Automated Assessment of Task Completion in Spontaneous Speech for Finnish and Finland Swedish Language Learners

Ekaterina Voskoboinik, Yaroslav Getman, Ragheb Al-Ghezi, Mikko Kurimo, Tamás Grósz
Department of Information and Communications Engineering
Aalto University, Finland
firstname.lastname@aalto.fi

Abstract

This study investigates the feasibility of automated content scoring for spontaneous spoken responses from Finnish and Finland Swedish learners. Our experiments reveal that pre-trained Transformer-based models outperform the tf-idf baseline in automatic task completion grading. Furthermore, we demonstrate that pre-fine-tuning these models to differentiate between responses to distinct prompts enhances subsequent task completion fine-tuning. We observe that task completion classifiers exhibit accelerated learning and produce predictions with stronger correlations to human grading when accounting for task differences. Additionally, we find that employing similarity learning, as opposed to conventional classification fine-tuning, further improves the results. It is especially helpful to learn not just the similarities between the responses in one score bin, but the exact differences between the average human scores responses received. Lastly, we demonstrate that models applied to both manual and ASR transcripts yield comparable correlations to human grading.

1 Introduction

The assessment of content is an important dimension of oral proficiency evaluation. It complements other areas like fluency, pronunciation, and the range and accuracy of grammar and vocabulary (Brown et al., 2005). This work examines the automatic evaluation of content by scoring task completion. A successful response should demonstrate both comprehension of the prompt and mastery in speech production, making task completion an important component of oral proficiency assessment.

The research in automated scoring of non-native English speech has shown that it is possible to automatically evaluate the content relevance of a response (Yoon and Lee, 2019). It was demonstrated that fine-tuning Transformer-based models is especially beneficial for this task (Wang et al., 2020).

The present study aims to evaluate the potential of BERT models (Devlin et al., 2019) for content scoring of non-native Finnish and Finland Swedish spontaneous speech. Additionally, we explore the effectiveness of fine-tuning BERT for task classification to enhance performance in subsequent fine-tuning for task completion. Given the multi-modal nature of our prompts, we find it challenging to map them to the same vector space as our responses for prompt awareness as in (Wang et al., 2021b). Consequently, we integrate task classification to inform the model about different tasks. Our choice to experiment with fine-tuning for an intermediate task is based on previous findings, which showcased improved robustness and effectiveness in the resulting target task model, particularly in low-resource scenarios (Phang et al., 2019). Our experiments reveal that this approach accelerates learning for task completion evaluation and leads to better correlations with human scores.

Due to the limited size and imbalance of our datasets, we further explore the use of similarity learning. We fine-tune BERT in a Siamese manner in two ways: first, to place responses that belong to the same task completion score bin closer together and those that belong to different score bins further away; second, to learn to position responses proportionately to the distance of their average task

Ekaterina Voskoboinik, Yaroslav Getman, Ragheb Al-Ghezi, Mikko Kurimo and Tamas Grosz. Automated Assessment of Task Completion in Spontaneous Speech for Finnish and Finland Swedish Language Learners. *Proceedings of the 12th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2023)*. Linköping Electronic Conference Proceedings 197: 102–110.

completion scores. Our results indicate that treating response scores as continuous numbers instead of bin categories leads to better correlation with human scores.

2 Related Work

The progress of research in content scoring of spontaneous non-native speech was initially hindered by the quality of ASR systems. Early approaches (Xie et al., 2012; Chen, 2013) explored techniques developed for automatic essay scoring. Typically, a vector space model like tf-idf, LSA (Landauer et al., 1998), or PMI (Turney, 2001) would be trained on a set of pre-graded responses for each prompt. The tasks would be represented by vectors for every score category. The to-be-graded response is then mapped to the same vector space and compared to the score vectors. The similarities between response and score vectors were used as content features for holistic grade prediction. However, this approach had several drawbacks. It relied on a large number of pre-graded responses to build a reliable vector space and did not take word relations into account. It was shown in (Loukina et al., 2014) that for tasks like giving a summary of a prompt material, ROUGE (Lin, 2004) would outperform tf-idf similarity and needed fewer reference responses. And (Evanini et al., 2013) demonstrated that comparing responses and prompts is a viable option even though it was slightly outperformed by comparison to pre-graded responses.

