KonTra at CMCL 2021 Shared Task: Predicting Eye Movements by Combining BERT with Surface, Linguistic and Behavioral Information

Qi Yu, Aikaterini-Lida Kalouli, Diego Frassinelli
Department of Linguistics, University of Konstanz
firstname.lastname@uni-konstanz.de

Abstract

This paper describes the submission of the team KonTra to the CMCL 2021 Shared Task on eye-tracking prediction. Our system combines the embeddings extracted from a fine-tuned BERT model with surface, linguistic and behavioral features, resulting in an average mean absolute error of 4.22 across all 5 eye-tracking measures. We show that word length and features representing the expectedness of a word are consistently the strongest predictors across all 5 eye-tracking measures.

1 Introduction

The corpora ZuCo 1.0 and ZuCo 2.0 by Hollenstein et al. (2018, 2019) contain eye-tracking data collected in a series of reading tasks on English materials. For each word of the sentences, five eye-tracking measures are recorded: 1) the number of fixations (*nFix*), 2) the first fixation duration (*FFD*), 3) the go-past time (*GPT*), 4) the total reading time (*TRT*), and 5) the fixation proportion (*fixProp*). Providing a subset of the two corpora, the CMCL 2021 Shared Task (Hollenstein et al., 2021) requires the prediction of these eye-tracking measures based on any relevant feature.

To tackle the task, we conduct a series of experiments using various combinations of BERT embeddings (Devlin et al., 2018) and a rich set of surface, linguistic and behavioral features (SLB features). Our experimental setting enables a comparison of the potential of BERT and the SLB features, and allows for the explainability of the system. The best performance is achieved by the models combining word embeddings extracted from a fine-tuned BERT model with a subset of the SLB features that are the most predictive for each eye-tracking measure. Overall, our model was ranked 8th out of 13 models submitted to the shared task.

Our main contributions are the following: 1) We show that training solely on SLB features provides better results than training solely on word embeddings (both pre-trained and fine-tuned ones). 2) Among the SLB features, we show that word length and linguistic features representing word expectedness consistently show the highest weight in predicting all of the 5 measures.

2 Describing Eye-Tracking Measures

To explore the impact of linguistic and cognitive information on eye-movements in reading tasks, we extract a set of surface, linguistic, behavioral and BERT features, as listed in Table 1.

Surface Features Given the common finding that surface characteristics, particularly the length of a word, influence fixation duration (Juhasz and Rayner, 2003; New et al., 2006), we compute various surface features at word and sentence level (e.g., word and sentence length).

Linguistic Features The linguistic characteristics of the words co-occurring in a sentence have an effect on eye movements (Clifton et al., 2007). Thus, we experiment with features of syntactic and semantic nature. The syntactic features are extracted using the Stanza NLP kit (Qi et al., 2020). For each word, we extract its part-of-speech (POS), its word type (content vs. function word), its dependency relation and its named entity type. According to Godfroid et al. (2018) and Williams and Morris (2004), word familiarity (both local and global) has an effect on the reader's attention, i.e., readers may pay less attention on words that already occurred in previous context. In this study, we treat familiarity as word expectedness and model it using three types of semantic similarity: a) similarity of the current word w_m to the whole sentence (*similarity*$_{w_m,s}$), b) similarity of the current word to its previous word (*similarity*$_{w_m,w_{m-1}}$), and c) similarity of the current word to all of its previous words within the current sentence (*similarity*$_{w_m,w_{1 \ldots m-1}}$). To compute these similarity measures, we use the BERT (base) (De-

Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, pages 120–124
Online Event, June 10, 2021. ©2021 Association for Computational Linguistics

Feature Category	Feature Name
Surface Features	word length, sentence length in tokens, sentence length in characters, word length-sentence length ratio
Linguistic Features	POS, word type, named entity type, dependency relation, surprisal score, frequency score, similarity$_{w_m,s}$, similarity$_{w_m,w_{m-1}}$, similarity$_{w_m,w_{1...m-1}}$
Behavioral Features	age of acquisition, prevalence score, valence score, arousal score, dominance score, concreteness$_{human}$, concreteness$_{auto}$
BERT Features	pre-trained BERT embedding, fine-tuned BERT embedding

Table 1: The complete set of surface, linguistic and behavioral (SLB) features and the BERT features.

vlin et al., 2018) pre-trained model[1] and map each word to its pre-trained embedding of layer 11. We chose this layer because it mostly captures semantic properties, while the last layer has been found to be very close to the actual classification task and thus less suitable for our purpose (Jawahar et al., 2019; Lin et al., 2019). Based on these extracted embeddings, we calculate the cosine similarities. To measure the similarity of the current word to the whole sentence (*similarity*$_{w_m,s}$), we take the CLS token to represent the whole sentence; we also experiment with the average token embeddings as the sentence embedding, but we find that the CLS token performs better. For measuring the similarity of the current word to all of its previous words (*similarity*$_{w_m,w_{1...m-1}}$), we average the embeddings of the previous words and find the cosine similarity between this average embedding and the embedding of the current word.

