Grammatical Error Detection on Spontaneous Children's Speech Using Iterative Pseudo Labeling

Christopher Gebauer[1], *Lars Rumberg*[1], *Lars Köhn*[1], *Hanna Ehlert*[2], *Edith Beaulac*[2], *Jörn Ostermann*[1]

[1]Institut für Informationsverarbeitung - L3S, Leibniz University Hannover, Germany
[2]Institut für Sonderpädagogik, Leibniz University Hannover, Germany
gebauer@tnt.uni-hannover.de

Abstract

Language acquisition is fundamental for the development of various skills in the early stage of a children's life. Unfortunately, developmental language disorder (DLD) is the most common developmental disorder during childhood. A common indicator of DLD is that children with such condition struggle to correctly use grammatical forms. Therefore, we focus in this work on automatic grammatical error detection on spontaneous children's speech. We extend the state of the art by an iterative pseudo labeling scheme to account for the ambiguity of grammatical error labels. Such ambiguity becomes obvious, when it is unclear which word is incorrect, e. g., for agreement errors. In terms of the F1 gain score (FG_1) we significantly improve upon the baseline on sentence- and word-level label. On automatic transcriptions of the kidsTALC corpus we increase the sentence-level FG_1 from 0.38 to 0.63. Further, our best performing system achieves a recall of 0.45, while maintaining a precision of 0.36.

Index Terms: Grammatical error detection, automatic speech recognition, spontaneous children's speech

1. Introduction

Language acquisition is one of the most fundamental prerequisites for the children's development in various areas of the early stages in life [1]. A delayed language acquisition, besides impacting the skill development, increases the risk for issues with mental health and social behavior [2]. Unfortunately, developmental language disorder (DLD) is the most common developmental disorder during childhood [3]. Further, for a successful intervention an early identification is necessary [2], which increases the need for an assessment tool suitable for everyday use. A common indicator for children in kindergarten that suffer from a DLD is that these children struggle with the correct use of grammatical forms [4]. Therefore, in this work we focus on automatic systems to detect grammatical errors (GED), which work on manually as well as automatically generated transcriptions of recordings from spontaneous children's speech.

One solution to children's speech assessment is to directly predict the final proficiency scores using end-to-end classifier, see also Fig. 1. Knill *et al.* [5] proposed a classifier that uses handcrafted features, like part of speech (POS)-tags and fluency metrics, to predict these proficiency scores. The authors especially investigated the impact of automatic speech recognition (ASR)-performance in terms of word error rate (WER) on the GED performance. Wang *et al.* [6] predict proficiency scores and compare different network structures as well as feature encoder. Context-aware, transformer-based models, like BERT [7], clearly performed best. Getman *et al.* [8] specialize on the subfield of phonological proficiency and predict scores

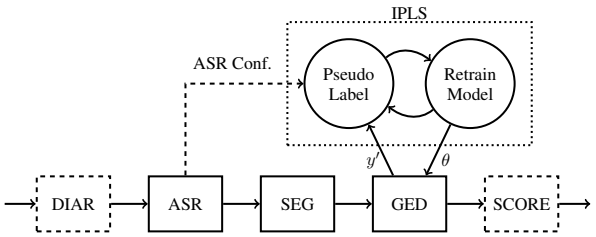

Figure 1: *Flowchart of our automatic grammatical proficiency scoring system. In this work we lay our focus on the usage of an ASR-system, the necessary segmentation (SEG) of its output, and the GED-model. The diarization (DIAR) and computation of a final proficiency score (SCORE) is left for future work. Our iterative pseudo labeling scheme (IPLS) requires a pretrained GED-model and, if automatically generated transcriptions are used, the confidence (Conf.) of the ASR-model to iteratively optimize the GED-model.*

based on Wav2Vec-based features [9]. Again, context-aware feature representations performed best. All those methods have a low degree of interpretability for the end-user, since the reasoning behind the final decision happens within the classifier. Further, the adaption to different scoring metrics or new settings requires finetuning.

