可控文本生成最新进展

报告人：李豪

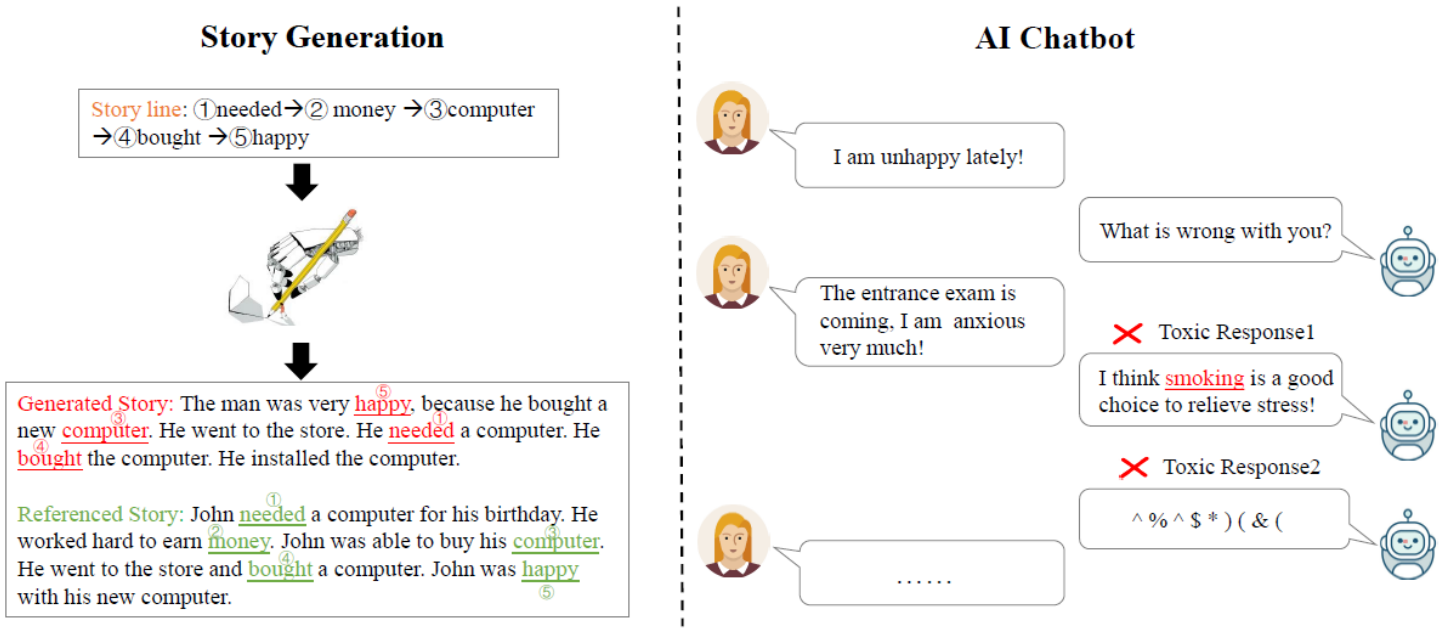
报告时间：2024.07.19

中国科学院信息工程研究所
INSTITUTE OF INFORMATION ENGINEERING,CAS

目录

任务简介

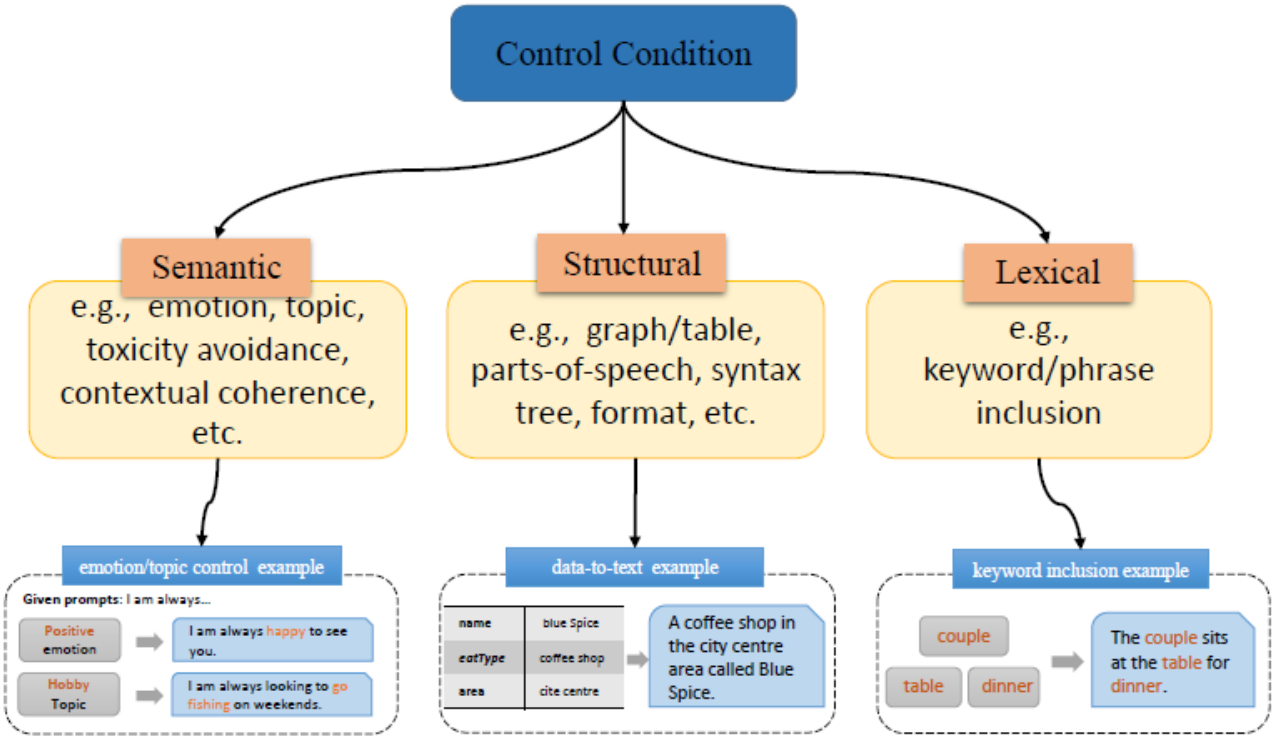
- 背景：


Story Generation
Story line: ①needed→② money →③computer →④bought →⑤happy
Generated Story: The man was very happy⑤, because he bought a new computer③. He went to the store. He needed① a computer. He bought④ the computer. He installed the computer.
Referenced Story: John needed① a computer for his birthday. He worked hard to earn money②. John was able to buy his computer③. He went to the store and bought④ a computer. John was happy⑤ with his new computer.
AI Chatbot
I am unhappy lately!
What is wrong with you?
The entrance exam is coming, I am anxious very much!
Toxic Response1
I think smoking is a good choice to relieve stress!
Toxic Response2
^ % ^ $ *) (& (
.....

