Automated Scoring of Clinical Expressive Language Evaluation Tasks

Yiyi Wang[†], Emily Prud'hommeaux[†], Meysam Asgari[‡], and Jill Dolata[‡]
[†] Boston College, Chestnut Hill MA, USA
[‡] Oregon Health & Science University, Portland OR, USA
{wangdil,prudhome}@bc.edu, {asgari,dolataj}@ohsu.edu

Abstract

Many clinical assessment instruments used to diagnose language impairments in children include a task in which the subject must formulate a sentence to describe an image using a specific target word. Because producing sentences in this way requires the speaker to integrate syntactic and semantic knowledge in a complex manner, responses are typically evaluated on several different dimensions of appropriateness yielding a single composite score for each response. In this paper, we present a dataset consisting of non-clinically elicited responses for three related sentence formulation tasks, and we propose an approach for automatically evaluating their appropriateness. Using neural machine translation, we generate correct-incorrect sentence pairs to serve as synthetic data in order to increase the amount and diversity of training data for our scoring model. Our scoring model uses transfer learning to facilitate automatic sentence appropriateness evaluation. We further compare custom word embeddings with pre-trained contextualized embeddings serving as features for our scoring model. We find that transfer learning improves scoring accuracy, particularly when using pre-trained contextualized embeddings.

1 Introduction

It is estimated that between 5% and 10% of the pediatric population will experience a speech or language impairment (Norbury et al., 2016; Rosenbaum and Simon, 2016). Diagnosing these impairments is complex, requiring the integration of structured assessments, medical history, and clinical observation, and there is evidence that language impairments are frequently misdiagnosed and underdiagnosed (Conti-Ramsden et al., 2006; Grimm and Schulz, 2014; Rosenbaum and Simon, 2016). As a result, there is a need for tools and technologies that can support efficient and remote screening for language impairment. However, developing methods for automatically scoring the subtests used to diagnose language disorders can be challenging because of the very limited amount of labeled data available for these subtests from these special populations.

In this paper, we focus on a task we have adapted from the Formulated Sentences (FS) subtest of the Clinical Evaluation of Language Fundamentals 4 (CELF-4), one of the most widely used language diagnostic instruments in the United States (Semel et al., 2003). In the CELF-4 Formulated Sentences task, a child is presented with a target word and an image, and must use that word in a sentence about that image. Poor performance on this subtest is strongly correlated with expressive language impairments. Responses are scored on a scale from 0 to 2; a sentence assigned a score of 2 must correctly use the target word, be a complete and grammatically correct sentence, and relate to the content and activities shown in the image. Reliable manual scoring can be difficult and time-consuming because of the large of number of factors that must be considered. This degree of subjectivity, together with the task's important role in identifying expressive language impairments, make automatic scoring of the formulated sentences subtest particularly worthwhile.

This paper makes the following contributions:

- We present a new data set of non-clinically elicited formulated sentence task responses, annotated for appropriateness evaluation (scores: 0, 1, and 2), which can be used as a benchmark and as a data source for future automated scoring of clinically elicited responses. The dataset includes 2160 sentences from three related sentence formulation tasks (Section 3).

Proceedings of the 15th Workshop on Innovative Use of NLP for Building Educational Applications, pages 177–185
July 10, 2020. ©2020 Association for Computational Linguistics

- We develop a neural machine translation model trained on second language learner data and generate two artificial datasets for training the formulated sentences scoring classifier.

- We demonstrate that our transfer learning model has benefits for scoring formulated sentences.

- We systematically compare the use of custom task-specific embeddings and pre-trained generic contextualized embeddings for scoring formulated sentences.

2 Related Work

Scoring formulated sentences in terms of syntactic correctness can be analogous to the more common task of Grammatical Error Detection (GED), in which points are deducted for each grammatical error detected in a sentence or text. The state-of-the-art approaches to GED use a supervised neural sequence labeling setup to detect errors trained on artificial data (Rei 2017; Kasewa et al. 2018). Performance on this task can generally benefit from using a large size of high-quality training data, but collecting large quantities of such data is expensive.