Therefore, we focus on an alternative approach and predict error labels on word-level, which allows the derivation of proficiency scores in an interpretable fashion. Morley *et al.* [10] predict per word error labels by training a dependency parser that incorporates error labels in the dependency tree. Besides high quality labels for error labels this requires also labels for the dependency trees and has not been tested on automatically generated transcriptions. He *et al.* [11] avoid the necessity for dependency label by training a transformer-based classifier to directly predict errors on word-level. Furthermore, Knill *et al.* [12] and Lu *et al.* [13] leverage the knowledge of pretrained feature embeddings from Word2Vec and use it as input to a bidirectional RNN-based classifier. The latter also uses ASR-generated transcriptions and compensates its errors by assuming words always as grammatically correct if the confidence of the ASR-model falls below a predefined threshold. Bell *et al.* [14] compare different word representations as input to a GED-model and show that BERT [7] performed best.

Lu *et al.* [15] investigate the impact of a disfluency classifier on GED performance if applied as additional post-processing step. The authors train the disfluency classifier end-to-end and assume likely disfluent words as grammatically correct. However, we noticed a larger source of errors originates in the am-

biguity of grammatical errors. Let's look at the sentence: *Our cat love food*. In this number agreement error, depending on the context, either the verb *loves* or the noun *cats* should be labeled as incorrect, but the respective other word is correct. The resulting error labels, therefore, depend on the context and are possibly inconsistent if multiple annotators are involved. This ambiguity of grammatical error labels has been first identified by Chodorow *et al.* [16]. Katinskaia *et al.* [17] address this problem in a computer-aided language learning systems (CALL)-system by integrating a learning-based GED-system and, on its failure, fall back to the rule-based baseline.

In the present work we aim to incorporate the aforementioned problem of ambiguity of manual grammatical error labels into the proposed GED-model of Bell *et al.* [14]. Specifically, we train a bidirectional GRU-based network and use BERT embeddings as input. To address the ambiguity problem, we propose an IPLS, which adds highly confident false positives to the set of training labels. Inspired by Xu *et al.* [18] in the domain of ASR, we use a trained instance of our GED-system to iteratively extend our set of training labels. In contrast to Xu *et al.*, we do not extend the training data with unlabeled data, but update the labels of the existing data to smooth label noise. The extended label set is again used to retrain a new instance of the GED-model. In our evaluation we demonstrate the significant gain in classification performance achieved by our IPLS, both on manual and automatic transcriptions.

2. Method

In this section we describe our pipeline for grammatical error detection including our IPLS.

2.1. Grammatical Proficiency Scoring

Grammatical proficiency scoring on spoken speech requires five essential steps: Speaker diarization, automatic speech recognition (ASR), text segmentation, grammatical error detection (GED), and the derivation of a proficiency score based on the predicted errors. All steps are summarized in Fig. 1. The diarization is only necessary, if spontaneous speech with multiple speakers is assessed. In this work we assume this step as manually given. The downstream proficiency scoring is highly dependent on the desired purpose. Therefore, we leave the investigation for suitable scoring methods to future work. Further, this flow chart includes our IPLS, which is described in Sec. 2.3 and summarized in Alg. 1. When relying on automatically generated transcriptions, an ASR is necessary and its confidence scores are passed to our IPLS. Details to the used ASR-system are given in Sec. 3.2. The text segmentation is always necessary and described in Sec. 3.4.

2.2. Preliminaries

Given a transcript $\mathcal{X}$ consisting of words x_i with $i < T$ and T being the number of words in $\mathcal{X}$. We define a sentence $s_j \in \mathcal{S}$ as a range $s_j = (m, n)$ with $m < n < T$. Given a manual word label $y_i \in \mathcal{Y}$ the sentence label $z_j \in \mathcal{Z}$ is positive if any label y_i for $i \in s_j$ is positive, i. e., z_j is positive if a grammatical error occurs in the corresponding sentence s_j. A feature encoder f maps the sequence of words $\mathcal{X}$ to a sequence of features of length T. A classifier $M(y_i = 1|x_i, f(\mathcal{X}), \theta)$ predicts the likelihood of sample x_i being positive given the full context of $\mathcal{X}$ in feature space. θ is a set of learnable parameters for classifier $\mathcal{M}$ and is optimized using word-level label $\mathcal{Y}$ in a supervised fashion with a given loss L.

