山西大学学报(自然科学版)47(5):1004—1014,2024
Journal of Shanxi University(Nat. Sci. Ed.)

DOI:10.13451/j.sxu.ns.2014112

大语言模型的汉语框架语义分析能力评估

高俊杰[1*],马博翔[1],闫智超[1],苏雪峰[1,2],李茹[1,3]
(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.计算智能与中文信息处理教育部重点实验室,山西 太原 030006;)

摘 要:大语言模型的出现对自然语言处理产生了广泛的影响,已有研究表明大语言模型在各类下游任务中具有出色的Zero-shot及Few-shot能力,而对于大语言模型的语义分析能力的评估仍然比较缺乏。因此,本文基于汉语框架语义分析中的三个子任务:框架识别、论元范围识别和论元角色识别,分别在Zero-shot及Few-shot设定下评估了ChatGPT、Gemini和ChatGLM三个大语言模型在CFN2.0数据集上的语义分析能力,并与目前基于BERT(Bi-directional Encoder Representations from Transformers)的SOTA模型进行了比较。在框架识别任务中,大语言模型的准确率仅比SOTA模型低0.04;但在论元范围识别与论元角色识别任务上,大语言模型表现不佳,与SOTA(State of the Art)模型相比,*F*1分数分别相差0.13和0.39。以上结果表明,大语言模型虽具备一定的框架语义分析能力,但进一步提升大语言模型的语义分析能力仍然是一个具有挑战性的工作。

关键词:大语言模型;框架识别;论元范围识别;论元角色识别
中图分类号:TP39 **文献标志码:**A **文章编号:**0253-2395(2024)05-1004-11

Evaluation of Chinese Frame Semantic Analysis Capabilities of Large Language Models

GAO Junjie[1*], MA Boxiang[1], YAN Zhichao[1], SU Xuefeng[1,2], LI Ru[1,3]
(1. *School of Computer and Information Technology, Shanxi University, Taiyuan* 030006*, China;*
2. *School of Modern Logistics, Shanxi Vocational University of Engineering Science and Technology, Jinzhong* 030609*, China;*
3. *Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Taiyuan* 030006*, China*)

Abstract: The emergence of large language models (LLMs) has a widespread impact on natural language processing. Studies have shown that the LLMs have excellent Zero-shot and Few-shot capabilities in various downstream tasks, but the evaluation of the semantic analysis capabilities of the LLMs is still lacking. Therefore, based on three subtasks in Chinese frame semantic analysis: frame identification, argument identification, and role identification, this paper evaluates the semantic analysis capabilities of three LLMs, namely ChatGPT, Gemini, and ChatGLM, on the CFN2.0 dataset under Zero-shot and Few-shot settings, and compares them with the current BERT-based SOTA model. In the frame identification task, the accuracy of the LLMs is only 0.04 lower than that of the SOTA model. However, in the argument identification and role identification task, the performance of the LLMs is suboptimal, with *F*1 scores differing by 0.13 and 0.39, respectively compared to the SOTA model. The above results show that although the LLMs have certain frame semantic analysis capabilities, further improving the semantic analysis capabilities of LLMs is still a challenging work.

Key words: large language model; frame identification; argument identification; role identification

收稿日期:2024-04-18;**接受日期:**2024-07-07
基金项目:山西省科技合作交流专项项目(202204041101016);山西省基础研究计划项目(202203021211286);国家自然科学基金重点项目(61936012)
* **通信作者:**高俊杰(1984-),男,山西浑源人,实验师,研究方向为自然语言处理。E-mail: gaojunjie@sxu.edu.cn
引文格式:高俊杰,马博翔,闫智超,等.大语言模型的汉语框架语义分析能力评估[J].山西大学学报(自然科学版),2024,47(5):1004-1014. DOI:10.13451/j.sxu.ns.2014112

0 引言

大语言模型凭借其庞大的参数化知识和强大的推理能力,在自然语言理解与生成和复杂推理等任务上成效显著,尤其是LLaMA[1-2]、ChatGLM[3]、GPT[4-6]等系列的新一代大语言模型的出现,在关系抽取[7]、文本摘要[8]等各种自然语言处理任务中展现出令人瞩目的效果。随着相关研究的不断深入,借助上下文学习、思维链等方法[9]进一步提升了大语言模型在许多下游任务中的性能,对自然语言处理领域产生了巨大的影响。

探究大语言模型在各种下游任务上的能力边界已成为一个热点问题。Bang等[10]聚合了覆盖8种不同自然语言处理任务的23个数据集,对ChatGPT在不同任务上的性能进行了全面的评估。Bian等[11]针对大语言模型所掌握的常识进行了评估,发现ChatGPT虽然拥有丰富的知识,但在解决问题的经验上存在欠缺。Gao等[12]、Wei等[13]、Li等[14]均在信息抽取任务上评估了大语言模型的能力,结果表明大语言模型在多数情况下具有良好的真实性,但偶尔会出现过度自信的问题。Yuan等[15]和EvEval[16]在事件抽取任务上评估了大语言模型的性能,其结论表明,尽管大语言模型对单一事件有所了解,但它们感知事件之间语义相似性的能力存在欠缺。Xie等[17]对大语言模型的实体识别能力进行了评估,并提出了包括句法分析在内的四种不同策略来增强其命名实体识别能力,并证明了额外引入的句法分析结果对大模型命名实体识别能力具有增强作用。以上研究虽然在一定程度上探索了大语言模型在各种下游任务上的能力边界,但缺少对大模型本身语义分析能力的评估,导致大语言模型所具备的语义理解能力尚不明确,这阻碍了大语言模型的进一步研究与应用,尤其是在中文上,由于训练语料相对英文较少,对其进行全面评估更是非常有必要的。为此,本文基于框架语义分析任务[18],在Chinese FrameNet 2.0(CFN2.0)数据集上对目前主流支持中文的大语言模型进行了一系列的评估,探究了大语言模型在语义分析任务上的能力边界。

