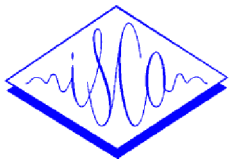
SLaTE 2013 - Grenoble, France - Proceedings

ISCA Archive
http://www.isca-speech.org/archive



Automated Content Scoring of Spoken Responses Containing Multiple Parts with Factual Information

Wenting Xiong[1], *Keelan Evanini*[2], *Klaus Zechner*[2], *Lei Chen*[2]

[1]Department of Computer Science, University of Pittsburgh, Pittsburgh, PA, USA

[2]Educational Testing Service, Princeton, NJ, USA

wex12@cs.pitt.edu,{kevanini,kzechner,lchen}@ets.org

Abstract

This paper presents approaches to automated content scoring of spoken language test responses from non-native speakers of English which contain multiple parts addressing factual information that the test taker has previously heard via auditory stimulus materials. While previous work relating to content scoring of spontaneous, unpredictable speech has focused only on entire responses and on general topic matching approaches, such as content vector analysis, the specific nature of spoken responses in our data requires response segmentation and extraction of features that indicate the relevance and correctness of the facts contained in the different parts of the response. Our best content features, based on similarity with key facts and concepts, achieve correlations of $r = 0.615$ (for speech recognition output) and $r = 0.637$ (using human transcriptions) with expert human rater scores. Furthermore, we show that these content features outperform traditional vector space based features. Finally, we demonstrate that the performance of a scoring model based on a combination of features developed previously and some of the newly designed content features improves significantly from $r = 0.624$ to $r = 0.664$ on an unseen evaluation set when using speech recognition output.

Index Terms: spoken language assessment, automated scoring, content appropriateness

1. Introduction

The research reported in this paper falls into the domain of automated scoring of non-native speech, and focuses on the scoring of the content of spoken responses in a test of English for non-native speakers. Determining the content accuracy of highly predictable speech (e.g. reading text aloud) is straightforward, in that one can use the output hypothesis of an automatic speech recognition (ASR) system as a proxy for what the speaker said, and then compute the edit distance to the stimulus passage or sentence to obtain an estimate for content correctness [1, 2]. Given that ASR systems for such highly predictable speech perform very well even for non-native speech (word error rates are generally below 10% [3]), these estimates of content accuracy are fairly reliable.

The situation is quite different, however, when processing spontaneous, highly unpredictable speech from non-native speakers. Typically, word error rates can range from 20-40% for this type of input. Furthermore, no particular text reference is available that can be used to compare the speaker's response to. For these reasons, most approaches to evaluating content correctness for spontaneous responses in spoken language tests have relied on methods originally proposed in the field of Information Retrieval, such as Content Vector Analysis (CVA) or Latent Semantic Analysis (LSA) [4]. In these approaches, the words in a spoken response are treated as an unordered list ("bag-of-words"), and comparisons between previously obtained training vectors and a spoken test response are made, e.g., by using the cosine similarity between weighted word vectors.

The test items used for this study are spontaneous in nature; however, the test taker is required to mention a set of specific facts in the response. Accordingly, generic approaches which regard one complete response as a single unit, such as CVA, may not be suitable, since responses for the test items in this study typically consist of multiple units (or text spans), each of which addresses a specific concept (content element) requested by the test item.

We therefore implement and evaluate several approaches that take into account the specific nature of these test items and their responses, in particular their localized content elements for specified concepts, to be able to assess the content accuracy more reliably and more specifically. Our aproach broadly consists of two main steps: (a) automated segmentation of the spoken test response into units of coherent content; and (b) computation of a set of content features based on scoring each text unit first separately and then comprehensively according to a set of content facts provided by test developers and further annotated by experts.

The remainder of this paper is organized as follows: Section 2 presents related work in the area of automated content scoring; Section 3 describes the data we use for this study; Section 4 provides details about our approach and the methods used in this research; Section 5 describes the experiments and evaluations we conducted; Section 6 discusses our findings; finally, Section 7 concludes the paper and provides an outlook into future work.

