MAUVE: Measuring the Gap Between Neural Text and Human Text using Divergence Frontiers

Колесников Георгий
Аюпов Шамиль
Котельников Аким

MAUVE

- Вводим MAUVE, меру сравнения между текстом нейронной сети и человеческим текстом.
- Эмпирически показываем, что MAUVE может более правильно и с меньшими ограничениями определять известные свойства сгенерированного текста, чем существующие метрики.
- С помощью человеческой оценки мы обнаруживаем, что MAUVE лучше коррелирует с человеческими суждениями о качестве текста.
- Наконец, обнаруживаем, что MAUVE может быть очень устойчивым к выбору квантования, встраивания и масштабирования.

Как устроена Open-Ended генерация текста?

Имеем текст

$$\boldsymbol{x} = (x_1, \ldots, x_{|\boldsymbol{x}|})$$

На нем модель имеет вероятностное распределение

$$\hat{P}(x_{t+1} | \boldsymbol{x}_{1:t})$$

Задача генерации - выдать:

$$\hat{\boldsymbol{x}}_{t+1:|\boldsymbol{x}|}$$

Два типа ошибок

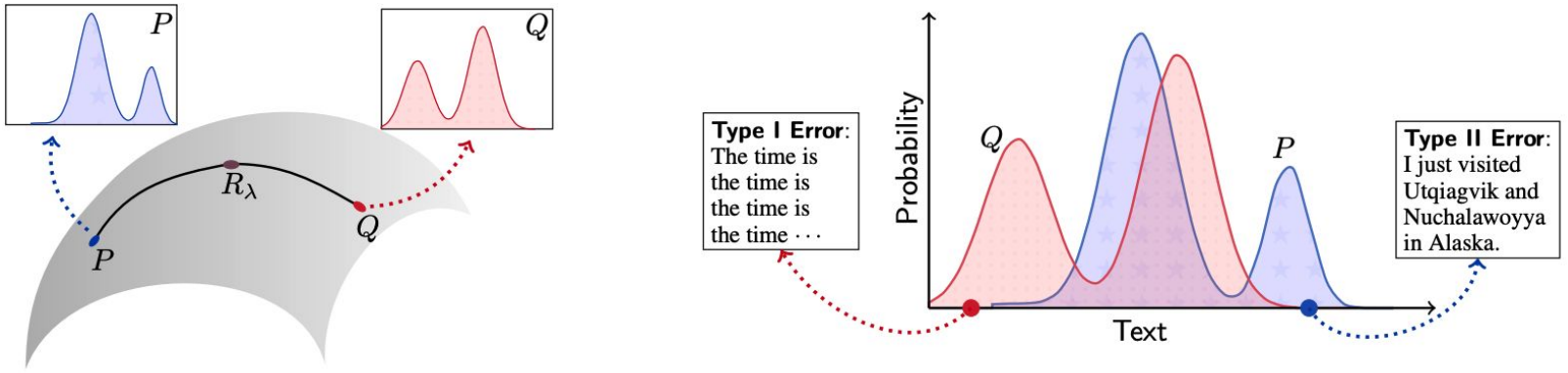

Figure 1: **Left**: MAUVE compares the machine text distribution Q to that of human text P by using the family of mixtures $R_\lambda = \lambda P + (1-\lambda)Q$ for $\lambda \in (0, 1)$. **Right**: Illustration of *Type I errors*, where Q produces degenerate, repetitive text which is unlikely under P, and, *Type II errors*, where Q cannot produce plausible human text due to truncation heuristics [26]. MAUVE measures these errors softly, by using the mixture distribution R_λ. Varying λ in $(0, 1)$ gives a divergence curve and captures a spectrum of soft Type I and Type II errors. MAUVE summarizes the entire divergence curve in a single scalar as the area under this curve.

Определение ошибки

Введем $\mathrm{KL}(Q|P)$ и $\mathrm{KL}(P|Q)$.

Тогда первая дивергенция штрафует модель, если есть x, что Q(x) велико, а P(x) мало - это первый тип ошибки, и наоборот второй.

