**文章编号**：1003-0077(2026)01-0060-10

# 基于预训练语言模型的汉语字形谜解谜模型

王涵玉，梁 循，关真宇，张森森，牛思敏，王孟威，宋世超，杨 丽

（中国人民大学 信息学院，北京 100872）

**摘 要**：该文提出了一种基于预训练语言模型的汉语字形谜自动解答系统。汉语字形谜是一类利用汉字结构和形态构造的谜语。为了实现字形谜的自动解答，该文整合了结构预测、关键词提取和谜底映射等多个模块。通过构建综合性的汉字数据集，并利用 BGE、BART 等 Transformer 架构的预训练语言模型进行训练和评估，验证了该方法在汉字谜语解答任务中的有效性。实验结果表明，该系统在解析汉字谜语和生成准确谜底方面达到了 96.90% 的准确率，有效提高了谜语解答的自动化水平。该文的研究成果有助于推动汉字处理技术的发展，并为智能化语言教学和文化传承应用提供了新的技术支持和思路。

**关键词**：字形谜；预训练语言模型；汉字谜语

**中图分类号**：TP391 **文献标识码**：A

# A Chinese Character Riddle Solving Model Based on Pre-trained Language Models

WANG Hanyu, LIANG Xun, GUAN Zhenyu, ZHANG Sensen,
NIU Simin, WANG Mengwei, SONG Shichao, YANG Li
(School of Information, Renmin University of China, Beijing 100872, China)

**Abstract**: This paper presents an automatic solution system for Chinese character riddles based on pre-trained language models. Chinese character riddles leverage the structure and morphology of Chinese characters. Accordingly, the proposed system integrates multiple modules, including structural prediction, keyword extraction and answer mapping. The pre-trained language models such as BGE and BART are employed and a comprehensive Chinese character dataset is constructed. Experimental results reveal an accuracy rate of 96.90% in parsing Chinese character riddles and generating accurate answers.

**Keywords**: Chinese character riddles; pre-trained language models; Chinese puzzles

## 0 前言

随着自然语言处理(NLP)技术的迅猛发展，深度学习模型在理解和生成人类语言方面取得了巨大的进步。特别是基于 Transformer[1] 架构的预训练语言模型的应用，使得文本分类[2-3]、文本生成和理解[4-5]等复杂文本任务得以实现，并在各类语言处理任务中展现出了强大的性能。在中文信息处理领域，汉字作为核心组成部分，以其独特的表意性、音韵结构以及深厚的文化底蕴，带来了新挑战和研究机遇。与拉丁字母等表音文字系统不同，汉字不仅代表音节，还承载着丰富的语义信息，并通过复杂的结构反映了悠久的文化历史和逻辑关系[6]。

汉字谜语作为一种历史悠久的语言游戏，融合了对汉字音、形、义的巧妙运用，它不仅是民间娱乐的方式，也是语言智慧和文化传承的重要载体。汉字谜语通过对汉字的创造性变换和语言艺术的表现，挑战解谜者的语言理解能力和逻辑思维能力，从而达到启迪思维、传播文化的目的。在这个过程中，谜语的谜面通常涉及字形、字音、字义等多个层面的综合考量，展现了汉字语言的独特美学和逻辑性。

**收稿日期**：2024-06-12 **定稿日期**：2024-09-23

**基金项目**：国家自然科学基金(62072463)；中国人民大学科学研究基金(24XNKJ31)；方正集团数字出版技术国家重点实验室开放课题

字形谜是一类特别的汉字谜语，其基于汉字的形态结构构造而成。解答字形谜需要识别汉字的组成部分及其关系，从而重构出谜底汉字。这类谜语强调汉字的形态和结构特征，因此研究自动解答字形谜的模型不仅需要具备语言理解能力，还需要深入掌握汉字的结构知识。

尽管自然语言处理（NLP）技术在处理复杂语言任务上取得了显著进展，现有模型主要集中在捕捉语义信息方面，对汉字的内部结构和字形特性的理解仍然不足。例如，先进的生成式大语言模型 GPT-4 在解答“上边是人，下边是玉”这一字谜时，错误地给出“玥”作为答案，而非正确答案“金”。这暴露了当前模型在解析汉字构成部件及其排列方式时的局限性。

一个重要挑战在于训练语料的缺乏，特别是关于汉字内部结构和字形谜的相关数据。现有语言模型的训练语料多以单个汉字为最小粒度，无法准确表示汉字的内部结构。现有数据只能通过句子序列间接描述汉字的结构，但这种方法既不精确也不全面。此外，现有语料中几乎没有专门描述汉字内部结构的句子数据，缺乏对字形谜任务的文本语料。这种数据的缺失使得模型难以有效应对字形谜的解析和解答任务。为提高自动解答汉字字形谜的能力，需要构建更细致的专门数据集，涵盖汉字的内部结构特征和丰富的字形谜语料。这些精细化的数据对于训练能够准确理解汉字结构和解答字形谜的模型至关重要。

研究汉字谜语自动解答的方法不仅有助于提升汉字处理技术，还能深化对汉字语言结构特性和文化内涵的理解，促进计算语言学、认知科学和中文信息处理等学科的交叉融合。基于 Transformer 的预训练模型（如 BERT[7]、GPT[8] 和 BART[9] 等）能够捕捉长距离的依赖关系，并在大规模文本数据上预训练，从而在特定任务中通过微调展现出优异性能。本研究基于这些技术进展，探索利用深度学习模型自动解答字形谜的可行性，旨在验证这些先进模型在理解和生成字形谜方面的有效性，进一步推动中文信息处理技术的发展。

本文首先分析了字形谜的结构特点，然后设计了一种基于深度学习的自动解谜框架。该框架包括结构预测、关键词提取和谜底映射三个主要部分，致力于自动识别和生成字形谜的答案。本文方法不仅考虑了谜语文本的语义信息，还融入了汉字的结构特性。通过结合现有的预训练模型和专门设计的网络结构，优化了解谜过程，实现了对字形谜的高效解析和对谜底的准确预测。

