FACE: Evaluating Natural Language Generation with Fourier Analysis of Cross-Entropy

Zuhao Yang\textsuperscript{1}, Yingfang Yuan\textsuperscript{2}, Yang Xu\textsuperscript{3,4}, Shuo Zhan\textsuperscript{1}, Huajun Bai\textsuperscript{4}, Kefan Chen\textsuperscript{2}
\textsuperscript{1} School of Computer Science and Engineering, Nanyang Technological University
\textsuperscript{2} School of Mathematical and Computer Sciences, Heriot-Watt University
\textsuperscript{3} Department of Computer Science, Southern University of Science and Technology
\textsuperscript{4} Genify

Abstract

Measuring the distance between machine-produced and human language is a critical open problem. Inspired by empirical findings from psycholinguistics on the periodicity of entropy in language, we propose FACE, a set of metrics based on Fourier Analysis of the estimated Cross-Entropy of language, for measuring the similarity between model-generated and human-written languages. Based on an open-ended generation task and the experimental data from previous studies, we find that FACE can effectively identify the human-model gap, scales with model size, reflects the outcomes of different sampling methods for decoding, correlates well with other evaluation metrics and with human judgment scores.

1 Introduction

The concept of entropy from Information Theory is broadly applied in Natural Language Processing (NLP) technology and computational linguistic studies. The most notable example is the use of cross-entropy in training and evaluating language models, where the exponentiation of cross-entropy, perplexity, is adopted to measure models’ performance in next-word (or masked-word) prediction task. However, low perplexity alone does not guarantee good performance in language generation tasks, which not only depend on model sizes but are also closely related to the sampling techniques used in decoding stage. The complexity of the generation task makes it especially important to have different metrics that can reflect the generation quality from multiple angles. One particular perspective is that the language generated from a good model should have a similar distribution of words/tokens as in the “natural” human language.

Recent advances in psycholinguistics put forward new directions for developing more sophisticated metrics other than Zipf’s coefficient. In particular, studies on temporal and spectral patterns in dialogue \cite{7,47} reveal that cross-entropy (or referred to as surprisal, information density in the psycholinguistics literature) changes periodically in natural language, which points out the potentials of using fine-grained transformation of cross-entropy to quantify the differences in language data (see Section\textsuperscript{5} for a detailed review). It motivates the basic idea of this study: Can we effectively quantify the periodic pattern of the cross-entropy, and use it as an indicator to distinguish human and model-generated languages?

We summarize our contributions as follows: 1. We propose a set of metrics based on the frequency spectra obtained from the Fast Fourier Transform (FFT) of the cross-entropy sequences of language data, named FACE (Fourier Analysis of Cross-Entropy). 2. We empirically show FACE’s performance on identifying human-model gap and how it scales with model sizes in Section\textsuperscript{4.1} 3. We explore

\* Equal contribution, in random order
\* Corresponding author, email: xuyang@sustech.edu.cn

We use a pre-trained language model \( m \) to estimate the cross-entropy of text in \( D_h \) and \( D_m \), resulting in two sequences of cross-entropy output, \( \varepsilon_h \) and \( \varepsilon_m \).

3. Obtain the frequency spectra for each cross-entropy sequence, \( \varepsilon_h \Rightarrow \mathcal{F}_h \) and \( \varepsilon_m \Rightarrow \mathcal{F}_m \).

4. Develop FACE metrics that quantify the spectral similarity between \( \mathcal{F}_h \) and \( \mathcal{F}_m \).

5. Evaluate FACE on different model types/sizes, sampling methods, and the correlations with other metrics for Natural Language Generation (NLG).

We describe the steps in detail from Section 2.1 to Section 2.3.

### 2.1 Estimate cross-entropy

We use a pre-trained language model \( m_{\text{est}} \) as the estimator for cross-entropy, which runs in the evaluation model (no gradients produced). It takes as input a sequence of \( T \) tokens, \([t_1, t_2, \ldots, t_T]\); for each position \( i = 1, \ldots, T \), it predicts the probability of the next token \( P(t_{i+1}|t_1, \ldots, t_i) \); the cross-entropy between this probability and the ground truth token \( t_{i+1} \) is then computed, resulting in the cross-entropy sequence that consists of \( T-1 \) real values \( \varepsilon = [c_1, c_2, \ldots, c_{T-1}] \), as the first token is not predicted:

\[
\varepsilon = [c_1, c_2, \ldots, c_{T-1}] \doteq [-\log P(t_2|t_1), -\log P(t_3|t_1, t_2), \ldots, -\log P(t_T|t_1, t_2, \ldots, t_{T-1})]
\]  

(1)

Note that \( \sum c_i = -\sum_{i=2}^{T} \log P(t_i|t_1 \ldots t_{i-1}) \) is exactly the definition of negative log-likelihood loss, i.e., cross-entropy loss, for training a language model, where \( c_i \) is the negative logarithm of the predicted probability for each token \( t_{i+1} \). In psycholinguistic studies, this \( c_i \) quantity is usually referred to several different terms, including surprisal \([16, 17]\), information density \([25, 20, 48]\), and entropy \([13, 14, 46, 47]\), each of which has a specific theoretical flavor. There have been debates over the justifiability of using “entropy” to denote the negative log-likelihood, because it is not a weighted summation as originally defined in \([37]\). Albeit, we decide to use cross-entropy as it is the most broadly communicated term and we believe it will not cause confusion as its mathematical form is clearly defined. Apparently, the choice for \( m_{\text{est}} \) will influence the next steps, because better...
language models produce lower perplexity scores, that is, lower cross entropy. Therefore, we discuss how different choices for $m_{rel}$ affect our metrics in Section 4.4.