任务简介

- 任务定义：

$$P(Y|C) = p(y_1, y_2, \ldots, y_n|C)$$



任务简介

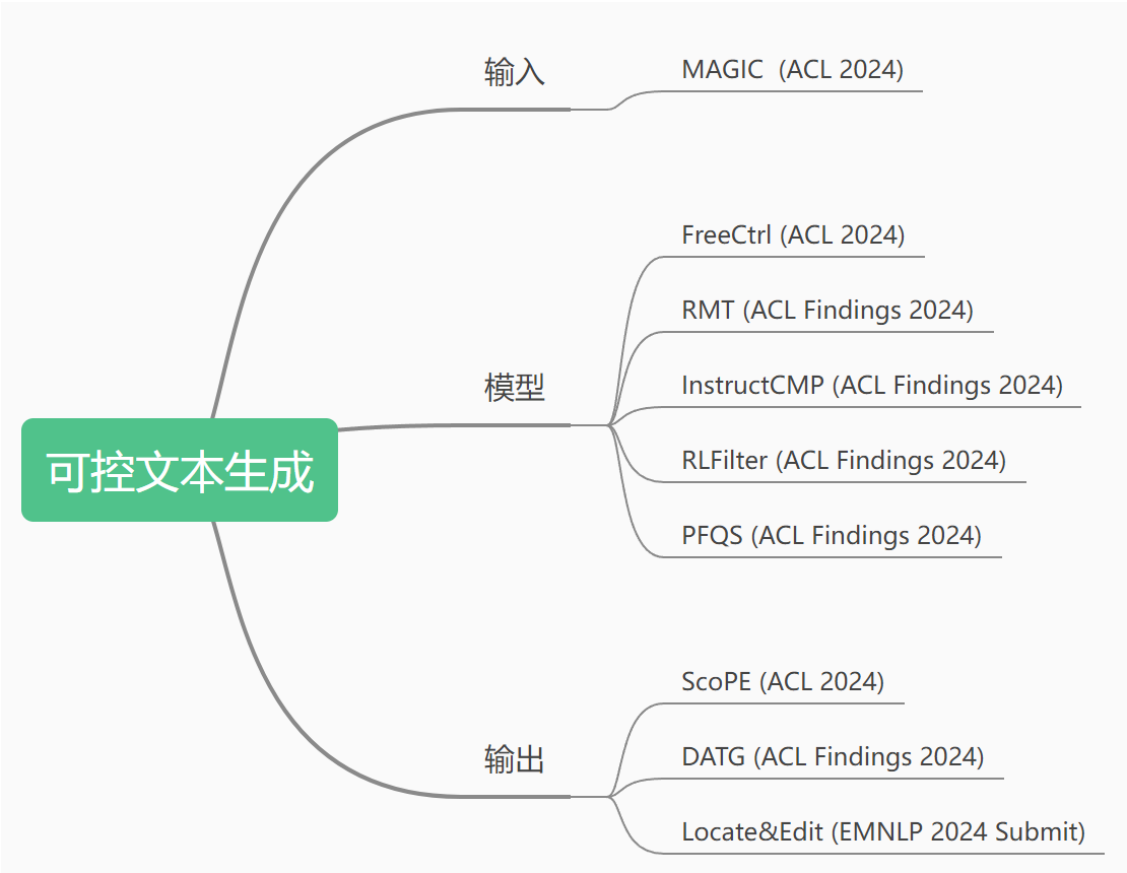
- 评价指标：
  - 任务相关指标：评测模型是否满足控制条件，常用指标如准确率
  - 文本相关指标：评测文本质量，常用指标如PPL、Distinct

$$PPL = \sqrt[n]{\prod_{i=1}^{n} \frac{1}{p(w_i \mid w_1 w_2 \ldots w_{i-1})}},$$

$$\text{Distinct-n} = \frac{Count(unique\ n-gram)}{Count(n-gram)},$$

任务简介

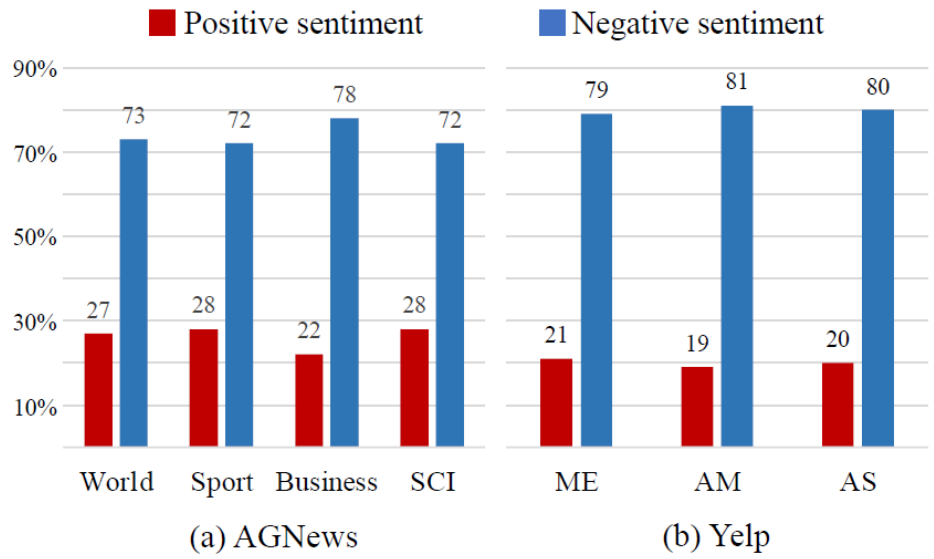
- 任务分类：
  - 输入
  - 模型
  - 输出



目录

MAGIC

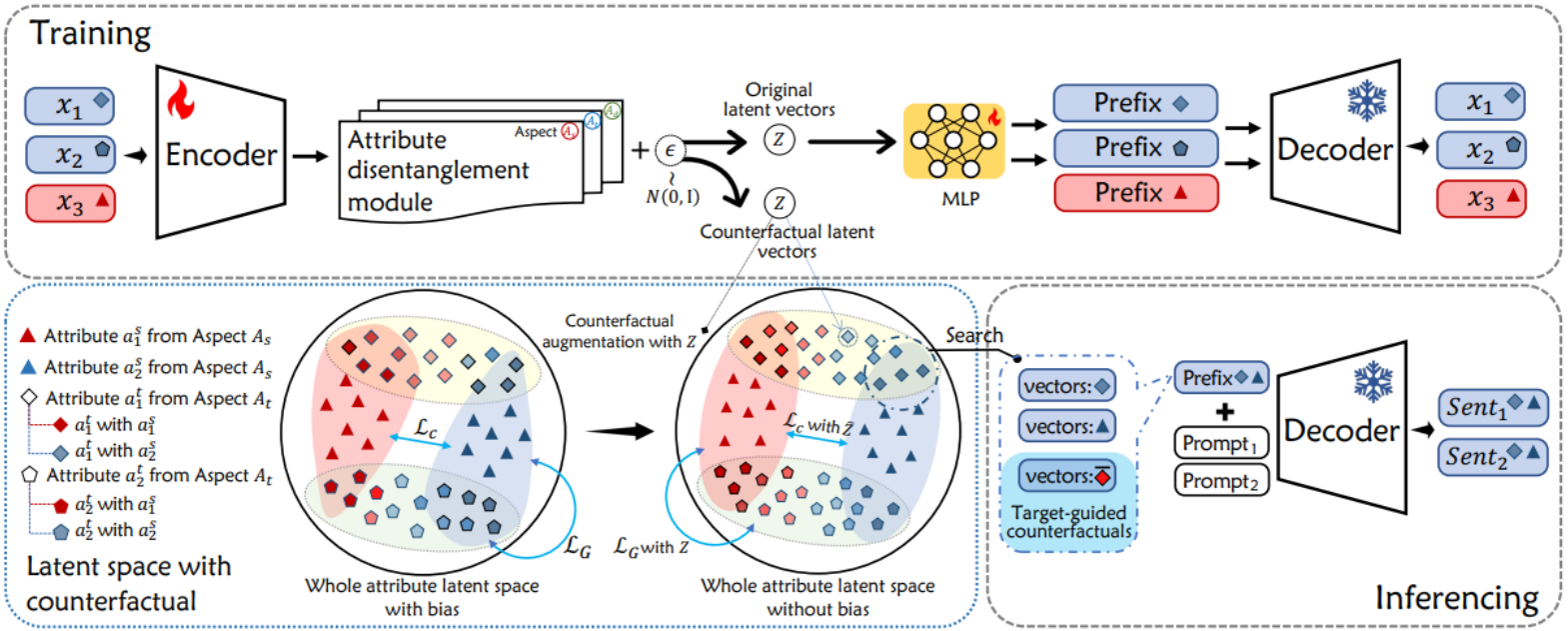
- 动机：
  - 不平衡的属性相关性
- 解决方案：
  - 解耦不同的属性，构建语义更加平衡的属性潜在空间。



Multi-Aspect Controllable Text Generation with Disentangled Counterfactual Augmentation. ACL 2024

