基于生成式语言模型的立场检测探究

张袁硕[1,2], 李澳华[1,2], 尹召宁[1,2], 王潘怡[1], 陈波[1,2,3,*], 赵小兵[1,2,3,*]
1.中央民族大学,信息工程学院,北京,100081
2.国家语言资源监测与研究少数民族语言中心
3.国家安全研究院语言信息安全研究中心
chenbomuc@muc.edu.cn

摘要

近年来，立场检测任务受到越来越多的关注，但相关标注数据在范围和规模上都有限，不能有效支撑基于神经网络的立场检测。为此，本文探索在零样本/少样本场景下生成式语言模型在立场检测任务上的能力。首先，构建了一个全新的面向立场检测的数据集，包含5个主题，共2500个人工标注样例；然后，在此数据集上进行了一系列探索实验，实验结果表明：生成式语言模型在零样本设定下，采用结构化的提示学习表现良好；增加额外信息能够显著提升模型性能；在少样本设定下，提供相同目标的示例能够明显提升模型性能，而不同目标示例产生了负面作用；使用思维链可以显著提升模型性能；受提示学习的启发，微调预训练语言模型进一步论证提供额外信息对立场检测的增益显著。

关键词： 生成式语言模型 ； 立场检测 ； 零样本/少样本 ； 提示学习

Research on Stance Detection with Generative Language Model

Yuanshuo Zhang[1,2], Aohua Li[1,2], Zhaoning Yin[1,2], Panyi Wang[1,2]
Bo Chen[1,2,3,*], Xiaobing Zhao[1,2,3,*]
1. School of Information Engineering, Minzu University of China
Beijing 100081, China
2. National Language Resource Monitoring and Research Center of Minority Languages
3. Language Information Security Research Center Institute of National Security MUC
chenbomuc@muc.edu.cn

Abstract

In recent years, stance detection has received increasing attention, but the relevant annotated data is limited in scope and scale, which cannot effectively support neural network-based stance detection methods. Therefore, this paper explores the ability of generative language models in stance detection in zero/few shot settings.Firstly, a new stance detection dataset is constructed, which contains 5 topics and a total of 2500 manually labelled samples; then, a series of exploratory experiments are carried out on this dataset, and the experimental results show that: the generative language model performs well in structured prompt learning under zero-shot setting; fusion of extra information can significantly improve model performance; under the setting of few-shot, providing examples with the same goal can significantly improve model performance, while examples with different goals have a negative impact; Chain-of-Thought can significantly improve model performance; Inspired by prompt learning experiments, this paper fine tunes the pre-trained language model to further demonstrate that providing additional information has a significant gain in stance detection.

Keywords: Generative Language Model , Stance Detection , Zero/Few Shot , Prompt Learning

(c) 2024 中国中文信息学会计算语言学专业委员会

1 引言

随着社交媒体的普及和信息时代的到来，人们在各种在线平台上表达对热门话题的不同观点的意愿上升，因此，产生了大量的用户生成内容。基于这些生成内容，相关组织或机构可以准确掌握公众对特定话题的态度和立场，为相关决策提供依据。面向此类应用情境的立场检测 (李洋et al., 2021)任务也因此吸引了广泛的关注，该任务目标是给定文本，判定其作者对特定话题、观点或事件的立场或态度（支持、反对或者中立）。例如：针对"深圳禁摩限电"这一政策法规，用户发表微博"广州的也给全部禁了吧，特别是摩托车，容易出事！" (Xu et al., 2016b)，从"全部禁了"、"出事"这类词语表达用户对该目标的支持立场。

立场检测是自然语言处理领域的一个重要研究任务，一般将其建模为分类问题。目前，立场检测主流方法是深度学习 (Igarashi et al., 2016)的方法。尽管现有的立场检测模型通过有监督学习可以在一些常见的任务设定，如单目标 (Mohammad et al., 2016)、多目标 (Sobhani et al., 2017)等设定上表现出色，但在少样本（Few-shot）和零样本（Zero-shot） (Allaway and McKeown, 2020)设定下，即模型在没有或只有少量针对特定话题的标注数据的情况下，其性能显得不尽人意。

随着生成式语言模型 (Radford and Narasimhan, 2018)的兴起，如GPT-3，因其在预训练阶段通过大规模无监督数据的学习，使其掌握了丰富的语言知识，具有强大的语言理解和生成能力，同时也具备很强的任务泛化能力，在新任务上表现出色。在零样本设定下采用合适的提示就能取得不错的效果 (Wu et al., 2021)，在少样本设定下采用示例或思维链（Chain of Thought，COT）的提示方式能进一步提升效果 (Wei et al., 2022)。Floridi等人(2020)讨论了GPT-3的性质、范围、限制和影响，并探讨了其在数学、语义（图灵测试）和伦理问题上的能力；Ye等人(2023)使用21个数据集评估了GPT-3和GPT-3.5在九个自然语言理解任务上比较了零样本和少样本场景下每个任务的表现。