框架语义分析是以Fillmore的框架语义学[19]为基础的语义分析任务,该任务旨在通过三个子任务:框架识别、论元边界识别和论元角色识别,从框架语义学的角度将句子解析为结构化的表示形式[20-22]。具体而言,框架语义学使用语义框架来表示事件的语义场景,使用框架元素来表示参与这一事件的语义角色,这种结构化形式更具表达力,对于阅读理解[23-25]、文本摘要[26-27]、关系抽取[28]和文本生成[29]等下游任务具有重要意义。如图1所示,在例句"他组织班级的同学明天参加由学院举办的学术研讨会"中,目标词"组织"激活了"安排"框架。"他"作为施动者,实施了安排的动作;"班级的同学"作为受益人,是被安排的对象;"明天"是事件发生的时间,"参加由学院举办的学术研讨会"是安排进行的具体事件。因此,整个句子的语义场景可以概括为:施动者安排受益人在特定时间进行某一事件。我们可以将句子中的短语与框架元素相匹配,得到其结构化表示。这种结构化表示全面地刻画了语义场景下的各个角色,对于语义理解具有重要作用[30-31]。此外,由图1中的示例可见,框架语义分析任务具有较细的粒度,需要从完整的句义中抽象出目标词所触发的语义场景,并细致分析句子中各个短语的划分、短语含义、短语之间的关系等,进而分析出这一语义场景下与目标词相关的各种语义角色。这种较细的粒度使得框架语义分析具有更细致的表达能力,能够更好地作用于下游任务,但同样也增强了汉语框架语义分析任务本身的难度。

以框架语义分析任务为背景,我们为了更好地评估大语言模型的语义分析能力,构建了一系列不同的提示模板,在Zero-shot和Few-shot两种设置下,基于框架识别、论元范围识别、论元角色识别三个框架语义分析的子任务,对大语言模型的框架语义分析能力进行了评估和测试,并对评估结果进行了分析。结果表明,大语言模型在框架语义分析能力和提示信息利用能力上存在不足,即使是在思维链的引导下仍然难以激发出其框架语义分析能力。

1 任务定义

1.1 框架识别

框架识别(Frame Identification, FI)任务需要

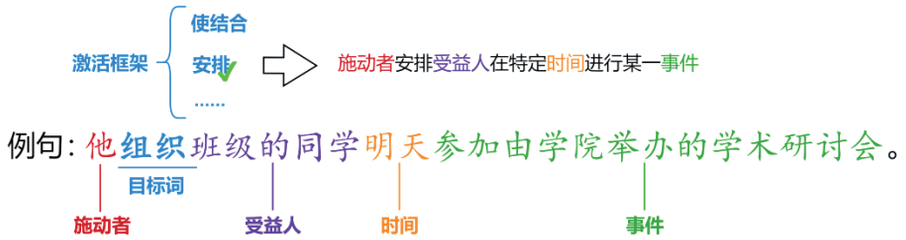

图1 框架语义分析示例

Fig. 1 Example of frame semantic analysis

大语言模型为句子中的目标词匹配最适合的语义框架,其主要的挑战是目标词通常是有歧义的,会激活多个语义框架,如图1中的"组织"在不同场景下可以激活"使结合"和"安排"等多个框架。该任务的定义为:给定一个句子$S=\{w_1,w_2,\cdots,w_n\}$,目标词$w_t\in S(1\leqslant t\leqslant n)$,要求通过目标词$w_t$在句子$S$中的上下文来理解语义场景,并从给定的框架集合$F=\{f_1,f_2,\cdots,f_m\}$中选择出最恰当的框架$f_t$。该任务主要评估大语言模型对句子中目标词语义场景的理解和概括能力。在图1的实例中,框架识别任务需要在"使结合""安排"等框架组成的集合中确定目标词"组织"在句子中所属的框架为"安排"。

1.2 论元范围识别

论元范围识别(Argument Identification, AI)任务需要大语言模型从给定的句子$S=\{w_1,w_2,\cdots,w_n\}$中找出目标词$w_t\in S(1\leqslant t\leqslant n)$所支配的全部论元$a_t\in\{a_1,a_2,\cdots,a_k\}$,其主要挑战在于论元的范围较长,数量也不确定。该任务能够评估大语言模型分析并寻找句子中与目标词在语义上相关的片段的能力。如图1所示,论元范围识别任务需要在句子中找出目标词"组织"所支配的所有论元:"他""班级的同学""明天""参加由学院举办的学术研讨会"。

1.3 论元角色识别

论元角色识别(Role Identification, RI)任务需要大语言模型将目标词所支配的全部论元与目标词所属框架的框架元素进行匹配,确定每个论元在其所属框架中对应的语义角色,主要的挑战在于角色数量多,框架语义知识库中的角色数量上千,其分类难度较高。该任务的定义为:给定一个句子$S=\{w_1,w_2,\cdots,w_n\}$,已知其中的目标词$w_t\in S(1\leqslant t\leqslant n)$及其所激活的框架$f_t$和目标词在句子中所支配的全部论元$a_t\in\{a_1,a_2,\cdots,a_k\}$,该任务需要大语言模型将这些论元映射到框架$f_t$具有的框架元素$R_f=\{r_1,r_2,\cdots,r_k\}$中。该任务是框架语义分析任务的最终步骤,需要大语言模型对目标词语义场景中所有参与者的具体角色进行分析,完整解析出目标词在句子中的框架语义信息。如图1所示,论元角色识别任务需要在句子中找出目标词"组织"所支配的所有论元,并将这些论元分别映射到"安排"框架下的"施动者""受益人""时间"和"事件"角色。

2 实验

2.1 实验设定

2.1.1 数据集

我们使用CFN2.0数据集[18]来评估大语言模型的框架语义分析能力。该数据集来源于山西大学中文信息处理团队,其测试集包含4 000个例句,覆盖了432个不同的框架、711种不同的框架元素。但受限于调用大语言模型的成本,我们从中随机采样了50个不同的框架,并为这50个框架各随机采样了2个例句,形成了包含100个例句、覆盖50个框架的数据集。

2.1.2 实验方案

我们的实验主要针对Zero-shot和Few-shot两种不同场景设定,采用的提示模板结构如图2所示。在Zero-shot场景下,我们在提示模板中不提供任何带有答案的信息,仅说明任务需求,要求大语言模型解决框架语义分析的相关问题。这样的场景设定主要评估大语言模型自身是否具有框架语义的相关知识,并分析其能否利用相关知识解决框架语义分析的相关问题。而在Few-shot场景下,我们在提示模板中