2. Related Work

There have been many previous studies about measuring the relevance and accuracy of content in the domains of automated assessment of essays and short textual responses. These efforts can be grouped into the following two sets. The first group relies on extracting patterns associated with the correct answers from the responses and matching them with pre-defined scoring rules [5, 6]. For example, the c-rater system described in [5] parses test-takers' responses and uses a pattern-matching algorithm to match the parsed constituents with manually written rules. The degree of match is used as a measure of content accuracy. To cope with the demand of manually generating these pre-defined patterns, [6] developed a bootstrapping method to generate patterns from a set of keywords and synonyms. Later, [7] compared several machine learning methods, e.g., decision trees and Bayesian learning, to the pattern matching method and

reported that the machine learning methods provide encouraging results.

The second type of method used for content scoring relies on a variety of text similarity measurements to compare a response with model responses [8]. Compared to the first group, such methods can bypass the labor intensive pattern-building step. A widely used approach to measuring text similarity between two text strings is to convert each text string into a vector of word counts and then use the angle between these two vectors as a similarity metric. For example, CVA has been successfully utilized to detect off-topic essays [9] and to provide content-related features for essay scoring [10]. For this group of methods, how best to measure the semantic similarity between two terms is a key question. A number of metrics have been proposed, including metrics derived from WordNet [11], a semantic ontology [12], and metrics related to the co-occurrence of terms in corpora or on the Web [13].

Recently, other novel NLP methods have been applied to the task of content scoring. For example, methods from the related NLP task of textual entailment [14], which attempts to find directional inference relations between two strings of text, have been applied to content scoring. [15] combined several graph alignment features with lexical semantic similarity measures using machine learning techniques and showed that answers can be more accurately scored in this way than by using semantic measures alone.

Compared to the research on content scoring for written text, there is only a small amount of research on scoring tests of spoken language based on content accuracy. In one example, [4] investigated using CVA, Pairwise Mutual Information (PMI), and Latent Semantic Analysis (LSA) to score spontaneous speech responses. In addition, [16] investigated the use of different semantic similarity measures, including PMI_IR from web queries, to score short spoken responses. The current study expands on these previous studies by introducing a new approach to segmenting the spoken response into discrete sections prior to assessing the content in each section and by introducing novel features to evaluate the accuracy of the content in the response.

3. Data

The data for this study was drawn from a pilot administration of the TOEFL® Junior™ Comprehensive test, an international assessment of English proficiency targeted at middle school students (aged from 11 to 15) which contains sections addressing the four components of English proficiency: Reading, Writing, Speaking, and Listening. The content-based Speaking test questions, or *items*, investigated in this study involve a task in which the test-taker first listened to an audio stimulus of a lecture or conversation containing several facts about a particular topic. The topics were drawn either from activities that are frequently done in middle school classes (e.g. writing a book report) or academic subjects appropriate for the targeted age range. While the test-taker is listening to the audio stimulus for each item, keywords are displayed on the computer screen to highlight the concepts that the test-taker will be asked to explain. After listening to the audio stimulus, the test-taker is given a fixed amount of time to prepare and then 60 seconds to provide a spoken response. The prompts emphasize that the response should contain information about each of the concepts highlighted during the presentation of the audio stimulus, and these keywords remain on the screen while the test-taker provides a spoken response.

During the course of developing the test material, additional resources were manually created for each item, including transcripts of the stimuli and "key points" (sample responses that contain the information that should be present in a high-scoring response). Each response was provided with a holistic score on a scale of 0 - 4 by expert human raters. The scoring rubrics addressed the following three main aspects of speaking proficiency: delivery (pronunciation, fluency, prosody), language use (grammar and lexical choice), and content.

The participants in this study were mostly students of middle school age residing in non-English speaking countries (average age = 13.1 years; s.d. = 2.2 years), and 15 different native language backgrounds were represented in the group. A total of 1700 participants are included in the study: 967 were used for the training corpus and 733 were used for the evaluation corpus. In the Speaking section, each participant responded to 3 content-based items, and a total of 6 different test forms were used. In this study, we focus only on a subset of 11 items that contain exactly four concepts (one for background information, which we refer to as *General*, plus three concrete fact-based concepts), and we exclude all responses that were flagged by raters as anomalous (e.g., because they contained a language other than English, the response was inaudible due to a technical difficulty, etc.). In total, 2048 spoken responses were available for training and 1568 were available for evaluation.