在本文中，我们针对立场检测任务，探索基于生成式语言模型的立场判别方法。我们先构建了一个新的人工标注的立场检测数据集，然后在此数据集上探究生成式语言模型的立场判别能力，特别是在零样本和少样本设定下，并尝试回答4个问题：Q1：生成式语言模型在零样本立场检测问题中的效果如何？Q2：补充额外的信息能否提升模型的表现？Q3：在少样本设定下，示例能否提升模型的表现？Q4：思维链能否提升模型的表现？

本文贡献如下：

- 针对立场检测任务，我们构建了一个新的数据集，该数据集包含来自微博的5个热点话题，每个话题收集并人工标注了500条用户生成内容，可作为中文立场检测单目标、多目标、零样本和少样本等不同设定下的标准数据集。
- 我们从多个角度对生成式语言模型基于提示的立场判别进行了探索，尝试在提示中增加不同的额外知识信息，并采用多样化的方式，以实验出面向生成式语言模型最有效的提示方法。
- 我们对预训练语言模型进行微调，进一步证实了明确的立场标签和事件背景信息对立场检测有重要的影响。

2 相关工作

2.1 立场检测

近年来，立场检测受到广泛的关注，并得到了应用，如舆情分析 (李洋et al., 2021)、社交媒体监测 (刘高勇et al., 2022)和政治舆论分析 (Thomas et al., 2006)等。立场检测方法的发展经历了基于传统机器学习的方法，到基于神经网络的方法 (Igarashi et al., 2016)，再到基于大规模预训练语言模型的方法的演变。基于传统机器学习的方法，主要依赖于手工定义特征，常用特征包括：文本特征、情感特征和混合特征 (奠雨洁et al., 2017)等，再利用分类模型进行立场分类，如支持向量机 (Xu et al., 2016a)。神经网络方法兴起后，在立场检测任务中，为了减少了人工构造特征的工作量，神经网络被引入来自动学习更丰富的特征。Mohtarami等人(2018)提

©2024 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

出了基于端到端的记忆网络模型。杨顺成等人(2020)提出了基于GCN和Bi-LSTM的微博立场检测方法。面向零样本/少样本情境，预训练语言模型可以从大规模的无标注语料中学习到丰富且高层次的语言知识，能很好地作用于下游任务，并具有一定的泛化能力。Hosseinia等人(2020)提出了一种基于BERT的问题表示与情绪信息相结合的立场检测的方法。Allaway等人(2020)基于BERT和Attention机制开发了Topic-Grouped Attention(TGA)模型，通过上下文聚类获得的广义主题表示来利用有关主题相似性的信息。He等人(2022)利用维基百科中关于目标的背景知识来增强立场检测，提出了具有两种变体的WS-BERT来编码此类知识。Li等人(2023a)提出了知识增强立场检测框架(KASD)，在社交媒体立场检测任务中引入了情景知识和话语知识，并利用ChatGPT对上述两种知识进行提取和注入，性能显著提高。Li等人(2023b)提出了目标立场提取（Target-Stance Extraction, TSE）任务，这种新任务有助于促进立场检测领域的未来研究。

2.2 生成式语言模型的零样本和少样本学习能力探究

生成式语言模型拥有优秀的任务泛化能力，可直接应用于各种新任务。在零样本方面,Kojima等人(2022)提出了一种基于零样本学习的思维链提示，在各种推理任务上的表现明显优于零样本大语言模型的性能；李燕等(2023)将ChatGPT作为辅助工具，利用提示和模型先验知识助力银行内部的渗透测试；卢宇等(2023)探究了ChatGPT在教育方面的题目生成、自动结题和辅助批阅的能力；寿建琪(2023)使用GPT助力信息检索，提出了一种结合OPAC类检索服务和基于LLM的智能化信息检索服务的自适应文献检索框架。在少样本方面，Brown等(2020)通过预训练大型语言模型并使用少量样本进行微调，在多种NLP任务上都表现出色，例如翻译、问答和完形填空任务，展示了GPT-3强大的少样本学习能力。冯广敬等人(2022)提出了基于DialoGPT的二阶段对话生成模型；马志强等人(2023)基于GPT-2构建情感导向对话回复生成实验系统EGRGM；言佳润等人(2023)在GPT、BERT、RoBERTa预训练模型上进行提示学习，通过P-tuning自动构建连续模板进行论辩挖掘。Tuckute等人(2024)基于GPT的编码模型来预测大脑对多样化句子的反应，并利用这些模型选择的句子在新个体中记录大脑反应验证模型的预测能力。Wachowiak等人(2023)探究了GPT-3 检测隐喻语言和预测隐喻源域的能力。Black等人(2022)介绍了一个在Pile 上训练的200亿参数自回归语言模型GPT-NeoX-20B，一个特别强大的少样本推理器。我们首次将生成式语言模型应用于立场检测，并通过系列对比实验探索最有效的提示方法。

