on ranked observations,  $p$  varied from  $p < 1 \times 10^{-10}$  to  $2.804 \times 10^{-4}$ ). The average correlations across the different types of text pre-processing for ASR and manual transcriptions are shown in Table 1.

Metrics	ASR output	Manual
ROUGE-1	0.616	0.637
ROUGE-SU4	0.592	0.608
ROUGE-SU3	0.595	0.609
ROUGE-SU2	0.594	0.613
ROUGE-SU1	0.598	0.619
ROUGE-ALL	0.523	0.527
ROUGE-2	0.553	0.560
ROUGE-3	0.468	0.461
ROUGE-4	0.366	0.357

Table 1: Average correlation coefficient with human scores (Spearman's  $\rho$ ) across different methods of computation for ROUGE based on  $n$ -grams of different lengths. The table shows the results for metrics computed based on ASR and manual transcriptions.

**The effect of text pre-processing** differed across the metrics: for metrics that relied on consecutive  $n$ -grams with  $n > 2$ , the removal of stop-words led to further drops in performance (Kruskal-Wallis test  $p = 4.4 \times 10^{-5}$ ). For ROUGE based on unigrams and skip-bigrams, counting only type frequencies led to a significant improvement in performance (Kruskal-Wallis test,  $p = 0.00017$ ). Correlations for the different types of pre-processing for the measures that performed the best are given in Table 2. Lemmatization did not make a significant difference to metric performance.

Pre-processing	ASR output	Manual
<i>All</i>	0.573	0.606
<i>Non-stop</i>	0.585	0.600
<i>Non-stop types</i>	0.601	0.617
<i>Types</i>	0.622	0.634

Table 2: Average correlation coefficient with human proficiency score (Spearman's  $\rho$ ) across ROUGE-1 and ROUGE-SU1-4 for different methods of text processing. The table shows the results for metrics computed based on ASR output and manual transcriptions.

Finally, a summarization metric performed better on tasks that required the test takers to summarize an announcement or lecture (average  $\bar{\rho} = 0.653$  for ROUGE-1 and ROUGE-SU1-4) rather than on tasks that required them to describe a picture sequence (average  $\bar{\rho} = 0.437$ , Mann-Whitney-Wilcoxon test, a non-parametric test for comparing two independent samples,  $p < 2.2 \times 10^{-16}$ ).

### 4.3 Evaluation of the final model

Analysis by prompt showed that the variants of ROUGE that included unigram counts (ROUGE and ROUGE-SU1-4) had the best correlations with human scores across all prompts. Further improvement in their performance was obtained by counting type frequencies only and by using several reference summaries. The optimal  $N$  references for these variants of ROUGE varied between prompts, but never exceeded four which was therefore selected as the optimal  $N$  references for this corpus.

Based on these results we computed ROUGE-1 metrics for all responses in the original training partition using four randomly selected, highly scored responses for each prompt and ‘types’ method of pre-processing. We then compared it with two baselines: (1) cosine distance (CVA) computed using type frequencies only and the same four references, and (2) naïve implementation of ROUGE-1 computed using one randomly selected reference summary and raw frequencies (tokens). The newly adjusted version of ROUGE-1 metrics performed significantly above the baselines (using Zou’s method for the comparison of overlapping correlations with confidence intervals constructed at  $\alpha = 0.001$ ). The correlation coefficients are shown in Table 3.

Metric	ASR output	Manual
New ROUGE-1	0.652	0.673
Base ROUGE-1	0.55	0.589
CVA	0.508	0.451

Table 3: Correlation coefficients with human scores (Spearman’s  $\rho$ ) for the entire training partition for the newly adjusted version of ROUGE and the baseline metrics. The table shows the results for metrics computed based on ASR and manual transcriptions.

We then trained a standard linear regression model using the human scores as the dependent variables and summarization metrics as indepen-

dent variables. The accuracy of prediction was evaluated using two metrics as suggested, for example, by Williamson et al. (2012): quadratic weighted kappa ( $\kappa$ ) and Pearson’s correlation coefficient ( $r$ ) between the observed and predicted scores. For computation of  $\kappa$ , the predicted scores were trimmed to the range of human scores and rounded to the nearest integer.

Repeated 10-fold cross-validation on the training partition showed that a model based on ROUGE-1 produced averages of  $\bar{r} = 0.65$  ( $\sigma = 0.031$ ) and  $\bar{\kappa} = 0.54$  ( $\sigma = 0.036$ ). The model based on a linear combination of several ROUGE variants using longer  $n$ -grams and a recursive feature elimination (Kuhn and Johnson, 2013, p. 480) did not show any improvement in the performance as compared to a model based on a single ROUGE-1.

