Project 1: Surprisal and RTs

Gony Idan
315817601

Hila Malka
313312753

Abstruct

In our project we will be working on project 1: surprisal and reading time, examining the relation between reading times and surprisal using different models and techniques. We will analyze the performance of RNN model, NGRAM model, and performance of neural networks models on different data-sets. Our main task will examine the effect of texts from different time periods on the model's ability to predict surprisal on modern text and it's relation to it's reading time.

Surprisal refers to the level of unexpectedness or unpredictability of a word or phrase within a given context. It plays a crucial role in measuring human reading comprehension and cognitive load. Reading time, on the other hand, reflects the duration required for an individual to process and comprehend a given text segment. Though out our tasks we will be examining the Smith and Levy (2013) self-paced reading data-set derives from each subject in the experiment reading a number of several-hundred-word passages selected from the Brown corpus (Kucera and Francis, 1967).

Link to Git repository of the project

Structured Task

In the structured task we examine and compared the results of the RNN model to the original N-Gram model from HW2.

In our work we trained a RNN model on the Penn Treebank data-set, a commonly used data-set in natural language processing research, in order to predict the surprisal of a given text. We examined the results of the trained model on the Brown corpus, matching each word with its corresponding average reading time.

RNN models excel at text tasks because they possess the ability to retain information from previous words in a sentence, enabling them to capture contextual dependencies and produce more coherent and accurate predictions. Given that, because surprisal is based on the context of the sentence, the RNN model has better understanding of the context than the N-Gram model. The N-Gram model only views the last N words for it's prediction, 5 words in our case, and so the context used in order to predict the surprisal is limited.

Also the RNN model can be better at learning more complex relations, and even syntactic relations, again because of its ability to retain information from previous parts of the sentence, unlike the N-Gram model which is a much simpler model that does not excel in such relations.

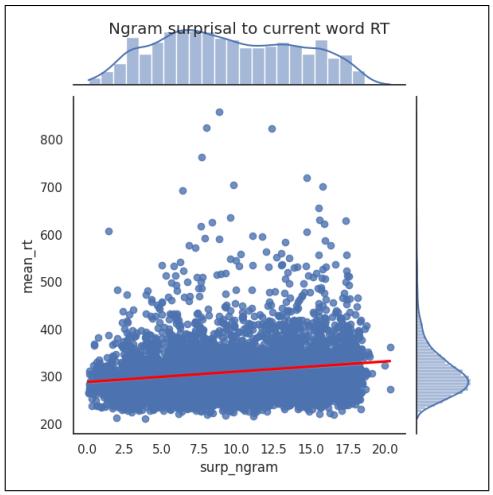
Results We observed the both models have a positive association between reading time and surprisal but both not very strong Fig. 1. The RNN received a slightly better correlation, 0.1881, than of the N-Gram, 0.1814. Also we did view that for high outliers the RNN model gave higher surprisal values than the N-Gram model, which led to a difference of 6 to 8 milliseconds between the two models for the high surprisal values, Fig. 2.

We examined a few sentences that includes a word that received a very high difference between the models prediction of surprisals, the outliered words marked in bold and the unkown words indicates words that the model did not learn since it didn't appear in it's vocabulary.

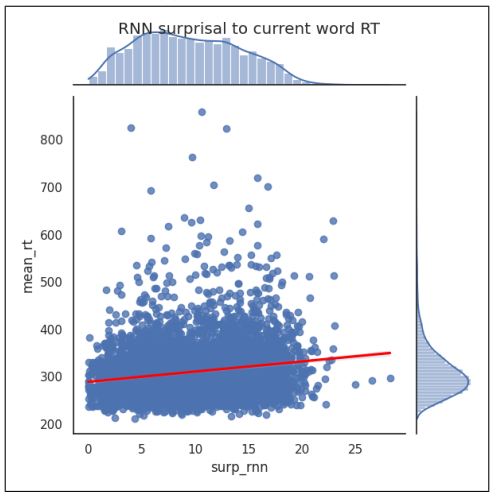
- Down in (unknown) New (unknown) was a flier in the right place at the right (unknown) Robert S. (unknown) a native New (unknown) had been a World War **I** flying (unknown) (unknown) and one of the original planners of the Concord (unknown)
- (unknown) been out with Pete **the** night before and her gay (unknown) about their date (unknown) my mood a (unknown)
- A ripple ran through the muscles of his (unknown) but he kept control upon his (unknown) (unknown) must be some water under (unknown) **He** tilted his (unknown) face toward the dry bed of the (unknown) (unknown) can get it if we (unknown) he said (unknown) (unknown) add fever to our (unknown) She (unknown) (unknown) do you want to see if I can stand (unknown) (unknown) (unknown) can (unknown) (unknown) he (unknown)