Data augmentation can increase the amount and diversity of training data, provide additional information about the representations of sentences, and improve performance on the GED task. The current state-of-the-art GED trains on artificially generated data produced via error induction. One traditional way is to use the patterns learned from annotated learner corpora and apply them to grammatically correct text to generate specific type of errors, such as noun errors (Brockett et al., 2006) and article errors (Rozovskaya and Roth, 2010). More recently, artificial training data is typically generated by using machine translation, where the source text is error-free text and the target is ungrammatical text (Rei 2017).

Deploying vectorized representation of word and sentence is now a ubiquitous technique in most NLP tasks. Incorporating word embeddings as features can provide another possible solution in low-resource scenarios. The current state of the art GED is achieved by using BERT embeddings to capture the contextual information. Bell et al. (2019) compare using ELMo, BERT and Flair embeddings on several publicly available GED datasets, and propose an approach to effectively integrate such representations to detect grammatical errors.

Our work is inspired by this prior research on using machine translation to generate artificial data and comparing the influence of task-specific versus generic embeddings. Although these methods are typically trained on second language learners' data in essay writing tasks, our goal is to seek a general representation of the syntactic and semantic representation of a single sentence in a constrained domain by children who are L1 speakers but may have deficits in expressive language. Given the very limited amount of clinical data, however, we make the assumption that the types of errors language learners make can be leveraged to evaluate formulated sentences responses, an assumption that will be empirically validated in future work with our clinical dataset.

3 Data

In this section, we describe the non-clinical formulated sentences dataset we have collected, as well as two other publicly accessible datasets, MS-COCO (Lin et al., 2014) and FCE (Yannakoudakis et al., 2011), which we use to train embeddings and generate artificial training data.

Formulated Sentences (FS) Dataset Using our own stimuli designed to mimic the properties of the CELF-4 Formulated Sentences (FS) stimuli, we collected 2160 sentences from Amazon Mechanical Turk workers and scored the responses according to the published guidelines for the CELF-4 FS task, which rely on syntactic grammaticality and semantic appropriateness given the image. Each of the 24 numbered stimulus words was manually selected by a speech language pathologist in order to be comparable to the corresponding CELF-4 FS stimulus word in terms of part of speech, age of acquisition, and phonological complexity. The participants were recruited on Amazon Mechanical Turk (AMT) and directed to take the test within the online survey platform Qualtrics (Barnhoorn et al., 2015), as required by the affiliated university's Institutional Review Board.

Our FS data collection effort is composed of three sub-tasks:

- Task 1: Formulating sentences from an image and a target word. Participants view an image and a target word and write a sentence using that word to describe that image.

- Task 2: Formulating sentence from target word only (no image). Participants are asked to write a sentence that includes the target word.

- Task 3: Formulating sentences from an image only (no target word). Participants are asked to write a sentence description of the image in their own words.

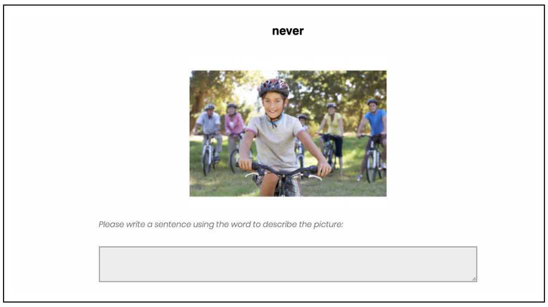
Each participant was randomly assigned to take one of the three sub-tasks. The participant was instructed to view a sample stimulus and response and was then asked to take two trial stimuli to ensure they were familiarized with the test format and environment. Since there are intra-relationship between the three tasks, a participant was able to complete only one sub-task to avoid covariate effects. There were 30 participants for each task, with 24 stimuli in each test, resulting in 2160 sentences (24 stimuli * 3 tasks * 30 participants) included in the dataset.

