零训练表驱动 Flex-Attn：可计算词法 + 有限状态索引的快速落地

- 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。

以 Catalog（域词库）+ 长度集合 U + 反向 Trie/AC 为主干，在**不训练神经网络**的前提下完成中文知识蒸馏与生成控制；语义门控（Jaccard/BM25-lite）+ IDF/Zipf 降权抑制高频堆词，L_h/L_p 由策略表管控成本与粒度。SAC/NN 被重定位为“**动态超参控制 + 学习型索引 + 隶属度快查**”的增值件（A→B→C 灰度演进）。优势：强可审计、可回放、TCO 友好；适合监管行业首发。按“业务可交付”的口径，文中将给出**Flex-Attn 在医疗问答场景**的**端到端应用演示**：从配置→一步一决策→命中与奖励→日志与KPI。输入是用户问句“请问什么是‘奥司他韦’？”，目标输出为专业定义句（含多组医学词法片段：磷酸/奥司他韦/神经氨酸酶/特异性/抑制剂…）。

1. 场景与配置（医疗Q&A蒸馏）

目标：把“大模型的医学定义能力”蒸馏到“小模型”，保证术语命中与可读性，抑制“堆词投机”。

关键参数

- 词法集合：$U = \{2, 3, 4, 5, 6, 8\}$（禁 1 字）；Catalog = {“磷酸奥司他韦”, “奥司他韦”, “神经氨酸酶”, “特异性”, “抑制剂”, “宿主细胞”, “流感病毒”, …}（只读、可热更）。
- Flex-Attn 两头：$L_h \in \{2, 3, 4, 6\}$、$L_p \in \{3, 4, 5, 6\}$（历史窗口/预测命中上限）。
- 训练期：**禁 Top-p**；演员侧**不可见** χ_t（目标字符，仅评论家/奖励可见）。
- 奖励门控：$\tau = 0.75$；IDF/Zipf 降权开启（高频词降权）。
- 长度成本：$\lambda_h = 0.08,\ \lambda_p = 0.10,\ \alpha_h = \alpha_p = 1.0$。

2. 一步一决策（示例回放）

> 逐字一步（one-char-per-step）。下列数值仅为“工程级示例”，用于说明信号流向与ROI账本。

Step 1（开篇定义名词）

- 状态：prev = “”（空），用户问句仅用于检索/语义对齐；χ_t（仅评论家可见）= “磷”。
- 采样：$L_h \sim \pi_{L_h} \Rightarrow 2$；$L_p \sim \pi_{L_p} \Rightarrow 6$。
- 动作：演员出字 **“磷”**（训练期无 Top-p）。
- 后缀拓扑（带上限）：从“磷”向右串接真实未来字符，反向Trie在 U 且 $L \le 6$ 上匹配到 **“磷酸奥司他韦”**（L=6，最长可用命中，命中即停）。
- 语义门控：similarity=0.88>τ=0.75，放行；IDF(“磷酸奥司他韦”)=0.90。
- 词法增益：$\delta = \lambda_{\text{lex}} \cdot 1 \cdot (0.88 - 0.75) \cdot 0.90$，取 $\lambda_{\text{lex}} = 0.30 \Rightarrow \delta \approx 0.035$。
- 长度成本：$\text{len\_cost} = \lambda_p (L_p / L_p^{\max}) = 0.10 \times (6/8) = 0.075$（示例）。
- 当步奖励：$R_1 = \text{base} + \eta_1 \chi^{soft} + \eta_2 \delta - \text{len\_cost}$（$\eta_1, \eta_2$ 按配置表给值）。

Step 2（定义结构“是一种”）

- $L_h \Rightarrow 3$，$L_p \Rightarrow 3$（策略自适应把预测上限收紧到短词）。
- 出字 **“是”**；后缀命中“**是一种**”（L=3 ∈ U）。
- 语义门控通过；IDF(“是一种”)=0.55（高频短语，降权）；δ 较小；len_cost 降低（$L_p = 3$）。

Step 3（机理短语“作用于…”）

- $L_h \Rightarrow 4$，$L_p \Rightarrow 3$。
- 出字 **“作”**；命中“**作用于**”（L=3）$\rightarrow$ δ 中等；继续进入机理对象。

Step 4（关键靶点“神经氨酸酶”）

- $L_h \Rightarrow 6$（需要更长历史以承接机理片段），$L_p \Rightarrow 4$。
- 出字 **“神”**；命中“**神经氨酸酶**”（L=4），IDF=0.87，δ 较高；语义门控强通过（similarity≈0.90）。

Step 5（性质“特异性”）

- $L_h \Rightarrow 4$，$L_p \Rightarrow 3$。
- 出字 **“特”**；命中“**特异性**”（L=3），IDF=0.8，δ 中高；合并上一部件成为“特异性抑制剂”的前缀。

Step 6（类别“抑制剂”）

- $L_h \Rightarrow 4$，$L_p \Rightarrow 3$。
- 出字 **“抑”**；命中“**抑制剂**”（L=3），IDF=0.82，δ 中高；完成第一分句核心谓词结构。

