A Contrastive Framework for Neural Text Generation

Yixuan Su, Tian Lan, Yan Wang, Dani Yogatama,
Lingpeng Kong, Nigel Collier

Motivations

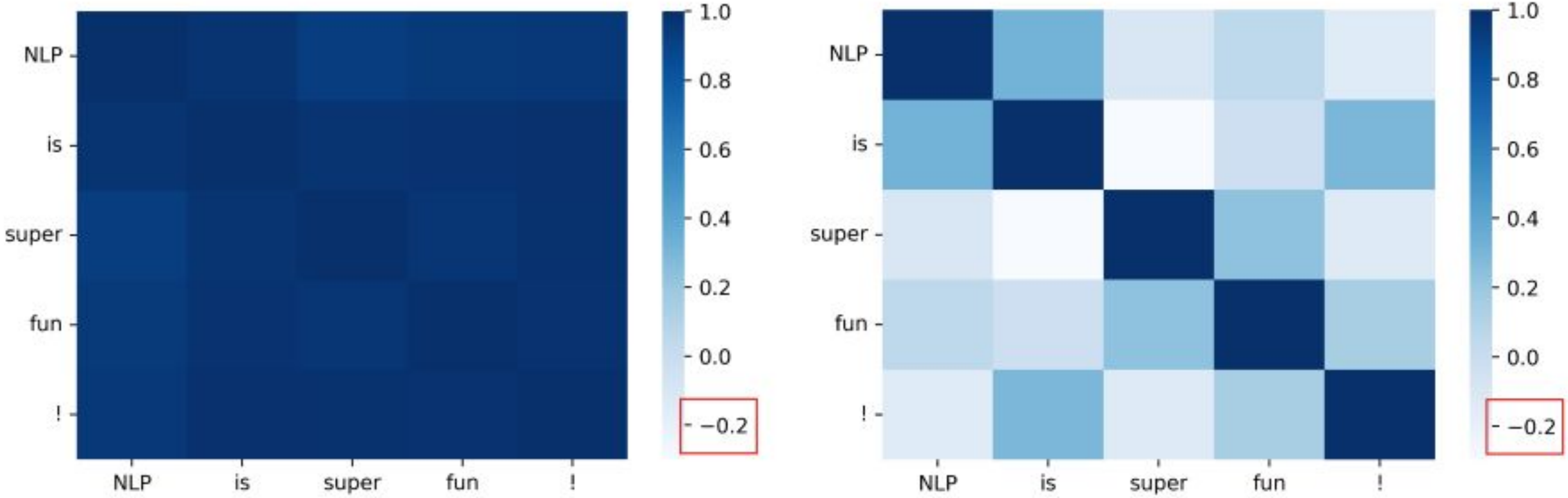
Text generation is of great importance to many natural language processing applications.

- Maximization-based decoding methods (e.g., beam search) often lead to degenerate solutions
  - the generated text is unnatural and contains undesirable repetitions.

Motivations

- Existing approaches introduce stochasticity via sampling modify training objectives to decrease the probabilities of certain tokens (e.g., unlikelihood training).
  - However, they often lead to solutions that lack coherence.
- This work shows that an underlying reason for model degeneration is the *anisotropic distribution of token representations*.

Motivations



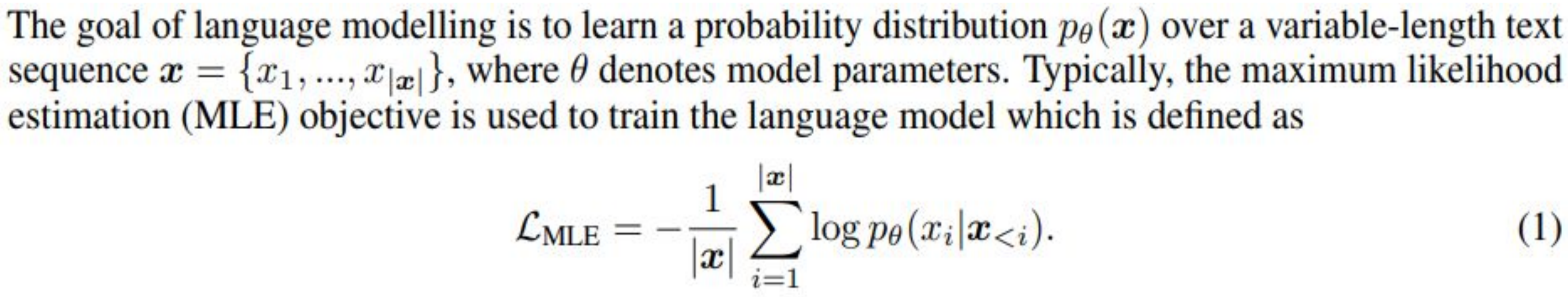
(a)

(b)

Contributions

- This work proposes a contrastive solutions to the mentioned problems.
- **Contrastive Training:** the sparseness of the token similarity matrix of the generated text should be preserved to avoid degeneration.
- **Contrastive Search:** at each decoding step, the output should be selected from the set of most probable candidates predicted by the model
  - allows the semantic coherence between the generated text and the human-written prefix

Language Modeling



The goal of language modelling is to learn a probability distribution $p_\theta(\boldsymbol{x})$ over a variable-length text sequence $\boldsymbol{x} = \{x_1, ..., x_{|\boldsymbol{x}|}\}$, where θ denotes model parameters. Typically, the maximum likelihood estimation (MLE) objective is used to train the language model which is defined as

$$\mathcal{L}_{\text{MLE}} = -\frac{1}{|\boldsymbol{x}|} \sum_{i=1}^{|\boldsymbol{x}|} \log p_\theta(x_i | \boldsymbol{x}_{<i}). \tag{1}$$

Contrastive Training

- Our goal is to encourage the language model to learn discriminative and isotropic token representations.

