Uncertainty, Bias and Interpretability for Deep Learning

Technical Milestone Report

Sanchit Gandhi, Queens' College
Supervisor: Prof. M.J.F. Gales

1 Introduction

In the last decade, the application of *deep neural networks* (DNNs) to long-standing problems in the fields of speech recognition and machine translation has brought about breakthroughs in performance and prediction power. Though these systems yield excellent performance, it is often challenging to understand how and why network decisions are made. Many of these models behave as black-boxes, failing to provide explanations for their predictions. This issue arises due to the essence of DNNs; the highly distributed nature of information and non-linearity of networks means they operate in a space which does not correspond directly to high-level concepts readily understood by humans. However, it is these same two features which enable DNNs to perform so well.

This project aims to address the problem of understanding by developing approaches for neural network *interpretability*, defined as the degree to which a human can understand the cause of a system's decision [1]. The particular application is neural machine translation (NMT) approaches for spoken language 'grammatical error correction' (GEC) and 'grammatical error generation' (GEG). Both of these tasks involve using a DNN to generate an output sequence given an input sequence, where the two sequences may be of different lengths. The purpose of a GEC system is to take an input sentence which contains grammatical errors and output a corrected sentence. The dual of GEC is GEG, in which corrected sentences are input into the model, and the output sentences contain newly formed grammatical errors.

Spoken language GEC is an important mechanism to help learners of a foreign language improve their spoken grammar, particularly in the context of automatic assessment of second language acquisition in computer assisted language learning (CALL). Improving network interpretability for GEC and GEG paves the way for a number of speech and language processing applications.

There are large amounts of labelled written text data which can be used to train GEC/GEG models. However, due to the differences between written and spoken communication, these text-based systems do not generalise well to speech transcripts. Written text tends to be more formal than spoken communication, with greater care taken over word choice, punctuation and structure. Furthermore, there are often disfluencies in spontaneous speech, caused by repetitions and hesitations, which are not found in written text. These differences alter the grammatical structure, which in turn affects the error correction/generation process. Since there is little labelled speech data available, direct training is not possible.

Uncovering insights into the operation of these DNNs provides a basis for developing methods relating to network *control*. This intention of this project is to establish a framework by which the network's predictions can be modified in a predictable and reliable way. By constructing the control parameters which govern this process as a direct product of the findings of an interpretability analysis, they themselves have interpretable meanings and significances. In doing so, a GEG model can be trained on labelled written text data, of which there is an abundance, and then controlled to generate outputs with grammatical errors which closely match low-resource speech data, a procedure termed *data augmentation*. Large amounts of this grammatically incorrect data can easily be produced, which can then go on to be used to directly train a GEC model, thus overcoming the issue of insufficient speech data.

Domain adaptation offers an alternative solution to the problem of a mismatch between source and target data. In this case, a GEC/GEG model is trained in the written text domain, followed by a form of transfer learning, in which the control parameters of the model are fine-tuned on the speech domain. Optimising just the control parameters returns substantial computation savings compared to directly adapting all the learnable parameters of the DNN. Furthermore, due to their interpretable essence, the changes to parameter values are indicative of the modifications made to the model in fine-tuning from the source to target domain, very much unlike the parameters of the DNN.

It is critical that GEC/GEG systems employed for spoken language assessment are unbiased in order to maintain trust in their results. These systems should be insensitive to factors that ought not affect the error correction process, such as gender, age and first language (L1). By nature, the data used to

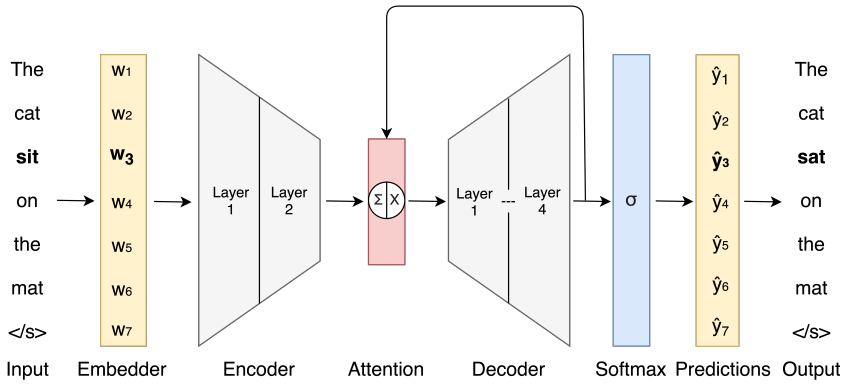


Figure 1: Sequence to sequence encoder-decoder RNN GEC model.

train these systems is uneven and may contain human bias, making these models susceptible to bias. The interpretable framework for model control developed in this project provides a mechanism by which biases in the training data can be compensated for.