2.2 Fast Fourier transform

We treat the estimated cross-entropy sequence $[c_1, \ldots, c_{T-1}]$ as a finite discrete signal in the time domain, where the sampling interval is approximated with the average duration of one token. With this simplified assumption, we find that the discrete Fourier transform (DFT) is the most suitable spectral analysis tool [39]. The formula for DFT is as follows:

$$X(\omega_k) \triangleq \sum_{n=0}^{N-1} x(t_n)e^{-j\omega_k t_n}, \quad k = 0, 1, \ldots, N - 1$$

in which $x(t_n)$ is the signal at time $t_n$, corresponding to the $n$-th cross-entropy value $c_n$ ($n = 1, \ldots, T - 1$ and $N \triangleq T - 1$). $X(\omega_k)$ is a complex number that reflects the magnitude (strength) of the $k$-th frequency component $\omega_k = 2\pi k/N$. In practice, DFT is implemented with an efficient algorithm known as Fast Fourier Transform [5] that runs in $O(n \log n)$ time.

We compared two methods, periodogram and vanilla FFT. The periodogram approach computes the Fourier transform after applying auto-correlation and time-averaging windows to the signal for de-noising purposes [43]. However, we think de-noising is inappropriate because our “signal” is a time series of cross-entropy, whose value reflects the sampling result at each time step from a large. Auto-correlation or time averaging will remove the distinctiveness of rare tokens. Therefore, we use vanilla FFT and take the real part of $X(\omega_k)$ to represent the magnitude spectrum for the frequency component $\omega_k$, which is written as $X(\omega_k)$ for brevity.

For an input cross-entropy sequence $\mathcal{E} = [c_1, \ldots, c_{T-1}]$ obtained from Section 2.1 the resulting frequency spectrum can be represented as a list of tuples of the same length, $\mathcal{F} = [(\omega_1, X(\omega_1)), \ldots, (\omega_{T-1}, X(\omega_{T-1}))]$, where $[\omega_1, \ldots, \omega_{T-1}]$ are the $T - 1$ sample frequencies, and $[X(\omega_1), \ldots, X(\omega_{T-1})]$ are the corresponding magnitudes.

2.3 Spectral similarity metrics

We develop four metrics to measure the similarity between spectra $\mathcal{F}_h$ and $\mathcal{F}_m$: Spectral Overlap (SO), Spectrum Angle Mapper (SAM), Pearson’s correlation (CORR), and Spearman’s correlation (SPEAR), as summarized in Figure 2. Before computing the metrics, two spectra $\mathcal{F}_h$ and $\mathcal{F}_m$ which are of different lengths $N_1$ and $N_2$, are first interpolated to the same length: $\mathcal{F}_h \in \mathbb{R}^{N_1} \Rightarrow \mathcal{F}_h' \in \mathbb{R}^{N_C}$, $\mathcal{F}_m \in \mathbb{R}^{N_2} \Rightarrow \mathcal{F}_m' \in \mathbb{R}^{N_C}$. Here, $N_C$ is the maximum length of the spectrum in our data. Thereafter, the computation of the subsequent metrics can commence.

Spectral Overlap (SO) is inspired by the power spectrum overlap proposed in [24], which is used in [47] for measuring the spectral similarity between dialogue participants. The frequency magnitudes in $\mathcal{F}_h'$ and $\mathcal{F}_m'$ are converted to absolute values, i.e., $\hat{X}(\omega_k) \Rightarrow |X(\omega_k)|$, and then compute the Area-Under-Curve (AUC) for the interaction $\mathcal{F}_h' \cap \mathcal{F}_m'$ and the union $\mathcal{F}_h' \cup \mathcal{F}_m'$, respectively. $SO$ is defined as the ratio of the two: $SO = \frac{\text{AUC}(\mathcal{F}_h' \cap \mathcal{F}_m')}{\text{AUC}(\mathcal{F}_h' \cup \mathcal{F}_m')}$. The procedure of converting

---

https://ccrma.stanford.edu/~jos/sasp/Fourier_Transforms_Continuous_Discrete_Time_Frequency.html
to absolute values is indispensable, since negative values in $X(\omega_k)$ will result in negative AUCs. SO has the range $[0, 1]$, and a higher value indicates a stronger resemblance between the two spectra.

Spectrum Angle Mapper (SAM) calculates the angles between $F'_h$ and $F'_m$, treating them as two vectors in a space [22]. The angle is measured in radians, which is calculated by the inverse function $\arccos(F'_h \cdot F'_m / \|F'_h\| \cdot \|F'_m\|)$, producing a value within $[0, \pi]$. We understand SAM is equivalent to the cosine similarity score, which is more commonly used in NLP, but here we just follow the conventions in [22]. A smaller SAM value indicates a greater similarity between $F'_h$ and $F'_m$.

Pearson’s correlation (CORR) can also be leveraged to measure spectral similarities as discussed in [22]. $\text{CORR} = \text{cov}(F'_h, F'_m) / \sigma(F'_h)\sigma(F'_m)$, with a $[-1, 1]$ range. A positive CORR value indicates high similarity (negative for dissimilarity), and 0 indicates weak correlation between $F'_h$ and $F'_m$.

Spearman’s correlation (SPEAR) [41] is commonly used to assess the monotonic relationship between the comparison and reference groups and to capture the presence of non-linear associations between the two. It has not been used for spectral similarity to the best of our knowledge, but we test it in our experiments. SPEAR also has the range $[-1, 1]$ with meanings similar to CORR.

3 Related Work

Entropy as a metric in psycholinguistics. The entropy of human language has long been a research interest in computational linguistics and psycholinguistics. The entropy of written text is estimated with the average per-word negative log-probability in sentences, and then used to validate the principle of entropy rate constancy (ERC) in human language [13] [14]. Similar studies were conducted in dialogue [46] [32]. Entropy is also defined in probabilistic grammars to describe the capacity of a language [40], and is used to develop complexity metrics to measure the cognitive load of processing syntactic expressions [16] [24] [17]. In the line of work on language production, a different term information density is the same mathematical formulation is used instead of entropy. It is found that speakers reduce syntactic complexity when the information density (or entropy) is high [25] [20]. In parallel with the concept of ERC, this line of work summarizes the tendency of distributing information evenly in human language with the term uniform information density (UID), which is commonly used as an equivalent term as ERC, for example, in [48] [15]. In conclusion, entropy is commonly used as a metric for essential properties of human language.