Не до конца подходит, так как если области определения не идентичны, одна или обе могут быть равны бесконечности. Тогда вводим:

$$R_\lambda = \lambda P + (1-\lambda)Q$$

$\mathrm{KL}(Q|R_\lambda)$ — первый тип ошибки

$\mathrm{KL}(P|R_\lambda)$ — второй тип ошибки

Сравнение моделей и декодеров

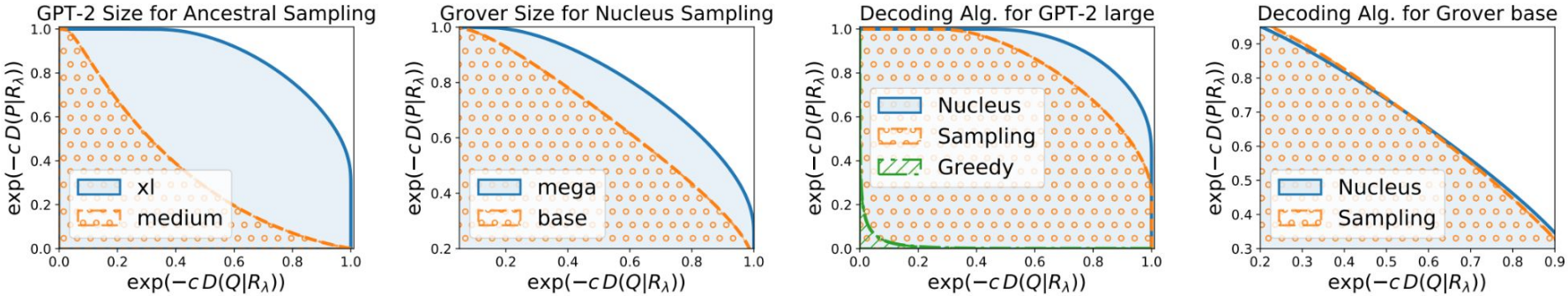

Figure 2: Divergence curves for different models (GPT-2 [45], Grover [61]) and decoding algorithms (greedy decoding, ancestral and nucleus sampling). MAUVE is computed as the area of the shaded region, and larger values of MAUVE indicate that Q is closer to P. In general, MAUVE indicates that generations from larger models and nucleus sampling are closer to human text. **Rightmost**: Nucleus sampling has a slightly smaller Type I error than ancestral sampling but a higher Type II error, indicating that ancestral sampling with Grover base produces more degenerate text while nucleus sampling does not effectively cover the human text distribution.

Подсчет с Монте-Карло

$$\mathcal{C}(P, Q) = \left\{ \left(\exp(-c\,\mathrm{KL}(Q|R_\lambda)), \exp(-c\,\mathrm{KL}(P|R_\lambda)) \right) \,:\, R_\lambda = \lambda P + (1-\lambda)Q,\ \lambda \in (0,1) \right\},$$

Computing MAUVE for Open-Ended Text Generation. Each point on the divergence curve $\mathcal{C}(P, Q)$ consists of a coordinate

$$\mathrm{KL}(P|R_\lambda) = \sum_{\boldsymbol{x}} P(\boldsymbol{x}) \log \frac{P(\boldsymbol{x})}{R_\lambda(\boldsymbol{x})}, \tag{2}$$

- Берем человеческий и компьютерный тексты.
- Через внешнюю модель M получаем их эмбединги.
- С помощью k-means находим приближения истинных распределений

$$\tilde{P}(j) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\left(\phi(\boldsymbol{x}_i) = j\right),$$

Квантизация

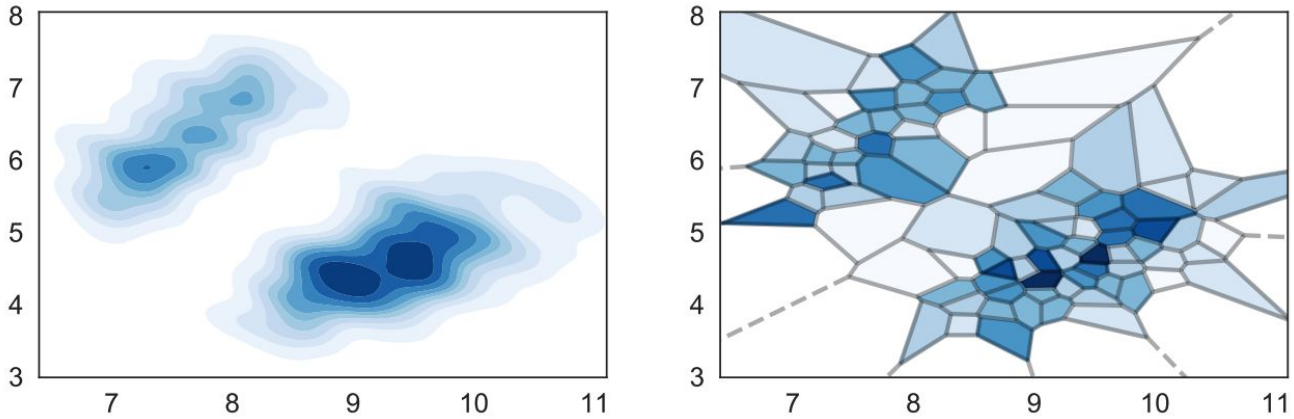

Figure 3: Illustration of the quantization. **Left**: A continuous two-dimensional distribution P. **Right**: A partitioning of the Euclidean plane $\mathbb{R}^2$ and the corresponding quantized distribution $\tilde{P}$.

