A B. Tech Project Report Submitted
in Partial Fulfillment of the Requirements
for the Degree of

Bachelor of Technology

by

Aryan Chauhan Rishikesh Songra
((180101012) (180101065))

under the guidance of

Amit Chintamani Awekar

to the

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI
GUWAHATI - 781039, ASSAM

CERTIFICATE

*This is to certify that the work contained in this thesis entitled "**Language Model Score for Grammatical Evaluation**" is a bonafide work of **Aryan Chauhan (Roll No 180101012)** , **Rishikesh Songra (Roll No 180101065)** carried out in the Department of Computer Science and Engineering, Indian Institute of Technology Guwahati under my supervision and that it has not been submitted elsewhere for a degree.*

Supervisor: **Dr. Amit Chintamani Awekar**

Associate Professor,

Nov, 2022

Department of Computer Science & Engineering,

Guwahati.

Indian Institute of Technology Guwahati, Assam.

Acknowledgement

We'd like to take this opportunity to thank our supervisor, **Dr.Amit Chintamani Awekar, IIT Guwahati CSE** department for his constant support, patience, enthusiasm, and encouragement during our B.Tech project. We appreciate his valuable input, insights, and guidance throughout the assignment, which aided us in gaining a theoretical and practical understanding of the subject. We appreciate his useful feedback on each problem or difficulty we encountered with our project, and he was always willing to help.

Sincerely,

Aryan, Rishikesh

Abstract

In the field of Natural Language Processing(NLP) rapid growth and progress has been seen in the recent years. Many manual tasks have been automated in the recent years with the help of Language Models. The accuracy of these models has been increasing given the developments in the compute ability of the hardware.

Motivated by recent findings on the probabilistic modeling of acceptability judgments, we have proposed few language model scores for reference-less grammar evaluation of natural language generation output at the sentence level.Using these scores we can harness a more compact language model potential. Our findings suggest that the current way of normalization of log-likelihood by the length of the sentences is not optimal. We show that WSNLL and KPPL yield a significantly higher correlation with human judgments than all other LM scores and Bleu(a reference based metric).

Contents

List of Figures

Chapter 1

Introduction

Artificial neural networks and deep learning approaches have been used in a variety of domains in recent years. Natural language processing (NLP) based on deep learning and machine learning has become a major topic among these domains. In NLP, grammar error correction (GEC) [CPL20] refers to identifying and correcting grammatical/spelling mistakes that appears in a sentence .

Currently for evaluating the grammatical-ness of sentences we have a lot of reference based metrics like MaxMatch , GLUE etc, these reference based require ground truth(grammatically correct sentences) sentences for computation of scores. There are few reference-less scores available like perplexity, but the downside being that they are not efficient on specific task like comparing grammar of sentences .

In this work we will present some novel reference less language model scores which enhances the performance of current LM's on specific tasks (In this paper we have chosen the task as grammatical sorting).

1.1 Motive for New Scores

We have observed that the current research in NLP is more focused on improving the language model themselves which is crucial part but we believe that choosing a optimal score metric is also very important to harness the full potential of LM's. Current LM's score are not sufficient for every use case this we have emperically demonstrated in our work.

Specifically, we test our hypothesis that is our score should be a suitable for evaluation of grammar which

- Does not rely on references(Sentences which are used as ground truth for the evaluation).
- Does not need human grammar annotations of any kind.

The first characteristic, notably, the fact that our scores do not require references, makes it a good candidate for automatic evaluation. Getting rid of human references is useful in a number of situations, such as when references are unavailable owing to a lack of annotation resources or when getting references is unfeasible.

1.2 Contributions

To summarize we have made the following contributions

- We have build novel scores for evaluating the grammatical correctness of sentences which we have tested on the task of grammatical sorting. **Grammatical sorting - Sorting a list of sentences such that more grammatically correct sentences appear at the beginning of the list**

Chapter 2

Literature Survey

We looked into a number of reference-based measures for GEC, including the F-score, precision, recall, MaxMatch, and GLUE. The F-score, precision, and recall are very well performance indicators in NLP. The F-score, which employs the harmonic mean of accuracy and recall as the final performance rating, was perhaps the most extensively utilized measure in the first GEC study. However, these measures have a flaw in that they can't analyze sentences that are longer than a phrase.

Furthermore, the F-score is incapable of distinguishing between "no mistakes" and "incorrect changes" in the GEC model. To overcome the limits of existing approaches, Dahlmeier and Ng [DN12] proposed the MaxMatch scorer, which could take into account modifications up to the phrase level.