Periodical change of cross-entropy in language. We draw inspiration from the following studies about the distribution of information in dialogue. Humans are sensitive to the peaks and troughs of entropy in speech, with evidence from human-system dialogues and crowd-sourced ratings from human judges [7]. The entropy of utterances from two speakers converge towards each other within the scope of topical segments in spontaneous dialogues [48]. They measure the entropy of utterances from two participants of a task-oriented dialogue, and have found that the frequency domain features – power spectrum overlap and phase delay – are useful predictors of task outcomes. Both works reviewed above suggest that the periodical up-and-downs of entropy are commonly observable in the human language. It naturally leads to the question of whether and to what extent model-generated language aligns with this empirical finding.

Automatic measures for text generation. Entropy and its related variants have already used as a metric for evaluating generated text, for instance, entropy provides good visualization for the difference between GPT2-generated text and human written ones [12]. Other than entropy, there is a rich body of existing metrics targeted on discriminating human-written text and model-generated text, which we summarize in three branches: (1) statistics-based; (2) language modeling; (3) reference-based. Table 1 gives a brief summary of these three categories, as well as our proposed frequency-based FACE.

Statistics-based measures compare the model-generated distribution $M$ with respect to the human-written distribution $H$ in terms of some statistic. The Zipf coefficient [30] is used in [19] to describe the distribution of word frequencies in text. Self-BLEU [52] is derived by calculating the BLEU [29] score for each generated text utilizing all other generations as references. Repetition measures the sequence-level degree of repetition on the basis of the percentage of duplicated n-grams in the generated continuations $x_{cont} \sim M$ [44]. Meanwhile, we aggregate the 2-gram, 3-gram, and 4-gram repetition rates to evaluate the lexical diversity in an inverse manner.
We consider such a text completion task in three domains: Wiki text, News, and Stories. Intuitively, writing style, which may reflect on metrics. We generate completions from large-scale language models (LMs). In particular, we adopt three representatives of state-of-the-art pre-trained autoregressive LMs: Generative Pre-trained Transformer 2 (GPT2) \[33\], Open Pre-trained Transformer LMs (OPT) \[49\], and BigScience Large Open-science Open-access Multilingual LM (BLOOM) \[35\].

We explore two sizes for each model to illustrate that our FACE metrics generalize across multiple LM families and sizes. Details regarding our task and input data are summarized in Table 2.

### Table 1: Summary of metrics (automatic & non-automatic) we employed for evaluating open-ended text generation.

<table>
<thead>
<tr>
<th>Type</th>
<th>Metric</th>
<th>Measure</th>
<th>Definition/Approximation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statistics</td>
<td>Zipf Coefficient</td>
<td>Unigram rank-frequency statistics</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Self-BLEU</td>
<td>N-gram diversity</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Repetition</td>
<td>Sequence-level percentage of repetition</td>
<td>[1 - \frac{\text{unique n-grams } (\text{cont})}{\text{total n-grams}}]</td>
</tr>
<tr>
<td></td>
<td>Diversity</td>
<td>Inverse of n-gram repetition rates (n = 2, 3, 4)</td>
<td>[\prod_{n=2}^{4} (1 - \text{Repetition})]</td>
</tr>
<tr>
<td>Language Modeling</td>
<td>Perplexity</td>
<td>Evaluation-set perplexity</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Coherence</td>
<td>LM quality (cosine similarity between sentence embeddings)</td>
<td></td>
</tr>
<tr>
<td>Divergence Curve</td>
<td>MAUVE</td>
<td>Quality &amp; diversity via the divergence frontiers</td>
<td>[C(H, M)] at all (\lambda \in (0, 1)) [31]</td>
</tr>
<tr>
<td>Frequency Domain</td>
<td>FACE (this work)</td>
<td>Quality &amp; diversity via the spectral similarities (four metrics)</td>
<td>[FACE(F_p, F_m)]</td>
</tr>
<tr>
<td>Human Judgment</td>
<td>Bradley-Terry Score</td>
<td>Human preference via the pairwise evaluation</td>
<td>[P(\text{wins}) = \frac{1}{\text{cont} + \text{pre} - 1}]</td>
</tr>
</tbody>
</table>

**Language modeling metrics** measure how un(certain) human text \(x \sim H\) follows the model distribution \(M\), using the probability distribution \(M(x)\). In our work, the perplexity is calculated upon the set of human texts to quantify how well the distribution \(M\) predicts a text continuation. Coherence is approximated by cosine similarity between the sentence embeddings of prompt \(x_{\text{pre}} \sim H\) and continuation \(x_{\text{cont}} \sim M\) as proposed in \[42\], where the embedding \(\text{EMB}(\cdot)\) is produced by the pre-trained SimCSE sentence embedding \[11\]. Metrics under this category never observe model-generated text samples, and hence, they cannot justify how likely \(x_{\text{cont}}\) is under the human distribution \(H\).

**Reference-based measures** assess the generated text with respect to a small set of reference text, rather than calculating over the full sequence distributions. Some recent reference-based approaches encompass: (1) \[4, 26, 38, 50\] aim to capture distributional semantic information in high-dimensional space; (2) \[51\] concerns Euclidean distance between vector representations of n-grams and their document frequencies; (3) \[31\] straightforwardly computes the similarity of one learned distribution from a text generation and the other distribution of human-written text using information divergence frontiers \[9, 23, 34\]. Reference-based metrics are well-suited for targeted generation tasks (e.g., machine translation). Nevertheless, they become unfavorable in the open-ended generation scenario where multiple reasonable and diverse continuations are preferred.

**Non-automatic metrics.** Recent works \[12, 31, 19, 26\] on evaluation metrics and decoding strategies for natural language generation rely on human judgments, assuming that human annotations are the gold standard. Considering the expense of Human Unified with Statistical Evaluation (HUSE) \[18\], we adopt a pairwise evaluation protocol based on human preferences, to serve as a non-automatic complement of FACE metrics. We leverage the Bradley-Terry model \[3\] to predict the outcome of a head-to-head comparison given \(n\) players with scores \(\beta_1, \ldots, \beta_n\).