本文的主要贡献包括：

（1）提出了一个创新的字形谜自动解答框架，整合了结构预测、关键词提取和谜底映射等多个模块。

（2）利用微调的预训练语言模型，实现了模型对复杂字形谜的理解。

（3）实验验证了本文方法的有效性，证明本研究提出的方法在字形谜自动解答任务上的能力。本研究代码已在 https://github.com/MarrytheToilet/zimi 开源。

## 1　相关工作

### 1.1　预训练语言模型

预训练语言模型在自然语言处理中的广泛应用显著提升了多种语言任务的性能。这些模型通过在大规模无监督文本上进行预训练，能够捕捉语义和上下文信息，从而在下游任务中展现出优异的性能。

Transformer 架构采用自注意力机制，使得模型能够并行处理长序列并捕获长距离依赖关系。这一架构极大地提升了自然语言处理任务的效率和效果，成为当前主流语言模型的基础架构。基于 Transformer 架构衍生出两类预训练语言模型，一类是以 BERT 为代表的嵌入模型，另一类是以 GPT 和 BART 为代表的生成式语言模型。

BERT 通过双向编码器捕捉每个词在上下文中的双向依赖关系。BERT 通过掩码预测和句间预测学习文本的精确语义表示。BGE[10] 是另一种应用于文本嵌入的开源模型，具有精确的语义表征能力。GPT 则采用自回归的生成策略，主要用于文本生成任务。与 BERT 专注于文本理解不同，GPT 在文本生成任务中具备出色的生成能力，并广泛应用于对话生成、文本补全等任务中。BART 采用了双向编码和自回归解码结构，通过去噪的预训练任务自回归地生成文本，能够在自然语言生成、翻译等任务中表现优异。

在许多中文领域独特问题的研究中，已经应用了各类预训练语言模型，显著提升了模型的任务解决能力。徐洋等人[11] 提出的模型通过强化语境和语义信息来提升对话幽默识别能力，为复杂对话任务提供了新的视角。陈潇等人[12] 提出了多通道

BERT,用于跨语言属性级情感分类,通过不同语言的BERT模型相互交互,提升了模型的语义理解能力。胡星武等[13]开发了事件预训练方法,通过跨语言事件对比预训练和事件要素掩码预训练,显著提升了跨语言事件检索的性能。桑杰端珠和才让加[14]开发了基于词典注入的藏汉机器翻译模型,通过词典预训练增强模型在低资源语言上的翻译效果。然而,这些方法的设计仍然主要聚焦于语义分析任务,而非汉字内部结构的解析,因此难以应用于字形谜的解谜任务。

### 1.2 汉语字形谜与汉语字形

汉语字形谜的构造主要利用了汉字的复杂形态和多样的构字方式。汉字形态的复杂性在字谜中得到了充分利用。汉字的构形不仅包含独立的部件,还包括这些部件的组合方式,如上下结构、左右结构等。在字谜解答过程中,解谜者需要识别这些部件之间的关系,理解它们如何组合成完整的汉字。赵云肖等人[15]通过融合汉字的形、音、义知识,提出了一种文本语义匹配模型,展示了多元知识对语义处理的增强效果。同时,刘梦迪等人[16]通过建立偏旁部首的知识图谱,探索了汉字字形相似度的计算方法,这对于识别和查找形似字有重要意义。吴少扬[17]等人则从跨语言事件检索的角度出发,探讨了汉越跨语言文本事件的检索技术,提出预训练方法以增强模型的语言适应性。

汉语字谜的语义和认知特性使其成为语言研究和心理学研究的一个重要方向。在谜语认知的研究领域,王艺臻[18]利用眼动追踪和事件相关电位技术深入分析了汉语字谜的处理机制,揭示了从字面意义到非字面意义的转换过程。卜小倩[19]从体认语言学的视角,整合了识解理论和概念整合理论,探讨了谜语解答的认知机制,展示了谜语认知的复杂性。朱新秤等人[20]则关注线索在解决汉语成语谜语问题中的作用,进一步丰富了研究对谜语解答过程中线索效用的理解。成语谜语与字形谜的本质不同在于前者主要涉及成语语义的联想,而后者需要深刻理解汉字的物理形态和结构特性,因此解决思路有所不同。

现代语言技术为汉语字谜的自动解答提供了新的方法。通过利用预训练语言模型,研究者能够捕捉字谜中的语义特征和结构特征,实现对汉字谜语的自动解析和解答。李建华等人[21]构建了一个中文大规模谜语问答数据集,并利用预训练语言模型进行了实验,取得了初步成果,但其研究主要集中于谜语问答中的文本理解和生成,而非对汉字结构的深入解析。由于这些研究主要聚焦于一般语义和文本生成任务,对字形谜这种需要处理汉字结构的任务存在显著的不足。因此,现有的方法在汉字字形谜解谜的特定挑战下无法充分发挥效力,这突显了需要专门针对汉字内部结构和形态特征进行深入分析的必要性。只有通过开发能够处理细粒度汉字结构信息的模型,才能有效应对汉字字形谜的复杂性。

## 2 数据构建

本文构建了一个综合性数据集,涵盖汉字的基础特性,并支持自动生成汉语字形谜。数据集的构建过程包括对汉字基础信息的收集和整理,以及对谜语生成机制的设计与实现。以下是数据集构建的具体步骤。

### 2.1 数据来源与预处理

本文构建的数据集主要基于两个数据源:一是包含汉字基础信息的数据库,其中记录了汉字的笔画数、拼音、偏旁部首等通用属性;二是详细记录汉字拆分方法的数据文件,这些文件包含了每个汉字的拆解部件信息。

基础信息的收集旨在获取汉字 $\chi$ 的通用属性,如间架结构 $S(\chi)$、笔画数 $B(\chi)$、拼音 $P(\chi)$、偏旁部首 $R(\chi)$等。这些属性为后续谜语的生成提供了必要的语言学特征。本文构建字谜数据库的基础信息从 mapull/chinese-dictionary① 项目中获取,汉字字典包含了约 3 500 个常用汉字的基本信息,如字符、笔画数、拼音等,为汉字提供了必要的描述属性。汉字的拆分方法反映其结构上的组成部件 $C(\chi)$,对于每个汉字 $\chi$,其拆分方法可以表示为一组部件的集合 $C(\chi)=\{c_1,c_2,\cdots,c_n\}$,其中,$n$ 为部件的数量。汉字的拆分方式从 kfcd/chaizi② 项目进行获取。预处理后的字典数据的部分内容如表 1 所示。