Algorithm 1 Iterative Pseudo-Labeling

Require: Transcripts $\mathcal{X}$, manual label $\mathcal{Y}$, model M, feature encoder f, threshold τ, loss $\mathcal{L}$, maximum iterations K and warm-up iterations K'.

```
Find initial parameter set θ_0 for M on f(X), Y using L
Initialize joined label set Y* = Y
for k = 1 to K do
    Predict labels Y' for X using M, f, θ_{k-1} and τ
    If k < K' and z_j = 0 with i ∈ s_j enforce y'_{i,k} = 0
    Expand joined set of labels Y* = Y* ∪ Y'
    Optimize θ_k for M on f(X), Y* using L
end for
return θ*
```

2.3. Iterative Pseudo Labeling

Our iterative pseudo labeling scheme (IPLS) has the purpose to refine the exiting labels and not to label unlabeled data. First, we optimize the initial set of parameters $\theta_{k=0}$ of the model M using the manual labels $\mathcal{Y}$ and loss L. k is the iteration index in the IPLS and K represents the maximal number of iterations. Using the model of the previous iteration we predict a set of pseudo label $y'_{i,k} \in \mathcal{Y}'_k$ for the k-th iteration as follows

$$y'_{i,k} = \begin{cases} 1, & \text{if } M(y_i = 1|x_i, f(\mathcal{X}), \theta_{k-1}) \geq \tau \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

The threshold τ is a hyperparameter. The resulting set is joined with the manual label set $\mathcal{Y}$, to a joined label set $\mathcal{Y}^*$. The joined label set $\mathcal{Y}^*$ is only extended and will not be reset during the entire training. We noticed a warm-up phase of K' iterations is beneficial for the overall training, in which the predicted label $y'_{i,k}$ in Eq. 1 is always negative if z_j for $i \in s_j$ is negative. The algorithm is summarized in Alg. 1. We do not extend the label set in the validation and test data. As usual, the final parameter set θ^* is selected based on the validation data.

2.4. Incorporation of ASR Confidence

The literature [5, 13] showed that transcription errors of the ASR-system lead to an increase in false positives. In the present work, we noticed a similar trend. Lu *et al.* [13] account for these errors by setting $M(y_i = 1|x_i, f(\mathcal{X}), \theta)$ to zero if the ASR-model confidence of word x_i falls below a threshold κ. This prevents the model being trained on erroneous words generated by the ASR. We also apply this masking during the computation of pseudo labels in our IPLS. An ablation study is carried out in Sec. 4.

3. Experimental Setting

In this section we outline the used speech corpus, namely kidsTALC, the ASR-system, the label transfer, the text segmentation, the GED-model, and the metrics used for evaluation. The diarization is done manually in our work and taken from kidsTALC.

3.1. Speech Corpus

For evaluation, we use kidsTALC [19], a corpus of monolingual, typically developing, German-speaking children in the age range from 3½ to 11 years. The corpus is extended by additional recordings of 40 children. All additional children are recorded in an identical setting as the ones in kidsTALC. The setting

always includes an adult or speech language therapist (SLT), which are also present in the transcriptions. In total, this results in 87 children distributed as followed across the age groups proposed alongside the kidsTALC corpus: 26 (AG1) - 41 (AG2) - 9 (AG3) - 11 (AG4). The original train set of kidsTALC is exclusively used for ASR training. All other recordings are split in train, validation, and test set for this work. Latter, is consistent across all experiments in Sec. 4 and contains 5 children.

For all children a manual utterance-level segmentation, transcription, and annotation of error labels are available. The annotations in kidsTALC contain three classes of error labels: Grammatical, lexical and phonological errors. Further details on the specific error type are not annotated. In this work we consider only grammatical errors. In total 1480 grammatical errors are present in the dataset, which contains in total 118,163 words. We define the positive label to refer to the presence of a grammatical error. This leads to fraction of positive labels π of $1.25\,\%$. Across all folds and splits this values varies in $\pi \in [1.15\,\%, 1.39\,\%]$.

3.2. Automatic Speech Recognition

Gebauer *et al.* [20] showed that sequence-to-sequence models outperform connectionist temporal classification (CTC)-based systems [21] on orthographic transcriptions. However, when we use a pretrained Wav2Vec 2.0 model from Hugging Face[1] as feature encoder and tune the dropout rate to 0.05 the WER generally decreased and CTC-based ASR-systems performed overall best. Therefore, in this work the feature encoder is combined with two small dense layers and trained using the CTC-loss [21]. We only finetune the contextual layer of Wav2Vec jointly with the classification head. As suggested by Rumberg *et al.* [19], we extend kidsTALC by the German Mozilla Common Voice (MCV)[2], version 16.1. The token set includes all characters of the German alphabet, is lower cased and excludes punctuations.

