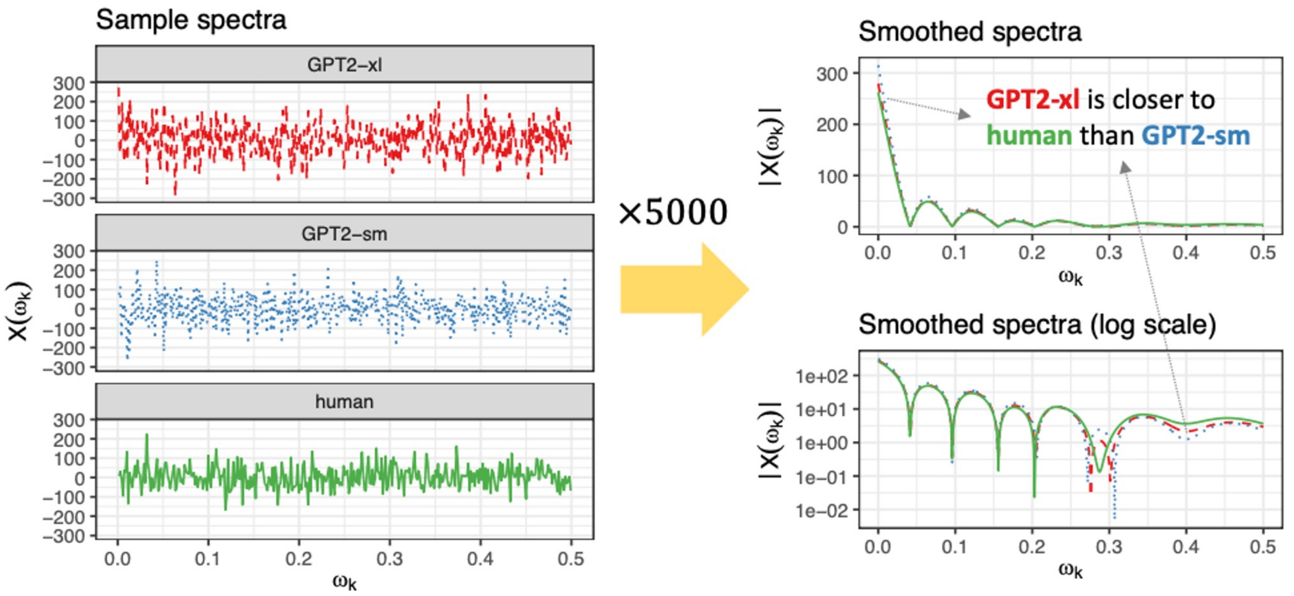
FACE: Evaluating Natural Language Generation with Fourier Analysis of Cross-Entropy