**表 1 预处理汉字字典示例**

| 字符 | 笔画 | 拼音 | 偏旁 | 拆字 | 结构 |
|---|---|---|---|---|---|
| 寄 | 11 | jì | 宀 | [宝 奇] | 上下 |
| 贷 | 9 | dài | 贝 | [代 贝] | 上下 |

① https://github.com/mapull/chinese-dictionary

② https://github.com/kfcd/chaizi

续表

| 字符 | 笔画 | 拼音 | 偏旁 | 拆字 | 结构 |
|---|---|---|---|---|---|
| 墨 | 15 | mò | 土 | [黑 土] | 上下 |
| 惨 | 11 | cǎn | 忄 | [心 参，忄 参] | 左右 |
| 粉 | 10 | fěn | 米 | [米 分] | 左右 |

### 2.2　谜语模板设计与生成

基于整合后的汉字数据集，本文设计了一套字形谜语模板，用于通过汉字信息自动生成字形谜语。

设 $\chi$ 表示待生成谜语的汉字，$S(\chi)$表示 $\chi$ 的结构特性(如上下、左右、全包围结构等)，而 $C(\chi)=c_1,c_2,\cdots,c_n$ 代表通过预处理步骤得到的 $\chi$ 的拆分部件集合，其中每个 $c_i$ 是一个拆分出的部件。

在预处理步骤获取每个汉字的基础信息和拆分部件信息之后，本文首先针对不同的汉字结构设计了一系列谜语模板 $T$，如式(1)所示。

$$T=\{T_{S_1},T_{S_2},\cdots,T_{S_m}\} \tag{1}$$

其中，每个模板 $T_{S_i}$ 对应一种特定的汉字结构特性 $S_i$，如上下结构、左右结构等，并且包含占位符来表示部件位置。选取模板的过程可以用函数 $\mathrm{Sel}_{\mathrm{temp}}$ 表示，该函数依据汉字结构 $S(\chi)$从模板集合 $T$ 中选择合适的字形谜语模板，如式(2)所示。

$$T_\chi=\mathrm{Sel}_{\mathrm{temp}}(S(\chi),T) \tag{2}$$

通过将描述汉字结构的字谜模板 $T_\chi$ 应用于数据集中的每个汉字 $\chi$，结合其拆解出的部分组成的集合 $C(\chi)$和结构标识符 $S(\chi)$，自动生成相应的汉字谜语。填充模板的过程可以通过一个函数 $\mathrm{Fill}_{\mathrm{rid}}$ 来实现，该函数通过替换模板中的占位符将拆分部件集合 $C(\chi)$映射到字形谜的具体内容上，如式(3)所示。

$$\mathrm{Riddle}=\mathrm{Fill}_{\mathrm{rid}}(T_\chi,C(\chi)) \tag{3}$$

例如，对于表 1 中的汉字“寄”，其拆分部件为 $C$(寄)＝[宝，奇]，结构标识符 $S$(寄)为上下结构，通过应用模板 $T_\chi$＝“上{0}下{1}，合为一体”，可生成字形谜语 Riddle＝“上宝下奇，合为一体”。

本文的研究目标是设计一个解密系统，即通过谜语获取原始汉字。因此可以定义一个解码函数 $\mathrm{Dec}_{\mathrm{rid}}$，对于生成的谜语进行解谜，如式(4)所示。

$$\hat{\chi}=\mathrm{Dec}_{\mathrm{rid}}(\mathrm{Riddle}) \tag{4}$$

其中，$\hat{\chi}$ 是通过解谜得到的预测汉字。此函数需要根据谜语的内容，在所有可能的模板中匹配并恢复出原始汉字及其结构特性。

整个字谜生成和解谜过程可以形式化地表示为式(5)的过程。

$$\hat{\chi}=\mathrm{Dec}_{\mathrm{rid}}(\mathrm{Fill}_{\mathrm{rid}}(\mathrm{Sel}_{\mathrm{temp}}(S(\chi),T),C(\chi))) \tag{5}$$

当 $\hat{\chi}=\chi$ 时，模型对于该字形谜解谜成功。

表 2 展示出一些具体的谜语生成示例。通过模板生成的谜语，能够有效训练并评估语言模型对汉字结构特性和拆分部件的理解和推理能力。

**表 2　字形谜构建模板(部分)及示例**

| 数据集 | 结构 | 模板示例 | 生成字形谜示例 |
|---|---|---|---|
| 训练集 | 上下结构 | 上{0}下{1}，合为一体 | 上宝下奇，合为一体 |
| | 上中下结构 | 上段{0}轻盈如风，中段{1}稳如山，下段{2}厚重如地 | 上段自轻盈如风，中段田稳如山，下段廾厚重如地 |
| | 左右结构 | 左岸{0}繁华，右岸{1}静谧 | 左岸心繁华，右岸参静谧 |
| | 左中右结构 | 左手{0}开拓，中间{1}协调，右手{2}收尾 | 左手手开拓，中间育协调，右手攵收尾 |
| | 全包围结构 | 外{0}庇护，中{1}藏匿 | 外口庇护，中人藏匿 |
| 测试集 | 上下结构 | 树冠{0}枝繁叶茂，根系{1}深埋大地 | 树冠代枝繁叶茂，根系贝深埋大地 |
| | 上中下结构 | 上部{0}光明，中部{1}坚实，下部{2}厚重 | 上部人光明，中部一坚实，下部口厚重 |
| | 左右结构 | 一肩{0}担当重任，一肩{1}守护安宁 | 一肩米担当重任，一肩分守护安宁 |
| | 左中右结构 | 左翼{0}策马奔腾，中翼{1}坚如磐石，右翼{2}翱翔天际 | 左翼辛策马奔腾，中翼丝坚如磐石，右翼辛翱翔天际 |
| | 全包围结构 | 外圈{0}环绕，中芯{1}坚守 | 外圈口环绕，中芯卷坚守 |