The collected sentences were scored by four native speakers of English by giving a score of 0, 1, or 2. A score of 2 indicates that the sentence fully expresses the content of the image by using the given word without any grammatical errors. If there is one grammatical error, or the sentence only represents unimportant details of the image, the sentences is marked as 1. If there are two or more grammatical errors, or if the content is unrelated to the image, it is assigned a score of 0.

	2	1	0	Total
Task 1	511	52	157	720
Task 2	658	29	33	720
Task 3	370	123	27	720
All Tasks	1739	204	217	2160

Table 1: Score distribution in 3 tasks.

Each grader was assigned to evaluate 2 sub-tasks, and the average pairwise kappa between the graders was 0.625. When there was a disagreement between two graders, the third grader was recruited, and the final score was the majority of the three graders.



never

Please write a sentence using the word to describe the picture:

Figure 1: Example of the formulated sentence task (stimulus 3 of Task 1). For Task 2, only the target word "never" is displayed, and for Task 3 only the image is displayed.

Task	Score-0 sentences
1	*Boys are never driving the bicycle.* *The boy has never let down his family.* *Run to cycle.* *The boy never driving the cycle.* *Three generation family on cycle ride in countryside.* *Never is a all boys cycle driving.* *Never give up the place.* *The boy doesn't ride his bike with the others, but he doesn't care.*
2	*Never get compromised for a second option.*
3	*This boy active for bike.* *The are cycling competition.* *The boy biking the cycle.*

Table 2: Score-0 sentences for stimulus 3, target word "never", image shown in Figure 2.

Integrating a target word into an image description requires more complicated linguistic competence than using the target word to make a sentence or having a free choice of vocabulary in describing an image. Therefore, Task 1 is considered to be more challenging than Task 2 or 3. As shown in Table 1 and Figure 2, there are significantly more sentences in Task 1 that are scored 0 or 1 than in Tasks 2 and 3. For example, for Stimulus 3 with target word "never" (shown in Figure 2), 8 sentences are assigned a score of 0 in Task 1, while only one sentence is given 0 in Task 2 and 3 sentences are marked as 0 in Task 3 (Table 2).

The final sentence-formulation dataset includes 5 columns: subject ID, task, stimulus, sentence and score. The participant's personal information is replaced by a randomly assigned 5-digit subject ID. The score distribution of the three tasks is summarized in Table 1 and the data is released for public access [1].

COCO COCO is a publicly released large-scale dataset for object detection, segmentation, and captioning (Lin et al., 2014). For each image, five hu-

[1] https://github.com/yiyiwang515/Automated-Scoring-of-Clinical-Expressive-Language-Evaluation-Tasks.git