### 4 Experiments

**Task formulation.** Given an input text passage as prefix, the open-ended generation aims to produce texts that form a fluent and coherent continuation. More formally, given a sequence of \(m\) tokens denoted \([x_1 \ldots x_m]\), as the **prompt**, the goal is to generate the next \(n\) **continuation** tokens to form a complete sequence \([x_1 \ldots x_{m+n}]\). The continuation probability at the decoding time by conditioning on the preceding context is defined as: \(P(x_{m+1} \ldots x_{m+n} | x_1 \ldots x_m) = \prod_{i=m+1}^{m+n} P(x_i | x_1 \ldots x_{i-1})\), where \(P(x_i | x_1 \ldots x_{i-1})\) is the next-token distribution.

**4.1 Model sizes**

We consider such a text completion task in three domains: Wiki text, News, and Stories. Intuitively, the generated texts involving different domain knowledge may have different language usages and writing style, which may reflect on metrics. We generate completions from large-scale language models (LMs). In particular, we adopt three representatives of state-of-the-art pre-trained autoregressive LMs: Generative Pre-trained Transformer 2 (GPT2) \[33\], Open Pre-trained Transformer LMs (OPT) \[49\], and BigScience Large Open-science Open-access Multilingual LM (BLOOM) \[35\]. We explore two sizes for each model to illustrate that our FACE metrics generalize across multiple LM families and sizes. Details regarding our task and input data are summarized in Table 2.
We applied the majority voting to determine the winner. OPT and BLOOM are postfixed with their values, the better, whereas ↓ indicates the larger the metric value, the better, whereas ↓ indicates the opposite. The vs. column indicates the better-performing model in each comparison, where L/S denotes the large/small model wins and E represents a tie.

In our investigations, the GPT2-xl model consistently outperforms its small counterpart among statistics-based and language modeling metrics as all relevant “vs.” columns indicate, apart from the Coherence in the Stories domain. In the GPT2 experimental group, it is astonishing that the small model always performs better when referring to the voting results from MAUVE and FACE rows.

models may generate vastly different numbers of continuations in each length interval (see Appendix). To ensure the fairness of investigating the correlation between FACE and other widely-used metrics with respect to different models (with different sizes), we compute the weighted arithmetic mean for every metric across five length intervals.

The evaluation metrics we are interested in are based on various motivations and principles. Specifically, MAUVE and FACE emphasize the parallels between human and machine-produced texts, as stated in Section 3. Therefore, we group MAUVE together with four FACE metrics. To further obtain intuitive results, we utilize the voting approach to explore the correlations between these metrics on large/small models across three task domains. The results are shown in Table 3.

In our investigations, the GPT2-xl model consistently outperforms its small counterpart among statistics-based and language modeling metrics as all relevant “vs.” columns indicate, apart from the Coherence in the Stories domain. In the GPT2 experimental group, it is astonishing that the small model always performs better when referring to the voting results from MAUVE and FACE rows. Across three task domains, the performances of OPT and BLOOM models in two sizes differ. Large models have better overall performance, and small models only win four out of twelve comparisons by voting. Nonetheless, it is noteworthy that four FACE metrics we proposed maintain a relatively high level of consistency with MAUVE across all models. At least two FACE metrics yield the same results (in eight out of nine sets of human-model language comparisons) with MAUVE. Concretely
speaking, $SO$ and $SAM$ show a higher positive correlation to MAUVE than $CORR$ and $SPEAR$, given that seven out of nine voting results (marked with yellow in Table 3) are identical.

To further evaluate model sizes, we apply FACE to the original GPT2 output data (webtext) generated from GPT2-sm and GPT2-xl. GPT2-xl has a higher $SO$ score than GPT2-sm, which is confirmed with the $t$-test, but non-significant effects are found on the other three metrics. Combining our generation task with the original GPT2 data, we illustrate the results for $SO$ in Figure 3.

To conclude, we discover three keypoints: (1) FACE is consistent with MAUVE in evaluating three different model types (two sizes for each); (2) the metrics estimating similarity between human-written and model-generated text generations (e.g., FACE, MAUVE) may produce opposite results to the text-centered metrics (e.g., Diversity, Coherence); (3) the four metrics of FACE show relatively homogeneous results, and using these metrics together helps to identify model-generated texts with a more comprehensive evaluation.

### 4.2 Sampling methods

Recent work [19, 26] has indicated three clear trends in open-ended text generation using auto-regressive LMs: (1) maximization-based decoding algorithms (e.g., beam search, greedy decoding, etc.) lead to copious repetition, while sampling with temperature may result in incoherence; (2) truncation-based sampling methods like nucleus sampling produce text with higher quality; (3) contrastive decoding outperform nucleus sampling in terms of both fluency and coherence. Accordingly, to demonstrate the effectiveness of our approach, FACE should follow the inequality: maximization-based/temperature-based $≺$ nucleus $≺$ contrastive in terms of the quality relationship.

Figure 4 visualizes the correlation between FACE scores and various decoding algorithms. The contrastive decoding approach yields the best performance among the four FACE metrics. It can be clearly observed that the maximization-based sampling methods behave worse than other algorithms. Moreover, adding the temperature parameter to top-$k$ sampling results in incoherent text generations.

---

[https://github.com/openai/gpt-2-output-dataset](https://github.com/openai/gpt-2-output-dataset)
Table 4: Results for comparing all sampling methods with selected parameters regarding the conditional generation. The values closest to human scores are bolded, except for our proposed FACE scores, where the highest (for SO, CORR, and SPEAR) or the lowest (for SAM) values are in bold.

