Coherence Modeling for the Automated Assessment of Spontaneous Spoken Responses

Xinhao Wang, Keelan Evanini, Klaus Zechner
Educational Testing Service
660 Rosedale Road
Princeton, NJ 08541, USA
xwang002,kevanini,kzechner@ets.org

Abstract

This study focuses on modeling discourse coherence in the context of automated assessment of spontaneous speech from non-native speakers. Discourse coherence has always been used as a key metric in human scoring rubrics for various assessments of spoken language. However, very little research has been done to assess a speaker's coherence in automated speech scoring systems. To address this, we present a corpus of spoken responses that has been annotated for discourse coherence quality. Then, we investigate the use of several features originally developed for essays to model coherence in spoken responses. An analysis on the annotated corpus shows that the prediction accuracy for human holistic scores of an automated speech scoring system can be improved by around 10% relative after the addition of the coherence features. Further experiments indicate that a weighted F-Measure of 73% can be achieved for the automated prediction of the coherence scores.

1 Introduction

In recent years, much research has been conducted into developing automated assessment systems to automatically score spontaneous speech from non-native speakers with the goals of reducing the burden on human raters, improving reliability, and generating feedback that can be used by language learners. Various features related to different aspects of speaking proficiency have been exploited, such as delivery features for pronunciation, prosody, and fluency (Strik and Cucchiarini, 1999; Chen et al., 2009; Cheng, 2011; Higgins et al., 2011), as well as language use features for vocabulary and grammar, and content features (Chen and Zechner, 2011; Xie et al., 2012). However, discourse-level features related to topic development have rarely been investigated in the context of automated speech scoring. This is despite the fact that an important criterion in the human scoring rubrics for speaking assessments is the evaluation of coherence, which refers to the conceptual relations between different units within a response.

Methods for automatically assessing discourse coherence in text documents have been widely studied in the context of applications such as natural language generation, document summarization, and assessment of text readability. For example, Foltz et al. (1998) measured the overall coherence of a text by utilizing Latent Semantic Analysis (LSA) to calculate the semantic relatedness between adjacent sentences. Barzilay and Lee (2004) introduced an HMM-based model for the document-level analysis of topics and topic transitions. Barzilay and Lapata (2005; 2008) presented an approach to coherence modeling which focused on the entities in the text and their grammatical transitions between adjacent sentences, and calculated the entity transition probabilities on the document level. Pitler et al. (2010) provided a summary of the performance of several different types of features for automated coherence evaluation, such as cohesive devices, adjacent sentence similarity, Coh-Metrix (Graesser et al., 2004), word co-occurrence patterns, and entity-grid.

In addition to studies on well-formed text, researchers have also addressed coherence modeling on text produced by language learners, which may contain many spelling and grammar errors. Utilizing LSA and Random Indexing methods, Higgins et al. (2004) measured the global

Proceedings of NAACL-HLT 2013, pages 814–819,
Atlanta, Georgia, 9–14 June 2013.

coherence of students' essays by calculating the semantic relatedness between sentences and the corresponding prompts. In addition, Burstein et. al (2010) combined entity-grid features with writing quality features produced by an automated assessment system of essays to predict the coherence scores of student essays. Recently, Yannakoudakis and Briscoe (2012) systematically analyzed a variety of coherence modeling methods within the framework of an automated assessment system for non-native free text responses and indicated that features based on Incremental Semantic Analysis (ISA), local histograms of words, the part-of-speech IBM model, and word length were the most effective.

In contrast to these previous studies involving well-formed text or learner text containing errors, this paper focuses on modeling coherence in spontaneous spoken responses as well as investigating discourse features in an attempt to extend the construct coverage of an automated speech scoring system. In a related study, Hassanali et al. (2012) investigated coherence modeling for spoken language in the context of a story retelling task for the automated diagnosis of children with language impairment. They annotated transcriptions of children's narratives with coherence scores as well as markers of narrative structure and narrative quality; furthermore they built models to predict the coherence scores based on Coh-Metrix features and the manually annotated narrative features. The current study differs from this one in that it deals with free spontaneous spoken responses provided by students at a university level; these responses therefore contain more varied and more complicated information than the child narratives.

