面向话题的讽刺识别：新任务、新数据和新方法

梁斌[1]，林子杰[1]，秦兵[2,3]，徐睿峰[1,3]*
[1]哈尔滨工业大学（深圳），计算机科学与技术学院，深圳，518055
[2]哈尔滨工业大学，社会计算与信息检索研究中心，哈尔滨，150006
[3]鹏城实验室，深圳，518055
bin.liang@stu.hit.edu.cn , lzjjeffery@163.com
qinb@ir.hit.edu.cn, xuruifeng@hit.edu.cn

摘要

现有的文本讽刺识别研究通常只停留在句子级别的讽刺表达分类，缺乏考虑讽刺对象对讽刺表达的影响。针对这一问题，本文提出一个新的面向话题的讽刺识别任务。该任务通过话题的引入，以话题作为讽刺对象，有助于更好地理解和建模讽刺表达。对应地，本文构建了一个新的面向话题的讽刺识别数据集。这个数据集包含了707个话题，以及对应的4871个话题-评论对组。在此基础上，基于提示学习和大规模预训练语言模型，提出了一种面向话题的讽刺表达提示学习模型。在本文构建的面向话题讽刺识别数据集上的实验结果表明，相比基线模型，本文所提出的面向话题的讽刺表达提示学习模型取得了更优的性能。同时，实验分析也表明本文提出的面向话题的讽刺识别任务相比传统的句子级讽刺识别任务更具挑战性。

关键词： 面向话题的讽刺识别 ； 讽刺识别 ； 提示学习

Topic-Oriented Sarcasm Detection: New Task, New Dataset and New Method

Bin Liang[1], Zijie Lin[1], Bing Qin[2,3], Ruifeng Xu[1,3]*
[1]School of Computer Science and Technology,
Harbin Institute of Technology, Shenzhen, 518055
[2]Research Center for Social Computing and Information Retrieval,
Harbin Institude of Technology, Harbin, 150006
[3] Peng Cheng Laboratory, Shenzhen, 518055
bin.liang@stu.hit.edu.cn , lzjjeffery@163.com
qinb@ir.hit.edu.cn, xuruifeng@hit.edu.cn

Abstract

Existing research on sarcasm detection generally attempts to recognize the sentence-level sarcastic expression from the context, which lacks of the consideration of the satirical object on the sarcastic expression. Therefore, This paper proposes a new topic-oriented sarcasm detection task, which can better understand and model the sarcastic expression by introducing the topics as the satirical objects. Correspondingly, this paper constructs a new dataset for topic-oriented sarcasm detection, which consists of 707 topics and 4871 topic-comment pairs. Based on this dataset, a topic-based prompt learning model is proposed to deal with the topic-oriented sarcasm detection based on the large-scale pre-trained language model and prompt learning. Experimental results on the proposed topic-oriented sarcasm dataset show that our proposed topic-based prompt learning model outperforms the baseline models. Simultaneously, the in-depth analysis show that the proposed topic-oriented sarcasm detection task is more challenging compared to the traditional sentence-level sarcasm detection.

Keywords: Topic-oriented sarcasm detection , Sarcasm detection , Prompt learning

* 通讯作者

第二十一届中国计算语言学大会论文集，第557页-第568页，南昌，中国，2022年10月14日至16日。

1 引言

讽刺是一种常见的语言现象，通常使用比喻、夸张等手法对人或事进行揭露、批评或嘲笑，在语言学、心理学和认知科学等领域都得到了广泛关注 (Gibbs, 1986; Gibbs, 2007; Kreuz and Glucksberg, 1989; Kreuz and Caucci, 2007)。韦氏词典（Merriam Webster）[0]将讽刺定义为“使用与你真正想说的意思相反的词语，尤其是为了侮辱某人、表示愤怒或搞笑情绪。”（*the use of words that mean the opposite of what you really want to say especially in order to insult someone, to show irritation, or to be funny.*）。从讽刺的定义可以看出，讽刺表达通常是针对人、事物等讽刺对象而做出的语言表达。但目前大多数文本讽刺检测的研究局限于句子级别的讽刺识别和分类 (Joshi et al., 2015a; Kumar Jena et al., 2020; Xiong et al., 2019; Lou et al., 2021)，而忽略了讽刺对象对讽刺表达的影响。随着社交媒体平台的飞速发展，越来越多网络用户会对热点事件发表想法和评论，包括大量讽刺表达。其中大量评论都是基于特定事件产生的。因此，仅从评论本身出发分析其中的讽刺信息，不足以准确全面地理解用户对特定事件的实际情感。