The exploration of more context-aware vector representations, such as doc2vec, demonstrated a higher correlation to holistic scores compared to tf-idf based approaches (Tao et al., 2016). The work in (Yoon et al., 2018) continued the research started in (Evanini et al., 2013) by comparing tf-idf and averaged word2vec embeddings for computing similarities between responses and prompts. The pre-trained embeddings proved more advantageous than tf-idf.

More recently, it was demonstrated that neural and pre-trained approaches are highly effective in scoring content relevancy. In one study (Qian et al., 2018), the authors used an attention LSTM-RNN model to directly score the proficiency level of a response based on its transcript. They found that conditioning the model on task prompts led to even better performance. Similarly, the authors of (Yoon and Lee, 2019) compared a Siamese CNN model to a tf-idf based one and found that the former outperformed the latter when predicting holistic proficiency scores based on the similarity between responses and a set of key points generated by experts for each task. Taking things further, (Wang et al., 2020) trained multi-task Transformer-based models that were able to detect missing key points or the spans of present key points and predict how well each present key point was communicated in a response. These models outperformed human agreement on these tasks. The success of Transformer-based models was further supported by experiments in (Wang et al., 2021b), which showed that fine-tuning BERT and XLNet for holistic proficiency scoring using only ASR response transcripts already surpassed human agreement. Additionally, augmenting the models with prompt awareness led to even better results.

Inspired by these findings, this study explores the capabilities of pre-trained BERT models for scoring content appropriateness of Swedish and Finnish learners' oral responses.

3 Data

This study investigates content relevancy scoring using two corpora of non-native spontaneous speech: Finnish and Finland Swedish (Al-Ghezi et al., 2021, 2023). The Swedish data was collected from upper secondary school students, while the Finnish data contains responses from both upper secondary school students and university students. The datasets include responses to semi-structured and open-ended tasks, such as reacting to a text or a picture prompt or simulating a phone call by answering pre-recorded questions.

Originally, the recordings were rated by humans across the following dimensions: holistic level, pronunciation, fluency, accuracy, range, and task completion (Al-Ghezi et al., 2023). The raters were asked to either assign a score for each dimension or mark a dimension as ungradable (zero). In our experiments, we include only the recordings that received non-zero scores from all raters across all criteria. Additionally, one task from the Swedish dataset was excluded, as it contained only two responses.

This work is focused on automatically assessing task completion (TC) criterion as a measure of content relevancy. Task completion was rated on a scale of 1 to 3, where 1 indicates that the as-

signment was answered only partially with many significant gaps in the response, and 3 signifies that the test-taker fulfilled the assignment excellently with no significant gaps in the response. The responses that received multiple human assessments were assigned an average of those assessments. We used binning to convert the average scores back to discrete classes. The range of scores from 1 to 3 was divided into three equal intervals, and each score was labeled based on the interval it fell into. In this study, we explore both continuous and binned scores. The data described in this study will be published in The Language Bank of Finland (FIN-CLARIN) [1].

To establish a reference for human agreement, we compared the scores of all recordings assessed by at least two raters. We report the Spearman correlation coefficient and Quadratic Weighted Kappa between two random raters in Table 1. The measures suggest a fair level of agreement. These numbers indicate that assigning task completion scores can be a challenging task for human raters. The Swedish samples were evaluated by 18 human raters, with 101 samples rated by one rater, 1358 samples rated by two raters, 42 samples rated by three raters, and 39 recordings rated by five raters. The Finnish recordings were rated by 25 raters, with 302 samples rated by one person, 1790 samples rated by two people, and 24 samples rated by three raters.

	cor	**kappa**
Swedish	0.372	0.377
Finnish	0.298	0.340

Table 1: Spearman correlation coefficient (cor) and Quadratic Weighted Kappa (kappa) between two random raters for Swedish and Finnish data.

