Lexical Predictability Analysis

Author: Minghua Wu

Date: 23/7/2024

Introduction

Language processing involves continuously predicting the next word based on the prior context, and the predictability of a word increases with the amount of preceding context available. However, when the context is disordered or incoherent, the task of predicting the next word becomes more challenging. Understanding how the level of disorder affects the relationship between context length and lexical predictability can provide valuable insights into the nature of disordered language and its impact on communication.

The goal of this study is to investigate the influence of disorder level on the relationship between context length and lexical predictability. It is hypothesized that as the level of disorder increases, the slope of the relationship between lexical predictability and context length will decrease. By quantitatively examining this relationship, I aim to contribute to our understanding of mental disorders and shed light on how incoherent speech affects individuals' communication.

In the following sections, I will present the results first, and then outline the methodology employed in this study, and lastly discuss the implications and significance of the findings.

The code in this report can be found in the Github

Results

The lexical predictability varies with both context length and disorder level. Figure 1 gives an overview of their relationship, which indicates that lexical predictability increased with longer context lengths but decreased as the disorder level increased. The specific effects of context length and disorder level on lexical predictability will be detailed in the following.

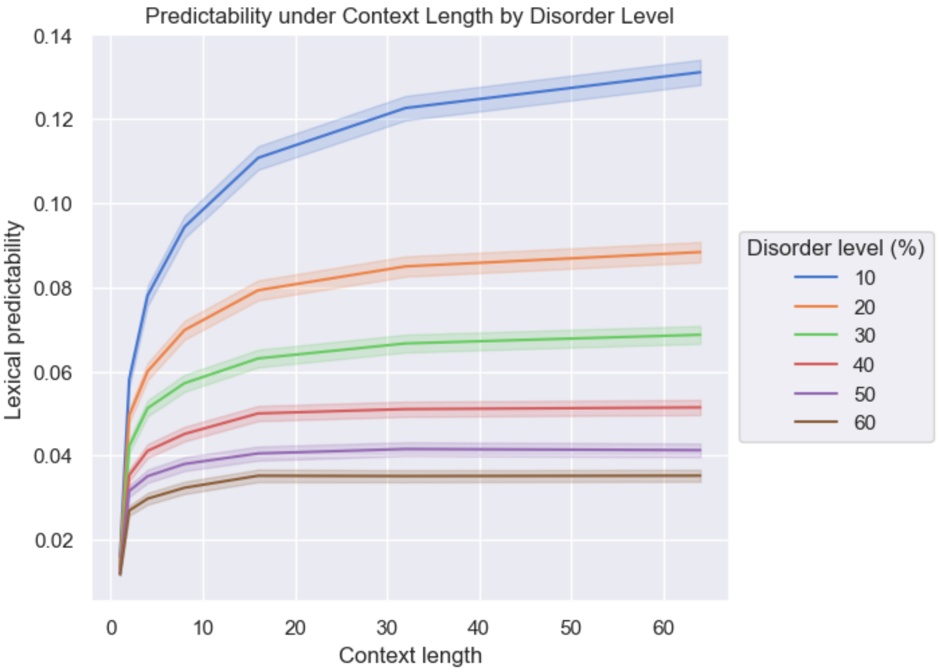


Figure 1: The relationship between the lexical predictability and context length, for different disorder levels.

Effects of context length

The effect of context length on lexical predictability is presented in Figure 2. The relationship between the mean of lexical predictability and context length for different disorder levels is fitted using logarithmic and exponential functions. The fitted curves alongside the measured data provide a direct assessment of the model's performance. Both models demonstrated a high degree of fit, with R^2 values approaching 1. The results indicated that lexical predictability increases significantly with context length. Specifically, when the context is relatively short, lexical predictability rapidly increases with increasing context length. However, as the context becomes longer, the rate of increase in lexical predictability slows down.