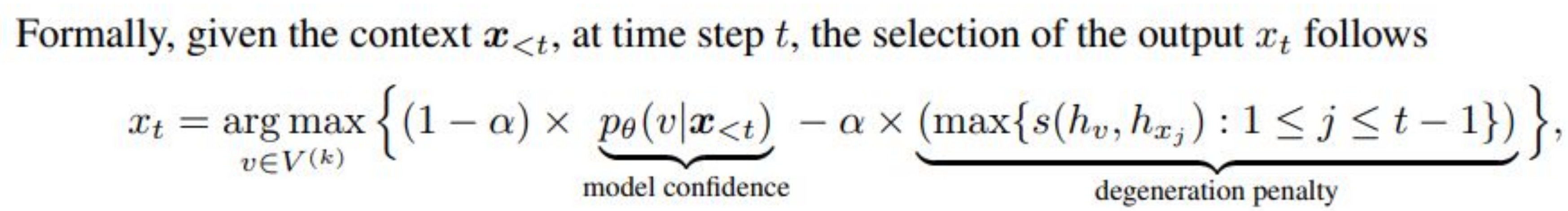
$$\mathcal{L}_{\text{CL}} = \frac{1}{|\boldsymbol{x}| \times (|\boldsymbol{x}| - 1)} \sum_{i=1}^{|\boldsymbol{x}|} \sum_{j=1, j \neq i}^{|\boldsymbol{x}|} \max\{0, \rho - s(h_{x_i}, h_{x_i}) + s(h_{x_i}, h_{x_j})\},$$

$$s(h_{x_i}, h_{x_j}) = \frac{h_{x_i}^{\top} h_{x_j}}{\|h_{x_i}\| \cdot \|h_{x_j}\|}.$$

$$\mathcal{L}_{\text{SimCTG}} = \mathcal{L}_{\text{MLE}} + \mathcal{L}_{\text{CL}},$$

Contrastive Search

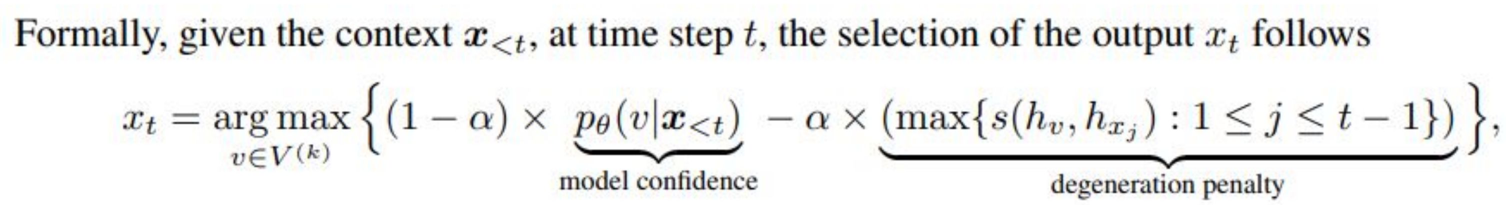
- **Model confidence**, is the probability of candidate ***v*** predicted by the model.
- **Degeneration penalty**, measures how discriminative of candidate ***v*** with respect to the previous context ***x<t*** and the similarity matrix, ***s***.



Formally, given the context $\boldsymbol{x}_{<t}$, at time step t, the selection of the output x_t follows

$$x_t = \underset{v \in V^{(k)}}{\arg\max} \left\{ (1-\alpha) \times \underbrace{p_\theta(v|\boldsymbol{x}_{<t})}_{\text{model confidence}} - \alpha \times \underbrace{(\max\{s(h_v, h_{x_j}) : 1 \le j \le t-1\})}_{\text{degeneration penalty}} \right\},$$

Contrastive Search



Formally, given the context $\boldsymbol{x}_{<t}$, at time step t, the selection of the output x_t follows

$$x_t = \underset{v \in V^{(k)}}{\arg\max} \Big\{ (1-\alpha) \times \underbrace{p_\theta(v|\boldsymbol{x}_{<t})}_{\text{model confidence}} - \alpha \times \underbrace{(\max\{s(h_v, h_{x_j}) : 1 \le j \le t-1\})}_{\text{degeneration penalty}} \Big\},$$

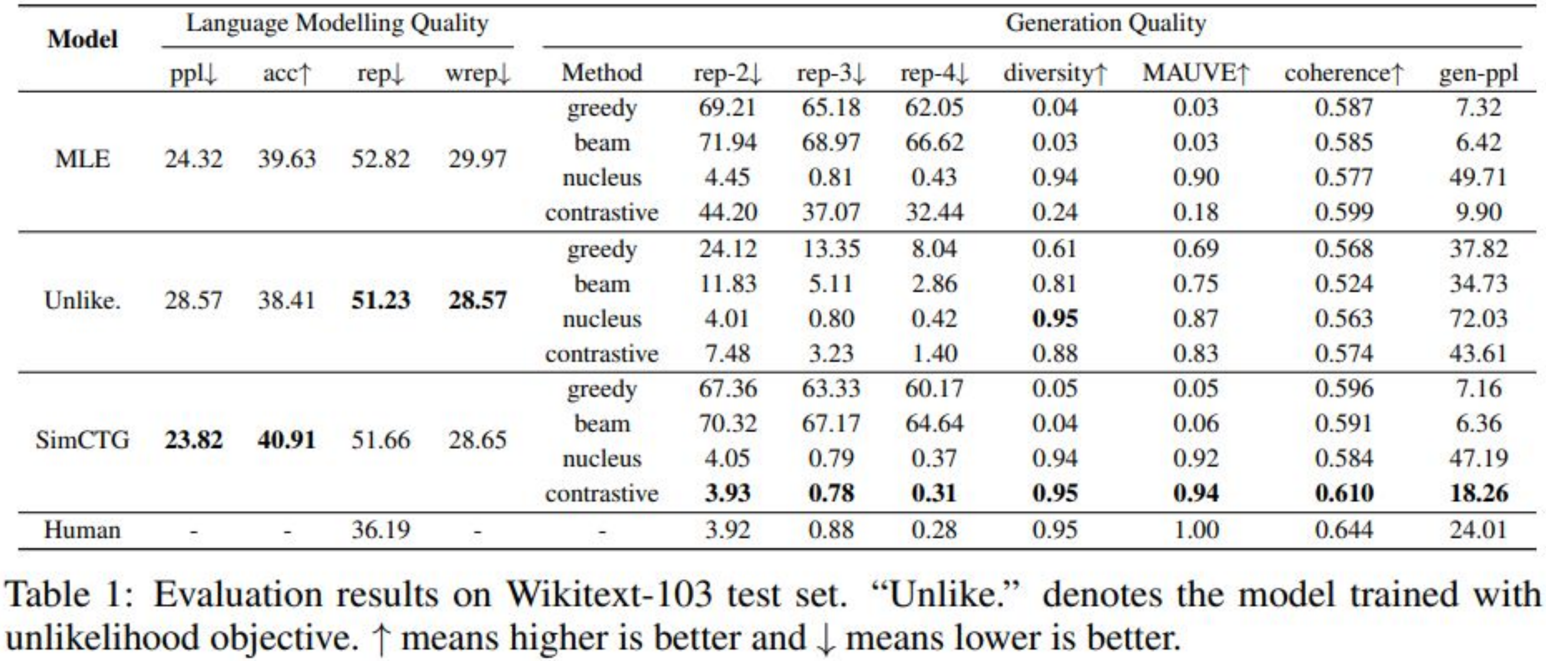
The two dogs are ______. P(running | x<t) = 0.5, P(playing | x<t) = 0.5