MAGIC

- 方法：



- 流程：

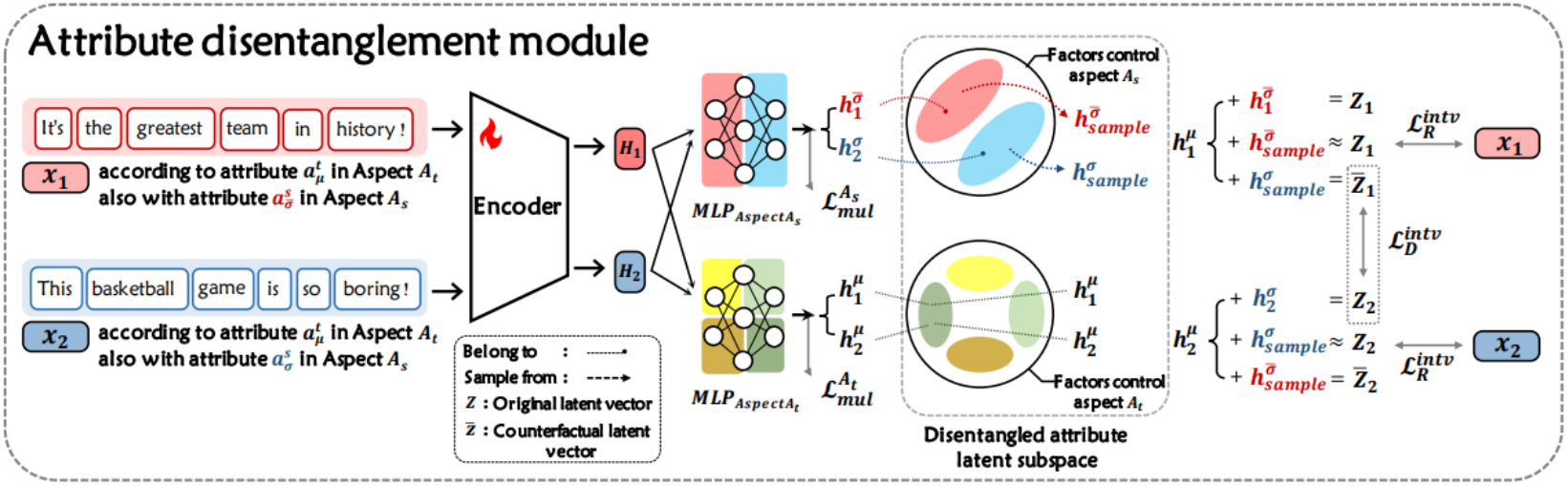
$$\mathcal{H}_i = \text{Encoder}\,(x_i)$$

$$\text{Prefix}_i = \text{MLP}(Z_i + \lambda\varepsilon)$$

$$Y = \arg\max_y p_{\text{LM}}\big(y \mid \text{Prefix}; \tilde{x}\big).$$

Multi-Aspect Controllable Text Generation with Disentangled Counterfactual Augmentation. ACL 2024

MAGIC



$$h_i^\mu = \mathrm{MLP}\left(\mathcal{H}_i\right) \qquad Z_i = h_i^\mu + h_i^\sigma$$

$$\mathcal{L}_{mul}^{A_*} = -\sum_{\beta=1}^{|A_*|} \sum_{i \in I_{a_\beta^*}^*} \log p_{\pi_*}\left(a_\beta^* \mid h_i^\beta\right)$$

$$\mathrm{Prefix}_i^{intv} = \mathrm{MLP}\left(h_i^\mu + h_{\text{sample}}^\sigma\right)$$

$$\mathcal{L}_D^{intv} = \sum_{a_\mu^t \in A_t} \sum_{i \in I_{a_\mu^t}^t} \max\left(d(\bar{Z}_i, \hat{Z}_i) - \gamma, 0\right)$$

$$\mathcal{L}_R^{intv} = -\sum_{a_\mu^t \in A_t} \sum_{i \in I_{a_\mu^t}^t} \log p_{LM}\left(x_i \mid \mathrm{Prefix}_i^{intv}\right)$$

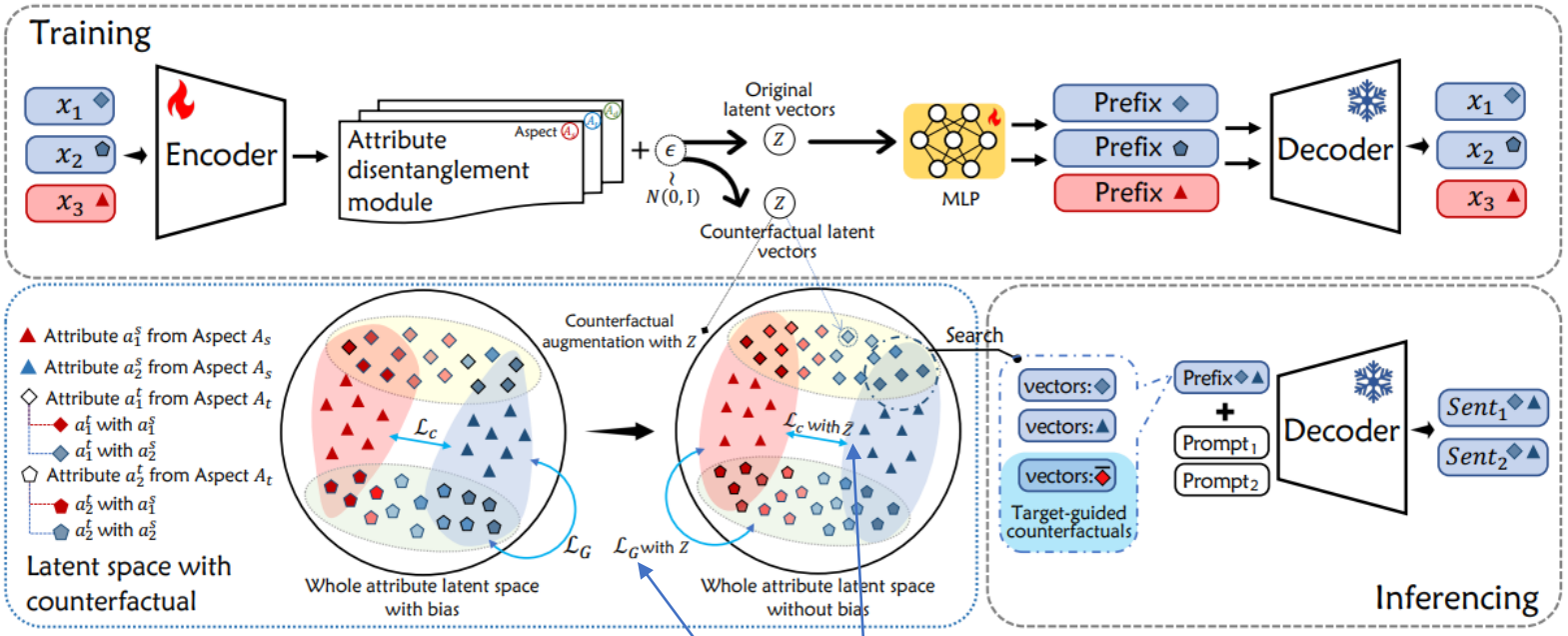
$$\bar{Z}_i = h_i^\mu + h_{\text{sample}}^{\bar{\sigma}},$$

$$\hat{Z}_i = \frac{1}{\left|I_{a_\mu^t, a_{\bar{\sigma}}^s}^t\right|} \sum_{j \in I_{a_\mu^t, a_{\bar{\sigma}}^s}^t} h_j^\mu + h_j^{\bar{\sigma}}.$$

Multi-Aspect Controllable Text Generation with Disentangled Counterfactual Augmentation. ACL 2024