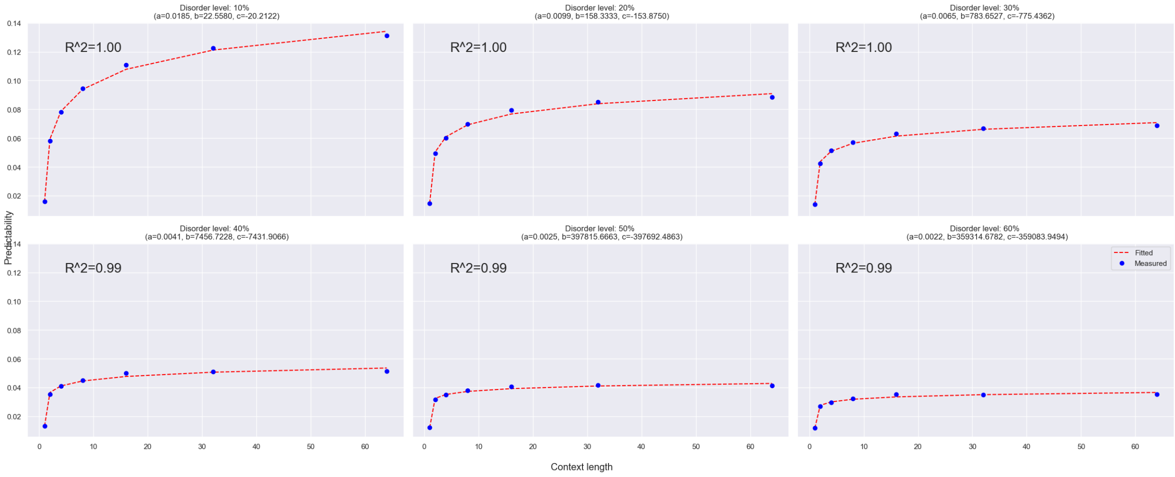


Figure 2a: Lexical Predictability fit as a function of context length with a logarithmic function (a*log(b*x+c)), for different disorder levels.

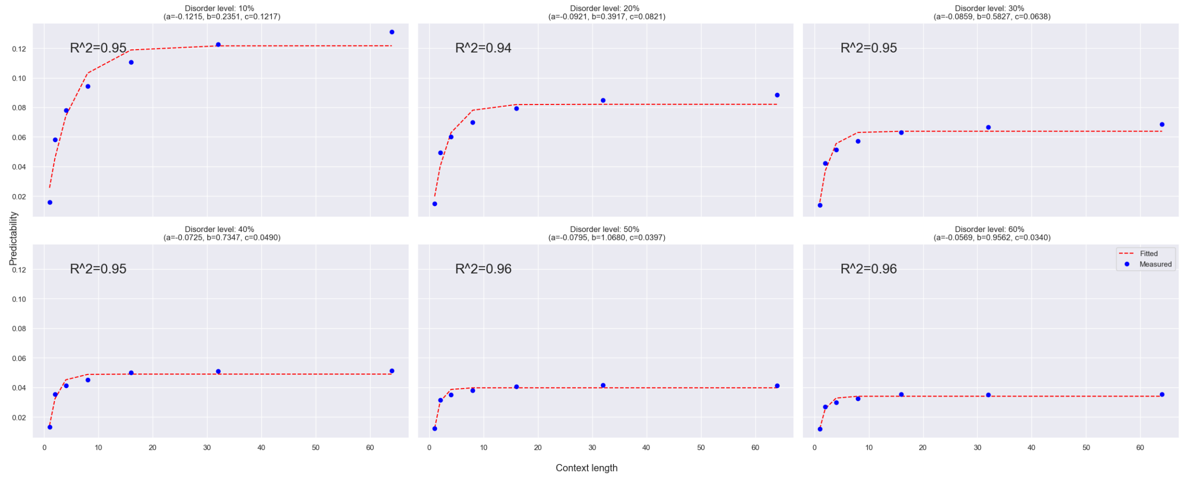
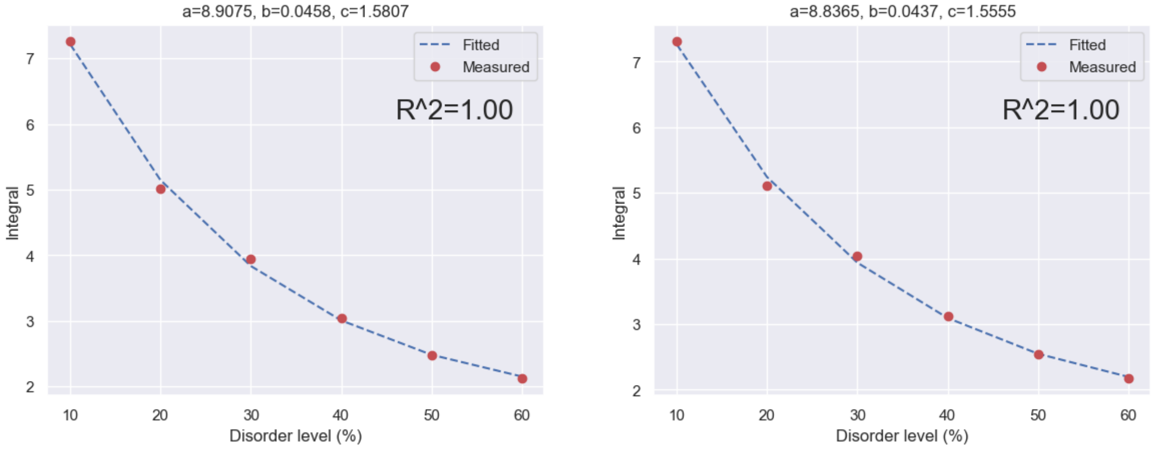