### 2.3 数据集划分

本文聚焦于汉字结构中最常见的5种类型：上下结构、左右结构、上中下结构、左中右结构和全包围结构。数据集按7:3的比例划分为训练集和测试集，以平衡模型的学习效率与泛化能力。

训练集包含258条上下结构谜语、769条左右结构谜语、5条上中下结构谜语、8条左中右结构谜语和12条全包围结构谜语。测试集则包含107条上下结构谜语、334条左右结构谜语、两条上中下结构谜语、5条左中右结构谜语和4条全包围结构谜语，用于独立验证模型在未见过的样本上的表现，评估其泛化能力和解谜准确性。上述分布符合不同结构的汉字在中文汉字中的数量分布。

为了解决数据不均衡的问题，本文在模型训练过程中构建了伪数据。通过使用已有的模板生成更多少数类结构类型的样本，这些伪数据并不以解答实际字形谜为训练目的，而是专门用于让模型学习到各种汉字结构的信息和关键字提取方法。这种方式确保模型能够在各类结构类型上均衡地学习，从而提升其对不同字形谜的解析能力。

测试集使用的字形谜模板与训练集使用的字形谜构建模板不存在交集，以保证训练过程与测试过程相互独立。数据集及模板数量见表3。

**表3 数据集字形谜数量及模板数量设置**

| 数据集 | 结构 | 字形谜数量 | 模板数量 |
|---|---|---|---|
| 训练集 | 上下结构 | 258 | 13 |
| | 左右结构 | 769 | 11 |
| | 上中下结构 | 5 | 10 |
| | 左中右结构 | 8 | 10 |
| | 全包围结构 | 12 | 10 |
| 测试集 | 上下结构 | 107 | 10 |
| | 左右结构 | 334 | 9 |
| | 上中下结构 | 2 | 10 |
| | 左中右结构 | 5 | 10 |
| | 全包围结构 | 4 | 10 |

## 3 模型

为了有效地解答汉字字形谜，本文提出了一种基于深度学习的自动解谜框架，如图1所示。该框架旨在综合利用预训练模型的语言理解能力和汉字结构特性，自动解析字形谜的结构和谜底。

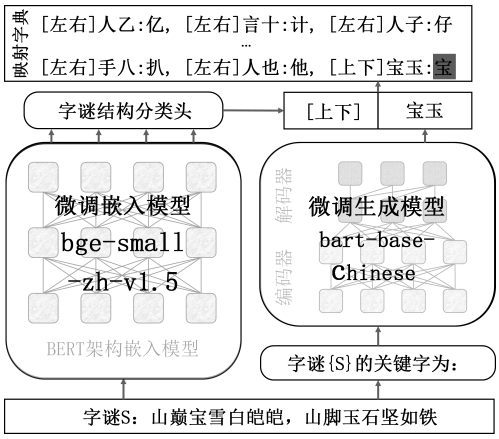



图1 字形谜解谜模型架构图

字形谜是一类通过拆解汉字的部件和结构来构建谜面的谜语，解谜者需要识别汉字的内部构造，如上下、左右等结构，从而推测谜底。这种类型的谜语不仅要求模型具备语言理解能力，还需要模型能够处理汉字的复杂形态和构造逻辑。

字形谜通过拆解汉字的部件和结构来构建谜面，解谜者需要识别汉字的内部构造，如上下、左右等结构，从而推测谜底。这种类型的谜语不仅要求模型具备语言理解能力，还需要模型能够处理汉字的复杂形态和构造逻辑。为此，本研究的框架首先通过字谜结构分析模型预测字形谜的结构（如“上下”），然后利用微调生成模型提取关键部件。基于字谜的结构和关键字信息，模型映射生成最终的谜底。例如，“山巅宝雪白皑皑，山脚玉石坚如铁”解析为“[上下]宝玉”，并映射为“宝”字。这种方法不仅考虑了字谜的语义信息，还有效地整合了汉字的结构特征，从而提升了模型对复杂谜语的理解和解析能力。

在模型设计过程中，本文区分了Encoder-Only和Encoder-Decoder架构的应用场景。Encoder-Only架构模型，如BERT和本文使用的BGE，擅长从输入文本中提取高质量的上下文语义表示。这种架构对文本进行双向编码，捕捉句子中各部分之间的相互关系，非常适用于分类任务。因此，本文选择Encoder-Only模型作为嵌入模型，用于结构分类任务。通过这种架构，本文能够将字形谜文本映射到一个高维向量空间中，并从中提取出反映汉字结构特征的有效特征表示。

另一方面，本文选择Encoder-Decoder架构的

模型，如 BART，作为关键词提取器。这种架构不仅包含一个用于理解输入文本的编码器，还包含一个用于生成输出文本的解码器。Encoder-Decoder 模型的设计使得其在处理生成任务时尤为有效，因为它能够在编码阶段充分理解输入内容，并在解码阶段基于上下文信息生成目标输出。Encoder-Decoder 架构模型在这种情况下表现优异，因为其解码器部分能够有效地生成与输入文本相关的关键字序列，确保模型准确地提取出重要的线索。

本文组合使用 Encoder-Only 模型进行结构分类，使用 Encoder-Decoder 模型进行关键词提取，充分利用了两种架构的各自优势，提高了整体系统对汉字字形谜的解析能力。

### 3.1　结构分析模型

在汉字字形谜中，谜语的语义往往与谜底汉字的结构特征直接相关，因此准确预测这些结构特征对解答字形谜至关重要。具体来说，汉字字形谜的语义线索通常暗示汉字的结构。例如，谜语中如果提到“高天”和“大地”，通常表明谜底汉字具有上下结构。类似“一面…一面…”的描述，则可能意味着谜底汉字的结构是左右关系。这样的语义与结构的关联为深度学习模型提供了丰富的学习信号，使其能够在大规模训练数据中捕捉这种隐含模式。