针对这一问题，本文从一种新的角度观察讽刺表达，提出一个面向话题的讽刺识别任务。由于目前尚未有面向话题的讽刺识别公开数据集，本文设计构建了一个面向话题的讽刺识别新数据集。该数据集包含707个话题以及对应的4871个样本。其中，每一个样本由一个话题和一个评论组成。讽刺识别模型需要针对特定话题从句子的上下文中判断该评论是否为讽刺表达（讽刺或非讽刺）。针对这一问题，本文基于提示学习（prompt learning），提出一种面向话题的讽刺表达提示学习（Topic-Oriented Sarcasm Prompt Learning，TOSPrompt）模型。这一模型通过针对话题设计提示模板，可以更好地从大规模预训练语言模型（pre-trained language model，PLM）中理解句子对于话题的讽刺表达信息，从而判断该句子是否为讽刺句子。

本文的主要贡献如下：

（1）本文首次以一种新的角度观察讽刺表达，并提出一种面向话题的讽刺识别任务。

（2）本文构建了一个新的数据集用于评估面向话题的讽刺识别任务。通过数据的开源，更好地推动这一问题的研究。

（3）本文提出了一种面向话题的讽刺表达提示学习模型。该模型能有效建模面向话题的讽刺识别任务，并取得比基线模型更优的性能。

2 相关工作

2.1 讽刺识别

句子级的文本讽刺识别旨在从句子中识别上下文的讽刺表达，判断句子是否为讽刺句 (Joshi et al., 2015b)。一些早期的研究工作使用特征工程方法来.提取了句子上下文中不一致的情感表达，例如在上下文中搜索一组积极情感的动词和消极情感的表达 (Riloff et al., 2013; Bamman and Smith, 2015)、或构建词汇特征来确定不一致性 (Davidov et al., 2010; González-Ibáñez et al., 2011; Lunando and Purwarianti, 2013)，从而识别上下文的讽刺表达。

随后，基于深度学习的方法被广泛应用于句子级讽刺识别任务中。例如 (Poria et al., 2016; Majumder et al., 2019)采用预先训练的卷积神经网络（convolutional neural networks，CNN）架构来提取上下文的情感和个性特征，用于文本讽刺识别。Zhang等人 (2016)利用双向门控递归神经网络和池化神经网络来捕获推特内容和上下文信息，从而识别推特的讽刺表达信息。孙等人 (2016)提出一种融合多特征的混合神经网络判别模型。该模型融合了CNN和长短期记忆网络（long short-term memory，LSTM），有效挖掘文本中深层次的语义信息，完成讽刺识别。Tay等人 (2018)引入注意力机制，结合神经模型对上下文的情感对比和不协调情感表达进行建模，识别上下文的讽刺信息。Xiong等人 (2019)提出一种结合自匹配网络、双向LSTM网络和低阶双线性池方法的神经网络模型，从而学习单词对之间的不协调情感表达。在基于图网络模型的讽刺识别研究中，Lou等人 (2021)提出了一个基于依赖树和情感知识的情感依赖图网络模型，可以学习上下文中复杂的情感依赖信息，挖掘讽刺表达。

此外，随着大规模预训练模型在自然语言处理任务中取得的成功，也有研究工作将强大的预训练模型应用于讽刺识别任务中。Babanejad等人 (2020)利用情感信息和上下文特征来改

[0]http://www.merriam-webster.com/dictionary/sarcasm

进BERT (Devlin et al., 2019)结构，使其可以有效识别文本中的讽刺表达。樊等人 (2021)针对讽刺识别中较少利用上下文语境信息和修辞表达信息的不足，通过结合大规模预训练模型ELMo (Peters et al., 2018)，提出了基于多语义融合的反讽识别方法。上述研究工作在句子级的讽刺识别任务中取得了较好性能，但是这些研究工作都是局限于句子级的文本讽刺识别，忽略了讽刺表达中的讽刺对象和话题信息。