Cos(x1, x2)	The	two	dogs	are
running	0.2	0.3	**0.4**	0.3
playing	0.2	0.1	0.2	**0.3**

The two dogs are **running** : 0.5 * 0.5 - 0.5 * 0.4 = 0.1

The two dogs are **playing**: 0.5 * 0.5 - 0.5 * 0.3 = 0.2

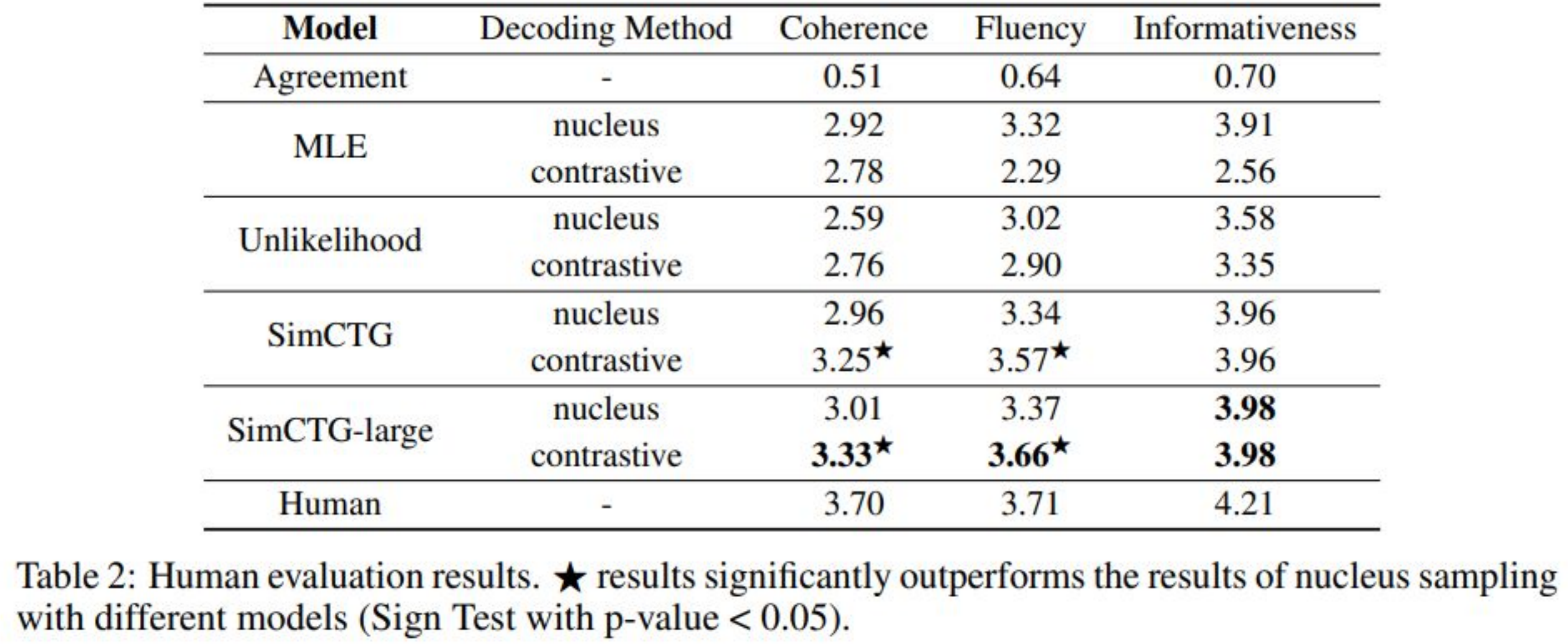
Experimental Results



Model	Language Modelling Quality				Generation Quality							
	ppl↓	acc↑	rep↓	wrep↓	Method	rep-2↓	rep-3↓	rep-4↓	diversity↑	MAUVE↑	coherence↑	gen-ppl
MLE	24.32	39.63	52.82	29.97	greedy	69.21	65.18	62.05	0.04	0.03	0.587	7.32
					beam	71.94	68.97	66.62	0.03	0.03	0.585	6.42
					nucleus	4.45	0.81	0.43	0.94	0.90	0.577	49.71
					contrastive	44.20	37.07	32.44	0.24	0.18	0.599	9.90
Unlike.	28.57	38.41	**51.23**	**28.57**	greedy	24.12	13.35	8.04	0.61	0.69	0.568	37.82
					beam	11.83	5.11	2.86	0.81	0.75	0.524	34.73
					nucleus	4.01	0.80	0.42	**0.95**	0.87	0.563	72.03
					contrastive	7.48	3.23	1.40	0.88	0.83	0.574	43.61
SimCTG	**23.82**	**40.91**	51.66	28.65	greedy	67.36	63.33	60.17	0.05	0.05	0.596	7.16
					beam	70.32	67.17	64.64	0.04	0.06	0.591	6.36
					nucleus	4.05	0.79	0.37	0.94	0.92	0.584	47.19
					contrastive	**3.93**	**0.78**	**0.31**	**0.95**	**0.94**	**0.610**	**18.26**
Human	-	-	36.19	-	-	3.92	0.88	0.28	0.95	1.00	0.644	24.01