通过构建和训练基于句子嵌入的文本表示模型，可以将谜语文本有效地转化为向量空间中的点。这些向量表示捕捉到了谜语中的语义和结构特征。接着通过将这些句向量输入到微调后的分类器中，从而预测汉字字形谜的结构类型。

#### 3.1.1　结构分析模型框架

本文采用基于句子向量表示的深度学习方法。具体而言，通过将谜语文本转化为向量空间中的点，可以利用这些向量来捕捉谜语的语义和结构特点。对于每条谜语 $r$，其向量表示为 $\boldsymbol{v}_r$，通过微调的 BAAI/bge-small-zh-v1.5 模型获取，该模型是一个经过大量中文文本预训练的文本嵌入模型，能够生成高质量的句向量。通过对该模型进行微调，可以使其更好地适应汉字谜语的特定任务需求。为了将谜语的向量表示映射到其结构类型，本文设计了一个基于神经网络的分类器。分类器首先通过一个线性变换层提取特征，然后通过一个输出层进行结构类型的分类。

首先，将谜语的向量表示 $\boldsymbol{v}_r$ 输入到一个线性变换层中，该层通过一个权重矩阵 $\boldsymbol{W}_1$ 和偏置向量 $\boldsymbol{b}_1$ 对输入进行变换，并使用 ReLU 激活函数得到隐藏层表示 $\boldsymbol{h}_r$，如式(6)所示。

$$\boldsymbol{h}_r = \mathrm{ReLU}(\boldsymbol{W}_1 \boldsymbol{v}_r + \boldsymbol{b}_1) \tag{6}$$

隐藏层表征 $\boldsymbol{h}_r$ 通过另一个线性变换层，该层由权重矩阵 $\boldsymbol{W}_2$ 和偏置向量 $\boldsymbol{b}_2$ 构成，生成输出层的原始得分 $\boldsymbol{z}_r$，如式(7)所示。

$$\boldsymbol{z}_r = \boldsymbol{W}_2 \boldsymbol{h}_r + \boldsymbol{b}_2 \tag{7}$$

最后使用 Softmax 函数对输出层的原始得分 $\boldsymbol{z}_r$ 进行归一化处理，得到每种结构类型的概率分布 $p_r$ 如式(8)所示。

$$p_r = \mathrm{Softmax}(z_r) \tag{8}$$

其中，$h_r$ 是第一层的隐藏表示，$\boldsymbol{W}_1$ 和 $\boldsymbol{b}_1$ 是第一层的权重和偏置，$\boldsymbol{z}_r$ 是输出层的原始得分，$\boldsymbol{W}_2$ 和 $\boldsymbol{b}_2$ 是输出层的权重和偏置，$\boldsymbol{p}_r$ 是每种结构类型的概率分布。该结构分类器能够从谜语的向量表示中提取有效特征，并根据特征预测字形谜的结构类型。

#### 3.1.2　训练与评估

模型的训练目标是最小化预测概率分布 $p_r$ 与实际标签分布 $y_r$ 之间的差异，这通常通过交叉熵损失函数来实现。具体地，损失函数 $L$ 定义如式(9)所示。

$$L = -\sum_r \sum_c y_{r,c} \log p_{r,c} \tag{9}$$

其中，$y_{r,c}$ 是实际标签分布，如果谜语 $r$ 属于类别 $c$，则 $y_{r,c}=1$，否则 $y_{r,c}=0$；$p_{r,c}$ 是模型预测的谜语 $r$ 属于类别 $c$ 的概率。$y_{r,c}$ 是如果谜语 $r$ 属于类别 $c$ 则为 1，否则为 0 的指示函数，$p_{r,c}$ 是分类器预测谜语 $r$ 属于类别 $c$ 的概率。

在字谜结构分析模型的训练过程中，为解决数据集中的类别不平衡问题，本文采用了过采样的策略，特别针对数据较少的上中下结构、左中右结构和全包围结构进行了数据扩展。本文使用已有的模板生成伪训练数据，这些数据并不用于解答字形谜的实际问题，而是专门用于让模型学习不同汉字结构的特征。增加了这些少样本结构类型的样本数量，使得模型能够在训练中均衡地学习各种结构类型的特征。

字谜结构分析模型的训练与评估过程中的损失值变化如图 2 所示。在训练过程中，利用 Adam 优化器进行参数更新，以实现快速收敛。

模型在验证集上的表现通过式(10)计算准确率(Accuracy)来评估。

$$\mathrm{Accuracy} = \frac{N_{\mathrm{correct}}}{N_{\mathrm{total}}} \tag{10}$$

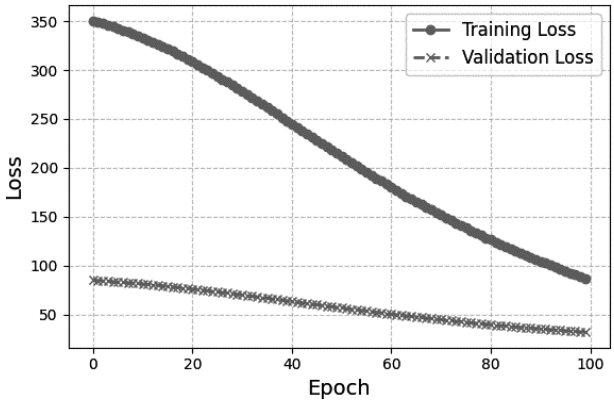



图2 结构分析模型损失值

字谜结构分析模型的验证准确率结果如图3所示。随着训练周期的增加,模型的训练损失和验证损失均呈现逐步下降的趋势,验证准确率则逐步提高,最终达到较高的准确率和较低的验证损失。模型在训练过程中逐渐学习到了有效的特征,能够准确预测字形谜的结构特性。