MaxMatch, on the other hand, demands specific mistake annotations. These measures have restrictions since they are reference-based metrics that will not operate without a ground truth reference.

With the development of deep learning, the GEC measure was primarily GLUE [CNT15] and the bilingual evaluation understudy (BLEU) [KPZ02]. BLEU is a machine translation (MT) statistic that analyzes the outcomes of MT with human translation. The n-gram is used to determine the measurement standards. This measure may be used in any language and has the benefit of being quick to calculate.

GLUE also requires human annotators to fix a sentence by recreating the original text, as recommended by Napoles. The distinction with GLEU is that it takes into account the source sentence and is a correction system-specific performance assessment statistic. This measure is used as the official GEC metric in the bulk of current research.

In today's world with new progressions in Deep Learning based Language Models, LM's performance has greatly improved, now LM's can generate score which can evaluate sentences without the need of any reference/annotation.

Therefore LM's are perfect candidates for automating NLP based task like grammar error correction. In our experiments, we have used GPT-2 variations.

One of the most often used measures for assessing language models is perplexity (PPL) [Per14]. It's important to note that the measure only applies to classical language models (also known as auto regressive or causal language models) and isn't adequately defined for masked language models like as BERT [JDT18].

Perplexity has an advantage over other metrics that it is a reference-less metric and does not require human based annotations, so we have used perplexity as a baseline comparison.

We have evaluated our metrics on the task of grammatical sorting. Calculating the number of inversions is the classical way of determining how much sorted an array is, by taking inspiration from this fact we have defined something similar to inversions in our work.

Chapter 3

Proposed Metrics

In this section, we first describe **MCP(Minimum Contextual Probability)**, **WSNLL(Weighted Sum of Negative log-likelihoods)** and **KPPL(Kth-perplexity)** and look at the intuition behind these metrics/scores.

We have tried two approaches to compute the contextual probabilities(defined in Section 3.1) vector

- We first remove the words which had contextual probability below a certain threshold and then for the rest of the words recalculated the contextual probabilities.
- In the second approach we have included all the words while calculating contextual probabilities.

3.1 Minimum Contextual Probability

Given a sentence X tokenized as [x0,....xn-1] , MCP is defined as the minimum of the contextual probabilities of all the tokens in the sentence.

$$MCP = min(p_\Theta(x_i|x_{<i})) \quad \forall \quad i \quad \epsilon \{1, 2..., n\}$$

where $p_\Theta(x_i|x_{i-1})$ is the contextual probability of the ith token conditioned on the preceding tokens $x_{<i}$.

The intuition behind this score is that the token with minimum probability will denote the most unlikely word in the sentence,this word is the most out of context word in the sentence. Hence we expect a sentence with lower minimum probability score will be more grammatically incorrect with respect to a sentence having higher minimum probability.

3.2 Weighted Sum of negative log-likelihoods

Like MCP here we take the contextual probability vector and construct the negative log-likelihood vector(NLLV) corresponding to it.Then we sort this NLLV in descending order.Finally we take a heuristical weight vector where ith term is $\alpha^i(\alpha < 1)$ using this weight vector and sorted NLLV we output their dot product as the score.

$$WSNLL = \sum_{i=1}^{n-1} - \ log(\ p'_{\Theta}(x_i|x_{<i})\ )\ *\ \alpha^i$$

The intuition behind this score is that for grammar correctness the word having the least probability should contribute more to the score. Hence the weight are less for more more in-context words.

3.3 Kth-perplexity

This score is almost identical to perplexity except that here we divide by n^k rather than n.

$$\log(KPPL) = \sum_{i=1}^{n-1} \frac{-\ log(\ p'_{\Theta}(x_i|x_{<i})\ )}{n^k}$$

The intuition is that say a sentence of length 20 has 2 out of context words and another sentence of length 10 has only 1 out of context word. Then perplexity (or 1PPL) for both the sentences would be similar but according to intuition the first sentence of length 20 is bad when it comes to the grammatical correctness as it has two errors as compared to the sentence of length 10. Therefore, some power $k < 1$ would be a better measure when it comes to the grammatical correctness of the sentence.

Chapter 4

Experimental Setup and Results

4.1 Dataset Construction

We experimented on the CONLL-13 [HTN13] dataset. CONLL-13 comprises of a many grammatical error types like ,including spelling mistakes , determinant , noun-number , subject-verb , verb-form , preposition and agreements error. CONLL-13 also comprises of sentences containing multiples interacting errors.