Furthermore, semantic surprisal, i.e., the negative log-transformed conditional probability of a word given its preceding context, provides a good measure of predictability of words in context and efficiently predicts reading times (Smith and Levy, 2013), N400 amplitude (Zhang et al., 2020) and pupil dilation (Frank and Thompson, 2012). We compute surprisal using a bigram language model trained on the lemmatized version of the first slice (roughly 31-million tokens) of the ENCOW14-AX corpus (Schäfer and Bildhauer, 2012). As an additional measure of word expectedness, we also include frequency scores based on the US subtitle corpus (SUBTLEX-US, Brysbaert and New, 2009).

Behavioral Features As discussed in Juhasz and Rayner (2003) and Clifton et al. (2007), behavioral measures highly affect eye-movements in reading tasks. For each word in the sentence, we extract behavioral features from large collections of human generated values available online: age of acquisition (Kuperman et al., 2012), prevalence (Brysbaert et al., 2019), valence, arousal, dominance (Warriner et al., 2013) and concreteness. For concreteness, we experiment both with human generated scores (*concreteness*$_{human}$, Brysbaert et al., 2014) and automatically generated ones (*concreteness*$_{auto}$, Köper and Schulte im Walde, 2017). All behavioral measures have been centered (mean equal to zero) and the missing values have been set to the corresponding mean value.

BERT Features Given the success of current language models for various NLP tasks, we investigate their expressivity for human-centered tasks such as eye-tracking: each word is mapped to two types of contextualized embeddings. First, each word is mapped to its BERT (Devlin et al., 2018) embedding extracted from the pre-trained base model. To extract the second type of contextualized embedding, we fine-tune BERT on each of the five eye-tracking measures. Specifically, the BERT base model[2] is fine-tuned separately 5 times, one for each of the eye-tracking measures to be predicted. Based on these fine-tuned models, we extract the embedding of each word as a fixed feature vector to be used for further experimentation. This means that in this step each word is in fact mapped to five distinct embeddings, one for each fine-tuned model. In the later experimentation, we use the respective embedding based on which measure is currently predicted (e.g., the embedding extracted from the model fine-tuned for nFix is used to predict nFix).

[1] https://github.com/google-research/bert

[2] We use the regression implementation from: https://github.com/fancyerii/bert

Measure	Feature Name
nFix	**word length** (0.81), **frequency score** (0.05), word length-sentence length ratio (0.01), **similarity**$_{w_m,w_{m-1}}$ (0.01), surprisal score (0.01), similarity$_{w_m,w_{1...m-1}}$ (0.01)
FFD	**word length** (0.80), **frequency score** (0.06), **similarity**$_{w_m,w_{m-1}}$ (0.02), word length-sentence length ratio (0.02), similarity$_{w_m,w_{1...m-1}}$ (0.02), surprisal score (0.01)
GPT	**word length** (0.40), surprisal score (0.27), word length-sentence length ratio (0.06), similarity$_{w_m,s}$ (0.04), **similarity**$_{w_m,w_{m-1}}$ (0.02), **frequency score** (0.02), stop word (0.02), similarity$_{w_m,w_{1...m-1}}$ (0.02), numeral token (0.02), age of acquisition (0.01), dominance (0.01)
TRT	**word length** (0.70), **frequency score** (0.11), word length-sentence length ratio (0.03), numeral token (0.01), **similarity**$_{w_m,w_{m-1}}$ (0.01), similarity$_{w_m,s}$ (0.01), sentence length in characters (0.01)
fixProp	**word length** (0.84), **similarity**$_{w_m,w_{m-1}}$ (0.04), **frequency score** (0.03), similarity$_{w_m,w_{1...m-1}}$ (0.02)

Table 2: SLB features with importance ≥ 0.01. Features in each row are sorted by their importance in descending order. Features that are strong predictors in all 5 measures are marked in **bold**.

3 Predicting Eye-Tracking Measures

We conduct three experiments using different feature combinations, and experiment with three model architectures. The models' parameters are experimentally defined. First, we train a Linear Regression model (LR). Second, we train a Decision Tree Regressor (DT) with the *mse* (*Mean Squared Error*) criterion and a maximum depth of 7. Last, we train a Random Forest Regressor (RF) with the *mse* criterion, 15 estimators and a maximum depth of 7. Before training the models, all categorical feature values are one-hot-encoded and all numeric values are normalized within the range $[0, 1]$.

