Team Ohio State at CMCL 2021 Shared Task: Fine-Tuned RoBERTa for Eye-Tracking Data Prediction

Byung-Doh Oh
Department of Linguistics
The Ohio State University
oh.531@osu.edu

Abstract

This paper describes Team Ohio State's approach to the CMCL 2021 Shared Task, the goal of which is to predict five eye-tracking features from naturalistic self-paced reading corpora. For this task, we fine-tune a pre-trained neural language model (RoBERTa; Liu et al., 2019) to predict each feature based on the contextualized representations. Moreover, motivated by previous eye-tracking studies, we include word length in characters and proportion of sentence processed as two additional input features. Our best model strongly outperforms the baseline and is also competitive with other systems submitted to the shared task. An ablation study shows that the word length feature contributes to making more accurate predictions, indicating the usefulness of features that are specific to the eye-tracking paradigm.

1 Introduction

Behavioral responses such as eye-tracking data provide valuable insight into the latent mechanism behind real-time language processing. Based on the well-established observation that behavioral responses reflect processing difficulty, cognitive modeling research has sought to accurately predict these responses using theoretically motivated variables (e.g. *surprisal*; Hale, 2001; Levy, 2008). Earlier work in this line of research has introduced incremental parsers for deriving psycholinguistically-motivated variables (e.g. Roark et al., 2009; van Schijndel et al., 2013), while more recent work has focused on evaluating the capability of neural language models to predict behavioral responses (Hao et al., 2020; Wilcox et al., 2020).

The CMCL 2021 Shared Task on eye-tracking data prediction (Hollenstein et al., 2021) provides an appropriate setting to compare the predictive power of different approaches using a standardized dataset. According to the task definition, the goal of the shared task is to predict five eye-tracking features from naturalistic self-paced reading corpora, namely the Zurich Cognitive Language Processing Corpus 1.0 and 2.0 (ZuCo 1.0 and 2.0; Hollenstein et al., 2018, 2020). These corpora contain eye-tracking data from native speakers of English that read select sentences from the Stanford Sentiment Treebank (Socher et al., 2013) and the Wikipedia relation extraction corpus (Culotta et al., 2006). The five eye-tracking features to be predicted for each word, which have been normalized to a range between 0 and 100 and then averaged over participants, are as follows:

- Number of fixations (nFix): Total number of fixations on the current word
- First fixation duration (FFD): The duration of the first fixation on the prevailing word
- Total reading time (TRT): The sum of all fixation durations on the current word
- Go-past time (GPT): The sum of all fixations before progressing to the right of the current word
- Fixation proportion (fixProp): The proportion of participants that fixated on the current word

In this paper, we present Team Ohio State's approach to the task of eye-tracking data prediction. As the main input feature available from the dataset is the words in each sentence, we adopt a transfer learning approach by fine-tuning a pre-trained neural language model to this task. Furthermore, we introduce two additional input features motivated by previous eye-tracking studies, which measure word length in characters and the proportion of sentence processed. Our best-performing model outperforms the mean baseline by a large margin in terms of mean absolute error (MAE) and is also competitive with other systems submitted to the shared task.

Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, pages 97–101

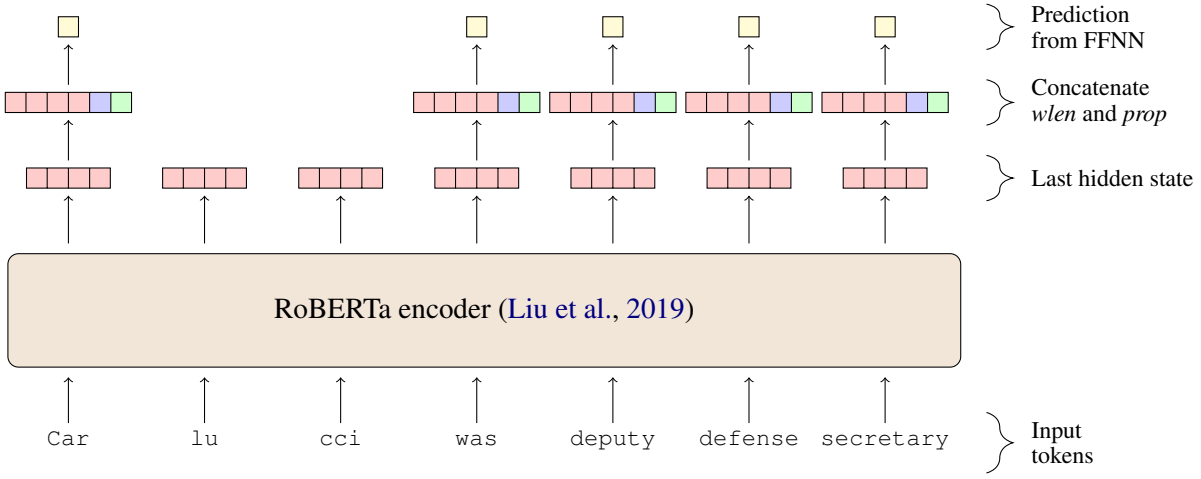

Figure 1: Model architecture for eye-tracking feature prediction.

2 Model Description

Our model relies primarily on the Transformer-based pre-trained language model RoBERTa (Liu et al., 2019) for contextualized representations of each word in the input sentence.[1] However, since RoBERTa uses byte-pair encoding (Sennrich et al., 2016) to tokenize each sentence, there is a mismatch between the number of output representations from RoBERTa and the number of words in each sentence. In order to address this issue, the model uses the representation for the first token associated with each word to make predictions. For example, if byte-pair encoding tokenizes the word `Carlucci` into `Car`, `lu`, and `cci`, the representation for `Car` is used to make predictions for the entire word `Carlucci`.[2]

Additionally, two input features based on previous eye-tracking studies are included in the model. The first is word length measured in characters (*wlen*), which captures the tendency of readers to fixate longer on orthographically longer words. The second feature is proportion of sentence processed (*prop*), which is calculated by dividing the current index of the word by the number of total words in each sentence. This feature is intended to take into account any "edge effects" that may be observed at the beginning and the end of each sentence, as well as any systematic change in eye movement as a function of the word's location within each sentence. These two features, which are typically treated as nuisance variables that are experimentally or statistically controlled for in eye-tracking studies (e.g. Hao et al., 2020; Rayner et al., 2011; Shain, 2019), are included in the current model to maximize prediction accuracy.[3]

A feedforward neural network (FFNN) with one hidden layer subsequently takes these three features (i.e. RoBERTa representation, *wlen*, and *prop*) as input and predicts a scalar value. To predict the five eye-tracking features defined by the shared task, this identical model was trained separately for each eye-tracking feature. An overview of the model architecture is presented in Figure 1.[4]

3 Training Procedures

3.1 Data Partitioning

Following the shared task guidelines, 800 sentences and their associated eye-tracking features from the ZuCo 1.0 and 2.0 corpora (Hollenstein et al., 2018, 2020) provided the data for training the model. However, a concern with using all 800 sentences to fine-tune the RoBERTa language model as described above is the tendency of high-capacity lan-

[1] Although other word representations could be used within our model architecture, the use of RoBERTa was motivated by its state-of-the-art performance on many NLP tasks. The $\text{RoBERTa}_{\text{base}}$ and $\text{RoBERTa}_{\text{large}}$ variants were explored in this work, which resulted in two different models. We used the implementation made available by HuggingFace (`https://github.com/huggingface/transformers`).

[2] Future work could investigate the use of more sophisticated approaches, such as using the average of all token representations associated with the word.

[3] Other variables typically examined in eye-tracking studies include frequency-based measures (e.g. token frequency) and prediction-based measures (e.g. various instantiations of surprisal). However, those variables were not included in our models as input features, as it was thought that the high-capacity RoBERTa model trained on a masked language modeling objective would implicitly encode such information.

[4] Code for model training and evaluation is available at `https://github.com/byungdoh/cmcl21_st`.