Table 1: Evaluation results on Wikitext-103 test set. “Unlike.” denotes the model trained with unlikelihood objective. ↑ means higher is better and ↓ means lower is better.

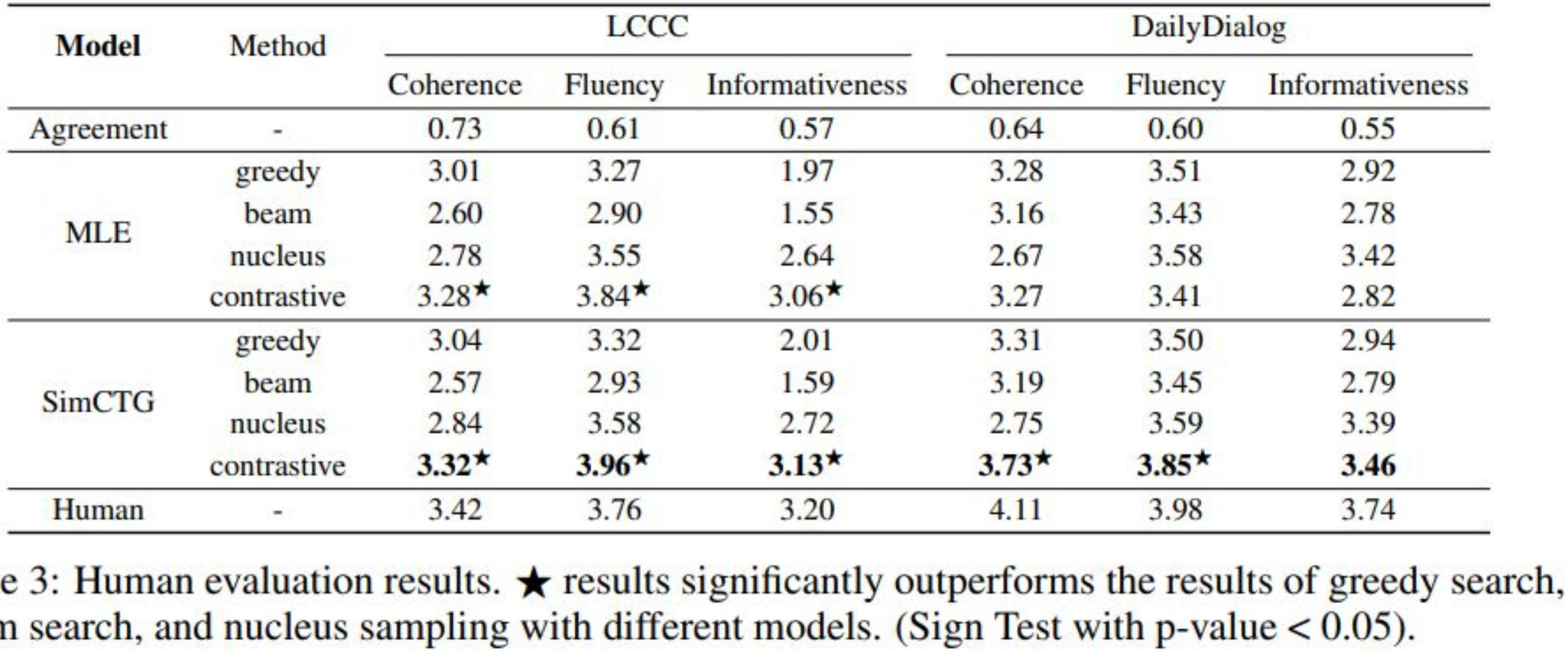
Human Evaluation



Model	Decoding Method	Coherence	Fluency	Informativeness
Agreement	-	0.51	0.64	0.70
MLE	nucleus	2.92	3.32	3.91
	contrastive	2.78	2.29	2.56
Unlikelihood	nucleus	2.59	3.02	3.58
	contrastive	2.76	2.90	3.35
SimCTG	nucleus	2.96	3.34	3.96
	contrastive	3.25★	3.57★	3.96
SimCTG-large	nucleus	3.01	3.37	**3.98**
	contrastive	**3.33**★	**3.66**★	**3.98**
Human	-	3.70	3.71	4.21

Table 2: Human evaluation results. ★ results significantly outperforms the results of nucleus sampling with different models (Sign Test with p-value < 0.05).

