面向对话场景的构式数据集

薛旭晶[1,‡]，李俊材[1,‡]，苏雪峰[1,2,‡]，杨沛渊[1,‡]，柴清华[3,⋆]，李茹[1,4,⋆,†]
[1]山西大学 计算机与信息技术学院，山西 太原 030006
[2]山西工程科技职业大学 现代物流学院，山西 晋中 030609
[3]山西大学 外国语学院，山西 太原 030006
[4]山西大学 计算智能与中文信息处理教育部重点实验室，山西 太原 030006
[‡]{497440161，1251972979，455375251，971859815}@qq.com
[⋆]{charles，liru}@sxu.edu.cn

摘要

大语言模型在多种自然语言处理任务中展现出强大的语义理解能力。现有研究通常基于各类语义解析数据集对大语言模型进行评估，然而，这些数据集难以覆盖对话语料中常见的口语化表达与特定结构表达语义的语言现象，无法有效评估大语言模型在对话场景中的细粒度语义理解能力。为此，本文面向对话语料构建了一个包含2146条语句、1748个构式的中文构式数据集，实现语义信息细粒度表达的同时有效覆盖了现有语义解析评估数据集的缺口。基于该数据集，本文选取了其中部分代表性构式，结合框架语义学理论，提出了构式识别与构式语义理解两项评测任务，以系统评估大语言模型在对话场景中识别构式与理解深层语义的能力。实验结果表明，当前大语言模型在构式识别方面仍存在明显不足；且在缺乏思维链推理的引导下，难以理解构式所承载的深层语义。

关键词： 框架语义学；对话场景；构式语法；大语言模型

Construction Dataset for Dialogue Scenarios

Xujing Xue[1,‡]，Juncai Li[1,‡]，Xuefeng Su[1,2,‡]，Peiyuan Yang[1,‡]，Qinghua Chai[3,⋆]，Ru Li[1,4,⋆,†]
[1]School of Computer and Information Technology，Shanxi University
[2]School of Modern Logistics, Shanxi Vocational University of Engineering Science and Technology
[3]School of Foreign Languages，Shanxi University
[4]Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education
[‡] {497440161，1251972979，455375251，971859815}@qq.com
[⋆] {charles，liru}@sxu.edu.cn

Abstract

Large language models (LLMs) show strong semantic understanding in many NLP tasks. However, existing evaluation datasets often overlook colloquial and constructional expressions common in dialogues, limiting their ability to assess fine-grained semantic understanding in conversational settings. To address this, we build a Chinese construction dataset for dialogue, containing 2,146 utterances and 1,748 constructions. It captures fine-grained semantics and fills gaps left by current semantic parsing benchmarks. Using this dataset, we propose two evaluation tasks：construction identification and constructional semantic understanding, based on frame semantics theory. Results indicate that LLMs struggle with recognizing constructions and understanding their deep semantics without chain-of-thought reasoning.

Keywords: Frame Semantics , Dialogue Scenarios , Construction Grammar , Large Language Model

第二十四届中国计算语言学大会论文集，第513页-第524页，济南，中国，2025年8月11日至14日。

1 引言

大语言模型在预训练阶段吸收了大量数据，在多种自然语言处理场景中广泛应用，展现出强大的语言理解能力。为了全面评估大模型的语言理解能力，现有的研究通过各类语义解析任务来评估大语言模型的语言理解能力，如开放信息抽取(Seitl et al., 2024)、框架语义解析(Li et al., 2023b)等。

然而，现有的语义解析数据在对话场景中存在语义表达粒度粗、语言现象覆盖率低等问题，难以深入评估模型的深层语义理解能力。一方面，泛用性较高的语义解析方法（如LTP(Che et al., 2021)）虽然在对话情景中有较高的覆盖率，但受限于其预定义的有限角色集合，无法支持对大语言模型细粒度语义理解能力的评估。另一方面，细粒度的语义解析方法（如框架语义解析(石佼et al., 2014)）侧重于分析书面语料，未能覆盖在对话场景中普遍存在的口语化表达和依赖特定结构传达语义的语言现象。

汉语框架网（Chinese FrameNet，CFN）(You and Liu, 2005)是以Fillmore (1976)的框架语义学理论为基础构建的汉语框架语义知识库。汉语框架语义解析(石佼et al., 2014)是基于汉语框架网提出的细粒度语义分析任务，其目标是从句中根据目标词提取出相应的框架及其语义角色，能够表达细粒度的语义信息，实现对句子中事件或情景的深层理解。如图1所示，框架语义解析以“想”为目标词，激活【渴望】框架，表达体验者渴望做某事的语义信息。该方法进一步将“我”标注为“体验者”，将“做球星”标注为“事件”，从语义角色的角度揭示句子成分间的深层语义关系。然而，其难以表达特定结构所对应的语义信息，导致在处理对话语料时的覆盖率不足，无法有效评估大语言模型在对话场景中的语义理解能力。

构式是形式和意义的对应体(Goldberg, 1995)，其提供了一种形式与意义直接关联的语义分析思路，能够较好捕捉语言形式所承载的语义功能。构式形式是由常项和变项组成的线性序列结构(詹卫东, 2017)，常项为固定成分，而变项则指可以替换或填充的成分，如动词、名词性短语等。如图1所示，在框架语义解析的基础上引入构式，可以触发“要是+能+v”构式，其中“要是”和“能”为常项，“v（动词）”为变项，从而更准确地捕捉句子蕴含的假设情境与更细粒度的情感色彩，实现更细粒度、更丰富的语义理解。山西大学在2024年推出了以构式为目标词的框架语义解析评测(Yang et al., 2024)，然而现有的特定的构式难以满足评估大语言模型在对话场景中的语义理解能力的需求。