DNNs tend to make over-confident predictions, especially on out-of-domain (OOD) data. A GEC/GEG system operating on a candidate with an L1 not seen during training is very likely to yield erroneous results. Estimating uncertainty in a model's predictions is a hugely important topic, as it enables for acting on the model's prediction in an informed manner. A certain degree of interpretability is required for treating network *uncertainty*, in order quantify the extent to which the model is behaving as intended.

2 Network Perturbation

The model architecture is an encoder-decoder based *recurrent neural network* (RNN) system, shown diagrammatically for a GEC system in Figure 1. It follows the standard model presented in [2]. It is assumed there is supervised training data $\mathcal{D}$, consisting of an input sequence of T words $\boldsymbol{w}_{1:T}$ and an K-length target sequence $\boldsymbol{y}_{1:K}$, from which the network parameters $\boldsymbol{\theta}$ are trained by assessing the network predictions $\hat{\boldsymbol{y}}_{1:L}$:

$$\mathcal{D} = \{\boldsymbol{w}_{1:T}, \boldsymbol{y}_{1:K}\}; \quad \hat{y}_i = \mathcal{F}\left(\boldsymbol{w}_{1:T}, \hat{\boldsymbol{y}}_{<i}, \boldsymbol{y}_{<i}; \boldsymbol{\theta}\right), \tag{1}$$

where $\mathcal{F}$ is a non-linear mapping performed by the DNN. The network's predictions are partitioned into two distinct stages. The first is a mapping from the t-th input word w_t to the hidden-layer activation from a particular layer $\boldsymbol{h}_t$. The recurrent nature of the model means $\boldsymbol{h}_t$ is also dependent on the hidden-layer activation for the previous word $\boldsymbol{h}_{t-1}$. The second mapping takes the sequence of hidden-layer activations $\boldsymbol{h}_{1:T}$ and the $i-1$ previous predicted words $\hat{\boldsymbol{y}}_{<i}$ to form the network's prediction of the i-th output $\hat{y}_i$:

$$\boldsymbol{h}_t = \mathcal{F}_{\text{h}}\left(w_t, \boldsymbol{h}_{t-1}; \boldsymbol{\theta}\right); \quad \hat{y}_i = \mathcal{F}_{\text{y}}\left(\boldsymbol{h}_{1:T}, \hat{\boldsymbol{y}}_{<i}; \boldsymbol{\theta}\right). \tag{2}$$

The global framework for controlling the model involves adding a perturbation $\Delta\boldsymbol{h}_t$ to the hidden-layer activation $\boldsymbol{h}_t$. The non-linear mappings $\mathcal{F}_{\text{h}}$ and $\mathcal{F}_{\text{y}}$ are unchanged between the original and perturbed models. Hence, the hidden-layer activation for the t-th word in the perturbed model $\tilde{\boldsymbol{h}}_t$ and the associated prediction of the i-th output $\tilde{y}_i$ are given by:

$$\tilde{\boldsymbol{h}}_t = \boldsymbol{h}_t + \Delta\boldsymbol{h}_t; \quad \tilde{y}_i = \mathcal{F}_{\text{y}}\left(\tilde{\boldsymbol{h}}_{1:T}, \tilde{\boldsymbol{y}}_{<i}; \boldsymbol{\theta}\right). \tag{3}$$

The aim of the project is to uncover perturbations $\Delta\boldsymbol{h}_t$ which enable the network's predictions $\tilde{\boldsymbol{y}}_{1:M}$ to be controlled in a predictable way. The nature of the perturbations is specific to the task for which the model is to be controlled. For instance, controlling a GEG model to generate fewer or more grammatical errors belonging to a particular part-of-speech (PoS), such as nouns, verbs, adjectives, etc., requires perturbations $\Delta\boldsymbol{h}_t$ specific to that PoS. Likewise, correcting for gender bias requires perturbations $\Delta\boldsymbol{h}_t$ which correspond to gender features.