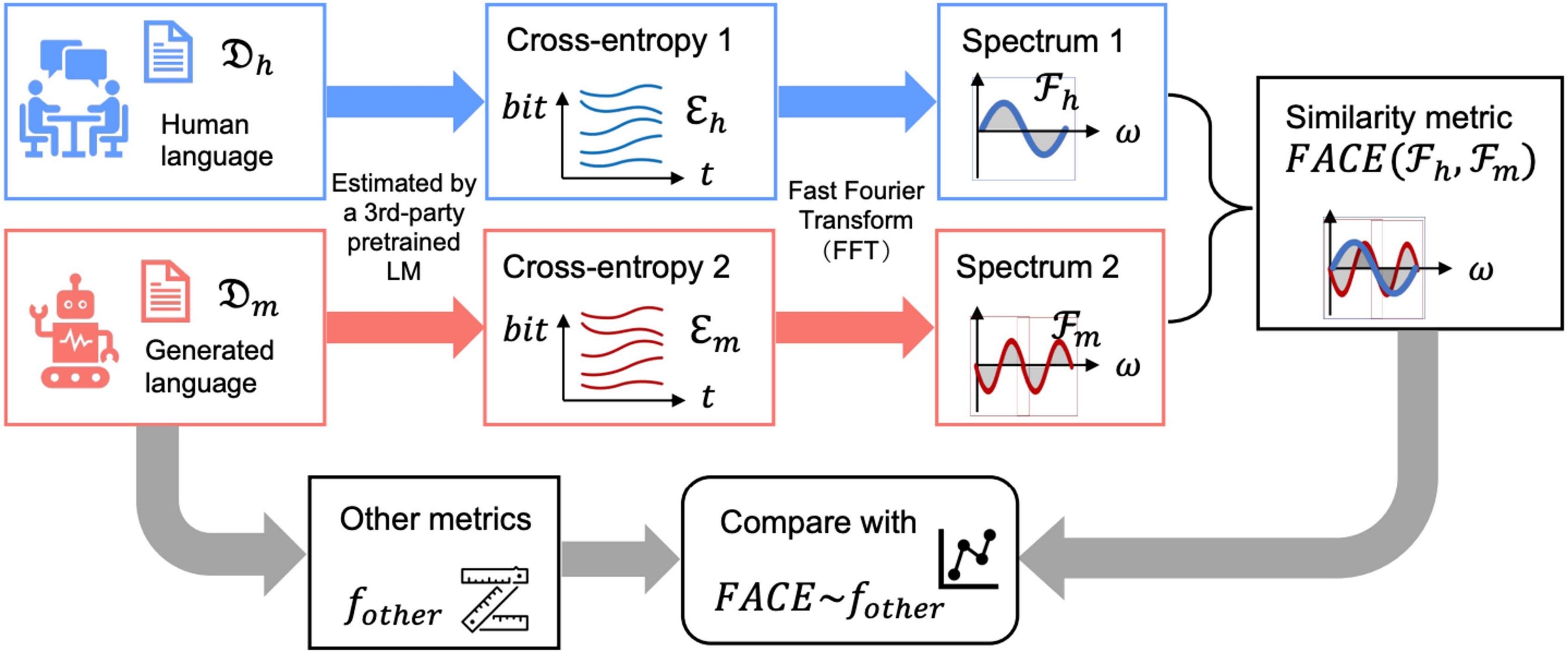
Zuhao Yang Yingfang Yuan Yang Xu
Shuo Zhan Huajun Bai Kefan Chen

SUSTech Southern University of Science and Technology
NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE
HERIOT WATT UNIVERSITY
Genify.ai

Motivation

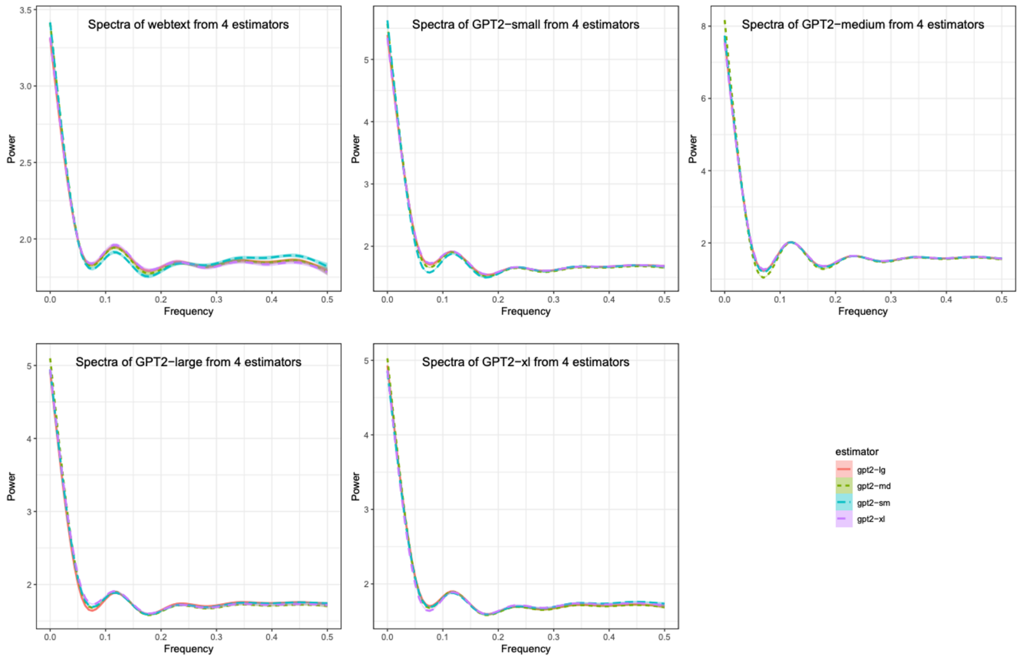

Sample spectra
GPT2–xl
GPT2–sm
human
$X(\omega_k)$
300
200
100
0
−100
−200
−300
0.0
0.1
0.2
0.3
0.4
0.5
ω_k
×5000
Smoothed spectra
$|X(\omega_k)|$
GPT2-xl is closer to
human than GPT2-sm
Smoothed spectra (log scale)
1e+02
1e+01
1e+00
1e−01
1e−02