3 生成式立场检测

为了探究生成式语言模型在立场检测任务上的效果，特别是在零样本和少样本的设定。我们设计了一系列实验，旨在回答Q1-Q4。

实验1：零样本设定下，使用常规提示和结构化提示，旨在探索最适合立场检测任务的提示，并分析在零样本设定下生成式语言模型立场检测的性能。（跟Q1有关）

实验2：在提示中加入额外的信息，包括：事件相关的背景和明确的立场标签。例如，针对社会热点事件“恶意殴打他人者妻女被网暴”，有关事件的背景知识：“唐山男子烧烤店无缘无故殴打女子，事后男子妻女遭到网暴。”和明确的立场标签：“活该被网暴（支持）、不该被网暴（反对）、中立。”实验2旨在探究通过补充额外的信息能否提升生成式语言模型在立场检测任务上的性能。（跟Q2有关）

实验3：少样本设定下，在提示中加入立场检测示例。实验3主要考察生成式语言模型在给定相同或者不同目标的示例后，能否提升立场判别的准确度。（跟Q3有关）

实验4：在提示信息中引入思维链，旨在验证引入推理过程后，能否提升立场判别的准确度。（跟Q4有关）

4 实验及分析

为了探索生成式语言模型在立场检测任务上的效果，我们先构建了一个新的立场检测数据集，接着选择代表性的生成式语言模型进行实验1-4，最后对实验结果进行详细分析。我们使用F1值作为立场检测的评价指标。对于没有给出明确立场的回答，我们将其视为判断错误。

(c) 2024 中国中文信息学会计算语言学专业委员会

4.1 微博立场检测数据集

我们从新浪微博平台选择了近期发生的5个具有争议的话题，包括“恶意殴打他人者妻女被网暴”、“女子不让6岁男童上女厕所遭痛骂”、“警方通告胡鑫宇为自杀”、“满江红起诉大V”和“泼水节女生选择原谅对方”，并将这5个话题作为立场检测的目标；然后，我们检索了5个话题的背景知识并爬取了对应评论，并利用启发式规则从目标中提取了明确的立场标签；最后，我们对评论进行立场标注，在标注过程中，采用多人交叉标注的方式，以确保数据标注的客观性和准确性。具体而言，三位标注员（均为NLP领域硕士研究生）对每一条评论文本根据目标话题进行独立的立场判断和标注。收集了三份独立的标注数据后，进行比对，若某条评论文本得到两人以上标注的结果相同，则该立场标注被视为正确标注。如果三位标注员对某条评论文本的立场标注存在不一致，即标注了三个不同的立场，那么该评论文本将被排除在数据集之外。

我们数据集的优势主要有两点：1. 数据集更具实用性，数据集全部针对微博平台的热点舆情事件，在该数据集上的研究能够快速对接到真实的微博平台立场分析的应用场景；2.数据集更具有挑战性，数据集中的评论针对的是一个复杂事件，包含多个子事件，如“恶意暴打他人者妻女被网暴”，包含了男子殴打女子，男子的妻女被网友网暴，要对评论给出准确的立场判定，需要有足够的事件知识信息和严谨的推理。

我们将该数据集命名为Weibo-SD并公开共享于github[0]。该数据集共有5个目标，每个目标有500个评论，每个目标下的背景、明确的立场标签和评论的立场分布如表1所示。

主题	明确的立场标签	背景知识	支持	反对	中立
恶意殴打他人者妻女被网暴	活该被网暴(支持),不该被网暴(反对),中立	唐山男子烧烤店无缘无故殴打女子，事后男子妻女遭到网暴。	260	62	178
女子不让6岁男童上女厕所遭痛骂	活该被骂(支持),不该被骂(反对),中立	女子不让6岁男童上女厕所，遭男童妈妈当众辱骂。	34	338	128
警方通告胡鑫宇为自杀	认为是自杀(支持),怀疑不是自杀(反对),中立	高一男生胡鑫宇傍晚从宿舍至教学楼路上离奇失踪，多点暗示胡鑫宇有自杀倾向。最终胡鑫宇在学校后山被找到，警方通报其为自缢身亡。	127	226	247
满江红起诉大V	应该起诉(支持),不该起诉(反对),中立	2023年春节电影《满江红》被指控“幽灵场”“抄袭”等。随后满江红官方发文，对造谣者沈逸、屠龙的胭脂井、平原公子赵胜、喵斯拉大王二号机四位大V提出诉讼。	61	195	224
泼水节女生选择原谅对方	应该原谅(支持),不该原谅(反对),中立	泼水节女子遭众男子围着泼水撕雨衣，事后女子选择原谅对方，不立案。	147	176	177

Table 1: Weibo-SD数据集

4.2 生成式语言模型

针对要探究的问题，我们选择了2个生成式语言模型（在本文中生成式语言模型与大语言模型（Large Language Model）两种表述是通用的，并采用LLM简称），分别是GPT-3.5-turbo和ChatGLM2-6B，分别简称为GPT和GLM。所有实验均采用API调用的方式完成，每个结果都是5次实验结果的平均值。

