INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Targeted Content Feedback in Spoken Language Learning and Assessment

Xinhao Wang[1], *Klaus Zechner*[2], *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, kzechner, chamill}@ets.org

Abstract

This study aims to develop automatic models to provide accurate and actionable diagnostic feedback within the context of spoken language learning and assessment, in particular, targeting the content development skill. We focus on one type of test question widely used in speaking assessment where test takers are required to first listen to and/or read stimulus material and then create a spontaneous response to a question related to the stimulus. In a high-proficiency response, critical content from the source material – referred to as "key points" – should be properly covered. We propose Transformer-based models to automatically detect absent key points or location spans of key points present in a response. Furthermore, we introduce a multi-task learning approach to measure how well a key point is rendered within a response (quality score). Experimental results show that automatic models can surpass human expert performance on both tasks: for span detection, the system performance reached an F1 score of 74.5% (vs. human agreement of 68.3%); for quality score prediction, system performance reached a Pearson correlation coefficient (r) of 0.744 (vs. human agreement of 0.712). Finally, the proposed key point-based features can be used to predict speaking proficiency scores with a correlation of 0.730.

Index Terms: spoken language learning and assessment, targeted content feedback, key point, Transformer

1. Introduction

When evaluating a language learner's spontaneous speech production, a wide range of speech dimensions must be evaluated, including aspects of fluency, pronunciation, rhythm, vocabulary range, grammatical accuracy, content appropriateness, and discourse organization [1]. Systems for automated scoring of speech have focused predominantly on aspects of fluency, pronunciation and prosody [2, 3], and to a lesser extent on aspects of vocabulary and grammar [4, 5], content appropriateness [6, 7, 8, 9], and discourse coherence [10, 11]. In the dimension of spoken content, features that measure the overall content appropriateness of a spoken response to a test question have been proposed [6, 8, 9], but more targeted assessment of content coverage and correctness that go beyond a generic measure of topicality has been underexplored. Nevertheless, assessing to what extent a test taker produces particular aspects of content in a spoken response would not only enable more precise measures of spoken proficiency in this content domain but furthermore also enable language learners to obtain diagnostic/targeted feedback automatically from such systems. These systems could indicate, for instance, which critical content (hereafter referred to as key points) is present or absent from the learner's response and provide interactive guidance to language learners to improve the content aspect of their responses to a particular question. Besides, key point-based features can be introduced into an automated spoken language assessment system to measure the content coverage and correctness of spoken responses.

In this paper, we are explicitly investigating the extent to which we can automatically determine the presence of key information in a test taker's response to a particular question, where the test developers determine ahead of time a set of essential key points that should be present in a high-proficiency response. Test questions and related responses are drawn from a large-scale standardized international language assessment. Each test question first presents a listening and/or reading passage to the test taker, then asks the test taker to formulate a one-minute spoken response by integrating relevant information from the provided listening and/or reading stimulus materials. Selected test takers' responses were annotated in order to identify which passages cover the necessary key points, and how well each was covered.

Related work has been conducted into automated ways of identifying whether a certain key point from a test question is rendered in spoken responses [12, 8]. The methods used included statistical models based on character and word n-grams [12], word embeddings [12], and siamese deep neural networks [8]. However, Yoon and Lee [8] still aimed to generate a generic measure of topical overlap between concatenated key points and a test response, which cannot identify missing key points and accordingly cannot generate desired diagnostic feedback. Alternatively, Yoon et al. [12] detected missing key points by building a binary classification model according to each individual key point, where n-grams and word embeddings were used to measure the topic relevance between a test response and a key point. In contrast to this related work, we formalize the task of automatic generation of targeted content feedback on spontaneous speech as 1) detecting the presence/absence of each predefined key point and locating the span of each one that appears in a test response; and 2) predicting quality score for each detected key point, which can indicate how well a key point is rendered in a test response.