As decoding scheme we use the flashlight beam search algorithm [22] with a beam size of 500 and constrain it to a lexicon as well as language model. As language model we deploy a 4-gram KenLM [23]. The lexicon and language model are based on the train-set from kidsTALC as well as MCV. Overall, we achieve a WER of 0.349 on the test-set of kidsTALC averaged across two seeds and two splits.

3.3. Label Transfer

When using ASR-based transcriptions no manual labels $\mathcal{Y}$ are present and have to be transferred from the manual transcriptions. First, the manual and automatic transcriptions are aligned on word-level. Given an aligned word pair, if the manual word is grammatically incorrect and the words match, the labels are always transferred to the automatic word. In case of a non-matching word pair two options for label transfer exist [13]. Unfortunately, both method are faulty in different ways. Either these labels are discarded. This will cause the GED-system to be falsely penalized, if the corresponding word for the discarded labels is identified as grammatically incorrect. This leads to wrong negative labels, but ensures that the set of positive labels is always correct. On the other hand, if these labels are transferred regardless of the matched word, incorrect positive labels are introduced. Further, the set of negative labels is still not ensured to be correct, because a grammatically correct word in the manual transcriptions could be aligned with a grammatically incorrect word from the automatic transcript. Therefore, we choose to only transfer the manual labels if the aligned word pair matches. This leads to a loss of about half of the manual labels of grammatically incorrect words.

For the sentence-level labels $\mathcal{Z}$ this effect is much less relevant, since misalignment usually does not cross sentence borders. Therefore, we keep all manual labels that have been aligned into a given automatic sentence regardless of whether the aligned words match. Further, we track manual words with grammatical error labels that are not assigned to any automatic sentence and pass them to the next sentence. This ensures that no manual labels $\mathcal{Y}$ are lost during their transfer to the automatic, sentence-level labels $\mathcal{Z}$.

3.4. Text Segmentation

Due to capacity limitations of model M and feature encoder f the full transcriptions have to be segmented into context windows. To maximize context, we keep the utterances from the adult speaker within the transcriptions and mask it during training as well as evaluation of the GED-model. We select context windows based on multiple consecutive sentences. Sentences are defined by punctuation, which is reconstructed using a pretrained, state-of-the-art punctuation model from Hugging Face[3]. It is specifically trained on the reconstruction of punctuation and is not allowed to further change the inferred text.

Due to varying length of sentences, we cannot define a fixed context window size and shift width. We constrain the windows to be close to a size of 50 words and shift the windows close to a length of 10 words. This leads to an average window size of 49.7 ± 6.8 and an average hop length of 13.7 ± 9.3. Due to the combination of window size and hop length each word can appear multiple times within the resulting dataset. On average each word appears 8.6 times within each epoch.

3.5. GED Model

In this section we describe the classifier M based on Bell *et al.* [14]. However, the feature encoder f is relevant for the outcome of the classifier and will be discussed as well. We use an uncased, German BERT model from Hugging Face[4]. For words that are split into multiple token, we aggregate by taking the mean feature vector, as this method is supposed to work best [24]. The classifier M is a small RDNN model consisting of roughly 1M trainable parameters. First, the features are passed to a two-layered, bidirectional GRU network. According to Chung *et al.* [25] these outperform LSTMs. Next, we have two fully-connected layers followed by the sigmoid function. For training efficiency, we keep the weights of the BERT model fix. The set of trainable parameters θ is optimized using the AdamW [26] optimizer with a learning rate of 0.001 and the binary cross entropy (BCE)-loss. We selected a dropout rate of 0.5 and a weight decay of 0.25. Both values are comparably high, because the model tends to overfit. In the automatic setting, we also use the focal loss [27] for stability, leading to a slight improvement. We set α to 0.5 and γ to 0.75.