4.3 实验结果及分析

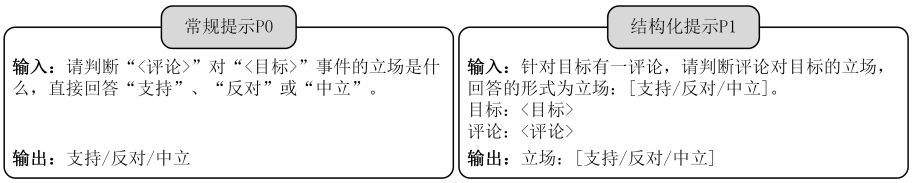

Figure 1: 提示模板P0、P1

[0] https://anonymous.4open.science/r/Weibo-SD-618D/

(c) 2024 中国中文信息学会计算语言学专业委员会

提示	GPT	GLM
P0	35.28	35.12
P1	**42.68**	**47.52**

Table 2: 在零样本设定下使用两种提示进行提问(单位%)

针对实验1：在零样本设定下，采用常规提示P0和结构化提示P1两种提示模板，提示模板如图1所示，立场检测的结果如表2所示。可以看到，使用常规提示时，GPT和GLM的F1值均在35.20左右。使用结构化提示时，GPT的F1从35.28提升到42.68，提升了7.40；GLM的F1从35.12提升到47.52，提升了12.40。GPT和GLM的效果均有显著提升，且GLM的效果优于GPT4.84个百分点。我们认为使用结构化提示较常规提示能显著提升LLM在立场检测任务上的性能。结构化提示可以引导模型关注于文本中的关键信息，帮助模型更好地理解任务需求，从而提高立场判别的准确性。因此，在后续的实验中我们均使用结构化提示进行提问。GLM（ChatGLM2-6B）发布于2023年，预训练语料中可能已经包含2022年的数据，而用于预训练GPT的数据是基于2021年之前的数据，因此在零样本且不提供额外信息的情况下，GLM的效果要优于GPT。

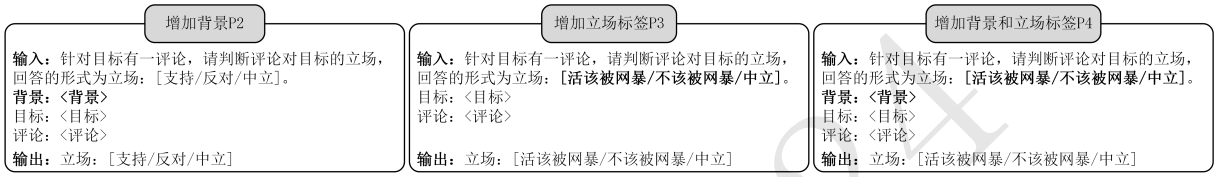

Figure 2: 提示模板P2、P3、P4

加入额外信息	提示	GPT	GLM
无额外信息	P1	42.68	**47.52**
背景	P2	44.20	47.04
立场标签	P3	50.32	40.52
背景+立场标签	P4	**51.44**	40.48

Table 3: 在提示中增加背景和明确的立场标签(单位%)

针对实验2：我们尝试在提示中增加了额外的信息，包括背景知识和明确的立场标签。提示模板如图2所示，实验结果表3所示。

对于GPT，在提升中增加额外信息对立场判别效果提升显著。具体来说，相比无额外信息的提示模板P1，增加了背景信息后F1值为44.20，提升了1.52；增加了明确立场标签后F1值为50.32，提升了7.64，提升显著；同时增加了背景和明确立场标签后F1值达到51.44，提升了8.76，取得了最好的效果。我们认为提供额外的背景知识能够帮助模型更深入地理解文本的上下文，从而提高立场检测的准确性；我们的立场检测数据集包含的都是复杂的话题，话题本身可能包含2个事件，如“恶意殴打他人者妻女被网暴”中就包含“某男子恶意殴打他人”（事件A）和“该男子的妻女被网暴”（事件B）两个事件，简单的立场标签，如“支持”、“反对”和“中立”，在零样本或少样本的情况下，很难让LLM学习到支持的立场到底是支持事件A还是B，特别是事件A和B往往是对立的情况下。明确的立场标签可以告诉模型支持的目标是什么，反对目标是什么，从而消除模型的混乱，提高立场检测性能。例如对于“恶意殴打他人者妻女被网暴”事件，支持的立场等同于“活该被网暴”，而不是支持打人者或者打人者的妻女。

对于GLM，增加了背景信息后F1有小幅下降，下降了0.48；增加了明确立场标签后F1大幅下降，从47.52下降到40.52，下降了7.00；同时增加了背景和明确立场标签和只增加明确立场标签F1值接近，下降了7.04。我们认为GLM的训练语料已经包含这些事件相关数据，因此增加背景信息对GLM影响不大；观察输出结果，我们发现GLM会因明确立场标签中“网暴”、“被骂”等敏感词汇产生误判，认为“网暴”、“骂人”是错误的行为，将立场判定为“不该网暴”、“不

