Evaluating Webcam-based Gaze Data as an Alternative for Human Rationale Annotations

Stephanie Brandl[1] Oliver Eberle[2] Tiago Ribeiro[3] Anders Søgaard[1] Nora Hollenstein[1,4]

[1] University of Copenhagen, [2] Technische Universität Berlin,

[3] IT University of Copenhagen, [4] University of Zurich

1. Motivation

💡 **Rationales**, i.e., manually annotated input spans, usually serve as **ground truth** when evaluating explainability methods in NLP. They are, however, **time-consuming** and often **biased** by the annotation process.

🤔 We debate whether **human gaze,** i.e., **webcam**-based eye-tracking recordings, poses a valid alternative when evaluating XAI

Contributions:

I. We evaluate additional information provided by gaze data, such as total reading times, gaze entropy, and decoding accuracy with respect to human rationale annotations

II. We compare WebQAmGaze, a multilingual dataset for information-seeking QA, with attention and explainability-based importance scores
   A. for 4 different multilingual Transformer-based language models (mBERT, distil-mBERT, XLMR, and XLMR-L) and
   B. 3 languages (English, Spanish, and German)

III. We find that gaze data offers **valuable linguistic insights** that could be leveraged to infer task difficulty and further show a comparable ranking of explainability methods to that of human rationales.

2. Datasets

⁉ **XQuAD (Artetxe et al., 2019)**
- professional translations of question-answer pairs from a subset of SQuAD v1.1 (Rajpurkar et al.,2016) into 11 languages.
- for each context paragraph, there is a set of questions that is annotated with the correct answer

👀 **WebQamGaze (Ribeiro et al., 2023)**
- multilingual webcam-based eye-tracking dataset
- participants read texts from XQuAD
- in English (N=126), Spanish (N=51) & German (N=19), N=participants

3. Models

👾 **mBERT, distil-mBERT, XLMR, XLMR-L**
- we finetune 4 multilingual LMs individually for each of the 3 languages after filtering out samples that have been used in WebQamGaze
- 90/10 train/validation split
- we finetune for 7 epochs and 3 different seeds

4. Experiments & Results

(a) Gaze Data

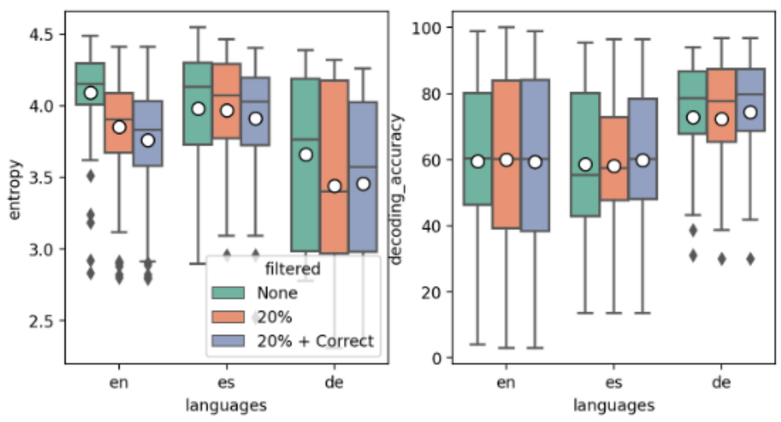


Figure 3: Entropy and decoding accuracy separated by all languages. Medians are displayed within the boxplots as a straight line whereas means are shown as white dots. Data has been filtered based on the WebGazer accuracy with a threshold of 20% (orange) and additionally we removed wrong answers (purple).