3.6. Performance Metrics

In our work we mainly use the F1 gain (FG_1) score [28], but also the non-gain recall and precision. The FG_1 score normalizes the

[1] facebook/wav2vec2-large-xlsr-53-german

[2] `https://commonvoice.mozilla.org/en/datasets`

[3] oliverguhr/fullstop-punctuation-multilang-large

[4] dbmdz/bert-base-german-uncased

general F_1 score with the proportion of positive labels π:

$$FG_1 = \frac{F_1 - \pi}{(1-\pi)F_1}.$$

The family of gain-scores has two advantages. First, they are linearized, which allows us to average across different tasks with varying fraction of positive labels π. This is necessary, as π varies across folds and depends on the underlying automatic transcriptions due to the varying number of lost labels (see Sec. 3.3). Secondly, the gain-scores have two consistent thresholds: 0.0 for outperforming a random classifier and 0.5 for outperforming an all positive classifier.

4. Results

In this section we evaluate the performance of our GED-system. We consider the first iteration as the baseline, since the design of our GED-model is state-of-the-art [14]. We start by comparing the systems on the manual transcription. As summarized in Tab. 1, using our IPLS significantly ($p < 0.05$) improves the FG_1 score. This holds for word- and sentence-level classification. In this work we measure significance using a dependent t-test for paired samples. The lower FG_1 on sentence- compared to word-level originate in the higher fraction of positive labels π, not in an overall worse performance. Absolutely speaking the system performs better on sentence-level, but according to the gain score normalization [28] the task itself is easier. All values are averaged across three splits and three seeds.

Table 1: *FG_1 on the manual transcriptions of kidsTALC. All values are averaged across three splits and three seeds. Both, on word- and sentence-level, the IPLS (indicated by +It.) outperforms the baseline significantly ($p < 0.05$).*

Word		Sentence	
-	+It.	-	+It.
0.97	**0.98**	0.89	**0.91**

Next, we take a look at the GED-system being evaluated on the automatic transcriptions. In Tab. 2 we visualize the results, when the ASR-model confidence is not taken into account. Our IPLS is able to significantly ($p < 0.05$) outperform the baseline in terms of FG_1 score. Applying the focal loss instead of BCE further improves the FG_1 score to 0.87 on word-level and 0.63 on sentence-level. Especially on sentence-level our IPLS is necessary to pass the relevant border of 0.5 with the FG_1.

Table 2: *FG_1 on automatically generated transcriptions of kidsTALC. We do not take the ASR-model confidence into account. All values are averaged across three splits and three seeds and transcriptions from four different ASR-systems. Both, on word- and sentence-level, the IPLS (indicated by +It.) outperforms the baseline significantly ($p < 0.05$). Further, the focal loss [27] leads to further improvement.*

Loss	Word		Sentence	
	-	+It.	-	+It.
BCE	0.83	0.85	0.38	0.54
Focal [27]	0.79	**0.87**	0.36	**0.63**

In Tab. 3 we visualize the results, when words with low ASR-model confidence are assumed to be grammatically correct and neglected during loss computation. Due to our IPLS the tuning of the confidence threshold κ is more expensive than for Lu *et al.* [13], because we influence the pseudo labels with this threshold. We tested two different values and could not achieve an improvement in terms of FG_1 score compared to not taking the ASR confidence into account. Not averaging across seeds, splits, and ASR-systems, but taking the best performing model in terms of FG_1 allows us to compare the precision or recall. Even on these metrics we could not find any improvement. This can have three distinct causes. Either the method is very sensible to the threshold κ and needs a more thorough hyperparameter search. Lu *et al.* [13] rely on hybrid hidden Markov model for ASR, which natively provide a more robust confidence score. While we use a state-of-the-art method for CTC-based ASR confidence [29], it is an estimation and not the true confidence. This could cause the decrease in performance. Lastly, the baseline GED-model tends to have a low recall in comparison to precision on kidsTALC, i. e., improving the number of true positives has a greater impact on the performance than reducing false positives.

