Looking for a Role for Word Embeddings in Eye-Tracking Features Prediction: Does Semantic Similarity Help?

Lavinia Salicchi
The Hong Kong Polytechnic University
lavinia.salicchi@connect.polyu.hk

Alessandro Lenci
University of Pisa
alessandro.lenci@unipi.it

Emmanuele Chersoni
The Hong Kong Polytechnic University
emmanuelechersoni@gmail.com

Abstract

Eye-tracking psycholinguistic studies have suggested that context-word semantic coherence and predictability influence language processing during the reading activity.

In this study, we investigated the correlation between the cosine similarities computed with word embedding models (both static and contextualized) and eye-tracking data from two naturalistic reading corpora. We also studied the correlations of surprisal scores computed with three state-of-the-art language models.

Our results show strong correlation for the scores computed with BERT and GloVe, suggesting that similarity can play an important role in modeling reading times.

1 Introduction

Eye-tracking data recorded during reading provide invaluable evidence about the factors influencing language comprehension. Research in computational modeling has particularly focused on two factors: i.) the semantic coherence of a word with the rest of the sentence (Ehrlich and Rayner, 1981; Pynte et al., 2008; Mitchell et al., 2010), measured via *semantic similarity* metrics and ii.) its predictability from previous context, as measured by *surprisal* (Hale, 2001; Levy, 2008). Intuitively, words that have low semantic coherence and low in-context predictability (i.e., high surprisal) induce longer reading times.

In distributional semantics (Lenci, 2018), words and their sentence contexts are represented with dense vectors called *embeddings* and produced by Distributional Semantic Models (DSM). In this paper, we modeled semantic coherence with the cosine similarity between the embeddings of words and their sentence contexts, and then we tested the correlation of the metric with the eye-tracking measures annotated on the GECO and Provo corpora. We analyzed the correlations for the similarity computed with 10 different embedding models (both static and contextualized), as well as for surprisal scores computed with several state-of-the-art neural language models. Among all the features under investigation, the similarity scores obtained with BERT and GloVe obtained the best correlations across features in both the benchmark corpora.

2 Related Work

Hollenstein et al. (2019) proposed a framework to evaluate six state-of-the-art word embedding models (GloVe, Word2Vec, WordNet2Vec, FastText, ELMo, BERT). The evaluation was based on the model capability to reflect semantic representations in the human mind, using cognitive data in different datasets for eye-tracking, EEG, and fMRI. Word embedding models were used to train neural networks on a regression task. While we aim at creating a computational model of the relationship between context processing and the integration of a new word during naturalistic reading, Hollenstein et al. (2019) evaluated embedding models on the prediction of out-of-context word features. The results of their analyses showed that BERT, ELMo, and FastText have the best prediction performances. On the other hand, approaches based on powerful Transformers language models were outperformed by a classifier using linguistic and psychometric features (Bestgen, 2021) in the recent CMCL 2021 Shared Task on Eye-Tracking Data Prediction (Hollenstein et al., 2021).

A series of contributions explored the role of surprisal in modeling reading times in naturalistic settings, coming to the general conclusion that the predictive power is strongly related to the language model quality, i.e. models with better perplexity perform better (Smith and Levy, 2013; Goodkind and Bicknell, 2018). Later work explored the most recent neural models, including LSTM (van Schijndel and Linzen, 2018), GRU (Aurnhammer and Frank, 2019), Transformers (Merkx and Frank, 2020) and GPT-2 (Wilcox et al., 2020), basically

Proceedings of the 14th International Conference on Computational Semantics, pages 87–92
June 17–18, 2021. ©2021 Association for Computational Linguistics

confirming this relationship.[1]

Early studies had also found correlations between semantic distance, computed by word embeddings, and eye-tracking features in reading processes (Pynte et al., 2008; Mitchell et al., 2010). However, the more recent work by Frank (2017) pointed out that, since word embeddings are based on co-occurrences, semantic distance may actually represent word predictability, rather than semantic relatedness, and that those early findings were actually due to a confound between these two concepts. To test this hypothesis, the author used linear regression models with and without surprisal, testing 5 surprisal measures. The results show that the effects of similarity on reading times disappear when surprisal is factored out, thereby proving the existence of a complex interplay between the two factors. Frank's experiments were carried out in a naturalistic reading setting and, to our knowledge, there have been no eye-tracking studies with controlled stimuli investigating a possible separate effect of the two components (for example, by comparing the fixation patterns of words that have low predictability, but different degrees of coherence with the sentence or with the discourse context).