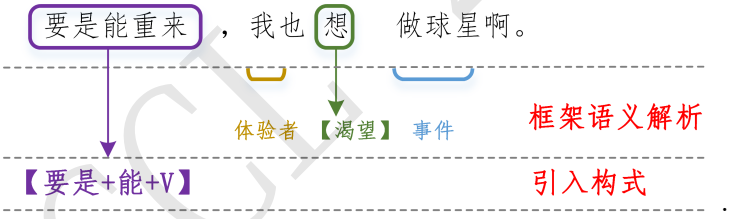

图 1: 框架语义解析引入构式示例

因此，为了全面评估大语言模型在对话场景中的语义理解能力，本文面向NaturalConv中文对话语料库 (Wang et al., 2021)，提出了基于思维链的构式提取方法，并结合人工筛选，最终构建了一个包含2146条对话语句、共计1748个有效构式的中文构式数据集。在框架语义解析及其他几种语义解析方法的基础上，融合该构式数据集对NaturalConv测试语料进行解析。通过引入词元覆盖率和压缩率两个指标，表明引入构式能够在保留语义信息的同时显著提升词元覆盖率并降低压缩率，验证了该数据集的有效性。

基于该数据集，本文选取了其中部分代表性构式，结合框架语义学理论，提出了构式（目标词）识别与构式语义理解两项评测任务，分别评估大语言模型在对话场景中发现构式语言结构和理解构式语义的能力。实验结果表明，在两个任务中，模型在不同构式上的表现存在显著差异，在理解难度较高的构式（如“a+都+来不及”构式和“有+什么+ap+的”构式）中，

* 基金项目：国家自然科学基金面上项目（62376144）；山西省科技合作交流专项项目(202204041101016)；山西省基础研究计划资助项目（202403021211092）；山西省重点研发计划项目课题(202102020101008)
† 通讯作者 Corresponding Author

未经深度推理的模型整体表现不佳。对于构式识别任务，即使经过深度推理的模型在“有+什么+ap+的”构式上仍然表现较差。因此，本文认为，尽管大模型在大规模语料的训练下已具备一定的构式语义理解能力，但在构式识别以及更细粒度的构式语义理解中仍存在一定挑战。

2 相关工作

语义解析方法 语义解析包括命名实体识别、关系抽取、事件抽取、语义角色标注等多个任务。Che et al. (2021)等人提出了语义解析工具集LTP，包含语义角色标注任务，然而其受限于语义角色数量导致语义粒度较为粗糙。框架语义解析任务旨在将自然语言文本与框架知识进行关联，是一种细粒度语义分析任务。该任务包含三个子任务：目标词识别、目标词所激起的框架识别，以及框架下语义角色的标注。近年来，基于CFN资源，涌现出一系列针对汉语框架语义解析的研究工作。张立凡 (2016)将目标词识别任务分成基于规则过滤与基于分类模型识别两个阶段，显著提升了目标词识别任务的准确率。Li et al. (2023c)等人提出基于旋转位置编码的分类模型，为框架识别任务贡献了新思路。王晓晖et al. (2022)等人提出了基于自注意力机制的句法感知汉语框架语义角色标注方法，有效提升了语义角色标注的F_1值。目前，山西大学已开源出汉语框架语义自动解析工具集[1]，推进了框架语义解析技术在实际中的应用。然而，现有的语义解析方法难以覆盖对话语料中通过特定结构表达语义的语言现象。

构式语法 Fillmore (1988)首次提出“语法构式”概念，并将构式分为实体构式和图示构式两类。其中，实体构式主要指词汇、复合词以及全固定的习语，而图示构式则指含有非固定成分的构式。本文所构建的构式数据集主要关注后者，即图示构式。Goldberg (1995)认为构式是一个形式与意义的对应体，构式应该整体表达意义，不能根据构式成分推知其所表达的意义，后汉语语法学界多采用这一概念定义构式。Dunn (2019)提出了基于频率和关联的构式提取模型，该模型能从大量文本数据中自动总结构式，但并不能保证所提取构式的正确性。詹卫东 (2017)探讨了国内构式知识库建设的相关问题，展望了将构式语法应用到计算机自动句法分析中的前景，提出构建现代汉语构式知识库并阐述其初步工作，目前北京大学构式库已收录2000余条构式。然而，现有研究仍缺少快速构建构式库的方法。

评估大语言模型语义理解能力 现有研究主要通过语义解析数据集对大语言模型的语义理解能力进行评估。Li et al. (2023a)等人评估了ChatGPT在7项信息抽取任务上的表现，结果表明，模型在开放信息抽取任务上表现出色，但在标准信息抽取场景下性能很差。Cai et al. (2024)等人进一步比较了ChatGPT与多种国内开源大语言模型在命名实体识别、关系抽取和事件抽取三项子任务上的性能，结果显示所有国内开源大模型在事件抽取任务中的表现均不及ChatGPT。Das et al. (2025)等人则聚焦于医疗领域的信息抽取任务，发现开源大模型存在显著的幻觉倾向，影响其在关键任务中的可靠性。山西大学2024年推出框架语义解析评测(Yang et al., 2024)，评测结果表明，大语言模型在框架识别任务中表现较差并且无法有效识别目标词所触发场景中的语义角色。然而，这些语义解析数据集难以覆盖对话语料中常见的口语化表达与特定结构表达语义的语言现象，无法有效评估大语言模型在对话场景中的细粒度语义理解能力。