2.2 提示学习

随着ELMo (Peters et al., 2018)、BERT (Devlin et al., 2019)、GPT(Radford et al., 2018; Radford et al., 2019; Brown et al., 2020)等大规模预训练语言模型的发展，自然语言处理的研究和应用很多趋向于以预训练语言模型为中心，在下游任务进行微调的方式来解决问题。近年来，基于提示学习（prompt learning）的研究也受到了越来越多学者的关注，其主要思想是通过改造下游任务、增加专家知识，使任务输入和输出适合原始语言模型，从而获得更好的任务效果。Petroni等人(2019)将关系抽取任务修改为填空题，在不修改预训练语言模型的情况下，得到了比融合知识库更好的关系抽取性能。Shin等人(2020)提出了一种基于梯度引导搜索的自动方法：AUTOPROMPT。该方法可以为一组不同的任务自动创建提示。与手动创建提示的方法相比，AUTOPROMPT的提示可以从掩码语言模型（MASKed language model，MLM）中获得更准确的事实知识。Schick和Schutze(2021)基于大规模预训练语言模型，引入模式利用训练（pattern-exploiting training，PET）方法。该方法是一种半监督的训练方法，将输入样本重新设置为完形填空式短语，以帮助语言模型理解给定的任务。然后使用这些短语为大量未标记的示例指定软标签。最后，在结果训练集上执行有监督训练。这类基于提示学习的方法通过模板的引入，使得下游任务更好地匹配大规模预训练语言模型，在很多自然语言处理任务中都取得了令人瞩目的效果。受这些工作的启发，本文提出一种面向话题的讽刺表达提示学习模型，该模型通过设计面向话题的提示模板，更好地解决面向话题的讽刺识别问题。

3 面向话题的讽刺识别任务

与句子级讽刺识别不同，本文提出了一种面向话题的讽刺识别任务。这一任务通过话题的引入，以话题为讽刺背景/对象，判断文本是否为讽刺表达。例如表1给出的讽刺表达示例。对于同一个句子“真的很优秀”，样例1是一个没有话题信息的文本句子，可以看出，单纯从句子的上下文信息，难以判断该句子是否为讽刺表达。样例2是带有话题的句子。可以看出，结合话题信息可以很容易地判断该句子为讽刺句。而在样例3给定的话题场景下，可以判断样例3不是讽刺句。这就意味着在面向话题的讽刺检测任务中，当针对的话题不同时，同一个句子也有可能是不一样的讽刺标签。这就意味着必须深入结合话题信息，才能更好地判断句子是否为讽刺表达。可以看到，与传统的句子级讽刺识别相比，面向话题的讽刺识别任务更贴切真实场景。

形式化定义：对于给定的输入$x = (t, s)$，面向话题的讽刺识别旨在从评论s中挖掘针对话题t的讽刺表达信息，从而判断s针对t的类别y为“讽刺”或“非讽刺”。

样例	话题	句子	标签
1	-	真的很优秀	?
2	美国两党被曝都曾花钱挖特朗普黑料	真的很优秀	讽刺
3	男孩曾多次获得国家级竞赛奖项	真的很优秀	非讽刺

Table 1: 与话题相关的讽刺表达示例

4 面向话题的讽刺识别数据集

目前尚未有公开的面向话题的讽刺识别标注数据。为此，本文基于中文社交媒体文本，设计并构建了一个新的面向话题的讽刺识别数据集。具体地，为了收集带有话题的讽刺表达文本数据，本文从“观察者”[1]爬取带有话题的中文评论文本，形成初始数据。“观察者”是一个集新闻传播、人文社会科学研究于一体的新闻与评论一体化网站，反映了当前中国与世界各种思潮的对抗，该网站具有新闻内容更新快、活跃用户多、用户对新闻事件评论多、用户之间讨论活跃等特点。接下来本文详细介绍数据集的处理和标注过程。

[1] https://www.guancha.cn/

4.1 数据处理

考虑到数据集的规范性、通用性以及可扩展性，本文根据以下方面筛选待标注数据：

- 为确保数据的规范性，在话题选取过程中屏蔽掉包含敏感词语、讽刺表达比例低、攻击性强等话题。
- 因为过长的句子会影响模型对讽刺表达的学习能力，因此过滤掉长度超过200个词语的长文本数据。
- 以话题-评论对来构建一个数据样本，并过滤掉重复的话题-评论对。
- 过滤掉数据中的特殊符号、网页地址等与语义信息无关的信息。

这样可以得到面向话题的讽刺识别初始数据。其中，每一条数据样例由一个评论和对应的话题组成。

4.2 数据标注

在数据标注过程中，针对每一条数据样例，由5名标注者进行独立标注。整个标注过程持续了4个月。对于标注结果不一致的样例，通过多数投票机制获得最终的类别标签。同时，在标注过程中，为了提升数据集的质量，标注者丢弃了原始数据中约20%的噪音数据。包括：