3 Experimental Setting

3.1 Datasets

Traditional corpora annotated with eye-tracking data consist of short isolated sentences (or even single words) with particular structures or lexemes, in order to investigate specific syntactic and semantic phenomena. In the present work, we used GECO (Cop et al., 2017) and Provo (Luke and Christianson, 2018), two eye-tracking corpora containing long, complete, and coherent texts. **GECO** is a monolingual and bilingual (English and Dutch) corpus composed of the entire Agatha Christie's novel *The Mysterious Affair at Styles*. The corpus is freely downloadable with a related dataset containing eye-tracking data of 33 subjects (19 of them bilingual, 14 English monolingual) reading the full novel text, presented paragraph-by-paragraph on a screen. GECO is composed of $54,364$ tokens. **Provo** contains 55 short English texts about various topics, with 2.5 sentences and 50 words on average, for a total of $2,689$ tokens, and a vocabulary of $1,197$ words. These texts were read by 85 subjects and their eye-tracking measures were collected in an available on-line dataset. GECO and Provo data are particularly interesting because they are recorded during naturalistic reading, instead of short selected stimuli.

For every word in the corpora, we extracted its mean *total reading time*, mean *first fixation duration*, and mean *number of fixations*, by averaging over the subjects. The choice of modeling mean eye-tracking measures is justified by the high inter-subject consistency of the recorded data. For instance, Cop et al. (2017) report an overall inter-subject correlation of 0.9 for the total reading times in GECO.

3.2 Word Embeddings

Table 1 shows the embeddings types used in our experiments, consisting of 6 non-contextualized, static DSMs and 4 contextualized DSMs. The former include predict models (**SGNS** and **FastText**) (Mikolov et al., 2013; Levy and Goldberg, 2014; Bojanowski et al., 2017) and count models (**SVD** and **GloVe**) (Bullinaria and Levy, 2012; Pennington et al., 2014). [2] Four DSMs are window-based and two are syntax-based (**synt**). Embeddings have 300 dimensions and were trained on a corpus of 3.9 billion tokens ca. (a concatenation of ukWaC and a 2018 dump of Wikipedia). Pre-trained contextualized embeddings include the 512-dimensional vectors produced by the three layers of the **ELMo** bidirectional LSTM architecture (Peters et al., 2018), the $1,024$-dimensional vectors produced by the 24-layers **BERT-Large** Transformer architecture (BERT-Large, Cased) (Devlin et al., 2019), the $1,600$-dimensional vectors by **GPT2-xl** (Radford et al.), and finally, the 200-dimensional vectors produced by the **Neural Complexity** model by van Schijndel and Linzen (2018).

3.3 Method

Our main goals were to investigate the potential contribution of cosine similarity in predicting eye-tracking features, to compare different word embedding models, and then to evaluate whether the information represented by cosine similarity is similar to the one represented by surprisal.

For each target word w in GECO and Provo, we measured the **cosine similarity** between the embedding of w and the embedding of the context

[1] Notice however that doubts have been raised on the reliability of perplexity as a metric for comparing large pretrained models, since it does not allow to compare models with different vocabularies (Hao et al., 2020).

[2] For the distinction between count and predict DSM, we refer to Baroni et al. (2014).

Model	Hyperparameters
Non-contextualized DSMs	
SVD.w2	count DSM with 345K window-selected context words, window of width 2, reduced with SVD
SVD.synt	count DSM with 345K syntactically typed context words reduced with SVD
GloVe	count DSM with context window of width 2, reduced with log-bilinear regression
SGNS.w2	Skip-gram with negative sampling, context window of width 2, 15 negative examples
SGNS.synt	Skip-gram with negative sampling, syntactically-typed context words, 15 negative examples
FastText	Skip-gram with subword information, context window of width 2, 15 negative examples
Contextualized DSMs	
ELMo	Pretrained ELMo embeddings on the 1 Billion Word Benchmark
BERT	Pretrained BERT-Large embeddings on the concatenation of the Books corpus and Wikipedia
GPT2-xl	Pretrained GPT2-xl embeddings on WebText
Neural Complexity	Pretrained Neural Complexity embeddings on Wikipedia

Table 1: List of the embedding models used for the study, together with their hyperparameter settings.

c formed by the previous words in the same sentence. We then computed the Spearman correlation between the cosine and the eye-tracking data for w (total reading time, first fixation duration, and number of fixations). To create context embedding, we used an **additive model**: the context vector is the sum of all its word embeddings.