The recent rapid progress in the field of natural language processing, in particular with the ubiquitous transformer architecture [13], makes it possible to generate reliable targeted content feedback that can be used by test takers to improve their content development performance. In this study, we use a setup similar to a question answering task with the purpose of key point span detection, and then use a multi-task learning strategy [14] to jointly optimize both key point span detection and quality score prediction. Two popular Transformer-based models were employed in this work, namely, BERT [15] and RoBERTa [16]. The motivation for using these Transformers is not only their high performance on many diverse natural language processing tasks, but also and in particular their ability to use only comparatively small annotated data sets for supervised

http://dx.doi.org/10.21437/Interspeech.2020-1766

fine-tuning after unsupervised pre-training on a large unlabeled data set. Furthermore, as our corpus contains spontaneous non-native speech, a particular key point can be rendered in many different ways by a test taker (or language learner), and hence we are also relying on Transformers' ability to generalize from particular example instances to semantically similar renderings in unseen evaluation data.

The contributions of this paper are: 1) formalizing a key point detection task towards the application of an automatic spoken language learning and assessment system; 2) building automatic detection models based on Transformers, which can significantly outperform human performance on the task of key point identification; 3) improving the language representation with more in-domain unlabeled data, which can further improve the target downstream key point detection performance; 4) using a multi-task learning approach to jointly optimize both key point span detection and quality score prediction; 5) with the ultimate goal of spoken language learning and assessment, designing targeted content feedback based on identified key points, and introducing key point-based features into an automatic spoken language assessment system.

2. Data and Annotation

2.1. Integrated Tasks and key points

In many large-scale English spoken language assessments, one type of widely used task is called an integrated test item. Such items ask test takers to first listen to and/or read stimulus materials, then construct a spoken response to a related test question. As the name suggests, these items require test takers to integrate multiple language skills (listening/reading and speaking) in a substantial way to complete the task. In the field of language testing, research has repeatedly shown that human raters pay considerable attention to speech content while scoring [17, 18]. Accordingly, this study focuses on providing a reliable way to generate targeted content feedback with the goal of automatic spoken language learning and assessment.

When test takers integrate stimulus materials to create a spoken response on an integrated item, a critical measure of content coverage and correctness is the degree to which the source materials can be accurately reflected/reproduced. Accordingly, key points can be defined as the critical content from the source materials that should be properly rendered in a high-proficiency response to a related test question. Research in language testing has shown a clear positive relationship between the number of key points covered and proficiency levels [19, 20].

2.2. Data Annotation

The data used in this study consisted of one-minute responses to four integrated test items from a large-scale standardized international language assessment. According to each test item, a list of six key points related to the listening and/or reading stimulus materials was identified in advance by test developers and English language learning experts. During the operational test, responses were scored by expert human raters on a four-point scale ranging from 1 (lowest proficiency) to 4 (highest proficiency). In total, 960 responses were collected and balanced according to test questions and proficiency scores; thus, there are 60 responses per item at each score level.

Two experts in the domains of language teaching and assessment then annotated the human transcriptions of these responses. The annotations fell into two categories: ratings and text spans. First, for each of the six key points, the annotators rated every response on a three-point scale, where 1 indicated full coverage of the relevant key point, 0.5 indicated partial coverage, and 0 indicated no coverage; these ratings served as the quality score for each key point. Second, the annotators identified the spans of text from a response which covered each key point. For missing key points, no spans were annotated.

Among the 960 responses, 400 were selected for double-annotation, i.e., 100 from to each test item, and the remaining 560 responses were split approximately evenly between the two annotators and received single-annotation from either annotator. In the following sections, the 560 single-annotated responses were taken as the training set, and the 400 double-annotated responses were taken as the test set. Regarding the development of key point detection models, since each response was annotated with six key points, in total, there are 3,360 and 2,400 samples in the training and test sets, respectively.

3. Method

3.1. Task Setup

Given a test response and a related key point, the task of automatic key point detection is to detect the span of the key point if it is covered in the response; otherwise, the key point's absence is detected. This can be analogous to a typical question answering task that has been widely studied in the field of natural language processing, i.e., SQuAD V2.0 [21]. SQuAD (Stanford Question Answering Dataset) [22] is a reading comprehension data set, where questions were asked on a set of Wikipedia articles, and the answer to every question is a segment of text (span) from the corresponding reading passage. Especially in SQuAD V2.0, some questions might be unanswerable.

In this work, we use a setup similar to SQuAD V2.0, where the macro-averaged F1 score was used as the evaluation metric [22]. F1 measures the average overlap between the predictions and ground truth, ignoring punctuation as well as articles, and with the prediction/ground truth taken as bags of tokens. Compared with the answers in SQuAD, key point spans in our task tend to be longer narrative sentences, and the average number of words within spans is around 18.3 (sd = 13.8). Furthermore, as described in Section 2.2, each identified key point was assigned a quality score in a range from 0 to 1 by human experts. Therefore, a regression model can also be built to measure how well a key point is rendered within a response, and the Pearson correlation coefficient of automated scores with manual scores can be used as the evaluation metric.