4. Methods

In our study we observed that the test-takers tended to organize the concepts in their responses in discrete segments of the response corresponding to the order in which they were discussed in the audio stimulus. Given this typical structure, we take an analytic approach to assessing the quality of the content contained in a given response: we propose to score a response's content with respect to each concept separately. This is accomplished by comparing the content of the response to pre-defined components that are expected from a proficient, on-topic response to each item (as described in Section 4.1) and an automatic segmentation of the transcriptions of the responses (as described in Section 4.2).

4.1. Item Content Analysis

To create a gold standard of the factual information that each item contains, we manually annotate the concepts for each item based on the relevant stimulus and response points provided by test developers. For each concept, we code its related factual information from four components: the *Name* of the concept (a keyword/phrase used in the prompt), the descriptive *Facts* (phrases that are likely to be included in a high quality response), the *Key Points* (a sample model response), and the *Context* (portions of the relevant stimulus that address this specific concept).

Take the *Frogs* item as an example.[1] After hearing a teacher present a lecture about the life cycle of frogs, the test-taker is presented with the following prompt: *"Talk about the physical changes a frog goes through. What happens at each stage? Be sure to include as many details as you can about each stage: tadpole, tadpole with legs, froglet, adult frog."* So, the student is expected to provide factual information about each of the four stages in a frog's life cycle that were discussed in the lecture, and these are the four concepts that constitute the item content

[1]The full content of this item can be viewed at *http://toefljr.caltesting.org/sampletest/s-frogs.html*.

Table 1: Item content analysis of the "Frogs" item.

Concept	Name	Facts
1	*General*	frog, physical changes, life, water, land, born, grows, moves
2	tadpole	first stage, water, little fish, tail, swim, gills, breath
3	tadpole with legs	second stage, little legs, like a frog, back, front
4	froglet	third stage, small frog, fully developed, tail, shorter, lungs, out of water, land
5	adult frog	last stage, adult, no tails, become, live on land, breathe air though lung

for this item. For a specific concept, e.g. "tadpole", the Name is *"tadpole"*; the Facts are *"first stage, water, little fish, tail, swim, gills, breath"*; and the Key Points regarding this concept are *"In the first stage, a frog is called a tadpole. A tadpole lives in water; it has a tail and breathes with gills."*; and the corresponding Context is *"In the first stage, when the frog is born in the water, it's called a tadpole. A tadpole looks a lot like a tiny little fish. Like a fish, it has a tail, and the tail is important because it helps it swim. It also has gills. That's another thing that a tadpole has that's like a fish. The gills are to help it breathe in the water. But this is just the first stage... The tadpole will go through more physical changes over the next few weeks."*.

The result of a complete concept analysis for this item is illustrated in Table1 for the *Name* and *Facts* content components. Due to space limitations, we do not present the *Key Points* and *Context* components.

4.2. Response Segmentation

We consider the segmentation problem within the context of our automated content scoring tasks as a pre-processing step on the transcripts. Before content feature extraction, we split responses into multiple self-contained text spans, each of which addresses a specific item concept (e.g. the concepts in Table 1). Specifically, we train a 1-order Hidden Markov Model[2] on manually transcribed and segmented responses to label each token of a response in terms of four labels (using indices from 1 to 4), and then split the response at the places where the labels of two successive tokens differ. Finally, each segment takes the same concept label as its tokens.

The automated segmentation model is trained on the features listed in Table 2;[3] we further tune the model for the best parameter settings on a held out sample set. In this study, three annotators segmented 625 human transcribed responses, which are used as the segmentation gold standard. For agreement analysis, all three annotators annotated a subset of 80 responses, in which $\kappa = .92$ on concept labels at the word level.

4.3. Content Feature Extraction

The proposed content scoring approach contains two steps for each response: first, we compute different features to measure the content information at the segment level using each of the four components of the item content analysis (as described in Section 4.1); second, we use various scoring functions to aggregate these segment-level content features to determine a score of content appropriateness for the entire response.

[2]We use the SVM representation of HMMs provided by SVM^{hmm} [17].

[3]To match a token with a n-gram fact, we compare all n-grams that contain that token against the fact to see if any of them match.