The main contributions of this paper can be summarized as follows: First, we obtained coherence annotations on a corpus of spontaneous spoken responses drawn from a university-level English language proficiency assessment, and demonstrated an improvement of around 10% relative in the accuracy of the automated prediction of human holistic scores with the addition of the coherence annotations. Second, we applied the entity-grid features and writing quality features from an automated essay scoring system to predict the coherence scores; the experimental results have shown promising correlations between some of these features and the coherence scores.

2 Data and Annotation

2.1 Data

For this study, we collected 600 spoken responses from the international TOEFL® iBT assessment of English proficiency for non-native speakers. 100 responses were drawn from each of 6 different test questions comprising two different speaking tasks: 1) providing an opinion based on personal experience (N = 200) and 2) summarizing or discussing material provided in a reading and/or listening passage (N = 400). The spoken responses were all transcribed by humans with punctuation and capitalization. The average number of words contained in the responses was 104.4 (st. dev. = 34.4) and the average number of sentences was 5.5 (st. dev. = 2.1).

The spoken responses were all provided with holistic English proficiency scores on a scale of 1 - 4 by expert human raters in the context of operational, high-stakes scoring for the spoken language assessment. The scoring rubrics address the following three main aspects of speaking proficiency: delivery (pronunciation, fluency, prosody), language use (grammar and lexical choice), and topic development (content and coherence). In order to ensure a sufficient quantity of responses from each proficiency level for training and evaluating the coherence prediction features, the spoken responses selected for this study were balanced based on the human scores as follows: 25 responses were selected randomly from each of the 4 score points (1 - 4) for each of the 6 test questions. In some cases, more than one response was selected from a given test-taker; in total, 471 distinct test-takers are represented in the data set.

2.2 Annotation and Analysis

The coherence annotation guidelines used for the spoken responses in this study were modified based on the annotation guidelines developed for written essays described in Burstein et al. (2010). According to these guidelines, expert annotators provided each response with a score on a scale of 1 - 3. The three score points were defined as follows: 3 = highly coherent (contains no instances of confusing arguments or examples), 2 = somewhat coherent (contains some awkward points in which the speaker's line of argument is unclear), 1 = barely

coherent (the entire response was confusing and hard to follow; it was intuitively incoherent as a whole and the annotators had difficulties in identifying specific weak points). For responses receiving a coherence score of 2, the annotators were required to highlight the specific awkward points in the response. In addition, the annotators were specifically required to ignore disfluencies and grammatical errors as much as possible; thus, they were instructed to not label sentences or clauses as awkward points solely because of the presence of disfluent or ungrammatical speech.

Two annotators (not drawn from the pool of expert human raters who provided the holistic scores) made independent coherence annotations for all 600 spoken responses. The distribution of annotations across the three score points is presented in Table 1. The two annotators achieved a moderate inter-annotator agreement (Landis and Koch, 1977) of $\kappa = 0.68$ on the 3-point scale. The average of the two coherence scores provided by the two annotators correlates with the holistic speaking proficiency scores at $r = 0.66$, indicating that the overall proficiency scores of spoken responses can benefit from the discourse coherence annotations.

	1	2	3
# 1	160 (27%)	278 (46%)	162 (27%)
# 2	125 (21%)	251 (42%)	224 (37%)

Table 1. Distribution of coherence annotations from two annotators

Furthermore, coherence features based on the human annotations were examined within the context of an automated spoken language assessment system, SpeechRater[SM] (Zechner et al., 2007; 2009). We extracted 96 features related to pronunciation, prosody, fluency, language use, and content development using SpeechRater. These features were either extracted directly from the speech signal or were based on the output of an automatic speech recognition system (with a word error rate of around 28%[1]). By utilizing a decision tree classifier (the J48 implementation from Weka (Hall et al., 2009)), 4-fold cross validation was conducted on the 600 responses to train and evaluate a scoring model for predicting the holistic proficiency scores. The resulting correlation between the predicted scores (based on the 96 baseline SpeechRater features) and the human holistic proficiency scores was $r = 0.667$.