(c) 2024 中国中文信息学会计算语言学专业委员会

该被骂”或保持中立。我们统计了去除包含敏感词汇立场后的F1值，分别为45.19和45.13，证实了我们的猜想是正确的。

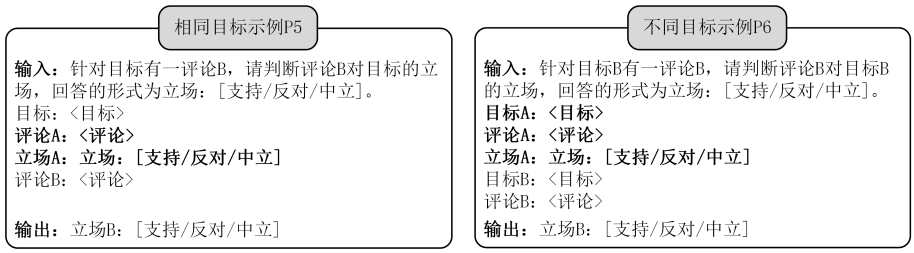

Figure 3: 提示模板P5、P6

示例	提示	GPT	GLM
无示例	P1	42.68	**47.52**
相同目标示例	P5	**47.92**	43.12
不同目标示例	P6	42.56	42.58

Table 4: 在少样本设定下在提示中添加示例(单位%)

针对实验3：探究少样本设定下，在提示中加入示例对立场判别的影响。提示模板如图3所示，结果如表4所示。

对于GPT，添加相同目标示例后F1有明显提升，从42.68提升到了47.92，提升了5.24；添加不同目标示例后F1小幅下降，从42.68下降到了42.56，下降了0.12。我们认为提供相同目标的示例能使模型更准确地捕捉到与指定目标特定立场相关的词汇、短语和句子结构，从而提高检测的准确性，而不同目标的示例可能会带来干扰，导致模型性能下降。

对于GLM，添加相同目标示例或不同目标示例后，模型的性能均有明显降低，平均降低了4.67。观察输出结果后，我们发现GLM有时会出现将示例评论与待检测评论混淆或将示例目标与待检测评论目标混淆的问题，例如忽略了评论B而回答示例评论A的立场，或将示例目标A当作待检测目标。我们认为加入示例会使提示复杂化，GLM不能完全理解示例的意图，导致性能下降。

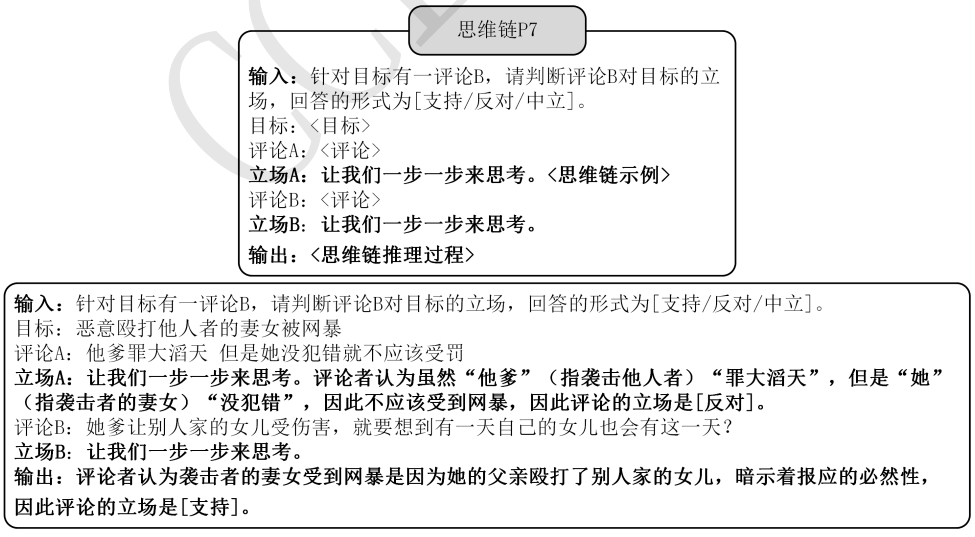

Figure 4: 提示模板P7及示例

针对实验4：探究引入思维链后，能否提升立场判别的准确度。为了更直观的展示思维链提示的过程，我们提供了一个示例。示例和提示模板如图4所示，结果如表5所示。

提问方式	提示	GPT	GLM
无思维链	P1	42.68	**47.52**
思维链	P7	**51.36**	28.84

Table 5: 在提示中添加思维链(单位%)

对于GPT，引入思维链后F1值显著提升，从42.68提升到了51.36，提升了8.68。我们认为，相比于直接给出评论的立场，思维链提供了中间的推导过程，减少了跳跃性推理，使模型不只关注于输入评论的表面特征，还能够捕捉到评论背后的复杂关系，如评论中的人物关系、与人物相关的信息等。例如评论“她爹罪大滔天，但她没有犯错就不该受到惩罚”，评论者认为虽然“他爹”（指袭击他人者）“罪大滔天”，但是“她”（指袭击者的妻女）“没犯错”，因此不应该受到惩罚，这表明评论者对袭击者的妻女感到同情并认为她们不应该因为袭击者的行为而受到处罚，因此评论的立场是[反对]。

