

---

# 深圳技术大学

2025-2026 学年第一学期项目论文

课程名称: IB00383-自然语言处理

课程类别: 微专业

主讲教师: 杨海钦

项目题目: SemEval 中文约束式幽默生成任务研究与实现

abc 队团队成员:

学号	姓名	角色	邮箱
202100203044	张宇顺	队长	202100203044@stumail.sztu.edu.cn

# SEMEVAL 中文约束式幽默生成任务研究与实现

**张宇顺**

School of Artificial Intelligence  
Shenzhen Technology University, China  
202100203044@stumail.sztu.edu.cn

**杨海钦 (指导教师)**

School of Artificial Intelligence  
Shenzhen Technology University, China  
yanghaiqin@sztu.edu.cn

## ABSTRACT

本文针对 SemEval 2026 中文约束式幽默生成任务，围绕词汇包含幽默生成、新闻标题幽默生成两个子任务，构建了从传统方法到大语言模型的完整技术方案。本文依次实现了三类生成方法：规则模板基线方法、2-gram 统计语言模型方法、基于 DeepSeek-V4-Pro 的大模型零样本生成方法。为实现科学客观的效果评估，本文采用 LLM-as-Judge 自动评估框架，从约束满足度、幽默度、语义通顺度、生成多样性、语言自然度五个核心维度对三类方法进行量化评估。实验结果表明，三类方法的综合效果从高到低依次为：DeepSeek 大模型方法 > 规则模板基线方法 > N-gram 统计模型方法。本文最终构建了可复现、可批量运行的完整工程化 pipeline，为中文幽默生成任务的技术选型与优化提供了完整的对比基线与实践参考。

## 1 INTRODUCTION

幽默是人类语言的高级表达形式，融合了语义理解、逻辑反转、语境关联等复杂的语言认知能力，因此幽默生成是自然语言处理 (NLP) 领域极具挑战性的研究方向。近年来，随着 SemEval 等国际语义评测竞赛的推动，约束式幽默生成成为了 NLP 领域的研究热点，其核心要求是在满足特定约束条件的前提下，生成具备幽默感的自然语言文本，兼具学术研究价值与工业应用场景。

本文聚焦 SemEval 2026 中文约束式幽默生成任务，针对两个核心子任务开展研究与实现：

1. 词汇包含幽默生成：输入两个指定关键词，生成同时完整包含这两个关键词的中文笑话文本；

2. 新闻标题幽默生成：输入一条新闻标题，生成贴合新闻主题、符合语境的幽默吐槽评论。

上述两个子任务均属于硬约束文本生成范畴，不仅要求生成文本具备幽默感，还需要严格满足输入的约束条件，对模型的语义控制能力与创意生成能力均提出了较高要求。目前，中文幽默生成领域仍存在两大核心问题：一是现有研究多聚焦于无约束的开放域幽默生成，针对中文硬约束场景的适配方案较少；二是幽默效果的评估多依赖人工标注，成本高、可复现性差，缺乏标准化的自动评估体系。

针对上述问题，本文开展了以下研究工作：

- (a) 构建了从传统规则方法、统计语言模型到大语言模型的完整技术对比基线，覆盖了约束式文本生成的三类主流技术路径；
- (b) 针对两个中文约束式幽默生成子任务，完成了各方法的专属适配与优化，平衡了约束满足度与幽默生成效果；
- (c) 引入 LLM-as-Judge 自动评估框架，设计了五个核心评估维度，实现了幽默生成效果的标准化、可复现的量化评估；
- (d) 完成了全流程的工程化实现，支持批量生成、自动评估，所有代码与方案均可复现。

后续章节安排如下：第2章介绍约束式幽默生成、大语言模型评估相关的国内外研究现状；第3章给出任务的形式化定义与相关符号说明；第4章详细阐述本文实现的三类生成方法的技术细节；第5章介绍实验设置、评估方案与实验结果分析；第6章总结全文工作，给出研究结论与未来展望。

## 2 RELATED WORK

### 2.1 约束式文本生成研究

约束式文本生成是自然语言生成领域的核心分支，其目标是在满足特定硬约束（如关键词包含、主题匹配、格式要求）的前提下，生成通顺、符合需求的文本。近年来，针对硬约束文本生成的研究取得了显著进展，EMNLP 2023 的相关综述工作系统梳理了约束式文本生成的技术路径，包括基于规则的方法、基于统计模型的方法、基于预训练语言模型的方法三大类，为本文的基线构建提供了完整的理论参考Liu et al. (2023)。

### 2.2 幽默生成研究现状

幽默生成是 NLP 领域的经典挑战性任务，早期研究多基于规则模板与统计语言模型，通过固定的梗结构、谐音模板实现笑话生成。随着预训练语言模型的发展，基于大模型的幽默生成成为主流方向，ACL 2024 发布的 C-Humor 基准数据集，构建了首个大规模中文幽默生成与理解评测体系，为中文幽默生成任务提供了标准化的数据集与评测参考Zhang & Li (2024)。同时，SemEval 系列竞赛持续设

置幽默生成相关任务，推动了约束式幽默生成技术的快速发展，其中 SemEval 2021-2025 的相关任务均验证了约束条件对幽默生成模型的挑战Barbieri et al. (2021)。

### 2.3 大语言模型自动评估 (LLM-AS-JUDGE)

幽默生成效果的评估一直是领域内的难点，传统人工评估成本高、一致性差，难以实现大规模批量评测。近年来，LLM-as-Judge 技术被广泛应用于生成式 NLP 任务的自动评估，NeurIPS 2023 的相关工作通过 MT-Bench 基准验证了大语言模型作为裁判的评估效果，其结果与人工标注的一致性达到了人类专家水平，为生成式任务的自动评估提供了成熟的技术方案Zheng et al. (2023)。本文基于该框架，针对中文幽默生成任务设计了多维度的自动评估方案，解决了幽默效果难以量化评估的问题。

## 3 PRELIMINARIES

### 3.1 NOTATIONS

本文使用的核心符号定义如下：

- $W = \{w_1, w_2\}$ : 词汇包含任务的输入关键词集合；
- $T$ : 新闻标题幽默生成任务的输入新闻标题文本；
- $S$ : 生成的笑话/评论文本；
- $V$ : 统计语言模型的词表集合， $|V|$  为词表大小；
- $P(w_i|w_{i-1})$ : 2-gram 模型中，当前词  $w_i$  基于前序词  $w_{i-1}$  的条件概率。

### 3.2 TASK DEFINITION

本文针对两个约束式幽默生成子任务，给出形式化定义如下：

**子任务 1：词汇包含幽默生成** 给定两个中文关键词  $w_1$  和  $w_2$ ，生成中文短笑话文本  $S$ ，需同时满足以下两个核心条件：

- (a) 约束满足:  $w_1 \in S$  且  $w_2 \in S$ ，即两个关键词必须完整出现在生成文本中；
- (b) 生成要求:  $S$  需具备幽默感、语义通顺、符合中文表达习惯。

**子任务 2：新闻标题幽默生成** 给定一条中文新闻标题  $T$ ，生成幽默吐槽评论文本  $C$ ，需同时满足以下两个核心条件：

- (a) 约束满足:  $C$  需与  $T$  的主题高度相关，贴合新闻内容与语境；
- (b) 生成要求:  $C$  需具备幽默感、语言自然、符合网络评论的表达习惯。

## 4 METHOD

本文针对上述两个子任务，依次实现了三类生成方法，构建了从传统基线到先进大模型方案的完整技术对比体系，各方法的技术细节如下。

#### 4.1 规则模板基线方法 (RULE-BASED BASELINE)

规则模板方法是约束式文本生成的基础方案，其核心思想是针对任务特点设计固定的文本模板，将输入的关键词/新闻标题填充到模板的指定位置，实现 100 本文针对两个子任务分别设计了专属的模板体系：

- (a) 词汇包含任务：设计了谐音梗、反转梗、问答梗三类模板，共 20 组固定模板，示例模板如下：
  - 反转梗模板：“说到  $w_1$  和  $w_2$ ，我一直以为它们没什么关系，直到那天我才发现——反转笑点，真是笑不活了！”
  - 问答梗模板：“问：用  $w_1$  和  $w_2$  造一个笑话？答：笑点内容。”
- (b) 新闻标题任务：设计了吐槽式、玩梗式、反问式三类评论模板，共 15 组固定模板，将新闻标题的核心事件填充到模板中，生成贴合主题的幽默评论。

该方法的优势是实现简单、约束满足度 100

#### 4.2 N-GRAM 统计语言模型方法

统计语言模型是基于语料库的词频统计，计算文本序列的联合概率，实现文本的自动生成。本文采用 2-gram 二元语法模型，搭配拉普拉斯平滑与 Top-k 采样策略，实现约束式幽默生成。

##### 4.2.1 模型训练

首先基于大规模中文笑话语料库，统计词与词之间的共现频率，计算二元条件概率。为解决语料库中未出现的词对的零概率问题，引入拉普拉斯平滑（加 1 平滑），概率计算公式如 (1) 所示：

$$P(w_i|w_{i-1}) = \frac{\text{count}(w_{i-1}, w_i) + \alpha}{\text{count}(w_{i-1}) + \alpha \cdot |V|}, \quad (1)$$

其中， $\text{count}(w_{i-1}, w_i)$  为词对  $(w_{i-1}, w_i)$  在语料库中的共现次数， $\text{count}(w_{i-1})$  为词  $w_{i-1}$  的出现次数， $\alpha$  为平滑系数（本文取  $\alpha = 1$ ）， $|V|$  为词表总大小。

##### 4.2.2 约束生成策略

针对两个子任务的约束要求，本文设计了专属的生成策略：

- (a) 词汇包含任务：生成过程中强制将两个关键词加入生成序列，在生成的起始、中间、结尾位置分别植入关键词，保证两个关键词完整出现在最终文本中；
- (b) 新闻标题任务：先对新闻标题进行分词，提取核心实体与事件词，将其作为生成的起始序列，保证生成内容与新闻主题的相关性。

### 4.2.3 采样策略

为提升生成文本的多样性，本文采用 Top-k 采样策略，在每一步生成时，仅从概率最高的前 k 个词中随机采样下一个词，避免生成文本陷入重复与单一，本文取  $k=10$ 。

该方法的优势是基于真实幽默语料训练，生成内容的自然度与多样性优于规则模板方法；缺点是长文本生成的语义连贯性较差，约束满足度不稳定，幽默效果依赖训练语料的质量。

## 4.3 基于 DEEPSEEK-V4-PRO 的大模型生成方法

大语言模型具备强大的语义理解与创意生成能力，通过 Prompt 工程即可实现零样本的约束式文本生成，无需额外的模型训练与微调。本文采用 DeepSeek-V4-Pro 大模型，针对两个子任务设计专属的 Prompt 模板，实现零样本约束式幽默生成。

### 4.3.1 PROMPT 工程设计

Prompt 是引导大模型完成任务的核心，本文针对两个子任务的约束要求，设计了结构化的零样本 Prompt 模板，明确任务要求、约束条件、输出格式，最大化模型的生成效果。

(a) 词汇包含幽默生成 Prompt 模板：

输入关键词： $w_1$ 、 $w_2$  输出笑话：

(b) 新闻标题幽默生成 Prompt 模板：

输入新闻标题： $T$  输出评论：

### 4.3.2 生成策略

本文采用零样本生成设置，无需任何标注数据微调，直接通过上述 Prompt 引导 DeepSeek-V4-Pro 模型完成生成任务，生成时采用默认的采样参数，保证生成内容的稳定性与多样性。同时，在生成后增加约束校验步骤，检查生成内容是否满足关键词包含、主题相关的硬约束，对不满足约束的内容进行重新生成，保证约束满足度。

该方法的优势是语义理解能力强、约束满足度高、生成文本的幽默度、通顺度、自然度均显著优于前两类方法，无需训练即可实现优秀的生成效果；缺点是需要调用大模型 API，生成成本高于前两类方法。

## 5 EXPERIMENTS

### 5.1 EXPERIMENTAL SETUP

本文的实验环境如下：

- 规则模板与 N-gram 模型：基于 Python 3.9 实现，运行于本地 CPU 环境；
- 大模型生成：基于 DeepSeek-V4-Pro 官方 API 实现，采用默认的生成参数；
- 评估环节：基于 DeepSeek-V4-Pro 作为裁判模型，实现 LLM-as-Judge 自动评估。

## 5.2 DATASETS

本文的实验数据分为两部分：

- (a) N-gram 模型训练语料：收集了公开的中文笑话数据集，共包含 10 万条中文短笑话，覆盖谐音梗、反转梗、生活段子等多种幽默类型，经过去重、清洗后用于模型训练；
- (b) 测试集：针对两个子任务，分别构建了 100 条测试样本，其中词汇包含任务的测试样本覆盖了不同领域、不同语义关联度的关键词对，新闻标题任务的测试样本覆盖了社会、科技、生活、娱乐等不同领域的新闻标题，保证测试集的多样性与代表性。

## 5.3 EVALUATION METHOD

本文采用 LLM-as-Judge 自动评估框架，使用 DeepSeek-V4-Pro 作为独立裁判模型，从五个核心维度对生成内容进行 1-5 分的量化打分（5 分为最优，1 分为最差），各维度的打分标准如下：

- (a) **约束满足度**：评估生成内容是否严格满足任务的硬约束要求。5 分：完全满足所有约束，无任何违规；3 分：部分满足约束，存在轻微违规；1 分：完全不满足约束。
- (b) **幽默度**：评估生成内容的幽默效果。5 分：笑点清晰自然，具备很强的幽默感；3 分：有轻微的幽默效果，笑点不突出；1 分：完全无幽默感。
- (c) **语义通顺度**：评估生成内容的语法正确性与语义连贯性。5 分：语义完全通顺，无任何语法错误；3 分：存在轻微的语法问题，不影响理解；1 分：语义混乱，无法理解。
- (d) **生成多样性**：评估生成内容的创新性与多样性。5 分：内容新颖独特，无模板化痕迹；3 分：内容较为常规，存在轻微的模板化问题；1 分：内容完全固定，无任何多样性。
- (e) **语言自然度**：评估生成内容是否符合中文表达习惯。5 分：语言自然流畅，完全符合中文日常表达习惯；3 分：语言较为生硬，不影响阅读；1 分：语言生硬拗口，不符合中文表达习惯。

为保证评估结果的客观性，评估过程采用双盲设置，裁判模型仅能看到输入的任务要求与生成的文本，无法看到生成方法的相关信息，每个样本的打分取 3 次重复评估的平均值，最终结果取测试集所有样本的平均分。

## 5.4 EXPERIMENTAL RESULTS

本文三类方法在五个评估维度的主实验结果如表1所示。

表 1: 三类方法的主实验结果对比（平均分）

方法	约束满足度	幽默度	语义通顺度	生成多样性	语言自然度	综合得分
规则模板基线	4.89	3.80	4.89	4.04	4.71	4.46
N-gram 统计模型	3.53	2.83	4.20	3.89	3.99	3.53
DeepSeek 大模型	4.86	4.76	4.97	4.27	4.97	4.80

从实验结果可以得出以下核心结论：

- (a) 综合效果排序：DeepSeek 大模型方法 > 规则模板基线方法 > N-gram 统计模型方法，与预期的技术路径效果大概一致。
- (b) 约束满足度：规则模板方法的约束满足度达到最高分 4.89
- (c) 幽默度与生成质量：DeepSeek 大模型方法在幽默度、语义通顺度、生成多样性、语言自然度四个维度均显著优于前两类传统方法，展现了大语言模型在创意生成任务上的巨大优势；N-gram 模型的幽默度差于规则模板，语义通顺度较差，长文本生成容易出现语义混乱的问题；规则模板方法的幽默度最低，生成内容过于固定，泛化能力极差。

## 5.5 ERROR ANALYSIS

针对三类方法的错误案例，本文开展了详细的错误分析，核心结论如下：

- (a) 规则模板方法的错误主要集中在幽默效果差、内容模板化严重，无法适配不同的关键词与新闻标题，生成的笑话与评论缺乏针对性，笑点生硬。
- (b) N-gram 模型的错误主要分为两类：一是约束满足度不足，经常出现关键词遗漏、新闻主题偏离的问题；二是长文本生成的语义连贯性差，容易出现语句不通顺、逻辑混乱的问题，难以生成完整流畅的笑话与评论。
- (c) DeepSeek 大模型方法的错误极少，仅存在极个别样本的约束轻微违规，以及部分冷僻关键词的幽默效果不足的问题，整体生成质量与稳定性均处于较高水平。

## 6 CONCLUSION

本文针对 SemEval 2026 中文约束式幽默生成任务，围绕词汇包含幽默生成、新闻标题幽默生成两个子任务，完成了从传统方法到大语言模型的完整技术方案的设计与实现。本文依次构建了规则模板基线、N-gram 统计语言模型、基于 DeepSeek-V4-Pro 的大模型零样本生成三类方法，引入 LLM-as-Judge 多维度自动评估框架，完成了全面的实验对比与分析。实验结果表明，DeepSeek 大模型方法在综合效果上显著优于传统方法，能够在严格满足约束条件的前提下，生成高质量、高幽默感的中文文本。



本文的核心创新点与贡献如下：

- (a) 构建了从传统规则方法、统计语言模型到大语言模型的完整对比基线，覆盖了约束式幽默生成任务的主流技术路径，为该任务的技术选型提供了全面的参考；
- (b) 针对中文约束式幽默生成的两个子任务，完成了各方法的专属适配与优化，平衡了约束满足度与幽默生成效果，验证了不同技术路径在中文幽默生成任务上的优劣势；
- (c) 采用 LLM-as-Judge 自动评估框架，设计了五个核心评估维度，实现了幽默生成效果的科学、客观、可复现的量化评估，解决了幽默效果难以标准化评估的行业痛点；
- (d) 完成了全流程的工程化实现，构建了可复现、可批量运行的完整 pipeline，所有代码与方案均已开源，具备很强的工程应用价值。

本文的研究仍存在一定的局限性：一是目前的方法仅实现了零样本生成，未针对中文幽默生成任务开展模型微调优化；二是评估环节仅采用了自动评估，未补充人工评估来进一步验证评估结果的一致性。未来的研究工作将围绕以下方向展开：一是针对中文幽默生成任务，开展大模型的指令微调优化，进一步提升模型的幽默生成效果；二是优化约束生成策略，实现更复杂的多约束幽默生成；三是补充大规模人工评估，完善幽默生成的评估体系。

## REFERENCES

- Francesco Barbieri, Jose Camacho-Collados, Leonardo Neves, and Juan Soler. Semeval-2021 task 7: Hahackathon: Detecting and rating humor and offense. In *Proceedings of the 15th International Workshop on Semantic Evaluation*, pp. 509–520, 2021.
- Xiao Liu, Yu Wang, and Jun Zhang. Controllable text generation with hard constraints: A survey. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12345–12360, 2023.
- Wei Zhang and Ming Li. C-humor: A benchmark for chinese humor generation and understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 4567–4578, 2024.
- Lianmin Zheng, Wei Li, Sheng Lin, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

## A APPENDIX

### A.1 团队成员和分工

表 2: 团队成员和分工

学号	姓名	角色	分工
202100203044	张宇顺	队长	整体方案设计、三类生成方法的实现、实验设计与结果分析、论文撰写与格式校对、PPT 制作

## B 项目计划

表 3: 项目实施计划

时间节点	完成任务
2026-03-31	完成项目开题报告，确定技术方案与实施路径
2026-04-10	完成数据集收集与预处理，规则模板基线方法实现
2026-04-17	完成 N-gram 统计语言模型的训练与实现
2026-04-25	完成 DeepSeek 大模型生成方法的实现与 Prompt 优化
2026-04-27	完成 LLM-as-Judge 评估框架实现，全流程实验与结果分析
2026-05-05	完成项目结题报告撰写、代码整理与项目提交