In all the sentences the words that received high different values between the surprisal predictions of the models were standards and common words. This indicates that the difference probably lies in reference to the context of the sentence and not in the level of surprisal of the word itself. It can be assumed that the difference lies in the length of the context taken into account in both models, as stated before.

Finally we examined the spillover effect and for both models we received similar results – that as the probability grows the reading time of the next word decreases slightly and the reading time of the current word decreases more significantly



(a) N-Gram model



(b) RNN model

Figure 1: Relation between reading time and surprisal calculated by the models.

Fig. 3. That is as the word is less common, has a higher surprisal value, the reading time of the next word increases.

Semi-Structured Task

Fit and plot the RT surprisal curve usinga General Additive Model (GAM)

In this part we have worked with GAM models. GAM models come to improve linear models, by understanding that the effect of each feature on the model may not always be linear. And so GAM model learn nonlinear weights for the features of the data - surprisal, word length and log of the frequency of the word.

Results In general we can view a non linear monotonically increasing relation between reading time and surprisal Fig. 4a. We did receive non linear relation between the features and the reading time, Fig. 5, as the model aims to achieve. Surprisingly, when the surprisal was high or the

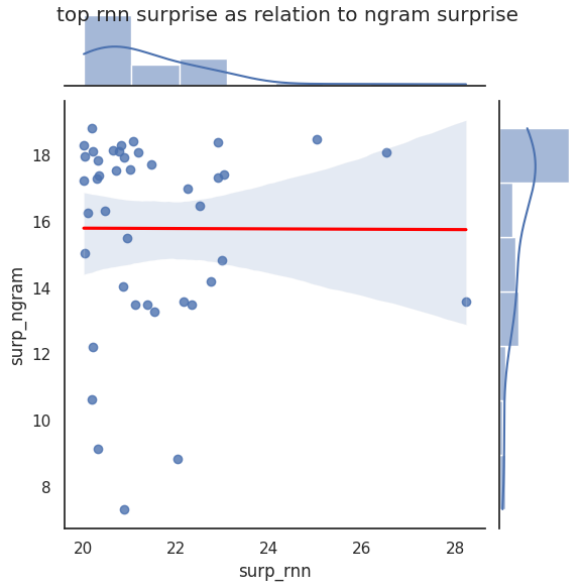


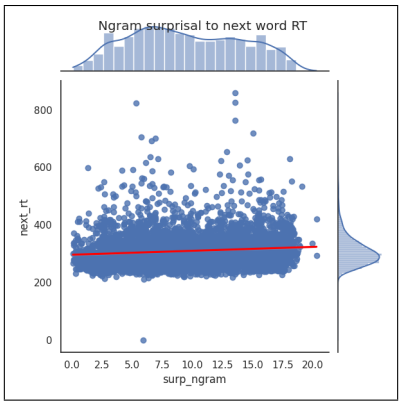
Figure 2: Relation between outliers of surprisal calculated by N-Gram model to RNN model

length of the word was long, the relation to the reading time decreased. This may result due to the fact that there are not many words with a large surprsial or long length which challenges the models ability to correctly match with the reading time. Also for this model it can be seen that the effect of spillover is significantly smaller than the effect of surprise on the reading time of the current word Fig. 4b.