Finally, we tested the performance of the metrics on an unseen test set that had not been used for any previous analyses. We tested both the model based solely on the content metric as well as on the performance of the content metrics in combination with 11 other features used for the automated scoring of spoken responses that measure pronunciation accuracy, prosody, fluency, and grammar. These results are presented in Table 4. Note that the performance of the content-only model based on the new ROUGE-1 was in line with the estimates obtained on the training set. Zou’s method for comparing overlapping correlations showed that in all cases, the difference between the model based on an adjusted ROUGE and the baselines was significant at  $\alpha = 0.001$ . In line with previous results, the models based on manual transcriptions showed better agreement with human scores than the models based on ASR output.

Table 4 shows that the addition of content metrics lead to relatively small increase in the performance of the integrated models. This is due to the fact that for most speakers different aspects of proficiency tend to be correlated. For example, more fluent speakers also achieve higher ROUGE scores (the correlation between ROUGE and pronunciation accuracy (Chen et al., 2009) is  $r = 0.62$ ). As a result, a model which measures only one aspect of performance such as fluency may sometimes reach near optimal performance and adding further predictors leads to a relatively small gain. When interpreting these results, it is important to bear in mind that empirical performance is only

Model	ASR		Manual	
	$r$	$\kappa$	$r$	$\kappa$
Content only				
CVA	0.492	0.340	0.469	0.303
Base ROUGE	0.587	0.440	0.632	0.489
New ROUGE	0.655	0.540	0.700	0.590
Integrated model				
No content	0.678	0.565	0.678	0.565
CVA	0.691	0.600	0.698	0.602
Base ROUGE	0.700	0.597	0.719	0.610
New ROUGE	0.715	0.617	0.738	0.652

Table 4: Performance of the linear regression model based on one content metric and an ‘integrated’ model based on 11 features that measure pronunciation, fluency, and grammar before and after the addition of ‘Base ROUGE,’ ‘CVA’ or ‘New ROUGE.’ The table shows the correlation coefficients (Pearson’s  $r$ ) and quadratic weighted kappa kappas ( $\kappa$ ) between the predicted scores and human ratings for the unseen test set. The agreement between the two expert raters on this dataset is  $\kappa = 0.69$ .

one aspect of evaluation of automated scoring systems. In addition to high agreement with human scores, operational automatic scoring systems also need to show good construct representation by covering different aspects of speaker performance (Williamson et al., 2012). This requirement ensures the validity of automated scores and prevents future test-takers from fine-tuning their performance to one particular feature measured by the scoring system. Therefore the addition of ROUGE to the automated scoring model serves both goals: it improves the agreement with human raters and also expands the construct coverage of the model.

## 5 Discussion

Summarization metrics can be successfully used to evaluate spoken summaries in the context of language assessment. Although the naïve implementation of ROUGE had good agreement with the scores assigned by human raters, several modifications led to a further increase in the performance.

Some of our findings show common patterns with what has previously been reported for written summaries. ROUGE-1, ROUGE-SU4 and ROUGE-2 are the three variants of ROUGE most commonly used for the evaluation of automatic text summaries. Our results showed that the first two

of these measures (ROUGE-1 and ROUGE-SU4) were also most suitable for content assessment of spoken responses. We note that both of these measures include unigram counts. More recently, Rankel et al. (2013) and Owczarzak et al. (2012) reported that metrics based on longer consecutive  $n$ -grams or linear combinations of different variants are more accurate. We did not find this for our data. Since our data represents abstractive summaries, poor performances of longer  $n$ -grams is not surprising. Finally, as in the case of written summaries, there was no effect of lemmatization while the removal of stop-words sometimes led to a decrease in performance.

Similar to written summaries, the use of more than one reference summary improved the performance. We found that the optimal number of reference summaries varied between prompts and metrics. For ROUGE-1, this number never exceeded four across all prompts in our corpus. Furthermore, we found that the choice of reference summaries from the pool of highly scored responses had no significant effect on the performance of the metric.