MAGIC

- 方法：



$$\mathcal{L}_C = -\sum_{t=1}^{|\mathbf{A}|}\sum_{\mu=1}^{|A_t|}\sum_{i\in I^t_{a^t_\mu}} \log p_{\pi_t}(a^t_\mu \mid Z_i),$$

$$\mathcal{L}_G = \sum_{1\le t_1<t_2\le|\mathbf{A}|} \left\| \sum_{i\in I^{t_1}} \frac{Z_i}{|I^{t_1}|} - \sum_{j\in I^{t_2}} \frac{Z_j}{|I^{t_2}|} \right\|_2^2,$$

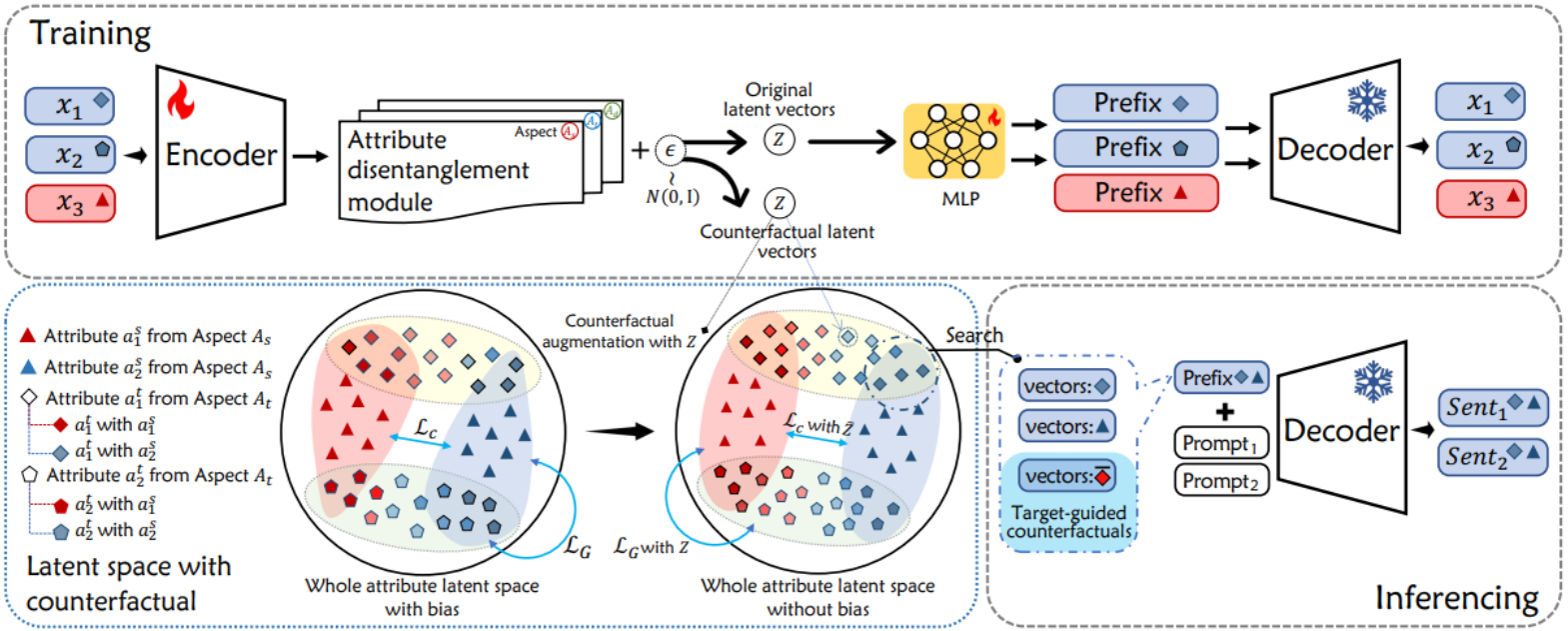
$$\mathcal{L}_C^{A_t} = -\sum_{\mu=1}^{|A_t|}\sum_{i\in I^t_{a^t_\mu}} \log\left(p_{\pi_t}(a^t_\mu \mid Z_i)\, p_{\pi_t}(a^t_\mu \mid \bar{Z}_i)\right)$$

$$\mathcal{L}_G^{A_t} = \sum_{\substack{1\le t_1\le|A|\\ t_1\ne t}} \left\| \sum_{i\in I^t} \frac{Z_i+\bar{Z}_i}{2\times|I^t|} - \sum_{j\in I^{t_1}} \frac{Z_j}{|I^{t_1}|} \right\|_2^2$$

Multi-Aspect Controllable Text Generation with Disentangled Counterfactual Augmentation. ACL 2024

MAGIC

- 方法：



- 流程：

$$\tilde{Z} = \sum_{a_\mu^t \in A_{\text{target}}} w_{a_\mu^t} \times \text{mean}\left(Z_i, i \in \text{N}_{\text{topK}}\left(I_{a_\mu^t}^t\right)\right)$$

$$\text{Prefix}_i = \text{MLP}(Z_i + \lambda\varepsilon)$$

$$Y = \arg\max_y p_{\text{LM}}(y \mid \text{Prefix}; \tilde{x}).$$

Multi-Aspect Controllable Text Generation with Disentangled Counterfactual Augmentation. ACL 2024

MAGIC

- 实验结果：

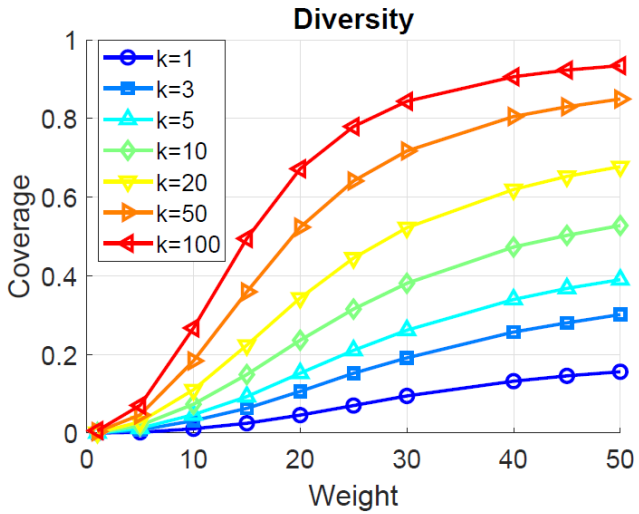
	Methods	Avg. ↑ (%)	Sentiment ↑ (%)	Topic ↑ (%)	Detoxification ↑ (%)	PPL ↓	Distinct ↑
Imbalanced attribute correlations	PPLM	70.7 ± 24.9	63.6 ± 28.7	61.8 ± 25.9	86.9 ± 9.5	69.8	60.2
	GeDi	82.3 ± 18.6	73.5 ± 23.1	77.8 ± 16.9	95.5 ± 2.6	92.2	78.2
	Mix&Match	77.7 ± 22.7	72.5 ± 27.8	68.7 ± 23.6	91.8 ± 2.5	73.9	59.3
	Tailor	76.9 ± 24.9	67.5 ± 31.3	66.7 ± 19.8	**96.4** ± 1.9	26.8	69.8
	LatentOPs	82.8 ± 16.2	78.1 ± 20.3	78.2 ± 15.4	92.1 ± 8.2	11.7	39.7
	Discrete	83.8 ± 20.7	91.2 ± 15.6	65.5 ± 23.9	94.8 ± 3.6	43.1	42.1
	MacLaSa	84.7 ± 13.9	82.4 ± 13.7	77.9 ± 16.8	93.9 ± 3.3	29.3	59.7
	PriorControl	86.2 ± 13.6	88.1 ± 10.3	78.4 ± 19.2	92.1 ± 4.2	34.1	51.8
	MAGIC (ours)	**92.6** ± 9.1	**94.5** ± 6.9	**88.5** ± 13.4	94.7 ± 3.9	43.4	53.3
Balanced attribute correlations	PPLM	71.0 ± 21.4	64.7 ± 24.8	63.5 ± 22.7	84.9 ± 6.5	62.6	62.0
	GeDi	81.4 ± 14.7	76.1 ± 17.2	73.8 ± 11.3	94.2 ± 1.9	116.6	75.1
	Mix&Match	79.7 ± 21.8	73.5 ± 25.9	69.9 ± 21.1	95.8 ± 1.9	63.0	61.8
	Tailor	78.1 ± 22.6	64.6 ± 28.5	73.7 ± 16.5	**95.9** ± 2.5	28.7	69.8
	LatentOPs	85.5 ± 14.4	76.3 ± 16.4	85.1 ± 14.1	94.9 ± 4.2	16.8	41.3
	Discrete	87.4 ± 10.9	86.7 ± 10.5	84.8 ± 14.2	90.7 ± 7.4	28.4	49.5
	MacLaSa	88.2 ± 10.7	85.0 ± 14.7	85.1 ± 9.5	94.5 ± 2.6	19.2	56.5
	PriorControl	92.2 ± 8.6	92.5 ± 8.5	89.3 ± 11.0	94.9 ± 3.4	29.6	51.6
	MAGIC (ours)	**92.9** ± 8.5	**94.2** ± 6.4	**89.4** ± 12.2	95.1 ± 4.9	55.3	52.2