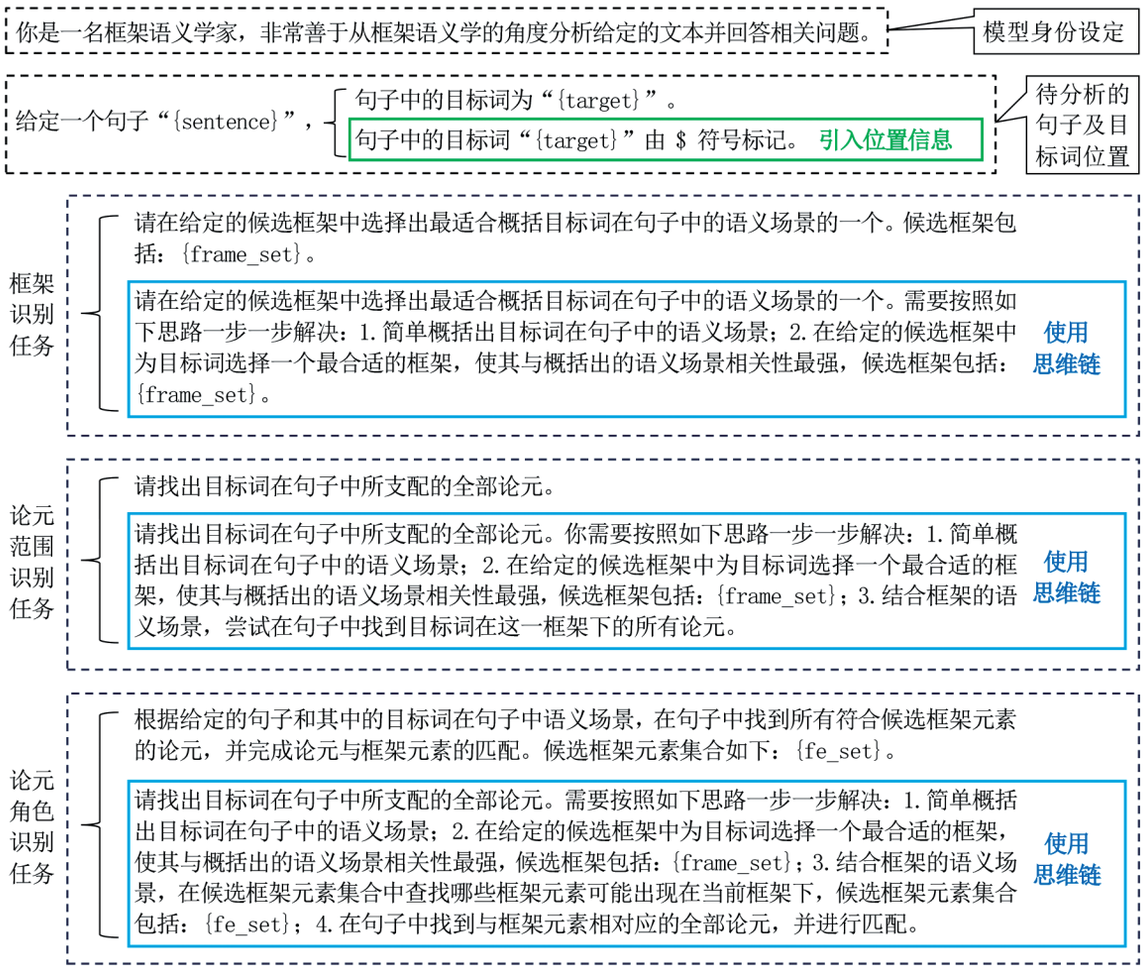

图2 提示模板示例

Fig. 2 Example of prompt template

引入了少量带有正确答案的示例样本，并要求大语言模型按照模板中提供的示例格式进行输出。这使得大语言模型能够通过给定样例中携带的正确答案更好地理解任务需求，从而评估大语言模型能否有效利用自身的上下文学习能力从少量的示例信息学习相关知识来提升其框架语义分析能力。此外，受到Wang等[32]的启发，我们注意到当同一目标词在句子中出现多次时，若不显式地标记出句子中的目标词，则大语言模型将无法确定需要进行框架语义分析的目标词的具体位置。因此，我们设计了不同的提示方法，在句子中使用“$”符号作为位置提示来标出目标词，评估大语言模型在引入目标词位置信息前后的框架语义分析任务性能差异。此外，我们也设计了基于思维链的提示模板，评估大语言模型在引入思维链前后的框架语义分析能力变化。

2.2 评价指标

根据上述三项框架语义分析任务的不同特点，我们使用了不同的评价指标来评估解析结果。

对于框架识别任务，我们将正确率Acc_{FI}作为评价指标。其定义为：给定n_{total}个句子和一个候选框架集合$F'=\{f_1,f_2,\cdots,f_m\}$，使用大语言模型为每一个句子中的目标词在$F'$中选择一个框架，将选择正确的句子数量记为$n_{\mathrm{correct}}$，则框架识别任务的$Acc_{\mathrm{FI}}$定义如下：

$$Acc_{\mathrm{FI}}=\frac{n_{\mathrm{correct}}}{n_{\mathrm{total}}}。\tag{1}$$

对于论元范围识别任务，我们统计每一个预测结果与所有真实标签的最大重合字数，并计算$F1$作为评价指标。具体来说，给定一组句子及其所包含的所有真实论元$a_{\mathrm{gold}}=\{a_{g1},a_{g2},\cdots,a_{gm}\}$，将模型预测出的所有论元记作$a_{\mathrm{pred}}=\{a_{p1},a_{p2},\cdots,a_{pn}\}$，则论元范围识别任务的$F1_{\mathrm{AI}}$计算如下：

$$P_{\mathrm{AI}}=\frac{\sum_{i=1}^{n}\max\left(len(a_{pi}\cap a_{g1}),len(a_{pi}\cap a_{g2}),\cdots,len(a_{pi}\cap a_{gm})\right)}{\sum_{i=1}^{n}len(a_{pi})},$$

$$R_{\mathrm{AI}}=\frac{\sum_{i=1}^{m}\max\left(len(a_{gi}\cap a_{p1}),len(a_{gi}\cap a_{p2}),\cdots,len(a_{gi}\cap a_{pn})\right)}{\sum_{i=1}^{m}len(a_{gi})},$$

$$F1_{\mathrm{AI}}=\frac{2\times P_{\mathrm{AI}}\times R_{\mathrm{AI}}}{P_{\mathrm{AI}}+R_{\mathrm{AI}}}。\tag{2}$$

对于论元角色识别任务,我们同样计算$F1$值作为评价指标。与论元范围识别任务不同的是,论元角色识别任务的评估只有在论元边界和论元角色均与标签完全一致的情况下才被认为是正确的。具体来说,给定一组句子及其所包含的所有真实论元$a_{\mathrm{gold}}=\{a_{g1},a_{g2},\cdots,a_{gm}\}$和对应的角色$r_{\mathrm{gold}}=\{r_{g1},r_{g2},\cdots,r_{gm}\}$,将模型预测出的所有论元记作$a_{\mathrm{pred}}=\{a_{p1},a_{p2},\cdots,a_{pn}\}$,预测出的对应角色记作$r_{\mathrm{pred}}=\{r_{p1},r_{p2},\cdots,r_{pn}\}$,完全正确的预测结果数量记为$k_{\mathrm{correct}}$,则论元角色识别任务的$F1_{\mathrm{RI}}$计算如下:

$$P_{\mathrm{RI}}=\frac{k_{\mathrm{correct}}}{n},\ R_{\mathrm{RI}}=\frac{k_{\mathrm{correct}}}{m},$$

$$F1_{\mathrm{RI}}=\frac{2\times P_{\mathrm{RI}}\times R_{\mathrm{RI}}}{P_{\mathrm{RI}}+R_{\mathrm{RI}}}。\tag{3}$$