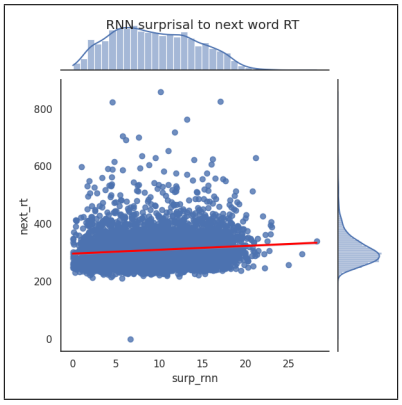
Train N-Gram and RNN Models on a Larger Dataset

We have trained N-Gram with $N = 5$ and RNN model on a larger data-set, Wikitext-2, and examine the effect on the relations between the surprisal and reading time. When examining the model's prediction of reading time for the Brown corpus we found that there were less unknown words. This makes sense since the unknown words are based on the vocabulary from the data that the model trained on, and it is no surprise that the Wikitext-2 data-set has a larger vocabulary than Penn Treebank data-set. And so in order to harmonized the new data and to be able to match it to the reading time we will match between the results of the previous model and its unknown words to the results from the new model.

Results We examined the two trained models, refered as Wiki-RNN and Wiki-NGRAM, and evaluated the surprisal predicted by them. We will examine the effect the different data-set has on models with the same architecture and also the effect the large data-set has on different models. We received a reasonable correlation of 0.8891 between Wiki-RNN and PTB-RNN, RNN trained on Penn Treebank, and a correlation of 0.744 between Wiki-NGRAM and PTB-NGRAM. We also received a lower correlation between Wiki-RNN and Wiki-NGRAM, 0.7499. These results are acceptable, as stated

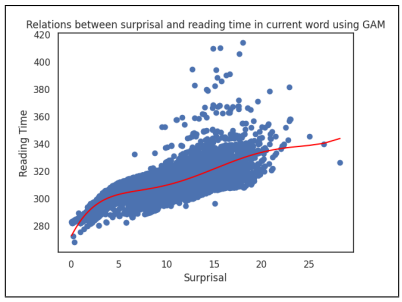


(a) N-Gram model

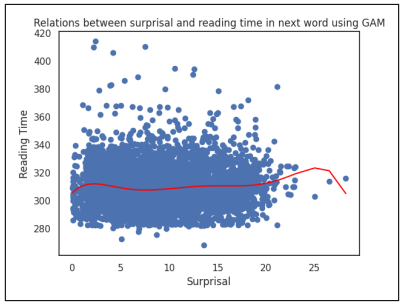


(b) RNN model

Figure 3: Relation between reading time of next word and surprisal of current word calculated by the models trained on Ptb-data.



(a) Relation calculated with GAM model between reading time and surprisal of current words



(b) Relation calculated with GAM model between reading time of next words and surprisal of current words

Figure 4: Relation between reading time of next word and surprisal of current word calculated by GAM model trained on Ptb-data.

above, due to the difference between the models and their use of previous parts of the sentence for context.

When examining the relation between surprisal and reading time we received a correlation of 0.1761 from Wiki-RNN and 0.0995 for Wiki-NGRAM. These results surprised us, we expected that since we trained the models on a significantly larger data-set the correlation would be larger than before. We assume that this occurs since the style of writing and the topics that appear in Wikipedia do not express well the surprise of continuous reading on other topics, and therefore do not express the reading time in a better way than the other model. For the N-Gram, we suspected that we received such small relation because the complexity of the Wiki-text data-set can not be fully captured in a simple 5-Gram model as only referring to the previous 5 words does not fully capture the complexity of the text, especially regarding theoretical text that is characterized with longer sentences.

We also examined the spill-over effect and recieved similar results as before. As the surprisal of the current word increases the reading time of the next word increases, but the effect is smaller than the effect of the current word reading time. In the graph we can see this in a smaller slope compared to the slope in the graph that expresses the relation between the surprise and the current word reading time.Fig. 6.

Open Task - Investigating the Impact of Training Models on Books from Different Eras: Predicting Surprisal and Reading Time

Introduction

In this project, we approach a different view of NLP models by training models on books from different eras and examining the influences it has on their predictive abilities, specifically in terms of surprisal and reading time. By training our language model on books from different historical periods, we aim to investigate how exposure to varied linguistic styles, vocabularies, and cultural contexts affects the model's ability to predict surprisal and estimate reading time accurately. With this exploration we hope to learn about the adaptability of the model and its potential to capture the nuances and intricacies of different eras in literature.