3.2. Transformer-based Models

Most of the state-of-the-art Transformers have been examined on SQuAD V2.0, which has advanced the state-of-the-art and achieved superior results compared to human performance[1]. Hence in this work, for the key point detection task, we adopted the Transformer architecture [13], which will not be reviewed in detail here due to space constraints. In particular, two Transformer-based models, i.e., BERT (Bidirectional Encoder Representations from Transformer) [15] and RoBERTa (Robustly Optimized BERT Approach) [16], were explored to build the automatic detection models.

BERT is an important milestone in natural language processing, which greatly boosts performance across almost all major tasks, including SQuAD 2.0 [15]. Different from earlier

[1] https://rajpurkar.github.io/SQuAD-explorer/

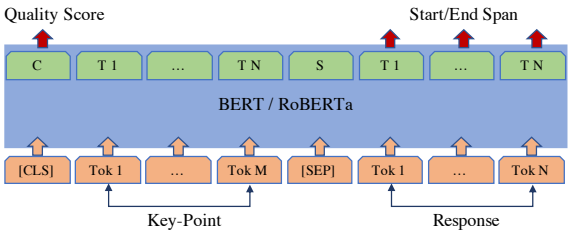

Figure 1: *BERT/RoBERTa in key point detection task.*

language representation models like ELMo [23] and OpenAI GPT [24], BERT can pre-train deep bidirectional representations from unlabeled texts by introducing a pre-training objective known as "masked language model" (MLM), which can alleviate the unidirectionality constraint and jointly condition on both left and right contexts in all layers.

With a pre-trained model, the self-attention mechanism in the transformer architecture makes it straightforward to further fine-tune on downstream tasks, such as key point detection. As shown in Figure 1, at the input, concatenated pairs of key points and test responses are plugged in. At the output, one additional layer is added to predict the span of a key point. Especially for a missing key point, a span which both starts and ends at the first special token [CLS] will be returned. In addition, since each labeled key point is associated with a quality score, we also model the additional layer to predict the quality scores using the aggregate representation from the final hidden vector C corresponding to the special token [CLS]. The detection model is first initialized with pre-trained parameters, and then all parameters are fine-tuned with task-specific labeled data.

RoBERTa (Robustly optimized BERT approach) [16] is an alternative version of BERT with an improved training recipe, including the application of a dynamic masking strategy on the input training data; dropping the NSP objective and training on longer sequences; increasing the size of mini-batches and training models longer; as well as some other changes to design choices and training strategies. The MLM objective is used to pre-train both BERT and RoBERTa models.

Recently another objective, such as permutation language modeling in XLNet [25], has been proposed to overcome the issues introduced by MLM, such as the neglect of dependency between masked positions and the discrepancy between pre-training and fine-tuning by introducing the artificial token [MASR]. However, the RoBERTa work [16] re-established that MLM is still competitive with other recently proposed methods, such as XLNet. Accordingly, we decided to explore BERT and RoBERTa with our key point modeling task.

4. Experiments and Discussion

4.1. Experimental Setup

We used the implementation from Hugging Face [26] to build the detection models, and experimented with BERT and RoBERTa models in both base and large sizes[2], which were pre-trained on a large amount of written texts from BooksCorpus [27], English Wikipedia, and other text corpora. With L as the number of layers (i.e., Transformer blocks); A as the number of self-attention heads; and H as the hidden size, the four experimented models are BERT_Base (L = 12; A = 12; H = 768; 110M parameters), BERT_Large (L = 24; A = 16; H = 1024; 340M parameters), RoBERTa_Base (L = 12; A = 12; H = 768; 125M parameters), and RoBERTa_Large (L = 24; A = 16; H = 1024; 355M parameters). All four models were fine-tuned with six epochs on the downstream task, and the number of warmup steps is set to be around 10% of the total steps. In order to make parallel comparisons with human experts' agreement, manual transcriptions were used as the input for model training and evaluation.

[2]https://huggingface.co/transformers/pretrained_models.html. The four experimented pre-trained models correspond to bert-base-uncased, bert-large-uncased-whole-word-masking, roberta-base, roberta-large.

