Interspeech 2018
2-6 September 2018, Hyderabad

Improvements to an Automated Content Scoring System for Spoken CALL Responses: The ETS Submission to the Second Spoken CALL Shared Task

Keelan Evanini[†], *Matthew Mulholland*[†], *Rutuja Ubale*[‡], *Yao Qian*[‡], *Robert Pugh*[‡], *Vikram Ramanarayanan*[‡], *Aoife Cahill*[†]

Educational Testing Service R&D
[†]660 Rosedale Rd., Princeton, NJ, USA
[‡]90 New Montgomery Street, Suite 1500, San Francisco, CA, USA
{kevanini,mmulholland,rubale,yqian,rpugh,vramanarayanan,acahill}@ets.org

Abstract

This paper describes the details of the ETS submission to the 2018 Spoken CALL Shared Task. We employed a system using word and character *n*-gram features in a random forest machine learning framework based on the system that achieved the second-highest score in the text processing track of the 2017 Spoken CALL Shared Task. This system was augmented with additional features based on comparing the learner's responses to language models trained on text written by both native English speakers and L1-German English learners. In addition, we developed a set of sequence-to-label models using bidirectional LSTM-RNNs with an attention layer. The RNN model predictions were combined with the other feature sets using feature-level and score-level fusion approaches resulting in a best-performing system that achieved a D score of 7.397 on the test set (ranking 5th out of 12 submissions to the text processing track of the Shared Task). Subsequent experiments resulted in higher D scores when the model parameters were optimized for D score instead of F-score, and the paper presents an error analysis of these models in an attempt to determine which metric is more appropriate for evaluating spoken CALL systems.

Index Terms: Spoken CALL Shared Task, automated content scoring, D score

1. Introduction

Many studies from the field of applied linguistics have demonstrated the effectiveness of corrective feedback provided to language learners by instructors in a classroom environment, e.g., [1, 2, 3, 4]. Recently, several Computer Assisted Language Learning (CALL) systems have been developed in an attempt to provide language learners with opportunities to practice their speaking skills and receive automated feedback when an instructor is not present. While the automated feedback provided by these systems is typically restricted to pronunciation quality, some speech-based CALL systems also attempt to provide automated grammar feedback (e.g., [5, 6, 7, 8, 9, 10, 11]). Since this field of research is relatively new, and since few shared resources exist for comparing various error detection methodologies on a common data set, a shared task for spoken CALL was held at the Workshop on Speech and Language Technology in Education at Interspeech 2017. In this shared task, spoken English responses produced by adolescent native speakers of German while using a CALL application were released to the community along with annotations about the grammatical and semantic correctness of each response that were used by the shared task participants to train models for predicting whether the responses are erroneous or not; the results of this shared task are presented in [12]. Based on the success of this shared task, the organizers released a new set of data from the same CALL application and organized a second edition of the shared task at Interspeech 2018.[1] This paper describes the system that ETS developed to participate in this shared task.

Our submission to the 2017 Spoken CALL Shared Task explored the use of a pre-existing automated content scoring system that was developed at ETS augmented with additional features related to grammar and content. The automated content scoring system has been applied to score content in a wide variety of tasks, including short answer writing tasks for the domains of elementary and secondary schools in areas such as science, English language arts, and math [13, 14], longer writing tasks from a standardized assessment for music teachers [15], and speaking tasks in the context of a standardized assessment of English speaking proficiency [16]. The submission based on that system finished 2nd out of 15 submissions in the text processing task for the 2017 Spoken CALL Shared Task; further details about the design of the system and analyses of its performance are presented in [17]. For the 2018 Spoken CALL Shared Task, we started with the feature sets from the 2017 submission and explored additional feature sets based on an RNN model and language models trained on essays written by native and non-native speakers of English.