In order to model a spoken response's coherence, three different features were extracted from the human annotations. Firstly, the average of the two annotators' coherence scores was directly used as a feature with a 5-point scale (henceforth Coh_5). Secondly, following the work in Burstein et al. (2010), we collapsed the average coherence scores into a 2-point scale to deal with the difficulty in distinguishing somewhat and highly coherent responses. For this second feature (henceforth Coh_2), scores 1 and 1.5 were mapped to score 1, and scores 2, 2.5, and 3 were mapped to score 2. Finally, the number of awkward points was also counted as a feature (henceforth Awk). As shown in Table 2, when these three coherence features were combined separately with the SpeechRater features, the correlations could be improved from $r = 0.667$ to $r > 0.7$. Meanwhile, the accuracy (i.e., the percentage of correctly predicted holistic scores) could be improved from 0.487 to a range between 0.535 and 0.543.

Features	r	Accuracy
SpeechRater	0.667	0.487
SpeechRater+Coh_5	0.714	0.540
SpeechRater+Coh_2	0.705	0.543
SpeechRater+Awk	0.702	0.535
SpeechRater+Coh_5+Awk	0.703	0.537
SpeechRater+Coh_2+Awk	0.701	0.542

Table 2. Improvement to an automated speech scoring system after the addition of human-assigned coherence scores and measures, showing both Pearson r correlations and the ratio of correctly matched holistic scores between the system and human experts

These experimental results demonstrate that the automatic scoring system can benefit from coherence modeling either by directly using a human-assigned coherence score or the identified awkward points. However, the use of both kinds of annotations does not provide further improvement. When collapsing the average scores into a 2-point scale, there was a 0.009 correlation drop (not statistically significant), but the accuracy was slightly improved. In addition, due to the relatively small

[1] Both the training and evaluation sets used to develop the speech recognizer consist of similar spoken responses drawn from the same assessment. However, there is no response overlap between these sets and the corpus used for discourse coherence annotation in this study.

size of the set of available coherence annotations, we adopted the collapsed 2-point scale instead of the 5-point scale for the coherence prediction experiments in the next section.

2.3 Experimental Design

As demonstrated in Section 2.2, the collapsed average coherence score can be used to improve the performance of an automated speech scoring system. Therefore, this study treats coherence prediction as a binary classification task: low-coherent vs. high-coherent, where the low-coherent responses are those with average scores 1 and 1.5, and the high-coherent responses are those with average scores 2, 2.5, and 3.

For coherence modeling, we again use the J48 decision tree from the Weka machine learning toolkit (Hall et al., 2009) and run 4-fold cross-validation on the 600 annotated responses. The correlation coefficient (r) and the weighted average F-Measure[2] are used as evaluation metrics.

In this experiment, we examine the performance of the entity-grid features and a set of features produced by the e-rater® system (an automated writing assessment system for learner essays) (Attali and Burstein, 2006) to predict the coherence scores of the spontaneous spoken responses, where all the features are extracted from human transcriptions of the responses.

2.4 Entity Grid and e-rater Features

First, we applied the algorithm from Barzilay and Lapata (2008) to extract entity-grid features, which calculated the vector of entity transition probabilities across adjacent sentences. Several different methods of representing the entities can be used before generating the entity-grid. First, all the entities can be described by their syntactic roles including S (Subject), O (Object), and X (Other). Alternatively, these roles can also be reduced to P (Present) or N (Absent). Furthermore, entities can be defined as salient, when they appear two or more times, otherwise as non-salient. In this study, we generated there basic entity grids: EG_SOX (entity grid with the syntactic roles S, O, and X), EG_REDUCED (entity grid with the reduced representations P and N), and EG_SALIENT (entity grid with salient and non-salient entities). In addition to these entity-grid features, we also used 130 writing quality features related to grammar, usage, mechanics, and style from e-rater to model the coherence.