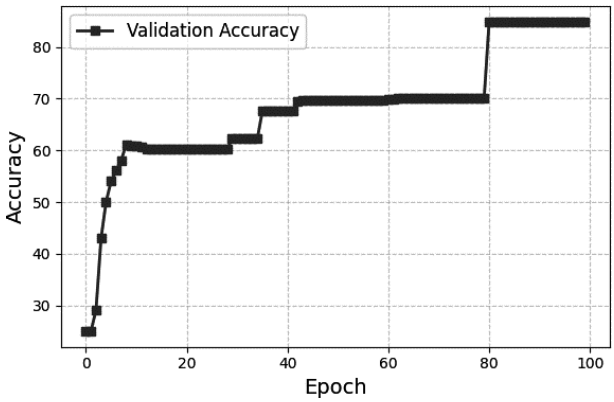



图3 结构分析模型验证准确率

### 3.2 关键字提取模型

关键字提取的主要任务是从谜语文本中识别出指向谜底的核心线索。在汉字字形谜中,谜面通常包含与谜底相关的关键信息,这些信息对于正确解答谜语至关重要。例如,在谜语"一口又一口——打一字:回"中,关键字是两个"口"字,它们直接构成了谜底"回"字的一部分。

为了准确提取这些关键信息,本文采用了一种微调后的预训练文本生成模型。该模型经过专门训练,能够适应汉字字形谜的特定需求。具体来说,模型首先接受谜语文本作为输入,通过编码层将其转化为高维向量表示。随后,解码层利用上下文信息生成与谜底相关的关键字。为提高模型性能,本文在少量汉字字形谜数据集上对预训练文本生成模型进行了微调,使其能够更精确地从谜语中提取出重要线索。

#### 3.2.1 关键字提取模型框架

为了实现字形谜的关键字的自动提取,本文构建了一个基于Transformer的预训练文本生成模型框架。该框架的输入为字形谜文本,经过模型处理后输出谜语中的关键字。并利用预训练的BART模型,结合谜语生成的任务特点对其进行了微调。

BART模型由编码器和解码器两部分组成,其中编码器负责对输入文本进行双向编码,捕捉上下文信息;解码器则以自回归的方式生成输出文本。BART模型通过预训练获得了强大的文本理解和生成能力,能够在微调后有效地完成关键字提取任务。本文使用了fnlp/bart-base-Chinese[①]模型,这是一种针对中文文本预训练的BART模型。该模型首先对汉字谜语文本进行分词和编码,将其转换为高维向量表示。编码器通过多层双向自注意力机制处理输入文本,生成包含丰富上下文信息的向量表示 $\boldsymbol{h}$。接下来,解码器基于编码器的输出,通过多层自注意力机制和编码器-解码器注意力机制,自回归地生成关键字序列 $y$。

模型的训练目标是最小化预测关键字序列与实际关键字序列之间的差异。这通常通过交叉熵损失函数来实现,具体如式(11)所示。

$$L=-\sum_{t=1}^{T}\log P(y_t \mid y_{<t},\boldsymbol{h};\theta) \tag{11}$$

其中,$y_t$是第$t$个关键字,$y_{<t}$表示在$t$时刻之前生成的所有关键字,$\boldsymbol{h}$是编码器生成的上下文向量表示,$\theta$是模型参数。通过对fnlp/bart-base-Chinese模型进行微调,能够使其更好地适应汉字谜语的具体任务需求,从而准确提取出谜语中的关键字。

#### 3.2.2 训练与评估

在关键字提取模型的训练过程中,每条谜语被表示为一个序列 $s=[w_1,w_2,\cdots,w_n]$,其中,$w_i$代表序列中的第$i$个词。对于模型的输入,本文构造了形式为"**请抽取字谜中的关键字:$s$。字谜中的关键字是:**"的提示语句$P$,以引导模型聚焦于谜语的关键信息抽取。模型的目标输出是一个或多个谜语中的关键字 $k=[k_1,k_2,\cdots,k_m]$。

在微调阶段,模型的目标是最小化预测关键字

① huggingface.co/fnlp/bart-base-chinese

序列与实际关键字序列之间的差异，可以通过式(12)定义的交叉熵损失函数来实现。

$$L = -\sum_{i=1}^{m} \log P(k_i \mid P, \theta) \tag{12}$$

其中，$P(k_i \mid P, \theta)$表示在给定输入序列 $s$ 和模型参数 $\theta$ 的条件下，正确预测第 $i$ 个关键字的概率。关键字提取模型的训练与评估过程中的损失值变化如图 4 所示。

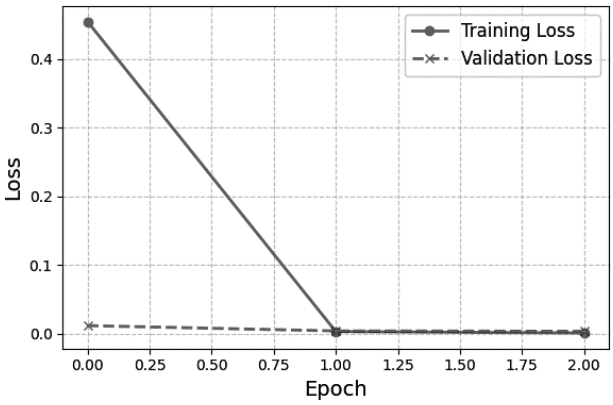



图 4　关键字提取模型损失值

与结构分类器训练过程一致，在关键字提取器的微调过程中，为应对数据不平衡问题，同样采取了过采样策略。通过使用已有的模板生成伪训练数据。

将构建的整个数据集分为训练集和验证集，分别用于模型的自监督训练和性能评估。在训练过程中，使用 Adam 优化器对模型参数进行更新，以最小化上述定义的损失函数。

字形谜关键字提取模型的验证准确率结果如图 5 所示，该准确率是通过将模型在验证集上的预测结果与真实关键字进行比较计算得出。将每个预测的关键字序列与相应的真实序列进行匹配，计算匹配正确的比例，以此评估模型在谜语关键字提取任务中的处理能力。关键字提取模型的训练过程中，随着训练周期的增加，模型的训练损失和验证损失均显著下降，而验证准确率逐步提高，模型在逐渐学习到有效特征以准确提取汉字谜语中的关键字。验证准确率的逐步提升和验证损失的逐步降低，验证了模型的有效性和鲁棒性。该模型能够高效、准确地从谜语文本中提取出关键信息。

