Content Modeling for Automated Oral Proficiency Scoring System

Su-Youn Yoon and Chong Min Lee
Educational Testing Service / 660 Rosedale road, Princeton, USA
{syoon,clee001}@ets.org

Abstract

We developed an automated oral proficiency scoring system for non-native English speakers' spontaneous speech. Automated systems that score holistic proficiency are expected to assess a wide range of performance categories, and the content is one of the core performance categories. In order to assess the quality of the content, we trained a Siamese convolutional neural network (Siamese CNN) to model the semantic relationship between key points generated by experts and a test response. The correlation between human scores and Siamese CNN scores was comparable to human-human agreement ($r = 0.63$), and it was higher than the baseline content features. The inclusion of Siamese CNN-based feature to the existing state-of-the-art automated scoring model achieved a small but statistically significant improvement. However, the new model suffered from score inflation for long atypical responses with serious content issues. We investigated the reasons of this score inflation by analyzing the associations with linguistic features and identifying areas strongly associated with the score errors.

1 Introduction

We developed an automated scoring model for an oral proficiency assessment of non-native English speakers. In particular, the system was designed to score spontaneous speech, elicited using questions where the test takers summarized the core content of a reading and/or listening passages. A system for scoring holistic proficiency of spontaneous speech is expected to assess a wide range of areas such as fluency (Cucchiarini et al., 2000; Zechner et al., 2009), pronunciation (Witt and Young, 1997), prosody, grammar (Chen and Zechner, 2011; Yoon and Bhat, 2018) and vocabulary (Yoon and Bhat, 2012). Content is also one of the core performance categories in holistic oral proficiency scoring. In particular, automated scoring systems without content scoring capabilities may show sub-optimal performance when scoring responses with mismatched proficiency levels between content and other areas. For instance, some responses have critical content issues but good delivery skills, while some responses have good content but issues in other areas. Furthermore, in large-scale oral proficiency assessments, some responses may have sub-optimal characteristics. The types of these problematic responses (hereafter, atypical responses) for the tests eliciting spontaneous speech frequently have severe content issues. For instance, some test takers may try to game the system by citing memorized responses for unrelated topics (e.g., off-topic responses). Even state-of-the-art automated scoring systems face challenges in scoring these atypical responses, and automated scoring systems without content scoring capability may assign inaccurate scores for these responses. To address these issues, more researchers started to actively explore content scoring in the context of oral proficiency scoring (Xie et al., 2012; Evanini et al., 2013; Yoon et al., 2018).

Recently, deep neural networks (DNN) and word embeddings have been applied successfully to various natural language processing tasks. In the automated scoring area, several researchers have explored the use of diverse neural networks for essay scoring (Farag et al., 2018; Alikaniotis et al., 2016; Dong and Zhang, 2016) and spontaneous speech scoring (Chen et al., 2018a; Qian et al., 2018a,b) and they achieved comparable or superior performance to the sophisticated linguistic feature-based system. In particular, Qian et al. (2018b) trained an automated scoring model covering the content aspect and achieved a further improvement over the generic model without content modeling.

Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications, pages 394–401
Florence, Italy, August 2, 2019. ©2019 Association for Computational Linguistics

The content relevance of a response is a concept relative to the question, and thus, it is important that the neural network learns the semantic relevance between a question-response pair. Siamese networks are characterized by shared weights between two subnetworks modeling inputs and are effective in calculating semantic similarity between sentence pairs (Mueller and Thyagarajan, 2016; Yin et al., 2015; Hu et al., 2014).

In order to address the strong need for content scoring and based on the promising performance of the Siamese CNN in the semantic relevance modeling, we developed a Siamese CNN-based content model. In particular, we make the following two contributions:

- We developed a new feature, based on the Siamese CNN by modeling the semantic distance between the core content and the test takers' responses. The new Siamese CNN-based feature outperformed the baseline content features, and the inclusion of the new feature further improved the performance of a state-of-the-art automated speech scoring model.

- We examined whether the automated scoring model including the new Siamese CNN-based feature could assign accurate scores for atypical responses. Differing from previous studies (Higgins and Heilman, 2014; Yannakoudakis and Briscoe, 2012; Lee et al., 2017) using synthesized atypical responses in their evaluations, we used authentic atypical responses collected from a large number of test administrations.