3.1 Experiment 1: Using Only SLB Features

In Experiment 1, we train the aforementioned model architectures on the full set of SLB features. Among the three models, the Random Forest Regressor achieves the best overall performance, with an average MAE across all 5 eye-tracking measures of $\overline{\text{MAE}}_{\text{RF}}$ = 4.059 , $\overline{\text{MAE}}_{\text{DT}}$ = 4.187, $\overline{\text{MAE}}_{\text{LR}}$ = 4.322. To shed light on the most predictive features for each of the eye-tracking measures, we perform feature selection based on the features' weight, i.e., the impurity-based feature importance (Gini importance) computed as the normalized total reduction of the criterion brought by that feature – the higher, the more important the feature. We select features with importance higher than 0.01, resulting in a reduced SLB feature set as shown in Table 2. This selected set is further used for Experiment 3 (see Section 3.3).

3.2 Experiment 2: Using Only BERT

Our second experiment aims at investigating the expressivity of the contextualized BERT embeddings. We experiment with the two variants of BERT embeddings (see Section 2). In the first variant, the three models use the pre-trained BERT embeddings, while in the second variant, the models use the fine-tuned BERT embeddings. The latter means that for each of the 5 eye-tracking measures, the extracted embeddings of the corresponding fine-tuned model are used and 3 models are trained for each measure, with a total of 15 models. We also experiment with the predictions directly resulting from the fine-tuning tasks, but we observe that these predictions show similar performance. This finding is in line with what is reported in Devlin et al. (2018).

3.3 Experiment 3: Enhancing BERT with SLB Features

Extracting BERT embeddings as fixed-length features instead of using the predictions directly out of the fine-tuned model allows us to extend the BERT vectors with further features. Thus, in the last experiment, we train the 3 regression models on an *extended* vector, comprising the extracted 768-dimensional BERT embedding and additional dimensions for the reduced SLB feature set of Experiment 1 (see Section 3.1). Again, two variants are tested: one using the pre-trained embeddings and the other one using the fine-tuned embeddings of the corresponding model.

4 Results and Discussion

Table 3 reports the results from all experimental settings on the development set and test set (80/20 split). Due to space limits, we only report the results of the best model in each configuration. Overall, combining the embeddings from the fine-tuned version of BERT with the surface, linguistic and behavioral features gives the best performance on

	nFix	FFD	GPT	TRT	fixProp
DEVELOPMENT SET					
SLB	4.126 (RF)	0.675 (RF)	2.682 (RF)	1.615 (RF)	11.198 (RF)
Pre-trained BERT	4.925 (LR)	0.769 (LR)	2.967 (LR)	1.888 (LR)	13.530 (LR)
Fine-tuned BERT	4.694 (LR)	0.753 (LR)	2.811 (LR)	1.805 (LR)	13.140 (LR)
Pre-trained BERT + SLB	4.086 (LR)	0.676 (RF)	2.625 (RF)	1.597 (RF)	11.150 (RF)
Fine-tuned BERT + SLB	**3.982 (LR)**	**0.676 (RF)**	**2.572 (RF)**	**1.555 (LR)**	**11.147 (RF)**
TEST SET (PRE-EVALUATION)					
Fine-tuned BERT + SLB	4.263 (LR)	0.698 (RF)	2.756 (RF)	1.682 (LR)	11.683 (RF)
TEST SET (POST-EVALUATION)					
Fine-tuned BERT + SLB	4.233 (LR)	0.700 (RF)	2.751 (LR)	1.673 (LR)	11.760(RF)

Table 3: Mean absolute errors on the development and the test set. The pre-evaluation test set results are the ones submitted to the competition. We obtained the post-evaluation results after further fine-tuning.

all 5 eye-tracking measures. When we compare the predictive power of the models including only SLB features against the models trained only on BERT, we see that the embeddings are less informative than the carefully selected set of SLB features.

A closer investigation of the selected SLB features in Table 2 provides interesting insights about the nature of the features and the task.

Surface Features Among all SLB features, *word length* is consistently the predictor with the highest weight across all 5 measures. Furthermore, *word length-sentence length ratio* is among the most important contributors in 4 of the 5 measures. This confirms the observation in Hollenstein et al. (2018, p. 10) that the probability of a word being skipped reduces as word length increases.