构式知识在口语化对话场景中的应用研究尚处于起步阶段。对话场景中，常见口语表达、依赖特定结构传达语义等语言现象对传统语义解析方法构成挑战，进而限制了对大语言模型在对话场景中语义能力的评估，而这些语言现象恰可由构式语法理论更自然地解释。基于此，本文以NaturalConv中文对话语料库(Wang et al., 2021)为基础，提出了思维链驱动的构式提取方法，并结合人工筛选构建了面向对话场景的构式数据集，实现语义信息细粒度表达的同时有效覆盖了现有语义解析方法的缺口。基于该数据集中部分代表性构式，结合框架语义学理论，进一步提出了构式识别和构式语义理解两项评测任务，分别评估大语言模型在对话场景中发现构式语言结构和理解构式语义的能力。一方面，为传统语义解析方法在处理对话中非规范语言现象时提供了新的补充机制；另一方面，为构式语法理论在自然语言处理领域中发挥作用提供了新思路。

3 构式数据集构建

NaturalConv是 Wang et al. (2021)等人提出的一个面向多轮话题驱动对话的中文语料库，共19.9K个会话和400K条语句。该语料库允许对话参与者围绕多个主题自由切换，呈现出深入

[1]https://github.com/SXUNLP/CFSP

探讨与自然过渡并存的特点，贴近真实人类交流的方式。因此，本文选取NaturalConv测试集中的4844条语句作为构式提取的基础数据集。

构式提取旨在从大规模语料库中自动识别并归纳构式。然而，现有基于统计的方法在构式提取时存在一定局限，既难以确保所归纳构式的正确性，也常常忽略低频但具有语言学价值的构式。尽管大语言模型在深层次语义理解上仍有缺陷(Yang et al., 2024)，但其在预训练阶段吸收了大规模语料的知识，在一定程度上可以辅助构式数据集的构建。因此，本文并未采用基于统计的方法进行构式提取，而是通过思维链提示大语言模型并结合人工筛选的方式来构建构式数据集。

3.1 基于思维链提示的构式提取方法

图2展示了构式提取任务所采用的基于思维链的提示模板，旨在引导大语言模型逐步识别句子中的固定表达模式并将其归纳为通用的语法构式。该模板主要包括三部分内容：任务指令、输出要求以及示例。任务指令明确模型需识别固定表达模式并抽象其形式；输出格式要求对构式中的变项使用统一的标签集，该标签集参考北京大学构式库[2]中对构式变项的统计信息；示例部分通过逐步推理过程，展示模型应如何分析句子并生成规范的构式表示。

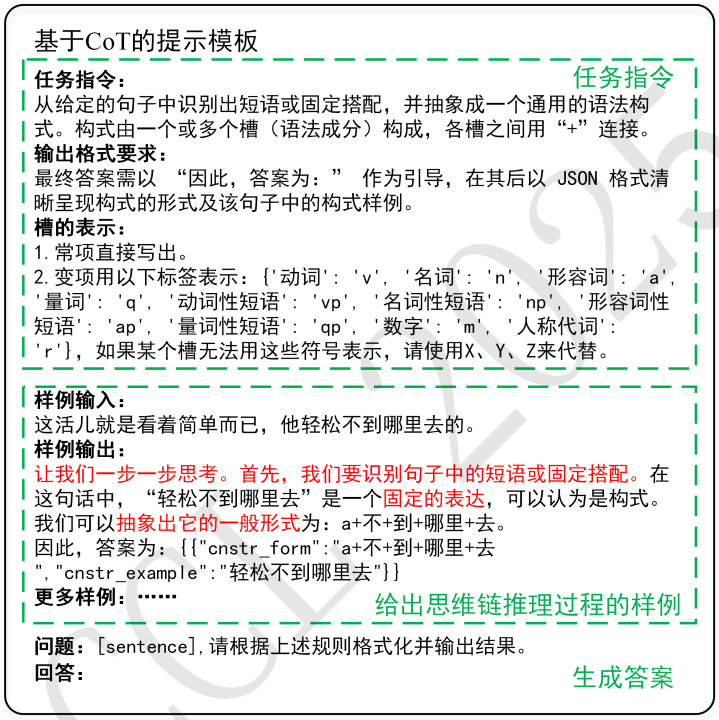

图 2: 构式提取任务提示模板

具体来说，我们选取了deepseek-v3 (Liu et al., 2024)作为构式提取的基础模型，对4844条对话语句进行处理，从3588条语句中提取出3398个候选构式。这些候选构式是大语言模型所生成的潜在构式，虽具有确定的形式，但并不能保证其符合语言学定义下构式的标准。因此，模型输出结果仍需人工进一步筛选。模型提取出的候选构式示例如表1所示。

句子	模型输出	构式
...就拿这次的比赛来说，他们队里面的两员大将...	就拿+np+来说	是
相信在之后的比赛中一定会越挫越勇的！	越+v+越+a	是
现在基本上都是朝九晚五的生活。	朝+m+晚+m	否

表 1: 模型提取出的候选构式示例

[2]http://ccl.pku.edu.cn/ccgd/

3.2 人工筛选

为提升筛选质量，在正式筛选开始前，我们选取了200条语句，共涉及232个候选构式，进行了预标注。标注人员包括一位语言学专家与三名团队中研究方向为自然语言处理的研究生。标注完成后就存在分歧的结果进行了充分讨论，在此基础上，总结了大语言模型结果的错误类型，为后续正式的筛选工作奠定了基础。