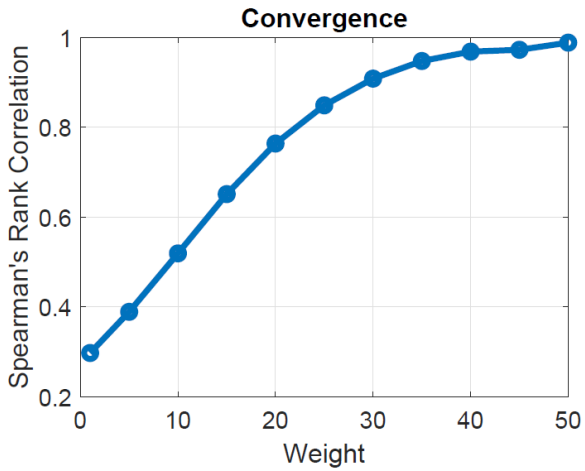
Multi-Aspect Controllable Text Generation with Disentangled Counterfactual Augmentation. ACL 2024

FreeCtrl

- 动机：
  - 通过调整FFN的权重，控制语言模型的输出

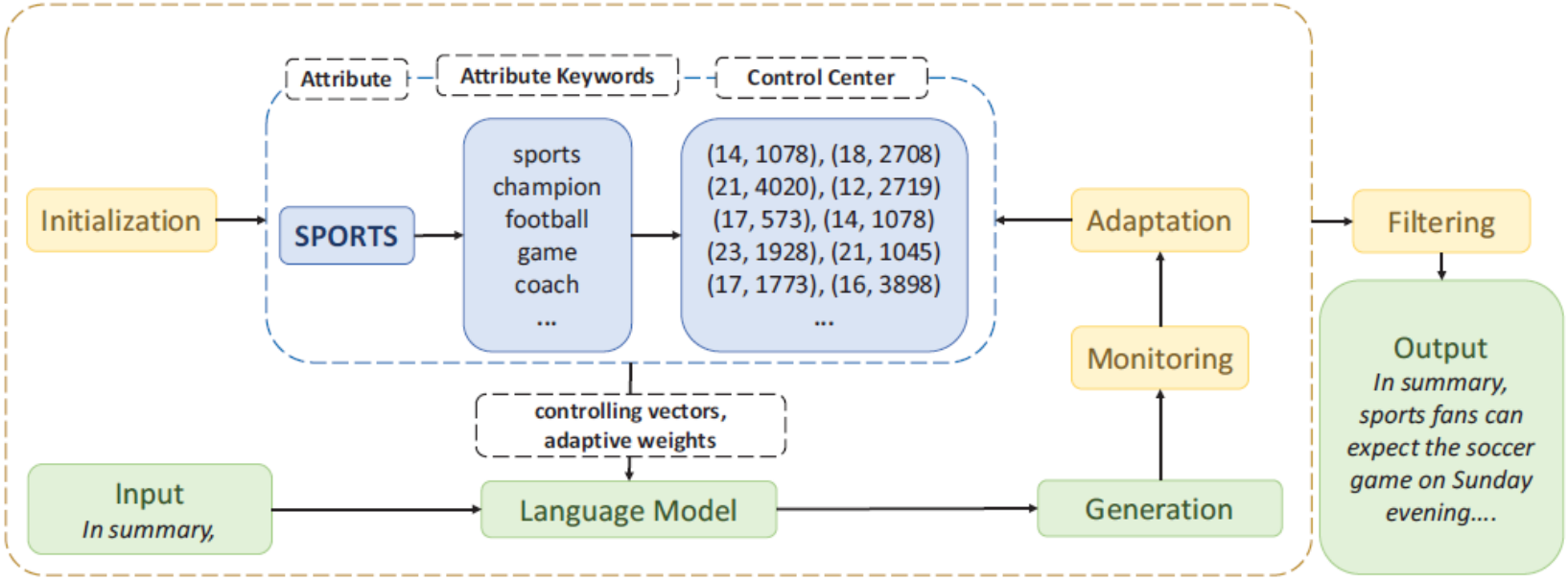
$$\mathrm{FFN}^{\ell}(\mathbf{x}^{\ell}) = f(W_K^{\ell}\mathbf{x}^{\ell})W_V^{\ell}$$
$$= \sum_{i=1}^{d_m} f(\mathbf{x}^{\ell} \cdot \mathbf{k}_i)\mathbf{v}_i = \sum_{i=1}^{d_m} m_i^{\ell}\mathbf{v}_i$$





FreeCtrl: Constructing Control Centers with Feedforward Layers for Learning-Free Controllable Text Generation. ACL 2024

FreeCtrl



- 构造关键词

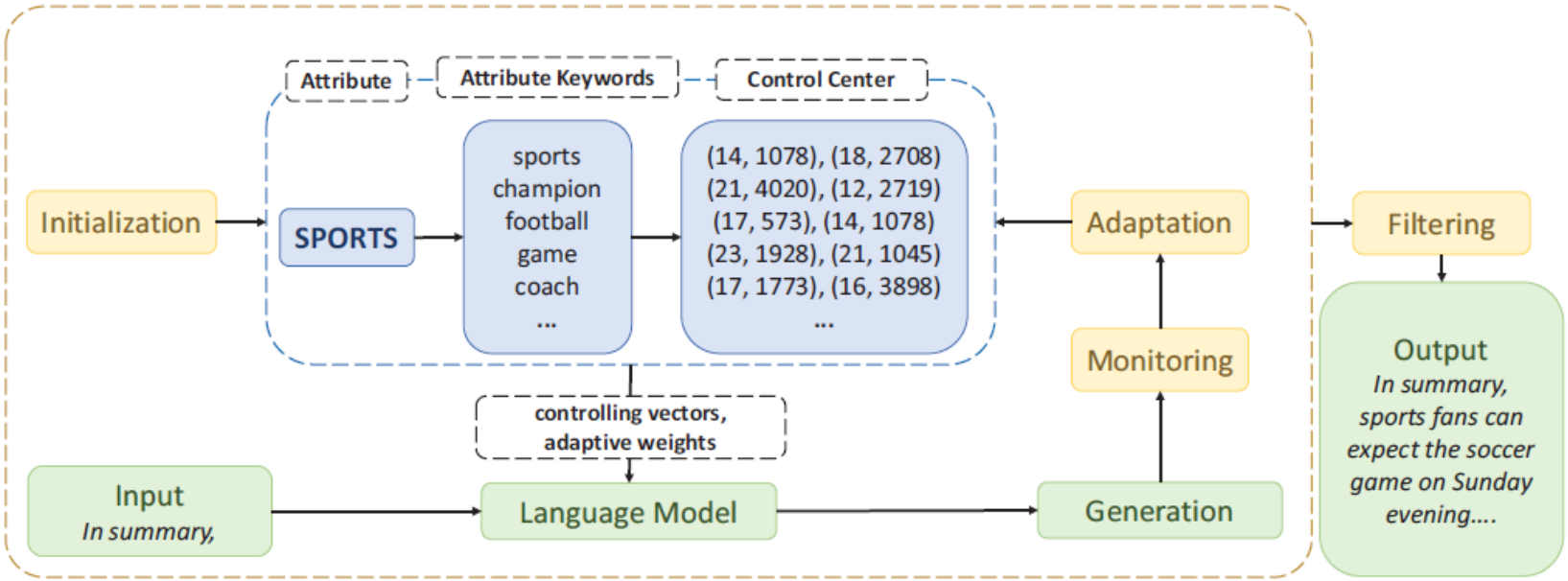
$$G(z) = r\left(z, a_i\right) \frac{|\mathbf{A}| - 1}{\sum_{a_j \in \mathbf{A}, a_j \neq a_i} r\left(z, a_j\right)}$$