(b) Model Explanations

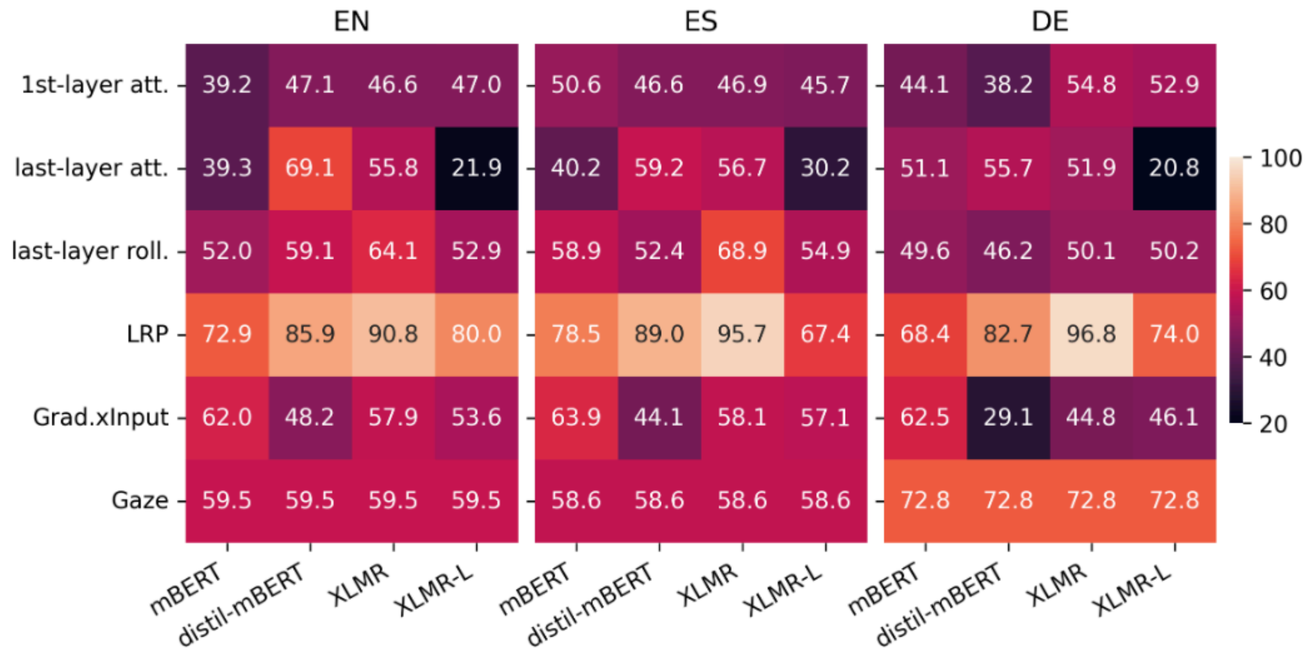
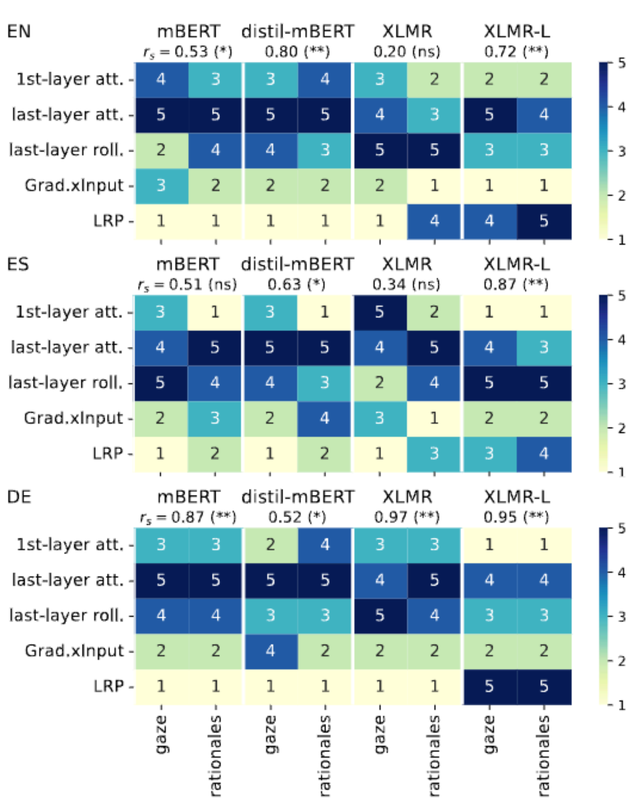


Figure 4: ROC-AUC scores for decoding rationales from attention-based and gradient-based model explanations, i.e., decoding accuracies, across all 3 languages. Results for Gaze are model-agnostic. Individual samples with an F1-scores below 50 have been filtered out per model and language.

(c) Gaze-XAI ranking

Figure 5: Comparison of gaze-based and rationale-based ranking of explanation methods for English (EN), Spanish (ES), and German (DE) – top to bottom. Ranks 1 to 5 indicate model explanations most to least aligned with human importance scores. Spearman rank correlation r_s at $p \leq 0.01$ (**), $p \leq 0.05$ (*), or not significant (ns). Results are based on text samples filtered by correct human answers.



📜 Paper

5. Takeaways

A. First look into the possibilities of low-cost gaze data as an alternative to human rationale annotations

B. We find that total reading times (all languages) and gaze entropy (Spanish and German) to strongly correlate with the error rate with negative coefficients

C. Relative position of the answer in the text as well as the text length and the number of tokens in the correct answer influence the ability to decode the gold label answer where longer texts and shorter answers lead to higher accuracies.

D. Rationales and gaze-based attention show comparable rankings, depending on the specific model and data

E. This pipeline can easily be applied to other tasks and languages.