Overall workflow



Key step 1: Cross-entropy Estimation

$$\mathcal{E} = [c_1, c_2, \ldots, c_{T-1}] \triangleq [-\log P(t_2|t_1), -\log P(t_3|t_1, t_2), \ldots, -\log P(t_T|t_1, t_2, \ldots, t_{T-1})] \quad (1)$$



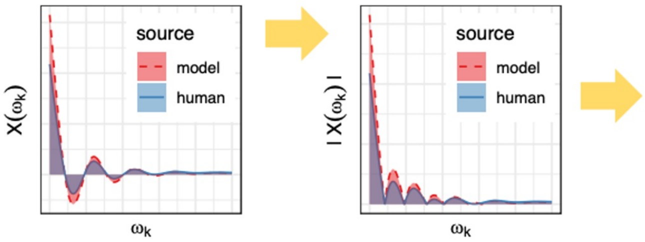
Key step 2. Fast Fourier Transform

$$X(\omega_k) \triangleq \sum_{n=0}^{N-1} x(t_n)e^{-j\omega_k t_n}, \; k = 0, 1, \ldots, N-1 \tag{2}$$

Key step 3. Spectral Similarity Metrics

Spectra:
$\mathcal{F}_h \in \mathbb{R}^{N_1}, \mathcal{F}_m \in \mathbb{R}^{N_2}$

Interpolated & absolute:
$\mathcal{F}'_h, \mathcal{F}'_m \in \mathbb{R}^{N_C}$



Four **FACE** metrics

$$\boxed{SO} = \frac{AUC(|\mathcal{F}'_h| \cap |\mathcal{F}''_m|)}{AUC(|\mathcal{F}'_h| \cup |\mathcal{F}'_m|)}$$

$$\boxed{SAM} = arccos\left(\frac{\mathcal{F}'_h \cdot \mathcal{F}'_m}{\sigma(\mathcal{F}'_h)\sigma(\mathcal{F}'_m)}\right)$$

$$\boxed{CORR} = \frac{cov(\mathcal{F}'_h, \mathcal{F}'_m)}{\sigma(\mathcal{F}'_h)\sigma(\mathcal{F}'_m)}$$

$$\boxed{SPEAR} = 1 - \frac{6\sum(|X_h(\omega_i)| - |X_m(\omega_i)|)^2}{N_C(N_C^2 - 1)}$$

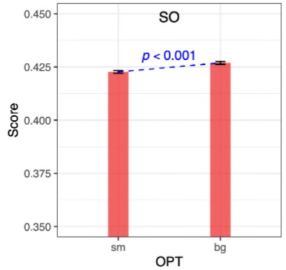
Experimental results - Model sizes

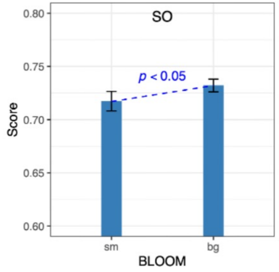
- Task: Open-ended text generation
- Three domains: wiki, news, stories
- Models tested:
  - GPT2: small, medium, large, x-large
  - BLOOM: 560m, 7b
  - OPT: 125m, 6.7b

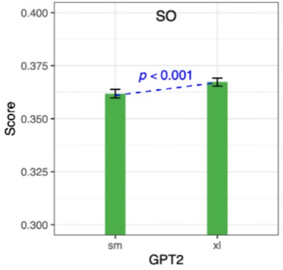
Domain	Model	Dataset	Prompt Length	Maximum Generation Length	Number of Generations
Wiki text	GPT2/OPT/BLOOM	WikiText-103	35 tokens	1024 tokens	5000
News	GPT2/OPT/BLOOM	RealNews	35 tokens	1024 tokens	5000
Stories	GPT2/OPT/BLOOM	WritingPrompts	varying	1024 tokens	5000

Model sizes (cont.)