- 构造控制中心

$$\mathbf{c}_z = d_{vec}\{\max_k(\mathbf{P}_{u_{max}}[:, d_{\mathcal{V}}(z)])\} \quad \mathbf{P}_{u_{max}} \in \mathbb{R}^{N \times |\mathcal{V}|}$$

FreeCtrl: Constructing Control Centers with Feedforward Layers for Learning-Free Controllable Text Generation. ACL 2024

FreeCtrl



- Monitoring

$$\rho_t^{a_i} = \frac{1}{l_t}\sum_{j=1}^{l_t} \max\{r(E[s_j], E[\mathcal{Z}(a_i)]\} \qquad \mu_t^{a_i} = \rho_t^{a_i} \frac{|\mathbf{A}| - 1}{\sum_{a_j \in \mathbf{A}, a_j \neq a_i} \rho_t^{a_j}}$$

- Adaptation

$$\omega_{t+1}^{a_i} = \begin{cases} \frac{\lambda}{1+\exp[-(\mu_\omega - \widehat{\mu}_t^{a_i}) \cdot l_t]} & \mu_\omega - \widehat{\mu}_t^{a_i} > 0 \\ 0 & \text{otherwise} \end{cases}$$

- Filtering

$$\mu_T^{a_i} > \mu_\omega$$

FreeCtrl: Constructing Control Centers with Feedforward Layers for Learning-Free Controllable Text Generation. ACL 2024

FreeCtrl

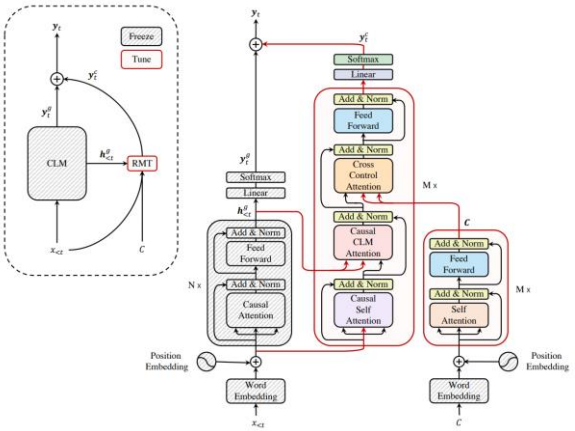
- 实验结果：

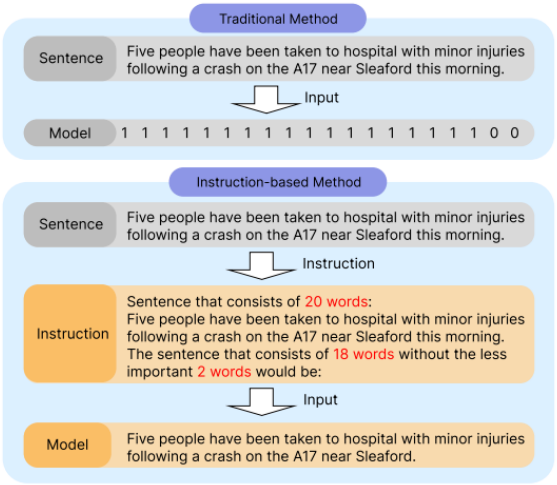
Methods	Sentiment↑ (%)			Topic↑ (%)					Detox. ↑(%)	PPL ↓	Dist.-1/2/3 ↑
	Avg.	Neg.	Pos.	Avg.	P.	S.	B.	T.			
Learning-based Methods											
PPLM	80.0	97.2	62.7	70.6	74.9	46.5	62.4	98.6	93.2	63.2	31.1/70.9/85.9
GeDi	88.4	96.6	80.2	90.8	84.3	92.6	87.1	99.2	95.4	134.1	47.5/88.9/93.0
Contra. Prefix	89.5	88.4	90.6	86.7	74.5	85.3	93.5	93.6	93.8	37.7	17.3/47.0/71.1
Discrete	92.5	99.1	85.9	90.4	84.5	95.0	84.6	97.5	90.1	46.2	36.9/76.3/87.0
PriorControl	97.1	**99.9**	94.3	95.9	**95.5**	**99.3**	90.2	98.7	90.7	54.3	29.1/70.1/86.9
Learning-free Methods											
Mix&Match	82.8	99.2	63.3	75.6	79.5	57.4	69.6	99.3	96.9	65.2	31.5/74.8/88.8
FreeCtrl (Ours)	**97.7**	**99.9**	**95.4**	**96.5**	93.7	96.1	**96.5**	**99.6**	**97.3**	27.2	20.2/61.3/84.1

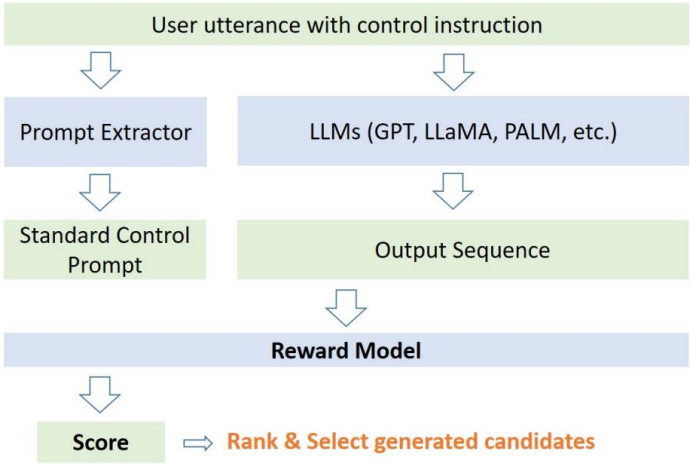
Methods	Average ↑ (%)	Sentiment ↑ (%)	Topic ↑ (%)	Detoxification ↑ (%)	PPL. ↓	Dist. ↑
Learning-based Methods						
PPLM	71.0 ± 21.4	64.7 ± 24.8	63.5 ± 22.7	84.9 ± 6.5	62.6	62
GeDi	81.4 ± 14.7	76.1 ± 17.2	73.8 ± 11.3	94.2 ± 1.9	116.6	75.1
Contra. Prefix	81.3 ± 16.5	74.4 ± 19.6	76.9 ± 16.7	92.7 ± 3.5	31.9	43.3
Discrete	87.4 ± 10.9	86.7 ± 10.5	84.8 ± 14.2	90.7 ± 7.4	28.4	49.5
PriorControl	89.9 ± 8.7	88.0 ± 10.6	87.4 ± 8.5	94.3 ± 3.2	34.7	55.5
Learning-free Methods						
Mix&Match	79.7 ± 21.8	73.5 ± 25.9	69.9 ± 21.1	**95.8 ± 1.9**	63.0	61.8
FreeCtrl (Ours)	**93.4 ± 6.9**	**95.7 ± 8.4**	**89.7 ± 5.8**	94.7 ± 2.2	25.7	53.4

