Universität Zürich[UZH]

Seminar
Cognitively Enhanced NLP
Fall semester 2024

Monolingual vs. Multilingual Language Models: Correlation of Model and Human Attention

Author: Luc Aggett
Student ID: 20-918-371

Course instructor: Lena Bolliger
Department of Computational Linguistics

Submission date: 09.01.2025

Abstract

This paper investigates the correlation between human attention and language model (LM) attention for monolingual and multilingual models. The paper employs the ZuCo 2.0 dataset, which comprises eye-tracking data for natural reading in English. The attention scores from the first and last layers of BERT (a monolingual transformer-based model) and mBERT (a multilingual BERT) are extracted and compared to the human total reading time metric using Spearman correlations. The paper examines the degree of alignment between model and human attention, providing insights into the cognitive plausibility of these models.

1 Introduction

The mechanisms underlying human language processing have long been a central focus in computational linguistics. With the advent of pre-trained LMs, such as BERT and multilingual BERT (mBERT) Devlin et al. (2019), significant advancements have been achieved in natural language processing (NLP). These models leverage attention mechanisms to encode contextual dependencies, which enhance performance, and also provide opportunities to explore parallels between human and model-based attention. Human attention during reading, as captured through cognitive data like eye-tracking, offers insights into which parts of a sentence are most significant during processing. Increasing attention has been directed toward integrating such data in NLP to improve the interpretability, performance, and alignment of the model with human cognition. Comparison of LM attention patterns with human attention has been proposed as a method to evaluate the cognitive plausibility of these models, potentially showing the extent to which LMs replicate human-like reading behaviors. Although previous studies have shown that monolingual LMs, such as BERT, correlate with human attention Eberle et al. (2022), the effects of multilinguality in LMs remain underexplored. In this paper, an investigation is conducted to determine whether multilingual LMs, such as mBERT, exhibit different correlations with human attention compared to their monolingual counterparts. Specifically, the ZuCo eye tracking dataset is used to quantify the correlations between human attention scores and attention from the first and last layers of BERT and mBERT.

2 Related Work

Transformer-based LMs have revolutionized NLP by capturing complex linguistic patterns through hierarchical representations in their layers.

2.1 Layer-wise Representations in Transformer Models

A key property of transformer-based models is their hierarchical encoding of linguistic information across layers. Lower layers typically capture surface-level and syntactic features, while higher layers encode task-relevant and semantic information. Rogers et al. shows BERT's representations are aligned with this hierarchy, progressing from syntax to semantics as information flows through the layers. This aligns with cognitive models of human language processing, where earlier stages involve syntactic parsing followed by semantic understanding. Other works, such as Clark et al., have demonstrated that specific attention heads in transformers correspond to syntactic dependency relations, with their significance varying across layers.

2.2 Monolingual vs. Multilingual LMs

The trade-offs between monolingual and multilingual LMs have previously been researched. Monolingual models, like BERT, are pre-trained exclusively on data in one language and often outperform

multilingual models in tasks specific to that language. For example, Conneau et al. observed that multilingual models such as multilingual BERT (mBERT) and XLM-R excel in cross-lingual transfer but occasionally lag behind monolingual models in tasks requiring deep linguistic nuance. This performance gap is attributed to the multilingual model's need to balance representations across multiple languages, potentially sacrificing language-specific optimization. Despite this, multilingual models hold significant advantages in low-resource language scenarios. Pires et al. demonstrated that mBERT, pre-trained on over 100 languages, can perform zero-shot transfer to languages it was not explicitly trained on, making it useful for cross-lingual applications.

2.3 Cognitive Alignment of Attention Mechanisms

Research on aligning LM attention with human attention mechanisms has grown substantially. Studies using eye-tracking data, such as Hollenstein et al., have compared human fixation patterns during reading with model attention scores. These investigations reveal that higher layers of BERT and mBERT correlate more strongly with human attention, emphasizing the importance of semantic-level processing. Additionally, Abnar and Zuidema introduced the concept of attention flow, showing that aggregated attention scores across layers better reflect cognitive processes compared to raw attention weights. This body of research highlights the strengths and limitations of monolingual and multilingual LMs and their potential for alignment with human cognition

3 Problem Setting

The focus of this paper is to investigate the correlation between human attention, derived from eye-tracking data, and model attention in monolingual and multilingual LMs. To achieve this, a hypothesis is formulated:

Multilingual LMs exhibit different correlations with human attention compared to monolingual models.

To evaluate this hypothesis, the attention scores of the models are compared with human attention scores on a per-token basis. Human attention scores are extrapolated from eye-tracking measures in the ZuCo dataset. Model attention scores are extracted from the first and last layers of BERT (monolingual) and mBERT (multilingual) by aggregating attention scores across attention heads.