<table>
<thead>
<tr>
<th>Sampling Method</th>
<th>Perplexity</th>
<th>Self-BLEU</th>
<th>Zipf Coefficient</th>
<th>Repetition</th>
<th>SO (↑)</th>
<th>CORR (↑)</th>
<th>SAM (↓)</th>
<th>SPEAR (↑)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>12.38</td>
<td>0.31</td>
<td>0.93</td>
<td>0.28</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Greedy</td>
<td>1.50</td>
<td>0.50</td>
<td>1.00</td>
<td>73.66</td>
<td>0.20</td>
<td>0.56</td>
<td>0.31</td>
<td>0.04</td>
</tr>
<tr>
<td>Beam (b=16)</td>
<td>1.48</td>
<td>0.44</td>
<td>0.94</td>
<td>28.94</td>
<td>0.21</td>
<td>0.31</td>
<td>0.40</td>
<td>0.04</td>
</tr>
<tr>
<td>Stochastic Beam (b=16)</td>
<td>19.20</td>
<td>0.28</td>
<td>0.91</td>
<td>32.32</td>
<td>0.37</td>
<td>0.49</td>
<td>0.33</td>
<td>0.04</td>
</tr>
<tr>
<td>Pure Sampling</td>
<td>22.73</td>
<td>0.28</td>
<td>0.93</td>
<td>0.22</td>
<td>0.41</td>
<td>0.63</td>
<td>0.28</td>
<td>0.03</td>
</tr>
<tr>
<td>Sampling (t=0.9)</td>
<td>10.25</td>
<td>0.35</td>
<td>0.96</td>
<td>0.66</td>
<td>0.42</td>
<td>0.61</td>
<td>0.29</td>
<td>0.03</td>
</tr>
<tr>
<td>Top-k (k=40)</td>
<td>6.88</td>
<td>0.39</td>
<td>0.96</td>
<td>0.78</td>
<td>0.40</td>
<td>0.64</td>
<td>0.28</td>
<td>0.03</td>
</tr>
<tr>
<td>Top-k (k=640)</td>
<td>13.82</td>
<td>0.32</td>
<td>0.96</td>
<td>0.28</td>
<td>0.42</td>
<td>0.63</td>
<td>0.28</td>
<td>0.03</td>
</tr>
<tr>
<td>Top-k (k=40, t=0.7)</td>
<td>3.48</td>
<td>0.44</td>
<td>1.00</td>
<td>8.86</td>
<td>0.34</td>
<td>0.61</td>
<td>0.29</td>
<td>0.03</td>
</tr>
<tr>
<td>Nucleus (p(t)=95)</td>
<td>13.13</td>
<td>0.32</td>
<td>0.95</td>
<td>0.36</td>
<td>0.42</td>
<td>0.63</td>
<td>0.28</td>
<td>0.03</td>
</tr>
<tr>
<td>Contrastive Decoding</td>
<td>14.39</td>
<td>0.54</td>
<td>1.04</td>
<td>0.24</td>
<td>0.44</td>
<td>0.75</td>
<td>0.23</td>
<td>0.17</td>
</tr>
</tbody>
</table>

which explains the gap between the red curve (top-k w/o temperature) and the gray curve (top-k w/ temperature). We also plot the correlation graphs of unconditional generation (in the Appendix) with fewer sampling methods involved. The trends and patterns in the visualization of unconditional generation are basically consistent with its conditional counterpart.

In Table F FACE scores on different decoding algorithms are summarized. FACE metrics correctly match the expected quality relationship of the sampling methods examined by assigning the best SO (.44), CORR (.75), SAM (.23), and SPEAR (.17) scores to contrastive decoding. Other evaluation metrics fail to capture the correct relationship, for example, the perplexity rates nucleus-sampled text as better than contrastive-decoded text, which is irrational suggested by Li et al. [26].

4.3 Human judgments

We also explore the correlation between FACE and human judgement scores, using the crowd-source dataset collected in [31] when human evaluation is available. The dataset contains model-generated continuations (by GPT2-sm, -md, -lg, and -xl with ancestral and nucleus sampling), human-written continuations using the same prefix, and the crowd-source workers’ answers on which completion is more human-like, interesting, and sensible. We follow the same experimental settings and protocol to verify whether the FACE scores of the text completions correlate well with the human quality judgements by computing the Spearman’s rank correlation coefficient. The results are presented in Table 5.

We observe a high and positive correlation between FACE-SO and human judgments scores, which outperforms five out of the six evaluation metrics reported in [31] and achieves a comparative performance against MAUVE. The remaining three FACE metrics have insignificant correlations. However, we consider human judgments to be subjective and sometimes biased. Including more fine-grained questions to perform human judgments may lead to more accurate correlation statistics. Additionally, we recomputed the correlations with human judgement scores to keep those pairs in which there are exactly one item from human and the other item from model (i.e., a subset of data used for the analysis in Table 5). As shown in SO-S and MAUVE-S columns, FACE-SO has a stronger correlation than MAUVE among two of these three dimensions.

4.4 Sanity tests

Sanity test on validity. We evaluate the validity of FACE by examining whether its scores on human-human split is lower than human-model groups – an expected result based on our assumption that the spectral difference between human’s “natural” language and models’ “artificial” ones should be amplified by FACE. Therefore, the sanity tests are conducted as follows: first, we evenly and randomly split the human data into two folds (across three domains) to serve as control groups. The FACE scores between these control folds are then computed. As for the human-to-model experimental group, we create other two folds using human data and the best model-generated data (from contrastive decoding [26]) in terms of text quality. If FACE can effectively capture the fundamental difference
### Table 5: Spearman’s rank correlation coefficients of SO and five other metrics with human judgments. Higher scores mean better correlation. All the numbers except the SO, SO-S, and MAUVE-S columns are sourced from [31]. “BT” denotes the Bradley-Terry score of the pairwise human evaluation, which is employed to compute the Spearman’s rank correlation with the scores of other metrics. Additionally, it is important to note that the original human judgments encompass certain pairs in which both texts are generated by models, albeit different models. Therefore, we refine the original human judgment dataset to only include judgments involving both human and model-generated languages, and the results are shown in SO-S and MAUVE-S.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Generation Perplexity</th>
<th>Zipf Coefficient</th>
<th>Repetition</th>
<th>Distinct-4</th>
<th>Self-BLEU</th>
<th>SO</th>
<th>MAUVE</th>
<th>SO-S</th>
<th>MAUVE-S</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human-like/BT</td>
<td>0.810</td>
<td>0.833</td>
<td>-0.167</td>
<td>0.738</td>
<td>0.595</td>
<td><strong>0.981</strong></td>
<td><strong>0.952</strong></td>
<td>0.357</td>
<td>0.214</td>
</tr>
<tr>
<td>Interesting/BT</td>
<td>0.643</td>
<td>0.524</td>
<td>-0.141</td>
<td>0.524</td>
<td>0.405</td>
<td><strong>0.762</strong></td>
<td><strong>0.810</strong></td>
<td>0.524</td>
<td>0.667</td>
</tr>
<tr>
<td>Sensible/BT</td>
<td>0.738</td>
<td>0.690</td>
<td>-0.071</td>
<td>0.595</td>
<td>0.524</td>
<td><strong>0.786</strong></td>
<td><strong>0.857</strong></td>
<td>0.995</td>
<td>0.706</td>
</tr>
</tbody>
</table>