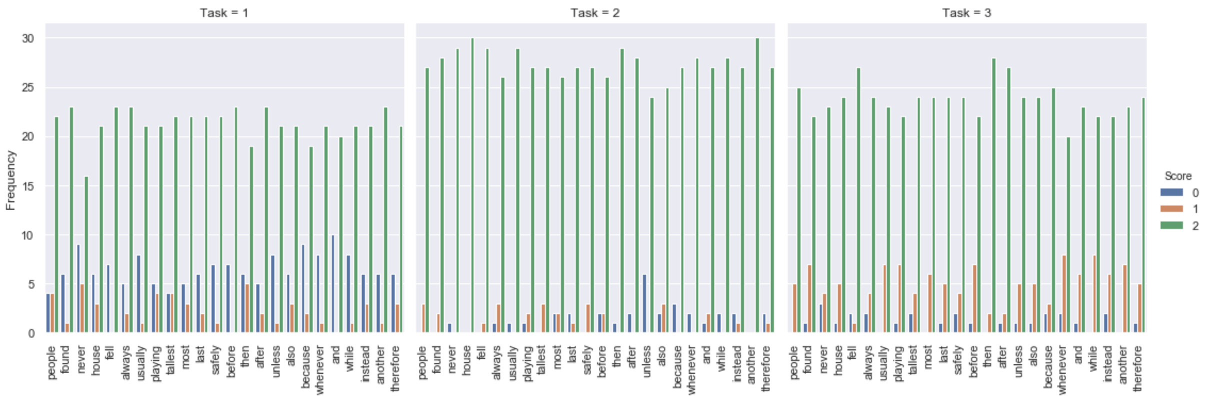


Figure 2: Score distribution for each stimulus in three tasks. Each stimulus is represented by its associated target word.

man generated captions were collected from AMT. In our work, we use the 2017 train dataset, which includes approximately 20k images with 600k captions. Since some of the captions were either empty or had likely incomplete sentences (fewer than 4 words), we exclude those captions resulting in a final dataset of 33,366 sentences.

The linguistic characteristics of COCO are analogous to our sentence formulation task regarding choice of lexicon, the use of syntactic structures, and the semantic context of the utterances. We therefore use COCO dataset for two purposes: (1) to train a task-specific Word2vec embedding to capture meaning-related and syntactic relationships; and (2) to use as score-2 source input to machine translation models that are trained to generate artificial errored (score-1 and score-0) sentences.

FCE First Certificate in English (FCE) (Yannakoudakis et al., 2011) is a publicly available dataset, including 1244 essays written by non-native learners of English with different first language backgrounds. The FCE exam is used to assess English proficiency of upper-intermediate level learners. The essays are annotated by language assessment experts with types of errors and their corresponding corrections in XML. The original incorrect sentences in the essay and their corrected counterparts are extracted by sentence pairs. The sentences containing no errors or with a length (including punctuation) less than 5 are excluded from the final dataset used for training our models.

A label is added for each sentence. For all the correct sentences, label of 2 is added to mark the appropriateness of the sentence. A sentence with one error is assigned a label of 1, while a sentence with two or more errors is assigned a label of 0. The final dataset includes 10799 correct sentences, 4810 sentences with one grammatical error, and 5989 sentences with two or more errors. The FCE data serves as training data for two neural machine translation models that we use to generate inappropriate (score-1 and score-0) sentences by taking appropriate (score-2) sentences as input.

4 Experiments

4.1 Sentence Embedding

Three types of sentence embeddings are used in this work: BERT, ELMo and Word2Vec embeddings (Devlin et al. 2018; Peters et al. 2017; Mikolov et al. 2013). For context sensitive embeddings BERT and ELMo, we use the publicly available pre-trained models. We trained a Word2cvec embedding on the 600,000 COCO captions.

BERT BERT can integrate information in raw corpora (BooksCorpus and English Wikipedia) while considering task-specific information contained in the target dataset. Kaneko and Komachi (2019) use BERT contextualized representation to achieve state-of-the-art results for word-based GED tasks. In addition to improving results in the GED task, BERT (Devlin et al., 2018) has been shown to be a powerful feature extractor for various other tasks. We employ BERT to generate pre-trained contextual representations. BERT pretrained embeddings have two versions. We use a lighter version of BERT which yields 768 dimensions for sentence embeddings. The DistilBERT is smaller but can roughly match the performance of using the full BERT (Sanh et al. 2019).

ELMo The ELMo pre-trained model we use is trained on the One Billion Word Benchmark corpus. The sentence representation is learned by a sequence labeler during training.

The BERT and ELMo models used here are trained on formal published writings, such as books and news articles. This is not a perfect domain or stylistic match for the evaluation of responses from the formulated sentences task. In order to better represent the linguistic nature of our task, we also train a task-specific Word2vec embedding.