Figure 2b: Lexical Predictability fit as a function of context length with an exponential function (a*exp(-b*x)+c), for different disorder levels.

Effects of disorder levels

This part focuses on investigating how the disorder level influences the relationship between lexical predictability and context length. Figure 2 clearly illustrates that lexical predictability increases as context length increases, regardless of the disorder level. However, the rate of increase varies across different levels of disorder. The growth rate is measured using the integral of fitted exponential and logarithmic models, respectively. Figures 3a and 3b depict the relationship between the disorder level and the calculated integral, revealing a logarithmic trend in the fitted curves with a remarkable R^2 value of 1. As the disorder level increases, the growth rate of lexical predictability decreases. Notably, the rate of decrease is initially rapid and then gradually slows down for higher levels of disorder. The results provide strong evidence supporting the hypothesized relationship.



a

b

Figure 3: The integral as a function of disorder level, for logarithmic (a) and exponential (b) functions.

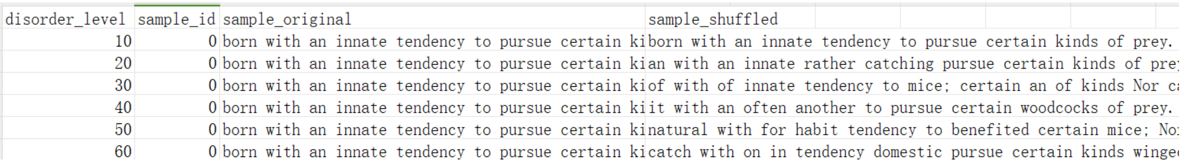
Method

Generate shuffled samples

To simulate varying levels of incoherent speech, I generated text samples with an increasing proportion of words randomly shuffled, ranging from 10% to 60%. For each disorder level, 100 samples were generated.