2. Data and Features

2.1. Data

The labels for the training data set released for the 2018 Spoken CALL Shared Task were obtained by first scoring them with four automated systems from the 2017 Spoken CALL Shared Task and subsequently obtaining up to three independent human judgments [18]. The 6,698 responses were divided into the following three categories with descending reliability based on the agreement statistics among the scores that were obtained: A (5,526 responses), B (873 responses) and C (299 responses). For our experiments, we combined the responses from all three categories together. In addition to this training set, the organizers also allowed the data that was released for the 2017 Spoken CALL Shared Task to be used for training. Since initial experiments indicated that the performance on the 2017 test set was lower than cross-validation performance on both the 2017 and 2018 training sets (this finding is consistent with the results from the 2017 shared task, in which the participating teams re-

[1] Further details about the Spoken CALL Shared Task, Second Edition are available here: https://regulus.unige.ch/spokencallsharedtask_2ndedition/.

10.21437/Interspeech.2018-2362

ported substantially higher performance on the training set than on the test set [12]), we decided not to include the 2017 test set in the training data. Therefore, the training data consisted of 11,919 responses from the 2017 and 2018 training sets.

2.2. Features from 2017 Submission

Labeling short responses provided by a spoken CALL application as "accept" or "reject" can be regarded as a binary classification problem. To accomplish this task, our experiments included the following feature sets that were initially explored in our submission to the 2017 Spoken CALL Shared Task (see [17] for further details about how the features in these feature sets were calculated).

- CHAR: Character n-grams for n = 2 to 5
- TOKEN: Token unigrams and bigrams
- SYN_DEP: Syntactic dependencies
- LENGTH: Features based on bins for values of the log of the number of characters in the response
- PROMPT: Prompt bias features
- WER: Similarity based on Word Error Rate between the response and sample correct responses for the given prompt contained in the reference grammar provided by the challenge organizers (`referenceGrammar.xml`)
- BLEU: Similarity based on BLEU score between the response and sample correct responses
- GRAMMAR: Grammatical errors detected using the open-source *LanguageTool* package[2]

In addition, we developed two new feature sets based on an RNN content model and language models trained on texts written in English by both L1 German speakers and native speakers of English; these feature sets are described in more detail in the following sections.

2.3. Attention BLSTM-RNN Feature

A straightforward way to carry out the binary classification task is to use a sequence-to-label function, which maps a sequence of input feature vectors to one of the two labels (either "accept" or "reject"). Motivated by the recent demonstrated success of deep learning technology in a variety of machine learning tasks, especially through the use of automatic feature extraction and feature engineering, we investigate whether it can reduce the effort that was required to develop the features listed in Section 2.2. Recurrent neural networks (RNNs) configured to process arbitrary-length input sequences have been successfully applied to solve a wide range of machine learning problems with sequence data. With long short-term memory (LSTM) cells, an RNN can overcome the vanishing gradient problem in training. A bidirectional LSTM-RNN (BLSTM-RNN) has two directions: the forward time direction and the backward time direction. The attention mechanism can be simply seen as a method for making the RNN focus on information that is of highest importance. It can significantly improve the performance of sequence-to-sequence models and has been used widely for the applications like machine translation and image captioning. Adding an attention layer into an LSTM-RNN model can be applied either to the input to the LSTM or to the output of the LSTM, which depends on the information required to propagate at every time step. The attention vector can also be dimension dependent if the input time series are multi-dimensional, i.e., one attention vector per dimension.

An attention BLSTM-RNN was constructed using the Keras package[3] along with `keras-attention-mechanism`.[4] The input word sequence was truncated to 50 tokens and converted to a 2D tensor using 300-dimensional word embedding vectors trained from Google News[5] resulting in a 50×300 input tensor. A stack of two BLSTM layers is used, and the attention layer is added either before the first BLSTM layer or after the second BLSTM layer. A softmax layer which contains the label "1" as "accept" and "0" as "reject" is used as the output layer of the BLSTM-RNN. The binary cross-entropy loss function and the Adam optimizer using the default parameters are applied to train the BLSTM-RNN parameters.

