

可变词数×注意力长度（Flex-Attn）方案：架构说明与落地路线图

- 作者：GaoZheng
- 日期：2025-09-26
- 版本：v1.0.0

摘要

阐述可变成本注意力（Flex-Attn）的动机、设计与实现：在合规约束与预算限制下，按需分配注意力计算资源。文中拆解组件与调用关系、关键超参与时间/显存开销，并给出与历史/状态缓存结合的工程实践与调优建议。

- 目标**：把“历史拓扑逐渐增加词数、预测增加词数、数量组合训练”产品化为可学习的历史窗口 L_h 与预测命中上限 L_p ，通过语义×词法门控 + 长度成本实现注意力灵活机制与可控的注意力长度，服务中文知识蒸馏与字符级 RL 的工程交付。
- 业务价值**：在不依赖分词稳定性的前提下，显著提升长词/OOV 边界对齐、训练信号密度与可解释性；蒸馏小模型更稳、更省钱。
- 技术抓手**：两路可学习长度控制 (L_h, L_p) + 词法拓扑 (U 集合) + 离散 SAC (训练期禁 Top-p) + 语义门控与 IDF 降权 + 长度成本正则。
- 上线口径**：先灰度 (10–20%)，同时关闭“单字奖励”、切断演员侧目标字符泄露、启用反向 Trie/Aho-Corasick 加速后缀命中；两周内合入 Auto-U 与一致性正则，形成 2.1 版本基线。

1. 背景与动机

中文字符级决策存在三大硬伤：**奖励稀疏**、**词法边界不稳**、**黑箱不可审计**。此前 v2.0.0 已把“两字命中”升级为“U 上最长命中”。下一步关键是：

- 让历史可见范围与预测命中上限都能学习与自适应；
- 用**数量组合训练 (Curriculum + Mix) **把不同长度的注意力切片学成“内生能力”，而非“写死的窗口”。

2. 变量、集合与数据结构

词法集合

- 词表并集： $\mathcal{C} = \text{Catalog} = \{ \text{chinese_name_frequency_word.json} \cup \text{chinese_frequency_word.json} \}$ (只读)。
- 长度集合： $U = \text{word_length_sets.json}[\text{union.lengths}] \subset \mathbb{N}$ ，建议剔除 $\{1\}$ 。

两类可学习长度

- 历史窗口 $L_h \in H \subseteq U$ ：控制可见上下文与前缀左扩阈值。
- 预测上限 $L_p \in P \subseteq U$ ：控制后缀最长命中的搜索上限（防长词投机）。

反向匹配索引（建议）

- 反向 Trie/DAWG 或 Aho-Corasick（反向）：尾部最长命中“命中即停”，均摊近 $O(1)$ 。
- 缓存策略：最近命中路径 + 最近失败尾段，减少重复匹配。

3. 拓扑与注意力的结构化设计

3.1 历史注意力： L_h 与局部遮罩

观测：

$$x_t(L_h) = [\text{<bos>}] \oplus \text{tail}(\text{prev}, L_h) \oplus [\text{<sep>}] \oplus \chi_t \oplus [\text{<eos>}]$$

- 编码器对 $\text{tail}(\cdot, L_h)$ **局部注意力遮罩**；更早历史压缩为 **sketch**（聚合 embedding/轻量摘要），避免显存炸裂。
- **前缀左扩条件**增加阈值：

$$\exists L \in U \cap [1..|\text{source}|], L \geq L_h, \text{prefix}(\text{source}, L) \in \mathcal{C}.$$

3.2 预测拓扑： L_p 与最长命中上限

后缀命中只在 $L \leq L_p$ 的长度集合内做降序匹配：

$$\exists L \in U \cap [1..|q|], L \leq L_p, \text{tail}(q, L) \in \mathcal{C}.$$

- 限制“越长越加分”的冲动；结合 IDF 降权防高频长词投机。

3.3 命中策略与日志

- **命中即停、最长优先、CJK 断言。**
- 日志结构化（JSONL）：`file,id,seg,len,freq,L_hit,pos,score_lex,score_sem`，支持离线回放与问题定位。

4. MDP 闭合与环境动态（逐字一步）

状态 $s_t = \langle \text{tail}(\text{prev}, L_h), \chi_t, \text{sketch} \rangle$

动作 $a_t \in \mathcal{A}_{\text{mask}}(s_t)$ （合法字符遮罩后集合）

成本 $c(a_t)$ （可为推断 token 成本/长度成本）

转移：写入 a_t 后构造 q ，以 L_p 上限在 U 上匹配，得到命中片段 seg 及注记，生成 s_{t+1}

缓存： $(s_t, L_h, L_p, a_t, r_t, s_{t+1}, d_t) \in \mathcal{D}$

5. 奖励函数与成本函数（ROI 口径）

语义/词法/纯净度（沿用你的 S_t ）

$$\mathcal{N}_\gamma(x) = 1 - (1 - x)^\gamma, \quad S_t = Q_t + L_t - P_t$$

词法增益（语义门控 + IDF）

$$\delta_t = \lambda_{\text{lex}} \cdot \mathbf{1}[\text{hit}(U; \leq L_p)] \cdot \max(0, \text{similarity} - \tau) \cdot w_{\text{IDF}}(\text{seg})$$

