Generating Content-oriented Feedback on Spontaneous Speech for Language Learning and Assessment

Xinhao Wang[1], *Christopher Hamill*[2]

Educational Testing Service
[1]90 New Montgomery St. #1450, San Francisco, CA 94105, USA
[2]660 Rosedale Rd., Princeton, NJ 08541, USA
{xwang002, chamill}@ets.org

Abstract

In spoken language learning and assessment, content appropriateness is an important dimension when assessing a language learner's spontaneous speech production. Although various measurements have been developed to evaluate content coverage and correctness, there is still a dearth of tools for automatically generating diagnostic feedback to assist language learners in improving their content development skills. In order to address the increasing demand for such a capability, we propose an effective way to automatically generate content-oriented feedback based on the detection of key points in an extended utterance. In this work, we focus on one type of widely used test question that requires test takers to first listen to and/or read stimulus materials, and then to construct a spontaneous response to a question related to the stimulus. We further define "key points" as the critical content from the stimulus materials that a high-proficiency response should properly cover. We build Transformer-based models to automatically detect the locations of key point spans within a response, or if no key points are covered, to detect their absence. We also introduce a multi-task learning approach for assigning a "quality score" to each key point span measuring how well the key point is rendered within the response. Experimental results demonstrate that the proposed models can automatically generate accurate and easily interpretable feedback that can provide interactive guidance to language learners in improving the content of their responses.

Index Terms: spoken language learning and assessment, content-oriented feedback, key point

1. Introduction

With the rapid progress of technological development in natural language and speech processing, the demand for automatic tools to effectively and reliably assist language learners is on the rise. In spoken language learning and assessment, automatic systems have been developed to provide valid scores so as to reduce the burden on teachers. These tools can assess a wide range of speech dimensions [1], including aspects of fluency, pronunciation and prosody [2, 3], as well as to a lesser extent aspects of vocabulary and grammar [4, 5], content appropriateness [6, 7, 8, 9], and discourse coherence [10, 11].

Of particular relevance to the spoken content dimension, features that measure the overall content appropriateness of a response to a test question have been proposed [6, 8, 12, 9]. The methods include both character and word n-gram based statistical models [12], word embeddings [12], and Siamese deep neural networks [8]. However, the systems generated by these studies are unable to generate diagnostic feedback for language learners, and a more targeted assessment of content coverage and correctness that goes beyond a generic measure of topicality has been underexplored.

This paper reports on an attempt to automatically generate diagnostic content feedback by determining the presence/absence of key points in language learners' responses. In many large-scale English spoken language assessments, one type of widely used task begins by presenting a listening and/or reading passage to the test taker, followed by a related question, where the test developers determine ahead of time a set of question-specific key points that high-proficiency responses should cover. The test taker must then formulate a one-minute spoken response by integrating relevant information from the provided listening and/or reading stimulus materials. In this work, test questions and related responses are collected from a large-scale standardized international language assessment. Selected responses were annotated to identify which key points were covered and how well.

This research reported in this paper is a continuation of earlier work. In our previous study [13], we proposed an automatic model to generate meaningful and easily interpretable content feedback for English learners by 1) detecting missing key points or the spans of present key points, and 2) predicting the quality score of each present key point, which can indicate how well the key point is rendered in the response. In this paper, we continue this line of research and make further contributions as follows: 1) we collect a much larger corpus to develop our models, involving 16 test questions and 2,540 spoken responses manually transcribed and annotated by human experts; 2) we use an automatic method for mapping manual annotations from human transcriptions onto automatic speech recognition (ASR) hypotheses; 3) under the goal of content-oriented feedback generation, multiple Transformer-based models are trained and evaluated for the key point detection task with both human transcriptions and ASR hypotheses.

2. Methodology

2.1. Key Points

In the field of language testing, research has repeatedly shown that human raters pay considerable attention to speech content while scoring [14, 15]. Accordingly, this study explores an effective way to automatically generate diagnostic content feedback with the goal of assisting language learners in improving their content development skills during speaking. A critical measure of content coverage and correctness is the extent to which the content of the listening and/or reading source materials can be accurately reflected/reproduced when learners integrate these stimuli in a spoken response. Therefore, we define key points as the critical content from the source materials that should be properly rendered in a high-proficiency response to a related test question. Language testing research has shown a clear positive relationship between the number of key points covered and proficiency levels [16, 17]. In the context of this research, the key points for each test question are decided in advance by the developers of the question. For the questions used in this study, each is paired with six distinct key points.