2.3 实验结果

2.3.1 使用基础提示模板的实验结果

我们在上述实验设置下对多个大语言模型进行了评估,包括ChatGPT-3.5[33]、Gemini-Pro[34]以及ChatGLM2-6B[3]。ChatGPT-3.5是由OpenAI于2022年推出的商用大语言模型,通过大量语料进行训练,并通过监督学习和强化学习技术进行微调,使其能够理解和生成自然语言,在自然语言处理领域的各项任务中展现出了较好的性能。Gemini系列商用大语言模型由Google于2023年推出,同样在各项自然语言处理领域的任务上具有出众的效果,其特点在于原生具备对多模态的支持,根据模型规模大小分为Nano、Pro和Ultra版本,本文采用的Gemini-Pro是中等规模的版本。而ChatGLM2则是由清华大学于2023年推出的开源大语言模型,其参数量仅为6 B,规模远小于ChatGPT-3.5和Gemini-Pro,但同样在自然语言处理领域的通用任务上具有不俗的表现,且具有便于部署的特点。作为对比的SOTA(State of the Art)模型[35]均为以BERT(Bidirectional Encoder Representations from Transformers)[36]为基础进行微调后的模型,该方法使用旋转矩阵对绝对位置进行编码,同时将显式的相对位置依赖性纳入自注意公式中,在框架识别等任务中达到了SOTA。主要实验结果见表1(Zero-shot场景)和表2(Few-shot场景)。表中加粗表示最高的分数,下划线表示第二高的分数。

由表1与表2中的主要实验结果可见,大语言模型在Few-shot场景下的框架语义分析能力要显著优于Zero-shot场景,这表明大语言模型具备的上下文学习能力在框架语义分析任务中能够发挥积极作用,使其能够根据提示样例更好地理解任务需求。然而,无论是在Zero-shot还是Few-shot场景下,大语言模型在框架语义分析任务中的表现与传统模型相比仍存在一定的差异,特别是随着任务粒度的细化,大语言模型的表现下滑明显。特别的,即使是本次评估中性能最佳的Gemini,也仅在框架识别任务上达到了与SOTA模型相近的性能,而在粒度较细的论元范围识别和论元角色识别任务上,其性能与SOTA模型的性能差异逐渐加大。总体而言,我们的评估结果表明现阶段的大语言模型具备了一定的框架语义分析能力,但是并不能很好地理解粒度较细的语义信息,因此其在框架语义分析任务上仍存在一定不足。

2.3.2 引入目标词位置信息的实验结果

为了分析目标词位置信息对大模型框架语义分析能力的影响,我们设计实验对比了在提示信息中引入目标词的位置信息对各项评估任务性能的影响。在两种少样本场景下,不同目标词位置信息设定的实验结果如表3(Zero-shot

表1 Zero-shot场景下的主要实验结果

Table 1 Main results in Zero-shot settings

模型	FI	AI			RI		
	Acc	Precision	Recall	*F*1	Precision	Recall	*F*1
SOTA	**0.74**	**0.908 3**	**0.828 8**	**0.866 8**	**0.595 8**	**0.574 1**	**0.584 7**
ChatGPT	0.66	0.475 7	0.803 8	0.597 6	0.130 2	0.210 5	0.160 9
Gemini	0.66	0.842 9	0.557 0	0.670 8	0.167 5	0.263 2	0.204 7
ChatGLM2	0.14	0.390 0	0.148 1	0.214 7	0.009 0	0.022 6	0.012 9

表2 Few-shot场景下的主要实验结果

Table 2 Main results in Few-shot settings

模型	FI	AI			RI		
	Acc	Precision	Recall	*F*1	Precision	Recall	*F*1
SOTA	**0.74**	**0.908 3**	**0.828 8**	**0.866 8**	**0.595 8**	**0.574 1**	**0.584 7**
ChatGPT	0.70	0.055 3	0.679 7	0.611 3	0.175 4	0.150 1	0.161 9
Gemini	0.70	0.730 0	0.739 2	0.734 6	0.222 2	0.180 5	0.199 2
ChatGLM2	0.36	0.373 0	0.741 8	0.496 4	0.064 3	0.165 4	0.092 6

场景)和表4(Few-shot场景)所示。

实验结果表明,即使是细微的位置信息的变动也会给其解析能力带来较为显著的扰动,且这种扰动是难以预知的,由此可见大语言模型在框架语义分析任务上对提示信息的利用是不稳定的。

2.3.3 引入思维链的实验结果

此外,为分析思维链对大语言模型框架语义分析任务的影响,我们分别使用传统提示构建方法以及基于思维链的构建方法来构建提示模板。在不同的提示模板构建方法上得到的实验结果见表5(Zero-shot场景)和表6(Few-shot场景)。

由实验结果可见,即使使用思维链引导,也并不能稳定地激发出大语言模型的框架语义分析能力,且对于一些参数量较小的模型(如ChatGLM2),甚至会直接导致输出内容不可控,这也体现了大语言模型在框架语义分析能力上的不足。

表3 Zero-shot场景下不同目标词位置信息设定的实验结果

Table 3 Results of different positional information of target word in Zero-shot settings

模型	是否引入位置信息	FI	AI			RI		
		Acc	Precision	Recall	*F*1	Precision	Recall	*F*1
ChatGPT	否	0.54	**0.482 1**	0.749 4	0.586 7	0.035 1	0.060 2	0.044 3
	是	**0.66**	0.475 7	**0.803 8**	**0.597 6**	**0.062 5**	**0.082 7**	**0.071 2**
Gemini	否	0.54	0.641 3	**0.681 0**	0.660 6	**0.167 5**	**0.263 2**	**0.204 7**
	是	**0.66**	**0.842 9**	0.557 0	**0.670 8**	0.143 5	0.255 6	0.183 8
ChatGLM2	否	**0.14**	0.315 2	0.073 4	0.119 1	0.005 2	0.007 5	0.006 2
	是	0.10	**0.390 0**	**0.148 1**	**0.214 7**	**0.009 0**	**0.022 6**	**0.012 9**