Word2vec We train our Word2vec model using the Gensim Python library (Rehurek and Sojka, 2010). We use skip-grams to train a word embedding model with 300 dimensions, again using the COCO captions. Words occurring fewer than 5 times are filtered out, and the maximum distance between a target word and its surrounding content is set as 4. The number of threads used is 5. A sentence embedding is calculated as the mean of the component word embeddings. Since COCO captions share similar linguistic features with the sentence formulation tasks, this custom sentence embedding is expected to capture a richer linguistic representation of the task.

4.2 Data Augmentation

In our FS dataset, score-1 and score-0 sentences account for around 20% of the total number of sentences. Since most of the classification algorithms are sensitive to imbalance in predicting classes, such a dataset can bias the classification model towards score-2. In this case, a baseline majority-class classifier, which predicts score-2 for all the sentences, can achieve 0.8 accuracy (Table 1). The unbalanced nature of this data requires us to synthesize additional score-1 and score-0 sentences to increase the size and variety of training set.

We implement two LSTM machine translation models using OpenNMT (Klein et al., 2017). The score-2 (corrected) sentences from the FCE dataset are used as the source data, and the score-0 or score-1 (incorrect) sentences serve as the target data. For example, for the source-2 to target-1 (2-1) model, we trained the model on sentences pairs from FCE dataset containing only one grammatical error. The LSTM model is a two-layer bidirectional LSTM with 500 hidden units with a global attention layer. We set an early stop if the training accuracy score dropped consistently for 10 epochs. Similarly, we train another source-2 to target-0 (2-0) model with the same settings to generate score-0 sentences.

Having trained a NMT model, we then "translate" 939 score-2 sentences from our formulated sentences dataset and convert the sentences into the same number of score-0/1 sentences by using the 2-0 and 2-1 machine translation models respectively. Eight hundred score-2 sentences were excluded for synthesizing data to reserve for testing. The final synthesized formulated sentence (SFS) dataset used for training includes 2817 sentences. The COCO captions used for training word embedding are further selected in this process serving as the error-free input. A large number of COCO captions are incomplete sentences with heavy noun phrases containing a long modifier. We remove such captions in final training set to preserve sentence-level grammaticality. For example, "Man in apron standing on front of oven with pans and bakeware" is excluded, because the root of the dependency parsing tree ("man") is not a verb. A subset of COCO captions containing 14017 sentences is selected using parse information extracted using the SpaCy library (Honnibal and Montani, 2017). The final synthesized COCO (SCOCO) dataset used for training includes 42051 sentences.

4.3 Transfer Learning

Transfer learning is a viable method for building NLP models in low-resource scenarios by leveraging data from other related sources. The two artificial data datasets, SFS and SCOCO, produced by applying machine translation to the FS dataset and the COCO captions, can provide a good basis for transfer learning (Section 4.2). We implement a Multilayer Perceptron model (MLP) by using Keras with tensorflow as the backend[2]. The MLP model has two hidden layers with five nodes in each layer and uses Relu as the activation function. The output layer has three nodes, corresponding to each score class with the softmax activation function. The categorical cross-entropy loss function is minimized, and stochastic gradient descent is used to learn the problem.

The two models are fit for 200 epochs on SFS and SCOCO datasets separately during training. We transfer the weights from the two standalone models to learn the initialize the weights for the formulated sentence evaluation tasks. The original FS data is split proportionally into a training and a test sets in terms of task, stimulus, and score distribution. The training set used for tuning the model includes 1160 sentences. The test set has

[2] https://github.com/keras-team/keras

1000 sentences, and the distribution of scores is presented in Table 3.

	Task 1	**Task 2**	**Task 3**	**Total**
2	235	303	262	800
1	25	14	61	100
0	70	15	15	100
Total	330	332	338	1000

Table 3: Score distribution in our formulated sentences test set.