In this work, the experimental data consists of 2,540 one-minute spoken responses to 16 test questions from a large-scale standardized international language assessment. Using human transcriptions of these responses, two language teaching and assessment experts conducted the annotation along two dimensions: 1) text span location and 2) quality score rating. First, for each response, the annotators identified which of the six key points are covered and where in the response the associated text spans are located; for key points deemed absent in a response, no text spans were annotated. Second, the annotators rated each key point on a three-point scale: 1 (full coverage of the relevant key point), 0.5 (partial coverage), and 0 (no coverage). These ratings serve as the quality score for each key point in a response.

After collecting the annotations, the annotated data was split into training, development, and test sets. The 1,693 single-annotated responses formed the training set, the 318 responses used for annotator calibration formed the development set, and the remaining 529 double-annotated responses formed the test set. Furthermore, the manual annotations on human transcriptions were automatically mapped onto ASR hypotheses based on edit distance. However, due to errors introduced by the speech recognizer (word error rate = 16.3% on a stand-alone test set from the same assessment) and the automatic span mapping procedure, we observe that the total number of present key points was reduced from 10,868 (on human transcriptions) to 9,712 (on ASR hypotheses).

2.2. Automatic Detection Models

Given a test response and a single related key point, the detection task consists of two components: 1) to automatically detect the location of the text span containing the key point if it is present in the response; otherwise, to detect its absence; 2) to automatically measure how well a key point is rendered within a response if it is present. The first component is analogous to a typical Question&Answering task that has been widely studied in the field of natural language processing, for example, SQuAD V2.0 [18] with unanswerable questions. In this work, we examine two Transformer-based models on the key point detection task, i.e., BERT (Bidirectional Encoder Representations from Transformer) [19] and RoBERTa (Robustly Optimized BERT Approach) [20], which has advanced the state-of-the-art on the Q&A task. The standard F1-score measuring the average overlap between the predictions and ground truth [21] is used as the evaluation metric for the span detection task. Second, with a quality score of 0, 0.5, or 1 assigned by human experts to each key point, a regression model can be built to measure quality scores, and the Pearson correlation coefficient between manual and automated quality scores is used as the evaluation metric.

In this study, we work with non-native spontaneous responses including various types of errors and at times unintelligible speech, particularly for low-proficiency responses. This data is mismatched with the large amount of well-written texts used to pre-train both the BERT and RoBERTa models, such as BooksCorpus [22], English Wikipedia, and other text corpora. Therefore, in order to obtain models with better language representation capabilities on non-native spontaneous speech, we collected a data set with human transcriptions on 58,291 spoken responses drawn from the same assessment and used it first to fine-tune BERT/RoBERTa. The fine-tuning was performed using masked language modeling (MLM), where the number of training epochs was set at 4, and around 10% of the total steps were used for warmup. Subsquently, the obtained in-domain models were further fine-tuned on the downstream key point detection task with the annotated data.

Table 1: *Model performance on the key point span detection (F1-score) and quality score prediction (Pearson correlation coefficients (r)) tasks. Human agreement levels are also listed for comparison.*

	Manual Transcripts		ASR Hypotheses	
	F1 (%)	*r*	F1 (%)	*r*
BERT	76.2	0.769	66.5	0.575
RoBERTa	76.1	0.755	66.6	0.579
Agreement	72.0	0.734	68.8	0.610

In addition, previous research has demonstrated that multi-task learning can benefit deep learning applications by jointly optimizing regression and/or classification objectives across multiple tasks [23, 24]. In this study, we employed a method proposed in [23] to automatically weight multiple loss functions by considering the homoscedastic uncertainty of each task. Therefore, the key point span detection model and the quality score prediction model can be jointly optimized in a single training process.

3. Experiments

We adapted the implementation of pre-trained Transformer-based models from Hugging Face [25] to build the key point detection models, and experimented with the large-size BERT and RoBERTa models[1]. Our detection models were fine-tuned with 6 epochs on the downstream task, and the number of warmup steps was set to be around 10% of the total steps. As shown in Table 1, based on the human transcriptions, both the BERT and RoBERTa models can greatly outperform human expert performance on both the span detection task and the quality score prediction task. In contrast, based on the ASR hypotheses, RoBERTa performs slightly better than BERT but still underperforms human expert performance.

4. Conclusion

This study explores an effective way to automatically generate accurate and actionable content-oriented feedback in spoken language learning and assessment. We address this challenge by formalizing a key point detection task to automatically identify the critical pieces of content information present in (or missing from) learners' spoken responses. Transformer-based models were built to detect missing key points, as well as text span locations and quality scores of the covered key points. Experimental results demonstrated the effectiveness of the proposed models on both the span detection task and the quality score prediction task. Accordingly, meaningful and easily interpretable feedback can be automatically generated to assist language learners in improving their content development skills during speaking.