We will create a diverse corpus comprising literary works from various periods, ranging from ancient classics to contemporary bestsellers. By incorporating texts from different eras, we hope to capture the evolution of language and writing styles, reflecting the cultural, social, and linguistic changes that have occurred over time.

Our work will involve training a RNN model on this comprehensive dataset and evaluating its performance using established metrics for surprisal and reading time prediction.

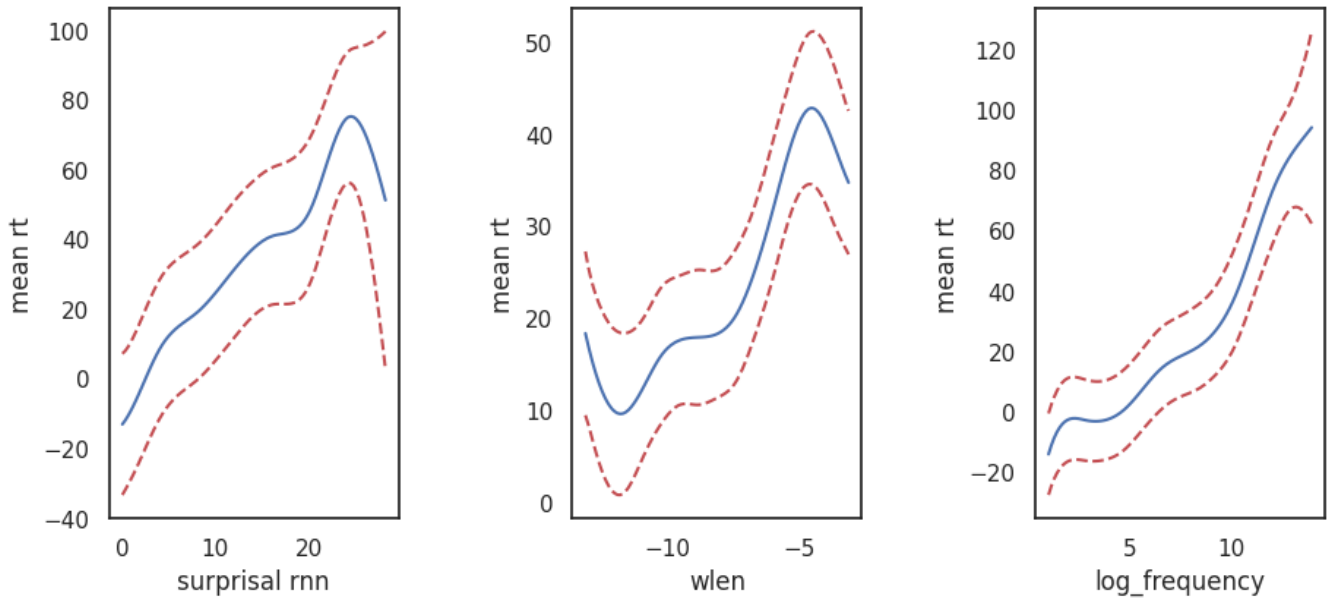


Figure 5: Non linear relations from GAM model on features.

By comparing the model's predictions against human reading experiments, as done before, we will be able to assess the models ability to capture the true surprisal of words based on contexts from different time periods.

Understanding how language models process and interpret texts from different eras can inform the development of more sophisticated algorithms and improve their generalization capabilities.

Books Used

All basic information of books used can be found in Table 1. For each of the following books we trained a different RNN models. By doing so each of the models used the context given to it, the sentences of the book, in order to learn the surprisal of the different words.

Crime and Punishment by Fyodor Dostoevsky (19th century - Realist era) Crime and Punishment a russian novel, following the story of Rodion Raskolnikov, a destitute former student who commits a heinous crime and grapples with the psychological turmoil of guilt and punishment, exploring themes of morality and the complexities of the human psyche in 1866 St. Petersburg.

1984 by George Orwell (20th century - Modernist era) A dystopian novel published in 1949, set in a totalitarian society ruled by Big Brother. It depicts a future where individualism and freedom are suppressed, and the protagonist, Winston Smith, rebels against the oppressive regime.