Thus far, the project has concentrated on hidden-layer activations at the encoder output layer, the second layer of the encoder in Figure 1. However, the approaches covered are not restricted to just this layer. They generalise to any number of hidden-layers within the model, potentially give rise to a greater degree of controllability. Likewise, the methods are not specific to the architecture of the model. Whilst

currently RNN models have been used, these methods could just as well be applied to different model architectures, such as *transformers*.

The simplest form of perturbation is one in which individual nodes in the hidden-layer are stimulated. This method amounts to discovering individual neurons with activation outputs that can be perturbed to yield a controllable model. The findings of the preceeding UROP [3] showed this to be an ineffective way of controlling the GEC model to correct errors for particular parts-of-speech and not others; stimulating individual nodes impacted the error correction performance of the model over many parts-of-speech to a similar degree. This is consequence of the highly distributed nature of information within DNNs, single neurons being responsible for the error correction process for numerous parts-of-speech.

An alternative scheme involves perturbing linear combinations, or directions, of nodes. Here, the activation output space is transformed into another vector space by means of a projection matrix. Perturbations are then applied in this transformed space, in an entirely equivalent way to node stimulation in the original activation space. The perturbed vector is then transformed back into the original space by inverting the projection operation, returning the final perturbation to be applied. The objective in this scheme is to find a projection matrix which transforms the hidden-layer activations into an interpretable space, whilst maintaining model controllability. The UROP showed this method to be far superior to individual node perturbation; the GEC/GEG models were successfully controlled to correct/generate errors for particular parts-of-speech and not others. It is, therefore, this line of direction orientated work which is pursued further in the project.

The UROP focused primarily on correlating hidden-layer activations $\boldsymbol{h}_t$ over ensembles of independently trained networks to uncover directions for network perturbations. This line of work, termed cross-model correlation analysis, is based on the assumption that important properties are captured in similar ways in multiple networks. It serves a particular use when supervised data for the perturbation directions is not available, requiring only pairs of input words and hidden-layer activations. Singular vector canonical correlation analysis is one such correlation method which was shown to yield controllable GEC and GEG models [3]. However, in many situations having multiple models may not be feasible, due to the computation expense of training an array of identical DNNs. The project addresses this point, by finding and developing a new, single model approach, explained in detail in Section 3.

3 Logistic Classification

The aim of this work is to devise a new, single model method for uncovering salient directions $\Delta\boldsymbol{h}_t$ to be used for model control in GEC/GEG systems. The task is set-up first by generating fully annotated (supervised) data specific to the control task. Each word w_t in the input sequence is assigned a label l_t from a label set $\mathcal{P}$, where the label set is specific to the control task of interest. The input set $\boldsymbol{w}_{1:T}$ is then fed into the network, and the set of hidden-layer activations $\boldsymbol{h}_{1:T}$ at the encoder output layer recorded, giving a sequence of $\{w_t, l_t, \boldsymbol{h}_t\}$ trios. The problem can then be treated as a classification task, where the goal is to predict the label for the t-th word l_t from the hidden-layer activation vector $\boldsymbol{h}_t$. This is based on the underlying assumption that if a classifier of simple linear architecture can predict the label assigned to a word, then the hidden-layer representations implicitly encode this information. A logistic classifier is trained on the $\{\boldsymbol{h}_t, l_t\}$ pairs to yield the maximum likelihood estimate (MLE) of parameters $\boldsymbol{\theta}^c$:

$$\hat{\boldsymbol{\theta}}^c = \underset{\boldsymbol{\theta}^c}{\text{argmax}} \left\{ \sum_{t=1}^{T} \log \left(P(l_t | \boldsymbol{h}_t; \boldsymbol{\theta}^c) \right) \right\} = \underset{\boldsymbol{\theta}^c}{\text{argmax}} \left\{ \sum_{t=1}^{T} \log \left(\frac{\exp(\boldsymbol{\theta}^c_{l_t} \cdot \boldsymbol{h}_t)}{\sum_{l \in \mathcal{P}} \exp(\boldsymbol{\theta}^c_l \cdot \boldsymbol{h}_t)} \right) \right\}, \tag{4}$$

where $\boldsymbol{\theta}^c_{l_t}$ is the vector of learned classifier weights for the t-th input word with label l_t. The full-form of the softmax probability in eq. (4) provides a clear interpretation for these weights: important directions in the logistic classifier are retained, whilst unimportant directions are diminished. The classifier's parameters therefore provide a direct measure of the importance of each direction. Multiple directions for each label are obtained by recursively using the classifier, and eliminating used directions from the hidden-layer activations.