2 Data

We used a large collection of spoken responses from an English proficiency assessment. It was composed of $109,894$ responses from $37,830$ speakers. For each question, test takers read and/or listened to a passage and then provided answers consisting of around one minute of spontaneous speech based on the given passage. We used 80 questions, covering a wide range of topics such as education, entertainment, health, and policies. For each question, the data included $1,374$ responses on average, but there were large variations ranging from 305 to $3,013$.

During the question generation, expert assessment developers first generated a list of key points to guide the creation of the reading and listening passages. These key points were provided to and used by human raters to evaluate content of the spoken responses. Three key points were generated for each question, and the responses with the perfect content coverage were expected to include all three key points. We concatenated three key points into one text and used it during the content model building. The key points contained on average 93 words.

All responses were scored by the trained raters using a 4-point scoring scale from 1 to 4 with 4 indicating the highest proficiency. In addition, raters provided a score of 0 when test takers did not show any intention of directly responding to the question. The rubrics consisted of three major performance categories: delivery (pronunciation, prosody, and fluency), language use (vocabulary and grammar), and topic development (content and coherence). Both the Pearson correlation and quadratic weighted kappa between two human raters based on 10% double-scored data[1] were 0.61.

The average of the human scores was 2.58, and the most frequent score was 3 (48%), followed by 2 (39%), 4 (8%), 1 (4%), and 0 (1%). The number of words in the transcriptions generated by an automated speech recognition (ASR) system (*numwds*) ranged from 11 to 248 (129 on average).

The characteristics of responses with score of 0 were widely varied, but some of the most frequent categories included (a) response in a non-target language; (b) off-topic; (c) canned responses[2]; (d) no-response including no speech other than fillers or simple sentences (e.g., "I don't know"); and (e) repetition of the question. These responses had serious problems in content. We used the responses with score of 0 as atypical responses and used them for an additional evaluation.

However, due to the low percentage of the score 0 responses, it was difficult to analyze the model accuracy for them. In order to address this issue, we constructed a separate atypical dataset by extracting a large number of responses with a score of 0 from the same English proficiency assessment, but much larger administrations. The size of dataset is presented in Table 1.

[1]The double-scored data included responses with scores of 1 to 4.

[2]Responses that only included memorized segments from external sources. The sources were irrelevant to the question, and the responses were likely to be off-topic.

Partition	Purpose	N. of responses
Train	Training of content features and a Siamese CNN model	54,051
LR Train	Training of linear regression models	25,706
Test	Evaluation	28,497
Atypical responses	Evaluation	1,640

Table 1: Number of responses for each partition

3 Method

3.1 Siamese Convolutional Neural Network (Siamese CNN)

We used a Siamese convolutional neural network (CNN) consisting of an input modeling step using two weight-sharing CNNs (one CNN was for modeling the key points and the other was for modeling responses), a similarity distance calculation layer, and a neural network layer. Figure 1 illustrates the overall architecture of our Siamese CNN.

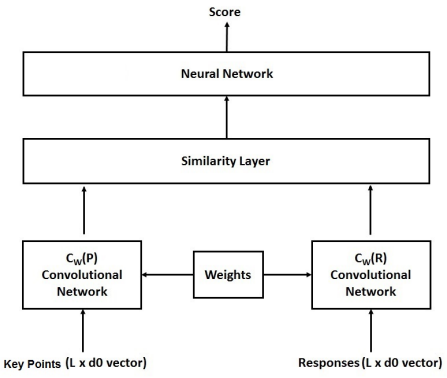

Figure 1: Diagram of Siamese Convolutional Neural Network

An input, a pair of a response and a text composed of three key points, was converted into a 2D tensor with a shape of $L \times d0$, where $L = 100$[3] and $d0 = 300$. $d0$ was the dimension of the word embedding vector, and we used Google word embedding vectors (Mikolov et al., 2013) with 300 dimensions.

The converted vector was fed into the convolution layer with the filter numbers $d1$ and filter width w; each filter created concatenated embedding vectors of w consecutive words. We trained models using different $d1$ (16, 64, 128, 256) and w (3, 4, 5, 6), and they were optimized using hyperopt software (Bergstra et al., 2013). This was followed by an averaging pooling layer, and two vectors (one for the key points and one for the response) were generated.

Next, a cosine similarity between the two vectors was calculated at the similarity layer. Finally, we stacked a neural network as the output layer, and it generated a score. The mean squared error (MSE) between the output scores and human scores was the learning metric and Adaptive Grad Optimizer was the optimizer. The model had a similar architecture to the 'basic Bi-CNN' model in Yin et al. (2015) with the final layer and different learning metric for the regression task.