A baseline system for this task would simply assign the majority class (high-coherent) to all of the responses; this baseline achieves an F-Measure of 0.587. Table 3 shows that the EG_REDUCED and e-rater features can obtain F-Measures of 0.677 and 0.726 as well as correlations with human scores of 0.20 and 0.33, respectively. However, the combination of the two sets of features only brings a very small improvement (from 0.33 to 0.34). In addition, our experiments show that by introducing the component of co-reference resolution for entity grid building, we can only get a very slight improvement on EG_SALIENT, but no improvement on EG_SOX and EG_REDUCED. That may be because it is generally more difficult to parse the transcriptions of spoken language than well-formed text, and more errors are introduced during the process of co-reference resolution.

	r	F-Measure
Baseline	0.0	0.587
EG_SOX	0.16	0.664
EG_REDUCED	**0.2**	**0.677**
EG_SALIENT	0.2	0.678
e-rater	**0.33**	**0.726**
EG_SOX +e-rater	0.30	0.714
EG_REDUCED +e-rater	**0.34**	**0.73**
EG_SALIENT + e-rater	0.26	0.695

Table 3. Performance of entity grid and e-rater features on the coherence modeling task

2.5 Discussion and Future Work

In order to further analyze these features, the correlation coefficients between various features and the average coherence scores (on a five-point scale) were calculated; Figure 1 shows the histogram of these correlation values. As the figure shows, there are a total of approximately 50 features with correlations larger than 0.1. Four of the entity-grid features have correlations between 0.15 and 0.29. As for the writing quality features, some

[2] The data distribution in the experimental corpus is unbalanced: 71% of the responses are high-coherent and 29% are low-coherent. Therefore, we adopt the weighted average F-Measure to evaluate the performance of coherence prediction: first, the F1-Measure of each category is calculated, and then the percentages of responses in each category are used as weights to obtain the final weighted average F-Measure.

of them show high correlations with the average coherence scores, despite the fact that they are not explicitly related to discourse coherence, such as the number of good lexical collocations.

Based on the above analysis, we plan to investigate additional superficial features explicitly related to discourse coherence, such as the distribution of conjunctions, pronouns, and discourse connectives. Moreover, based on the research on well-formed texts and learner essays, we will attempt to examine more effective features and models to better cover the discourse aspects of spontaneous speech. For example, local semantic features related to inter-sentential coherence and the ISA feature will be investigated on spoken responses. In addition, we will apply the features and build coherence models using the output of automatic speech recognition in addition to human transcriptions. Finally, various coherence features or models will be integrated into a practical automated scoring system, and further experiments will be performed to measure their effect on the performance of automated assessment of spontaneous spoken responses.

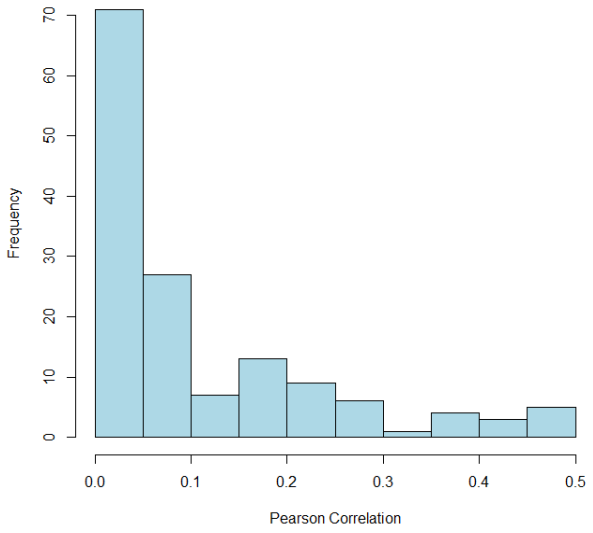

Figure1. Histogram of entity-grid and writing quality features based on their correlations with coherence scores

3 Conclusion

In this paper, we present a corpus of coherence annotations for spontaneous spoken responses provided in the context of an English speaking proficiency assessment. Entity-grid features and features from an automated essay scoring system were examined for coherence modeling of spoken responses. The analysis on the annotated corpus showed promising results for improving the performance of an automated scoring system by means of modeling the coherence of spoken responses.