Step 7（逗号）

- 标点“，”；**不计词法命中**，但进入“洁净度/可读性”分量（P_t 轻微惩罚），保持句法节奏。

Step 8（重复靶点与作用——解释因果链）

- 继续“其抑制神经氨酸酶的活性，可以抑制成熟的流感病毒脱离宿主细胞，从而抑制流感病毒在人体内的传播”。
- 命中片段包括：**“神经氨酸酶”**（再次命中，去重加权）、**“流感病毒”**（L=3，高IDF）、**“宿主细胞”**（L=4）等；
- 语义门控在机理-结果链路上给出更高 similarity（>0.85），δ 按 IDF 加权；标点“、”“。”不计词法命中。

> 小结：策略通过 L_p 上限**动态收紧/放宽**来“控粒度”：遇**核心术语**放宽到 4–6 命中“长词”，遇**功能词/句式**收紧到 3 控制成本并避免“长词投机”。L_h 在结构转换点自适应上调以承接前后语义。

3. 结构化日志（可审计）

> 产线落地必须能“回放”每一步的命中与门控。示例 JSONL：


```
{"step":1,"Lh":2,"Lp":6,"a":"磷","seg":"磷酸奥司他韦","len":6,"idf":0.90,"sim":0.88,"delta":0.03
{"step":2,"Lh":3,"Lp":3,"a":"是","seg":"是一种","len":3,"idf":0.55,"sim":0.82,"delta":0.011,"len
{"step":3,"Lh":4,"Lp":3,"a":"作","seg":"作用于","len":3,"idf":0.70,"sim":0.84,"delta":0.019,"len
{"step":4,"Lh":6,"Lp":4,"a":"神","seg":"神经氨酸酶","len":4,"idf":0.87,"sim":0.90,"delta":0.039,
{"step":5,"Lh":4,"Lp":3,"a":"特","seg":"特异性","len":3,"idf":0.80,"sim":0.86,"delta":0.026,"len
{"step":6,"Lh":4,"Lp":3,"a":"抑","seg":"抑制剂","len":3,"idf":0.82,"sim":0.85,"delta":0.025,"len
{"step":7,"Lh":4,"Lp":3,"a":", ","seg":"—","len":0,"idf":0.00,"sim":0.00,"delta":0.000,"len_cos
{"step":12,"Lh":6,"Lp":4,"a":"流","seg":"流感病毒","len":3,"idf":0.88,"sim":0.87,"delta":0.032,"
{"step":18,"Lh":6,"Lp":4,"a":"宿","seg":"宿主细胞","len":4,"idf":0.86,"sim":0.86,"delta":0.030,"
{"step":28,"Lh":4,"Lp":3,"a":"。 ","seg":"—","len":0,"idf":0.00,"sim":0.00,"delta":0.000,"len_co
```

4. 机制到KPI的映射（为什么有效）

- **词法对齐**↑：长词（6/4/4）在关键位点命中，**降低解释错误**（如把“神经氨酸酶”误写为“神经酸酶”）。
- **分布稳定**↑：训练禁 Top-p + 动作遮罩，熵目标与可行动作集合一致；上线仅推理侧开 Top-p 保可读性。
- **成本受控**：L_p 在术语处放宽、功能词处收紧 + 长度成本，**吞吐/QPS 波动可控**。
- **投机抑制**：IDF/Zipf 降权 + 语义门控（similarity>τ）杜绝“堆高频长词”。

5. 可直接拷贝的最小骨架（伪码）

```
# 训练期（无 Top-p、演员看不到 chi_t）
Lh ~ pi_Lh(s_t);                x_t = build_obs(prev, Lh)
# 局部注意力 + sketch
Lp ~ pi_Lp(s_t, Lh);            a_t ~ pi_char(. | x_t)

q, seg = longest_suffix_hit_with_cap(a_t, future, U, cap=Lp)
# 反向Trie/AC，命中即停
delta = lambda_lex * 1[hit(seg)] * max(0, similarity - tau) * idf(seg)
R_t = base + eta1*chi_soft + eta2*delta - len_cost(Lh, Lp)

update_critics_and_policies_with_discrete_SAC_masked()
# H_tgt = kappa*log|A_mask(s)|
```

6. 上线口径（验收阈值）

- 语义：BERTScore/ROUGE-L ≥ 基线，医学术语召回率（术语覆盖）≥ +10%。
- 词法：`word_noncompliance` ↓≥30%；“错误术语/错别字率”显著下降。
- 稳定：收敛步数 ↓≥15%，多次训练方差 ↓≥20%。
- 生产：训练 tok/s ≥ 基线90%；Eval-w/o-Top-p 与线上指标偏差<阈值；日志 I/O 不成瓶颈。

一句话总结：把“奥司他韦”类问答做稳，关键不是堆词，而是**让 L_p 在术语处放宽、句法处收紧**，用**语义门控×IDF**给奖励“加闸”，再配合**训练期禁 Top-p**与**演员去泄露**保证分布一致性。这样蒸馏出来的小模型既能**命中医学长词**，又能**控成本、可审计**，适合规模化上线。

许可声明 (License)