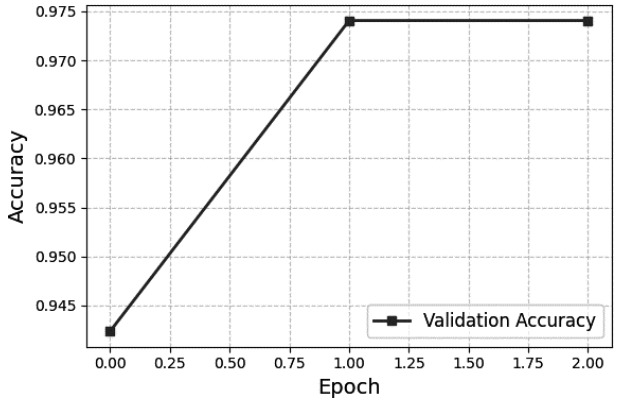



图 5　关键字提取模型验证准确率

## 4　实验与结果分析

本文对包含 455 个汉字谜语的测试集进行了结构分析和关键字提取模型的性能评估，结果如表 4 所示。同时部分模型预测成功或失败的例子如表 5 所示。

**表 4　模型测试准确率**　　(单位：%)

| 测试集 | 结构准确率 | 关键字准确率 | 综合准确率 | 字符准确率 |
|---|---|---|---|---|
| 准确率 | 98.92 | 97.30 | 96.90 | 96.90 |

**表 5　字形谜解谜模型预测结果(部分)**

| 字符 | 字谜 | 真实结构 | 真实关键字 | 预测结构 | 预测关键字 | 预测谜底 | 预测字符 |
|---|---|---|---|---|---|---|---|
| 皱 | 左侧刍浓烈，右侧皮淡雅 | [左右] | 刍皮 | [左右] | 刍皮 | [左右]刍皮 | 皱 |
| 宝 | 山巅宝雪白皑皑，山脚玉石坚如铁 | [上下] | 宝玉 | [上下] | 宝玉 | [上下]宝玉 | 宝 |
| 圃 | 外壳口坚硬，内里甫温暖 | [全包围] | 口甫 | [全包围] | 口甫 | [全包围]口甫 | 圃 |
| 浏 | 左部水撑起，中部文连接，右部刀收束 | [左中右] | 水文刀 | [左中右] | 水文刀 | [左中右]水文刀 | 浏 |
| 叫 | 一肩口担当重任，一肩丩守护安宁 | [左右] | 口丩 | [左右] | 口美 | [左右]口美 | 失败 |
| 耍 | 树冠而枝繁叶茂，根系女深埋大地 | [上下] | 而女 | [上下] | 女女 | [上下]女女 | 失败 |
| 合 | 上方人如日，中间一如月，下方口如星 | [上中下] | 人一口 | [上下] | 人口 | [上下]人口 | 失败 |

### 4.1 结构预测准确率分析

结构谜底的预测准确率达到了98.92%，表明模型在汉字谜语结构解析任务中表现出色。例如，在面对谜语“山巅宝雪白皑皑，山脚玉石坚如铁”时，模型通过理解“山巅与山脚”的空间关系，成功提取出“上下”结构的构型特征，准确预测了该谜语的结构谜底。类似地，对于谜语“左侧刍浓烈，右侧皮淡雅”，模型通过理解“左侧”和“右侧”这一结构对称关系，识别出“左右”结构。模型正确地将“刍”和“皮”分别对应到汉字“皱”的左部和右部，成功预测了结构谜底“左右”，并生成了正确的谜底“刍皮”。这表明模型能够有效利用语言指示信息，捕捉汉字谜语中的结构模式，说明其在处理汉字谜语结构特征上具有较高的精度和较好的泛化能力。

在对谜语“上方人如日，中间一如月，下方口如星”进行预测时，模型未能正确识别出“上中下”结构，而是错误地预测为“上下”结构，导致最终谜底“合”字的预测失败。分析其原因，可能在于上中下结构和上下结构在句向量的特征表示上过于相似。例如，二者可能都包含“太阳”和“星星”的元素，而上中下结构只比上下结构多包含“月亮”的元素。这种特征的相似性使得模型在结构分析和关键词提取过程中容易混淆，无法准确区分这些微妙的差异。

要完善这一不足，可能需要依赖更先进和更多参数的文本嵌入模型，这些模型能够更好地捕捉和区分句子中不同元素的特征。此外，丰富训练数据中的多样性和精细度，特别是针对少见的结构类型进行专门的数据增强，也将有助于提高模型在不同结构类型间的区分能力。通过这些改进，模型可以更加准确地处理和理解汉字字形谜中的复杂结构特征。

### 4.2 关键字预测准确率分析

关键字谜底的预测准确率为97.30%，说明该模型在处理复杂语义和部件识别上比较有效。然而，由于该任务需要深入解析谜语的语义并精确地识别部件，部分谜语的关键字提取失败。问题可能来自两方面：首先，关键字谜底预测固有的复杂性要求模型不仅具备强大的文本理解能力，还需深入剖析汉句子的结构来确定关键词所在的位置。例如，在谜语“树冠而枝繁叶茂，根系女深埋大地”中，模型虽然能够识别上下结构，但在解析细节时可能由于“而”字虚词的特性将其忽略，将“而女”错误预测为“女女”。

其次，生僻或罕见部件的识别失败率较高，例如，在谜语“一肩口担当重任，一肩刂守护安宁”中，模型正确识别“左右”结构，但由于右半部分“刂”的生僻性，模型未能有效学习其特性，错误预测为“口美”。这些错误主要源于预训练阶段语料库和词汇表对这些特殊部件的覆盖不足，影响了模型对其的有效学习和提取。

### 4.3 综合准确率与字符准确率

综合谜底的预测准确率为96.90%，此指标代表模型同时成功预测结构和关键字的概率字符的预测准确率，字符准确率与综合谜底准确率一致。说明模型在对结构和关键字进行预测之后，能够正确通过映射字典映射到正确的谜底。

