字符模式 SAC 的工程实现与数学化描述v3.0.1

- 作者：GaoZheng

- 日期：2025-09-27
- 版本：v1.0.0

摘要

在 v3.0.0 基于“拓扑词包（向前）+ 多字符迭代（向后）”的框架上，v3.0.1 进一步强调“尾缀的可词包性”：不仅向前拓扑在 $s = \chi_t \oplus q$ 的尾部可匹配词包，向后的“迭代尾缀”也允许直接对“后缀词包”命中，从而以统一的“词包语义”覆盖前后两个方向。本文给出后缀词包的形式化定义、与多字符迭代的融合伪代码、配置与日志扩展，以及回滚与评审要点，确保升级在可观测性、稳定性与合规治理下落地。

- 新增“后缀词包命中”：在行为前缀轨迹 q 的尾部（后缀）对“词包短语”进行最长命中；
- 多字符迭代算子支持“直接命中词包或 Catalog”，`\mathcal H` 可配置为“Catalog/词包/并集”；
- 配置新增/扩展：在向后路径中引入 `hit_mode` 与词包归一设置，默认回退行为与 v3.0.0 一致。

---

1. 形式化

- 词表与长度集合（承接 v2/v3.0.0）：

  $$\mathcal{C} = \texttt{Catalog} = \texttt{chinese\_name\_frequency\_word.json} \cup \texttt{chinese\_frequency\_word.json}, \quad U = \texttt{union.lengths} \subset \mathbb{N}.$$

- 词包族（承接 v3.0.0）：

  $$\mathfrak{P} = \{P_1, \ldots, P_M\}, \;\; P_i = \{\omega_{i,1}, \ldots, \omega_{i,k_i}\}, \; \omega_{i,j} \in \Sigma^+.$$

  $\mathrm{hit}(s, \omega)$ 表示“短语 ω 是否在 s 的指定作用域中命中（尾部/子串）”。
- 文本片段：目标章节 χ_t，上一轮摘要 prev_t，源 $\mathrm{source}_t = \mathrm{prev}_t \oplus \chi_t$，行为前缀轨迹 q。

---

2. 后缀词包算子（强调“尾缀可词包性”）

在 q 的尾部，以 U 约束的长度候选上对词包短语进行命中：

$$\exists L \in U \cap [1..|q|], \; \exists P \in \mathfrak{P}, \; \exists \omega \in P, \; \text{s.t. } \mathrm{tail}(q, L) = \omega.$$

```
function SUFFIX_PACK_HIT(q, U, Packs):
    for L in sort_desc(U ∩ {1..len(q)}):
        seg = tail(q, L)
        if is_cjk(seg) and PACK_HIT(seg, Packs):
            return True, seg
    return False, ""

function PACK_HIT(seg, Packs):
    for P in Packs:
        for ω in P:
            if match_phrase(seg, ω):  # 非交换短语，顺序敏感
                return True
    return False
```

注：`match_phrase` 可配置“最长可用/别名归一/大小写/全半角”，与向前拓扑一致。

---

3. 多字符迭代 × 后缀词包融合

将 v3.0.0 的迭代扩展与“后缀词包命中”合流：

```
function ITER_BACKWARD_EXTEND_WITH_PACKS(initial_char, sample_next, U, Packs, Catalog,
                                         K_max, hit_mode="union", stop_on_hit=True):
    # hit_mode ∈ {"catalog", "packs", "union"}
    def in_H(seg):
        if hit_mode == "catalog":
            return in_catalog(seg)
        elif hit_mode == "packs":
            return PACK_HIT(seg, Packs)
        else:  # union
            return in_catalog(seg) or PACK_HIT(seg, Packs)

    q = dedup_head_repeat(initial_char)
    for step in range(K_max - 1):
        ch = sample_next(q)
        q += ch
        # 尾缀候选（受 U 约束）
        for L in sort_desc(U ∩ {1..len(q)}):
            seg = tail(q, L)
            if is_cjk(seg) and in_H(seg):
                if stop_on_hit:
                    return q, seg, L, step+1
                else:
                    break  # 命中但继续累计，取更长命中
    # 回退到“最长 CJK 尾部”以保读性
    return q, longest_cjk_tail(q, max(U)), min(max(U), len(q)), min(step+1, K_max)
```

接口返回命中短语 `seg`、命中长度 `L` 与迭代步数 `step+1`，便于日志与可视化。

---

4. 配置扩展（向后路径）

在 `res/config.json/config_template.json` 中扩展：

```jsonc
{
  "backward_iter": {
    "enabled": true,
    "k_max": 4,
    "stop_on_hit": true,
    "hit_mode": "union",                // catalog | packs | union
    "packs_path_back": "data/topology_word_packs.json",  // 为空回退为 catalog
    "normalize": { "alias": true, "casefold": true, "fullwidth": true }
  }
}
```

兼容性：`hit_mode="catalog"` 与未配置 `packs_path_back` 时，行为退化到 v3.0.0（乃至 v2）等价路径。

---

5. 日志与奖励对接

- 日志新增（或沿用字段名但含义更广）：
  - `suffix_topo_hit`、`suffix_hit_phrase`、`suffix_pack_id`、`iter_k`；
  - 统计：命中率/命中长度分布/迭代步分布/包覆盖率。
- 奖励：延续 v2/v3 的 δ_t 设计（对“词/词包”命中给 1.0，加权仍可复用现有系数），不改变基础成本与预算项；
  可选在“词包命中”时施加细粒度正则或温度校正作为实验分支（A/B）。

---

6. 复杂度与实现建议

- 词包匹配建议构建 Trie/AC 自动机或哈希桶，均摊近似 $\mathcal{O}(1)$；
- `hit_mode="union"` 场景下先尝试词包再 Catalog（或反之）以减少负载；
- 归一化（大小写/全半角/别名）在装载时完成，热缓存最近命中短语。

---

7. 评审要点与回滚

- 正确性：作用域、早停、命中优先级、归一策略一致；
- 可观测性：新增字段在 CSV/HTML 可见且阈值告警生效；
- 性能：QPS、P95 延迟与内存占用回归在阈值内；
- 回滚：`hit_mode="catalog"` 及 `enabled=false` 立即回退；
- A/B：不同 `hit_mode` 与 `k_max` 的组合对比（命中率、收敛速度、稳定性）。

---

8. 验收指标（建议）

$$
\begin{aligned}
\Delta\text{hit} &:= \text{HitRate}^{\text{pack}}_{v3.0.1} - \text{HitRate}^{\text{word}}_{v3.0.0} \;\geq\; \tau_1, \\
\Delta\text{stability} &:= \text{Var}(\text{reward})_{v3.0.0} - \text{Var}(\text{reward})_{v3.0.1} \;\geq\; \tau_2, \\
\Delta\text{eff} &:= \frac{\text{avg\_reward}}{\text{cost}}\Big|_{v3.0.1} - \frac{\text{avg\_reward}}{\text{cost}}\Big|_{v3.0.0} \;\geq\; \tau_3.
\end{aligned}
$$

未达标需回滚并复盘“词包定义/归一/命中优先级/早停阈值”的配置。