对于GLM，引入思维链后F1值大幅降低，降低了18.68。观察输出结果，发现有近20%的输出答非所问，不能输出明确的立场。我们认为原因有两点，一是GLM的参数规模只有60亿,本身的上下文理解能力和推理能力有限；二是思维链的引入可能会使模型在处理问题时过度复杂化，使本就能力受限的模型更难以推理。

4.3.1 立场判别偏向性

我们对LLM的立场判别的偏向性进行了评估。实验结果如表6所示。可以发现，GPT对于中立立场的判别准确度整体很高，基于提示P1的准确度达到了最高的83.52，平均值也高达71.61，而对支持判定的准确度较低，平均值分别为27.92。在加入了明确立场标签的P3、P4和引入了思维链的P7中，GPT对反对立场的判别明显高于其他提示模板，分别为49.65、50.65、40.02，远高于平均值。我们认为GPT对中立立场的偏向性较高，当它不能确定评论的立场时，可能会采取保守的策略而选择中立，明确的立场标签和思维链都能明显提高模型的推理能力。而GLM对反对立场的准确度更高，基于P6的准确度达到了最高的83.55，平均值为49.06，但对于支持和中立的准确度较低，分别为33.09和36.27。我们认为GLM对一些中文词汇敏感度较高，当发现提示中出现“网暴”之类词汇时，倾向于将立场判定为反对，因此对反对的偏向性较高。另外，LLM整体对支持立场的判别的准确度不高，平均值在35%左右，但增加了额外信息或引入思维链后，这一现象有所改善。

提示	支持	反对	中立
P0	15.42/38.63	5.52/54.36	83.52/10.64
P1	36.72/19.08	15.35/80.44	78.15/30.43
P2	41.65/22.26	20.96/70.61	72.54/37.99
P3	32.75/42.45	49.65/20.16	63.73/62.36
P4	33.70/42.29	50.65/20.16	65.10/62.36
P5	53.42/46.10	26.28/39.92	68.65/44.62
P6	43.24/14.79	14.94/83.55	73.57/13.96
P7	46.74/39.11	40.02/23.27	67.62/27.80
AVG	37.96/33.09	27.92/49.06	71.61/36.27

Table 6: 不同提示下检测不同立场的准确度(GPT/GLM(单位%))

4.3.2 主题的差异性

我们评估了LLM在各个主题上立场判别差异性。结果如表7所示。可以发现，“警方通告胡鑫宇为自杀、“满江红起诉大V”和“泼水节女生选择原谅对方”这三个主题的F1值整体较高，均在45%左右；而对于“恶意殴打他人者妻女被网暴”和“女子不让6岁男童上女厕所遭痛骂”主题的F1值较低。我们分析其原因有二：一是后三个主题的事件逻辑相对简单，只涉及到“胡鑫宇是否是自杀”、“满江红是否该起诉大V”以及“女生是否该原谅对方”，而其前两个主题基本都包含两个事件，如看似只包含1个事件的“女子不让6岁男童上女厕所遭痛骂”也包含“女子不让6岁

(c) 2024 中国中文信息学会计算语言学专业委员会

男童上女厕所”和“女子遭痛骂”这两个明显立场对立的事件；二是后三个主题下的用户评价多数立场鲜明，如“警方通告胡鑫宇为自杀”这个主题下的评论“就我而言，失踪106天截至到山上发现，且是自缢，这期间遇害地点和学校如此之近，警方进行地毯式搜索无果。现在突然间找到，没有自杀一说，我直接拍板，就是他杀。”这都有利于LLM对立场进行相对准确的判别。

提示	恶意殴打他人者妻女被网暴	女子不让6岁男童上女厕所遭痛骂	警方通告胡鑫宇为自杀	满江红起诉大V	泼水节女生选择原谅对方
P0	40.29/33.10	23.96/33.47	32.39/39.71	48.45/37.40	37.04/37.72
P1	44.49/20.20	27.48/65.40	38.88/53.60	52.05/55.80	50.90/42.60
P2	45.60/23.80	25.85/57.80	54.80/52.60	47.20/52.40	47.74/48.60
P3	41.40/39.88	45.80/27.40	59.20/46.94	57.80/47.20	47.40/41.40
P4	43.20/39.92	46.80/27.40	60.80/46.94	57.00/47.05	49.40/41.40
P5	55.40/30.43	32.40/39.80	51.80/50.20	51.40/51.20	48.65/44.00
P6	49.20/15.07	28.20/64.84	37.31/49.33	52.00/45.89	46.29/40.76
P7	51.60/23.36	42.53/26.53	57.00/41.33	54.05/35.05	51.90/30.98
AVG	46.40/28.22	34.13/42.83	49.02/47.58	52.49/46.50	47.42/40.93