The perturbation directions $\Delta\boldsymbol{h}_{t,l}$ associated with the label l are formed very simply from the classifier directions $\boldsymbol{\theta}^c_l$. The classifier weights are normalised in the l_2-norm, and a zero-mean version of $\boldsymbol{h}_t$ projected onto the directions spanned by the weights $\boldsymbol{\theta}^c_l$. The level of perturbation is set in this transformed space by multiplying the projected activation by a scalar factor α_l. The overall perturbation $\Delta\boldsymbol{h}_t$ is formed as

the summation over all labels in the label set $\mathcal{P}$:

$$\Delta \boldsymbol{h}_t = \sum_{l \in \mathcal{P}} \alpha_l \Delta \boldsymbol{h}_{t,l} = \sum_{l \in \mathcal{P}} \alpha_l \left(\left(\boldsymbol{h}_t - \bar{\boldsymbol{h}} \right) \cdot \boldsymbol{\theta}_l^c \right) \frac{\boldsymbol{\theta}_l^c}{\|\boldsymbol{\theta}_l^c\|_2}. \tag{5}$$

The interpretation of the scalar factors α_l is as follows. A value of $\alpha_l = 0$ produces no perturbation, the $\Delta \boldsymbol{h}_{t,l}$ term disappearing from the expression in eq. (5), whilst $\alpha_l < 0$ and $\alpha_l > 0$ correspond to under and over-stimulating directions respectively. The value $\alpha_l = -1$ is of particular significance; it results in complete erasure of the component of the hidden-layer layer activation in the direction $\boldsymbol{\theta}_l^c$, thus eliminating it completely from the perturbed vector $\tilde{\boldsymbol{h}}_t$. This is equivalent to ablating an individual node in the transformed vector space. With this insight, a clear mechanism for model control arises, in which the size and magnitude of the scalar factors α_l are varied in order to under or over-stimulate particular directions.

4 Results

The GEC and GEG systems were trained on data taken from the Cambridge Learner Corpus (CLC) [4]. This corpus consists of written examinations of candidates at different proficiency levels with 86 different L1s. Grammatical errors in the erroneous learner text have been carefully annotated and correct native responses transcribed, giving paired training data of erroneous and correct sentences. The GEC model was trained with the erroneous sentences as the inputs and correct sentences as the target translations. A GEG model was trained simply by copying the GEC architecture, and swapping the input and target files for training: the correct sentences were made the inputs and erroneous sentences the targets. Thus, if perfectly operational, the GEG model can be made to exactly undo the corrections made by the GEC model.

Four different test sets, of varying similarity to the CLC training set, were used for evaluation. The FCE test set is a held-out subset of the CLC, and as such is in-domain (ID) with the training data. NICT-JCE [5] is a publicly available non-native speech corpus. It is based on manual transcriptions of English oral proficiency interviews provided by Japanese learners of varying abilities. BULATS [6] is a corpus derived from a free speaking business English test consisting of prompted responses of up to one minute in length. Both manual (BLT) and automatic speech recognition (BLT-ASR) transcriptions were used. The learners of this corpus are from six L1s and span the full range of English speaking abilities. Other than FCE test, the evaluation sets are based on speech transcripts, and as such are out-of-domain (OOD). The predictive performance of the baseline GEC and GEG models was quantified through the use of the Generalized Language Evaluation Understanding metric (GLEU) [7], values for which are shown in Table 1. From top to bottom, the domain mismatch between training and test sets increases.

Table 1: GLEU scores for baseline and fine-tuned models.

Test Set	**GEC**			**GEG**		
	Baseline	`LClass`	`SVCCA`	Baseline	`LClass`	`SVCCA`
FCE test	68.05	68.22	68.24	56.15	56.19	56.17
NICT	44.59	44.86	46.34	42.41	42.45	42.39
BLT	47.96	47.92	48.01	41.19	42.34	42.38
BLT-ASR	28.99	29.32	29.20	22.70	22.79	25.97