4 Methods

4.1 Language Models

The models evaluated in this paper are BERT (a monolingual transformer-based model) and multilingual BERT (mBERT). BERT is trained on English text and optimized for capturing contextual relationships within a single language. mBERT, in contrast, is pre-trained on over 100 languages, balancing cross-lingual generalization with per-language optimization. For both models, attention scores are extracted from the first and last layers to explore layer-wise differences in alignment with human attention.

4.2 Dataset

The study employs the ZuCo dataset, which comprises eye-tracking data for natural reading in English. The dataset provides metrics like total reading time, nFixations, and first-pass duration at the token level, capturing human attention during reading. The analysis uses word-level averages of total reading time

across participants as proxies for human attention. Due to issues with the data import from the raw ZuCo dataset, the data of two of the twelve participants had to be discarded.

4.3 Attention Scores

4.3.1 Machine Attention Scores

To compute attention scores, sentences from the dataset are tokenized using both monolingual and multilingual tokenizers. For each tokenized sentence, attention scores are extracted from the **first** and **last layers** of both models.

- **Tokenization**: Sentences are tokenized into subword units, with special tokens (e.g., `[CLS]` and `[SEP]`) included.
- **Attention Extraction**: For each layer of interest (first and last):
  - The attention weights for the token `[CLS]` are extracted from all attention heads.
  - Attention scores are averaged across heads to produce a single vector per token.
- **Token-to-Word Alignment**: Attention scores are mapped from subword tokens back to the original words using an alignment procedure:
  - The tokens are concatenated until they reconstruct a word in the sentence.
  - Attention scores for words are computed by averaging the attention values of their aligned subword tokens.

4.3.2 Human Attention Scores

Human attention scores in this study are derived from the ZuCo 2.0 dataset, which provides a rich set of eye-tracking metrics. These metrics reflect human reading behavior and cognitive processing during natural text comprehension. This study utilizes the total reading time metric as a proxy for human attention. The total reading time represents the sum of all fixation durations on a word, capturing the overall cognitive effort allocated to that word.

This representation of human attention provides a basis for comparing cognitive attention patterns with the attention mechanisms of monolingual and multilingual LMs. By using eye-tracking data as a direct measure of human cognitive engagement during reading, the study establishes a link between natural human behaviors and computational attention mechanisms.

4.4 Analysis

The analysis focuses on measuring the alignment between model-generated attention scores and human attention scores derived from the dataset:

- **Human Attention Scores**: Human attention scores, represented by total reading time, is retrieved at the word level from the eye-tracking data.
- **Spearman Correlation**: For each sentence, the correlation between human and model attention scores is computed using the **Spearman correlation coefficient**, which measures monotonic relationships:
  - Attention scores for each word are compared to its corresponding human attention value.

- **Results Aggregation**: Correlations are computed for:
  - First and last layers of the **monolingual model**.
  - First and last layers of the **multilingual model**.
- This is done for each sentence and participant, and results are aggregated into a tabular format for further analysis.

By examining these correlations, this paper quantifies the degree of alignment between model and human attention, providing insights into the cognitive alignment of attention mechanisms across monolingual and multilingual models.

5 Experiments

5.1 Data Preparation

The dataset used in this paper requires preprocessing to align with the input requirements of the LMs. Preprocessing steps include:

- **Tokenization**: Sentences are tokenized using the Hugging Face tokenizers for both monolingual and multilingual models. This step ensures consistency between the input format and the models' pre-trained tokenization schemes.
- **Token-to-Word Alignment**: Tokens generated by the tokenizer are aligned back to the original words using a custom alignment algorithm. This step handles cases where words are split into multiple subwords during tokenization.
- **PAD Token Handling**: During tokenization, sentences are padded to the maximum sequence length for batch processing. PAD tokens are excluded from the computation of attention scores and correlations to prevent artifacts in the results.

These steps are essential to ensure compatibility between the cognitive data and the outputs of the models for meaningful comparisons.

5.2 Experimental Design

The experimental setup involves processing sentences through both monolingual and multilingual LMs to extract attention scores and compare them with human attention:

- **Sentence Processing**: Each sentence is passed through both the monolingual (BERT) and multilingual (mBERT) models. Tokenized inputs are generated, and attention scores are extracted from the first and last layers of the models.
- **Attention Score Aggregation**: Attention scores across heads are averaged for each token. These token-level scores are then aligned to words using the token-to-word alignment algorithm.
- **Human Attention Comparison**: Human attention scores derived from eye-tracking data (e.g., total reading time) are compared to the model's attention scores using correlation metrics.
- **Evaluation Levels**: Correlations are computed at both the token level and the sentence level to capture patterns in the alignment of human and model attention.