Model	Dev (MSE)					Test (MAE)				
	nFix	FFD	GPT	TRT	fixProp	nFix	FFD	GPT	TRT	fixProp
RoBERTa$_{\text{base}}$	28.307	**0.757**	14.780	**4.234**	**198.917**	**3.987**	0.682	**2.364**	**1.540**	11.311
RoBERTa$_{\text{large}}$	**28.023**	0.762	**14.669**	4.502	200.352	4.079	**0.668**	2.407	1.544	**11.210**
Mean baseline	91.783	2.062	35.509	13.838	662.309	7.303	1.149	3.782	2.778	21.775

Table 1: MSE on the held-out dev set and MAE on the test set for the two models.

Model	Test (MAE)				
	nFix	FFD	GPT	TRT	fixProp
Full model	3.987	0.682	2.364	1.540	11.311
-prop	3.987	0.681	2.364	1.540	11.315
-wlen	3.997	0.681	2.376	1.543	11.424
-prop,wlen	3.998	0.681	2.377	1.543	11.431

Table 2: MAE on the test set for full RoBERTa$_{\text{base}}$ model and its ablated variants.

guage models to agressively overfit to the training data (Howard and Ruder, 2018; Jiang et al., 2020; Peters et al., 2019). To prevent such overfitting, the last 80 sentences (10%; 1,546 words) were excluded from training as the dev set and were used to conduct held-out evaluation. This partitioning resulted in the final training set, which consists of 720 sentences (90%; 14,190 words).

3.2 Implementation Details

For each eye-tracking feature, the two models were trained to minimize mean squared error (MSE, Equation 1),

$$\mathcal{L}(\theta) = \frac{1}{N}\sum_{i=1}^{N}(y_i - f(\mathbf{x}_i;\theta))^2 \qquad (1)$$

where $f(\cdot;\theta)$ is the model described in Section 2, $\mathbf{x}_i$ is the concatenation of three input features, y_i is the target value associated with the eye-tracking feature, and N is the number of training examples in each batch. The AdamW algorithm (Loshchilov and Hutter, 2019) with a weight decay hyperparameter of 0.01 was used to optimize the model parameters. The learning rate was warmed-up over the first 10% of training steps and was subsequently decayed linearly. The number of nodes in the hidden layer of the FFNN was fixed to half of that of the input layer. Additionally, dropout with a rate of 0.1 was applied before both the input layer and the hidden layer of the FFNN. Finally, to avoid exploding gradients, gradients with a norm greater than 1 were clipped to norm 1.

The optimal hyperparameters were found using grid search based on MSE on the held-out dev set. More specifically, the learning rate was explored within the set of $\{1 \times 10^{-5}, 2 \times 10^{-5}, 3 \times 10^{-5}, 5 \times 10^{-5}\}$, batch size was explored within the set of $\{4, 8, 16, 32, 64\}$ sentences, and the maximum number of training epochs was explored within the set of $\{8, 16, 32, 64, 128, 192\}$. During training, the model was evaluated on the dev set after every training epoch.

4 Results and Discussion

Table 1 shows the MSE on the dev set and MAE[5] on the test set for the two models. Both models strongly outperformed the baseline approach that predicts the mean value of the training set, resulting in a ∼40% decrease in MAE for all five features. Additionally, although the difference is small, the RoBERTa$_{\text{base}}$ model tended to perform better than the RoBERTa$_{\text{large}}$ model on the test set.[6] This suggests that models with higher capacity may not necessarily be preferable for this task, especially in light of the small amount of training data available.

To evaluate the contribution of the *wlen* and *prop* features, an ablation study was conducted using the RoBERTa$_{\text{base}}$ model. In addition to showing how useful *wlen* and *prop* information is for predicting eye-tracking features, the analysis was also thought to reveal whether or not such information is already contained within the RoBERTa representations. The two input features were ablated by simply replacing them with zeros during inference, which allowed a clean manipulation of their contribution to the final predictions.

The results in Table 2 show that the ablation of the *prop* feature made virtually no difference in the model predictions. This is most likely due to the fact that the Transformer (Vaswani et al., 2017), which the RoBERTa models are based on, includes positional encodings that allow the model to be sen-

[5]The official evaluation metric, $\frac{1}{N}\sum_{i=1}^{N}|y_i - f(\mathbf{x}_i;\theta)|$.