Examples of such interacting errors

- Noun number and determiner error
  - "that cars" $->$ "that car" / "those cars".

- preposition and verb form
  - "an interest to study" $->$ "an interest in studying".

The above dataset is first converted to a list of pairs. Where every pair contains a sentence and its label. The label is a boolean value denoting if the sentence is correct or not. Our dataset comprises of approximately 1400 sentences of which 260 of grammatically correct.

4.2 Optimal Parameters Selection

We first divided the sentences dataset into two parts namely validation and testing. 200 sentences for validation and 1181 for testing. We have used grid search for finding the optimal parameters using the validation dataset. We picked the parameters which maximize the pearson correlation coefficient between the scores array and the labels array .Then we have tested these parameters on the testing dataset.

4.3 Pre-Trained Language Model

We have used the following pre-trained LMs :

1. GPT2 [AR19] Small
2. GPT2 Medium
3. GPT2 Large
4. OpenAI-GPT [ARS18]

5. GPT-Neo

We have used library's default Hyper-parameters.

4.4 Baseline Metrics

We are comparing MCP, WSNLL and KPPL Metrics with Perplexity and BLEU as baseline metrics.

1. Perplexity : Our first baseline is perplexity, which is commonly used for evaluating LM's,which corresponds to the exponentiated cross-entropy:

$$\log(PPL) = \sum_{i=1}^{n-1} \frac{-\ log(\ p'_{\Theta}(x_i|x_{<i})\ )}{n}$$

2. BLEU : We wanted to compare our score with some reference based baseline, Hence as BLEU is a common reference based metric so we have used it in our comparison. BLEU's performance is directly tied to the similarity between the given sentence and its closest match present in the reference text .

4.5 Correlation and Evaluation Scores

We have evaluated the scores on the task of grammatical sorting as defined previously using the following metrics.

1. **TOP k-Error** : It is the number of grammatically correct sentences among the sentences having top k score,here higher score means that the

sentence is more grammatically incorrect.

2. **Number Of Inversion** : An inversion is a pair (i,j) such that i¡j and ith sentence is grammatically incorrect while the jth sentence is grammatically correct.

3. **Bottom k-Error** : It is the number of grammatically incorrect sentences among the sentences having least k scores.

4.6 Results

All the values in the tables below have been calculated using optimal parameters for KPPL and WSNLL.

GPT Results				
Metrics	Inversion	Correlation	Top-200 Error	Bottom-200 Error
WSNLL	55343	0.45	2	40
KPPL	55165	0.42	2.5	41
MCP	84024	0.3	8	57
Perplexity	92327	023	9	58.5

GPT-2 Results				
Metrics	Inversion	Correlation	Top-200 Error	Bottom-200 Error
WSNLL	52580	0.45	1.5	43
	50678	0.46	2	40.5
	51160	0.46	1	41.5
KPPL	54573	0.41	2.5	44.5
	52117	0.43	1.5	44
	51542	0.43	1	42.5
MCP	83411	0.28	7.5	56
	85077	0.27	9	54.5
	85696	0.27	8	58.5
Perplexity	129676	0.05	20.5	72.5
	127197	0.065	18.5	73
	129312	0.06	19	72.5

*** Each Cell contains value for GPT2-Small , GPT2-Medium, GPT-2Large Respectively.**

GPT-Neo Result				
Metrics	Inversion	Correlation	Top-200 Error	Bottom-200 Error
WSNLL	51362	0.46	2.5	39.5
KPPL	53546	0.42	2	42
MCP	86071	0.26	6.5	57.5
Perplexity	13500	0.03	20	75