表4 Few-shot场景下不同目标词位置信息设定的实验结果

Table 4 Results of different positional information of target word in Few-shot settings

模型	是否引入位置信息	FI	AI			RI		
		Acc	Precision	Recall	*F*1	Precision	Recall	*F*1
ChatGPT	否	**0.70**	0.528 4	**0.696 2**	0.601 4	0.100 7	**0.105 3**	**0.102 9**
	是	0.64	**0.555 3**	0.679 7	**0.611 3**	**0.107 4**	0.097 7	0.102 3
Gemini	否	**0.70**	**0.759 8**	0.708 9	0.733 5	**0.222 2**	**0.180 5**	**0.199 2**
	是	0.68	0.730 0	**0.739 2**	**0.734 6**	0.167 9	0.172 9	0.170 4
ChatGLM2	否	**0.32**	0.373 0	**0.741 8**	**0.496 4**	0.050 2	0.112 8	0.069 4
	是	0.30	**0.379 6**	0.668 4	0.484 2	**0.064 3**	**0.165 4**	**0.092 6**

表5 Zero-shot场景下有关思维链引入的实验结果

Table 5 Results on the use of Chain-of-Thought in Zero-shot settings

模型	是否使用思维链	FI	AI			RI		
		Acc	Precision	Recall	*F*1	Precision	Recall	*F*1
ChatGPT	否	0.52	**0.512 4**	0.627 8	0.564 3	0.117 2	**0.225 6**	0.154 2
	是	**0.66**	0.475 7	**0.803 8**	**0.597 6**	**0.130 2**	0.210 5	**0.160 9**
Gemini	否	0.52	**0.842 9**	0.597 5	**0.670 8**	**0.167 5**	**0.263 2**	**0.204 7**
	是	**0.66**	0.641 3	**0.681 0**	0.660 6	0.146 3	0.180 5	0.161 6
ChatGLM2	否	0.10	**0.390 0**	**0.073 4**	**0.214 7**	**0.009 0**	**0.022 6**	**0.012 9**
	是	**0.14**	0	0	0	0	0	0

表6 Few-shot场景下有关思维链引入的实验结果

Table 6 Results on the use of Chain-of-Thought in Few-shot settings

模型	是否使用思维链	FI	AI			RI		
		Acc	Precision	Recall	*F*1	Precision	Recall	*F*1
ChatGPT	否	0.62	0.473 6	**0.739 2**	0.577 4	0.154 4	**0.157 9**	0.156 1
	是	**0.70**	**0.555 3**	0.679 7	**0.611 3**	**0.175 4**	0.150 4	**0.161 9**
Gemini	否	0.62	**0.730 0**	0.739 2	**0.734 6**	**0.222 2**	**0.180 5**	**0.199 2**
	是	**0.70**	0.591 3	**0.807 6**	0.682 7	0.167 9	0.172 9	0.170 4
ChatGLM2	否	0.32	0.339 2	**0.783 5**	0.473 4	**0.064 3**	**0.165 4**	**0.092 6**
	是	**0.36**	**0.373 0**	0.741 8	**0.496 4**	0.034 0	0.037 6	0.035 7

2.3.4 不同温度系数下的实验结果

除提示模板外，由于大语言模型自身生成时具有一定的随机性，且这一随机性与温度系数呈正相关，在不同的温度系数下其输出结果存在很大的差异。因此我们分析了这一随机性对其框架语义分析能力的影响。对于每一个任务的不同模板设定，我们均分别测试了大语言模型在 $T=0.1, 0.3, 0.5, 1.0$ 四种不同温度系数下的表现，并统计了其在不同的温度系数下取得最好成绩的次数，实验结果如图3所示。

结果表明，大语言模型更偏向于在较低的温度系数下完成框架语义分析任务，且仅有极少数的实验在温度系数为1.0时取得了最好的效果。具体而言，在Few-shot场景下，温度系数为0.5时有最多的实验达到了最好效果，而在Zero-shot场景下时这一数值为0.1。我们认为，在Zero-shot场景下，较大的温度系数导致模型随机性过高，在没有充足示例的情况下导致输出偏离了任务本身；而由于Few-shot场景下的输入中含有示例，因此在相对较高的温度系数下其输出仍然可控，且相对较高的温度系数更好地激发出大语言模型的表达能力。由此可见，温度系数的改变同样对大语言模型的框架语义解析能力具有非常显著的影响。

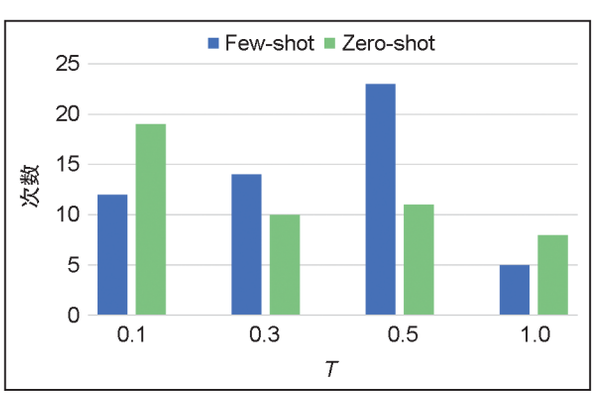

图3 大语言模型在不同温度(T)系数下取得最好成绩的次数分布

Fig. 3 The distribution of the number of times that LLMs achieve the best results at different temperature settings T

3 分析与讨论

3.1 语义消歧能力分析

我们通过上述大量实验结果可以看出，大语言模型并不能很好地完成框架语义分析任务，其性能与传统模型相比仍有较大差距。一方面，框架语义分析需要具有一定的框架语义专业知识，而大语言模型虽然经过了海量文本数据的训练，但这些数据中包含的框架语义专业知识的数量、质量是难以确定的，这会严重影响大语言模型对框架语义信息的理解和处理。另

一方面,框架语义分析任务粒度较细,其中包含了许多细致、严谨的定义,且不同定义之间的差异可能是非常细微的。如图4所示,“供应”与“提供”两个框架均表示转移体在两者之间转移的场景,其定义上的区分仅在于更强调“提供者”的“提供意愿”还是“接收者”的“接受意愿”,而这种区分是非常细致的。这使得汉语框架语义能够更细致地刻画语义场景,但对于大语言模型而言则极大地增强了其进行框架语义分析的难度。大语言模型虽然擅长对输入的上下文进行处理,但它们在理解词语和概念之间的复杂关系和识别歧义能力方面存在局限性。