Parameters such as the number of layers, the number of nodes per layer, etc. were optimized using a 10-fold cross validation on the training set. The results of this optimization showed that two stacked BLSTM layers with 512 nodes each, i.e., 256 nodes per direction, achieved the best performance in terms of prediction accuracy. Table 1 presents further results from these cross-validation experiments indicating that 1) word embeddings initialized using pre-trained Google News and refined in the training of the sequence-to-label function can slightly outperform fixed embeddings; 2) an attention applied after the second BLSTM layer achieves slightly higher accuracy than one applied before the first BLSTM layer; 3) an attention vector added per input dimension is almost on par with a single vector across all input dimensions. The optimal structure and the corresponding parameters were used in the subsequent model for the shared task submission.

Word Embedding	Attention Vector	Attention Layer	Accuracy
trainable	per dim.	after BLSTM	88.8%
fixed	per dim.	after BLSTM	88.5%
fixed	per dim.	before BLSTM	87.9%
fixed	single	after BLSTM	88.4%

Table 1: *BLSTM-RNN prediction accuracy with different parameter settings using cross-validation on the training set.*

2.4. LM-based Features

In order to better model the grammatical and linguistic correctness of the spoken responses, we used features based on language models trained on text written in English by both L1 German speakers (since the English learners in the Spoken CALL Shared Task are L1 German speakers) and native speakers of English. Using responses from a large-scale standardized assessment of academic English proficiency, we trained trigram language models using the KenLM tool [19] on essays in English that received the highest and lowest possible scores (5 and 1, respectively); Table 2 provides the number of words that were used to train each of the four language models.

[2] We used the *language-check* Python wrapper for *LanguageTool* available from `https://pypi.python.org/pypi/language-check`.

[3] `https://keras.io`

[4] `https://github.com/philipperemy/keras-attention-mechanism`

[5] `https://code.google.com/archive/p/word2vec`

		L1	
		German	English
Score	5	22,172,498	13,894,091
	1	799,662	1,285,206

Table 2: *Number of words used to train the language models from essays written in English by L1 German and English speakers received high (5) and low (1) scores*

For each spoken response in the Spoken CALL Shared Task data set, we calculated the negative log-probability and perplexity of the response for each of the language models listed above; these values were then included in the LM feature set. The intuition behind these features is that correct responses would be expected to be closer matches to language models trained on responses from the native English speakers and high-scoring L1 German speakers whereas incorrect responses would be expected to be closer matches to the language model trained on responses from low-scoring L1 German speakers.

3. Model Training

3.1. RNN-only

In order to evaluate the performance of the BLSTM-RNN model using word embeddings on its own, one of the submissions used the scores produced by the RNN model described in Section 2.3 without any of the other features. This submission was given the label PPP by the shared task organizers.

3.2. Feature-level Fusion

A feature-level fusion model was trained using the combination of all of the feature types described in Section 2 (including the label posterior produced by the BLSTM-RNN model) fused together at the feature-level. Under this approach, features from all feature types are computed for each input and were combined for training as opposed to using a stacking configuration. Several different machine learning models were explored through cross-validation on the training set using the `scikit-learn` package[6] with hyper-parameter optimization conducted using F-score as the objective metric. The Random Forest classifier obtained the best result and was used to score the test set for the QQQ submission.

3.3. Score-level Fusion

Since most of the feature sets included in the feature-level fusion model are based on raw word tokens whereas the attention BLSTM-RNN model was trained with word sequences represented by embeddings, the two approaches may compensate for each other in predicting the labels jointly. Therefore, we also explored a score-level fusion approach by using the label posterior generated from a Random Forest classifier trained without the RNN feature and the label posterior from the attention BLSTM-RNN as the input to another classifier to predict the final label. Again using the `scikit-learn` toolkit (via the SKLL[7] interface), we experimented with many classifiers (including Support Vector Machine, Random Forest, Logistic Regression, AdaBoost Decision Tree, Multilayer Perceptron, among others) to train the score-level fusion model through cross-validation on the training set and using accuracy as the objective metric to optimize the hyper-parameters of the classifiers. Among these models, the AdaBoost classifier achieved the highest performance and it was subsequently used to score the test set for the RRR submission.