### 4.4 小结

本文采用了一个专为汉字字形谜设计的模型，对测试集的数据进行了深入分析，旨在评估其在结构解析和关键字提取方面的性能。研究结果显示，该模型在结构预测上取得了98.92%的准确率，展示了其在理解和解析汉字空间布局特征方面的卓越能力。例如，通过解析“山巅与山脚”的空间关系，模型能够准确捕捉“上下”结构，或通过“左侧与右侧”的对称性，识别“左右”结构，从而成功预测谜底的结构要素。关键字预测的准确率为97.30%，体现了模型在处理语言深层语义和识别部件方面的强大能力。综合谜底预测准确率为96.90%，字符预测准确率与之保持一致，表明模型在整合结构分析和关键字匹配方面表现出高效性和稳定性，能够正确引导至最终谜底。

综上所述，本文构建的汉字字形谜解析模型在评估中展现了优异的性能，特别是在结构和关键字预测的准确率上达到了较高水平。这不仅验证了模型对汉字结构特征的深刻理解和在复杂语境中对关键字的高效识别能力，还体现了其在整合多维度信息处理上的稳定性和准确性。

## 5 结论与展望

本文提出了一种基于预训练语言模型的汉字字形谜自动解答模型，并验证了其在实际应用中的有效性。通过构建综合性的汉字字形数据集，设计了一个结合结构预测、关键词提取和谜底映射的解谜框架，有效解决了汉字字形谜中的结构解析和谜底生成问题。研究通过微调BGE和BART等预训练语言模型，使其适应汉字字形谜的解答任务，显著提升了模型对复杂谜语的理解和解析能力。实验结果显示，这一解谜框架在自动解答汉字字形谜方面表

现出色，结构预测和关键词提取的准确率分别达到了 98.92%和 97.30%，综合准确率为 96.90%。这些结果表明，利用预训练语言模型和深度学习技术，可以有效捕捉汉字谜语中的语义线索和结构特征，从而实现对复杂谜语的高效解析和准确解答。

在汉字字谜领域未来的研究中，还可以在几个方面进一步深化。首先，扩展数据集的规模和多样性，纳入更多类型的汉字字形谜和汉字变体，以增强模型的泛化能力和适应性。此外，需要提升模型对生僻字和罕见部件的理解能力，以减少在处理这些特殊字符时的误差。未来的工作还应结合更多的文化背景和语境信息，以增强模型对谜语中隐含线索和背景知识的识别能力。同时，探索跨语言谜语解答的应用潜力，有助于推动多语言处理技术的发展。这些拓展和优化不仅将进一步提升汉字字形谜自动解答系统的性能，还将为智能语言教学、文化传承及娱乐互动等应用领域提供新的技术支持和研究思路。

## 参考文献


[1] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017: 5998-6008.

[2] 陆晓蕾，倪斌. 基于预训练语言模型的 BERT-CNN 多层级专利分类研究[J]. 中文信息学报，2021，35(11)：70-79.

[3] 安波，龙从军. 基于预训练语言模型的藏文文本分类[J]. 中文信息学报，2022，36(12)：85-93.

[4] 马志强，王春喻，贾文超，等. 情感导向对话回复生成模型[J]. 中文信息学报，2023，37(8)：104-114.

[5] 杨锦锋，梁先桂，王刘安，等. 基于 Prompt 策略的医疗对话生成[J]. 中文信息学报，2023，37(4)：118-125.

[6] 金嘉祺. 汉字与中国传统文化[J]. 赤子：上中旬，2014，(15)：63-64.

[7] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of NAACL-HLT. 2019: 4171-4186.

[8] BROWN T B, MANN B, RYDER N, et al. Language models are few-shot learners[C]//Proceedings of the 34th International Conference on Neural Information Processing Systems, 2020: 1877-1901.

[9] LEWIS M, LIU Y, GOYAL N, et al. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension [J/OL]. arXiv preprint arXiv/1910.13461, 2019.

[10] XIAO S, LIU Z, ZHANG P, et al. C-Pack: Packaged resources to advance general Chinese embedding [J/OL]. arXiv preprint arXiv: 2309.07597, 2024.

[11] 徐洋，蒋玉茹，张禹尧. 基于 BERT 的强化语境与语义信息的对话幽默识别模型[J]. 中文信息学报，2022，36(4)：73-80.

[12] 陈潇，王晶晶，李寿山，等. 基于多通道 BERT 的跨语言属性级情感分类方法[J]. 中文信息学报，2022，36(2)：121-128.

[13] 胡星武，桂韬，张奇，等. 基于对比学习的表达式偏序关系建模[J]. 中文信息学报，2023，37(4)：166-174.

[14] 桑杰端珠，才让加. 基于词典注入的藏汉机器翻译模型预训练方法[J]. 中文信息学报，2023，37(8)：43-51.

[15] 赵云肖，李茹，李欣杰，等. 基于汉字形音义多元知识和标签嵌入的文本语义匹配模型[J]. 中文信息学报，2024，38(3)：42-55.

[16] 刘梦迪，梁循. 基于偏旁部首知识表示学习的汉字字形相似度计算方法[J]. 中文信息学报，2021，35(12)：47-59.

[17] 吴少扬，余正涛，黄于欣，等. 面向汉越跨语言事件检索的事件预训练方法[J]. 中文信息学报，2024，38(4)：78-85.

[18] 王艺臻. 汉语字谜理解的认知神经机制研究[D]. 杭州：浙江大学博士学位论文，2021.

[19] 卜小倩. 体认语言学视域下汉语谜语的认知机制研究[D]. 成都：四川外国语大学硕士学位论文，2021.

[20] 朱新秤，李瑞菊，周治金. 谜语问题解决中线索的作用[J]. 心理学报，2009，41(5)：397-405.

[21] 李建华，陈涛，贾旭东，等. 中文大规模谜语问答数据集构建与评测方法研究[J]. 五邑大学学报(自然科学版)，2023，37(4)：38-46.





王涵玉(2001—)，博士，主要研究领域为自然语言处理。
E-mail：hy.wang@ruc.edu.cn





关真宇(1993—)，博士，主要研究领域为自然语言处理。
E-mail：zhenyu_guan@ruc.edu.cn





梁循(1965—)，通信作者，博士，教授，博士生导师，主要研究领域为自然语言处理、商务智能、社会计算。
E-mail：xliang@ruc.edu.cn