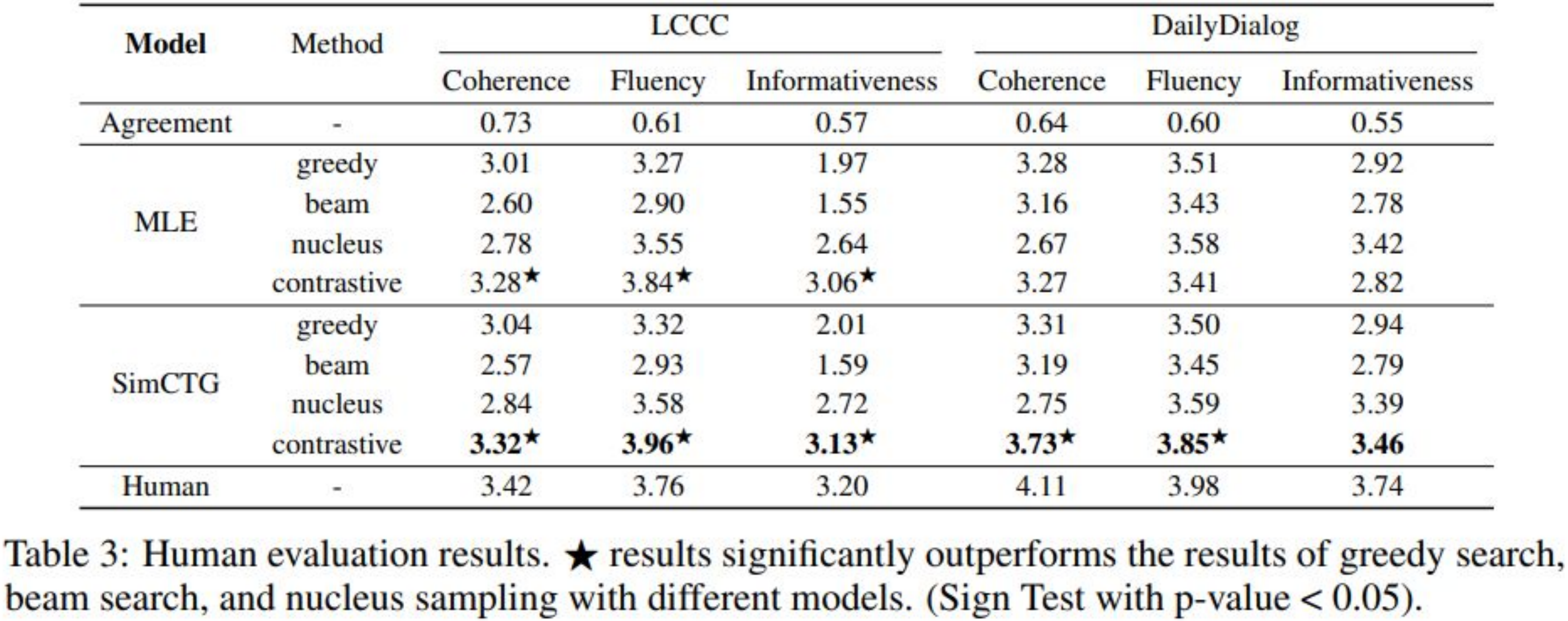
Open-domain Dialogue Generation



Model	Method	LCCC			DailyDialog		
		Coherence	Fluency	Informativeness	Coherence	Fluency	Informativeness
Agreement	-	0.73	0.61	0.57	0.64	0.60	0.55
MLE	greedy	3.01	3.27	1.97	3.28	3.51	2.92
	beam	2.60	2.90	1.55	3.16	3.43	2.78
	nucleus	2.78	3.55	2.64	2.67	3.58	3.42
	contrastive	3.28★	3.84★	3.06★	3.27	3.41	2.82
SimCTG	greedy	3.04	3.32	2.01	3.31	3.50	2.94
	beam	2.57	2.93	1.59	3.19	3.45	2.79
	nucleus	2.84	3.58	2.72	2.75	3.59	3.39
	contrastive	**3.32**★	**3.96**★	**3.13**★	**3.73**★	**3.85**★	**3.46**
Human	-	3.42	3.76	3.20	4.11	3.98	3.74

Table 3: Human evaluation results. ★ results significantly outperforms the results of greedy search, beam search, and nucleus sampling with different models. (Sign Test with p-value < 0.05).

Open-domain Dialogue Generation



Model	Method	LCCC			DailyDialog		
		Coherence	Fluency	Informativeness	Coherence	Fluency	Informativeness
Agreement	-	0.73	0.61	0.57	0.64	0.60	0.55
MLE	greedy	3.01	3.27	1.97	3.28	3.51	2.92
	beam	2.60	2.90	1.55	3.16	3.43	2.78
	nucleus	2.78	3.55	2.64	2.67	3.58	3.42
	contrastive	3.28★	3.84★	3.06★	3.27	3.41	2.82
SimCTG	greedy	3.04	3.32	2.01	3.31	3.50	2.94
	beam	2.57	2.93	1.59	3.19	3.45	2.79
	nucleus	2.84	3.58	2.72	2.75	3.59	3.39
	contrastive	**3.32**★	**3.96**★	**3.13**★	**3.73**★	**3.85**★	**3.46**
Human	-	3.42	3.76	3.20	4.11	3.98	3.74

Table 3: Human evaluation results. ★ results significantly outperforms the results of greedy search, beam search, and nucleus sampling with different models. (Sign Test with p-value < 0.05).

Open-domain Dialogue Generation

$$\text{self-similarity}(\boldsymbol{x}) = \frac{1}{|\boldsymbol{x}|\times(|\boldsymbol{x}|-1)}\sum_{i=1}^{|\boldsymbol{x}|}\sum_{j=1,j\neq i}^{|\boldsymbol{x}|}\frac{h_{x_i}^{\top}h_{x_j}}{\|h_{x_i}\|\cdot\|h_{x_j}\|},$$

- In the intermediate layers, the self-similarity of different models are relatively the same.
- the output layer (layer 12), SimCTG’s self-similarity becomes notably lower than other baselines.
- the Unlikelihood model also yields more discriminative representations than MLE, but its language model accuracy is lower than MLE and SimCTG