Table 2 describes the overall statistics of the corpora. However, these numbers vary from task to task. For instance, the duration of responses is highly task dependent. In the Swedish dataset, the task that elicits the longest answers has responses averaging 26.4 seconds, while the task with the shortest answers has responses averaging about 4.2 seconds. In the Finnish dataset, the task eliciting the shortest answers on average has responses of 3.2 tokens, and the task eliciting the longest answers has an average response length of 91 tokens. The distribution of scores varies between the tasks as well. In the Swedish data, the task with the highest-scored responses has an average score of 2.8, while the task with the lowest-scored responses has an average score of 1.5. In the Finnish data, the lowest average score for task completion in a task is 2.1, and the highest average score in a task is 2.9.

	Swedish	**Finnish**
# of samples	1540	2112
# of students	178	308
# of tasks	21	25
avg. TC score	2	2.6
total duration (h)	5.6	14.1
# of samples per task		
min.	30	6
max.	110	173
avg.	73.3	72.8
Response duration		
min. (s)	1.1	2
max. (s)	30.7	91
avg. (s)	13	24
Response length (words)		
min.	1	1
max.	49	228
avg.	9.4	31.6

Table 2: Dataset statistics.

The distribution of task completion scores is quite unbalanced. This problem is the most pronounced for the Finnish dataset: the average task completion score is 2.6, which indicates the prevalence of high-scoring responses. Moreover, there are five tasks with no responses in the lowest score bin. In total, 17 out of 29 tasks have less than 5% of responses with the lowest score bin. The distribution of scores in the datasets can be found in Table 3.

	1	**2**	**3**
Swedish	517	368	655
Finnish	134	339	1639

Table 3: Score bin distributions of Swedish and Finnish data.

[1] https://www.kielipankki.fi

4 Methods

4.1 Baselines

First, we evaluate the ability of out-of-the-box BERT and tf-idf-based vector spaces to represent the differences between high and low-scoring responses. We will use their performance as our baselines.

For training tf-idf models, we generated task documents from all the responses to each prompt and derived the inverse document frequency (idf) from them. Each response in the dataset was then mapped to a vector by weighing its word counts (tf) by the idf. To obtain response representations using BERT models, we applied mean pooling to the outputs of the final layer, since (Reimers and Gurevych, 2019) demonstrated that it produces better representations than other pooling strategies.

4.2 Task classification fine-tuning

In our first experiment, we fine-tuned the model to classify the recordings according to the tasks they were answering using Siamese fine-tuning. We opted for this approach due to its efficiency, as it enabled us to leverage the weights already learned by the model rather than requiring it to learn the weights for a classification head from scratch. The goal of this fine-tuning stage is to place the responses to the same prompt closer to each other and further away from the responses to other prompts. While we were not primarily interested in the model's performance for this problem, we focused on adjusting the final embeddings. We measured the changes in cosine distances between task centroids and in the properties of task clusters. To establish how well different categories of responses are represented in a vector space we use the Calinski-Harabasz score (Caliński and Harabasz, 1974). It measures the ratio of between-cluster dispersion to within-cluster dispersion. The score gets higher when data points are close to each other within the same cluster and are far from other clusters' centroids. In other words, the Calinski-Harabasz score measures the separation of vector classes in a space. We would like to have a high Calinski-Harabasz score when measuring the distance between responses belonging to different tasks.

We trained the models using positive and negative examples of responses to the same task. Each response in our dataset was paired with one positive example and five negative examples. The positive example was randomly selected, while negative examples were chosen based on their level of "hardness" (closest responses from other tasks were selected). Similarly to our BERT baseline, we embed a response in a vector space using mean pooling.

