# Automated Scoring of Picture-based Story Narration

**Swapna Somasundaran**[1], **Chong Min Lee**[1], **Martin Chodorow**[2] and **Xinhao Wang**[1]
[1]Educational Testing Service, 660 Rosedale Road, Princeton, NJ 08541, USA
[2]Hunter College and the Graduate Center, CUNY, New York, NY 10065, USA
{ssomasundaran,clee001,xwang002}@ets.org
martin.chodorow@hunter.cuny.edu

## Abstract

This work investigates linguistically motivated features for automatically scoring a spoken picture-based narration task. Specifically, we build scoring models with features for story development, language use and task relevance of the response. Results show that combinations of these features outperform a baseline system that uses state of the art speech-based features, and that best results are obtained by combining the linguistic and speech features.

## 1 Introduction

Story-telling has been used in evaluating the development of language skills (Sun and Nippold, 2012; McKeough and Malcolm, 2011; Botvin and Sutton-Smith, 1977). It has also been incorporated into assessment of English language proficiency in tests such as ETS's TOEFL Junior Comprehensive Test[1], where English language skills of non-native middle-school students are tested on a task designed to elicit stories based on pictures. The Six-Picture Narration task presents a series of six pictures (similar to a comic strip) to the test taker, who must orally produce a story which incorporates the events depicted in the pictures. As the scoring guide[2] for this task indicates, in addition to fluidity of speech and few pronunciation errors, high scoring responses must also show good command of language conventions, including grammar and word usage, and must also be relevant to the task.

Previous work (Evanini and Wang, 2013) explored automated assessment of the speech component of the spoken responses to the picture narration task, but the linguistic and narrative aspects of the response have not received much attention. In this work, we investigate linguistic and construct-relevant aspects of the test such as (1) relevance and completeness of the content of the responses with respect to the prompt pictures, (2) proper word usage (3) use of narrative techniques such as detailing to enhance the story, and (4) sequencing strategies to build a coherent story.

The contribution of this work is three-fold. First, we improve the construct coverage of the automated scoring models by incorporating evaluation of elements prescribed in the scoring rubric. Second, our linguistically motivated features allow for clear interpretation and explanation of scores, which is especially important if the automated scoring is to be employed for educational purposes. Finally, our results are promising – we show that the combination of linguistic and construct-relevant features which we explore in this work outperforms the state of the art baseline system, and that the best performance is obtained when the linguistic and construct-relevant features are combined with the speech features.

## 2 Related Work

Evanini et al. (2013; 2014) use features extracted mainly from speech for scoring the picture narration task. They employ measures capturing fluency,

[1] Details of the task and sample can be found at https://toefljr.caltesting.org/sampleQuestions/TOEFLJr/s-movietheater.html

[2] https://www.ets.org/s/toefl_junior/pdf/toefl_junior_comprehensive_speaking_scoring_guides.pdf

*Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications, 2015*, pages 42–48,
Denver, Colorado, June 4, 2015. ©2015 Association for Computational Linguistics

prosody and pronunciation. Our work explores the other (complementary) dimensions of the test such as language use, content relevance and story development.

Somasundaran and Chodorow (2014) construct features for awkward word usage and content relevance for a written vocabulary test which we adapt for our task. Discourse organization features have been employed for essay scoring of written essays in the expository and argumentative genre (Attali and Burstein, 2006). Our discourse features are focused on the structure of spoken narratives. Our relevance measure is intended to capture topicality while providing leeway for creative story telling, which is different from scoring summaries (Loukina et al., 2014). King and Dickinson (2013) use dependency parses of written picture descriptions. Given that our data is automatically recognized speech, parse features are not likely to be reliable. We use measures of n-gram association, such as pointwise mutual information (PMI), that have a long history of use for detecting collocations and measuring their quality (see Manning and Schütze (1999) and Leacock et al. (2014) for reviews). Our application of a large n-gram database and PMI is to encode language proficiency in sentence construction without using a parser.

Picture description tasks have been employed in a number of areas of study ranging from second language acquisition to Alzheimer's disease (Ellis, 2000; Forbes-McKay and Venneri, 2005). Picture-based story narration has also been used to study referring expressions (Lee et al., 2012) and to analyze child narratives (Hassanali et al., 2013).