		P	R	F1	Acc
BL1 Majority Class		0.80	0.80	0.80	0.80
BL2 Target Word		0.82	0.82	0.82	0.82
W2V	SFS	0.69	0.59	0.63	0.59
	SFS-FS	0.74	0.80	0.76	0.80
	SCOCO	0.67	0.70	0.68	0.70
	SCOCO-FS	0.72	0.73	0.72	0.73
ELMo	SFS	0.78	0.63	0.68	0.63
	SFS-FS	0.79	0.75	0.77	0.75
	SCOCO	0.77	0.72	0.74	0.72
	SCOCO-FS	0.80	0.79	0.79	0.79
BERT	SFS	0.80	0.78	0.77	0.78
	SFS-FS	0.76	0.85	0.81	0.85
	SCOCO	0.77	0.79	0.78	0.79
	SCOCO-FS	**0.82**	**0.85**	**0.83**	**0.85**

Table 4: Sentence evaluation precision, recall, F1 and accuracy on the full three-task test set.

5 Results

Table 4 presents the results of integrating contextual embeddings with and without transfer learning, evaluating all the three sentence formulation tasks together. The method of using transfer learning incorporating BERT contextual embeddings achieves the best performance in most of the cases, except for transferring from SFS with BERT (Figure 3(a)).

The experiments demonstrate a marginal increase in precision and substantial improvement in recall for the sentence formulation tasks. Although the three tasks all involve making sentences, each task tests different aspects of linguistic knowledge. Therefore it is more valuable to investigate the model performance on individual tasks, presented in Table 5.

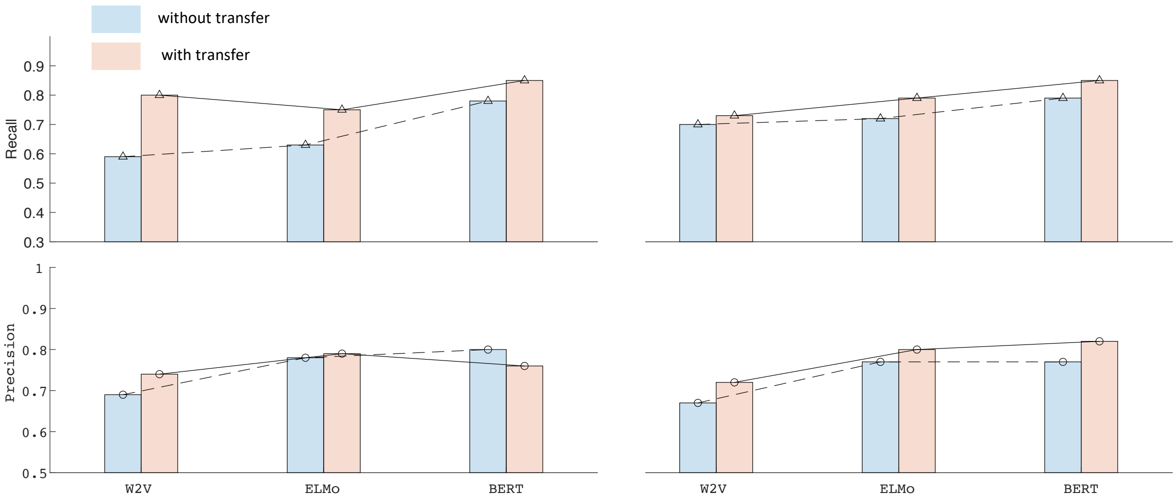
On the test sets of Task 1 and Task 2, the best performing model is transferred SFS (Task 1: F1 = 0.81; Task 2: F1 = 0.92) with BERT. For Task 3, the best model is trained on the larger SCOCO dataset (F1 = 0.82) with BERT embeddings. Task 2 and Task 3 have marginal improvement compared with the baseline. However, Task 1, which is the actual parallel to the CELF-4 Formulated Sentence task used to diagnose language impairments in children, shows a substantial improvement in performance. The improvement by incorporating BERT embedding achieves the best performance in all the three tasks individually.

These experiments demonstrate that transfer learning provides a beneficial addition for evaluating formulated sentence tasks. The language representations trained on a large general dataset allow the model to acquire a better representation of linguistic knowledge. Overall, we find that the model with transfer learning and BERT embeddings achieves the largest improvement across all three tasks.