Table 1: *Key point detection performance in terms of F1 score on fine-tuned base/large BERT/RoBERTa models. The human agreement is also listed for comparison.*

Models	Base	Large
BERT	69.7	71.7
RoBERTa	70.0	71.9
Human Agreement	68.3	

4.2. Key Point Detection Results

4.2.1. Span Detection

We first focus on the task to detect the spans of key points without predicting quality scores. Table 1 shows that in terms of F1 score, all four models outperform human agreement (F1 = 68.3%), and that the large models generally perform better than the base models. Therefore, in the following experiments we only report results with the large models. In addition, RoBERTa shows slightly higher performance than BERT on this task, i.e., 71.9% vs. 71.7%.

4.2.2. Improvement in Language Representation

In this study, we work with human transcriptions of non-native spontaneous speech, which is quite different from the written texts used to pre-train BERT and RoBERTa models. This mismatch may result in less satisfying language representation while applying these pre-trained models on speech data, especially for low-proficiency responses. Therefore, in order to obtain models with better language representation capabilities on speech, we collected a data set with human transcriptions on 58,291 spoken responses drawn from the same assessment[3], and used it to first fine-tune BERT/RoBERTa with MLM, where the number of training epochs was set at four, and around 10% of the total steps were used for warmup. Afterwards, the obtained in-domain models were further fine-tuned on the downstream span detection task with labeled data. The experimental results indicate that adding more in-domain unlabeled data can greatly benefit the downstream task; the F1 scores can be improved from 71.7% to 74.5% for BERT, and from 71.9% to 73.3% for RoBERTa respectively. Therefore, the models fine-tuned with in-domain data were adopted in the following experiments.

4.2.3. Multi-task Learning

Previous research has demonstrated that multi-task learning can benefit deep learning applications by jointly optimizing regression and/or classification objectives across multiple tasks [14, 28]. In this work, as shown in Figure 1, a Transformer-

[3]There was no overlap in test takers between this large data set and the annotated data in our study.

Table 2: *Performance improvement by introducing multi-task learning, where both the span detection and quality score prediction tasks are jointly optimized. F1 scores for span detection and Pearson correlation coefficients (r) between automatic and manual scores are provided.*

Models	F1 (%)	r
BERT_Large_inDomain	74.5	None
BERT_Large_inDomain_Multi	74.5	0.739
RoBERTa_Large_inDomain	73.3	None
RoBERTa_Large_inDomain_Multi	74.1	0.744
Human_Agreement	68.3	0.712

based model can be built to complete both the automatic detection of key point spans and the automatic prediction of key point quality scores at the same time, where the span detection task uses cross-entropy loss, and the scoring task uses mean square error loss. Since each task's loss may range on a different scale, it is important to weight relatively between losses of multiple tasks. However, tuning these weights by hand is expensive; thus, we followed the method proposed in [14] to automatically weight multiple loss functions by considering the homoscedastic uncertainty of each task. Table 2 shows that by conducting multi-task learning, the performance on span detection can be improved with RoBERTa from 73.3% to 74.1%, but no further improvement can be obtained with BERT. Moreover, the Pearson correlations coefficients between automatic scores and manual key point quality scores are 0.739 for BERT and 0.744 for RoBERTa respectively, which are higher than the correlation with human agreement of 0.712.

4.3. Targeted Content Feedback

In order to develop an automatic tool that can provide actionable diagnostic feedback used by language learners, it should meet several requirements, such as they can accurately identify errors of learner performance, they should be meaningful, easily interpretable, and actionable to users, and they can lead to gains in targeted areas of language ability [29]. In this study, we developed such a tool with the capability to provide targeted content feedback. The Transformer-based models can detect the missing pieces of key points within test takers' responses. They can also identify the locations of presented key points and determine whether they are properly rendered in the spoken response. Experimental results have demonstrated that automatic models can outperform human experts' agreement on this task. Hence the tool proposed in this paper meets the requirements of "accurate" and "actionable". In the future, a user study will be conducted to verify how much gain can be obtained in improving language learners' speaking skills, in particular, related to the content/topic elaboration/development.