Given the bidirectional nature of BERT, the input to this model needed a special pre-processing: To prevent that the vectors representing words within the context were computed using the target word itself, we passed to BERT a list of sub-sentences, each of which were composed of context words only. So given the sentence *The dog chases the cat*:
S[0] = [”The”]
S[1] = [”The dog”]
S[2] = [”The dog chases”]
S[3] = [”The dog chases the”]
S[4] = [”The dog chases the cat”]
Starting from the second sub-sentence, the cosine similarity was computed between the last word vector and the sum of words vectors belonging to the previous sub-sentence (list element). So, to compute the cosine similarity between *cat* and the previous context, we selected *cat* from S[4] and $The + dog + chases + the$ from S[3].

For BERT we used as context also the embedding produced by the model for the special token **CLS**, which is created using a weighted additive model. As for the $simple$ additive model, BERT was fed with sub-sentences, and for each target word the CLS-context-vector was the one computed at the previous list element. In the previous example, given *cat* as target word, we used the CLS vector representing all the S[3] elements.

Given the positive effect of semantic coherence on language processing, we expected that the eye-tracking data for w had a *negative correlation* with its cosine similarity with c: **The higher the cosine, the lower the reading time of w measured by eye-tracking**.

We used BERT, GPT2-xl and Neural Complexity to compute word-by-word surprisal. Like with cosine similarity, the input sentences for BERT were organized in sub-sentences, and the last token (i.e., the target word), was replaced with the special tag [MASK]. Finally, we computed the Spearman correlation between the **surprisal** of w, and the eye-tracking data for the target word. Differently from the cosine, we expected the surprisal to be *positively correlated* with the word reading time: **The less predictable a word is, the slower its processing will be**.

The analyses have been performed with the following models: 6 values of cosine similarity between non-contextualized vectors, 51 values of cosine similarity between contextualized vectors (48 from 24 layers of BERT in two different ways to compute the context vector, and 3 from ELMo, GPT2-xl and Neural Complexity), 3 values of surprisal from BERT, GPT2-xl, Neural Complexity.

4 Results and Discussion

Looking at the correlations results, it is clear that every model performed better on Provo. One possible explanation for this difference is that GECO eye-tracking data are recorded on participants reading a literary text, while Provo materials are online news articles, science magazines and only partially short text from works of fiction. The consequence is a difference in the syntactic complexity of sentence structure and in the frequency of words. This gap implies that the modeling of GECO contexts is less directly reducible to an additive fashion of processing, and, most importantly, is more likely to find *Out Of Vocabulary* words in GECO, rather than in Provo.

Corpus	Model	total reading time	1st fix. duration	number fixations
GECO	BERT Additive (22)	-0.54	-0.53	-0.55
	BERT CLS (22)	**-0.57**	**-0.56**	**-0.58**
	ELMo (1)	-0.35	-0.34	-0.36
	FastText	-0.39	-0.38	-0.40
	GloVe	-0.45	-0.44	-0.46
	SGNS.w2	-0.40	-0.39	-0.40
	SGNS.synt	-0.30	-0.29	-0.30
	SVD.w2	*-0.07*	*-0.06*	*-0.07*
	SVD.synt	-0.24	-0.23	-0.24
	GPT2-xl	-0.05	-0.05	-0.05
	NC	-0.12	-0.11	-0.12
Provo	BERT Additive (22)	-0.65	-0.66	-0.66
	BERT CLS (22)	**-0.71**	**-0.72**	**-0.71**
	ELMo (1)	-0.36	-0.36	-0.37
	FastText	-0.57	-0.56	-0.57
	GloVe	-0.65	-0.65	-0.66
	SGNS.w2	-0.60	-0.60	-0.60
	SGNS.synt	-0.42	-0.42	-0.43
	SVD.w2	*-0.03*	*-0.02*	*-0.03*
	SVD.synt	-0.32	-0.32	-0.32
	GPT2-xl	-0.37	-0.38	-0.38
	NC	-0.16	-0.17	-0.17

Table 2: Spearman correlations between the target-context cosine and the eye-tracking measures. Numbers in parenthesis indicate models' layers.

Corpus	Model	total reading time	1st fixation duration	number fixations
GECO	BERT	0.28	0.26	0.28
	GPT2-xl	0.41	0.39	0.41
	NC	0.31	0.30	0.32
Provo	BERT	0.25	0.24	0.24
	GPT2-xl	0.44	0.43	0.44
	NC	0.46	0.48	0.46

Table 3: Spearman correlations between surprisal and eye-tracking measures.