[1] https://huggingface.co/transformers/pretrained_models.html.

5. References

[1] K. Zechner and K. Evanini, *Automated Speaking Assessment: Using Language Technologies to Score Spontaneous Speech*. Routledge, 2019.

[2] H. Franco, H. Bratt, R. Rossier, V. Rao Gadde, E. Shriberg, V. Abrash, and K. Precoda, "Eduspeak®: A speech recognition and pronunciation scoring toolkit for computer-aided language learning applications," *Language Testing*, vol. 27, no. 3, pp. 401–418, 2010.

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

[4] J. Bernstein, J. Cheng, and M. Suzuki, "Fluency and structural complexity as predictors of l2 oral proficiency," in *Proceedings of INTERSPEECH*, 2010.

[5] L. Chen, K. Zechner, S.-Y. Yoon, K. Evanini, X. Wang, A. Loukina, J. Tao, L. Davis, C. M. Lee, M. Ma *et al.*, "Automated scoring of nonnative speech using the speechrater sm v. 5.0 engine," *ETS Research Report Series*, vol. 2018, no. 1, pp. 1–31, 2018.

[6] Y. Qian, R. Ubale, M. Mulholland, K. Evanini, and X. Wang, "A prompt-aware neural network approach to content-based scoring of non-native spontaneous speech," in *Proceedings of the IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 979–986.

[7] S. Xie, K. Evanini, and K. Zechner, "Exploring content features for automated speech scoring," in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2012, pp. 103–111.

[8] S.-Y. Yoon and C. Lee, "Content modeling for automated oral proficiency scoring system," in *Proceedings of the Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, 2019, pp. 394–401.

[9] A. Loukina and A. Cahill, "Automated scoring across different modalities," in *Proceedings of the Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, 2016, pp. 130–135.

[10] X. Wang, B. Gyawali, J. V. Bruno, H. R. Molloy, K. Evanini, and K. Zechner, "Using rhetorical structure theory to assess discourse coherence for non-native spontaneous speech," in *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking*, 2019, pp. 153–162.

[11] X. Wang, K. Evanini, K. Zechner, and M. Mulholland, "Modeling discourse coherence for the automated scoring of spontaneous spoken responses," in *Proceedings of the ISCA Workshop on Speech and Language Technology in Education (SLaTE)*, 2017, pp. 132–137.

[12] S.-Y. Yoon, C.-N. Hsieh, K. Zechner, M. Mulholland, Y. Wang, and N. Madnani, "Toward automated content feedback generation for non-native spontaneous speech," in *Proceedings of the Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, 2019, pp. 306–315.

[13] X. Wang, K. Zechner, and C. Hamill, "Targeted content feedback in spoken language learning and assessment," in *Proceedings of Interspeech*, 2020, pp. 3850–3854.

[14] T. Sato, "The contribution of test-takers' speech content to scores on an english oral proficiency test," *Language Testing*, vol. 29, no. 2, pp. 223—241, 2012.

[15] A. Brown, N. Iwashita, and T. McNamara, "An examination of rater orientations and test-taker performance on english-for-academic-purposes speaking tasks," *ETS Research Report Series*, no. 1, pp. 1–157, 2005.

[16] K. Frost, C. Elder, and G. Wigglesworth, "Investigating the validity of an integrated listening-speaking task: A discourse-based analysis of test takers' oral performances," *Language Testing*, vol. 29, no. 3, pp. 345—-369, 2012.

[17] C.-N. Hsieh and Y. Wang, "Speaking proficiency of young language students: A discourse-analytic study," *Language Testing*, vol. 36, no. 1, pp. 27—-50, 2019.

[18] P. Rajpurkar, R. Jia, and P. Liang, "Know what you don't know: Unanswerable questions for SQuAD," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, Jul. 2018, pp. 784–789.

[19] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2019, pp. 4171–4186.

[20] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[21] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "Squad: 100,000+ questions for machine comprehension of text," *arXiv preprint arXiv:1606.05250*, 2016.

[22] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler, "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 19–27.

[23] A. Kendall, Y. Gal, and R. Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7482–7491.

[24] X. Liu, P. He, W. Chen, and J. Gao, "Multi-task deep neural networks for natural language understanding," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019, pp. 4487–4496.

[25] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, and J. Brew, "Huggingface's transformers: State-of-the-art natural language processing," *ArXiv*, vol. abs/1910.03771, 2019.