大语言模型结果错误类型主要包括：(1)构式形式未按规定输出。例如，候选构式："已经+有+数字+年"中，"数字"并未按规定标注为"m"；(2)过度抽象导致结果错误。例如，将构式"吃不吃得消"过度抽象为"v+不+v+得+消"；(3)将尚未固化的语言误认为是构式。例如，模型将"只怪我当初不够努力"误认为是构式。

此外，考虑到候选构式实例的出现频次，我们在BCC对话语料库(荀恩东et al., 2016)中检索候选构式实例，认为实例出现频次低于15次的候选构式并不是有效构式。

基于上述标准，两名研究生进行了候选构式的筛选工作。为保证数据集质量，我们仅将两名标注人员都认为是有效构式的构式纳入最终的构式数据集。

3.3 构式数据集相关信息

本文共对3588条语句中的3398个构式候选项进行了人工标注，最终构建了包含2146条语句、共计1748个有效构式的构式数据集。构式数据集样例如图3所示。其中`cnstr`字段用于标识各个词元是否属于构式实例的组成部分：标记为`O`表示该词元不属于任何构式，`BX`表示该词元为构式实例的起始词元，`IX`则表示其位于构式实例内部。本文还对构式集中各个常项与变项的出现频次及频率进行了统计，并依据构式中所包含的变项数量对构式数据集进行了分类。结果显示，变项数量分别为1至4时，对应的构式数量分别为1175条、528条、42条和3条。


```
{
"sentence":确实是这样的，但是我还是觉得上港是缺少突破点的，就拿这次的比赛来说，它们队里面的两员大将巨人都没有上场。",
"text": ["确实","是","这样","的","但是", "我", "还是","觉得", "上","港", "是", "缺少","突破点", "的", "就", "拿","这次", "的","比赛", "来说", "它们","队", "里面","的", "两", "员", "大将","巨人","都", "没有", "上场"],
"cnstr":["O","O","O","O","O","O","O", "O", "O", "O","O","O","O", "O", "O","O","BX",
"IX","IX","IX","IX", "IX","O","O", "O", "O", "O", "O","O","O", "O", "O", "O","O"],
"cnstr_form": [["就拿","np","来说"]],
"cnstr_example":[{"cnstr_example": "就拿这次的比赛来说","start": 14,"end": 19}]
}
```


图 3: 构式数据集样例

4 构式数据集有效性说明

4.1 词元覆盖率与压缩率

词元覆盖率指采用语义解析方法后语义解析所覆盖的词元比例。除此以外，我们还采用Dunn (2019)提出的压缩率（Compression Ratio）作为衡量构式数据集有效性的另一评估标准。该指标衡量在引入语法规则与不引入语法规则两种情形下，对语料库进行编码所需比特数的比值，反映语法规则对语料压缩效率的贡献。计算方法如式(1)所示。

$$\text{Compression Ratio} = \frac{\text{MDL}_{\text{Grammar}}}{\text{MDL}_{\text{Base}}} \tag{1}$$

下文以构式语法为例，具体说明压缩率指标的计算方式。为便于后续表述，现对相关符号做如下定义：$W = \{w_1, w_2, w_3, \ldots, w_n\}$表示语料库中的所有词元集合，$C = \{c_1, c_2, c_3, \ldots, c_m\}$表示语料库中出现的所有构式集合，$W_{\text{new}} = \{w_1, w_2, w_3, \ldots, w_{n_{\text{new}}}\}$表示语义解析后未覆盖的词元集合，$C_{\text{new}} = \{c_1, c_2, c_3, \ldots, c_{m_{\text{new}}}\}$表示所采用语义解析方法使用到的构式集合。

$\mathrm{MDL_{Base}}$为不采用任何语义解析方法时对语料库进行编码所需比特数，编码的基本单元为词元，核心思想为：出现频次越高的词元应该使用更少的编码数量。计算方法如式(2)所示，其中$P(w)$表示词元w在W中的出现概率，$\mathrm{freq}(w)$表示词元w出现频次。

$$\mathrm{MDL_{Base}} = \sum_{w \in W} \left(-\log_2 P(w) \times \mathrm{freq}(w)\right) \tag{2}$$

$\mathrm{MDL_{CxG}}$表示在使用构式数据集对语料库进行编码的情况下所需的比特数。计算公式如式(3)所示。其中$L_1(G)$ 表示对构式本身进行编码所需的比特数。L_2表示在使用构式数据集对语料进行解析的基础上对语料进行编码所需的比特数。L_2由两部分组成，一部分为对语料中含有的构式进行指示编码所需的比特数，即$L_2(C)$；一部分为对语料库中未被构式所覆盖的词元进行即时编码所需的比特数，即$L_2(R)$。

$$\mathrm{MDL_{CxG}} = L_1(G) + L_2(C) + L_2(R) \tag{3}$$

$L_1(G)$：它表示为解析过程中使用到的所有构式自身编码所需的比特数，计算公式如下：

$$L_1(G) = \sum_{c \in C_{\mathrm{new}}} c \tag{4}$$

构式由一系列的槽位（slot）填充构成，槽位可分为常项和变项两类。如图4 所示，每个构式的编码代价取决于其各个槽位的类型及其在该类型中的出现概率。构式的整体编码代价可由各槽位的编码代价之和给出，计算公式如式(5)所示，其中$\mathrm{N_{slot}}$表示构式中的槽位总数，$P(s_i)$表示第i个槽位在其所属类型(常项或变项)中出现的概率。

$$c = \sum_{i=1}^{N_{\mathrm{slot}}} -\log_2 P(s_i) \tag{5}$$

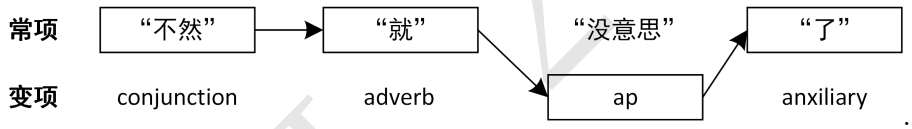

图 4: 构式(不然+就+ap+了)