Harry Potter and the Order of the Phoenix by J.K. Rowling (20th century - Modernist era) The fifth book in the immensely popular Harry Potter series written by J.K. Rowling. Published in 2003, the story continues to follow Harry Potter, a young wizard attending Hogwarts School of Witchcraft and Wizardry.

Results

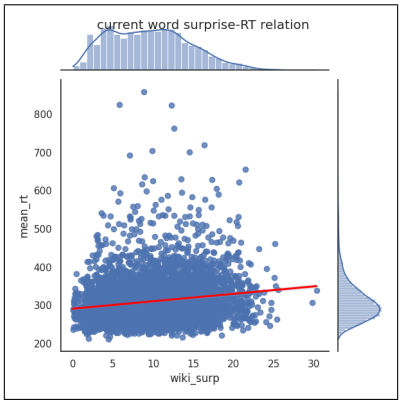
Relation Between Surprisal and Reading Time Given the three models trained for each of the three books we calculated the surprisal each model predicts on the Brown corpus in order to compare to the reading time achieved in the experiment. We achieved interesting results of correlation to the reading time - Crime and Punishment with correlation of 0.1689, 1984 with correlation of 0.0037 and Harry Potter and the Order of the Phoenix with correlation of 0.2018. These results mostly match our expectations.

The fact that 1984 recieved such a low correlation is with a straight relation to its size, only 88,000 words, compared to the Penn Tree bank data-set that has 5,124,439 words. Also when comparing between the models, the one trained on the Harry Potter book received the highest correlation. This matches our expectations since the Harry Potter book is closest in time to the experiement measuring the reading time, at 2013, and so the book represented more closely the modern day surprisal of words.

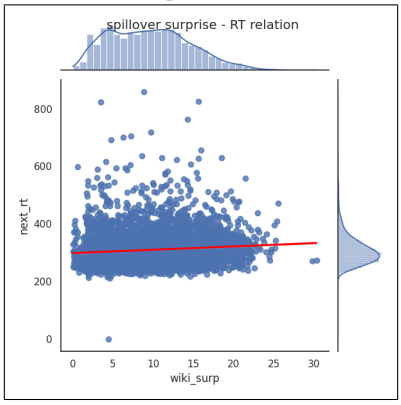
Comparison between books from different time eras In this next section we wanted to examine how a model that trained on a book from a certain time era predicts the surprisal from a book of different time era. For each trained model we computed the surprisal on the two other books and compared it to the surprisal computed for the book by the model trained on the matching book, ground truth, Table 2. In the results we can see that books from closer time eras, 1984 and Harry Potter No.5, received better correlation on one another than Crime and Punishment. Also it is intrest-

Book	Author	Time Published	Number of words
Crime and Punishment	Fyodor Dostoevsky	1866	210,000
1984	George Orwell	1949	88,000
Harry Potter No. 5	J.K. Rowling	2003	257,000

Table 1: Information per book.



(a) Relation between surprisal of current words and reading time trained on Wikipedia

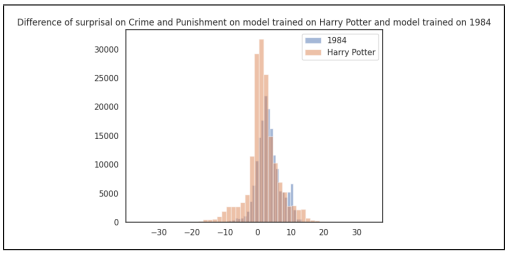


(b) Relation between surprisal of current words reading time of next word trained on Wikipedia

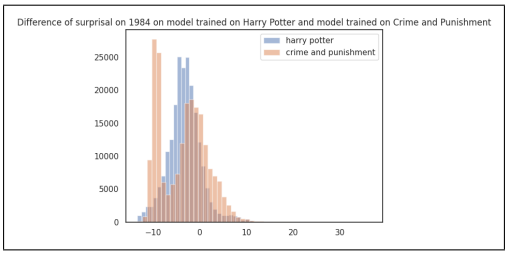
Figure 6: Relation between surprisal of current word and reading time of current and next word calculated by RNN model trained on Wikipedia.

ing to view that the model trained on Crime and Punishment recieved significantly worst results.