Table 3: *FG_1 on automatically generated transcriptions of kidsTALC. Words with low ASR-model confidence are assumed to be grammatically correct. All values are averaged across three splits and three seeds and transcriptions from four different ASR-systems. Both, on word- and sentence-level, the IPLS (indicated by +It.) outperforms the baseline significantly ($p < 0.05$). The focal loss [27] leads to further improvement.*

Loss	Word		Sentence	
	-	+It.	-	+It.
BCE	0.72	0.77	0.12	0.3
Focal [27]	0.73	**0.78**	0.11	**0.52**

While the gain scores allow averaging across different tasks, the intuition for the absolute performance is lost. Therefore, we present precision and recall on sentence-level for our best performing systems according to the FG_1 score. On manual transcriptions we achieve a recall of 0.4, while maintaining a precision of 0.66. On ASR-based transcriptions we slightly increase the recall to 0.45, but the precision drops to 0.36. While the absolute values seem low, this is impressive due to the low π of about 1.38 % on our test set.

5. Conclusion

In this work, we focus on automatic GED for children's speech assessment in spontaneous speech recordings. We extend Bell *et al.* [14] by an iterative pseudo labeling scheme to account for the ambiguity problem of grammatical error labels. The ambiguity of grammatical error labels is obvious for, e. g., agreement errors, when it is unclear which of the two related words is incorrect. Our proposed iterative pseudo labeling scheme accounts for this problem by iteratively extending the training labels. We significantly improve upon the baseline on both, sentence- and word-level label, of kidsTALC. On automatic transcriptions we achieve an FG_1 score of 0.87 on word-level and 0.63 on sentence-level. Further, our best performing system according to the FG_1 score achieves a recall of 0.45, while maintaining a precision of 0.36. On manual transcriptions the precision increases to 0.66.

6. References

[1] K. Berendes, "Sprache als Schlüsselkompetenz für Bildungsprozesse," *Forum Logopädie*, vol. 38, no. 2, pp. 8–14, 2024.

[2] C. Kiese-Himmel, "Früherkennung primärer Sprachentwicklungsstörungen – zunehmende Relevanz durch Änderung der Diagnosekriterien?" *Bundesgesundheitsblatt - Gesundheitsforschung - Gesundheitsschutz*, vol. 65, no. 9, pp. 909–916, 2022.

[3] C. F. Norbury, D. Gooch, C. Wray, G. Baird, T. Charman, E. Simonoff, G. Vamvakas, and A. Pickles, "The impact of nonverbal ability on prevalence and clinical presentation of language disorder: Evidence from a population study," *Journal of Child Psychology and Psychiatry*, vol. 57, no. 11, pp. 1247–1257, 2016.

[4] L. H. Finestack, E. Ancel, H. Lee, K. Kuchler, and M. Kornelis, "Five Additional Evidence-Based Principles to Facilitate Grammar Development for Children With Developmental Language Disorder," *American Journal of Speech-Language Pathology*, vol. 33, no. 2, pp. 552–563, 2024.

[5] K. Knill, M. Gales, K. Kyriakopoulos, A. Malinin, A. Ragni, Y. Wang, and A. Caines, "Impact of ASR Performance on Free Speaking Language Assessment," in *Proceedings INTERSPEECH 2018 – 19th Annual Conference of the International Speech Communication Association*. ISCA, 2018, pp. 1641–1645.

[6] X. Wang, K. Evanini, Y. Qian, and M. Mulholland, "Automated Scoring of Spontaneous Speech from Young Learners of English Using Transformers," in *2021 IEEE Spoken Language Technology Workshop (SLT)*, 2021, pp. 705–712.

[7] 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)*, J. Burstein, C. Doran, and T. Solorio, Eds. Association for Computational Linguistics, 2019, pp. 4171–4186.

[8] Y. Getman, R. Al-Ghezi, K. Voskoboinik, T. Grósz, M. Kurimo, G. Salvi, T. Svendsen, and S. Strömbergsson, "Wav2vec2-based Speech Rating System for Children with Speech Sound Disorder," in *Proceedings INTERSPEECH 2022 – 23st Annual Conference of the International Speech Communication Association*. ISCA, 2022, pp. 3618–3622.

[9] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33. Curran Associates, Inc., 2020, pp. 12 449–12 460.

[10] E. Morley, A. E. Hallin, and B. Roark, "Data Driven Grammatical Error Detection in Transcripts of Children's Speech," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2014, pp. 980–989.

[11] Z. He, "English Grammar Error Detection Using Recurrent Neural Networks," *Scientific Programming*, vol. 2021, no. Scientific Programming for Smart Internet of Things, p. 7058723, 2021.