规范：**不允许单字奖励**（移除 $L = 1$ 分支）；双字命中降权，高频词按 Zipf/IDF 降权。

长度成本

$$\text{cost}_{\text{len}} = \lambda_h \left(\frac{L_h}{L_h^{\max}} \right)^{\alpha_h} + \lambda_p \left(\frac{L_p}{L_p^{\max}} \right)^{\alpha_p}$$

总奖励（字符模式）

$$R_t = f(C_t) - \lambda_t - \psi_t + S_t + \eta_1 \chi_t^{\text{soft}} + \eta_2 \delta_t - \text{cost}_{\text{len}}$$

$\eta_1, \eta_2, \lambda_{\text{lex}}, \tau, \lambda_h, \lambda_p, \alpha_h, \alpha_p$ 统一入参治理；删除历史混用的 $B_{\text{char}}, \Delta_{\text{char}}$ 影子变量。

6. 策略与价值网络（分层 + 离散 SAC）

层级策略（三头）

- $L_h \sim \pi_{L_h}(L|s) \rightarrow$ 构建 $x_t(L_h)$
- $L_p \sim \pi_{L_p}(L|s, L_h)$
- $a_t \sim \pi_{\text{char}}(\cdot|x_t(L_h))$ (训练期**禁用 Top-p**)

评论家 (可分解 Q)

$$Q(s, L_h, L_p, a) \approx Q_0(s, a) + Q_h(s, L_h) + Q_p(s, L_p)$$

降维、可解释，便于做长度灵敏度分析。

SAC 目标 (离散、遮罩)

$$V(s) = \sum_{a \in \mathcal{A}_{\text{mask}}(s)} \pi(a|s) (\min(Q_1, Q_2) - \alpha \log \pi(a|s))$$

$$J_Q = \mathbb{E} \frac{1}{2} \sum_{i=1}^2 (Q_i(s, a) - [r + \gamma(1-d)V(s')])^2$$

$$J_\pi = \mathbb{E}_s \sum_{a \in \mathcal{A}_{\text{mask}}(s)} \pi(a|s) (\alpha \log \pi(a|s) - Q_1(s, a))$$

$$J_\alpha = -\alpha \cdot \mathbb{E}_s (H_{\text{tgt}} + \sum_a \pi(a|s) \log \pi(a|s)), \quad H_{\text{tgt}} = \kappa \log |\mathcal{A}_{\text{mask}}(s)|$$

同步对 π_{L_h}, π_{L_p} 加熵正则，避免长度崩塌到单点。

一致性要求

- 训练期禁用 Top-p/温度截断**；推理期可开启以保证可读性。
- 演员侧禁用 χ_t 明文**（仅在评论家/奖励侧作为特权信息）。

7. 数量组合训练 (Curriculum + Mix)

阶段化 Curriculum

- Stage-A** (短窗口稳定)：采样较小 L_h, L_p ，先学稠密信用分配。
- Stage-B** (混合过渡)：短中长混合，加入域分布先验。
- Stage-C** (业务贴合)：按域内长度直方图做重要性采样，覆盖真实分布。

混采分布

$$(L_h, L_p) \sim (1 - \rho) \cdot \text{Zipf}(H) \times \text{Zipf}(P) + \rho \cdot \text{Empirical}(H, P), \quad \rho \uparrow$$

稳健性正则 (跨长度不敏感)

同一样本采样两组 (L_h, L_p) 生成两次输出，做一致性损失：

$$\mathcal{L}_{\text{stab}} = \lambda_{\text{stab}} \cdot \|\text{stopgrad}(y^{(1)}) - y^{(2)}\|_1$$

8. 工程实现 (关键流程与配置)

核心伪代码

```
# step t
Lh ~ pi_Lh(s_t)
x_t = build_obs(prev, chi_t, Lh)          # 局部注意力遮罩 + sketch
Lp ~ pi_Lp(s_t, Lh)
a_t ~ pi_char(. | x_t)                   # 训练期禁 Top-p

# 词法拓扑 (反向 Trie/AC)
q, seg = longest_suffix_hit_with_cap(a_t, future_chars, U, cap=Lp)
delta = gated_delta(seg, similarity, IDF, tau) # 语义门控 + IDF/Zipf
R_t = base_reward + eta1*chi_soft + eta2*delta - len_cost(Lh, Lp)

store_transition(s_t, Lh, Lp, a_t, R_t, s_{t+1})
update_critics()
update_policies_with_entropy_targets()
```