Table 7: 不同提示下五种主题的F1(GPT/GLM(单位%))

5 微调预训练语言模型

基于生成式语言模型的立场检测实验表明：1）明确的立场标签具有很大的影响；2）补充背景信息能增强立场检测。为了探究LLM与流行立场检测模型的性能差异并进一步论证上述结论，我们微调了预训练语言模型（BERT、RoBERTa和mT5），在基础的设置中只输入目标和评论（P1），对比设置中，分别加入事件背景（P2）、明确的立场标签（P3），还有组合加入的设置（P4）。我们将Weibo-SD数据集中的80%作为微调数据集，剩余的20%作为测试集。实验结果如表8所示。

输入	BERT	RoBERTa	mT5	GPT	GLM
目标和评论（P1）	63.6	61.4	60.0	42.68	**47.52**
+事件背景（P2）	64.8	62.2	60.2	44.20	47.04
+明确标签（P3）	65.4	62.4	60.8	50.32	40.52
+明确标签和事件背景（P4）	**66.6**	**65.4**	**62.6**	**51.44**	40.48

Table 8: 微调预训练语言模型F1(单位%)

我们发现:

(1)微调BERT和RoBERTa的结果总是优于微调mT5。我们认为，mT5在处理生成类任务时表现出色，但在分类任务上可能不如专注于理解文本深层含义的BERT和RoBERTa。

(2)当同时融合明确的立场标签和事件背景信息时，微调BERT取得了最好的效果，F1值达到66.6。我们认为，明确的立场标签和事件背景信息缓解了复杂的评论-目标-事件-立场关系所带来的混乱，从而提高立场检测的效果。

(3)LLM的立场检测能力能达到微调小规模语言模型的60%到80%，与在特定任务的数据集上进行训练的模型相比还有一定差距。

(4)事件背景可以提供足够的有用信息，同时减少噪声的引入。例如评论：“故意撕雨衣就过分了，必须立案追究责任”，其中“撕雨衣”、“立案”的信息在主题中并未提及，而是出现在事件背景中，这些信息表达了评论者对“原谅”持反对态度。

(5)明确的立场标签能增强立场检测。明确的立场标签使得模型在训练过程中能够聚焦于与立场相关的关键特征，例如针对主题“泼水节女生选择原谅对方”，明确的立场标签使模型聚焦于“应该或不该原谅”。

(c) 2024 中国中文信息学会计算语言学专业委员会

6 总结

本文探究了生成式语言模型在立场检测任务中的表现。实验结果表明，在零样本设定下，使用结构化的提示学习取得了更好的效果，尤其是在提供额外信息时。在少样本设定下，模型的性能也能得到不错的提升，尤其在引入思维链后，但当给的学习样本与测试样本不属于相同目标时，效果出现小幅下降。另外，受限于模型的规模大小，当提示过于复杂时，模型的理解能力反而变差。额外的实验和分析还表明：生成式语言模型存在一定的立场判别倾向性；事件逻辑越简单立场判别越容易。最后，我们通过微调预训练语言模型（BERT、RoBERTa和mT5），发现生成式语言模型在立场检测任务中展现出一定的潜力，并进一步证实了事件背景和明确的立场标签能增强立场检测。

本文的发现揭示了生成式语言模型在立场检测中的潜在价值，也强调了在特定情境下其局限性。在未来，我们将1）进一步扩充数据集，并且在运用大模型做立场检测时，预先将数据中的敏感词处理成等同语义的非敏感词，以减少此因素对大模型的干扰；2）研究探索更有效的提示方式和更全面的信息融合及更好的信息融合机制，如简化提示，同时保留有效信息，或分步提示，以提升模型对复杂提示的理解能力，以进一步提升生成式语言模型在这一领域的应用价值。

致谢

感谢所有匿名审稿人的宝贵意见。本项研究成果受国家社会科学基金重大项目（22&ZD035）、国家语委重点项目（ZDI145-61）、中央民族大学项目(GRSCP202316, 2023QNYL22, 2024GJYY43)资助。

参考文献

Emily Allaway and Kathleen McKeown. 2020. Zero-Shot Stance Detection: A Dataset and Model using Generalized Topic Representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8913–8931.

Sidney Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, Usvsn Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. 2022. GPT-NeoX-20B: An open-source autoregressive language model. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 95–136.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *ArXiv*, abs/2005.14165.

Luciano Floridi and Massimo Chiriatti. 2020. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30(4):681–694.

Zihao He, Negar Mokhberian, and Kristina Lerman. 2022. Infusing knowledge from Wikipedia to enhance stance detection. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 71–77.

Marjan Hosseinia, Eduard Dragut, and Arjun Mukherjee. 2020. Stance prediction for contemporary issues: Data and experiments. In *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*, pages 32–40.