In addition to good agreement with human scores, metrics used for automated scoring also need to match the construct of interest, as defined by the assessment program (Williamson et al., 2012). The scoring guidelines for the tasks used in this paper ask raters to judge whether the key information contained in the prompt has been conveyed accurately. A notable difference between ROUGE and previously used metrics is that as a recall measure, ROUGE does not penalize for the lack of precision. Our results suggest that a recall-oriented approach has better agreement with human judgments than cosine distance which combines both precision and recall.

Recall-based approaches are sensitive to the length of candidate responses. In the case of automatic summary evaluation, the length of the summaries is limited to a predefined number of words. In this data, the length of the responses is limited more loosely by the time available to record the response and the actual number of words varied between the responses. Therefore, a recall-based approach may produce inflated scores by assigning higher metric values to a response which contains multiple repetitions of the same  $n$ -gram as long as the  $n$ -gram occurs several times in the reference response. The common occurrence of re-

pairs and repetitions in spoken speech further aggravates this problem further. We addressed this issue by only counting type frequencies, which also improved agreement with human judgments.

The adjusted metric had better agreement with human judgments than other “bag-of-words” approaches such as the cosine-based measure commonly used for content scoring that requires a much larger set of model responses than ROUGE. It also performed equally well on human and ASR transcriptions and did not require any manual annotation of the data. We also found that the performance of ROUGE depended on the task: we obtained better agreement for tasks that required the student to summarize a stimulus rather than tasks that required the student to describe a sequence of pictures. While in both cases the students produced short summary-like texts, the picture description task allowed for greater variability between the responses than the summarization task and, therefore, recall-oriented comparisons with highly-scored responses showed less agreement with human scores.

As a “bag-of-words” approach, ROUGE-1 has the same shortcomings as other methods discussed in Section 2.2 in that it doesn’t distinguish between syntactic roles. While variants based on longer  $n$ -grams could in theory address this, our results showed that neither a linear nor a geometric combination of these variants with ROUGE-1 improved agreement with human scores. This issue has also been acknowledged in the context of non-extractive text summarization and new metrics such as AutoSummEng (Giannakopoulos and Karkaletsis, 2011) have been developed to address it. Future research will include the conceptualization and development of metrics that can address the content accuracy of spoken summaries beyond the ‘bag-of-words’ approach.

## 6 Conclusion

In this paper we applied ROUGE, a recall-based metrics for evaluation of written summaries to the automatic assessment of spoken summaries produced by non-native speakers of English. We performed a thorough evaluation of different types of ROUGE by varying the length of  $n$ -grams, various methods of frequency computation, and text-preprocessing. We also explored the effect of the number of reference summaries. We found that the standard baseline implementation of ROUGE-1

computed over the output of the automated speech recognizer showed good agreement with expert ratings and performed better than the cosine similarity measure commonly used for the evaluation content of spoken responses. A further increase in agreement with human ratings could be achieved by using types instead of tokens for the frequency computation of both candidate and reference summaries. We also found that the use of several reference summaries improves the performance of the metric, but only four reference summaries were necessary to achieve reliable results.

## Acknowledgments

We would like to thank Keelan Evanini, Nitin Madnani, Xinhao Wang, Derrick Higgins and three anonymous reviewers for their helpful comments and suggestions and René Lawless for editing assistance.

## References

- Yigal Attali and Jill Burstein. 2006. Automated essay scoring with e-rater V. 2. *The Journal of Technology, Learning and Assessment*, 4(3):1–30.
- Thomas Baguley. 2012. *Serious Stats: A guide to advanced statistics for the behavioral sciences*. Palgrave Macmillan.
- Douglas Biber, Susan M. Conrad, Randi Reppen, Pat Byrd, Marie Helt, Victoria Clark, Viviana Cortes, Eniko Csomay, and Alfredo Urzua. 2004. *Representing language use in the university: analysis of TOEFL 2000 Spoken and Written academic language corpus*. Educational Testing Service, Princeton.
- 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 Building Educational Applications Using NLP*, pages 86–94, Stroudsburg, PA. Association for Computational Linguistics.
- Lei Chen, Klaus Zechner, and Xiaoming Xi. 2009. Improved pronunciation features for construct-driven assessment of non-native spontaneous speech. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, NAACL ’09*, pages 442–449, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Lei Chen. 2013. Applying unsupervised learning to support vector space model based speaking assessment. *Proceedings of the 8th Workshop on Innovative Use of NLP for Building Educational Applications, Atlanta, Georgia*, pages 58–62.