Acknowledgments

The authors wish to express our thanks to the discourse annotators Melissa Lopez and Matt Mulholland for their dedicated work and our colleagues Jill Burstein and Slava Andreyev for their support in generating entity-grid features.

References

Yigal Attali and Jill Burstein. 2006. Automated essay scoring with e-rater® V.2.0. *Journal of Technology, Learning, and Assessment*. 4(3): 159-174.

Regina Barzilay and Lillian Lee. 2004. Catching the drift: Probabilistic content models, with applications to generation and summarization. *Proceedings of NAACL-HLT*, 113-120.

Regina Barzilay and Mirella Lapata. 2005. Modeling local coherence: An entity-based approach. *Proceedings of ACL*, 141-148.

Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, *34*(1):1-34.

Jill Burstein, Joel Tetreault and Slava Andreyev. 2010. Using entity-based features to model coherence in student essays. *Proceedings of NAACL-HLT*, 681-684. Los Angeles, California.

Lei Chen, Klaus Zechner and Xiaoming Xi. 2009. Improved pronunciation features for construct-driven assessment of non-native spontaneous speech. *Proceedings of NAACL-HLT*, 442-449.

Miao Chen and Klaus Zechner. 2011. Computing and evaluating syntactic complexity features for automated scoring of spontaneous non-native speech. *Proceedings of ACL*, 722-731.

Jian Cheng. 2011. Automatic assessment of prosody in high-stakes English tests. *Proceedings of Interspeech* , 27-31.

Peter W. Foltz, Walter Kintsch and Thomas K. Landauer. 1998. The measurement of textual coherence with Latent Semantic Analysis. *Discourse Processes*, 25(2&3):285-307.

Arthur C. Graesser, Danielle S. McNamara, Max M. Louwerse and Zhiqiang Cai. 2004. Coh-Metrix: Analysis of text on cohesion and language. *Behavior*

Research Methods, Instruments, & Computers, 36(2):193-202.

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

Khairun-nisa Hassanali, Yang Liu and Thamar Solorio. 2012. Coherence in child language narratives: A case study of annotation and automatic prediction of coherence. *Proceedings of the Interspeech Workshop on Child, Computer and Interaction.*

Derrick Higgins, Jill Burstein, Daniel Marcu and Claudia Gentile. 2004. Evaluating multiple aspects of coherence in student essays. *Proceedings of NAACL-HLT*, 185-192.

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

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics, 33*(1):159-174.

Emily Pitler, Annie Louis and Ani Nenkova. 2010. Automatic evaluation of linguistic quality in multi-document summarization. *Proceedings of ACL*. 544–554. Uppsala.

Helmer Strik and Catia Cucchiarini. 1999. Automatic assessment of second language learners' fluency. *Proceedings of the 14th International Congress of Phonetic Sciences*, 759-762. Berkeley, CA.

Shasha Xie, Keelan Evanini and Klaus Zechner. 2012. Exploring content features for automated speech scoring. *Proceedings of NAACL-HLT*, 103-111.

Helen Yannakoudakis and Ted Briscoe. 2012. Modeling coherence in ESOL learner texts. *Proceedings of the 7th Workshop on the Innovative Use of NLP for Building Educational Applications*, 33-43. Montreal.

Klaus Zechner, Derrick Higgins and Xiaoming Xi. 2007. SpeechRater^{SM}: A construct-driven approach to scoring spontaneous non-native speech. *Proceedings of the International Speech Communication Association Special Interest Group on Speech and Language Technology in Education*, 128-131.

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.