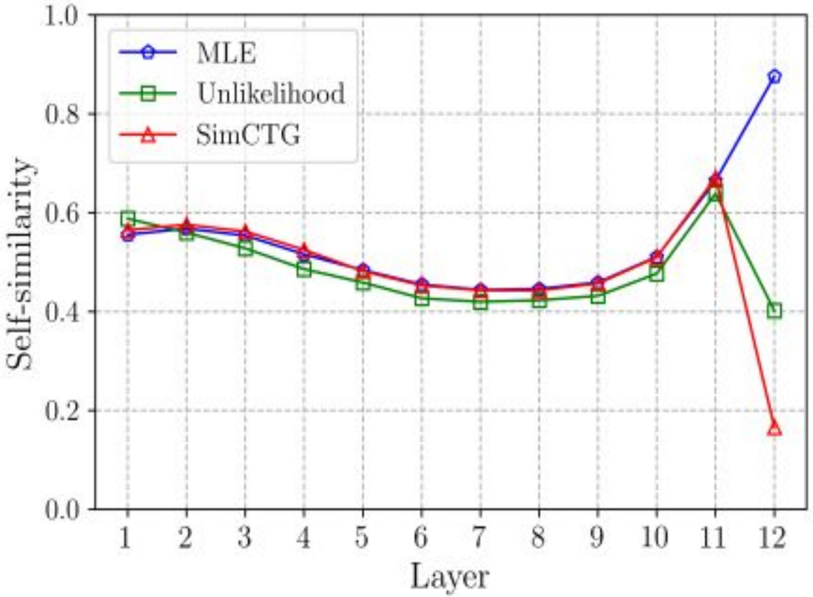


Figure 2: Layer-wise representation self-similarity.

Open-domain Dialogue Generation

- when ρ = 0, **SimCTG** is equivalent to **MLE**.
- The contrastive training always helps to improve the perplexity as compared with **MLE**.
- However, when ρ is either too small (e.g., 0.1) or large (e.g., 1.0), the learned representation space of the model would be either less or too isotropic, leading to a sub-optimal perplexity.
- In the experiments, the most suitable margin ρ = 0.5.

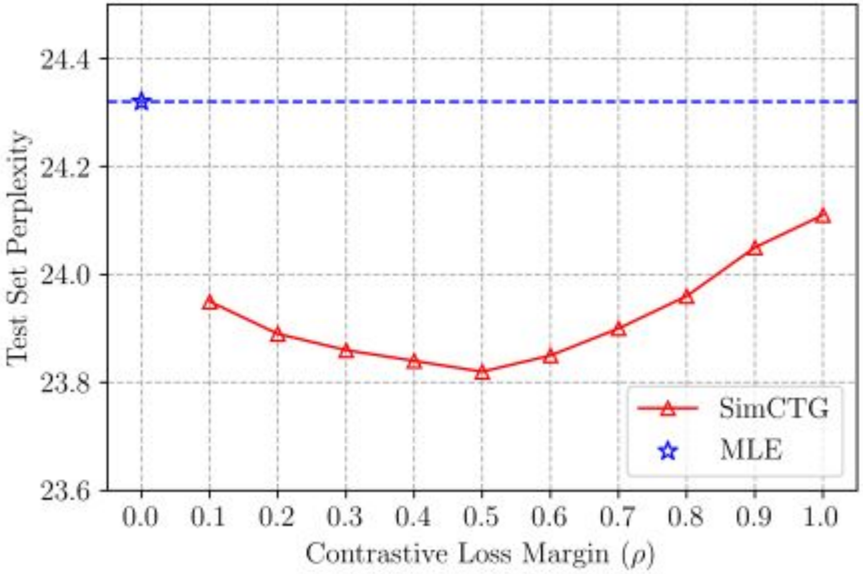


Figure 3: The effect of contrastive margin ρ.

Open-domain Dialogue Generation

- when p is small (i.e., p ≤ 0.7), its generation perplexity is comparable to that of human.
    - However, the diversity is notably lower than human performance, meaning it stuck in undesirable repetition loops.
- when p is large (i.e., p ≥ 0.95), the generation diversity is close to that of human but with higher generation perplexity.
    - means the generated text is very unlikely, therefore being low quality.

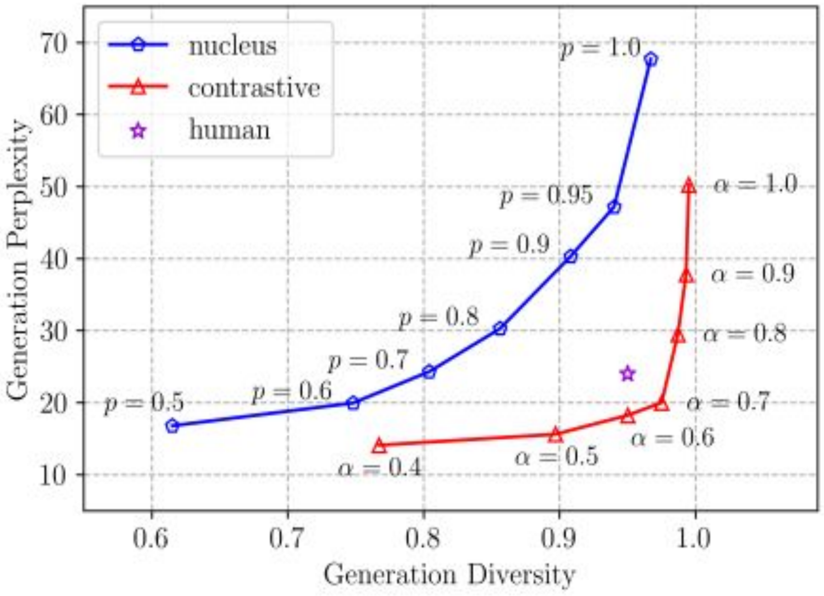


Figure 4: Contrastive search vs nucleus sampling.

Open-domain Dialogue Generation

- As for contrastive search, when α ∈ [0.5, 0.8], it yields generation diversity and perplexity that are both comparable to human performance.
    - These results demonstrate the superiority of contrastive search as it better balances the trade-off between the generation diversity and perplexity.