3.2 Features from an automated proficiency scoring system

We used 38 features generated by a state-of-the-art automated proficiency scoring system for non-native speakers' spontaneous speech (Chen et al., 2018b). For a given spoken response, the system performed speech processing including speech recognition and forced-alignment and generated 38 features in five groups: (a) speech rate features, (b) pronunciation quality features[4], (c) pause pattern features, (d) prosody features[5], and (e) content features.

In particular, we generated 3 content features to assess the content accuracy and completeness. The first feature was designed to assess lexical similarity with high-scoring responses. It calculated a term frequency-inverse document frequency (tf-idf) weighted cosine similarity score between a test response vector and the question-specific tf vector. The question-specific tf vector was a vector whose elements were the frequency of each word in the entire sample responses with a score of 4 that answered the same question. The question-specific vector was trained on the Train partition.

[3]Typically, L is the maximum length of the input, but we used $L = 100$ due to the consistently superior performance in the pilot experiments using varying L. For key points or responses shorter than 100 words, we added zeros to the end. On the contrary, for key points or responses longer than 100, we selected the initial 100 words.

[4]This group of features measures how much the test takers' pronunciation deviates from the native norms.

[5]This group of features measures patterns of variation in time intervals between syllables or phonemes.

The remaining two features were designed to measure similarity with the key points created by the assessment developers. We created an average embedding vector and an idf weighted average embedding vector for the key-points. Next, we created two vectors for a test response using the same process. Finally, we calculated two cosine similarity scores between the key-point embedding vector and response embedding vector: one score for the average embedding vectors and one score for the idf weighted average embedding vectors. The detailed description of features used in this study is provided in Yoon et al. (2018).

3.3 Scoring Model Training

We trained linear regression models to generate a proficiency score for each response. In order to evaluate the impact of the Siamese CNN based feature, we classified features into 4 groups:

- content: three content features in Section 3.2
- all-features: all 38 features in Section 3.2
- Siamese CNN: output score of the Siamese CNN model
- CMB: combination of all-feature and the output score of the Siamese CNN model

Finally, we trained 4 linear regression models (one model for each feature group) using a human score as a dependent variable using the RSMTool (Madnani and Loukina, 2016).

4 Experiment

We generated transcriptions of a spoken response using an ASR system composed of a gender-independent acoustic model and a trigram language model trained on 800 hours of spoken responses extracted from the same English proficiency test using the Kaldi toolkit (Povey et al., 2011). The ASR system achieved a Word Error Rate of 23% on 600 held-out responses (Tao et al., 2016).

Next, we normalized both key points and ASR-based transcriptions by tokenizing and removing stop words and disfluencies. After the normalization process, the length of the key points and responses were reduced to 60% and 40% of the original texts.

We trained a Siamese CNN model using the normalized texts of the Train partition and the key-points. The model was implemented using Tensorflow (Abadi et al., 2015). The parameters were optimized using the hyperopt software, and the final model used $L = 100$, $d = 300$, $d1 = 256$, $w = 4$, and the learning rate $l = 0.0001$. In addition, the automated scoring system using the same ASR engine generated 38 features. Finally, we trained linear regression models on the LR Train partition.

5 Results

5.1 Scoring of normal responses

We first evaluated the performance of the automated scoring models on the Test partition in terms of its strength in the associations with proficiency scores assigned by human raters. Table 2 presents the agreement between the human scores and the automated scores for each model.

	Correlation	κ	RMSE
Siamese CNN	0.634	0.588	0.601
content	0.452	0.499	0.663
all-feature	0.672	0.620	0.565
CMB	0.686	0.631	0.555

Table 2: Correlations, quadratic weighted kappas (κ), and root mean squared error (RMSE) between the automated scores and human scores

The performance of the Siamese CNN model was substantially better than the content feature-based model; the correlation and quadratic weighted kappa increased approximately 0.18 and 0.09, respectively. On the contrary, the performance of the Siamese CNN model was significantly lower than the performance of the all-feature model, and this difference was also statistically significant ($p < 0.01$) based on the Steigers Z-test for dependent correlations.

The combination of the Siamese CNN and all-feature achieved a small improvement. The correlation and quadratic weighted kappa of the CMB model were 0.686 and 0.631, respectively. There was approximately 0.01 increase over the best performing individual model (all-feature model). This improvement was statistically significant at 0.01 level ($p < 0.01$).