- Anthony C. Davison and David V. Hinkley. 1997. *Bootstrap Methods and their Application (Cambridge Series in Statistical and Probabilistic Mathematics)*. Cambridge University Press.
- Keelan Evanini and Xinhao Wang. 2013. Automated speech scoring for non-native middle school students with multiple task types. *Proceedings of Interspeech 2013, Lyon, France*, pages 2435–2439.
- Peter W. Foltz, Darrell Laham, and Thomas K. Landauer. 1999. Automated essay scoring: applications to educational technology. In B. Collis and R. Oliver, editors, *Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications 1999*, pages 939–944.
- George Giannakopoulos and Vangelis Karkaletsis. 2011. AutoSummENG and MeMoG in Evaluating Guided Summaries. In *TAC 2011 Workshop*, Gaithersburg, MD, USA. NIST.
- 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, April.
- Makoto Hirohata, Yousuke Shinnaka, Koji Iwano, and Sadao Furui. 2005. Sentence extraction-based presentation summarization techniques and evaluation metrics. In *Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on*, volume 1, pages 1065–1068.
- Max Kuhn and Kjell Johnson. 2013. *Applied Predictive Modeling*. Springer.
- Chin-Yew Lin and Marina Rey. 2004. ROUGE: A package for automatic evaluation of summaries. In Stan Szpakowicz Marie-Francine Moens, editor, *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Chin-Yew Lin. 2004. Looking for a few good metrics: Automatic summarization evaluation - how many samples are enough. In *Proceedings of the NTCIR Workshop*, pages 1765–1776, Tokyo.
- Annie Louis and A Nenkova. 2013. Automatically assessing machine summary content without a gold standard. *Computational Linguistics*, 39(2):267–300.
- Nitin Madnani, Jill Burstein, John Sabatini, and Tenaha O’Reilly. 2013. Automated scoring of a summary-writing task designed to measure reading comprehension. In *Proceedings of the 8th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 163–168, Atlanta, Georgia. Association for Computational Linguistics.
- Ani Nenkova and Annie Louis. 2008. Can you summarize this? Identifying correlates of input difficulty for generic multi-document summarization. In *Proceedings of the ACL-08: HLT*, pages 825–833. Association for Computational Linguistics.
- Ani Nenkova and Kathleen McKeown. 2011. Automatic summarization. *Foundations and Trends in Information Retrieval*, 5(2-3):103–233.
- Karolina Owczarzak, John M. Conroy, Hoa Trang Dang, and Ani Nenkova. 2012. An assessment of the accuracy of automatic evaluation in summarization. In *Proceedings of workshop on evaluation metrics and system comparison for automatic summarization.*, pages 1–9, Stroudsburg, PA. Association for Computational Linguistics.
- Art B. Owen. 2007. The pigeonhole bootstrap. *The Annals of Applied Statistics*, 1(2):386–411.
- Kishore Papineni, Salim Roukos, Todd Ward, and Weijing Zhu. 2002. BLEU : a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the ACL*, pages 311–318, Philadelphia, PA. Association for Computational Linguistics.
- Gerald Penn and Xiaodan Zhu. 2008. A Critical Reassessment of Evaluation Baselines for Speech Summarization. In *in Proceedings of RANLP workshop on Crossing Barriers in Text Summarization Research*, number June, pages 470–478, Columbus, Ohio, June. Association for Computational Linguistics.
- Peter A. Rankel, John. M. Conroy, Hoa Trang Dang, and Ani Nenkova. 2013. A decade of automatic content evaluation of news summaries: reassessing the state of the art. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, Sofia, Bulgaria, August 4-9, 2013*, pages 131–136, Sofia. Association for Computational Linguistics.
- Rand R. Wilcox. 2009. Comparing Pearson correlations: dealing with heteroscedasticity and nonnormality. *Communications in Statistics - Simulation and Computation*, 38(10):2220–2234.
- David M. Williamson, Xiaoming Xi, and F. Jay Breyer. 2012. A framework for evaluation and use of automated scoring. *Educational measurement: issues and practice*, 31(1):2–13.
- 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.
- Wenting Xiong, Keelan Evanini, Klaus Zechner, and Lei Chen. 2013. Automated content scoring of

spoken responses containing multiple parts with factual information. In Pierre Badin, Thomas Hueber, Gérard Bailly, Didier Demolin, and Françoise Raby, editors, *Proceedings of SLaTE 2013, Grenoble, France*, pages 137–142, Grenoble.

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.

Guang Yong Zou. 2007. Toward using confidence intervals to compare correlations. *Psychological methods*, 12(4):399–413.