Table 2: Segmentation features

#	Description
1	Token index
4	Indicator of the presence of each concept name
4	Indicator of the presence of any fact from each concept
4	Signed distance to the closest occurrence of each concept name
4	Signed distance to the closest occurrence of any fact of each concept
32	Part-Of-Speech tags

4.3.1. Segment content features

We develop four segment content features (one for each component in the item content analysis) to measure the content information about concepts *c* contained in a segment, *s*, of a response, listed in order of increasing amount of information:

- $f_{name}(s, c)$: number of occurrences of the *Name* of concept *c* in the segment *s*.
- $f_{fact}(s, c) = f^{abs}_{fact}(s, c)/f^{abs}_{fact}(c_{points}, c)$: the scoring of the segment *s* with respect to the set of facts of concept *c* and then normalized by the score computed from the *Key Points* of *c*, where $f^{abs}_{fact}(s, c)$ computes the number of the *Facts* of *c* occurred in *s* weighted by the fact token length, plus the sum of the average unigram frequency of each *Fact*.
- $f_{points}(s, c)$: text to text similarity between *s* and the *Key Points* of *c* using WordNet.
- $f_{context}(s, c)$: text to text similarity between *s* and the *Context* of *c* using WordNet.

$f_{name}(s, c)$ captures whether a particular concept is mentioned in a response, while $f_{fact}(s, c)$ gives credit to relevant descriptive details including phrases that are matched exactly as well as their variations, thus adding more robustness to ASR errors in comparison to features based on exact string matching or syntactic features. In addition, we also consider traditional content scoring methods based on text to text similarity metrics derived from WordNet [18], where we use $f_{points}(s, c)$ and $f_{context}(s, c)$ to compare a segment against our "model response" and the concept's context, respectively.

4.3.2. Response content features

For aggregation, we first group the segments of a response by concept labels (denoted as $r(c_i)$), and compute the maximum feature value within each $r(c_i)$ for each concept c_j from each

component a:

$$h_a(i,j) = \max_{s \in r(c_i)} f_a(s, c_j) \quad (a \in \{name, concept, points, context\}) \tag{1}$$

For a given component a, we then summarize all $h_a(i,j)$ into a n by n matrix H_a, where n is the number of related item concepts.[4] Intuitively, H models how much information of concept c_j is covered by the part of the response $r(c_i)$ that is supposed to address concept c_i, as indicated by the value at position (i,j):

$$H_{n,n} = \begin{pmatrix} h(1,1) & h(1,2) & \cdots & h(1,n) \\ h(2,1) & h(2,2) & \cdots & h(2,n) \\ \vdots & \vdots & \ddots & \vdots \\ h(n,1) & h(n,2) & \cdots & h(n,n) \end{pmatrix}$$

We propose two scoring functions over H to aggregate the segment content features into response content features for any component. The first scoring function S_{mean} computes the mean of the sum of every column, which measures on average how each concept is addressed in the whole response.

$$A_{n,n} = \begin{pmatrix} 1 & 1 & 0 & \cdots & 0 \\ 1 & 1 & 1 & \cdots & 0 \\ 0 & 1 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}$$

In contrast, the second scoring function S_{matrix} (Formula 2) takes the segment difference into consideration: it constrains the scoring of the response's content for c_j to be based on the corresponding segments which address c_j, which are the numbers on the diagonal ($H(j,j)$). Considering possible segmentation errors and the transition from one concept to another in responses, we relax the constraint to also enable scoring on the segments of the preceding concept ($H(j-1,j)$) and the following one ($H(j+1,j)$), when they are available, in the process of scoring the response's content on c_j. This scoring matrix is denoted as A. In addition, we introduce a penalty component, which penalizes when a response regarding a specific concept carries more information about another concept (indicating that the response's content is inaccurate).[5] Finally, we denote the penalty matrix as $B-I$, where B is an indicator matrix of which concept is developed most in the segments of each concept, and I is an identity matrix. Ideally, $B = I$ and thus $B - I = 0$.

$$S_{matrix}(H) = H \cdot A - 0.5H \cdot (B - I) \tag{2}$$

We compute $S_{mean}(H)$ and $S_{matrix}(H)$ for each component to measure the response content quality for automated scoring. Note that our construction of the content features as described above can be easily adapted to items of any number of concepts, as long as responses can be segmented correspondingly.