4.4. Automated Speech Scoring

In this work, we designed a set of key point-based features to measure the content coverage and correctness of non-native spoken responses within an automated speech assessment system. There are six key points defined for each integrated test question; accordingly, six features can be defined as the six quality scores, one for each key point (with 0 for absent key points); furthermore, the quality scores can be summed together as an additional feature to measure the overall quality. As shown in Table 2, the *RoBERTa_Large_inDomain_Multi* model can achieve a relatively higher correlation with quality scores; thus, automatic features were extracted with predictions generated by this model and evaluated in terms of Pearson correlation coefficients with human proficiency scores. Experimental results show that the features corresponding to six key points can achieve correlations with human proficiency scores in a range from 0.356 to 0.628. In particular, the feature for the last[4] key point can obtain a correlation of 0.628, since it generally contained more elaborated content depending on the nature of the test item. Finally, the summed feature (i.e., the sum of individual quality scores across all six key points) can achieve a correlation as high as 0.670.

We examined the proposed key point features within an automated spoken English assessment system, SpeechRater® [3]. The task is to build effective scoring models, which can automatically predict holistic proficiency scores by measuring different aspects of non-native speaking proficiency. The baseline scoring model was built with 28 automatic features extracted from the SpeechRater system, which can measure the pronunciation, prosody, fluency, rhythm, vocabulary, grammar, and cohesion of spontaneous speech. We used the Random Forest Regression method from the machine learning tool scikit-learn[5] [30] to build the scoring models, and 10-fold cross-validation was conducted on the test partition with 400 responses. The baseline system using only SpeechRater features can achieve a correlation of 0.832 with human proficiency scores. In comparison, the automatic model using only key point features can achieve a correlation of 0.730. Furthermore, by combining both SpeechRater and key point features, the correlation can be improved to 0.843. All experimental results have demonstrated the effectiveness of the proposed key point-based features in an automated speech scoring system.

5. Conclusion

In this paper, a key point detection task was proposed with the purpose of generating accurate and actionable targeted feedback for English language learners to improve their speaking skills in terms of content development. Transformer-based models were built to detect missing key points, as well as text span locations and quality scores of the covered key points. The developed models can outperform human agreement on these detection tasks. Moreover, the derived key point-based features can improve an automated speech scoring system by measuring content coverage and correctness. We will continue this line of work in the future, focusing on verifying the generalizability and robustness of the proposed key point detection models by 1) using a larger data set with more test items so that models can be examined on both seen and unseen test items; and 2) examining model robustness by using automatic speech recognition outputs instead of manual transcriptions as inputs to the model.

6. Acknowledgements

We would like to thank Yuan Wang, Florencia L. Tolentino, and Ching-Ni Hsieh from Educational Testing Service (ETS). They greatly supported our research in the manual annotation of our data set.

[4]Based on the order in which particular key points appear in a test item's stimulus materials.

[5]SKLL, a Python tool that simplifies the carrying out of scikit-learn experiments, was used. Downloaded from `https://github.com/EducationalTestingService/skll`.

7. 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 *Eleventh annual conference of the international speech communication association*, 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 *2018 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 2012 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies*, 2012, pp. 103–111.

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

[9] 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*, 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*, 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 *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 Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 2019, pp. 306–315.

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems 30*, 2017, pp. 5998–6008.

[14] 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*, 2018, pp. 7482–7491.

[15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Jun. 2019, pp. 4171–4186.

[16] 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.

[17] 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.

[18] 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*, vol. 2005, no. 1, pp. 1–157, 2005.

[19] C. E. Kellie Frost 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.

[20] 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.

[21] P. Rajpurkar, R. Jia, and P. Liang, "Know what you don't know: Unanswerable questions for SQuAD," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Jul. 2018, pp. 784–789.

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

[23] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in *Proceedings of NAACL-HLT*, 2018, pp. 2227–2237.

[24] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding with unsupervised learning," *Technical report, OpenAI*, 2018.

[25] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," in *Advances in neural information processing systems*, 2019, pp. 5754–5764.

[26] 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.

[27] 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.

[28] X. Liu, P. He, W. Chen, and J. Gao, "Multi-task deep neural networks for natural language understanding," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 4487–4496.

[29] X. Xi, "Automated scoring and feedback systems: Where are we and where are we heading?" *Language Testing*, vol. 27, no. 3, pp. 291–300, 2010.

[30] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *Journal of machine learning research*, vol. 12, no. Oct, pp. 2825–2830, 2011.