- Download the book "On the Origin of Species" utilizing the `requests` library from the URL `https://www.gutenberg.org/files/1228/1228-h/1228-h.htm`.
- Parse the downloaded HTML content using the `BeautifulSoup` library and save it to a .txt file.
- Preprocess the data with regular expressions (`re`) and get the chapter list.
- Randomly select one chapter from the list and exact 100 samples from the selected chapter with `Random`, each consisting of 100 tokens. Since the purpose was to shuffle around the word order and later calculate lexical predictability based on a fixed context length rather than complete language fragments, the sample can be extracted from any position within the chapters.
- Shuffle around the word order: for each sample, randomly select a proportion of word tokens corresponding to the disorder levels ranging from 10% to 60% and place them at random positions within them. At last, 600 shuffled samples were generated.

Here are examples of shuffled samples at each disorder level for the same text:



disorder_level	sample_id	sample_original	sample_shuffled
10	0	born with an innate tendency to pursue certain ki	born with an innate tendency to pursue certain kinds of prey.
20	0	born with an innate tendency to pursue certain ki	an with an innate rather catching pursue certain kinds of pre
30	0	born with an innate tendency to pursue certain ki	of with of innate tendency to mice; certain an of kinds Nor c
40	0	born with an innate tendency to pursue certain ki	it with an often another to pursue certain woodcocks of prey.
50	0	born with an innate tendency to pursue certain ki	natural with for habit tendency to benefited certain mice; No
60	0	born with an innate tendency to pursue certain ki	catch with on in tendency domestic pursue certain kinds winge

Compute lexical predictability

To examine the impact of context length and disorder level on lexical predictability, I conducted lexical predictability computations on the dataset of 600 shuffled samples using the GPT2 model in the `Transformer` library. Due to the computational intensity of this task, I utilized GPU resources available on the Kaggle platform.

Initially, I computed the lexical predictability for a single sample, gradually increasing the context length from the previous word up to the maximum available context. This process took approximately 2 minutes per sample. Considering that calculating word predictability for all available contexts across the 600 samples would take over 1000 minutes, I decided to compute predictability for fixed context lengths.

After plotting the computed lexical predictability for one sample, the results revealed a logarithmic growth pattern characterized by rapid initial growth followed by a gradual slowdown. Consequently, I selected 7 context lengths to ensure logarithmically uniform distribution: 1, 2, 4, 8, 16, 32, and 64.

To ensure an equal number of word predictabilities under different context lengths, I computed the lexical predictability starting from the 65th word. This allowed each word to be evaluated with all 7 context lengths. For each shuffled sample, I encoded the entire text using the GPT2Tokenizer in the `Transformer` framework and extracted the encoded context, which served as input to generate predictions for the next word. Subsequently, I extracted the predicted logits from the output and computed probabilities using Softmax.

Analyze the effects of the disorder level and context length

To explore the relationship between lexical predictability and context length, I conducted curve fitting using a logarithmic function (a*log(b*x+c)) and an exponential function (a*exp(-b*x)+c) to capture the pattern of predictability means with varying context lengths.

Next, to investigate whether the growth rate of word predictability decreases as the disorder level increases, I calculated the integral of the fitted models mentioned above to obtain the growth rate of predictability. Subsequently, I performed further curve fitting with an exponential function to examine the relationship between the growth rate and the disorder level.

Conclusion

In this study, I simulated varying levels of incoherent speech and computed word-by-word predictability using the GPT2 model with different context lengths. I then analyzed the relationship between lexical predictability, context length, and disorder level. The results provide insights into the impact of disorder level on the predictability of words and support the initial hypothesis.

The findings demonstrate that as the amount of context increases, word predictability also increases. However, the rate of increase in predictability slows down as the disorder level increases. This suggests that when speech becomes more incoherent, more context is needed to accurately predict the next word.

Overall, this study enhances our knowledge of the relationship between context length, lexical predictability, and disorder level. It underscores the importance of considering disorder level when studying language processing and provides valuable insights into the challenges faced by individuals with disordered speech. Further research in this area can contribute to the development of effective interventions and strategies to improve communication and language processing for individuals with mental disorders.