Yuki Igarashi, Hiroya Komatsu, Sosuke Kobayashi, Naoaki Okazaki, and Kentaro Inui. 2016. Tohoku at SemEval-2016 task 6: Feature-based model versus convolutional neural network for stance detection. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 401–407.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Proceedings of Neural Information Processing Systems*, volume 35, pages 22199–22213.

(c) 2024 中国中文信息学会计算语言学专业委员会

Ang Li, Bin Liang, Jingqian Zhao, Bowen Zhang, Min Yang, and Ruifeng Xu. 2023a. Stance detection on social media with background knowledge. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15703–15717.

Yingjie Li, Krishna Garg, and Cornelia Caragea. 2023b. A new direction in stance detection: Target-stance extraction in the wild. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10071–10085.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. SemEval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41.

Mitra Mohtarami, Ramy Baly, James Glass, Preslav Nakov, Lluís Màrquez, and Alessandro Moschitti. 2018. Automatic stance detection using end-to-end memory networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 767–776.

Alec Radford and Karthik Narasimhan. 2018. Improving language understanding by generative pre-training.

Parinaz Sobhani, Diana Inkpen, and Xiaodan Zhu. 2017. A dataset for multi-target stance detection. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 551–557.

Matt Thomas, Bo Pang, and Lillian Lee. 2006. Get out the vote: Determining support or opposition from congressional floor-debate transcripts. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 327–335.

Greta Tuckute, Aalok Sathe, Shashank Srikant, Maya Taliaferro, Mingye Wang, Martin Schrimpf, Kendrick Kay, and Evelina Fedorenko. 2024. Driving and suppressing the human language network using large language models. *Nature human behaviour*, 8(3):544–561.

Lennart Wachowiak and Dagmar Gromann. 2023. Does GPT-3 grasp metaphors? identifying metaphor mappings with generative language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1018–1032.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *CoRR*, abs/2201.11903.

Shaohua Wu, Xudong Zhao, Tong Yu, Rongguo Zhang, Chong Shen, Hongli Liu, Feng Li, Hong Zhu, Jiangang Luo, Liang Xu, and Xuanwei Zhang. 2021. Yuan 1.0: Large-scale pre-trained language model in zero-shot and few-shot learning. *arxiv:2110.04725[cs.CL,cs.AI]*.

Jiaming Xu, Suncong Zheng, Jing Shi, Yiqun Yao, and Bo Xu. 2016a. Ensemble of feature sets and classification methods for stance detection. In *Proceedings of Natural Language Understanding and Intelligent Applications*, pages 679–688.

Ruifeng Xu, Yu Zhou, Dongyin Wu, Lin Gui, Jiachen Du, and Yun Xue. 2016b. Overview of nlpcc shared task 4: Stance detection in chinese microblogs. In *Proceedings of the ICCPOL 2016*, pages 907–916.

Junjie Ye, Xuanting Chen, Nuo Xu, Can Zu, Zekai Shao, Shichun Liu, Yuhan Cui, Zeyang Zhou, Chao Gong, Yang Shen, Jie Zhou, Siming Chen, Tao Gui, Qi Zhang, and Xuanjing Huang. 2023. A comprehensive capability analysis of gpt-3 and gpt-3.5 series models. *arxiv:2303.10420[cs.CL]*.

冯广敬, 刘箴, 刘婷婷, 许根, 庄寅, 王媛怡, and 柴艳杰. 2022. 基于情感变量的二阶段对话生成模型. 中文信息学报, 36:102–111.

刘高勇, 黄靖钊, and 艾丹祥. 2022. 融合立场检测和主题挖掘的突发公共事件网络舆情演化研究. 广东工业大学学报, 39:32–40+48.

卢宇, 余京蕾, 陈鹏鹤, and 李沐云. 2023. 生成式人工智能的教育应用与展望——以chatgpt系统为例. 中国远程教育, 43:24–31+51.

奠雨洁, 金琴, and 吴慧敏. 2017. 基于多文本特征融合的中文微博的立场检测. 计算机工程与应用, 53:77–84.

寿建琪. 2023. 走向"已知之未知": gpt大语言模型助力实现以人为本的信息检索. 农业图书情报学报, 35:16–26.

(c) 2024 中国中文信息学会计算语言学专业委员会

李洋, 孙宇晴, and 景维鹏. 2021. 文本立场检测综述. 计算机研究与发展, 58:2538–2557.

李燕, 赖胜枢, and 蒋泽宁. 2023. Chatgpt在银行业信息安全建设的应用. 金融科技时代, 31:14–18.

杨顺成, 李彦, and 赵其峰. 2020. 基于gcn和bi-lstm的微博立场检测方法. 重庆理工大学学报*(*自然科学*)*, 34:167–173.

言佳润and 鲜于波. 2023. 面向中文网络对话文本的论辩挖掘——基于微调与提示学习的大模型算法. 中文信息学报, 37:139–148.

马志强, 王春喻, 贾文超, and 杜宝祥. 2023. 情感导向对话回复生成模型. 中文信息学报, 37:104–114.

CCL 2024