5. Experiments

To evaluate our analytic-based content features, we conduct two intrinsic evaluations based on a sample set of three items (200 responses), and one extrinsic evaluation by means of a scoring model on all 11 items selected for analysis (see Section 3 for details). For the intrinsic evaluations, we used the responses from only the training set of the scoring model: we first investigate the utility of our content features, based on their correlation with scores provided by human experts across four content components between two scoring functions; we then compare our best content feature based on automated segmentation with the best CVA-based feature. For the extrinsic evaluation, we evaluate our best content features in the context of a scoring model containing both the proposed content features and features related to other components of a non-native speaker's proficiency.

5.1. Analysis of content features on human transcriptions

First we compare different content features on 200 manually transcribed responses for 3 items. We randomly select 92 responses to train the HMM segmentation model, and compute the content features $S_{mean}(H)$ and $S_{matrix}(H)$ on 108 testing responses based on various segmentation output: 1) human segmented results (S-manual), 2) HMM model predictions (S-auto) and 3) no segmentation, i.e., considering the whole response as one single segment (S-no). Note that we evaluate our segmentation model (S-auto) within our analysis of the utility of the content scoring features, as it is designed to be the pre-processing step performed prior to content feature extraction. For comparison, we visualize the performance of the content features computed from three segmentation models in a group, across 4 components for the two features $S_{mean}(H)$ (Figure 1)[6] and $S_{matrix}(H)$ (Figure 2).

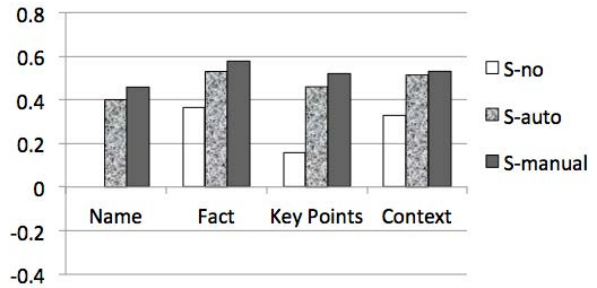


Figure 1: *Features' correlation with response scores using scoring method $S_{mean}(H)$. The segmentation models used for each component are S-no, S-auto and S-manual, from left to right.*

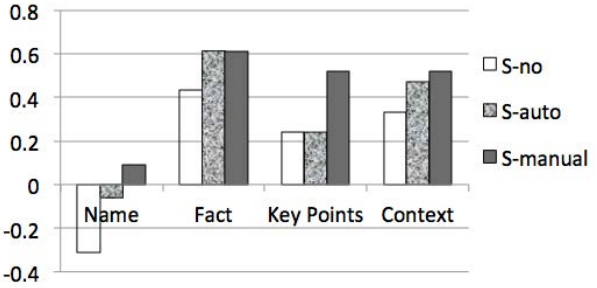


Figure 2: *Features' correlation with response scores using scoring method $S_{matrix}(H)$.*

[4] n=4 in this study.

[5] As a preliminary study, we set the coefficient of the penalty item (0.5) based on our intuition.

[6] The value of the $S_{matrix}(H_{name})$ feature is undefined for the S-no segmentation condition and is thus not included in Figure 1.

As Figure 1 and Figure 2 show, the *Fact* component yields the best performance regardless of the segmentation model and the scoring function. Although $S_{matrix}(H)$ results show a less consistent pattern across components (Figure 2), it yields better *Fact*-based content features compared to $S_{mean}(H_{fact})$ for all segmentation models. With respect to the impact of segmentation, in general, manual segmentation always works better than machine-based segmentation models, and segmentation improves the features' performance for all components. More importantly, for the best feature component, the performance of $S_{matrix}(H_{fact})$ generated by manual segmentation and machine segmentation are almost the same. This suggests that even though automated segmentation may introduce errors, we can still assess response content based on factual information quite reliably, by applying S_{matrix}. (In other words, $S_{matrix}(H_{fact})$ can work equally well when moderate segmentation errors exist.)

5.2. Human transcriptions vs. ASR output

To test whether our approach works equally well on ASR output, we compute the content features based on the automated segmentation of the ASR output of the sample testing set, which contains 3 items.[7] Here we retrain S-auto on all available gold-standards ($N = 625$).[8] Due to space limitations, we only present the results of $S_{mean}(H_{fact})$ and $S_{matrix}(H_{fact})$ – the two best-performing features identified in previous sections.

Table 3: Comparison of the *fact*-based content features between human transcriptions and ASR output.