[12] K. Knill, M. Gales, P. Manakul, and A. Caines, "Automatic Grammatical Error Detection of Non-native Spoken Learner English," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 8127–8131.

[13] Y. Lu, M. J. Gales, K. M. Knill, P. Manakul, L. Wang, and Y. Wang, "Impact of ASR Performance on Spoken Grammatical Error Detection," in *Proceedings INTERSPEECH 2019 – 20th Annual Conference of the International Speech Communication Association*. ISCA, 2019, pp. 1876–1880.

[14] S. Bell, H. Yannakoudakis, and M. Rei, "Context is Key: Grammatical Error Detection with Contextual Word Representations," in *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics, 2019, pp. 103–115.

[15] Y. Lu, M. J. F. Gales, K. M. Knill, P. Manakul, and Y. Wang, "Disfluency Detection for Spoken Learner English," in *Proc. 8th ISCA Workshop on Speech and Language Technology in Education (SLaTE 2019)*, 2019, pp. 74–78.

[16] M. Chodorow, M. Dickinson, R. Israel, and J. Tetreault, "Problems in Evaluating Grammatical Error Detection Systems," in *Proceedings of International Conference on Computational Linguistics (COLING)*, M. Kay and C. Boitet, Eds. The COLING 2012 Organizing Committee, 2012, pp. 611–628.

[17] A. Katinskaia and R. Yangarber, "Assessing Grammatical Correctness in Language Learning," in *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, 2021, pp. 135–146.

[18] Q. Xu, T. Likhomanenko, J. Kahn, A. Hannun, G. Synnaeve, and R. Collobert, "Iterative Pseudo-Labeling for Speech Recognition," in *Proceedings INTERSPEECH 2020 – 21st Annual Conference of the International Speech Communication Association*. ISCA, 2020, pp. 1006–1010.

[19] L. Rumberg, C. Gebauer, H. Ehlert, M. Wallbaum, L. Bornholt, J. Ostermann, and U. Lüdtke, "kidsTALC: A Corpus of 3- to 11-year-old German Children's Connected Natural Speech," in *Proceedings INTERSPEECH 2022 – 23rd Annual Conference of the International Speech Communication Association*. ISCA, 2022, pp. 5160–5164.

[20] C. Gebauer, L. Rumberg, H. Ehlert, U. Lüdtke, and J. Ostermann, "Exploiting Diversity of Automatic Transcripts from Distinct Speech Recognition Techniques for Children's Speech," in *Proceedings INTERSPEECH 2023 – 24rd Annual Conference of the International Speech Communication Association*. ISCA, 2023, pp. 4578–4582.

[21] A. Graves, S. Fernandez, F. Gomez, and J. Schmidhuber, "Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks," in *International Conference on Machine Learning (ICML)*, 2006, p. 8.

[22] J. Kahn, V. Pratap, T. Likhomanenko, Q. Xu, A. Hannun, J. Cai, P. Tomasello, A. Lee, E. Grave, G. Avidov, B. Steiner, V. Liptchinsky, G. Synnaeve, and R. Collobert, "Flashlight: Enabling Innovation in Tools for Machine Learning," in *Proceedings of the 39th International Conference on Machine Learning*, vol. 162. PMLR, 2022, pp. 10 557–10 574.

[23] K. Heafield, "KenLM: Faster and Smaller Language Model Queries," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, 2011, pp. 690–696.

[24] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, vol. 3982–3992. Association for Computational Linguistics, 2019.

[25] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *NIPS 2014 Workshop on Deep Learning*, 2014.

[26] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," in *International Conference on Learning Representations*, 2017.

[27] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal Loss for Dense Object Detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 318–327, 2018.

[28] P. Flach and M. Kull, "Precision-Recall-Gain Curves: PR Analysis Done Right," in *Advances in Neural Information Processing Systems*, vol. 28. Curran Associates, Inc., 2015.

[29] L. Rumberg, C. Gebauer, H. Ehlert, M. Wallbaum, U. Lüdtke, and J. Ostermann, "Uncertainty Estimation for Connectionist Temporal Classification Based Automatic Speech Recognition," in *Proceedings INTERSPEECH 2023 – 24rd Annual Conference of the International Speech Communication Association*. ISCA, 2023, pp. 4583–4587.