Another aspect that is quite evident are the similar correlation values among different eye-tracking features. This aspect is not surprising: in the original datasets of GECO and Provo, it can be noticed that many words show the same value for the total reading time and the first fixation duration. This happens when i) the word is not read (0 ms for both the features); ii) the word is read only once (total reading time and first fixation duration overlap). Also regarding the similar values of the correlations between similarity and number of fixations and between similarity and total reading times, taking into account the original data gives us an explanation of the results: since the total reading time is computed summing the duration of all the multiple fixations, the higher the number of fixation, the higher the total reading time, leading to a similar tendency in the values of the two features. For these reasons, the total reading time may be considered as a "bridge" field, that holds close relations with both first fixation duration and number of fixations, justifying the similar correlation values in our results.

Comparing word embedding models, we may notice that correlations can reach very high values, up to -0.71 for the total reading time (by BERT CLS layer 22), suggesting that semantic coherence -modeled as cosine similarity between context and target- can be a strong predictor of eye-tracking measures of reading process. GloVe (mean correlation over eye-tracking features on GECO: -0.45, on Provo: -0.65) and BERT (mean correlation over eye-tracking features on GECO: -0.57, on Provo: -0.71) score the best results on both corpora, and in the latter case the [CLS] context model brings some advantage over the simple additive one. The lower BERT layers show a steadily decreasing performance (see Figure 1). This was expected because, as it was pointed out in the layers analysis by Tenney et al. (2019), the BERT architecture reproduces the classical NLP pipeline: the lower layers process mainly the syntactic information, while the highest ones give a more precise representation of semantic relations. We also notice a strong variability among the embedding models, which is orthogonal to the contextualized vs. non-contextualized dichotomy. The ELMo contextualized vectors perform much worse than BERT ones, probably because they have a lower degree of contextualization, and syntax-based count models are not significantly worse than predict DSMs.

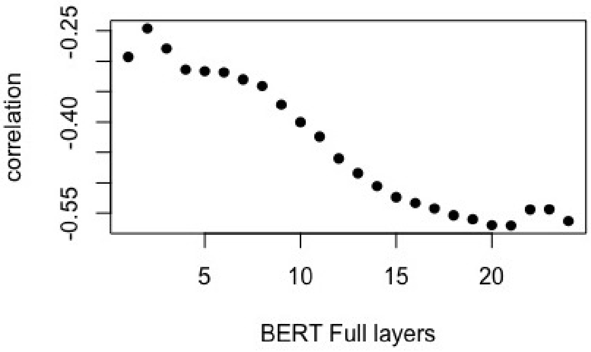

Figure 1: Spearman's correlation of different layers of BERT on GECO.

Regarding the correlations between the target word surprisal computed with BERT, GPT2-xl and Neural Complexity (NC) and the eye-tracking measures (see Table 3), the first striking fact is that the absolute values are generally lower than the scores obtained with the cosine (higher correlations are reached by GPT2-xl con GECO, mean correlation = 0.40, and by NC on Provo, mean correlation = 0.47). This might prompt us to conclude that surprisal is a much weaker predictor than semantic coherence. However, a significant negative correlation between cosine similarity and surprisal (e.g. with BERT it is -0.40 on GECO and -0.32 on Provo) supports the hypothesis by Frank (2017) that there is a strong overlap between semantic coherence and surprisal. Factoring out the contribution of these two factors on eye-tracking features will be the next step of our research work.

5 Conclusions and ongoing work

In this paper, we have used contextualized and non-contextualized DSMs to compute the cosine between a target word and the previous sentence context. Our results show that cosine similarity is able to achieve very high correlations with the eye-tracking metrics of GECO and Provo, especially with the BERT and GloVe models, providing further evidence that semantic coherence is potentially very useful in modeling reading times. Furthermore, we computed word-by-word surprisal using BERT, GPT2-xl, and Neural Complexity.

Among the language models, the best results have been achieved by GPT2-xl, confirming the previous findings that Transformers are very good at modeling sentence processing metrics (Wilcox et al., 2020; Hao et al., 2020; Merkx and Frank, 2021). However, the absolute value of correlation is lower than the one obtained with cosine similarity scores: for example, the mean correlation achieved on Provo with the cosine similarity between vectors produced by BERT is -0.71, while the correlation between eye tracking features and the surprisal computed by the same model is 0.24. The comparison between correlations reached by cosine similarity and surprisal may lead us to the conclusion that semantic coherence is a stronger predictor of eye-tracking features than word predictability. However, given the significant degree of correlation between cosine similarity and surprisal, further investigations are needed to disentangle the two factors.

Our next step will be to include Transformers-based surprisal and vector-based cosine similarity in a large-scale regression study to predict eye tracking features, in order to ensure a close comparison with the experimental setting of Frank (2017), and to investigate if semantic similarity models can actually play a distinct role from surprisal in the prediction of reading times. Differently from Frank (2017), we plan to test with several regression models, from a simple linear regression to more advanced regression models (e.g. Gradient Boosting, Multilayer Perceptron etc.), and with different word embedding models, in order to account for the different types of semantic similarity computed by static and contextualized embeddings.