5.2 Scoring of the atypical responses

Next, we evaluated whether the automated scoring models assign accurate scores for atypical responses using the Atypical response set. Table 3 compares the mean and standard deviation (STD) of the automated scores for each model. In general, the models with the lower average score are more accurate than those with the higher average score because the human scores for all responses in this set were 0.

	Mean	STD
Siamese CNN	1.129	0.344
content	0.545	0.500
all-feature	0.969	0.845
CMB	0.732	0.490

Table 3: Comparison of the automated scores for the atypical responses

In general, the average scores of the automated models were low. The average scores of the feature-based models (both content and all-feature) were lower than 1.0, and this was lower than the lowest scale score for the normal responses; our scoring scale for normal responses (excluding atypical responses) ranged from 1 to 4 with 1 indicating the lowest proficiency. The average score of the Siamese CNN model was slightly higher, at 1.13. Finally, the average score of the CMB model was lower than both Siamese CNN and all-feature models. The combination of the two groups of features resulted in assigning more accurate scores for the atypical responses and improved the robustness of the automated scoring system.

In general, automated scoring models tend to assign high scores for long responses, and thus the automated models in this study may assign even higher scores for the long atypical responses. In the Atypical response set, the percentage of short responses was high (atypical responses with less than 20 words was 59%). Therefore, despite the low average score, there was a possibility that the automated models assigned high scores for a subset of atypical responses. Figure 2 presents the average automated scores by the response length.

The automated scores for the Siamese CNN model were relatively low for the short responses, and they increased substantially as the response length increased; it sharply increased

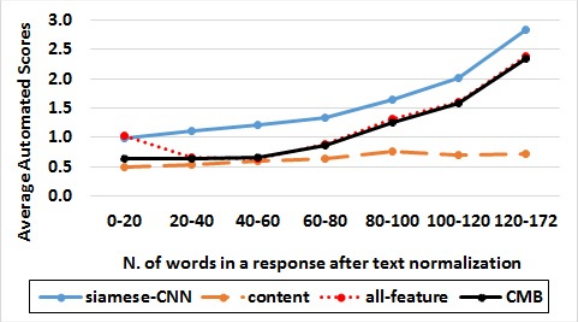

Figure 2: Average score predicted by the scoring models.

when the responses contained more than 60 words. The average Siamese CNN score for the responses longer than 120 words was 2.84. In contrast, the automated scores for the content model were consistently lower than 1.0.

The all-feature model also showed a similar trend to the Siamese CNN model; as the response length increased, the automated scores increased substantially. However, the average scores were substantially lower than those of the Siamese CNN model. Finally, the combination of the Siamese CNN and the all-feature (CMB model) resulted in the improvement in the short atypical responses; the average score of the responses containing 0-20 words was 0.64, and it was 0.4 lower than the all-feature model. However, no large difference was found from the longer atypical responses.

6 Discussion

From the atypical response scoring experiment, we found that the Siamese CNN model had a tendency to inflate scores for the long atypical responses. The long atypical responses in this study tended to be associated with the salient content issues such as off-topic responses from the test takers who cited answers for unrelated topics. The score inflation of these responses suggested that the Siamese CNN model may not have strong power in identifying responses with the severe content abnormality.

In order to get better understanding about which performance areas (e.g., content, fluency, vocabulary) the Siamese CNN model assessed mainly, we analyzed the relationships between the Siamese CNN score and three features from the automated proficiency scoring system: (a) speaking rate (fluency), (b) an average of the frequencies of the

words used in a response (vocabulary), and (c) a cosine similarity score between a question-specific content vector and a response (content). These features assess fluency, vocabulary, and content skills. Table 4 presents the correlation analysis calculated from the Atypical response set.

	vocabulary	content	fluency
r	0.395	0.430	0.706

Table 4: Correlation of the Siamese CNN score with the features from the oral proficiency scoring system: Pearson correlation coefficients in absolute values

All human scores in this set were 0. Therefore, the associations between the features and the holistic proficiency had no effect on the correlations showed in Table 4.

All correlations were statistically significant at 0.01 level. However, the Siamese CNN score showed the strongest correlation with the fluency feature, and the correlation with the content feature was much weaker than that with the fluency feature.