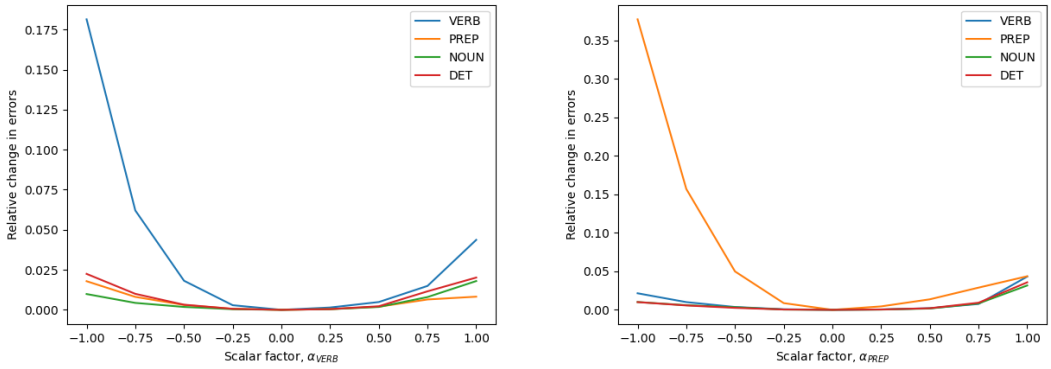
An investigation was performed into control of the baseline GEG model on the FCE test set, with the aim of perturbing the model to generate grammatical errors for certain parts-of-speech and not others. PoS labels for the input sequence in the CLC training set were generated using the **Stanford Neural Network Dependency Parser**[1], a piece of software that reads the input text and assigns a label from a specified PoS set to each word. Having trained the classifier to obtain the MLE of the classifier weights, the perturbation directions were constructed as in eq. (5).

As an unconventional NMT task, a bespoke metric for quantifying the degree of network control was formed. This metric was defined as the relative change in grammatical errors generated between the perturbed and baseline models for each PoS. In doing so, a controllable model can easily be identified as one in which the additional errors generated by the perturbed model can be increased for specific parts-of-speech and not others.

[1]Stanford Neural Network Dependency Parser: `https://nlp.stanford.edu/software/nndep.html`

For each PoS, in turn, the value of the scalar factor α_l was varied, and the level of perturbation in every other PoS direction set to zero. The output of the perturbed model $\tilde{\mathbf{y}}_{1:M}$ was recorded for each value of α_l. The number of grammatical errors generated in this output was compared to that of the original model by measuring the Levenshtein distance between the two translations for each PoS, yielding the relative change in errors generated by perturbing the model. This operation was realised through the use of the **grammatical ERRor ANnotation Toolkit** (ERRANT)[2], a toolkit which automatically extracts and classifies edits between original and modified sentence pairs.

Figure 2(a) plots the relative change in number of errors as a function of the level of verb perturbation, and Figure 2(b) as a function of the level of preposition perturbation. In both cases, under-stimulating ($\alpha_l < 0$) the PoS direction of focus drastically increases the number of grammatical errors generated by the perturbed GEG model for that PoS, whilst keeping the number of additional grammatical errors for other parts-of-speech low. This is exactly the behaviour required of a controllable model, thus demonstrating that salient directions have in fact been returned by the logistic classifier, and that a functional framework for control has been established.



(a) VERB **(b)** PREP

Figure 2: Relative change in grammatical errors as a function of the level of perturbation. Panel (a) is for the perturbation of verb directions. Panel (b) is for the perturbation of preposition directions.

By under-stimulating PoS directions, the representation of hidden-layer activations associated with that PoS is reduced, resulting in the model generating more grammatical errors of that type. The same is not true for over-stimulation. In Figures 2(a) and 2(b), as α_l is increased above zero, the number of errors increases similarly for all four parts-of-speech. This universal increase is attributed to the nature of the hyperbolic tangent activation function used in the encoder output layer; the function saturates when its argument is very positive or very negative, meaning the function becomes very flat and insensitive to small changes in its input. Over-stimulating pushes the hidden-layer activations into the saturated region, resulting in fewer grammatical errors being generated compared to when operating in the linear region, as with under-stimulation.

Having trained the GEC/GEG models in the written text domain on the CLC training data, an exploration into domain adaptation was performed by fine-tuning the models to the target speech domain. In this case, fine-tuning consisted of the optimisation process which involved finding the set of scalar factors $\boldsymbol{\alpha} = \{\alpha_l\}_{l \in \mathcal{P}}$ which maximise the probability of observing the target data under the perturbed model. The optimal scalar factors were found for the top 29 parts-of-speech by frequency in the training set. Covering a broad range of parts-of-speech enables for adaptation throughout the distribution of data. Table 1 reports the GLEU scores for the baseline and fine-tuned GEC and GEG models. Scores are shown for the fine-tuning of both the logistic classifier (`LClass`) and singular vector canonical correlation analysis (`SVCCA`) directions. On average, across both models, there was a 0.490% improvement in GLEU score following the optimisation of `LClass` directions, and 1.51% increase for the optimisation of `SVCCA` directions, with the greatest changes observed when the model was adapted to OOD data sets. Although these results demonstrate a degree of domain adaptation, the improvements are somewhat low. A number of alterations and developments are proposed in Section 5. Incorporating these changes might well improve the extent to which the fine-tuned models outperform the baseline ones.