3.2 论元边界识别能力分析

本节以论元范围识别任务为例,对大模型的论元边界识别能力进行样例分析,重点关注了目标词位置信息的影响。图5为论元范围识别任务中表现最好的Gemini模型在不同位置信息提示设定下的识别结果样例。在例句中,目标词“丰富”出现了两次,因此在没有明确指定目标词位置信息的情况下,大语言模型会受到无关词语的影响而误判或遗漏论元,而在提示中加入位置信息则缓解了这一问题。特别地,正如示例中所示,我们在Zero-shot场景下的绝大多数实验中引入位置信息都达到了相对更好的效果,但在Few-shot场景中却相反。这是由于在Zero-shot场景中不存在提示样例,导致大语言模型对任务需求的理解有限,而额外引入的位置信息提示作为任务需求的补充,对于大语言模型更好地理解

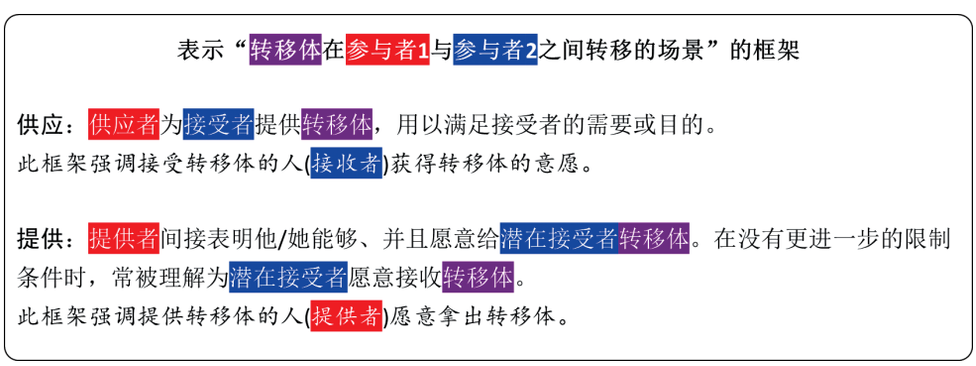

图4 细粒度框架定义示例

Fig. 4 Examples of fine-grained frame definition

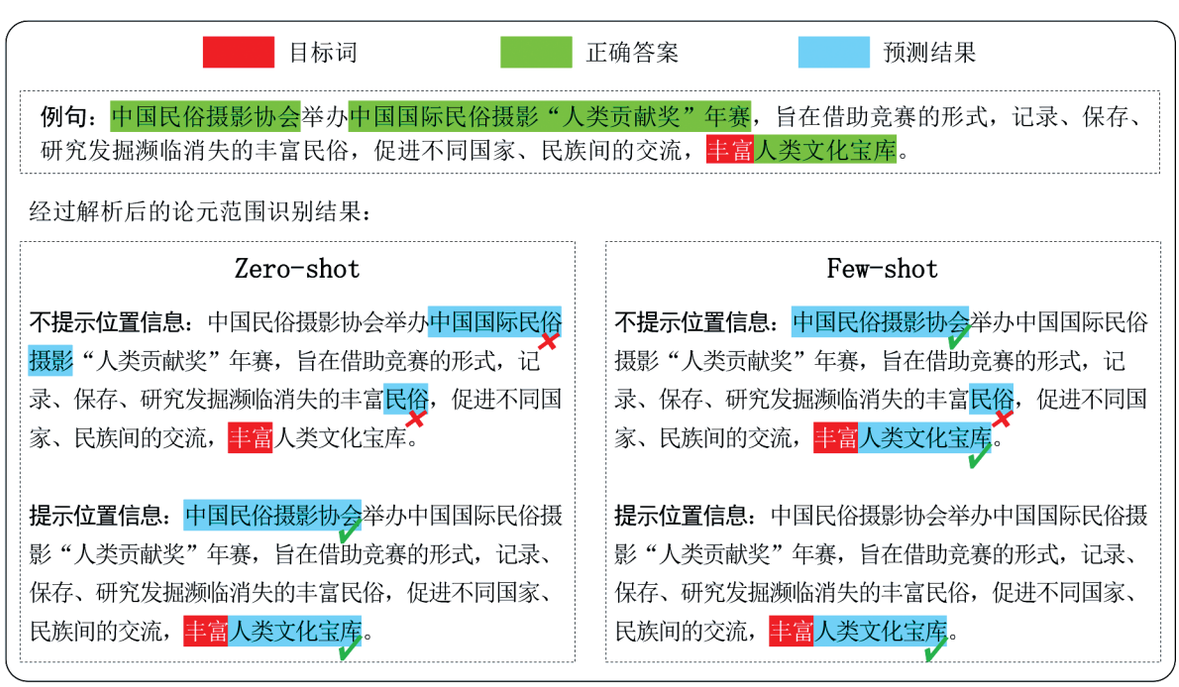

图5 位置信息提示对Gemini论元范围识别影响的样例

Fig. 5 Examples of the impact of positional information prompts on arguments identification of Gemini

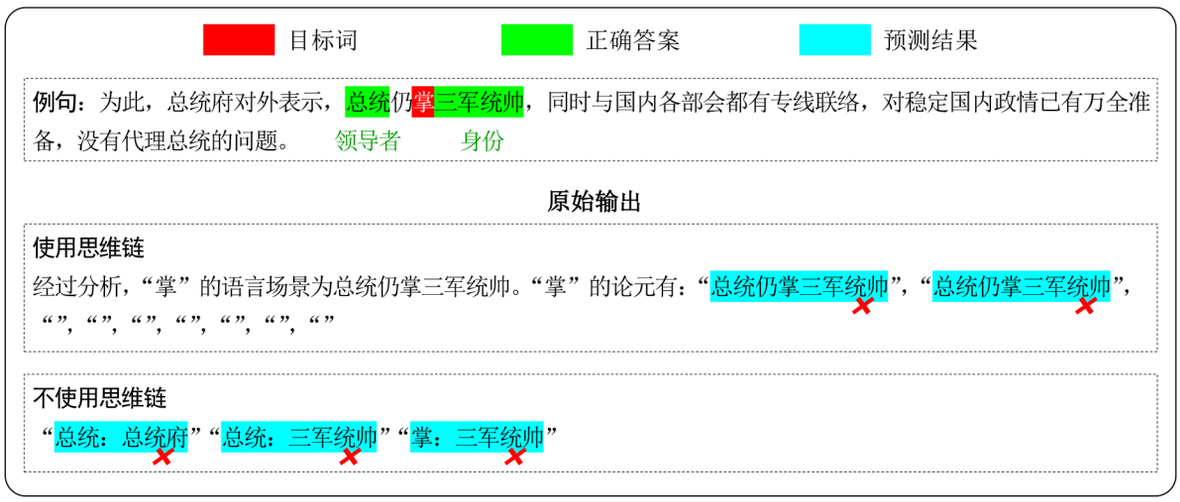

图6 ChatGLM2语义角色识别结果错误的样例

Fig. 6 An example of incorrect results on the role identification of ChatGLM2

任务具有积极作用,进而使其能够更准确地进行推理;而Few-shot场景下的提示样例已经能够清晰地描述任务所需的输入和输出,此时引入的位置信息可能会被大语言模型理解为噪音,对其推理过程造成干扰。