	Trans_human		Trans_ASR	
	S_{mean}	S_{matrix}	S_{mean}	S_{matrix}
S-no	.367	.437	.403	.459
S-auto	.443	.637	.545	.615

5.3. Comparison with CVA

To validate the approach involving manual item content analysis, we compare our best content features computed using automated segmentation with the features generated by Content Vector Analysis (CVA). CVA was chosen as a baseline since it is a widely used alternative for text scoring and it only requires the response texts and the associated gold-standard scores for training.

In this experiment, we compute the two proposed content features and the CVA features from using human transcriptions from the sample data set (200 responses), in order to be able to make a fair comparison based on the upper bound of the features' performance. Furthermore, the CVA features are computed as leave-one-out evaluation, which includes the similarity between a given response and the model vector at each score level (e.g. $SimToScore\_1$), as well as the score level of the model vector that it is closest to. For a given response, the model vectors are computed on all the other responses of the specific item.[9]

[7]Similar patterns were observed when we extended the comparison to the other 8 items.

[8]The word-level accuracy of the updated segmentation model is 77% when trained and tested on all available gold-standards.

[9]Because the response is item dependent, we only consider the responses to the same item when computing the model vectors.

Table 4 compares the features' performance on the testing set. It shows that our best content feature ($S_{matrix}(H_{fact})$, $r = 0.612$) outperforms the best CVA feature ($SimToScore\_3$, $r = 0.440$) by 40% relative.

Table 4: Comparison between the best proposed content feature and the best CVA baseline feature.

	Proposed features	CVA features
Best r	.612	.440

5.4. Automated Scoring

Finally, we evaluate the proposed content features by examining their influence on a scoring model that is designed to predict holistic English speaking proficiency scores provided by expert raters. The raters provided a single discrete score for each response on a scale of 1 - 4, and were instructed to take into account detailed rubrics for the following components of speaking proficiency while providing each score: fluency, pronunciation, stress, intonation, grammar, word choice, and content. Based on a set of 2048 responses with scores, a linear regression model was trained with 8 features extracted using SpeechRater, an existing automated scoring system designed for spontaneous speech [19]. These features included measurements of a speaker's rate of speech (fluency), words per breath group (fluency), rate of long silences (fluency), acoustic model score (pronunciation), phone duration score (pronunciation), rate of stressed syllables (stress), rate of lexical types (fluency and word choice), and language model score (grammar and word choice). The model's correlation on the unseen set of 1568 responses (no speaker overlap) is $r = 0.624$.

To evaluate the contribution of the proposed content scoring approach to this model, the two top-performing content features, $S_{mean}(H_{fact})$ and $S_{matrix}(H_{fact})$, were added to the model. After the addition of these two features, the updated linear regression model obtained a correlation of $r = 0.664$ on the same evaluation set of 1568 responses, representing a statistically significant improvement of 0.04 over the baseline model with no content features ($t = 6.3$, $p < 0.001$).[10]

6. Discussion

In this paper, we propose a method for assessing the quality of the content of spoken responses based on segmenting the response and evaluating how the content contained in each of the segments relates to the test question at various levels of detail. Our experimental results show that the content features based on factual information relating to each concept perform best with both automated and manual concept-based segmentation, although the performance of the other features is typically within around 10% of the fact-based features. As described in Section 4.1, we define Facts to be short keywords or phrases that provide supporting details for the content related to each concept in a response. One drawback of this approach is that these facts must be manually extracted from the test questions prior to automated scoring; however, one can imagine that such information could be automatically constructed given the relevant resources using NLP techniques. In this paper we mainly focus on content feature engineering, and thus use these manually provided resources in our experiments.

[10]We use William's test of significance for dependent correlations.

Another important finding from this study is the fact that the proposed content features performed better after the segmentation procedure, and the automated segmentation model works comparable to a human gold standard in terms of their impact on $S_{matrix}(H_{fact})$ (our best content feature). This finding suggests a structured nature of the spoken responses which is induced by the way in which the test question is asked. Thus, the design of this speaking task may enable the use of a more targeted approach to content evaluation than has typically been employed in the past.