## 3 Data

The TOEFL Junior Comprehensive assessment is a computer-based test intended for middle school students around the ages of 11 - 15, and is designed to assess a student's English communication skills. As mentioned above, we focus on the Six-Picture Narration task. Human expert raters listen to the recorded responses, which are about 60 seconds in duration, and assign a score to each on a scale of 1 - 4, with score point 4 indicating an excellent response. In this work, we use the automatic speech recognition (ASR) output transcription of the responses (see (Evanini and Wang, 2013) for details).

|  | Total | —Score Distribution— | | | |
|---|---|---|---|---|---|
|  |  | 1 | 2 | 3 | 4 |
| Train | 877 | 142 | 401 | 252 | 82 |
| Eval | 674 | 132 | 304 | 177 | 61 |

Table 1: Number of responses and score distributions for training and evaluation datasets.

The data consists of 3440 responses to 6 prompts, all of which were scored by human raters. Table 1 shows the data size and partitions for the experiments as well as the score distributions. An ASR partition (with 1538 responses) was created and used for training the speech recognition models and was used also for our linguistic feature development. *Train* was used for cross validation experiments as well as for training a final model that was evaluated on *Eval* evaluation dataset. Quadratic Weighted Kappa (QWK) between human raters for Train is 0.69 and for Eval is 0.70. Responses containing anomalous test taker behavior (such as non-English responses or non-responses) and responses with severe technical difficulties (such as static or background noise) receive separate ratings and are excluded from this study. This filtering resulted in a total of 874 responses in Train and 672 responses in Eval data sets.

## 4 Features

We explore five different feature sets to help us answer the following questions about the response: Did the test taker construct a story about the pictures in the prompt (or did he/she produce an irrelevant response instead?) (*Relevance*); Did the test taker use words appropriately in the response? Proper usage of words and phrases is characterized by the probabilities of the contexts in which they are used (*Collocation*); Did the test taker adequately organize the narrative? (*Discourse*); Did the test taker enhance the narrative by including details (*Detailing*); and Did the test taker develop the story through expression of emotion and character development? (*Sentiment*)

### 4.1 Relevance

In order to test if a given response tells a story that is relevant to the pictures in the prompt, we calculate the overlap of the content of the response and the content of the pictures similar to (Somasundaran and Chodorow, 2014). To facilitate this, each prompt is associated with a reference corpus containing a detailed description of each picture, and also an overall narrative that ties together the events in the pictures. Each reference corpus was created by merging the picture descriptions and narratives that were generated independently by 10 annotators.[3] To calculate overlap, stop words were first removed from lemmatized versions of the response and the reference corpus.

Because test-takers often use synonyms and other words related to the prompt, we expanded the content words in the reference corpus by adding their synonyms, as provided in Lin's thesaurus (Lin, 1998) and in WordNet, and also included their WordNet hypernyms and hyponyms. This gave us the following 6 features which measure the overlap, or coverage, between the lemmatized response and the lemmatized (i) reference corpus (*lemmas*), (ii) reference corpus expanded using Lin's thesaurus (*cov-lin*), (iii) reference corpus expanded using WordNet Synonyms (*cov-wn-syns*), (iv) reference corpus expanded using WordNet Hypernyms (*cov-wn-hyper*), (v) reference corpus expanded using WordNet Hyponyms (*cov-wn-hypo*), and (vi) reference corpus expanded using all of the above methods (*cov-all*).

## 4.2 Collocation

Inexperienced use of language is often characterized by inappropriate combinations of words, indicating the writer's lack of knowledge of collocations. In order to detect this, we calculate the Pointwise Mutual Information (PMI) of all adjacent word pairs (bigrams), as well as all adjacent word triples (trigrams) in the Google 1T web corpus (Brants and Franz, 2006). The higher the value of the PMI, the more common is the collocation for the word pair/triple in well formed texts. On the other hand, negative values of PMI indicate that the given word pair or triple is less likely than chance to occur together. We hypothesized that this would be a good indicator of awkward usage, as suggested in Chodorow and Leacock (2000).