[6] `http://scikit-learn.org/`

[7] `https://github.com/EducationalTestingService/skll`

4. Results

In this section, we present the results of the three systems that we officially submitted to the 2018 Spoken CALL Shared Task along with several additional analyses that were conducted after the conclusion of the submission deadline.

4.1. Shared Task Submissions

Table 3 presents the following evaluation metrics for the three official submissions on the test set (PPP, QQQ, and RRR): precision, recall, F-score, accuracy, and D score. D score was the official metric used to rank submissions to the shared task and is defined as the ratio of the relative corect reject rate to the relative false reject rate [20]. As Table 3 shows, the score-level fusion system (RRR) that combined the predictions from the Random Forest model based on all feature sets minus the RNN feature and the attention BLSTM-RNN model using word embeddings achieved the highest performance using D score as the evaluation metric. This system ranked 5th out of the 12 systems that submitted results for the text processing task, and outperformed the baseline system in terms of D score, but not in terms of accuracy and F-score.

4.2. Objective Function

As described in Sections 3.2 and 3.3, the hyper-parameters of the fusion systems were optimized using either F-score or accuracy as the objective function, since these were available as built-in options in `scikit-learn`. Since the shared task used D score as the main evaluation metric, we extended the `scikit-learn` code base to enable the use of D score as an additional objective function and then retrained the feature-level fusion model using D score as the objective function and evaluated the results on the test set after the scores were released; as shown in Table 3 this system resulted in a substantially higher D score of 14.317 and an F-score (0.891) that was higher than the baseline (0.884). As shown in the table, this system had an exceptionally high recall value (0.988), which is similar to the recall value obtained by the highest performing system in the shared task (0.984 from the LLL system).

We also experimented with using recall and precision as the objective functions; however, the performance of the precision-based system was identical to the performance of the system optimized using F-score and the performance of the recall-based system was not meaningful since it accepted all responses (and therefore had an undefined D score, since it produced no false rejects).

4.3. Feature Comparisons

In order to determine the relative contributions of the different feature sets included in the models, we conducted an ablation study in which separate models were trained on the training set using each of the feature sets individually with D score as the objective function; these results are presented in Table 4. The results for models based on the SYN_DEP and TOKEN feature sets are not presented in the table since their D scores on the

System	Prec.	Rec.	F	Acc.	D
D score optimized	0.812	0.988	0.891	0.819	14.317
RRR (Score-level fusion)	0.842	0.920	0.880	0.823	7.397
QQQ (Feature-level fusion)	0.840	0.916	0.876	0.818	7.001
PPP (RNN-only)	0.802	0.912	0.853	0.784	5.648
Baseline	0.916	0.855	0.884	0.834	5.343

Table 3: *Evaluation results for the three official submissions on the test set (PPP, QQQ, RRR) compared to the baseline and a feature-level fusion system optimized based on the D score*

test set were undefined (due to the fact that these models did not produce any false rejects on the test set).

Feature Set	D
CHAR	**11.769**
LENGTH	10.274
LM	8.028
WER	7.075
RNN	5.648
BLEU	5.155
PROMPT	3.321
GRAMMAR	1.32

Table 4: *Results obtained on the test set with models trained using each of the feature sets optimizing for D score*

Subsequently, we conducted a step-wise ablation experiment in which each of the individual feature sets were added to the model in the order of the performance of the individual models. These results are presented in Table 5. As the table shows, a model that contained the CHAR, LENGTH, LM, WER, RNN, and BLEU features resulted in the highest performance, with a D score of 15.24.