Если коротко

MAUVE - площадь под кривой дивергенции, суммирующая все ошибки первого и второго типа с помощью эффективного приближения, созданного для генерации текста.

Другие методы оценки

Type	Metric	Measures	Approximates
Statistics	Zipf Coefficient [26]	Unigram rank-frequency statistics	–
	Self-BLEU [65]	N-gram diversity	–
	Generation Perplexity [18]	Generation quality via external model R	$\lvert\mathbb{E}_Q[\log R(\boldsymbol{x})] - \mathbb{E}_P[\log R(\boldsymbol{x})]\rvert$ (a single point inside $\mathcal{C}(P, Q)$)
Language Modeling	Perplexity	Test-set perplexity	$\mathbb{E}_P[\log Q(\boldsymbol{x})]$
	ε-perplexity [39]	Perplexity w/ Laplace smoothing	$\mathbb{E}_P[\tilde{Q}(\boldsymbol{x})]$
	Sparsemax Score [39]	LM quality (sparsemax loss [38])	$\mathbb{E}_P[\tilde{Q}(\boldsymbol{x})]$
	Token JS-Div. [39]	LM quality (JS divergence)	$\mathbb{E}_P[\tilde{Q}(\boldsymbol{x})]$
Divergence Curve	MAUVE (this work)	Quality & diversity via the divergence curve	$\mathcal{C}(P, Q)$ at all λ

Table 1: Summary of automatic distributional metrics for evaluating open-ended text generation. MAUVE provides a summary of all points along the divergence curve, rather than a single point. The summary is based on comparisons in a joint embedding space, rather than a statistic computed independently on each distribution. $\tilde{Q}$ informally refers to a quantity related to Q.

Эксперименты

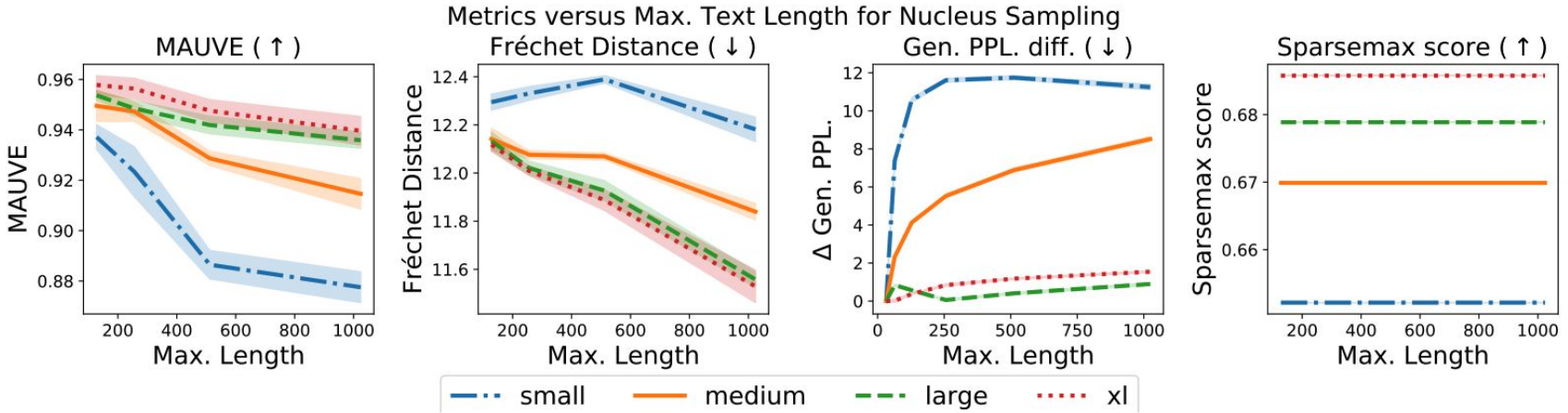

Figure 4: Generation quality versus maximum generation length according to MAUVE and three alternative measures (web text, GPT-2). MAUVE is the only comparison measure which identifies that generation quality decreases monotonically with increasing text length. The shaded area shows one standard deviation over generations from 5 random seeds.