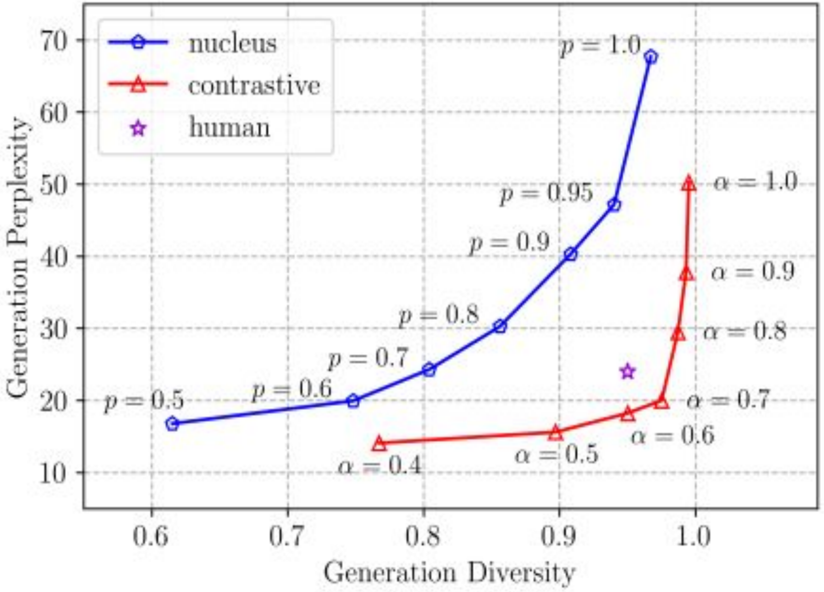


Figure 4: Contrastive search vs nucleus sampling.

Open-domain Dialogue Generation

- Greedy search is the fastest method and the latency of different methods are generally comparable with each other.
- Comparing contrastive search with beam search, when b and k are small (i.e., ≤ 6), their latency are nearly identical.
- When b and k gets larger (i.e., > 6), contrastive search becomes faster.
- In summary, these comparison results further verify the practical usage of contrastive search.

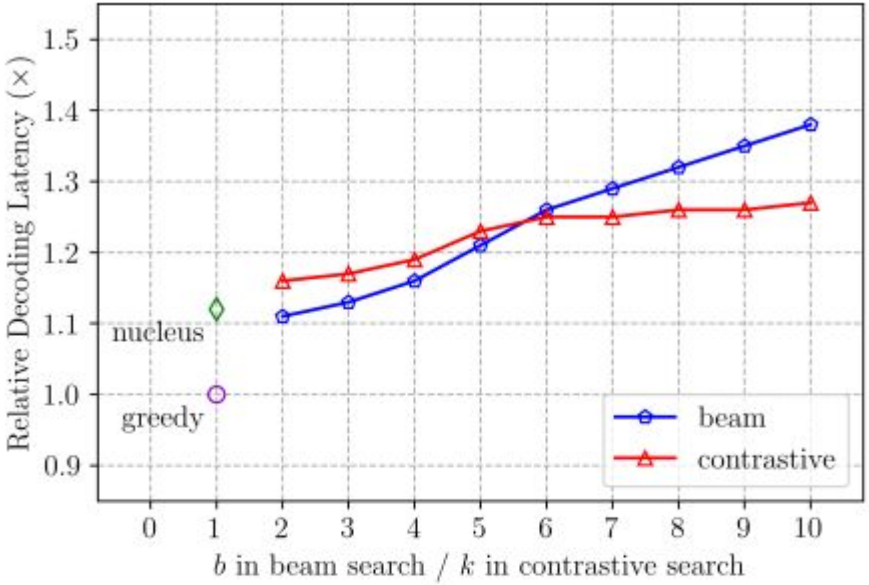
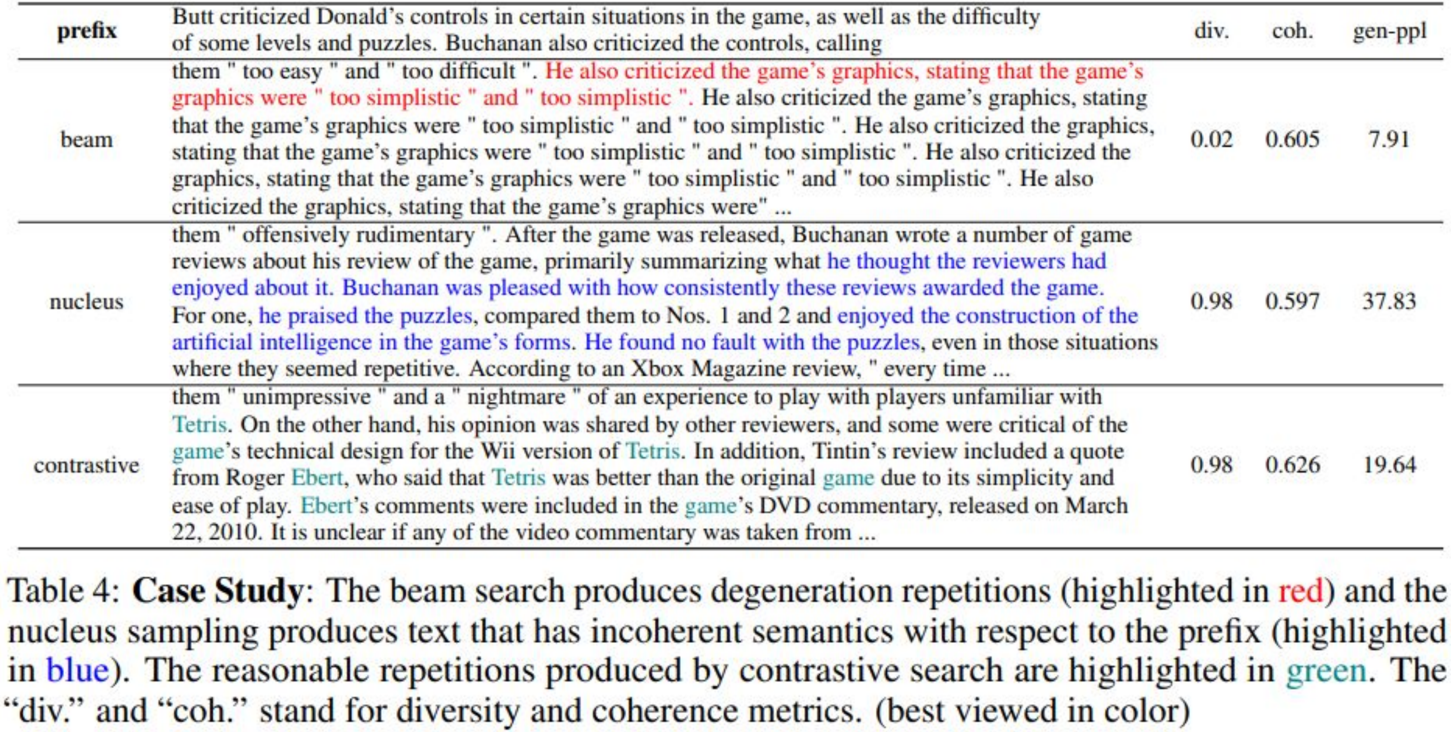