The PMI values for adjacent words obtained over the entire response are then assigned to bins, with 8 bins for word pairs and another 8 for word triples following the procedure from (Somasundaran and Chodorow, 2014). Each of the 8 bins represents a range of PMI : $p > 20, 10 < p \leq 20, 1 < p \leq 10, 0 < p \leq 1, -1 < p \leq 0, -10 < p \leq -1, -20 < p \leq -10, p \leq -20$.

We generate two sets of features based on the proportions of bigrams/trigrams falling into each bin, resulting in a total of 16 features. In addition to binning, we also encode as features the maximum, minimum and median PMI value obtained over all bigrams and trigrams. These encode the best and the worst word collocations in a response as well as the overall general quality of the response.

## 4.3 Discourse

Stories are characterized by events that are related (and ordered) temporally or causally. In order to form a coherent narrative, it is often necessary to use proper transition cues to organize the story. Intuitively, coherent responses are more likely to have these cues than less coherent responses.

In order to detect discourse organization cues, we use two lexicons. The first was obtained from the Penn Discourse Treebank (PDTB) annotation manual (Prasad et al., 2008). The second was developed by manually mining websites giving advice on good narrative writing. The two lexicons gave us a total of over 550 cues. From the PDTB and our lexicon, we extracted the number of times each connective was encountered in a particular sense (sense information such as "Temporal" or "Cause" is directly provided in the PDTB manual, and we added similar information to our manually collected lexicon) and used the frequencies to construct a probability distribution over the senses for that cue. Then, for each response, we produced the following features: the number of cues found in the response (*totalCuesCount*), the number of cues found in the response divided by the number of words in the response (*normalizedCuesCount*), the number of cues belonging to the temporal category (*temporalCuesCount*), the number of cues belonging to the causal category (*causalCuesCount*), the sum of the probabilities of belonging to the temporal category for each cue found in

[3] We do not calculate agreement as producing different descriptions and having variety was the goal of the task of reference corpus creation.

the response (*temporalCuesScore*), the sum of the probabilities of belonging to the causal category for each cue found in the response (*causalCuesScore*).

### 4.4 Detailing

We hypothesized that better responses would show evidence of effective narrative techniques, such as providing vivid descriptions of the events and providing depth to the story. For example, one could say *"In the afternoon a boy and a man went to the library."*, or make the story more interesting by assigning names to the characters and places as *"One day John went to the Central Public Library because he wanted to do some research for his science project. An old man was walking behind him; his name was Peter. "*

We observed that certain syntactic categories, such as adjectives and adverbs, come into play in the process of detailing. Also, detailing by providing names to the characters and places results in a higher number of proper nouns (NNPs). Thus our detailing feature set consists of the following features: a binary value indicating whether the response contains any proper nouns (*presenceNames*), the number of proper nouns in the response (*countNames*), a binary value indicating whether the response contains any adjectives (*presenceAdj*), the number of adjectives in the response (*countAdj*), a binary value indicating whether the response contains any adverbs (*presenceAdv*), the number of adverbs in the response (*countAdv*). We use separate features for counts and presence of the syntactic category in order to balance the trade-off between sparsity and informativeness. The count features are more informative, but they can be sparse (especially for higher counts).

### 4.5 Sentiment

One common technique used in developing a story is to reveal the character's private states, emotions and feelings. This requires the use of subjectivity and sentiment terms.

We use lexicons for annotating sentiment and subjective words in the response. Specifically, we use a sentiment lexicon (*ASSESS*) developed in previous work in assessments (Beigman Klebanov et al., 2013) and the MPQA subjectivity lexicon (Wilson et al., 2005). ASSESS lexicon assigns a positive/negative/neutral polarity probability profile to its entries, and MPQA lexicon associates a positive, negative or neutral polarity category to its entries. We consider a word from the ASSESS lexicon to be polar if the sum of positive and negative probabilities is greater than 0.65 (we arrived at this number after manual inspection of the lexicon). This gives us the subjectivity feature set comprised of the following features: A binary value indicating whether the response contains any polar words from the ASSESS lexicon (*presencePolarProfile*), the number of polar words from the ASSESS lexicon found in the response (*cntPolarProfile*), a binary value indicating whether the response contains any polar words from the MPQA lexicon (*presenceMpqaPolar*), the number of polar words from the MPQA lexicon found in the response (*cntMpqaPolar*), a binary value indicating whether the response contains any neutral words from the MPQA lexicon (*presenceMpqaNeut*), the number of neutral words from the MPQA lexicon found in the response (*cntMpqaNeut*).