References

Christoph Aurnhammer and Stefan L Frank. 2019. Evaluating Information-theoretic Measures of Word Prediction in Naturalistic Sentence Reading. *Neuropsychologia*, 134.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't Count, Predict! A Systematic Comparison of Context-counting vs. Context-predicting Semantic Vectors. In *Proceedings of ACL.*

Yves Bestgen. 2021. LAST at CMCL 2021 Shared Task: Predicting Gaze Data During Reading with a Gradient Boosting Decision Tree Approach. In *Proceedings of the NAACL Workshop on Cognitive Modeling and Computational Linguistics*.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with

Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

John A Bullinaria and Joseph P Levy. 2012. Extracting Semantic Representations from Word Co-Occurrence Statistics: Stop-Lists, Stemming, and SVD. *Behavior Research Methods*, 44(3):890–907.

Uschi Cop, Nicolas Dirix, Denis Drieghe, and Wouter Duyck. 2017. Presenting GECO: An Eye-Tracking Corpus of Monolingual and Bilingual Sentence Reading. *Behavior Reseach Methods*, 49(2):602–615.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL*.

Susan E. Ehrlich and Keith Rayner. 1981. Contextual Effects on Word Perception and Eye Movements During Reading. *Journal of Verbal Learning and Verbal Behavior*, 20:641–65.

Stefan L Frank. 2017. Word Embedding Distance Does not Predict Word Reading Time. In *Proceedings of CogSci*, pages 385–390.

Adam Goodkind and Klinton Bicknell. 2018. Predictive Power of Word Surprisal for Reading Times is a Linear Function of Language Model Quality. In *Proceedings of the LSA Workshop on Cognitive Modeling and Computational Linguistics*.

John Hale. 2001. A Probabilistic Earley Parser as a Psycholinguistic Model. In *Proceedings of NAACL*.

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 EMNLP Workshop on Cognitive Modeling and Computational Linguistics*.

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

Nora Hollenstein, Antonio de la Torre, Nicolas Langer, and Ce Zhang. 2019. CogniVal: A Framework for Cognitive Word Embedding Evaluation. In *Proceedings of CONLL*.

Alessandro Lenci. 2018. Distributional Models of Word Meaning. *Annual Review of Linguistics*, 4:151–171.

Omer Levy and Yoav Goldberg. 2014. Dependency-Based Word Embeddings. In *Proceedings of ACL*.

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

Steven G Luke and Kiel Christianson. 2018. The Provo Corpus: A Large Eye-tracking Corpus with Predictability Norms. *Behavior Research Methods*, 50(2):826–833.

Danny Merkx and Stefan L Frank. 2020. Comparing Transformers and RNNs on Predicting Human Sentence Processing Data. *arXiv preprint arXiv:2005.09471*.

Danny Merkx and Stefan L Frank. 2021. Human Sentence Processing: Recurrence or Attention? In *Proceedings of the NAACL Workshop on Cognitive Modeling and Computational Linguistics*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *arXiv preprint arXiv:1301.3781*.

Jeff Mitchell, Mirella Lapata, Vera Demberg, and Frank Keller. 2010. Syntactic and Semantic Factors in Processing Difficulty: An Integrated Measure. In *Proceedings of ACL*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global Vectors for Word Representation. In *Proceedings of EMNLP*.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *Proceedings of NAACL*.

Joel Pynte, Boris New, and Alan Kennedy. 2008. Online Contextual Influences During Reading Normal Text: A Multiple-Regression Analysis. *Vision research*, 48(21):2172–2183.

A. Radford, Jeffrey Wu, R. Child, David Luan, Dario Amodei, and Ilya Sutskever. Language Models are Unsupervised Multitask Learners. In *Open-AI Blog*.

Marten van Schijndel and Tal Linzen. 2018. A Neural Model of Adaptation in Reading. In *Proceedings of EMNLP*.

Nathaniel J Smith and Roger Levy. 2013. The Effect of Word Predictability on Reading Time Is Logarithmic. *Cognition*, 128(3):302–319.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT Rediscovers the Classical NLP Pipeline. *arXiv preprint arXiv:1905.05950*.

Ethan Gotlieb Wilcox, Jon Gauthier, Jennifer Hu, Peng Qian, and Roger Levy. 2020. On the Predictive Power of Neural Language Models for Human Real-Time Comprehension Behavior.