На этих графиках наблюдается два преимущества: Чем больше длина сгенерированного, тем хуже оценка (с длиной текста качество в реальности падает), чем больше модель, тем лучше показатели.

Цифры

	Adv.	Greedy	Sampling	Nucleus
Gen. PPL(↓)	**0.05**	11.3	19.3	1.54
Zipf(↓)	0.03	0.02	0.02	**0.01**
Self-BLEU(↓)	0.07	0.03	**0.02**	0.03
SP(↑)	–	0.50	**0.69**	0.69
JS(↓)	–	**0.35**	0.37	0.36
ε-PPL(↓)	–	497	**11.4**	13.7
MAUVE (↑)	0.06	0.02	0.88	**0.94**
Human(↑)	–	–	9.0	**15.7**

Table 3: Generation quality w.r.t different **decoding algorithms** (web text, GPT-2 xl) under various metrics, and humans. MAUVE correctly captures the relationship greedy $\prec$ ancestral $\prec$ nucleus, and rates the adversarial decoder's text as low quality. Results are consistent across model sizes and random seeds. Boldfaced/highlighted entries denote the best decoding algorithm under each metric.

	Small	Medium	Large	XL
Gen. PPL(↓)	11.2	8.5	**0.9**	1.5
Zipf(↓)	0.06	**0.00**	0.02	0.01
Self-BLEU(↓)	0.05	**0.02**	0.03	0.03
SP(↑)	0.65	0.67	0.68	**0.69**
JS(↓)	0.41	0.39	0.37	**0.36**
ε-PPL(↓)	25.9	18.8	14.9	**13.7**
MAUVE (↑)	0.878	0.915	0.936	**0.940**
Human(↑)	−15.9	−3.4	12.6	**15.7**

Table 4: Generation quality w.r.t different **model sizes** (web text, nucleus sampling) under various metrics, as well as human evaluators. MAUVE captures the relationship between model size and generation quality, agreeing with human-evaluated quality. Results are consistent across random seeds and decoding algorithms. Boldfaced/highlighted entries denote the best model size under each metric.

Здесь наблюдается третье преимущество: правильная иерархия декодеров greedy < sampling (ancestral) < nucleus

Различные эмбединги

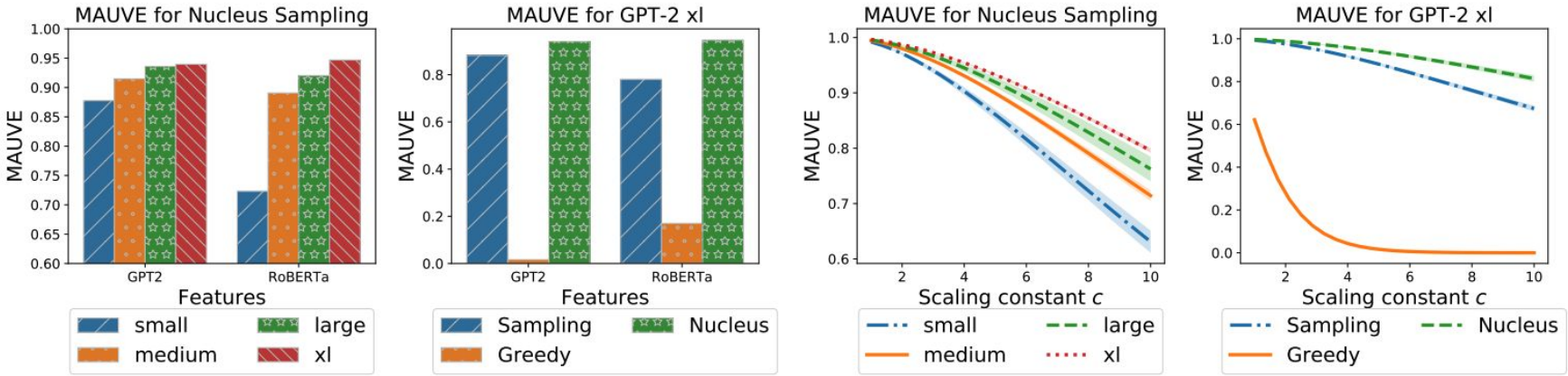

Figure 5: **Left**: MAUVE computed using GPT-2 (default) and RoBERTa [34] embeddings, across model sizes and decoding algorithms; see Table 12 in the Appendix for further results. The Spearman rank correlation between the two is **0.993** across model sizes and decoding algorithms. **Right**: Effect of the scaling constant c on MAUVE. Choice of c does not affect the relative order of the curves but only the numerical value. We use $c = 5$ to get interpretable values with both nucleus and greedy decoding.