4.3 BERT with a classification head

To investigate the impact of pre-fine-tuning for task classification on subsequent task completion fine-tuning, we compared BERT models trained for task completion before and after task classification fine-tuning. We employed a linear classification head preceded by dropout. The head receives a vector obtained by mean-pooling, as this was the representation learned during task classification.

4.4 BERT Siamese

We further sought to experiment with similarity learning as an alternative to classic fine-tuning for our limited and imbalanced datasets, following previous findings of its potential benefits (Schroff et al., 2015). Our goal was to adjust the vector space so it would place higher scored responses further away from lower scored responses. For these means, we experiment using both score bins and average scores to learn similarities between the responses.

To learn response similarity using score bins, we generated pairs of samples from each response within a task. A pair received a label of 1 if both samples belonged to the same score bin and 0 if they originated from different bins. To train using average grades, we assigned the desired cosine distances in the range of 0-1 based on the differences between the samples' scores. For instance, a pair consisting of a sample with a score of 1 and a sample with a score of 3 would be assigned a cosine distance label of 1. On the other hand, a pair with samples having scores of 1 and 2 would receive a cosine distance label of 0.5.

5 Experiments and Results

5.1 Speech-to-text

For the experiments, we employed a 4-fold cross-validation strategy to evaluate our models. In this approach, each model was trained on three folds and evaluated on the remaining fold. The folds were designed by creating four non-overlapping

student sets. Furthermore, we stratified the folds by tasks and holistic levels, ensuring that every task was represented in each split.

In this work, we used wav2vec 2.0 models (Baevski et al., 2020) to produce ASR transcripts for the responses. For L2 Finland Swedish, we used a monolingual Swedish model that was pretrained on 11.5K hours of unlabeled speech from the collections of the National Library of Sweden (Malmsten et al., 2022), such as local radio broadcasts and audiobooks, and fine-tuned on the Common Voice (Ardila et al., 2020) and the NST (Birkenes, 2020) corpora. For Finnish ASR experiments, we used a multilingual model pretrained on the Uralic (Finnish, Estonian, and Hungarian) subset of the European parliamentary session recordings collection called Voxpopuli (Wang et al., 2021a) and fine-tuned on a 100-hour subset of the Finnish colloquial speech dataset Lahjoita Puhetta (Donate Speech) (Moisio et al., 2022). The models were further fine-tuned on the target data with 4-fold cross-validation mentioned above. After aggregating the test set outputs produced by each of the 4 sub-systems, the total word and character error rates are 17.71% / 9.08% and 21.89% / 7.06% for the L2 Finland Swedish and the L2 Finnish data, respectively (Al-Ghezi et al., 2023).

5.2 Baselines

For tf-idf models, we utilized the TfidfVectorizer from the scikit-learn Python package (Pedregosa et al., 2011). As for BERT representations, we used FinBERT[2] trained by (Virtanen et al., 2019) for the Finnish part of the data and a BERT model trained by National Library of Sweden[3] for the Swedish part.

We evaluate the models using simple k-NN classifiers, where a response is assigned a score based on its similarity to reference vectors. We compare two approaches for selecting these reference vectors: either using bin centroids (CTR) or all historical responses to a task prompt (1-NN). In the first approach, each score bin in a task is represented by the mean embedding of its responses. A new response is then assigned a score based on its closest score bin vector. In the second approach, a test response is compared to all prior responses given to a prompt and assigned the score of the nearest one. Due to data imbalance, we opted for only one nearest neighbor in this experiment, as selecting more than one neighbor could prevent our system from recognizing underrepresented score intervals.

	Human		ASR	
	cor	kappa	cor	kappa
Swedish				
tf-idf CTR	0.381	0.360	0.392	0.373
tf-idf 1-NN	0.561	0.491	0.537	0.462
BERT CTR	0.451	0.439	0.445	0.431
BERT 1-NN	**0.580**	**0.524**	**0.560**	**0.500**
Finnish				
tf-idf CTR	0.213	0.242	0.253	0.275
tf-idf 1-NN	0.170	0.196	0.199	0.220
BERT CTR	**0.286**	**0.313**	**0.279**	**0.305**
BERT 1-NN	0.259	0.232	0.277	0.248

Table 4: Spearman correlation coefficient (cor) and Quadratic Weighted Kappa (kappa) of Baseline Models.