It can be seen from Table 6 that the control groups show significantly better FACE scores than the experimental group: FACE-SO and FACE-CORR are higher in human-to-human folds, while FACE-SAM scores are lower. The only exception is FACE-SPEAR, while we will show it is a good metric in later sections. Nonetheless, these tabulated results have proved the validity of FACE in effectively capturing human-to-model spectral differences.

### Table 6: Results of sanity test on FACE’s validity. The top three rows are the control groups, and “h-h” stands for human-to-human folds. The last row is the experimental group, where “h-m” is for human-to-model fold. Better FACE scores are in bold. The scores in the bottom row are retrieved from Table 4.

<table>
<thead>
<tr>
<th></th>
<th>SO (%)</th>
<th>CORR (%)</th>
<th>SAM (%)</th>
<th>SPEAR (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>h-h (wiki)</td>
<td>0.45</td>
<td>0.76</td>
<td>0.22</td>
<td>0.05</td>
</tr>
<tr>
<td>h-h (news)</td>
<td>0.45</td>
<td>0.76</td>
<td>0.22</td>
<td>0.07</td>
</tr>
<tr>
<td>h-h (stories)</td>
<td>0.47</td>
<td>0.79</td>
<td><strong>0.21</strong></td>
<td>0.05</td>
</tr>
<tr>
<td>h-m</td>
<td>0.44</td>
<td>0.75</td>
<td>0.23</td>
<td><strong>0.17</strong></td>
</tr>
</tbody>
</table>

### 4.5 Interpretation of spectra

As the frequency spectrum reflects the key characteristics of a signal, we attempt to interpret the spectra to see if they tell how the “signals” – entropy of human and machine languages – differ. Without aggregation, the raw spectra of single cross-entropy sequence look indistinguishable between GPT2-sm, GPT2-xl, and human (see the left plot in Figure 5). By aggregating 5,000 spectra from each group and smoothing the curves, it can be seen that GPT2-xl’s curve is closer to human than the GPT2-sm curve (readers can find this by zooming in the middle plot in Figure 5). Here, the smoothing is done with generalized additive models (GAMs) [45]. Results from other models are included in the Appendix.
When plotted separately, the aggregated spectra from human and different models have similar shapes: First, the majority of components exist in the low-frequency range \((\omega < 0.05)\). In addition, the locations of peaks and troughs are almost the same between groups. For instance, \(\omega_1 = 0.06\) is the first trough, and \(\omega_2 = 0.12\) is the first peak (see the right plots in Figure 5). Thus, roughly speaking, the main difference between human and model spectra is not in the locations of peak and trough frequencies but in the relative magnitudes of those frequencies.

We propose a simple way to interpret the peaks in spectra: the reciprocal of a frequency component \(T_k = 1/\omega_k\) denotes the corresponding cycle in the time domain. Because the time interval (i.e., sampling interval) of an entropy sequence is not measured in seconds but fixed as one token, the measurement unit of \(T_k\) is also in number of tokens. For example, the first frequency peak in Figure 5 (right plot) implies \(\omega_2 = 0.12 \Rightarrow T_2 = 1/0.12 \approx 8.3\) (tokens), which approximately means that tokens of the same cross-entropy levels tend to recur every 8.3 tokens. This pattern is consistent in both human and model data. However, the degree of this recurrence can mark the difference between the human and model languages. We leave more detailed interpretations of spectra to future work.

5 Conclusion and Limitations

We propose FACE, a set of metrics based on the Fourier analysis of cross-entropy, which is able to distinguish human and model-generated language with satisfactory performance in the open-ended generation task. The metrics scale with model sizes; reflect the effect of various sampling methods; correlate well with other existing metrics and outperform most of them in alignment with human judgement scores. Among the four implementation methods of FACE experimented, Spectral Overlap (SO) has the best overall performance.

FACE is computationally efficient with easy-to-interpret output. As a method inspired by psycholinguistic studies on the predictability (entropy/surprisal/information density) of human language, we believe FACE is a good example of incorporating knowledge from different fields for better human-centered AIs. We can generally conclude that better language models can produce spectral representations of information that are more similar to human.

Our current work has several limitations: Firstly, for open-ended generation experiments (Section 4.1), a broader set of sampling methods other than top-\(k\) can be used. Secondly, larger models (with more than 100 billion parameters) need to be included for more comprehensive comparisons. We will improve from these aspects in future work.

Acknowledgement

This material is based upon work supported by the National Science Foundation under Grant No. (2105192).
References


Appendix

1 Broader Impacts

FACE measures the distance between human and model-generated languages, therefore it is technically possible to be used for designing or augmenting systems that mimic humans. We acknowledge the risks of FACE (and other metrics) being utilized in applications that deliberately confuse human-authored and model-produced text. We call for the collective efforts from the community to come up with a systematic framework that unifies different metrics, for developing more reliable and natural language generation systems.