$L_2(C)$和$L_2(R)$的计算方法如下，其中$P(c)$表示构式c在C_{new}中的出现概率，使用概率的负对数表示指示该构式所需的编码数量，核心思想为构式出现概率越高则其所需编码数量越少，$\mathrm{freq}(c)$表示该构式在待解析语料中的出现频次。$L_2(C)$代表了对语料库中构式解析到的词元进行编码的数量总和。同理，$P(w)$与$\mathrm{freq}(w)$分别表示词元w在W_{new}中出现的概率和频次，$L_2(R)$则代表了对语料库中构式未解析到的词元进行即时编码的数量总和。其余几种语义解析方法皆采用类似的思想进行计算，此处不再赘述。

$$L_2(C) = \sum_{c \in C_{\mathrm{new}}} \left(-\log_2 P(c) \times \mathrm{freq}(c)\right), \quad L_2(R) = \sum_{w \in W_{\mathrm{new}}} \left(-\log_2 P(w) \times \mathrm{freq}(w)\right) \tag{6}$$

4.2 语义粒度

本文提出了语义粒度评价指标，用于衡量语义解析方法对语料库解析的细致程度。该指标通过比较使用语义解析方法后语义解析方法覆盖词元部分的编码数量与未使用语义解析方法时该部分的编码数量之间的比值进行计算，核心思想为：语义解析后覆盖词元部分编码数量与之前相比越少，则其语义粒度越粗糙。因此，该指标越大说明语义解析方法的粒度越细。计算公式如下，其中β表示所使用的语义解析方法的词元覆盖率。

$$\mathrm{Granularity} = \frac{L_1(G) + L_2(C)}{\mathrm{MDL_{Base}} \times \beta} \tag{7}$$

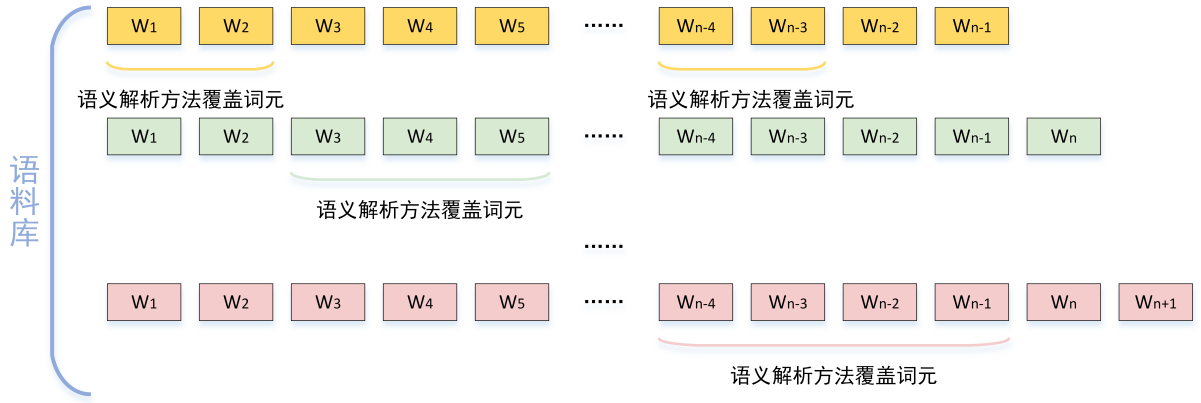

图 5: 语义解析结果

4.3 实验结果及分析

本文在NaturalConv的测试语料库(Wang et al., 2021)上对CFN[3]、LTP (Che et al., 2021)、OmniEvent (Peng et al., 2023)三种方法以及在其解析方法的基础上继续融入构式分别做了实验。实验结果如表2所示，CxG表示融入构式数据集。

方法	词元覆盖率(%)	压缩率(%)	使用到的构式数量	语义粒度(%)
CxG	13.26	95.95	1748	84.28
CFN	47.09	63.41	——	32.46
CFN + CxG	**50.93**	**62.05**	564	——
LTP	54.91	47.49	——	12.90
LTP + CxG	**56.90**	**46.88**	343	——
OmniEvent	3.16	96.79	——	30.63
OmniEvent + CxG	**15.85**	**93.05**	1691	——

表 2: 在NaturalConv测试集上的实验结果

实验结果表明，引入构式后，各种语义解析方法在原有基础上均实现了词元覆盖率的提升，同时压缩率有所下降，进一步验证了构式数据集在语义解析中的有效性与实用价值。值得注意的是，在使用语义解析方法完成初步分析后，再融入构式数据集，仍然能够识别出大量构式，这表明构式能够捕捉到之前的语义解析方法所无法解析的内容，进一步证实了引入构式的必要性。