Metric	Task	Gen. PPL	Zipf Coef.	REP	Distinct-4	Self-BLEU	MAUVE
Human-like/BT	Web text	0.810	0.833	−0.167	0.738	0.595	**0.952**
Interesting/BT	Web text	0.643	0.524	−0.143	0.524	0.405	**0.810**
Sensible/BT	Web text	0.738	0.690	−0.071	0.595	0.524	**0.857**
% Disc. Acc.	News	0.468	0.595	0.792	0.653	0.516	**0.956**
% Disc. Acc.	Stories	0.643	0.643	0.250	0.750	0.857	**0.893**

Table 5: Correlation of various similarity measures with human judgments when available, and the accuracy of a trained discriminator otherwise. "BT" denotes the Bradley-Terry score for a pairwise human evaluation (§ 4.3). Boldfaced/highlighted numbers indicate highest correlation in each row. We observe that MAUVE has the highest correlation with human evaluation and discriminator accuracy.

Вывод

Мы представили MAUVE, автоматическую меру разницы между текстом нейронной сети и человеческим текстом для генерации открытого текста. MAUVE измеряет площадь под кривой дивергенции, формализуя и обобщая спектр ошибок, которые охватывают явления, присутствующие в машинном и созданном человеком тексте. MAUVE также коррелирует с человеческими суждениями и выявляет различия в качестве из-за длины сгенерированного текста, алгоритма декодирования и размера модели, которые с трудом удается зафиксировать в предыдущих метриках. Автоматизированные метрики способствовали развитию компьютерного зрения и многих других областей машинного обучения. Принципиальная основа MAUVE и высокая эмпирическая производительность предлагают аналогичный путь вперед для открытых систем генерации текста. Расширения MAUVE для закрытых задач, таких как обобщение текста и перевод, где сгенерированный текст должен сравниваться с фиксированным набором правильных ответов, являются многообещающими направлениями для будущей работы.

Рецензия. Вклад

В статье предложен метод оценивания качества открытой генерации текста, который:

- лишен необходимости ручной разметки
- хорошо коррелирует с человеческим восприятием

Рецензия. Сильные стороны

Простота

Актуальность, т.к. нет общепринятой метрики, хорошо коррелирующей с человеческим восприятием

Достаточно количество **экспериментов**, которые показывают, что метод:

- очень хорошо коррелирует с человеческим восприятием
- согласуется с известными свойствами сгенерированных текстов
- устойчив к изменению внутренних составных частей (почти всех)

Супер **текст**

Рецензия. Слабые стороны

Использование внешней языковой модели

Пересекающиеся обучающие данные это плохо или хорошо?

OpenReview: на большую часть замечаний авторы ответили и добавили в Appendix.

Рецензия. Оценка

Оценка: 9 (Top 15% of accepted NeurIPS papers. An excellent submission; a strong accept)

Уверенность: 4 (You are confident in your assessment, but not absolutely certain)

Статья

- NeurIPS 2021 Oral (21 May 2021 submitted)
- NeurIPS 2021 Outstanding Paper
- arxiv[v1]: 2 Feb 2021
- Оказали наибольшее влияние: “Precision-Recall Curves Using Information Divergence Frontiers”. Также статьи по предыдущим метрикам: FID, perplexity, BLEU, etc.
- Авторы в основным из “Allen School of CS & Eng., University of Washington” и “Allen Institute for Artificial Intelligence”
- Развить идею метрики для применения в машинном переводе или суммаризации

Авторы

Krishna Pillutla, Ph.D candidate in the Paul G. Allen School of Computer Science & Engineering at the **University of Washington.** Master's at CMU.

- Самая цитируемая (121): "Robust aggregation for federated learning", 2019
- Всего цитирований: 182
- Соавтор еще двух статей, попавших на NeurIPS 2021.

Авторы

Swabha Swayamdipta, postdoctoral researcher at the Allen Institute for AI**.** Master’s at Columbia University. PhD at CMU.

- Самая цитируемая (616): “Annotation artifacts in natural language inference data”, 2018
- Всего цитирований: 2592
- Соавтор многих статей по NLP, попавших на EMNLP, ICLR предыдущих лет

Авторы

Rowan Zellers, a final year PhD candidate at the University of Washington. Part time at the Allen Institute for Artificial Intelligence.

- Самая цитируемая (477): “Neural motifs: Scene graph parsing with global context”, 2018
- Всего цитирований: 2481
- Соавтор еще одной NeurIPS 2021 Oral: “MERLOT: Multimodal N… Knowledge Models”

🎉New (Nov 2021): I'm on the academic job market!🎉