We construct separate features from the ASSESS lexicon and the MPQA lexicon because we found that the neutral category had different meanings in the two lexicons – even the neutral entries in the MPQA lexicon are valuable as they may indicate speech events and private states (e.g. view, assess, believe, cogitate, contemplate, feel, glean, think etc.). On the other hand, words with a high probability of being neutral in the ASSESS lexicon are non-subjective words (e.g. woman, undergo, entire, technologies).

## 5 Experiments

For our experiments, we used a supervised learning framework, with the data described above, to build scoring models based on our feature sets. We evaluated several different learning algorithms and found that a Random Forest Classifier consistently produced the best results in cross-validation experiments on the training data when we used our features as well as when we used the baseline set of features. Hence, all of our results in this section are reported using this Random Forest learner. Performance was calculated using Quadratic Weighted Kappa (QWK) (Cohen, 1968), which is the standard evaluation metric used in automated scoring. QWK measures the agreement between the system score and the

| Feature set | CV | Eval |
|---|---|---|
| Relevance | 0.43 | 0.46 |
| Collocation | 0.48 | 0.40 |
| Discourse | 0.25 | 0.27 |
| Details | 0.18 | 0.21 |
| Subjectivity | 0.17 | 0.16 |
| EW13 baseline | 0.48 | 0.52 |
| All Feats | 0.52 | 0.55 |
| All Feats + EW13 | 0.58 | 0.58 |

Table 2: Performance of different feature sets.

human-annotated score, correcting for chance agreement and penalizing large disagreements more than small ones.

### 5.1 Baseline

We use the previous state-of-the-art features from Evanini and Wang (2013) as our baseline *(EW13)*. They are comprised of the following subsets: fluency (rate of speech, number of words per chunk, average number of pauses, average number of long pauses), pronunciation (normalized Acoustic Model score, average word confidence, average difference in phone duration from native speaker norms), prosody (mean duration between stressed syllables), and lexical choice (normalized Language Model score).

### 5.2 Results and Analysis

We performed cross validation on our training data (Train) and also performed training on the full training dataset with evaluation on the Eval data. Table 2 reports our results on 10-fold cross validation experiments on the training data (CV), as well results when training on the full training dataset and testing on the evaluation dataset (Eval). The first 5 rows report the performance of the individual feature sets described in Section 4. Not surprisingly, each individual feature set is not able to perform as well as the EW13 baseline, which is comprised of an array of many features that measures various speech characteristics. One exception to this is the collocation feature set that performs as well as the EW13 baseline in the cross validation experiments. Notably, the combination of all five feature sets proposed in this work (*All Feats*), performs better than the EW13 baseline, indicating that our relevance and linguistic features are important for scoring for this spoken response item type. Finally the best performance is obtained when we combine our features with the speech-based features. This improvement of All Feats + EW13 over the baseline is statistically significant at $p < 0.01$, based on 10K bootstrap samples (Zhang et al., 2004). Somewhat surprisingly, the testing on the evaluation dataset showed slightly better performance for most types of features than the cross validation testing. We believe that this might be due to the fact that, for the Eval results, all the training data were available to train the scoring models.

| Feature set | Performance |
|---|---|
| EW13 baseline | 0.48 |
| EW13 + Relevance | 0.54 |
| EW13 + Collocation | 0.57 |
| EW13 + Discourse | 0.49 |
| EW13 + Details | 0.50 |
| EW13 + Subjectivity | 0.50 |

Table 3: Performance of the Baseline when each individual feature set is added to it.

We also performed analysis on the Train set to see if the baseline's performance is impacted when each of our individual feature sets is added to it. As shown in Table 3, each of the feature sets is able to improve the baseline's performance (of 0.48 QWK). Specifically, Discourse and Subjectivity produce a slight improvement while Relevance produces modest improvement. However, only the improvement produced by the Collocation features was statistically significant ($p < 0.01$)

## 6 Conclusions

In this work, we explored five different types of linguistic features for scoring spoken responses in a picture narration task. The features were designed to capture language proficiency, story development and task relevance. Our results are promising: we found that each feature is able to combine well with a state of the art speech feature system to improve results. The combination of the linguistic features achieved better overall performance than the speech features alone. Finally the best performance was achieved when linguistic and speech features were combined.