2 Implementation Details

**Preprocessing.** We utilize three raw datasets: WritingPrompts, WikiText-103, and RealNews. For WritingPrompts, the prompt set has already been well-curated, so we just extracted the first 5,000 prompts (the length may vary) for our generation task. WikiText-103 and RealNews contain many complete texts. For each complete text, we further truncate it corresponding to the first 35 tokens as a prompt. To fairly evaluate the performance of metrics, we also divide text generations according to five predefined length (from 0 up to 1024) intervals for each dataset. Thereby, the human-written texts and model-produced texts used to evaluate the performance of metrics may be generated by different prompts (i.e., unpaired comparison).

**Hyper-parameters.** We have several hyper-parameters during the text generation and evaluation phases. For both conditional and unconditional generation, we preset a random seed integer (32 by default). Furthermore, the maximum length of each text (1024 by default) as well as the batch size (which varies according to GPUs capacity) for perplexity computation have to be determined before automatic evaluation.

3 Miscellaneous Details

**Software.** Our experiments were performed on Ubuntu 20.04.1 system with Python 3.9.16. The versions of key Python libraries include: Transformers 4.27.4, PyTorch-CUDA 11.6, PyTorch 1.13.1, Scipy 1.5.4.

**Hardware.** For the text generation task, we use the remote workstation that has two NVIDIA RTX A6000 graphics cards. It should be noted that all models were run in parallel when available.

**Computation time for text generation.** We spent 10 and 25 hours or so obtaining 5,000 text continuations by GPT2-sm, -xl, respectively. OPT-125m, -6.7b cost our GPU resources roughly 11 and 44 hours to output the same number of text continuations, respectively. When it comes to BLOOM-560m, -7b, they took approximately 18 and 48 hours, respectively, to generate 5,000 continuations per task domain.

**Evaluation time for FACE.** Computation time of four FACE metrics for a single pair of references are: $5.96 \times 10^{-8}$ seconds for $SO$, $5.01 \times 10^{-8}$ seconds for $CORR$, $4.53 \times 10^{-8}$ seconds for $SAM$, and $4.29 \times 10^{-8}$ seconds for $SPEAR$, respectively. The cross-entropy, which should be calculated beforehand, takes $5.65 \times 10^{-2}$ seconds. All of the above measurements take place on an AMD Ryzen Threadripper PRO 3995WX 64-Cores CPU (frequency range ∈ [2200.00MHz, 4308.40MHz]). Users can leverage more advanced GPU resources to perform the whole computation process with a faster speed.

4 Additional Experimental Results

4.1 Model sizes (generation length)

It should be emphasized that LMs have diverse designs and were pre-trained using different strategies on different datasets, giving them distinct preferences on the generation length. The numbers of text generations in each length interval are summarized in Table 7.
### Table 7: Domain-specific generation length with respect to different models (GPT2/OPT/BLOOM) and model sizes (one large model and one small model) using top-\(k\) (\(k = 50\)) sampling corresponding to five continuous length intervals.

<table>
<thead>
<tr>
<th>Domain</th>
<th>Length Interval</th>
<th>GPT2-sm</th>
<th>GPT2-xl</th>
<th>OPT-125m</th>
<th>OPT-6.7b</th>
<th>BLOOM-560m</th>
<th>BLOOM-7b</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wiki text</td>
<td>0-200</td>
<td>403</td>
<td>485</td>
<td>964</td>
<td>1522</td>
<td>4928</td>
<td>803</td>
</tr>
<tr>
<td></td>
<td>201-400</td>
<td>571</td>
<td>672</td>
<td>888</td>
<td>929</td>
<td>61</td>
<td>599</td>
</tr>
<tr>
<td></td>
<td>401-600</td>
<td>251</td>
<td>316</td>
<td>441</td>
<td>417</td>
<td>8</td>
<td>388</td>
</tr>
<tr>
<td></td>
<td>601-800</td>
<td>260</td>
<td>310</td>
<td>268</td>
<td>285</td>
<td>1</td>
<td>316</td>
</tr>
<tr>
<td></td>
<td>801-1024</td>
<td>3515</td>
<td>3217</td>
<td>2439</td>
<td>1847</td>
<td>2</td>
<td>2894</td>
</tr>
<tr>
<td>News</td>
<td>0-200</td>
<td>750</td>
<td>836</td>
<td>844</td>
<td>1119</td>
<td>4978</td>
<td>1371</td>
</tr>
<tr>
<td></td>
<td>201-400</td>
<td>1222</td>
<td>1336</td>
<td>1220</td>
<td>1325</td>
<td>20</td>
<td>917</td>
</tr>
<tr>
<td></td>
<td>401-600</td>
<td>824</td>
<td>759</td>
<td>1194</td>
<td>939</td>
<td>1</td>
<td>628</td>
</tr>
<tr>
<td></td>
<td>601-800</td>
<td>584</td>
<td>678</td>
<td>764</td>
<td>593</td>
<td>0</td>
<td>427</td>
</tr>
<tr>
<td></td>
<td>801-1024</td>
<td>1620</td>
<td>1391</td>
<td>978</td>
<td>1024</td>
<td>1</td>
<td>1657</td>
</tr>
<tr>
<td>Stories</td>
<td>0-200</td>
<td>549</td>
<td>745</td>
<td>2731</td>
<td>3588</td>
<td>4924</td>
<td>1608</td>
</tr>
<tr>
<td></td>
<td>201-400</td>
<td>625</td>
<td>757</td>
<td>715</td>
<td>501</td>
<td>63</td>
<td>688</td>
</tr>
<tr>
<td></td>
<td>401-600</td>
<td>296</td>
<td>404</td>
<td>241</td>
<td>176</td>
<td>9</td>
<td>410</td>
</tr>
<tr>
<td></td>
<td>601-800</td>
<td>241</td>
<td>324</td>
<td>160</td>
<td>95</td>
<td>4</td>
<td>271</td>
</tr>
<tr>
<td></td>
<td>801-1024</td>
<td>3289</td>
<td>2770</td>
<td>1153</td>
<td>640</td>
<td>0</td>
<td>2023</td>
</tr>
</tbody>
</table>

Figure 6: FACE scores of GPT2 (our generated data), OPT, and BLOOM with different model sizes.