Next, we randomly selected an atypical response with a high Siamese CNN score; the Siamese CNN score was 3.3 while the score of the all-feature model was 2.2. Thus, the Siamese CNN model showed a stronger score inflation than the all-feature model. The response included 53 words after the text normalization. The response was clearly off-topic; the question was in the "entertainment life at the university" domain, while the answer was about "science, nature." Similar to Zeiler and Fergus (2014)'s occlusion experiment, we systematically removed n-words ($n = 1, 2, ..., 5$) from the response and generated scores for the new responses by the Siamese CNN model to identify the areas associated with high score inflation. Figure 3 presents the relationship between the score changes and the removed n-words.

There were approximately 5 points with substantial score drops (marked with red square in the Figure). The words at these points were "plankton," "swam," "microsoft," "semester," and "nice." These words were strongly associated with the score inflation and removal of these words resulted in substantially lower scores. Among them, first three words were relatively low frequency words but not topically relevant. The word frequencies in language learners' responses have been consistently identified as one of the strong predictors of vocabulary skill. These analyses supported the notion that the current Siamese CNN model might be paying strong attention to the fluency and vocabulary aspect.

7 Conclusion

We trained a Siamese CNN to model the semantic distance between the key points generated by the experts and the test takers' responses. The Siamese CNN model achieved a high performance without sophisticated feature engineering. For scoring normal responses, it achieved substantially better performance than the model using the content features from the existing automated speech scoring system. The inclusion of the Siamese CNN based feature to the existing state-of-the-art automated speech scoring system resulted in a small but statistically significant improvement. Furthermore, it improved the validity and robustness of the automated scoring system by assigning more accurate scores for short atypical responses. However, the Siamese CNN model suffered from score inflation during scoring long atypical responses. In the current human scoring scenario, the percentage of these long atypical responses was extremely low and they were correctly scored by human raters. However, this may be an important challenge that we need to overcome for the use of an automated scoring model as a sole scorer.

In this study, we explored the linear combination of the Siamese-CNN and linguistic features. The reviewers commented that there may be a further improvement by using non-linear algorithms to combine them. In particular, one of the reviewer suggested a possibility to train a Siamese CNN with linguistic features as additional inputs. In a future study, we will explore these points. In addition, we will also explore developing separate binary classifiers to filter out atypical responses and prevent an automated scoring model from generating erroneous scores.

Acknowledgments

We thank Rene Lawless, Patrick Houghton, Klaus Zechner, Michael Flor, Xinhao Wang, and anonymous reviewers for their comments and suggestions.

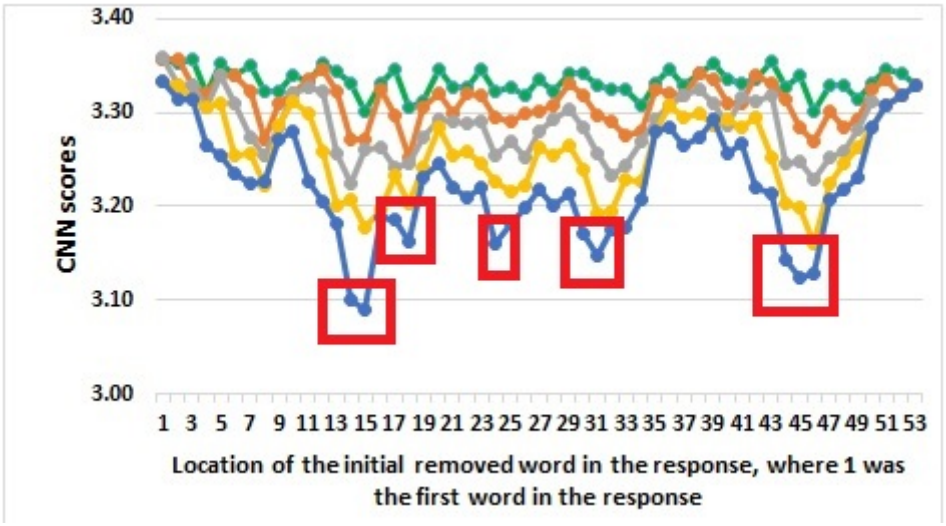

Figure 3: Siamese CNN scores for the responses excluding n-words by 1-word (green), 2-words (orange), 3-words (grey), 4-words (yellow), and 5-words (blue)

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Dimitrios Alikaniotis, Helen Yannakoudakis, and Marek Rei. 2016. Automatic text scoring using neural networks. *arXiv preprint arXiv:1606.04289*.

James Bergstra, Daniel Yamins, and David Daniel Cox. 2013. Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures.