[6]The RoBERTa$_{\text{base}}$ model ranked 11th out of 29 submissions on the shared task (6th out of 13 participating teams).

sitive to the position of each token in the sequence. Therefore, in order to fully examine the contribution of positional information on this task, a variant of the current model using RoBERTa representations trained without positional encodings would have to be evaluated.

The ablation of the *wlen* feature resulted in a more notable difference in four out of five eye-tracking features. This indicates that information about orthographic length is both useful for eye-tracking data prediction and also orthogonal to the information captured by the RoBERTa representations. This may partially be explained by RoBERTa's use of byte-pair encoding, which can result in many short tokens for a given word (e.g. tokens `Car`, `lu`, `cci` for the word `Carlucci`). Since only the first token was used by the current models to represent each word, explicitly including information about word length seems to have contributed to making more accurate predictions. More generally, this highlights the utility of incorporating features that are specific to eye-tracking, which may not be inherent in high-capacity language models trained for a different objective.

5 Conclusion

In this paper, we present our approach to the CMCL 2021 Shared Task on eye-tracking data prediction. Our models primarily adopt a transfer learning approach by employing a feedforward neural network to predict eye-tracking features based on contextualized representations from a pre-trained language model. Additionally, we include two input features that have been known to influence eye movement, which are word length in characters (*wlen*) and proportion of sentence processed (*prop*). Our best model based on $\text{RoBERTa}_{\text{base}}$ strongly outperforms the mean baseline and is also competitive with other systems submitted to the shared task. A follow-up ablation study shows that the *wlen* feature contributed to making more accurate predictions, which indicates that explicitly incorporating features specific to the eye-tracking paradigm can complement high-capacity language models on this task.

Acknowledgments

We would like to thank the anonymous reviewers for their helpful comments.

References

Aron Culotta, Andrew McCallum, and Jonathan Betz. 2006. Integrating probabilistic extraction models and data mining to discover relations and patterns in text. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 296–303.

John Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies*, pages 1–8.

Yiding Hao, Simon Mendelsohn, Rachel Sterneck, Randi Martinez, and Robert Frank. 2020. Probabilistic predictions of people perusing: Evaluating metrics of language model performance for psycholinguistic modeling. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 75–86.

Nora Hollenstein, Emmanuele Chersoni, Cassandra Jacobs, Yohei Oseki, Laurent Prévot, and Enrico Santus. 2021. CMCL 2021 Shared Task on Eye-Tracking Prediction. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*.

Nora Hollenstein, Jonathan Rotsztejn, Marius Troendle, Andreas Pedroni, Ce Zhang, and Nicolas Langer. 2018. ZuCo, a simultaneous EEG and eye-tracking resource for natural sentence reading. *Scientific Data*, 5.

Nora Hollenstein, Marius Troendle, Ce Zhang, and Nicolas Langer. 2020. ZuCo 2.0: A dataset of physiological recordings during natural reading and annotation. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 138–146.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 328–339.

Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Tuo Zhao. 2020. SMART: Robust and efficient fine-tuning for pre-trained natural language models through principled regularized optimization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2177–2190.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization.

Matthew E. Peters, Sebastian Ruder, and Noah A. Smith. 2019. To tune or not to tune? Adapting pretrained representations to diverse tasks. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 7–14.

Keith Rayner, Timothy J. Slattery, Denis Drieghe, and Simon P. Liversedge. 2011. Eye movements and word skipping during reading: Effects of word length and predictability. *Journal of Experimental Psychology: Human Perception and Performance*, 37(2):514–528.

Brian Roark, Asaf Bachrach, Carlos Cardenas, and Christophe Pallier. 2009. Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 324–333.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1725.

Cory Shain. 2019. A large-scale study of the effects of word frequency and predictability in naturalistic reading. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4086–4094.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642.

Marten van Schijndel, Andy Exley, and William Schuler. 2013. A model of language processing as hierarchic sequential prediction. *Topics in Cognitive Science*, 5(3):522–540.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30.

Ethan Gotlieb Wilcox, Jon Gauthier, Jennifer Hu, Peng Qian, and Roger P. Levy. 2020. On the predictive power of neural language models for human real-time comprehension behavior. In *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society*, pages 1707–1713.