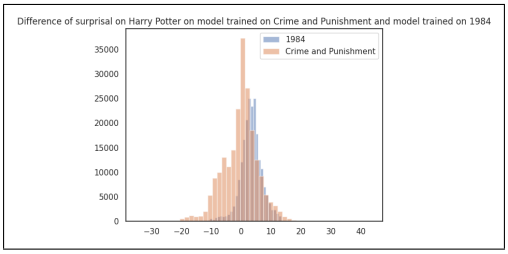
In order to further explore these results we examined the difference between the ground truth of the surprisal, the surprisal given for a book by a model trained on the same book, and the surprisal computed by models trained on different books, the results can be viewed Fig. 7. When looking at Fig. 7c we can see that the model trained on Crime and Punishment gave lower surprisal values than the ground truth for the Harry Potter book, can be viewed with a wider range on negative values in the graph. This can also be viewed when examining Fig. 7b for the model trained on Crime and Punishment again gave lower surprisal values than the ground trough for the 1984 book .A possible explanation for this phenomenon can be the complexity of the text in the 19th century - Realist era compared to the other eras we view. This can lead that the model trained on Crime and Punishment was more exposed to complex sentences and so when encountering sentences from more modern eras, that are less complex, would predict a lower surprisal.



(a) Crime and-Punishment



(b) 1984



(c) Harry Potter

Figure 7: Difference between surprisal from model trained on different book to surprisal of the ground truth model trained on the examined book

Summary

In our work we explored the abilities of different models to predict the surprisals of words and view its correlation to real reading time measured from experiments. We viewed two main models, RNN and N-Gram with $N = 5$. Over all we received that the RNN model was better at computing and predicting the surprisal that matches the reading time. As RNN

Test Correlation On	Model trained on		
	Crime and Punishment	**1984**	**Harry Potter No. 5**
Crime and Punishment	1	0.5332	0.4020
1984	0.2188	1	0.5872
Harry Potter No. 5	0.1978	0.5872	1

Table 2: Correlation on different books.

are known for being able to capture the whole context of the sentence which is very important for predicting the surprisal of the word, unlike the N-Gram model which was only exposed to the previous 5 words. We hoped to view that the size of the data the models were trained on would improve the relation between surprisal and reading time for both models, but when training the models on Wikitext-2 we did not receive better results. This led us to the understanding that it is not only the size of the dataset that matters but it's context as well. Wikitext-2 is more of a theoretical and scientific dataset, based on Wikipedia pages whose purpose are to teach and transfer knowledge, which is different that the style of our test data, Brown corpus, from which the reading time was extracted, which consists of a diverse range of texts from various genres, including fiction, non-fiction, news articles, and conversations.

We also explored the way texts from different eras are able to predict surprisal to match reading time, by training RNN models of books from different eras. The model trained on modern book received the best correlation to the reading time. The Brown Corpus was created in 1967, so it may not capture more recent linguistic phenomena or changes in the English language, So our results were a little surprising. We hoped to view that the book written closest to the time of Brown would receive better correlation to reading time though it ended up achieving the worst correlation. One explanation is the size of the book, which is significantly smaller than the other books tested. Also, we saw in the cross comparison that out two most modern books performed well on each other, so probably the evolving of the literature language wasn't too dramatic in the time that pasted between the publication of these books.

We also viewed the spill-over effect, how the surprisal of the current word effects the reading time of the next word. In all of our different models and different datasets we saw a stronger effect of the surprisal on the reading time of the current word than the next. When taking in to consideration the experiment that was conducted in order to measure the reading time, where the subjects needed to click to the keyboard in order to move to the next word, perhaps it does not evaluate the reading time to it's true nature. The subject would finish analyzing the word before moving on to the next, that will cause limiting the effect of the surprisal to the current word read and less to the next word.

In general, in all of our models and experiments we did not view a strong correlation between the predicted surprsial and the reading time. Further research and exploration around these relations is needed. Evaluating more elements that effect the reading time, usage of part-of-speech tags and syntactic tree structures or multi modal data such as eye movements to improve measuring reading time. Also experimenting the use of more advance models, LLM's, that are better on capturing the full context of the text which has a high effect on the surprisal and reading time.