Lei Chen, Jidong Tao, Shabnam Ghaffarzadegan, and Yao Qian. 2018a. End-to-end neural network based automated speech scoring. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6234–6238. IEEE.

Lei Chen, Klaus Zechner, Su-Youn Yoon, Keelan Evanini, Xinhao Wang, Anastassia Loukina, Jidong Tao, Lawrence Davis, Chong Min Lee, Min Ma, et al. 2018b. Automated scoring of nonnative speech using the speechrater sm v. 5.0 engine. *ETS Research Report Series*, 2018(1):1–31.

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

Catia Cucchiarini, Helmer Strik, and Lou Boves. 2000. Quantitative assessment of second language learners fluency by means of automatic speech recognition technology. *The Journal of the Acoustical Society of America*, 107(2):989–999.

Fei Dong and Yue Zhang. 2016. Automatic features for essay scoring–an empirical study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1072–1077.

Keelan Evanini, Shasha Xie, and Klaus Zechner. 2013. Prompt-based content scoring for automated spoken language assessment. In *Proceedings of the eighth workshop on innovative use of NLP for building educational applications*, pages 157–162.

Youmna Farag, Helen Yannakoudakis, and Ted Briscoe. 2018. Neural automated essay scoring and coherence modeling for adversarially crafted input. *arXiv preprint arXiv:1804.06898*.

Derrick Higgins and Michael Heilman. 2014. Managing what we can measure: Quantifying the susceptibility of automated scoring systems to gaming behavior. *Educational Measurement: Issues and Practice*, 33(3):36–46.

Baotian Hu, Zhengdong Lu, Hang Li, and Qingcai Chen. 2014. Convolutional neural network architectures for matching natural language sentences. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2042–2050. Curran Associates, Inc.

Chong Min Lee, Su-Youn Yoon, Xihao Wang, Matthew Mulholland, Ikkyu Choi, and Keelan Evanini. 2017. Off-topic spoken response detection using siamese convolutional neural networks. In *INTERSPEECH*, pages 1427–1431.

Nitin Madnani and Anastassia Loukina. 2016. Rsmtool: collection of tools building and evaluating automated scoring models. *The Journal of Open Source Software*, 2016.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Jonas Mueller and Aditya Thyagarajan. 2016. Siamese recurrent architectures for learning sentence similarity. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, pages 2786–2792. AAAI Press.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. 2011. The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*, EPFL-CONF-192584. IEEE Signal Processing Society.

Yao Qian, Rutuja Ubale, Matthew Mulholland, Keelan Evanini, and Xinhao Wang. 2018a. A prompt-aware neural network approach to content-based scoring of non-native spontaneous speech. In *Proceedings of the Spoken Language Technology Workshop*.

Yao Qian, Rutuja Ubale, Matthew Mulholland, Keelan Evanini, and Xinhao Wang. 2018b. A prompt-aware neural network approach to content-based scoring of non-native spontaneous speech. In *Proceedings of the 2018 Workshop on Spoken Language Technology*.

Jidong Tao, Shabnam Ghaffarzadegan, Lei Chen, and Klaus Zechner. 2016. Exploring deep learning architectures for automatically grading non-native spontaneous speech. In *Proceedings of ICASSP*, pages 6140–6144.

Silke Witt and Steve Young. 1997. Performance measures for phone-level pronunciation teaching in CALL. In *Proceedings of STiLL*, pages 99–102.

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

Helen Yannakoudakis and Ted Briscoe. 2012. Modeling coherence in ESOL learner texts. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 33–43. Association for Computational Linguistics.

Wenpeng Yin, Hinrich Schütze, Bing Xiang, and Bowen Zhou. 2015. ABCNN: Attention-based convolutional neural network for modeling sentence pairs. *arXiv preprint arXiv:1512.05193*.

Su-Youn Yoon and Suma Bhat. 2012. Assessment of esl learners' syntactic competence based on similarity measures. In *Proceedings of EMNLP*, pages 600–608.

Su-Youn Yoon and Suma Bhat. 2018. A comparison of grammatical proficiency measures in the automated assessment of spontaneous speech. *Speech Communication*, 99:221–230.

Su-Youn Yoon, Anastassia Loukina, Chong Min Lee, Matthew Mulholland, Xinhao Wang, and Ikkyu Choi. 2018. Word-embedding based content features for automated oral proficiency scoring. In *Proceedings of the Third Workshop on Semantic Deep Learning*, pages 12–22.

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.

Matthew D Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer.