

CCL25-Eval任务9系统报告：中医辨证辨病及处方生成中的少样本数据增强方法

左梓呈^{1,2} 任佳敏^{1,2} 吐尔地·托合提^{1,2*}

¹新疆大学，计算机科学与技术学院/ 中国，乌鲁木齐，830017

²新疆信号检测与处理重点实验室/ 中国，乌鲁木齐，830017

turdy@xju.edu.cn

摘要

中医药在临床诊断和治疗中发挥了不可或缺的作用。中医辨证辨病及中药处方生成任务包含两个富有挑战性的问题，包括中医多标签辨证辨病和中药处方推荐。由于缺乏高质量的标注数据，之前的方法大多需要引入外部数据，容易出现知识滞后的问题。因此，我们提出了一种融合大模型与可控文本生成的混合增强策略。具体而言，通过设计基于词汇独立性的数据增强，与微调大模型进行可控文本生成，在少量标注样本的基础上构建高质量扩展数据集。然后采用LoRA微调技术适配此任务。实验结果表明，该方案分别获得了0.553和0.4515的得分。在不需要引入额外数据的情况下，也能获得较好的效果。

关键词： 中医辨证辨病；中药处方生成；可控文本生成；大语言模型；数据增强

System Report for CCL25-Eval Task 9: Few-shot Data Augmentation Methods in Traditional Chinese Medicine Syndrome and Disease Differentiation and Prescription Generation

Zicheng Zuo^{1,2} Jiamin Ren^{1,2} Turdi Tohti^{1,2*}

¹School of Computer Science and Technology, Xinjiang University / Urumqi 830017, China

²Xinjiang Key Laboratory of Signal Detection and Processing / Urumqi 830017, China

turdy@xju.edu.cn

Abstract

Traditional Chinese Medicine has played an indispensable role in clinical diagnosis and treatment. The tasks of syndrome differentiation and disease diagnosis, as well as Traditional Chinese Medicine prescription generation, involve two challenging problems: multi-label syndrome-disease classification and herbal prescription recommendation. Due to the lack of high-quality annotated data, previous methods often rely on external resources, which may lead to outdated or inconsistent knowledge. To address this, we propose a hybrid augmentation strategy that integrates large language models with controllable text generation. Specifically, by designing data augmentation based on vocabulary independence and fine-tuning large models for controllable text generation, a high-quality extended dataset is constructed on the basis of a small number of annotated samples. We then apply LoRA-based fine-tuning to adapt the model to the task. Experimental results demonstrate that our approach achieves scores of 0.553 and 0.4515, respectively, showing competitive performance without the need for additional external data.

Keywords: Syndrome-disease classification, TCM prescription generation, controllable text generation, large language models, data augmentation

*通讯作者

© 2025 中国计算语言学会/ 中国计算语言学大会论文集, 第347页-第354页, 济南, 中国, 2025年8月11日至14日。

根据《Creative Commons Attribution 4.0 International License》许可出版
(c) 2025 中国中文信息学会/ 中国计算语言学专业委员会

1 任务概述

中医辨证辨病及中药处方生成旨在通过给定的患者临床文档（文档内容见附录表6所示），判断患者所患的证型和疾病，为患者推荐合适的中药处方。涉及两项子任务：中医多标签辨证辨病和中药处方推荐。

1.1 子任务一：中医多标签辨证辨病

利用自然语言处理技术从电子病历及症状叙述中自动抽取医学信息，通过语义解析实现主症、兼症与疾病表征的关联性分析，判断患者所患的证型和疾病。包括10种中医基础证型（下称证型）、4种中医疾病（下称疾病）。当一位患者同时存在两种证型时，证型可由两种基础证型组合得到，疾病和证型如表1所示：

疾病	胸痹心痛病, 心衰病, 眩晕病, 心悸病
证型	气虚血瘀证, 痰瘀互结证, 气阴两虚证, 气滞血瘀证, 肝阳上亢证, 阴虚阳亢证, 痰热蕴结证, 痰湿痹阻证, 阳虚水停证, 肝肾阴虚证

表 1: 中医疾病和证型分类

1.2 子任务二：中药处方推荐

构建基于自然语言理解的智能中药处方推荐模型，输入包含患者现病史、主诉及四诊信息的结构化文本，输出符合中医证候特征的中药配伍方案（不含剂量），共381种中药。

中医多标签辨证辨病任务需要分析患者的特征匹配合适的疾病和证型。训练数据中，疾病和证型的占比情况如图1和图2所示。可以看出，各标签之间呈现出明显的分布不平衡状况。具体而言，在所有疾病和证型的标签中，部分常见疾病和证型占据了绝大部分的比例，呈现在证型数据中出明显的长尾分布特征。使得模型倾向于过度学习这些高频标签的特征和模式，而对低频标签的学习不足。之前的工作(Hu et al., 2025), (Yan et al., 2025), (Liu et al., 2023), (Dong et al., 2022), (Yang et al., 2024)大多使用外部数据增强模型性能，但外部数据不总是可用的，且需要长期维护以避免知识滞后。因此，我们通过数据增强的方式，平衡各标签的数量。

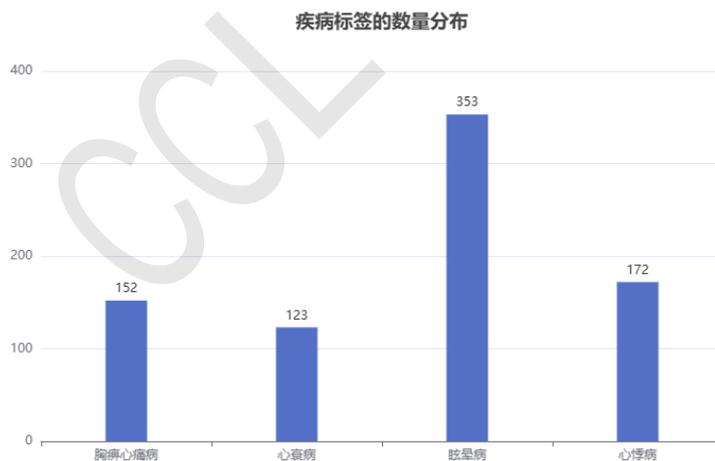


图 1: 疾病标签分布

2 数据增强方法

由于疾病标签类别较少，观察患者信息(症状，中医望闻切诊)可以发现，内容由专业性较强的若干词汇构成，且每个词汇都包含重要信息。之前的方法如反向翻译(Zaheri et al., 2020)、基于词典替换(Zhang et al., 2015)、Mask、随机噪声注入等都不适合此任务。但正是由于患者

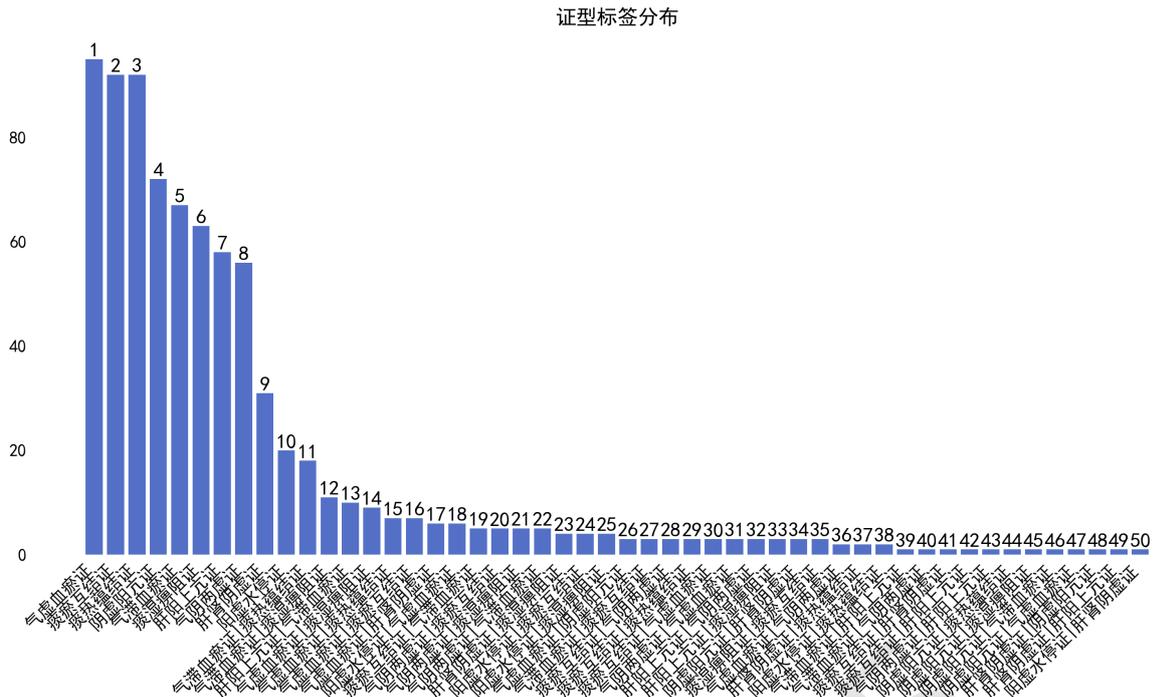


图 2: 证型标签分布, 柱子顶部显示了每个标签的编号

信息由词汇构成, 语序的改变并不影响语义, 所以使用语序替换即可完成数据增强。证型分类体系繁杂多样, 所涵盖的具体类别数量众多。训练数据中包含的基础证型共有10种, 按照常规的数学逻辑, 若将这10种基础证型进行两两排列组合, 所能生成的复合证型数量理论上应为90种。然而, 从图2中可以看出, 其中实际存在的复合证型数量仅有40种。然而, 当面对训练数据集中未曾出现过的复合证型, 仅仅依靠简单的语序替换方法, 显然已不足以满足实际需求。此时, 借助大模型所展现出的强大涌现能力, 能够更为出色地达成数据增强的目标, 生成更具价值和多样化的数据样本。

2.1 疾病数据增强: 词汇独立性

通过观察患者的临床文档可以发现, “症状”和“中医望闻切诊”字段的内容由专业的表达构成, 如“舌暗红、苔黄腻、舌下脉络曲张、脉弦。”其中每个描述词都具有独立的语义, 即词汇独立性, 只要这些关键特征词汇不变, 即便调整顺序, 整体语义依旧不变。因此, 我们通过改变“症状”和“中医望闻切诊”的顺序进行数据增强, 统一各种疾病的样本数量。

2.2 证型数据增强: 可控文本生成

大模型具有广泛的知识储备, 使得它在进行数据增强时能够引入更多的背景知识和细节。我们借助大模型的这一特性, 依托已有的基础证型和复合证型信息, 模拟生成大量合乎逻辑与实际复合证型数据。具体来说,

①首先构建训练样本, 以复合证型为准, 随机抽取两个基础证型的患者信息(如患者A复合证型是气滞血瘀证—痰热蕴结证, 故抽取证型为气滞血瘀证的患者B和证型为痰热蕴结证的患者C)。由于除患者信息中的“主诉”、“症状”和“中医望闻切诊”外, 其余信息包含极大的差异性, 所以仅采用上述三个条目作为输入特征, 输出为证型。

②使用低秩自适应技术(LoRA), 使大模型精准学习基础证型向复合证型转换的特征映射关系。

③推理阶段生成训练数据中未出现的复合证型。

上述阶段使用的提示模板如下:

User:

任务: 根据患者1和患者2的[主诉],[症状],[中医望闻切诊],生成患者3的[主诉],[症状],[中医望闻切诊]。

[患者1]: [主诉]:XXX[症状]:XXX[中医望闻切诊]:XXX

[患者2]: [主诉]:XXX[症状]:XXX[中医望闻切诊]:XXX

Assistant:

[患者3]: [主诉]:XXX[症状]:XXX[中医望闻切诊]:XXX

经过上述过程,我们得到了90种复合证型,使用词汇独立性,平衡各个标签的数量。最终得到的数据作为训练数据。

2.3 处方数据增强: 配伍理论

语言中句子的核心语义由关键成分决定,只要核心成分不变,语序调整语义可基本不变。中医处方也有类似的整体性,其核心功效是由处方中所有草药的综合作用来实现的。

中医有独特的配伍理论,如君臣佐使。君药是针对主病或主证起主要治疗作用的药物;臣药是辅助君药加强治疗主病或主证的药物;佐药是协助君臣药以加强治疗作用,或用以消除或减缓君臣药的毒性与烈性的药物;使药是引导诸药直达病所,或调和诸药的药物。在一个处方中,各草药按照配伍理论组合在一起,它们之间的关系是相对固定的,而不是由顺序决定的。比如在四君子汤(人参、白术、茯苓、甘草)中,人参为君药,白术为臣药,茯苓为佐药,甘草为使药,无论怎样调整它们的顺序,四君子汤健脾益气的功效不会改变,因为其配伍关系决定了整体的功效。

综上所述:中药的配伍主要取决于药物之间的药性、功效、归经以及它们在方剂中的君、臣、佐、使关系,而不是药物在处方中书写的先后顺序。因此,我们在不改变患者信息的情况下,随机打乱各草药之间的顺序,生成新的样本作为扩充后的训练数据。

3 模型构建及训练

在处理上述两个子任务时,我们选用大模型文本生成的方法来开展研究。具体而言,使用训练数据,对基础模型进行指令微调操作,以此构建起一个以HuatuoGPT2-7B(Chen et al., 2023)为基础的中医辨证辨病及处方生成模型。为满足不同子任务的需求,提升模型的适应能力,我们针对各个子任务设计了不同的提示模板。这些提示模板的具体内容可详见表2。

4 实验

本章介绍了增强后数据集的情况,评价指标以及消融实验。

4.1 增强数据集

经过数据增强后的数据集情况如表3所示,其中每种疾病的数量统一为350条;每种基础证型的数量统一为95条;复合证型的数量统一为20条;对原始处方的草药位顺序进行最多5次随机替换,不同的任务使用对应的增强子集。

使用微调模型生成的数据如表4所示,总体来看,生成的数据与真实数据在结构、内容和语义方面具有较高的一致性。具体来说,生成数据与真实数据都有三部分组成(主诉,症状和中医望闻切诊)。说明模型掌握了医疗记录的结构逻辑。生成数据在语义与内容上与真实数据高度对齐,准确还原了主诉、现病史及体征的结构,涵盖了胸痛相关的常见症状与否认信息,细节描述较为丰富;语言风格上,仿写语句通顺、用词规范,贴合临床病历表达习惯;在医学合理性方面,症状组合和体征描述总体逻辑清晰,虽部分辨证细节略显不一致,但整体表现出较强的医学知识掌握与仿写能力。

4.2 评价指标

中医多标签辨证辨病任务采用准确率作为评价指标,中药处方推荐采用Jaccard相似系数、Recall、Precision、F1分数和药物平均数量作为评价指标。其中药物平均数量用于衡量模型推荐的中药方剂数量与真实标签数量的接近程度。计算方法是通过对模型推荐的中药数量和真实标签的中药数量,并计算它们的匹配度。匹配度越高,表示模型推荐的中药数量越接近真实标签的数量。

子任务	任务提示模板
中医多标签辨证辨病	<p>任务: 根据患者的[主诉],[中医望闻切诊],[病史]信息,在[证型]中辨别患者的[主证和兼证] (辨病任务时, 将证型替换为疾病, 不包含要求2)。</p> <p># 要求1: 证型局限在下方列表中。 [证型]: ‘气虚血瘀证’, ‘痰瘀互结证’, (辨病任务时为‘胸痹心痛病’, ‘心衰病’,)</p> <p># 要求2: 有一个或两个证型, 第二个是兼证, 兼证可能没有, 每个证型之间用竖线分隔。</p> <p># 要求3: 没有的证型不要多写。</p> <p># 要求4: 输出中仅需要输出证型名称,不需要给出任何解释和其他信息, 数量控制在1-2个左右。</p>
中药处方推荐	<p>任务: 根据患者的[基本信息],[主诉],[症状],[中医望闻切诊]信息,在[草药]中为患者推荐需要使用的[推荐草药]。</p> <p># 要求1: 草药局限在下方列表中。 [草药]: ‘冬瓜皮’, ‘沉香’, ‘茜草炭’,)</p> <p># 要求2: 有多个中草药, 每个中草药之间用逗号分隔。</p> <p># 要求3: 没有的中草药不要多写。</p> <p># 要求4: 输出中仅需要输出草药名称,不需要给出任何解释和其他信息, 数量控制在10-15个左右。</p> <p>[基本信息]患者性别为XXX 年龄为XX 发病节气在XX。[主诉]:XX[症状]:XX[中医望闻切诊]:XX[推荐草药]:XXX</p>

表 2: 不同子任务的提示模板

数据集	疾病数量	证型数量	处方数量
原始数据集	800	800	800
增强后数据集	1400	2750	4000

表 3: 增强后数据对比

数据来源	主诉	症状	中医望闻切诊
模型生成	发作性胸痛1个月余, 加重1周。	发作性胸痛, 伴胸闷气短, 乏力倦怠, 偶有头晕, 时有恶心, 无头痛, 无口干口苦, 纳眠差, 大便稀溏, 日一次, 尿色深黄, 夜尿1次。	表情自然, 面色红润, 形体正常, 动静姿态, 语气清, 气息平; 无异常气味, 舌质红、苔薄白, 脉沉细。
真实数据	胸痛6天余。	患者胸痛, 胸闷, 阵发性心慌, 出汗, 乏力, 双下肢水肿, 无头晕头痛, 无恶心呕吐, 无口干口苦, 纳眠差, 大便调, 小便频数。	表情自然, 面色少华, 形体正常, 动静姿态, 语气清, 气息平, 无异常气味, 舌暗红、苔白, 脉沉细。

表 4: 模型生成与真实数据对比

4.3 消融实验

我们在训练数据上进行了多组实验，结果如表5所示，验证了我们提出数据增强方法的有效性。其中 $LoRA_Rank = 8$, $LoRA_alpha = 32$ ，使用Huatuogpt2-7B在两张3090上共使用150卡时训练。

模型	子任务1分数	子任务2分数
Baseline	0.45	0.41
w/o LoRA(Huatuogpt2-7B)	0.4075	0.3736
w/o DA(Huatuogpt2-7B)	0.415	0.435
w/o CTG(Huatuogpt2-7B)	0.527	0.4515
w/o Char-IND(Huatuogpt2-7B)	0.513	0.4515
Qwen3-4B	0.513	0.4152
Qwen2-7B	0.535	0.425
R=16(Huatuogpt2-7B)	0.537	0.444
alpha=16(Huatuogpt2-7B)	0.5425	0.45
Ours(Huatuogpt2-7B)	0.553	0.4515

表 5: 模型性能对比

表5展示了不同配置下模型的表现。其中w/o LoRA代表未使用微调模型，该对比旨在验证模型经微调前后的性能差异。w/o DA表示对于辨证辨病未使用任何一种数据增强方法，仅增强了处方数据。w/o Char-IND表示未使用词汇独立性替换，w/o CTG表示未使用受控文本生成。由于使用受控文本生成增加了训练数据中证型的种类，所以优于仅使用语序替换的方法。使用Qwen3-4B(Team, 2025)和Qwen2-7B(qwe, 2024)作为不同大模型的对比实验。与Qwen系列大模型相比，Huatuogpt2-7B在同等参数规模下表现更优，表明领域适配性优化较单纯扩大模型规模更具实践价值。同时还增加了两组参数对比实验，R=16表示 $LoRA_Rank = 16$, $alpha = 16$ 表示 $LoRA_alpha = 16$ 。

5 结论

本文针对中医多标签辨证辨病与中药处方推荐中高质量标注数据匮乏的问题，提出了一种融合大模型与可控文本生成的混合增强策略。通过引入基于词汇独立性的可控生成框架，有效扩展了有限的标注样本，并结合LoRA微调技术实现了对任务的高效适配。实验结果验证了该方法在无需引入外部数据的前提下，依然能够取得具有竞争力的性能，分别达到了0.553和0.4515的得分。该研究为中医药智能化诊疗系统的构建提供了新的思路，也为解决医学文本生成中的小样本问题提供了具有通用性的解决方案。但本工作仍存在局限，如无法量化判断生成数据与真实数据之间的差距，从而不断优化生成过程；对于生成的数据缺少人类评估；复合证型生成依赖随机组合，可能导致低频标签噪声较大。后续可通过引入对抗学习框架量化生成数据质量，并结合主动学习机制融入人工评估，提升生成样本的临床合理性；同时探索跨模态数据融合，将舌象、脉象等特征与文本生成结合，进一步增强诊疗决策的准确性。