6 Discussion

Transfer learning is generally used in situations in which the related task has more training data than the problem of interest, and the two tasks share similarities in structure. In this paper, we train an MLP model on two artificial datasets (SFS and SCOCO) and improve the performance of formulated sentence scoring task. The two datasets are similar in terms of lexicon variations, syntactic structures, and semantic expressions. They are generated by using the same machine translation models. The difference between these two sets is in the number of sentences contained. Since the number of correct sentences used for generated SCOCO is 14 times larger than that used to generated the SFS set, the final data of SCOCO is much larger than SFS.

Although the SCOCO dataset is larger, the best performance on both Task 1 and Task 2 is achieved by using transfer learning from in-domain SFS data. One of the similarities between the two tasks is the requirement of applying stimuli words in the generated sentences, whereas the test participants have a free choice of words in Task 3. This requirement influences the grading of the sentences. If the target word is not included in a sentence, no matter how grammatical the sentence is, it is marked with a label 0 in Task 1 and Task 2. For all the score-2 sentences in SFS, the target stimuli words are included; however this is not always the case in SCOCO. The lack of target word representation information in SCOCO may cause the transfer learning results to be inferior to the SFS model.



(a) SFS

(b) SCOCO

Figure 3: Precision and recall on full dataset using (a) SFS and (b) SCOCO with and without transfer learning.

		Task 1				Task 2				Task 3			
		P	R	F1	Acc	P	R	F1	Acc	P	R	F1	Acc
BL1 Majority Class		0.71	0.71	0.71	0.71	0.91	0.91	0.91	0.91	0.78	0.78	0.78	0.78
BL2 Target Word		0.77	0.77	0.77	0.77	0.92	0.92	0.92	0.92	0.78	0.78	0.78	0.78
W2V	SFS	0.66	0.63	0.64	0.63	0.84	0.43	0.56	0.43	0.69	0.71	0.70	0.71
	SFS-FS	0.71	0.75	0.71	0.75	0.84	0.86	0.85	0.86	0.75	0.78	0.73	0.78
	SCOCO	0.56	0.65	0.59	0.65	0.84	0.69	0.76	0.69	0.69	0.75	0.70	0.75
	SCOCO-FS	0.70	0.68	0.67	0.68	0.85	0.8	0.82	0.80	0.68	0.69	0.69	0.69
ELMo	SFS	0.75	0.68	0.71	0.68	0.89	0.54	0.65	0.54	0.78	0.67	0.71	0.67
	SFS-FS	0.77	0.76	0.76	0.76	0.88	0.71	0.78	0.71	0.81	0.79	0.79	0.79
	SCOCO	0.74	0.72	0.73	0.72	0.86	0.69	0.76	0.69	0.79	0.75	0.76	0.75
	SCOCO-FS	0.80	0.78	0.78	0.78	0.87	0.80	0.83	0.80	0.80	0.80	0.80	0.80
BERT	SFS	0.79	0.83	0.81	0.83	0.88	0.73	0.79	0.73	0.83	0.78	0.76	0.78
	SFS-FS	0.79	**0.85**	**0.81**	**0.85**	0.88	**0.92**	**0.90**	**0.92**	0.68	0.79	0.73	0.79
	SCOCO	0.78	0.77	0.76	0.77	0.86	0.80	0.83	0.80	0.76	0.79	0.76	0.79
	SCOCO-FS	**0.80**	0.82	0.80	0.82	**0.89**	0.91	0.90	0.91	**0.84**	**0.82**	**0.80**	**0.82**

Table 5: Sentence evaluation precision, recall, F1 and accuracy on individual tasks.