3.3 论元角色识别能力分析

本节以论元角色识别任务为例,对大语言模型的论元角色识别能力进行样例分析,并重点关注了思维链推理对其的影响。如图6所示,参数量较小的ChatGLM2在该任务上产生了完全不可控的输出,无论是否使用思维链进行引导,均出现了$F1$值为0的情况,即其输出内容完全错误。由此可见,一些大语言模型在语义论元识别能力上仍存在欠缺,虽然其能够理解上下文内容的语义,但是在分析和分辨句子中不同片段所充当的语义角色时,其难以分辨这些粒度较细的语义信息,即使在思维链的引导下其表现仍然不佳。我们认为这是由于论元角色识别任务更倾向于考验大模型对语义的理解能力,但思维链通常用于强化大语言模型的推理能力,而这种推理能力并不能很好地作用于语义理解能力的增强。此外,思维链的引入在一定程度上增加了模型需要理解和生成的上下文长度,这也使得一些参数量较小的模型在输出中出现不可控的情况。

4 结论

尽管大语言模型在许多自然语言处理领域的任务上取得了令人印象深刻的效果,但经过我们的评估,与传统模型相比,大语言模型在框架语义分析任务上仍存在不足。这种不足主要体现在:(1)语义分析任务上解析能力严重不足;(2)无法稳定利用额外提示信息提升解决此类问题的能力;(3)在思维链的引导下仍然难以激发出其框架语义分析能力。此外,我们也通过实验发现了大语言模型温度系数设定对其框架语义分析能力的影响,并分析了温度系数设定与示例样本量之间存在的关联。这些评估结果为我们今后的研究提供了新的方向,一方面,我们将在今后的工作中有针对性地改进大语言模型存在的问题,使其能够在框架语义分析任务中展现出更具竞争力的表现;另一方面,现阶段大语言模型的不足表明我们继续基于传统模型来解决框架语义分析任务仍然是非常有意义的。

参考文献:

[1] TOUVRON H, LAVRIL T, IZACARD G, *et al*. LLaMA: Open and Efficient Foundation Language Models[EB/OL]. arXiv Preprint: 2302.13971, 2023. https://arxiv.org/abs/2302.13971.

[2] TOUVRON H, MARTIN L, STONE K, *et al*. Llama 2: Open Foundation and Fine-Tuned Chat Models[EB/OL]. arXiv Preprint: 2307.09288, 2023. https://arxiv. org/abs/2307.09288.

[3] DU Z, QIAN Y, LIU X, *et al*. GLM: General Language Model Pretraining with Autoregressive Blank Infilling [C]//Proceedings of the 60th Annual Meeting of the Asso-

ciation for Computational Linguistics (Volume 1: Long Papers). Stroudsbarg, PA: ACL, 2022: 320−335. DOI: 10.18653/v1/2022.acl-long.26.

[4] RADFORD A, NARASIMHAN K, SALIMANS T, *et al*. Improving Language Understanding by Generative Pre-Training[EB/OL]. (2018−06−09) [2024−02−15]. https://cdn.openai.com/research−covers/language−unsupervised/language_understanding_paper.pdf.

[5] RADFORD A, WU J, CHILD R, *et al*. Language Models are Unsupervised Multitask Learners[EB/OL]. (2019−02−15) [2024−03−10]. https://cdn. openai. com/better−language−models/language_models_are_unsupervised_multitask_learners.pdf.

[6] BROWN T, MANN B, RYDER N, *et al*. Language Models are Few-Shot Learners[J]. *Adv Neural Inform Process Syst*, 2020, **33**: 1877−1901.

[7] WADHWA S, AMIR S, WALLACE B C. Revisiting Relation Extraction in the era of Large Language Models [C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Stroudsburg, PA: ACL, 2023: 15566−15589. DOI: 10.18653/v1/2023.acl-long.868.

[8] LUO Z, XIE Q, ANANIADOU S. ChatGPT as a Factual Inconsistency Evaluator for Text Summarization[EB/OL]. arXiv Preprint: 2303.15621, 2023. https://arxiv.org/abs/2303.15621.

[9] WEI J, WANG X, SCHUURMANS D, *et al*. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models[J]. *Adv Neural Inform Process Syst*, 2022, **35**: 24824−24837.

[10] BANG Y, CAHYAWIJAYA S, LEE N, *et al*. A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity[C]//Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia−Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers). Romania: AACL, 2023: 675−718. DOI: 10.18653/v1/2023.ijcnlp-main.45.

[11] BIAN N, HAN X, SUN L, *et al*. ChatGPT is a Knowledgeable but Inexperienced Solver: An Investigation of Commonsense Problem in Large Language Models[EB/OL]. arXiv Preprint: 2303.16421, 2024. https://arxiv.org/abs/2303.16421.

[12] GAO J, ZHAO H, YU C, *et al*. Exploring the Feasibility of ChatGPT for Event Extraction[EB/OL]. arXiv Preprint: 2303.03836, 2023. https://arxiv. org/abs/2303.03836.

[13] WEI X, CUI X, CHENG N, *et al*. Zero-Shot Information Extraction via Chatting with ChatGPT[EB/OL]. arXiv Preprint: 2302.10205, 2023. https://arxiv.org/abs/2302.10205.

[14] LI B, FANG G, YANG Y, *et al*. Evaluating ChatGPT's Information Extraction Capabilities: An Assessment of Performance, Explainability, Calibration, and Faithfulness[EB/OL]. arXiv Preprint: 2304.11633, 2023. https://arxiv.org/abs/2304.11633.

[15] YUAN C, XIE Q, ANANIADOU S. Zero-shot Temporal Relation Extraction with ChatGPT[C]//The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks. Stroudsbarg, PA: ACL, 2023: 92−102. DOI: 10.18653/v1/2023.bionlp-1.7.