# References

Yigal Attali and Jill Burstein. 2006. Automated essay scoring with e-rater v. 2.0. *Journal of Technology, Learning, and Assessment*, 4:3.

Beata Beigman Klebanov, Jill Burstein, and Nitin Madnani. 2013. Sentiment profiles of multi-word expressions in test-taker essays: The case of noun-noun compounds. In *ACM Transactions on Speech and Language Processing*, volume 10(3).

Gilbert J. Botvin and Brian Sutton-Smith. 1977. The development of structural complexity in children's fantasy narratives. *Developmental Psychology*, 13(4):377 – 388.

Thorsten Brants and Alex Franz. 2006. Web 1T 5-gram Version 1. In *Linguistic Data Consortium, Philadelphia*.

Martin Chodorow and Claudia Leacock. 2000. An unsupervised method for detecting grammatical errors. In *Proceedings of the Conference of the North American Chapter of the Association of Computational Linguistics (NAACL)*, pages 140–147.

Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4).

Rod Ellis. 2000. Task-based research and language pedagogy. *Language teaching research*, 4(3):193–220.

Keelan Evanini and Xinhao Wang. 2013. Automated speech scoring for non-native middle school students with multiple task types. In *Proceedings of Interspeech*, pages 2435–2439.

Keelan Evanini, Michael Heilman, Xinhao Wang, and Daniel Blanchard. 2014. Automated scoring for TOEFL Junior comprehensive writing and speaking. Technical report, ETS, Princeton, NJ.

KE Forbes-McKay and Annalena Venneri. 2005. Detecting subtle spontaneous language decline in early Alzheimers disease with a picture description task. *Neurological sciences*, 26(4):243–254.

Khairun-nisa Hassanali, Yang Liu, and Thamar Solorio. 2013. Using Latent Dirichlet Allocation for child narrative analysis. In *Proceedings of the 2013 Workshop on Biomedical Natural Language Processing*. Association for Computational Linguistics.

Levi King and Markus Dickinson. 2013. Shallow semantic analysis of interactive learner sentences. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 11–21, Atlanta, Georgia, June. Association for Computational Linguistics.

Claudia Leacock, Martin Chodorow, Michael Gamon, and Joel Tetreault. 2014. *Automated Grammatical Error Detection for Language Learners*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool.

Choonkyu Lee, Smaranda Muresan, and Karin Stromswold. 2012. Computational analysis of referring expressions in narratives of picture books. In *Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature*, pages 1–7, Montréal, Canada, June. Association for Computational Linguistics.

Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of the 17th international conference on Computational linguistics-Volume 2*, pages 768–774. ACL.

Anastassia Loukina, Klaus Zechner, and Lei Chen. 2014. Automatic evaluation of spoken summaries: the case of language assessment. In *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 68–78, Baltimore, Maryland, June. Association for Computational Linguistics.

Christopher Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press.

Anne McKeough and Jennifer Malcolm. 2011. Stories of family, stories of self: Developmental pathways to interpretive thought during adolescence. *New Directions for Child & Adolescent Development*, 2011(131):59 – 71.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse Treebank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May. European Language Resources Association (ELRA).

Swapna Somasundaran and Martin Chodorow. 2014. Automated measures of specific vocabulary knowledge from constructed responses (use these words to write a sentence based on this picture). In *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–11. Association for Computational Linguistics.

Lei Sun and Marilyn A Nippold. 2012. Narrative writing in children and adolescents: Examining the literate lexicon. *Language, speech, and hearing services in schools*, 43(1):2–13.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT-EMNLP)*, pages 347–354. Association for Computational Linguistics.

Ying Zhang, Stephan Vogel, and Alex Waibel. 2004. Interpreting bleu/nist scores: How much improvement do we need to have a better system? In *Proceedings of the International Conference on Language Re-*

*sources and Evaluation (LREC)*. European Language Resources Association (ELRA).