Linguistic Features Two features for word expectedness, i.e., *frequency score* and *similarity*$_{w_m,w_{m-1}}$, also show a high predictive power for all 5 measures. This confirms previous findings by Godfroid et al. (2018) and Williams and Morris (2004). Likewise, *similarity*$_{w_m,w_{1\ldots m-1}}$ ranks among the most important features for 4 of the 5 measures, and *surprisal score* for 3 of the 5 measures. Most importantly, surprisal score shows a much higher importance in predicting GPT, which indicates that encountering an unexpected word may cause a regressive reading to re-inspect previous words and thus increases the go-past time. On the other hand, the syntactic properties of a word (e.g., POS, dependency relation and named entity type) do not show any strong effect in our results. The only exception is that *numeral tokens* are among the most important features in predicting GPT and TRT. After a closer look into the data, we found that a majority of the numeral tokens are information about date (e.g. *November 28*; *1826-1905*). The effect of such numeral tokens could probably be explained by the nature of the data, where a majority of the sentences are biographical sentences from Wikipedia (Hollenstein et al., 2018, 2019). In such data, this numeral information is highly relevant for the context.

Behavioral Features *Dominance* and *age of acquisition* also play a significant role in predicting GPT: as indicated in the literature (Juhasz and Rayner, 2003), such behavioral measures have a strong impact on the processing time of words in context.

5 Conclusion

We presented a system of eye-tracking feature prediction which combines BERT with a rich set of surface, linguistic and behavioral (SLB) features. Overall, our three studies indicate that including not only semantic properties that can be directly extracted from text, such as embeddings and surprisal score, but also measures reflecting behavioral (e.g., dominance and age of acquisition) and surface properties (word and sentence length) has a positive impact on the performance of our models in predicting eye-tracking data.

References

Marc Brysbaert, Paweł Mandera, Samantha F McCormick, and Emmanuel Keuleers. 2019. Word prevalence norms for 62,000 English lemmas. *Behavior Research Methods*, 51(2):467–479.

Marc Brysbaert and Boris New. 2009. Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, 41(4):977–990.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, 46(3):904–911.

Charles Clifton, Adrian Staub, and Keith Rayner. 2007. Eye movements in reading words and sentences. In Roger P.G. Van Gompel, Martin H. Fischer, Wayne S. Murray, and Robin L. Hill, editors, *Eye Movements*, pages 341–371. Elsevier, Oxford.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *preprint arXiv:1810.04805*.

Stefan Frank and Robin Thompson. 2012. Early effects of word surprisal on pupil size during reading. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 34.

Aline Godfroid, Jieun Ahn, Ina Choi, Laura Ballard, Yaqiong Cui, Suzanne Johnston, Shinhye Lee, Abdhi Sarkar, and Hyung-Jo Yoon. 2018. Incidental vocabulary learning in a natural reading context: An eye-tracking study. *Bilingualism*, 21(3):563.

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(1):1–13.

Nora Hollenstein, Marius Troendle, Ce Zhang, and Nicolas Langer. 2019. ZuCo 2.0: A dataset of physiological recordings during natural reading and annotation. *preprint arXiv:1912.00903*.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657.

Barbara J Juhasz and Keith Rayner. 2003. Investigating the effects of a set of intercorrelated variables on eye fixation durations in reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29(6):1312.

Maximilian Köper and Sabine Schulte im Walde. 2017. Improving verb metaphor detection by propagating abstractness to words, phrases and individual senses. In *Proceedings of the 1st Workshop on Sense, Concept and Entity Representations and their Applications*, pages 24–30.

Victor Kuperman, Hans Stadthagen-Gonzalez, and Marc Brysbaert. 2012. Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, 44(4):978–990.

Yongjie Lin, Yi Chern Tan, and Robert Frank. 2019. Open sesame: Getting inside BERT's linguistic knowledge. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 241–253.

Boris New, Ferrand Ludovic, Pallier Christophe, and Brysbaert Marc. 2006. Reexamining the word length effect in visual word recognition: New evidence from the English Lexicon Project. *Psychonomic Bulletin & Review*, 13(1):45–52.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Roland Schäfer and Felix Bildhauer. 2012. Building large corpora from the web using a new efficient tool chain. In *Eighth International Conference on Language Resources and Evaluation (LREC)*, pages 486–493.

Nathaniel J Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319.

Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, 45(4):1191–1207.

Rihana Williams and Robin Morris. 2004. Eye movements, word familiarity, and vocabulary acquisition. *European Journal of Cognitive Psychology*, 16(1-2):312–339.

Ye Zhang, Diego Frassinelli, Jyrki Tuomainen, Jeremy I Skipper, and Gabriella Vigliocco. 2020. More than words: Word predictability, prosody, gesture and mouth movements in natural language comprehension. *preprint BioRxiv*.