[16] TAO Z, JIN Z, BAI X, *et al*. EvEval: A Comprehensive Evaluation of Event Semantics for Large Language Models[EB/OL]. arXiv Preprint: 2305.15268, 2023. https://arxiv.org/abs/2305.15268.

[17] XIE T, LI Q, ZHANG J, *et al*. Empirical Study of Zero-Shot NER with ChatGPT[C]//Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Stroudsbarg, PA: ACL, 2023: 7935−7956. DOI: 10.18653/v1/2023.emnlp-main.493.

[18] LI J, YAN Z, SU X, *et al*. Overview of CCL23-Eval Task 3: Chinese FrameNet Semantic Parsing[C]//Proceedings of the 22nd Chinese National Conference on Computational Linguistics (Volume 3: Evaluations). Beijing: CIPS, 2023: 113−123.

[19] FILLMORE C J. Frame Semantics[M]//GEERAERTS D. Cognitive Linguistics: Basic Readings. Berlin, New York: De Gruyter Mouton, 2006: 373−400. DOI: 10.1515/9783110199901.373.

[20] 闫智超, 李茹, 苏雪峰, 等. 融合目标词上下文序列与结构信息的框架识别方法[J]. 中文信息学报, 2024, **38**(1): 86−96.
YAN Z C, LI R, SU X F, *et al*. Integrating Contextual and Structural Information of Target Words for Frame Identification[J]. *J Chin Inf Process*, 2024, **38**(1): 86−96.

[21] SU X F, LI R, LI X L, *et al*. A Span-based Target-aware Relation Model for Frame-semantic Parsing[J]. *ACM Trans Asian Low-Resour Lang Inf Process*, 2023, **22**(3): 1−24. DOI: 10.1145/3569581.

[22] YAN Z C, SU X F, CHAI Q H, *et al*. Multiple POS Dependency-aware Mixture of Experts for Frame Identification[J]. *IEEE Access*, 2023, **11**: 25604−25615. DOI: 10.1109/ACCESS.2023.3253128.

[23] GUO S, GUAN Y, LI R, *et al*. Incorporating Syntax and Frame Semantics in Neural Network for Machine Read-

ing Comprehension[C]//Proceedings of the 28th International Conference on Computational Linguistics. America: ICCL, 2020: 2635−2641. DOI: 10.18653/v1/2020.coling-main.237.

[24] GUO S, LI R, TAN H, *et al*. A Frame-based Sentence Representation for Machine Reading Comprehension [C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsbarg, PA: ACL, 2020: 891−896. DOI: 10.18653/v1/2020.acl-main.83

[25] 王智强, 李茹, 梁吉业, 等. 基于汉语篇章框架语义分析的阅读理解问答研究[J]. 计算机学报, 2016, **39**(4): 795−807. DOI: 10.11897/SP.J.1016.2016.00795.
WANG Z Q, LI R, LIANG J Y, *et al*. Research on Question Answering for Reading Comprehension Based on Chinese Discourse Frame Semantic Parsing[J]. *Chin J Comput*, 2016, **39**(4): 795−807. DOI: 10.11897/SP.J.1016.2016.00795.

[26] GUAN Y, GUO S, LI R, *et al*. Frame Semantic-Enhanced Sentence Modeling for Sentence-level Extractive Text Summarization[C] //Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Stroudsbarg, PA: ACL, 2021: 4045−4052. DOI: 10.18653/v1/2021.emnlp-main.331.

[27] GUAN Y, GUO S, LI R, *et al*. Integrating Semantic Scenario and Word Relations for Abstractive Sentence Summarization[C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Stroudsbarg, PA: ACL, 2021: 2522−2529. DOI: 10.18653/v1/2021.emnlp-main.196.

[28] ZHAO H Y, LI R, LI X L, *et al*. CFSRE: Context-aware Based on Frame-semantics for Distantly Supervised Relation Extraction[J]. *Knowl Based Syst*, 2020, **210**: 106480. DOI: 10.1016/j.knosys.2020.106480.

[29] 谭红叶, 闫真, 李茹, 等. 迈向创造性语言生成: 汉语幽默自动生成的探索[J]. 中国科学: 信息科学, 2018, **48**(11): 1497−1509. DOI: 10.1360/N112018-00158.
TAN H Y, YAN Z, LI R, *et al*. Towards Creative Language Generation: Exploring Chinese Humor Generation[J]. *Sci Sin Informationis*, 2018, **48**(11): 1497−1509. DOI: 10.1360/N112018-00158.

[30] 郝晓燕, 刘伟, 李茹, 等. 汉语框架语义知识库及软件描述体系[J]. 中文信息学报, 2007, **21**(5): 96−100. DOI: 10.3969/j.issn.1003-0077.2007.05.018.
HAO X Y, LIU W, LI R, *et al*. Description Systems of the Chinese FrameNet Database and Software Tools[J]. *J Chin Inf Process*, 2007, **21**(5): 96−100. DOI: 10.3969/j.issn.1003-0077.2007.05.018.

[31] 刘开瑛. 汉语框架语义网构建及其应用技术研究[J]. 中文信息学报, 2011, **25**(6): 46−52. DOI: 10.3969/j.issn.1003-0077.2011.06.007.
LIU K Y. Research on Chinese FrameNet Construction and Application Technologies[J]. *J Chin Inf Process*, 2011, **25**(6): 46−52. DOI: 10.3969/j.issn.1003-0077.2011.06.007.

[32] WANG S, SUN X, LI X, *et al*. GPT-NER: Named Entity Recognition *via* Large Language Models[EB/OL]. arXiv Preprint: 2304.10428, 2023. https://arxiv.org/abs/2304.10428.

[33] OUYANG L, WU J, JIANG X, *et al*. Training Language Models to Follow Instructions with Human Feedback[J]. *Adv Neural Inform Process Syst*, 2022, **35**: 27730−27744.

[34] ANIL R, BORGEAUD S, ALAYRAC J, *et al*. Gemini: A Family of Highly Capable Multimodal Models[EB/OL]. arXiv Preprint: 2312.11805, 2024. https://arxiv.org/abs/2312.11805.

[35] LI Z, GUO X, QIAO D, *et al*. System Report for CCL23-Eval Task 3: Application of Entity Classification Model Based on Rotary Position Embedding in Chinese Frame Semantic Parsing[C]//Proceedings of the 22nd Chinese National Conference on Computational Linguistics (Volume 3: Evaluations). Beijing: CIPS, 2023: 94−104.

[36] DEVLIN J, CHANG M W, LEE K, *et al*. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Stroudsbarg, PA: ACL, 2019: 4171−4186. DOI: 10.18653/v1/N19-1423.