7. Conclusion

We have shown in this paper that it is feasible to evaluate the content of spontaneous spoken responses in a language test automatically, using features related to factual information in the responses and automated speech recognition to obtain transcriptions of test takers' spoken responses. While human experts are needed to generate the phrases containing factual information for each test item, this effort only needs to be undertaken once for each new test form, is not very time consuming, and could conceivably be done in a semi-automated manner in the future.

We also demonstrated that these content features are beneficial for developing automated scoring systems for spontaneous speech since they improve the agreement with expert human ratings and expand the aspects of speaking proficiency that are covered by the scoring system.

In future research, we plan to investigate methods of automatically extracting the Facts contained in the stimulus materials and extending the approach to additional types of structured, content-based spoken test responses, such as narrative retellings.

8. References

[1] P. Price, J. Tepperman, M. Iseli, T. Duong, M. Black, S. Wang, C. Boscardin, M. Heritage, P. David Pearson, S. Narayanan *et al.*, "Assessment of emerging reading skills in young native speakers and language learners," *Speech Communication*, vol. 51, no. 10, pp. 968–984, 2009.

[2] M. Eskenazi, "An overview of spoken language technology for education," *Speech Communication*, vol. 51, no. 10, pp. 832–844, 2009.

[3] J. Balogh, J. Bernstein, J. Cheng, and B. Townshend, "Automatic evaluation of reading accuracy: assessing machine scores," in *Proceedings of the ISCA Tutorial and Research Workshop on Speech and Language Technology in Education (SLaTE)*, 2007, pp. 1–3.

[4] S. Xie, K. Evanini, and K. Zechner, "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*. Montréal, Canada: Association for Computational Linguistics, 2012, pp. 103–111.

[5] C. Leacock and M. Chodorow, "C-rater: Automated scoring of short-answer questions," *Computers and the Humanities*, vol. 37, no. 4, pp. 385–405, 2003.

[6] J. Z. Sukkarieh, S. Pulman, and N. Raikes, "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*, 2004.

[7] S. Pulman and J. Z. Sukkarieh, "Automatic short answer marking," in *Proceedings of the 3rd NAACL-HLT Workshop on Innovative Use of NLP for Building Educational Applications*, 2005, pp. 9–16.

[8] M. Mohler and R. Mihalcea, "Text-to-text semantic similarity for automatic short answer grading," in *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, 2009, pp. 567–575.

[9] D. Higgins, J. Burstein, and Y. Attali, "Identifying off-topic student essays without topic-specific training data," *Natural Language Engineering*, vol. 12, 2006.

[10] Y. Attali and J. Burstein, "Automated essay scoring with e-rater v.2.0." in *Presented at the Annual Meeting of the International Association for Educational Assessment*, 2004.

[11] C. Corley and R. Mihalcea, "Measuring the semantic similarity of texts," in *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, 2005, pp. 13–18.

[12] C. Fellbaum, *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press, 1998.

[13] P. D. Turney, "Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL," in *Proceedings of the Twelfth European Conference on Machine Learning (ECML)*, Freiburg, Germany, 2001, pp. 491–502.

[14] I. Dagan, O. Glickman, and B. Magnini, "The PASCAL recognising textual entailment challenge," in *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, J. Quiñonero-Candela, I. Dagan, B. Magnini, and F. d'Alché Buc, Eds. Springer, 2006, pp. 177–190.

[15] M. Mohler, R. Bunescu, and R. Mihalcea, "Learning to grade short answer questions using semantic similarity measures and dependency graph alignments," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 2011, pp. 752–762.

[16] F. Huang and L. Chen, "Scoring spoken responses based on content accuracy," in *Proceedings of the 7th Workshop on Innovative Use of NLP for Building Educational Applications, NAACL-HLT*. Montréal, Canada: Association for Computational Linguistics, 2012.

[17] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun, "Large margin methods for structured and interdependent output variables," *Journal of Machine Learning Research*, vol. 6, no. 2, p. 1453, 2006.

[18] C. Leacock and M. Chodorow, "Combining local context and wordnet similarity for word sense identification," in *WordNet: An Electronic Lexical Database*, C. Fellbaum, Ed. MIT Press, 1998, pp. 305–332.

[19] K. Zechner, D. Higgins, X. Xi, and D. M. Williamson, "Automatic scoring of non-native spontaneous speech in tests of spoken English," *Speech Communication*, vol. 51, no. 10, pp. 883–895, 2009.