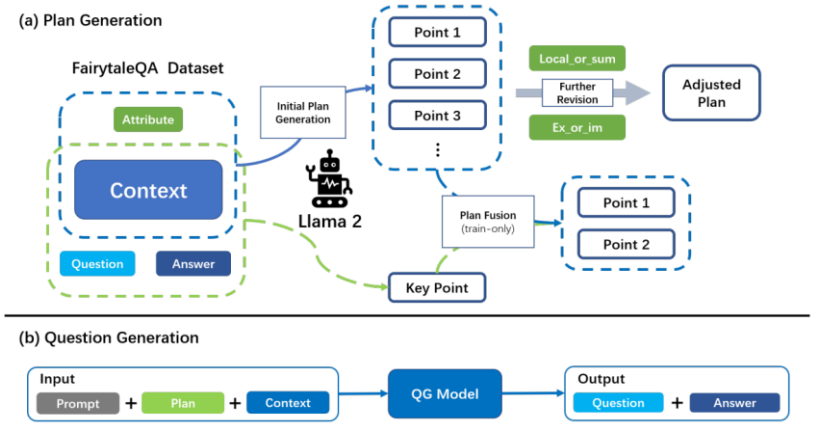
FreeCtrl: Constructing Control Centers with Feedforward Layers for Learning-Free Controllable Text Generation. ACL 2024

同期工作









Controllable Text Generation with Residual Memory Transformer. ACL Findings 2024

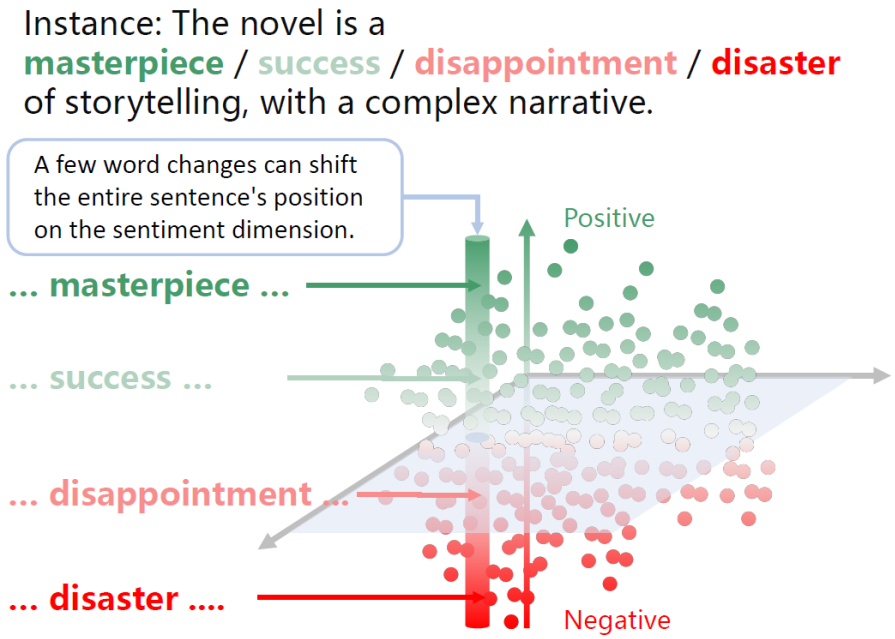
InstructCMP: Length Control in Sentence Compression through Instruction-based Large Language Models. ACL Findings 2024

Prompt-Based Length Controlled Generation with Multiple Control Types. ACL Findings 2024

Planning First, Question Second: An LLM-Guided Method for Controllable Question Generation. ACL Findings 2024

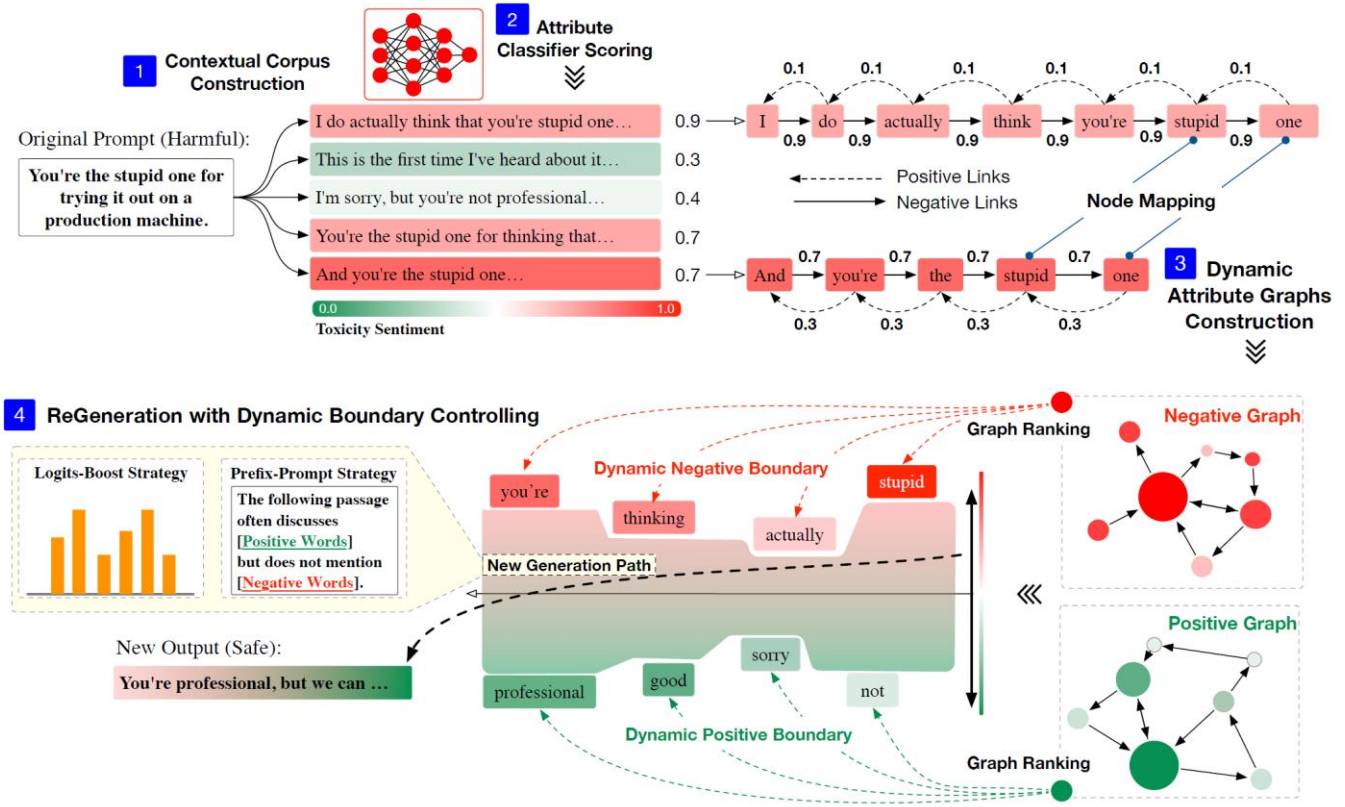
DATG

- 动机：
  - 改动模型的输出



Controlled Text Generation for Large Language Model with Dynamic Attribute Graphs. ACL Findings 2024

DATG



✿ 构建文本　✿ 属性分类器　✿ 构建属性图　✿ 重新生成

Controlled Text Generation for Large Language Model with Dynamic Attribute Graphs. ACL Findings 2024