We assess performance by comparing the predicted scores with human scores using two metrics: the Spearman correlation coefficient between average human scores and predicted scores, and the Quadratic Weighted Kappa between binned average human scores and binned machine scores. The results can be found in Table 4. Here, we see that BERT models outperformed tf-idf models for both Swedish and Finnish. The strategy of assigning a score based on a single nearest neighbor proved to be more effective for Swedish, but it was less successful than using bin centroid vectors for Finnish. Finally, models applied to ASR transcripts demonstrated results comparable to those of human transcripts, with the correlations to human scores being only marginally lower for the best-performing approaches.

5.3 Task Classification

The models were trained with SentenceTransformers Python package (Reimers and Gurevych, 2019), using Contrastive loss (Chopra et al., 2005) with a margin of 0.5. To achieve vector spaces with similar properties in order to keep the models comparable in the subsequent experiments, the Swedish model was trained for 4 epochs, and the Finnish model was trained for 5 epochs. Each fold was trained with 50 warm-up steps for every new epoch. We used a batch size of 16. The prop-

[2]https://hf.co/TurkuNLP/bert-base-finnish-cased-v1
[3]https://hf.co/KBLab/bert-base-swedish-cased-new

	BC distance	Task cluster score
SWE	0.11	20
SWE ft	0.66	1676
FIN	0.18	58
FIN ft	0.66	1762

Table 5: Properties of out-of-the-box models vs the models fine-tuned (ft) for task classification. We report average cosine distances between bin centroids (BC) and Calinski-Harabasz score (Task cluster score).

erties of the resulting vector spaces are described in Table 5. The task cluster scores have significantly improved from 20 to 1676 for Swedish, and from 58 to 1762 for Finnish. The average cosine distance between the task centroids also went up from 0.11 to 0.66 for Swedish, and from 0.18 to 0.66 for Finnish.

5.4 Task completion with a classification head

For this experiment, we either trained the models described in the previous subsection or used the models explored as BERT baselines. We then fine-tuned the models with HuggingFace's Transformers library (Wolf et al., 2020), using dropout with 0.1 probability, a learning rate of 2e-5, and a batch size of 4. For the models initialized with a baseline BERT, we used 15 epochs for Swedish, and 9 epochs for Finnish. For the models that were pre-trained with task classification, we used 3 epochs for Swedish and 4 epochs for Finnish. Here and in the next section the number of reported epochs indicates the epoch after which the performance stopped improving with more training. One can notice that pre-fine-tuning results in fewer epochs needed for further fine-tuning.

The results of fine-tuning BERT for task completion classification with (cls_task) and without (cls_no_task) task classification pre-fine-tuning showed strong favor for task classification pre-fine-tuning. The results can be found in Table 6.

5.5 Task completion Siamese

In this part, we continue to fine-tune the models trained on task classification problems. For learning score bin similarity we have applied Contrastive loss with 0.5 margin. For learning distances between average task completion, mean squared-error loss was employed as the objective function. We used a batch size of 16 and 50 warm-up steps for every fold in every new epoch. All models were trained for 2 epochs. For task completion scoring, we used 1-NN approach.

	Human		ASR	
	cor	kappa	cor	kappa
Swedish				
cls_no_task	0.530	0.507	0.507	0.486
cls_task	0.603	0.584	0.601	0.583
S_bins	0.656	0.617	0.658	0.611
S_cosine	**0.714**	**0.650**	**0.679**	**0.623**
Finnish				
cls_no_task	0.271	0.336	0.242	0.299
cls_task	0.295	0.325	0.286	0.308
S_bins	0.291	0.328	0.286	**0.357**
S_cosine	**0.390**	**0.365**	**0.368**	0.354

Table 6: Results of task completion fine-tuning. cls stands for BERT with classification head, task stands for task classification pre-finetuning, S is short for Siamese.