就语义粒度而言，构式数据集的语义粒度最细，说明其在语义保留方面表现优异，能够较好地捕捉文本的语义信息。与OmniEvent相比，CxG不仅具有更低的压缩率，同时还展现出更细致的语义粒度，说明了构式在语义解析上具有一定的优势。CFN在压缩与语义保留方面表现较为平衡，既能够在一定程度上压缩信息，又能保留一定的语义信息。相比之下，虽然LTP的压缩率(47.49%)较低，但其语义粒度最粗(12.90%)，表明其语义损失较为严重，甚至可能导致关键信息的完全丢失。

5 评估大语言模型对构式的理解能力

我们从构式数据集中选取了4个代表性构式，分别为：“a+都+来不及”（构式1）、“有+什么+ap+的”（构式2）、“v+得+了+np”（构式3）以及“以+X+为+Y”（构式4），基于框架语义学理论，提出了构式（目标词）识别与构式语义理解两个评测任务，分别评估大语言模型在对话场景中发现构式语言结构和理解构式语义的能力。本文针对两个任务分别构建了一定规模的数据集，在deepseek-v3(Liu et al., 2024)、glm-4-plus(Zeng et al., 2024)、qwen2.5-32B(Tahmid and Sarker, 2024)等多个国内主流大模型以及gpt-4o(Hurst et al., 2024)上进行了测

[3]https://github.com/SXUNLP/CFSP

试，此外我们还选取了deepseek-r1(Guo et al., 2025)、glm-z1-air(Zeng et al., 2024)等四个具有深度推理能力的模型进行评测。

5.1 大语言模型对构式的识别能力

5.1.1 任务说明及数据集构建

对于每一个确定的构式，语料库中的语句可分为三类：(1) 不匹配构式形式；(2) 匹配构式形式但不表达构式含义；(3) 匹配构式形式且表达构式含义。如表3所示，以构式“有+什么+ap+的”为例，展示语料分类情况。

类别	样例
不匹配构式形式	同学，咱这里的食堂可以用手机银行支付吗?
匹配构式形式但不表达构式含义	很有道理，好期待我最喜欢的有道词典会有什么有趣的更新呢，嘿嘿。
匹配构式形式且表达构式含义	一起发和分开发都一样的钱，有什么高兴的。

表 3: 语料分类

我们将匹配构式形式且表达构式含义的语句视作正类，例如：(1)一起发和分开发都一样的钱，有什么高兴的。(2)英冠?这个有什么好看的，还不如看欧冠来的刺激一点。这两句话都匹配了构式“有+什么+ap+的”形式并且“有什么高兴的”和“有什么好看的”表达了该构式固有的否定含义。相反，将匹配构式形式但不表达构式含义的语句视作反类，例如：(1)很有道理，好期待我最喜欢的有道词典会有什么有趣的更新呢，嘿嘿。(2)有什么好听的歌快推荐一下哥在下载歌呢!快。这两句话虽在形式上符合“有+什么+ap+的”构式，但并未表达出与该构式对应的语义。

基于上述标注标准，我们从NaturalConv语料库(Wang et al., 2021)的训练集和验证集以及BCC对话语料库(荀恩东et al., 2016)中进行筛选，最终为每个构式构建了正反例各100条，共计800条的构式识别数据集。在此基础上，设计实验评估大语言模型在区分正类与反类语句方面的能力。具体而言，我们通过提示引导模型完成构式识别任务，并统计模型的分类正确率，从而评估大语言模型的构式识别能力，提示模板和示例如表4所示。

提示模板:
请判断以下句子中是否包含给出的构式，只需回答是或否即可，无需回答其他内容。\n 句子：{sentence} \n 构式：{cxg}
示例:
请判断以下句子中是否包含给出的构式，只需回答是或否即可，无需回答其他内容。 句子：一起发和分开发都一样的钱，有什么高兴的。\n 构式：(有+什么+ap+的)，其中ap表示形容词性短语。

表 4: 构式识别任务提示模板

5.1.2 实验结果及分析

表5的实验结果显示，模型在不同构式上的表现存在显著差异。整体来看，对于“a+都+来不及”和“以+X+为+Y”等结构相对简单的构式，具备思维链推理能力的模型表现尤为出色，部分模型的正确率甚至超过95%。值得注意的是，在处理“有+什么+ap+的”构式时，具备推理能力的模型反而表现不佳，正确率甚至低于普通模型。在表6中，我们展示了deepseek-r1在该构式上推理失败的案例：模型误将“有+什么+ap+的”（构式语义：不+ap）构式与“有+什么+ap+的+n”这种更常见的名词短语结构混淆，导致推理结果错误。这一现象表明，尽管具备推理能力的模型在多数构式中具有优势，但在面对句法歧义或与常规表达存在重叠的构式时，模型仍可能受到训练语料中频率偏差或分布错位的影响，导致判断错误。

模型	构式1	构式2	构式3	构式4
deepseek-v3	56.5	**81.5**	**74.5**	91.5
glm-4-plus	51.0	54.5	55.0	81.0
doubao-1.5-pro	70.0	53.5	61.0	**96.5**
ernie-4.5-8k-preview	74.5	65.0	57.5	91.5
gpt-4o	51.0	49.5	42.5	76.0
qwen2.5-32b	**77.5**	51.5	56.0	93.0
deepseek-r1	92.0	**76.0**	80.0	**98.0**
glm-z1-air	90.5	69.5	75.0	96.0
doubao-1.5-thinking-pro	**95.0**	52.5	**91.5**	**98.0**
ernie-x1-turbo-32k	92.0	63.0	67.5	97.5