Embedding representations can capture underlying meanings and relationships. Different embeddings trained on distinct datasets may focus on particular aspects of linguistic and context information. We use two pre-trained contextual embedding and a customized embedding trained by using the COCO captions, which is much more similar to the data we intend to evaluate. The results show that BERT generally outperforms others in all tasks. Word2vec embeddings achieve results comparable to those of using BERT and outperform ELMo on Task 2. For the tasks involving sentence-level semantic meanings, however, its performance is inferior to the two contextualized representations.

The results we have presented here point the way to new approaches for automatically scoring tasks used in clinical diagnosis of language impairments, where labeled data for training models is typically scarce. In our future work, we will apply these models to data we are currently collecting from children with language disorders, autism spectrum disorder, ADHD, and typical development. Although there are differences between the populations studied (MTurk workers vs. children) and the modalities in which the responses were (written vs. spoken), our findings demonstrate the robustness of our methods even in the presence of domain or modality mismatch.

Acknowledgements

We thank Beth Calamé, Julie Bird, Kristin Hinton, Christine Yang, and Emily Fabius for their contributions to data collection and annotation. This work was supported in part by NIH NIDCD awards R01DC012033 and R21DC017000. Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the NIH or NIDCD.

References

Jonathan S Barnhoorn, Erwin Haasnoot, Bruno R Bocanegra, and Henk van Steenbergen. 2015. Qrtengine: An easy solution for running online reaction time experiments using qualtrics. *Behavior research methods*, 47(4):918–929.

Samuel Bell, Helen Yannakoudakis, and Marek Rei. 2019. Context is key: Grammatical error detection with contextual word representations. *arXiv preprint arXiv:1906.06593*.

Chris Brockett, William B Dolan, and Michael Gamon. 2006. Correcting esl errors using phrasal smt techniques. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 249–256. Association for Computational Linguistics.

Gina Conti-Ramsden, Zoë Simkin, and Nicola Botting. 2006. The prevalence of autistic spectrum disorders in adolescents with a history of specific language impairment (sli). *Journal of Child Psychology and Psychiatry*, 47(6):621–628.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Angela Grimm and Petra Schulz. 2014. Specific language impairment and early second language acquisition: the risk of over-and underdiagnosis. *Child Indicators Research*, 7(4):821–841.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Masahiro Kaneko and Mamoru Komachi. 2019. Multi-head multi-layer attention to deep language representations for grammatical error detection. *arXiv preprint arXiv:1904.07334*.

Sudhanshu Kasewa, Pontus Stenetorp, and Sebastian Riedel. 2018. Wronging a right: Generating better errors to improve grammatical error detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4977–4983, Brussels, Belgium. Association for Computational Linguistics.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Courtenay Frazier Norbury, Debbie Gooch, Charlotte Wray, Gillian Baird, Tony Charman, Emily Simonoff, George Vamvakas, and Andrew Pickles. 2016. The impact of nonverbal ability on prevalence and clinical presentation of language disorder: evidence from a population study. *Journal of Child Psychology and Psychiatry*, 57(11):1247–1257.

Matthew Peters, Waleed Ammar, Chandra Bhagavatula, and Russell Power. 2017. Semi-supervised sequence tagging with bidirectional language models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1756–1765, Vancouver, Canada. Association for Computational Linguistics.

Radim Rehurek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Citeseer.

Marek Rei. 2017. Semi-supervised multitask learning for sequence labeling. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2121–2130, Vancouver, Canada. Association for Computational Linguistics.

Sara Rosenbaum and Patti Simon. 2016. *Speech and Language Disorders in Children: Implications for the Social Security Administration's Supplemental Security Income Program.* ERIC.

Alla Rozovskaya and Dan Roth. 2010. Generating confusion sets for context-sensitive error correction. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pages 961–970. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Eleanor Semel, Elizabeth Wiig, and Wayne Secord. 2003. *Clinical Evaluation of Language Fundamentals–Fouth Edition (CELF-4)*. NCS Pearson, Bloomington, MN.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A new dataset and method for automatically grading ESOL texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 180–189, Portland, Oregon, USA. Association for Computational Linguistics.