Feature Sets	D
CHAR	11.769
+ LENGTH	12.766
+ LM	11.619
+ WER	11.905
+ RNN	13.167
+ BLEU	**15.24**
+ PROMPT	13.761
+ GRAMMAR	14.565
+ SYN_DEP	11.954
+ TOKEN	14.317

Table 5: *Results obtained on the test set with models trained based on the step-wise addition of the feature sets optimizing for D score*

5. Discussion and Conclusion

In this paper we presented the results of a system that automatically accepts or rejects responses submitted to an English spoken CALL application by L1 German speakers. Using a combination of features targeting content and grammatical accuracy and an attention BLSTM-RNN model based on word embeddings, our highest performing system achieved a D score of 7.397 on the test set of the 2018 Spoken CALL Shared Task. Additional experiments examined the impact of using different objective functions to optimize the hyper-parameters of the models. These experiments demonstrated that the D score result on the test set can be improved substantially when D score is used specifically as the objective function (in comparison to F-score)—the system optimized using D score with all features resulted in a D score of 14.317. A more detailed analysis of these results demonstrates that this system accepted 913 out of the 1000 responses in the test set (as evidenced by the high recall presented for this system in Table 3) and that it correctly rejected 31.2% of responses labeled as incorrect. This system therefore satisfied the constraint placed on submissions to the 2018 Spoken Call Shared Task: *In order to prevent "gaming" of the metric, entries are required to reject at least 25% of all incorrect responses.* However, other models that were trained using D score as the objective function did not satisfy this 25% threshold; for example, the model based on the CHAR feature set with a D score of 11.769 shown in Table 4 only detected 59 out of the responses labeled as incorrect by the human annotators (since the test set includes 250 responses labeled as incorrect, the 25% threshold is 63 responses). These results raise questions about whether it is appropriate to use the D score alone as the evaluation metric or whether it would be best to combine it with other metrics to produce a more robust and meaningful overview of the system's performance.

In addition to the step-wise feature ablation studies presented in Section 4.3, we also conducted full ablation studies using all combinations of feature sets with D score as the objective function. The highest performing system from these experiments had an exceptionally high D score of 101.351; however, it only rejected 60 of the test responses, and therefore would not meet the 25% threshold (24%). In fact, the vast majority of high-performing models from this full feature ablation study do not meet the 25% requirement. Therefore, when the D metric is used to evaluate spoken CALL systems, additional constraints should be placed on the model during the training phase to ensure that a valid model is learned. Taking into consideration the D metric and the 25% constraint, the best-performing model from the ablation experiment achieves a D score of 60.174 and it correctly rejects 27.6% of rejected responses. The shared task organizers discuss the instability of the D score as an evaluation metric further in their summary paper [18] and propose a new metric, D_{full}, which is the harmonic mean between the D score and a version of the D score that is based on the ratio of the relative corect accept rate to the relative false accept rate.

6. Acknowledgements

The authors would like to acknowledge Eugene Tsuprun's contributions to the development of the feature extraction pipeline that was initially used in the 2017 Spoken CALL Shared Task.

7. References

[1] R. Lyster, "Negotiation of form, recasts, and explicit correction in relation to error types and learner repair in immersion classrooms," *Language Learning*, vol. 48, no. 2, pp. 183–218, 1998.

[2] F. A. Morris, "Negotiation moves and recasts in relation to error types and learner repair in the foreign language classroom," *Foreign Language Annals*, vol. 35, no. 4, pp. 395–404, 2002.

[3] E. Kartchava and A. Ammar, "The noticeability and effectiveness of corrective feedback in relation to target type," *Language Teaching Research*, vol. 18, no. 4, pp. 428–452, 2014.

[4] D. Brown, "The type and linguistic foci of oral corrective feedback in the L2 classroom," *Lantuage Teaching Research*, vol. 20, no. 4, pp. 436–458, 2016.