To ensure the consistency of our experiments, we run six LMs separately (using their own tokenizers) with the same prompt sets and settings as described in Table 2 to generate 5,000 pieces of continuations for each interval, which allows us to compare FACE scores with other metrics more objectively, as we believe it is unfair to explicitly compare texts of varying lengths. Then, we compute weighted arithmetic mean to evaluate a model in each domain, by

\[ s' = \frac{\sum_{i=1}^{n} m_i s_i}{\sum_{i=1}^{n} m_i}, \]

where \( s' \) denotes the weighted mean; \( n \) denotes the number of length intervals; \( m_i \) is the number of generated continuations in the length interval \( i \); \( M = \sum_{i=1}^{n} m_i \), and \( s_i \) means a certain metric value in the interval \( i \).

Figure 6 conveys a more intuitive representation (via bar plots) of Table 3.

### 4.2 Sampling methods (unconditional generation)

We also carried out experiments on unconditional text generation. Here, the prompt is not required as we generate continuations from a random seed (set to 32 empirically). Four sampling methods, which are greedy decoding, beam search, stochastic beam search, and contrastive decoding, are not involved in this set of experiments.

The results are displayed in Figure 7. The overall trends are same as its conditional counterpart, where the previous quality relationship (maximization-based/temperature-based\( \prec \) nucleus\( \prec \) contrastive) is satisfied. Yet, it is crucial to note that the advantages of top-\(k\) sampling w/o temperature become more obvious compared to the conditional case.

https://github.com/ari-holtzman/degen
Figure 7: FACE scores (unconditional generation) on original experimental data. Five sampling (decoding) methods are compared: pure sampling, temperature, top-\textit{k}, top-\textit{k} with temperature, and nucleus. Note that logarithmic normalization on parameter values as well as an enlarged marker for pure sampling are adopted for better visualization.

<table>
<thead>
<tr>
<th>Model</th>
<th>Sampling Method (parameter)</th>
<th>SO</th>
<th>CORR</th>
<th>SAM</th>
<th>SPEAR</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPT2-xl</td>
<td>Nucleus Sampling (p=0.95)</td>
<td>0.481</td>
<td>0.821</td>
<td>0.191</td>
<td>0.359</td>
</tr>
<tr>
<td></td>
<td>Ancestral Sampling</td>
<td>0.472</td>
<td>0.807</td>
<td>0.199</td>
<td>0.331</td>
</tr>
<tr>
<td>GPT2-lg</td>
<td>Nucleus Sampling (p=0.95)</td>
<td>0.480</td>
<td>0.819</td>
<td>0.193</td>
<td>0.356</td>
</tr>
<tr>
<td></td>
<td>Ancestral Sampling</td>
<td>0.472</td>
<td>0.814</td>
<td>0.196</td>
<td>0.338</td>
</tr>
<tr>
<td>GPT2-md</td>
<td>Nucleus Sampling (p=0.9)</td>
<td>0.478</td>
<td>0.815</td>
<td>0.194</td>
<td>0.358</td>
</tr>
<tr>
<td></td>
<td>Ancestral Sampling</td>
<td>0.462</td>
<td>0.813</td>
<td>0.197</td>
<td>0.310</td>
</tr>
<tr>
<td>GPT2-sm</td>
<td>Nucleus Sampling (p=0.9)</td>
<td>0.476</td>
<td>0.817</td>
<td>0.194</td>
<td>0.359</td>
</tr>
<tr>
<td></td>
<td>Ancestral Sampling</td>
<td>0.468</td>
<td>0.816</td>
<td>0.195</td>
<td>0.319</td>
</tr>
</tbody>
</table>

Table 8: FACE results based on MAUVE’s original experimental data.

4.3 Human judgments

Table 8 shows the FACE scores based on the output texts from MAUVE. Each column of FACE scores is used to compute the Spearman’s rank correlation coefficient between a specific FACE metric and Bradley-Terry scores (4 model sizes × 2 sampling methods = 8 scores in total) from one criterion (three criteria correspond to three questions in total).

4.4 Choice of estimator model

We examine how different choices of estimator model \( m_{est} \) affect the resulting spectra of cross-entropy. Five input data sources are examined (webtext plus four GPT2 original output datasets), on which four different estimator models are applied: \( m_{est} \in \{ \text{GPT2-sm, GPT2-md, GPT2-lg, GPT2-xl} \} \), resulting in \( 5 \times 4 = 20 \) aggregated spectra curves in Figure 8. It can be found that on the same input data, the spectra from four estimators largely overlap. It indirectly suggests that FACE should be stable across different \( m_{est} \). We leave the full inspection for future work.

4.5 Intuitive interpretation of spectra

As pointed out in Section 4.5, the aggregated spectral shapes from human and different models are nearly identical. A set of higher resolution plots from GPT-xl, OPT, BLOOM and human (webtext) are shown in Figure 9. It can be seen that although the \( X(\omega_k) \) has different ranges on y-axis, the x coordinates of the peaks and troughs are the same.

4.6 Corner Cases

Two examples highlight the difference between our proposed FACE-SO and MAUVE in their ability to recognize human-generated and model-generated texts. In the filtered datasets for human judgments, the average values for FACE-SO and MAUVE are 0.4738 and 0.9549, respectively. In Case 1, human evaluators noted a high level of similarity between the model-generated text and human text, resulting
in ties for human-like, interesting, and sensible aspects. However, the MAUVE score in Case 1 is lower than the average value, while the FACE-SO score surpasses its mean. This discrepancy suggests that SO aligns more consistently with human opinions. Conversely, in Case 2, human judgement indicates a significant dissimilarity between the model-generated text and human text, making them easily distinguishable. However, the MAUVE score exceeds its mean, suggesting the two texts are similar to each other. Our FACE-SO score is lower than its mean, indicating better alignment with human opinion.
Figure 10: Example of two corner cases. For each case, the prompt text, model-generated text, human text, MAUVE and FACE-SO scores, as well as the results from human judgments are tabulated.