4.7 Graph Comparisions

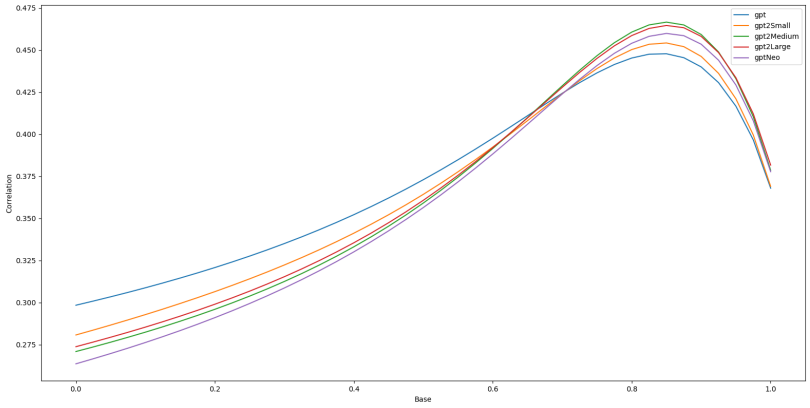

Fig. 4.1: Correlation vs WSNLL-Base α

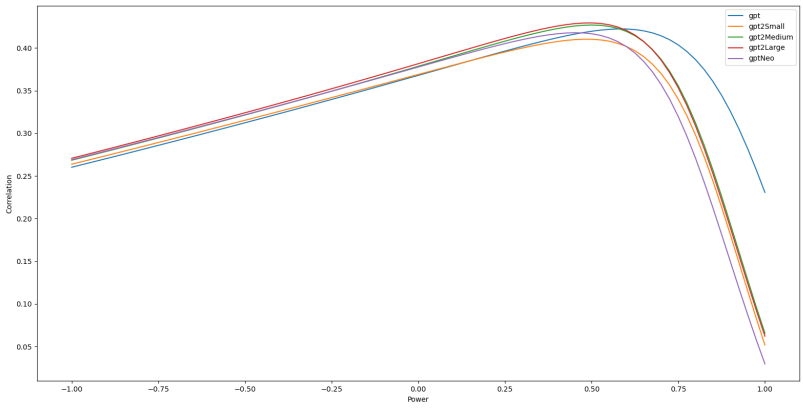

Fig. 4.2: Correlation vs KPPL Power K

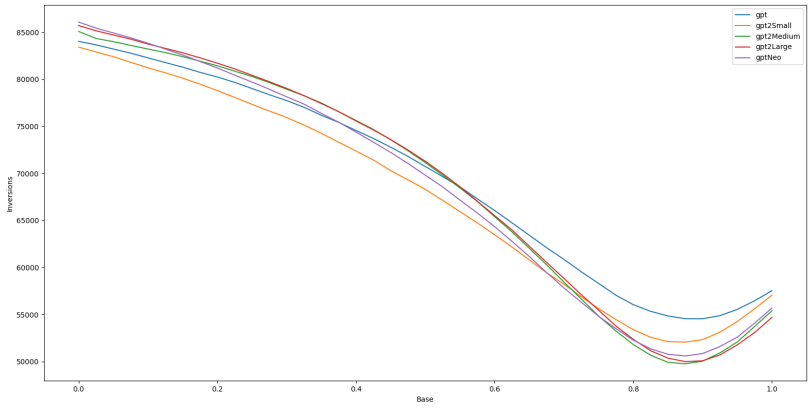

Fig. 4.3: No. Of Inversions vs WSNLL-Base α

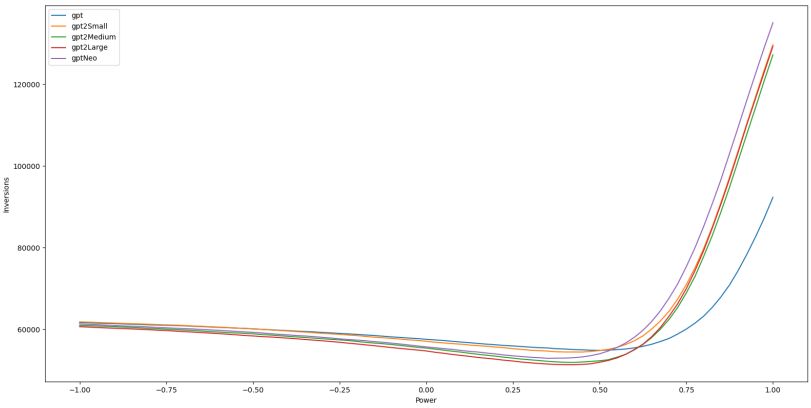

Fig. 4.4: No. Of Inversions vs KPPL Power K

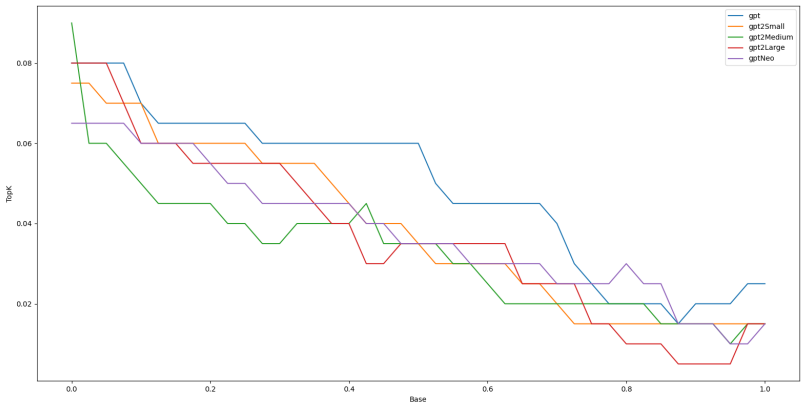

Fig. 4.5: Top 200 error vs WSNLL-Base α

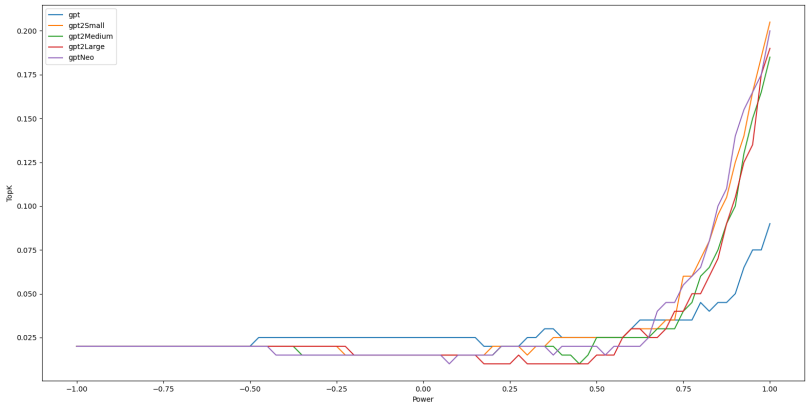

Fig. 4.6: Top 200 error vs KPPL Power K

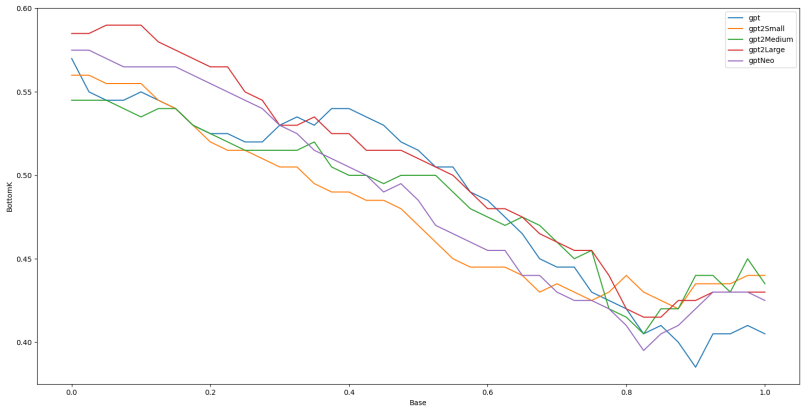