6 致谢

本工作受到以下项目资助：国家自然科学基金（62166042），新疆自然科学基金（2021D01C076），新疆大学优秀研究生创新项目（XJDX2025YJS092）

参考文献

- Junying Chen, Xidong Wang, Ke Ji, Anningzhe Gao, Feng Jiang, Shunian Chen, Hongbo Zhang, Dingjie Song, Wenya Xie, Chuyi Kong, Jianquan Li, Xiang Wan, Haizhou Li, and Benyou Wang. 2023. Huatuogpt-ii: One-stage training for medical adaption of llms. *arXiv preprint arXiv:2311.09774*.
- Xin Dong, Yi Zheng, Zixin Shu, Kai Chang, Jianan Xia, Qiang Zhu, Kunyu Zhong, Xinyan Wang, Kuo Yang, and Xuezhong Zhou. 2022. Tcmpr: Tcm prescription recommendation based on subnetwork term mapping and deep learning. *BioMed Research International*, 2022(1):4845726.
- Hailong Hu, Yaqian Li, Zeyu Zheng, Wenjun Hu, Riyang Lin, and Yanlei Kang. 2025. A traditional chinese medicine prescription recommendation model based on contrastive pre-training and hierarchical structure network. *Expert Systems with Applications*, 268:126318, April.
- Jingjin Liu, Hankz Hankui Zhuo, Keping Jin, Jiamin Yuan, Zhimin Yang, and Zhengan Yao. 2023. Sequential condition evolved interaction knowledge graph for traditional chinese medicine recommendation.
2024. Qwen2 technical report.
- Qwen Team. 2025. Qwen3 technical report.
- Yehan Yan, Tianhao Ma, Ruotai Li, Xinhan Zheng, Guodong Shan, and Chisheng Li. 2025. Jingfang: A traditional chinese medicine large language model of expert-level medical diagnosis and syndrome differentiation-based treatment.
- Kuo Yang, Zecong Yu, Xin Su, Fengjin Zhang, Xiong He, Ning Wang, Qiguang Zheng, Feidie Yu, Tiancai Wen, and Xuezhong Zhou. 2024. PrescDRL: Deep reinforcement learning for herbal prescription planning in treatment of chronic diseases. *Chinese Medicine*, 19(1):144.
- Sara Zaheri, Jeff Leath, and David Stroud. 2020. Toxic comment classification. *SMU Data Science Review*, 3(1):13.
- Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28.

附录

字段	说明	样例
ID	患者入院的唯一id	35
性别	男或女	女
职业	患者的职业信息，如职员、退（离）休人员等	退休
年龄	患者的年龄	66岁
婚姻	描述婚姻状况，如已婚、未婚等	已婚
病史陈述者	入院时描述患者身体状况的人员与患者本人的关系	本人
发病节气	患者出现病情时的节气	立夏
主诉	患者就诊时的主要不适或症状	发作性胸闷20年，加重伴胸痛3月余
症状	入院时主要症状和体征概述	胸部疼痛，呈针刺样，胸闷不舒，心慌不安，气短乏力，眼干眼涩，口干口苦，纳可，食后反酸烧心，眠可，二便调。
中医望闻切诊	医师“望、闻、切”后对患者状态的描述	表情自然，面色暗红，形体正常，动静姿态，语气低，气息平；无异常气味，舌暗红、苔黄腻，舌下脉络曲张，脉弦。
病史	现病史、既往史、个人史、婚育史、家族史	患者20年前因劳累后出现胸闷……（此处省略，全文较长）
体格检查	患者的体格检查内容	体温：36.5℃ 脉搏：61次/分……（此处省略，全文较长）
辅助检查	其他检查（如CT、心电图报告等）	2020-4-29 冠脉CT示：LM轻度狭窄……
疾病	对应中医疾病（如心悸病、胸痹心痛病等）	胸痹心痛病
证型	中医证型（如气虚血瘀证、痰热蕴结证等）	气虚血瘀证—痰热蕴结证
处方（不含剂量）	中药处方（如黄芪、白芷等）	丁香，广藿香，黄芪，檀香，砂仁，木香，草豆蔻，附片，花椒，制川乌，细辛，桔梗，麸炒枳壳，葛根

表 6：患者临床文档