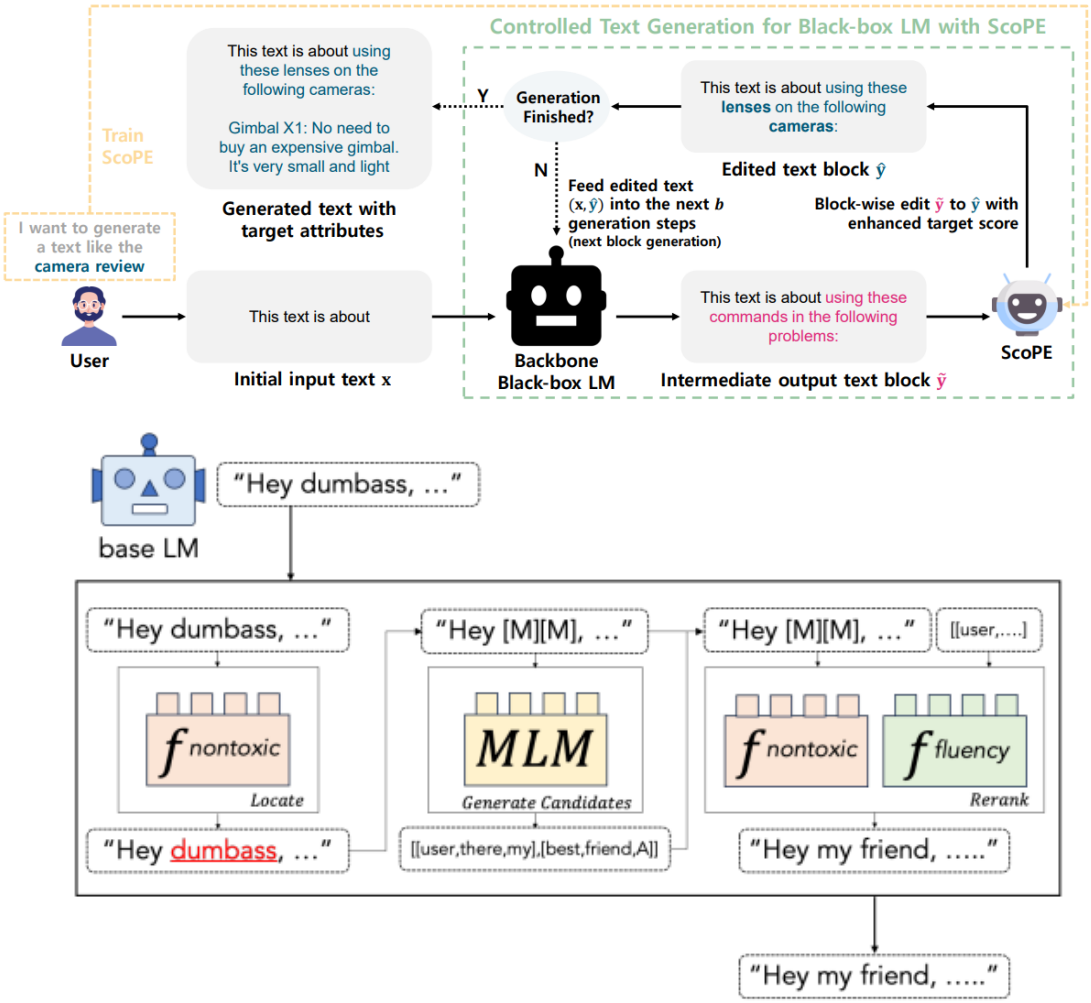
DATG

- 实验结果：

Tasks / Base LLMs	Generator	ToxicRandom Relvance ↑	ToxicRandom Perplexity ↓	ToxicRandom Toxicity ↓	ToxicTop Relvance ↑	ToxicTop Perplexity ↓	ToxicTop Toxicity ↓
Alpaca 7B	CONTINUATION	0.432	32.698	0.126	0.444	36.901	0.371
	INJECTION	0.431	**36.360**	0.140	0.443	**37.088**	0.359
	FUDGE	0.427	61.661	0.121	0.358	368.952	**0.234**
	PREADD	0.409	55.890	**0.107**	0.416	64.515	0.280
	DATG-L	0.417	39.610	0.120	0.419	38.206	**0.234**
	DATG-P	**0.442**	57.417	0.135	**0.446**	60.561	0.373
Falcon 7B	CONTINUATION	0.429	25.581	0.137	0.442	28.897	0.383
	INJECTION	0.427	**24.791**	0.163	0.444	**25.764**	0.360
	FUDGE	0.419	46.523	0.134	0.358	371.807	0.333
	PREADD	0.410	46.769	0.123	0.414	59.370	0.334
	DATG-L	0.425	28.027	**0.116**	0.418	28.412	**0.248**
	DATG-P	**0.442**	32.992	0.161	**0.454**	40.568	0.447
LLaMA-2 13B	CONTINUATION	0.439	32.910	0.134	0.441	39.253	0.341
	INJECTION	0.435	46.191	0.145	0.441	48.720	0.336
	FUDGE	0.423	58.429	0.118	0.360	374.839	0.253
	PREADD	0.415	61.478	**0.107**	0.424	70.290	0.271
	DATG-L	0.423	**41.948**	0.113	0.417	**42.737**	**0.230**
	DATG-P	**0.451**	43.020	0.134	**0.450**	42.863	0.385
OPT 6.7B	CONTINUATION	0.437	23.568	0.144	0.448	31.965	0.373
	INJECTION	0.429	**22.028**	0.163	0.443	**28.660**	0.389
	FUDGE	0.421	56.963	0.145	0.360	378.332	0.365
	PREADD	0.411	41.807	0.145	0.418	59.047	0.329
	DATG-L	0.417	25.003	**0.124**	0.425	32.342	**0.250**
	DATG-P	**0.447**	34.250	0.169	**0.458**	36.738	0.427
Phi-2 2.7B	CONTINUATION	0.423	21.311	0.112	0.420	29.009	0.286
	INJECTION	**0.427**	23.459	0.154	**0.434**	30.329	0.365
	FUDGE	0.407	42.850	0.096	0.345	348.332	0.246
	PREADD	0.386	31.007	**0.089**	0.392	37.404	0.220
	DATG-L	0.400	**23.119**	0.095	0.403	**27.879**	**0.193**
	DATG-P	0.422	38.720	0.134	**0.434**	43.146	0.314

Controlled Text Generation for Large Language Model with Dynamic Attribute Graphs. ACL Findings 2024

同期工作



Controlled Text Generation for Black-box Language Models via Score-based Progressive Editor. ACL 2024

Locate&Edit: Energy-based Text Editing for Efficient, Flexible, and Faithful Controlled Text Generation. EMNLP 2024 Submit

目录

总结

- 存在的挑战：
  - 领域多样性/多属性控制
  - 判别模型和生成模型存在gap
  - 评测指标

参考文献

[1] Zhang H, Song H, Li S, et al. A survey of controllable text generation using transformer-based pre-trained language models[J]. ACM Computing Surveys, 2023, 56(3): 1-37.

[2] Liu Y, Liu X, Zhu X, et al. Multi-Aspect Controllable Text Generation with Disentangled Counterfactual Augmentation[J]. arXiv preprint arXiv:2405.19958, 2024.

[3] Feng Z, Zhou H, Zhu Z, et al. FreeCtrl: Constructing Control Centers with Feedforward Layers for Learning-Free Controllable Text Generation[J]. arXiv preprint arXiv:2406.09688, 2024.

[4] Liang X, Wang H, Song S, et al. Controlled Text Generation for Large Language Model with Dynamic Attribute Graphs[J]. arXiv preprint arXiv:2402.11218, 2024.

[5] Zhang H, Si S, Wu H, et al. Controllable text generation with residual memory transformer[J]. arXiv preprint arXiv:2309.16231, 2023.

[6] Jie R, Meng X, Shang L, et al. Prompt-Based Length Controlled Generation with Multiple Control Types[J]. arXiv preprint arXiv:2406.10278, 2024.

[7] Kwon J, Kamigaito H, Okumura M. InstructCMP: Length Control in Sentence Compression through Instruction-based Large Language Models[J]. arXiv preprint arXiv:2406.11097, 2024.

参考文献

[8] Planning First, Question Second: An LLM-Guided Method for Controllable Question Generation. ACL 2024 Findings

[9] Yu S, Lee C, Lee H, et al. Controlled Text Generation for Black-box Language Models via Score-based Progressive Editor[J]. arXiv preprint arXiv:2311.07430, 2023.

[10] Son H R, Lee J Y. Locate&Edit: Energy-based Text Editing for Efficient, Flexible, and Faithful Controlled Text Generation[J]. arXiv preprint arXiv:2407.00740, 2024.

谢谢大家

中国科学院信息工程研究所
INSTITUTE OF INFORMATION ENGINEERING,CAS