表 5: 构式识别任务实验结果

句子：好呀，最近有什么好看的电影吗?
思考过程：...题目中的构式写作“(有+什么+ap+的)”，可能“的”作为结构助词，后面可以接名词，因此整个结构是存在的...
答案：是

表 6: deepseek-r1识别“有+什么+ap+的”构式失败案例

5.2 大语言模型对构式语义的理解能力

5.2.1 任务说明及数据集构建

为便于评估，我们将构式语义理解任务形式化为一个二分类问题，即判断给定的两个句子是否表达相同的语义。我们为每个构式提供了正例释义模板和反例释义模板，其中正例释义模板参考了北京大学构式库提供的释义模板。具体释义模板如表7所示。

构式编号	构式	正例释义模板	反例释义模板
1	a+都+来不及	非常+a	来不及+a
2	有+什么+ap+的	不+ap	挺+ap
3	v+得+了+np	能够+v+np	v+了+np
4(1)	以+X+为+Y	把+X+当作+Y	不把+X+当作+Y
4(2)	以+X+为+Y	把+X+当作+Y	把+X+变成+Y

表 7: 构式释义模板

具体而言，我们从5.1节构建的数据集中选取400条含有构式的语句，针对每条语句中的构式部分，分别按照正例释义模板与反例释义模板进行转换，而保留其余非构式部分不变。例如，句子“我怎么会紧张呢？我高兴都来不及”，分别按其正例释义模板和反例释义模板进行转换，生成与其语义相同和不同的句子“我怎么会紧张呢？我非常高兴”和“我怎么会紧张呢？我来不及高兴”。通过这一方法，为每个句子生成了一对语义相同和不同的正反例样本。

为了进一步检验模型是否真正掌握构式的深层语义，我们在“以+X+为+Y”构式中设置了两种类型的反例释义模板，一种与原构式语义完全相反，另一种在语义上与原构式相近但又有所区别。我们假设，若模型真正具备构式语义理解能力，其在这两种不同设定下的实验结果应该相近。通过测试大语言模型能否准确区分正例与反例，评估其在构式语义理解任务中的表现，提示模板及示例如表8所示。

第二十四届中国计算语言学大会论文集，第513页-第524页，济南，中国，2025年8月11日至14日。

提示模板:

请判断以下两个句子是否表达了相同的语义，只需回答是或否即可，无需回答其他内容。\n 句子1：{sentence1} \n 句子2：{sentence2}

示例:

请判断以下两个句子是否表达了相同的语义，只需回答是或否即可，无需回答其他内容。

句子1：我怎么会紧张呢？我高兴都来不及。\n 句子2：我怎么会紧张呢？我非常高兴。

表 8: 构式语义理解任务提示模板

5.2.2 实验结果及分析

模型	构式1	构式2	构式3	构式4(1)	构式4(2)
deepseek-v3	68.0	55.0	67.0	82.0	**78.5**
glm-4-plus	69.0	66.0	**80.0**	88.0	76.5
doubao-1.5-pro	69.0	**82.5**	65.0	96.5	63.0
ernie-4.5-8k-preview	69.0	80.5	57.0	**97.5**	65.0
gpt-4o	**78.5**	74.0	58.5	92.5	64.5
qwen2.5-32b	76.0	66.5	61.5	**97.5**	63.0
deepseek-r1	86.5	81.5	79.0	95.5	**82.5**
glm-z1-air	61.0	75.5	84.0	97.0	79.5
doubao-1.5-thinking-pro	**94.0**	85.5	82.5	**98.5**	81.5
ernie-x1-turbo-32k	88.0	**86.5**	**89.0**	98.0	77.5

表 9: 构式语义理解任务实验结果

表9的实验结果表明，具备推理能力的模型显著优于其他缺乏思维链支撑的模型，这说明思维链对于构式语义理解具有积极促进作用。整体来看，大多数模型在“以+X+为+Y”构式上表现最好，这说明该构式形式稳定、语义清晰，较易被模型捕捉和理解。相反，模型在面对歧义性较强的“a+都+来不及”构式时表现不佳，反映出当前大模型在缺乏推理引导的情况下，仍难以准确把握构式背后所蕴含的深层语义。

构式4(2)使用与“以+X+为+Y”的语义相近但又有所区别的反例，对比构式4(1)与构式4(2)的正确率，可以看出大多数模型在构式4(2)上正确率下降非常明显。例如，gpt-4o从92.5%降至64.5%，ernie-4.5-8k-preview从97.5%降至65.0%，表明模型容易受到语义近似实例的干扰，并未理解构式更深层的语义信息。gpt-4o在部分构式上表现不及其他大模型，显示出在中文构式理解方面仍有提升空间。

6 结论

本文基于NaturalConv对话语料库，采用思维链提示大模型与人工筛选相结合的方法，构建了一个包含2146条对话语句、1748个构式的中文构式数据集。在现有语义解析方法中融入该数据集，能够在保留语义信息的同时有效提升词元覆盖率并降低压缩率。基于该数据集，选取了其中部分代表性构式，基于框架语义学理论，提出了构式（目标词）识别与构式语义理解两项任务，用于评估大语言模型对构式的理解能力。通过对目前多个国内主流大模型与gpt-4o进行评估，研究发现即使是具有深度推理能力的大模型在面对部分构式时，构式识别能力不佳；且当前大模型在缺乏思维链推理的引导下，仍难以把握构式背后所蕴含的深层语义。

然而，本文构建的中文构式数据集仍存在一定局限性，未来工作将从以下几个方面对现有数据资源进行扩展：（1）引入更多元的的中文对话语料库提取构式，丰富构式来源，增加构式数据集中构式的数量，从而增强数据集的适用性；（2）为构式数据集中的构式增加更多高质量的实例；（3）为构式数据集构建语义场景，实现形式到意义的对应。

第二十四届中国计算语言学大会论文集，第513页-第524页，济南，中国，2025年8月11日至14日。

参考文献

Yida Cai, Hao Sun, Hsiu-Yuan Huang, and Yunfang Wu. 2024. Assessing the performance of chinese open source large language models in information extraction tasks. *arXiv preprint arXiv:2406.02079*.