In Table 6, we demonstrate that employing similarity learning further enhances the results of task completion scoring. It is particularly advantageous to organize the space not only by score bins of the responses but also by the distance proportional to the difference in task completion scores between the responses. Again, while the correlation to human scores is higher when using manual transcripts for the best-performing approach, the results for ASR transcripts are close.

For a more comprehensive understanding of the technical aspects involved in our experiments, we encourage interested readers to examine our scripts[4].

6 Discussion

In this work, we explore different approaches to content scoring of spontaneous spoken responses of non-native Finnish and Finland Swedish learners.

As was expected, pre-trained BERT models have shown to be more efficient for our data than tf-idf baseline since they already contain language knowledge. We demonstrate that training BERT models to separate responses to different tasks before fine-tuning directly for task completion brings similar benefits to prompt awareness. The models subsequently achieve higher correlations to human scores while requiring fewer training epochs. This improvement can likely be attributed to several

[4] https://github.com/katildakat/NLP4CALL_TC

factors. Firstly, in order to accurately score task completion, a model must comprehend the typical responses associated with a specific prompt. Secondly, the data utilized for task classification fine-tuning is the same data subsequently employed for task completion fine-tuning, thereby facilitating domain adaptation.

We have also shown that similarity learning was more helpful than fine-tuning with the classification head. We believe that it happens because we can translate our data into a larger labeled set this way. It was especially beneficial not to limit the similarities between responses to their score bins, but to organize the space in accordance with how different the scores are.

Additionally, we show the applicability of our approach not only for manual transcripts but for ASR transcripts as well. Although the results of ASR transcripts are generally slightly behind the manual transcripts, they are not far off. This is an important finding since using human transcripts is not feasible in real-life applications.

Finally, we should address the differences in performance between the Swedish and Finnish models. The predictions of Swedish models correlated better with human scores than those of Finnish models. We believe that there might be several reasons for this behavior. The first one is that inter-human agreement between the raters was lower for Finnish responses than for Swedish as reported in Table 1. The second reason is that the Finnish corpus is considerably more imbalanced than the Swedish one with most of the scores receiving the highest score. For many tasks, it is impossible or almost impossible to get a score of 1, so the models, in turn, favor higher score bins.

7 Conclusions

In conclusion, this study demonstrates the effectiveness of pre-trained Transformer-based models in automated content scoring for spontaneous spoken responses from non-native Finnish and Finland Swedish learners. Our findings show that pre-fine-tuning these models to differentiate between responses to distinct prompts significantly improves task completion fine-tuning, resulting in faster learning and stronger correlations to human grading. Additionally, we discovered that similarity learning, compared to traditional classification fine-tuning, further enhances the results. It is especially useful to learn not only the similarities within responses of the same score bin but also the exact differences between the average human scores received.

Importantly, our work highlights that the performance of models applied to both manual transcripts and ASR transcripts is comparable, suggesting the feasibility of using this approach in real-life scenarios. The ability to obtain similar results with ASR transcripts enables the potential deployment of automated scoring systems in various educational contexts without the need for manual transcription, increasing efficiency and reducing costs.

For future work, we would like to explore the applicability of similarity learning in text and audio Transformers for automatic scoring of other dimensions in our assessments.

Acknowledgments

This work has been funded by the Academy of Finland grant number 322625 "Digital support for training and assessing second language speaking". The computational resources were provided by Aalto ScienceIT.

References

Ragheb Al-Ghezi, Yaroslav Getman, Aku Rouhe, Raili Hildén, and Mikko Kurimo. 2021. Self-Supervised End-to-End ASR for Low Resource L2 Swedish. pages 1429–1433.

Ragheb Al-Ghezi, Yaroslav Getman, Ekaterina Voskoboinik, Mittul Singh, and Mikko Kurimo. 2023. Automatic Rating of Spontaneous Speech for Low-Resource Languages. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 339–345.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common Voice: A Massively-Multilingual Speech Corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations.