配置示例 (片段)

```
{
  "U": [2,3,4,5,6,7,8,9,10,11,13],
  "heads": {"pi_Lh": [2,3,4,6,8], "pi_Lp": [2,3,4,5,6,8]},
  "ban_single_char_reward": true,
  "idf_weighting": true,
  "tau_semantic_gate": 0.72,
  "lambda_lex": 0.25,
  "eta1": 0.15, "eta2": 0.30,
  "lambda_h": 0.1, "lambda_p": 0.12,
  "alpha_h": 1.2, "alpha_p": 1.0,
  "sac": {"gamma": 0.997, "tau_ema": 0.005, "kappa": 0.9, "train_top_p": false}
}
```

日志与回放

- JSONL 单行事件: `step,s, a, Lh, Lp, seg, L_hit, idf, delta, S_t, R_t, file:id, pos`
- 回放脚本: 按 `step` 重建命中链, 定位“词法投机/语义退化”根因。

9. 指标体系与 A/B 设计

核心 KPI

- 语义: ROUGE-L / BERTScore (\geq 基线显著提升或持平不降)。
- 词法: `word_noncompliance` $\downarrow \geq 30$; 合法词覆盖 \uparrow 。
- 稳定: 收敛步数 $\downarrow \geq 15$; 多次训练方差 $\downarrow \geq 20$ 。
- 生产: QPS 不降 > 10 ; 显存可控; 日志写入不成瓶颈。

关键 A/B

- δ_t : 硬奖励 vs **语义门控 + IDF**
- $U = \{2\}$ vs $U = \text{union.lengths}$
- 训练期 Top-p: 开 vs **关**
- 演员可见 χ_t : 是 vs **否**
- 单头策略 vs **层级三头**
- 一致性正则: 无 vs 有

10. 风险评估与防护

风险	触发机制	对策	监控/验收	备注		
长度偏置 (越长越有利)	长词命中奖励放大、策略倾向选择大 L_p	设定 L_p 上限; 加入长度成本 λ_h, λ_p ; 语义门控 (similarity> τ) ; IDF/Zipf 降权	平均 L_p 与长词占比受控; δ_t 占总奖励比 \leq 阈值; ROUGE/BERTScore 持平或提升	禁止单字奖励; 二字命中降权		
词典投机	高频词堆砌触发“命中即停”	禁单字奖励; 二字降权; 动态调优门控阈值 τ ; 停用词/黑词表惩罚	Top-100 高频词占比下降; IDF 加权得分 \uparrow ; word_noncompliance $\downarrow\geq$ 30%	Catalog 定期热更; 命中仅作增益、非硬判		
训练/ 推理分布偏移	训练期使用 Top-p/ 温度截断导致熵目标与可行动作错配	训练禁 Top-p; 遮罩全量期望; 上线前做 Eval-w/o-Top-p 一致性校验; Top-p 仅推理侧启用	训练/推理 KL 差<阈值; $H_{\text{emp}} \approx H_{\text{tgt}}$; 线下指标不劣化	配置审计与门禁 (CI 兜底)		
信息泄露	χ_t 进入演员输入, 学习成复制	χ_t 仅评论家/奖励可见; 演员则剔除; 特权信息门控	Ablation (无/有 χ_t) 差异显著性降低; 复制率/抄写比下降	代码审计规则: 禁止 χ_t 进入策略前向		
吞吐下降	U 线性扫描、日志 I/O 放大	反向 Trie/Aho-Corasick; 命中状态缓存; 日志批量/采样写入; 异步 I/O	tok/s \geq 基线90%; I/O 等待占比<10%; 缓存命中率>90%	Profiling 与压测纳入发布门槛		
长度崩塌	π_{L_h}, π_{L_p} 收敛到单点	熵正则; 长度混采; 一致性正则; Curriculum 分阶段训练	L_h/L_p 熵>阈值; 长度直方图覆盖>80%; 跨长度泛化跌幅<5%	监控长度分布漂移并告警		

11. 上线策略与版本路线图

- **v2.1 (两周)** : 合入 P0 (禁单字奖励/禁训练 Top-p/去泄露/MDP 闭合) + P1 (反向 Trie、Auto-U、一致性正则、结构化日志)。
- **灰度**: 10–20% 流量, 域内高频长词业务优先 (法规/医疗/政务)。
- **退路**: 命中权重热更、 L_p 软上限热更; 异常即切回 v2.0.0 匹配策略。
- **v2.2**: 多目标调度 (语义/词法/可读性 Pareto)、域自适应门控参数 ($\lambda_{\text{lex}}, \tau$)。

12. 小结 (一句话)

把**长度**也做成**一等公民的决策变量**: 用 L_h 管历史注意力、用 L_p 管预测拓扑, 把“语义正确”与“词法成立”通过**门控与成本**拉到同一 ROI 帐本里, 这样学出来的注意力才**灵活、可控、能审计**, 并能稳定地把中文教师知识**蒸馏**到字符级学生里。下一步, 补齐 Auto-U 与一致性正则, 打穿训练-推理闭环, 把能力固化为可复用的产线资产。

许可声明 (License)

Copyright (C) 2025 GaoZheng

本文档采用[知识共享-署名-非商业性使用-禁止演绎 4.0 国际许可协议 \(CC BY-NC-ND 4.0\)](#)进行许可。