Domain	Metric	GPT2-sm	GPT2-xl	*vs.*	Voting	OPT-125m	OPT-6.7b	*vs.*	Voting	BLOOM-560m	BLOOM-7.1b	*vs.*	Voting
Wiki text	Diversity (↑)	0.733	0.753	L		0.645	0.789	L		0.533	0.732	L	
	Coherence (↑)	0.595	0.624	L	L	0.614	0.634	L	L	0.926	0.819	S	E
	Zipf Coefficient (↓)	0.990	0.975	L		0.989	1.016	S		1.092	0.980	L	
	Self-BLEU (↓)	0.459	0.424	L		0.423	0.379	L		0.280	0.422	S	
	MAUVE (↑)	0.677	0.186	**S**		0.169	0.265	**L**		0.517	0.184	S	
	SO (↑)	0.414	0.406	**S**		0.424	0.436	**L**		0.426	0.432	L	
	CORR (↑)	0.806	0.781	S	S	0.771	0.769	S	L	0.675	0.789	L	L
	SAM (↓)	0.199	0.213	**S**		0.216	0.217	S		0.258	0.208	L	
	SPEAR (↑)	0.022	0.023	L		0.026	0.029	L		0.059	0.023	S	
News	Diversity (↑)	0.890	0.897	L		0.853	0.876	L		0.740	0.870	L	
	Coherence (↑)	0.613	0.640	L	L	0.663	0.663	S	L	0.897	0.785	S	S
	Zipf Coefficient (↓)	0.961	0.958	L		0.965	0.968	L		0.964	0.966	S	
	Self-BLEU (↓)	0.619	0.573	L		0.611	0.543	L		0.384	0.501	S	
	MAUVE (↑)	0.393	0.281	**S**		0.162	0.130	**S**		0.014	0.095	**L**	
	SO (↑)	0.424	0.412	**S**		0.438	0.440	L		0.436	0.437	**L**	
	CORR (↑)	0.757	0.723	S	S	0.746	0.732	S	S	0.615	0.733	S	L
	SAM (↓)	0.224	0.240	**S**		0.229	0.236	**S**		0.281	0.234	**L**	
	SPEAR (↑)	0.021	0.019	S		0.017	0.021	L		0.048	0.019	S	
Stories	Diversity (↑)	0.743	0.785	L		0.769	0.875	L		0.527	0.830	L	
	Coherence (↑)	0.421	0.420	S	L	0.440	0.388	S	L	0.880	0.660	S	S
	Zipf Coefficient (↓)	1.097	1.085	L		1.021	1.003	L		0.999	1.058	S	
	Self-BLEU (↓)	0.617	0.565	L		0.587	0.511	L		0.180	0.455	S	
	MAUVE (↑)	0.504	0.121	**S**		0.025	0.013	**S**		0.006	0.008	**L**	
	SO (↑)	0.411	0.402	**S**		0.406	0.405	**S**		0.350	0.418	**L**	
	CORR (↑)	0.813	0.787	S	S	0.737	0.705	S	S	0.573	0.772	L	L
	SAM (↓)	0.195	0.209	**S**		0.231	0.245	**S**		0.300	0.214	**L**	
	SPEAR (↑)	0.023	0.022	S		0.036	0.041	L		0.050	0.027	S	







Experimental results - Sampling methods

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

Experimental results - Human judgments

Metric	Generation Perplexity	Zipf Coefficient	Repetition	Distinct-4	Self-BLEU	*SO*	MAUVE	*SO*-S	MAUVE-S
Human-like/BT	0.810	0.833	−0.167	0.738	0.595	0.881	**0.952**	**0.357**	0.214
Interesting/BT	0.643	0.524	−0.143	0.524	0.405	0.762	**0.810**	0.524	**0.667**
Sensible/BT	0.738	0.690	−0.071	0.595	0.524	0.786	**0.857**	**0.995**	0.706

- We examined the correlation between FACE and human judgment scores, using data collected from MAUVE's paper (Pillutla et al., 2021)
- **FACE-SO** (spectral overlap) has high correlation with human judgments.
- SO has higher correlations than MAUVE on 2 out 3 dimensions (on the subset of data strictly comparing human vs. model text)

Conclusion and Limitations

- FACE – Metrics for NLG based on the Fourier analysis of cross-entropy
- FACE can distinguish human and model-generated language with good performance in open-ended generation tasks.
- FACE is computationally efficient with easy-to-interpret output.
- FACE also carries intuitive cognitive meanings of language, that is, better language models should produce similar spectral representations as human, which reflects the cognitive load of language processing.
- More models/generation tasks will be tested in the future.
- Code and data are available at: https://github.com/CLCS-SUSTech/FACE