5.3 Metrics and Tools

The following metrics and tools are used in this study:

- **Metrics**:
  - **Spearman Correlation Coefficient**: Measures the monotonic relationship between human attention scores and model attention scores.
- **Tools**:
  - **PyTorch**: Used for implementing and running the BERT and mBERT models, as well as managing tensors and computations.
  - **Hugging Face Transformers**: Provides pre-trained models and tokenizers for BERT and mBERT, enabling efficient sentence processing and attention extraction.
  - **Pandas**: Used for data manipulation, including aligning and aggregating attention scores.
  - **SciPy**: Provides statistical functions for computing Spearman correlations and handling edge cases in the data.

These tools and metrics ensure robust and reproducible experiments for evaluating the relationship between human and model attention.

6 Results

This study examined the alignment between human attention and the attention mechanisms of monolingual (BERT) and multilingual (mBERT) LMs using Spearman correlations. The highest correlation observed was 0.833. Multilingual BERT exhibited a higher average correlation (-0.076) compared to monolingual BERT (-0.129). Layer-specific analysis revealed that the last layer had a higher average correlation (0.054) compared to the first layer (-0.259). Variability in correlation across sentences was observed, with standard deviations as follows:

- `Mono_First_Corr`: 0.257
- `Mono_Last_Corr`: 0.261
- `Multilingual_First_Corr`: 0.264
- `Multilingual_Last_Corr`: 0.259

6.1 Key Findings

The results revealed notable differences in the correlation patterns:

- Multilingual BERT (mBERT) demonstrated a slightly higher average correlation with human attention compared to monolingual BERT, suggesting that multilinguality may enhance cognitive alignment.
- Correlationss were consistently higher in the last layers of both models, indicating the importance of semantic-level processing for aligning with human attention patterns.

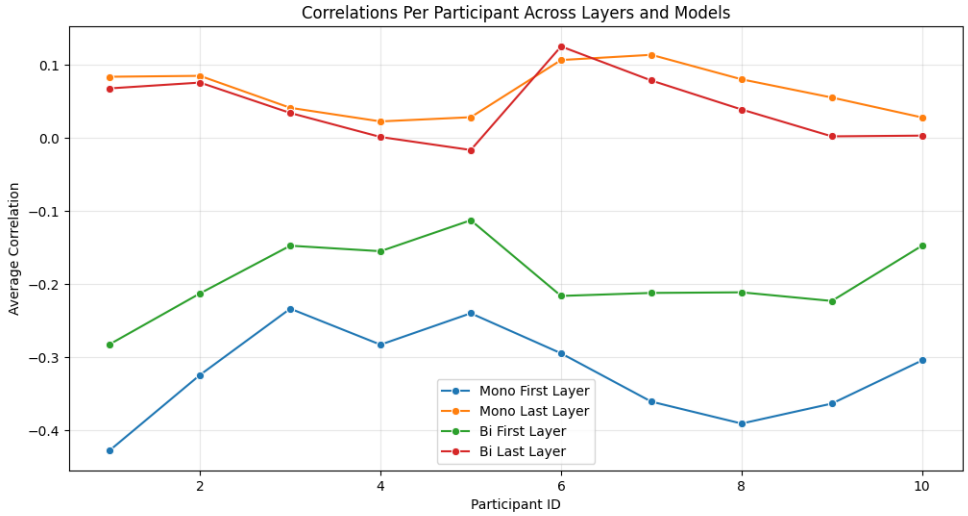

Figure 1: Distribution of attention values across participants and layers

7 Discussion

7.1 Implications of Findings

The higher correlation observed in the last layers of both BERT and mBERT aligns with prior research emphasizing the semantic processing capabilities of these layers. This suggests that deeper layers capture information more closely related to human cognitive processes during reading. The findings also indicate the role of multilingual training in facilitating broader linguistic representations, which may explain the marginally higher alignment with human attention.

7.2 Limitations

While the results are promising, there are several limitations to consider:

- This paper focuses solely on English data, limiting the generalizability of findings to other languages or multilingual contexts.
- Human attention was approximated using the total reading time metric, which, while robust, may not capture all facets of cognitive processing.
- Correlations were computed on a per-sentence basis, which may obscure finer-grained patterns of alignment at the token level or across diverse syntactic structures.

7.3 Future Work

Future research could address these limitations by:

- Expanding the analysis to include multilingual datasets, enabling a deeper exploration of cross-lingual cognitive alignment.

- Incorporating additional eye-tracking metrics, such as fixation counts or first-pass durations, to provide a more nuanced view of human attention.
- Exploring the role of intermediate layers and individual attention heads to identify specific mechanisms driving alignment with human cognition.

References

Samira Abnar and Willem Zuidema. 2020. Quantifying attention flow in transformers.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT‘s attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6022–6034, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Oliver Eberle, Stephanie Brandl, Jonas Pilot, and Anders Søgaard. 2022. Do transformer models show similar attention patterns to task-specific human gaze? In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4295–4309, Dublin, Ireland. Association for Computational Linguistics.

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 Twelfth Language Resources and Evaluation Conference*, pages 138–146, Marseille, France. European Language Resources Association.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.