Figure 5: Inference latency comparison.

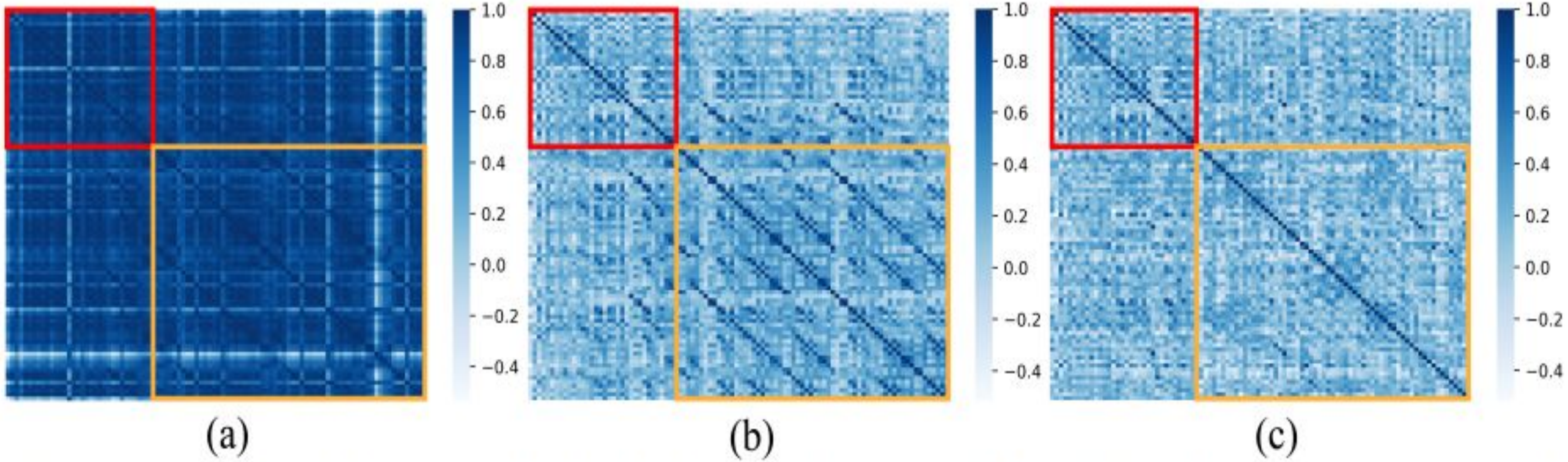
Case Study



prefix	Butt criticized Donald's controls in certain situations in the game, as well as the difficulty of some levels and puzzles. Buchanan also criticized the controls, calling	div.	coh.	gen-ppl
beam	them " too easy " and " too difficult ". He also criticized the game's graphics, stating that the game's graphics were " too simplistic " and " too simplistic ". He also criticized the game's graphics, stating that the game's graphics were " too simplistic " and " too simplistic ". He also criticized the graphics, stating that the game's graphics were " too simplistic " and " too simplistic ". He also criticized the graphics, stating that the game's graphics were " too simplistic " and " too simplistic ". He also criticized the graphics, stating that the game's graphics were" ...	0.02	0.605	7.91
nucleus	them " offensively rudimentary ". After the game was released, Buchanan wrote a number of game reviews about his review of the game, primarily summarizing what he thought the reviewers had enjoyed about it. Buchanan was pleased with how consistently these reviews awarded the game. For one, he praised the puzzles, compared them to Nos. 1 and 2 and enjoyed the construction of the artificial intelligence in the game's forms. He found no fault with the puzzles, even in those situations where they seemed repetitive. According to an Xbox Magazine review, " every time ...	0.98	0.597	37.83
contrastive	them " unimpressive " and a " nightmare " of an experience to play with players unfamiliar with Tetris. On the other hand, his opinion was shared by other reviewers, and some were critical of the game's technical design for the Wii version of Tetris. In addition, Tintin's review included a quote from Roger Ebert, who said that Tetris was better than the original game due to its simplicity and ease of play. Ebert's comments were included in the game's DVD commentary, released on March 22, 2010. It is unclear if any of the video commentary was taken from ...	0.98	0.626	19.64

Table 4: **Case Study**: The beam search produces degeneration repetitions (highlighted in red) and the nucleus sampling produces text that has incoherent semantics with respect to the prefix (highlighted in blue). The reasonable repetitions produced by contrastive search are highlighted in green. The "div." and "coh." stand for diversity and coherence metrics. (best viewed in color)

Comparison of Token Similarity Matrix



(a) (b) (c)

Figure 6: (a) MLE + beam search; (b) SimCTG + beam search; (c) SimCTG + contrastive search. The token similarity matrix of the prefix and the generated text are highlighted in red and yellow.

Summary

- This work proposes a contrastive solutions to the mentioned problems.
- A **Contrastive Training** objective to calibrate the model's representation space.
- A decoding method—**Contrastive Search**—to encourage diversity while maintaining coherence in the generated text.

Thank You