M. B. Birkenes. 2020. NST Swedish Dictation (22 kHz). https://www.nb.no/sprakbanken/en/resource-catalogue/oai-nb-no-sbr-17/.

Annie Brown, Noriko Iwashita, and Tim McNamara. 2005. An examination of rater orientations and test-taker performance on English-for-academic-purposes speaking tasks. *ETS Research Report Series*, 2005(1):i–157.

Tadeusz Caliński and Jerzy Harabasz. 1974. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1):1–27.

Lei Chen. 2013. Applying Unsupervised Learning T Support Vector Space Model Based Speaking Assessment. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 58–62, Atlanta, Georgia. Association for Computational Linguistics.

S. Chopra, R. Hadsell, and Y. LeCun. 2005. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 539–546 vol. 1.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

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, Atlanta, Georgia. Association for Computational Linguistics.

Thomas K Landauer, Peter W Foltz, and Darrell Laham. 1998. An introduction to latent semantic analysis. *Discourse processes*, 25(2-3):259–284.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

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

Martin Malmsten, Chris Haffenden, and Love Börjeson. 2022. Hearing voices at the National Library – a speech corpus and acoustic model for the Swedish language. *arXiv preprint. arXiv:2205.03026*.

Anssi Moisio, Dejan Porjazovski, Aku Rouhe, Yaroslav Getman, Anja Virkkunen, Ragheb Al-Ghezi, Mietta Lennes, Tamás Grósz, Krister Lindén, and Mikko Kurimo. 2022. Lahjoita puhetta: a large-scale corpus of spoken Finnish with some benchmarks. *Language Resources and Evaluation*.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Jason Phang, Thibault Févry, and Samuel R. Bowman. 2019. Sentence Encoders on STILTs: Supplementary Training on Intermediate Labeled-data Tasks.

Yao Qian, Rutuja Ubale, Matthew Mulholland, Keelan Evanini, and Xinhao Wang. 2018. A Prompt-Aware Neural Network Approach to Content-Based Scoring of Non-Native Spontaneous Speech. pages 979–986.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. FaceNet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Jidong Tao, Lei Chen, and Chong Min Lee. 2016. DNN Online with iVectors Acoustic Modeling and Doc2Vec Distributed Representations for Improving Automated Speech Scoring. In *Proc. Interspeech 2016*, pages 3117–3121.

Peter D Turney. 2001. Mining the web for synonyms: PMI-IR versus LSA on TOEFL. In *Machine Learning: ECML 2001: 12th European Conference on Machine Learning Freiburg, Germany, September 5–7, 2001 Proceedings 12*, pages 491–502. Springer.

Antti Virtanen, Jenna Kanerva, Rami Ilo, Jouni Luoma, Juhani Luotolahti, Tapio Salakoski, Filip Ginter, and Sampo Pyysalo. 2019. Multilingual is not enough: BERT for Finnish.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021a. VoxPopuli: A Large-Scale Multilingual Speech Corpus for Representation Learning, Semi-Supervised Learning and Interpretation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, volume 1, pages 993–1003. Association for Computational Linguistics.

Xinhao Wang, Keelan Evanini, Yao Qian, and Matthew Mulholland. 2021b. Automated Scoring of Spontaneous Speech from Young Learners of English Using Transformers. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 705–712.

Xinhao Wang, Klaus Zechner, and Christopher Hamill. 2020. Targeted Content Feedback in Spoken Language Learning and Assessment. In *INTERSPEECH*, pages 3850–3854.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Shasha Xie, Keelan Evanini, and Klaus Zechner. 2012. 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*, pages 103–111, Montréal, Canada. Association for Computational Linguistics.

Su-Youn Yoon and Chong Min Lee. 2019. Content modeling for automated oral proficiency scoring system. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 394–401, Florence, Italy. Association for Computational Linguistics.

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, Santa Fe, New Mexico. Association for Computational Linguistics.