Wanxiang Che, Yunlong Feng, Libo Qin, and Ting Liu. 2021. N-ltp: An open-source neural language technology platform for chinese. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 42–49.

Anindya Bijoy Das, Shibbir Ahmed, and Shahnewaz Karim Sakib. 2025. Hallucinations and key information extraction in medical texts: A comprehensive assessment of open-source large language models. *arXiv preprint arXiv:2504.19061*.

Jonathan Dunn. 2019. Frequency vs. association for constraint selection in usage-based construction grammar. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 117–128.

Charles J Fillmore. 1976. Frame semantics and the nature of language. *Annals of the New York Academy of Sciences*, 280(1):20–32.

Charles J Fillmore. 1988. The mechanisms of" construction grammar". In *Annual Meeting of the Berkeley Linguistics Society*, pages 35–55.

Adele E Goldberg. 1995. *Constructions: A construction grammar approach to argument structure*. University of Chicago Press.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Bo Li, Gexiang Fang, Yang Yang, Quansen Wang, Wei Ye, Wen Zhao, and Shikun Zhang. 2023a. Evaluating chatgpt's information extraction capabilities: An assessment of performance, explainability, calibration, and faithfulness. *arXiv preprint arXiv:2304.11633*.

Juncai Li, Zhichao Yan, Xuefeng Su, Boxiang Ma, Peiyuan Yang1, and Ru Li. 2023b. CCL23-eval 任务3总结报告:汉语框架语义解析评测(overview of CCL23-eval task 1:Chinese FrameNet semantic parsing). In Maosong Sun, Bing Qin, Xipeng Qiu, Jing Jiang, and Xianpei Han, editors, *Proceedings of the 22nd Chinese National Conference on Computational Linguistics (Volume 3: Evaluations)*, pages 113–123, Harbin, China, August. Chinese Information Processing Society of China.

Zuoheng Li, Xuanzhi Guo, Dengjian Qiao, and Fan Wu. 2023c. Ccl23-eval 任务3 系统报告: 基于旋转式位置编码的实体分类在汉语框架语义解析中的应用(system report for ccl23-eval task 3: Application of entity classification model based on rotary position embedding in chiness frame semantic parsing). In *Proceedings of the 22nd Chinese National Conference on Computational Linguistics (Volume 3: Evaluations)*, pages 94–104.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Hao Peng, Xiaozhi Wang, Feng Yao, Zimu Wang, Chuzhao Zhu, Kaisheng Zeng, Lei Hou, and Juanzi Li. 2023. Omnievent: A comprehensive, fair, and easy-to-use toolkit for event understanding. *arXiv preprint arXiv:2309.14258*.

Filip Seitl, Tomáš Kovářík, Soheyla Mirshahi, Jan Kryštfek, Rastislav Dujava, Matúš Ondreička, Herbert Ullrich, and Petr Gronat. 2024. Assessing the quality of information extraction. *arXiv preprint arXiv:2404.04068*.

Saad Tahmid and Sourav Sarker. 2024. Qwen2. 5-32b: Leveraging self-consistent tool-integrated reasoning for bengali mathematical olympiad problem solving. *arXiv preprint arXiv:2411.05934*.

Xiaoyang Wang, Chen Li, Jianqiao Zhao, and Dong Yu. 2021. Naturalconv: A chinese dialogue dataset towards multi-turn topic-driven conversation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14006–14014.

Peiyuan Yang, Juncai Li, Zhichao Yan, Xuefeng Su, and Ru Li. 2024. Chinese frame semantic parsing evaluation. In *Proceedings of the 23rd Chinese National Conference on Computational Linguistics (Volume 3: Evaluations)*, pages 32–42.

Liping You and Kaiying Liu. 2005. Building chinese framenet database. In *Natural Language Processing and Knowledge Engineering, 2005. IEEE NLP-KE '05. Proceedings of 2005 IEEE International Conference on*.

Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *CoRR*.

张立凡. 2016. 基于规则和分类模型的核心目标词识别研究. 太原师范学院学报*(*自然科学版*)*, 15(03):32–38.

王晓晖, 李茹, 王智强, 柴清华, and 韩孝奇. 2022. 基于self-attention的句法感知汉语框架语义角色标注. 中文信息学报, 36(10):38–44.

石佼, 李茹, and 王智强. 2014. 汉语核心框架语义分析. 中文信息学报, 28(06):48–55.

荀恩东, 饶高琦, 肖晓悦, and 臧娇娇. 2016. 大数据背景下bcc语料库的研制. 语料库语言学, 3(01):93–109+118.

詹卫东. 2017. 从短语到构式:构式知识库建设的若干理论问题探析. 中文信息学报, 31(01):230–238.