[5] H. Morton and M. A. Jack, "Scenario-based spoken interaction with virtual agents," *Computer Assisted Language Learning*, vol. 18, no. 3, pp. 171–191, 2005.

[6] W. Johnson and A. Valente, "Tactical Language and Culture Training Systems: Using AI to teach foreign languages and cultures," *AI Magazine*, vol. 30, no. 2, pp. 72–83, 2009.

[7] K. Lee, S.-O. Kweon, S. Lee, H. Noh, and G. G. Lee, "POSTECH immersive English study (POMY): Dialog-based language learning game," *IEICE Transactions on Information and Systems*, vol. 97, no. 7, pp. 1830–1841, 2014.

[8] B. Penning de Vries, S. Bodnar, C. Cucchiarini, H. Strik, and R. van Hout, "Spoken grammar practice in an ASR-based CALL system," in *Proceedings of the Workshop on Speech and Language Technology in Education (SLaTE)*, Grenoble, France, 2013, pp. 60–65.

[9] H. Strik, J. van Doremalen, J. Colpaert, and C. Cucchiarini, "Development and Integration of Speech technology into COurseware for language learning: The DISCO project," in *Essential Speech and Language Technology for Dutch*, P. Spyns and J. Odijk, Eds. Berlin, Heidelberg: Springer, 2013, pp. 323–338.

[10] H. Strik, P. Drozdova, and C. Cucchiarini, "GOBL: Games online for basic language learning," in *Proceedings of the Workshop on Speech and Language Technology in Education (SLaTE)*, Grenoble, France, 2013, pp. 48–53.

[11] V. Timpe-Laughlin, J. Lee, K. Evanini, J. Bruno, and I. Blood, "*Can you guess who I am?*: An interactive task for young learners to practice *yes/no* question formation in English," in *Proceedings of the Workshop on Child Computer Interaction*, Glasgow, Scotland, 2017, pp. 62–67.

[12] C. Baur, C. Chua, J. Gerlach, M. Rayner, M. Russell, H. Strik, and X. Wei, "Overview of the 2017 Spoken CALL Shared Task," in *Proceedings of the Workshop on Speech and Language Technology in Education (SLaTE)*, Stockholm, Sweden, 2017, pp. 71–78.

[13] M. Heilman and N. Madnani, "The impact of training data on automated short answer scoring performance," in *Proceedings of the 10th Workshop on Innovative Use of NLP for Building Educational Applications*, 2015, pp. 81–85.

[14] O. L. Liu, J. A. Rios, M. Heilman, and M. C. Linn, "Validation of automated scoring of science assessments," *Journal of Research in Science Teaching*, vol. 53, pp. 215–233, 2016.

[15] N. Madnani, A. Cahill, and B. Riordan, "Automatically scoring tests of proficiency in music instruction," in *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, 2016, pp. 217–222.

[16] A. Loukina and A. Cahill, "Automated scoring across different modalities," in *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications, San Diego, California*, 2016, pp. 130–135.

[17] K. Evanini, M. Mulholland, E. Tsuprun, and Y. Qian, "Using an automated content scoring system for spoken call responses: The ets submission for the spoken call challenge," in *Proceedings of the ISCA Workshop on Speech and Language Technology in Education (SLaTE)*, 2017, pp. 97–102.

[18] C. Baur, A. Caines, C. Chua, J. Gerlach, M. Qian, M. Rayner, M. Russell, H. Strik, and X. Wei, "Overview of the 2018 Spoken CALL Shared Task," in *Proceedings of Interspeech*, Hyderabad, India, 2018.

[19] K. Heafield, I. Pouzyrevsky, J. H. Clark, and P. Koehn, "Scalable modified Kneser-Ney language model estimation," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, Sofia, Bulgaria, August 2013, pp. 690–696.

[20] C. Baur, J. Gerlach, M. Rayner, M. Russell, and H. Strik, "A shared task for spoken CALL?" in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, 2016.