Fig. 4.7: Bottom 200 error vs WSNLL-Base α

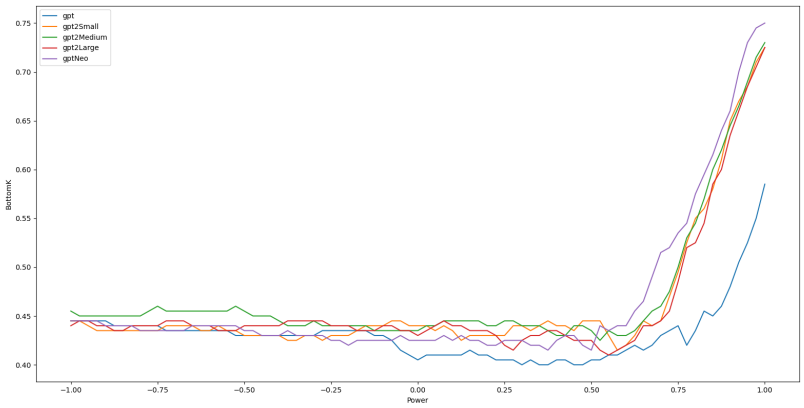

Fig. 4.8: Bottom 200 error vs KPPL Power K

4.8 Discussion

As from above result we can observe

- Following are the results of Bleu:

  Inversions Count: 133173

  Correlation: 0.076

  top 200 error: 0.24

  bottom 200 error: 0.6 The above are calculated in the following way:

  - First the most similar match of the sentence is found in the reference-text.
  - Then Bleu score is computed for the sentence by using the above found sentence as the ground truth.