[2]ERRANT Toolkit: `https://github.com/chrisjbryant/errant`

5 Future Work

One potential reason why the gains for optimising the models in Table 1 are not more significant is due to overlap in perturbation directions for different parts-of-speech. Consider two distinct parts-of-speech with labels l and l', with corresponding perturbation directions $\Delta \boldsymbol{h}_{t,l}$ and $\Delta \boldsymbol{h}_{t,l'}$. Should there be shared representations between $\Delta \boldsymbol{h}_{t,l}$ and $\Delta \boldsymbol{h}_{t,l'}$, then the interaction between perturbation directions will hinder the extent to which the level of perturbation for each PoS can be optimised. A solution to this is to enforce an orthogonality constraint in the offsets. For the logistic classifier, this amounts to adding a regularisation term which penalises overlaps in the directions of classifier weights, a modification which is currently being trialled.

The distribution of parts-of-speech between data sets from different domains is only one aspect by which they differ. Optimisation of directions relating to punctuation, tenses, sentence structure etc. would likely result in greater improvements in predictive performance. The key question in tackling this task is finding a means by which these features can be identified in a data set, such that supervised data can be generated for training the logistic classifier. This was not much of an issue for parts-of-speech, the availability of the Stanford Parser making it a trivial process. Other grammatical features require more refined software or hand-labelling of training data. Uncovering salient directions for more features and expanding domain adaptation is a potential avenue of work which could be explored further.

Alongside domain adaptation, another control related task is bias correction. The framework established in this project for model control provides a mechanism by which biases in the training data can be corrected. Directions relating to features for which there is a bias could be identified, and the model perturbed as to eliminate the biases. Once again, difficulty arises in labelling these features to produce sufficient amounts of training data.

Having uncovered a certain degree of interpretability, the project has the scope of shifting more towards treating model uncertainty. A scheme of work which bridges from interpretability to uncertainty is currently being investigated. Here, the hypothesis is that in instances when the GEC/GEG model accurately corrects/generates a grammatical error for the input word w_t with a PoS label l, the hidden-layer activation $\boldsymbol{h}_t$ aligns with the perturbation directions associated with that PoS $\Delta \boldsymbol{h}_{t,l}$. Preliminary results for the experiments testing this hypothesis appear promising, with more work required to certify that the model is in fact operating as expected. This work has the potential to be developed, with the goal of devising a perturbation-inspired metric which quantifies the degree of uncertainty in a DNN.

References

[1] T. Miller, "Explanation in Artificial Intelligence: Insights from the Social Sciences," *arXiv e-prints*, p. arXiv:1706.07269, June 2017.

[2] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, Ł. Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, and J. Dean, "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation," *arXiv e-prints*, p. arXiv:1609.08144, Sept. 2016.

[3] S. Gandhi, "ALTA Report," 2020. [Online]. Available: `http://mi.eng.cam.ac.uk/raven/esol/ind_reports/sg836/2020-09.pdf`.

[4] D. Nicholls, "The Cambridge Learner Corpus - error coding and analysis for lexicography and ELT," *Proc. of the Corpus Linguistics 2003 conference; UCREL technical paper number 16*, 2003.

[5] E. Izumi, K. Uchimoto, and H. Isahara, "The NICT JLE Corpus: Expliting the language learners' speech database for research and education," *International Journal of the Computer, the Internet and Management*, vol. 12, no. 2, pp. 119–125, 2004.

[6] L. Chambers and K. Ingham, "The BULATS online speaking test," *Research Notes*, p. 21–25, 2011. [Online]. Available: `https://www.cambridgeenglish.org/Images/23161-research-notes-43.pdf`.

[7] C. Napoles, K. Sakaguchi, M. Post, and J. Tetreault, "Ground Truth for Grammatical Error Correction Metrics," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, (Beijing, China), pp. 588–593, Association for Computational Linguistics, July 2015.