- We can observe that MCP result were comparable to perplexity and BLEU. Here Inversion count,k-errors(Top-K and Bottom-k errors) were comparable to perplexity and BLEU.We believe inversion are less because MCP handles length based normalization of sentences better than baseline metrics, but the results are not significantly better because MCP discards a lot of information as it considers only one token.

- Regarding WSNLL and KPPL we have significantly better result than baseline metrics.Here Inversion count, k-errors are significantly better.Thus KPPL and WSNLL handle normalization and contextual information very well which results in low inversions and k-errors.

- We have observed that removing bad words approach leads to degraded performance.The reason why the performance degrades after removing the bad words is that our initial hypothesis of having the rest of the part wrong after encountering a bad word is not correct, as it is evident from the NLL vector. To mark bad words we have used a cutoff value, above which the word is a bad word. Eg.

  – Sentence - Some people started to think if electronic products can

be further operated to more advanced utilization and replace human beings for better performances.

– NLL's - [6.1, 2.8, 6.8, 1.5, 3.0, 5.8, 12.2, 5.8, 3.0, 0.9, 9.5, 11.4, 4.7, 5.8, 3.7, 10.6, 2.6, 7.8, 3.3, 2.0, 4.2, 5.8, 6.3, 7.5].

– Here we have kept cutoff to 14.

Chapter 5

Conclusion and Future works

5.1 Conclusion

From above experiments we can conclude the following :

- From empirical results we have found that the variation of K and alpha with correlation and number of inversions is unimodular in nature.
- We have discovered powerful non-referential language model scores like WSNLL,kPPL.
- We empirically confirmed the effectiveness of kth-perplexity and LWSNLL, LM score which better accounts for the effects of sentence length and individual unigram probabilities, as a score for grammatical correctness of sentences.

- The normalization of the LM score by simply dividing by the length of the sentence is not optimal for all tasks as it is evident from the above results for the task of grammatical sorting.
- These scores better harness the power of LM's, hence smaller models can also be used in place of larger models(GPT2-medium's performance is almost identical with that of GPT2-large's).

5.2 Limitations

There were some following limitations to our model

- These Metrics does not perform effectively when the sentences contains non-frequent words as the contextual probability of these words are very low.
- These Metrics only on the prefix context of the sentence and does not take into account the suffix of the sentence.

5.3 Future Work

Some future works related to our work are

- We would like to make the metrics to take into account non-frequent words.

- Try to work on different languages, currently our dataset contains only English sentences.
- Currently these metrics haven't been tested on the degree of error which we plan to do in future.

References

[AR19] Rewon Child David Luan Dario Amodei Ilya Sutskever Alec Radford, Jeffrey Wu. Language models are unsupervised multitask learners. 2019.

[ARS18] Tim Salimans Alec Radford, Karthik Narasimhan and Ilya Sutskever. Improving language understanding by generative pre-training. 2018.

[CNT15] M. Post C. Napoles, K. Sakaguchi and J. Tetreault. "ground truth for grammatical error correction metrics," in proc. 53rd annu. meeting assoc. comput. linguistics 7th int. joint conf. natural language process., vol. 2, 2015, pp. 588593. 2015.

[CPL20] C. Lee C. Park, Y. Yang and H. Lim. "comparison of the evaluation metrics for neural grammatical error correction with over-correction," in ieee access, vol. 8, pp. 106264-106272, 2020, doi: 10.1109/access.2020.2998149. 2020.

[DN12] D. Dahlmeier and H. T. Ng. "better evaluation for grammatical error correction," in proc. conf. north amer. chapter assoc. comput. linguistics: Hum. lang. technol., 2012, pp. 568572. 2012.

[HTN13] Yuanbin Wu Christian Hadiwinoto Joel Tetreault Hwee Tou Ng, Siew Mei Wu. The conll-2013 shared task on grammatical error correction. in proceedings of the seventeenth conference on computational natural language learning: Shared task, pages 1–12, sofia, bulgaria. association for computational linguistics. 2013.

[JDT18] K. Lee J. Devlin, M.-W. Chang and K. Toutanova. "bert: Pre-training of deep bidirectional transformers for language understanding," , arxiv:1810.04805. [online]. available: http://arxiv.org/abs/1810.04805. 2018.

[KPZ02] T. Ward K. Papineni, S. Roukos and W.-J. Zhu. "bleu: A method for automatic evaluation of machine translation," in proc. 40th annu. meeting assoc. comput. linguistics, 2002, pp. 311318. 2002.

[Per14] Perplexity. https://huggingface.co/docs/transformers/perplexity), pages 271–281. 2014.