- 评论和话题内容不相关的数据。
- 评论带有敏感词语、攻击性词语等表达的数据。
- 5名标注者都难以通过话题和评论内容判断类别标签的数据。

此外，因为话题通常是有明确的主题，并且有规范的信息表达，而从社交平台中爬取到的某些原始话题会存在噪音。因此，本文在标注过程中对原始话题进行了修正，包括：

- 删除不合适的断句，例如表2话题1中的“外交部回应”。
- 删除话题中的冗余信息，例如表2话题2中的“又一个！”。
- 重新组织话题表述，使话题更通顺，例如表2话题3中前后两个分句缺少的因果关系连接词“造成”。

Id	原话题	修改后话题
1	新西兰禁止运营商使用华为5G技术 外交部回应	新西兰禁止运营商使用华为5G技术
2	又一个！萨尔瓦多与台湾“断交”	萨尔瓦多与台湾“断交”
3	英国曼彻斯特发生恐怖爆炸袭击 22死59伤	曼彻斯特发生恐怖爆炸袭击造成22死59伤

Table 2: 话题处理示例

样例	话题	评论	类别
1	美国驱逐舰又撞船了	真会玩，在海上开“碰碰船”	讽刺
2	中国首款RISC-V高性能家电芯片在青岛诞生	加油	非讽刺

Table 3: 讽刺标注结果示例

受现有的句子级讽刺识别研究工作的启发 (Joshi et al., 2015b; Xiong et al., 2019; Lou et al., 2021)，本文针对每一个话题-评论对组成的样例标注为“讽刺”或“非讽刺”类别，类别标签标注举例如表3所示。因此，面向话题的讽刺识别任务中每一条标注样例可以表示为(t, s, y)，其中y为该样例的类别标签。

为了保证面向话题的讽刺识别任务具有更合理的评估结果，本文尽可能使每一个话题都包含讽刺样本和非讽刺样本。同时，为了避免因样本类别分布过于不平衡而导致无效的模型训练，本文从标注好的数据集中随机挑选样本，构建平衡“讽刺”和“非讽刺”两个类别数据分布的数据集。最终得到一个面向话题的讽刺识别（Topic-Oriented Sarcasm Detection，ToSarcasm）标注数据集。该数据集包含4871个由话题-评论对组成的样本，其中话题有707个。数据集在各个类别下的样本分布情况如表4所示。

	讽刺样本	非讽刺样本	总计
样本数量	2436	2435	4871
样本占比（%）	50.01	49.99	100

Table 4: ToSarcasm数据集的数据分布

5 面向话题的讽刺表达提示学习模型

面向话题的讽刺识别任务的主要难点是需要结合特定话题，判断评论针对该话题是否为讽刺表达。因此，本文借助大规模预训练语言模型的优势，设计面向话题的讽刺表达提示模板，目的是让模型更好地从预训练语言模型中学习面向话题的讽刺表达知识。如图1所示，本文提出的面向话题的讽刺表达提示学习（TOSPrompt）模型主要由三个模块组成：1）面向话题的讽刺表达模板构造。根据输入的话题和评论构造面向话题的讽刺表达提示模板；2）标签词预测。通过预训练语言模型给模板中的“$[MASK]$”位置预测标签词；3）模型训练。根据训练数据对模型进行训练优化。接下来，本节将详细介绍TOSPrompt模型的各个模块。

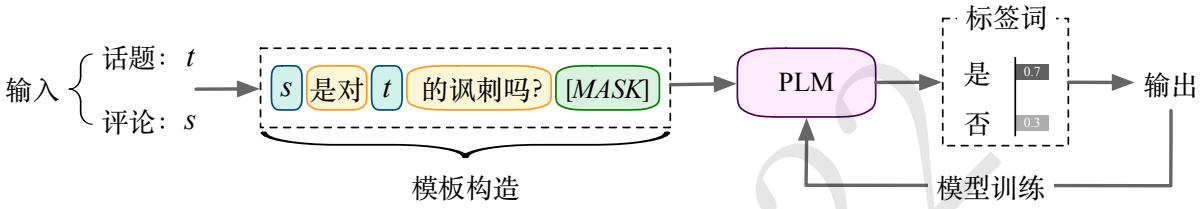

Figure 1: 面向话题的讽刺表达提示学习（TOSPrompt）模型框架图

5.1 面向话题的讽刺表达模板构造

为了使面向话题的讽刺识别任务更适合预训练语言模型，本文基于 (Shin et al., 2020; Schick and Schütze, 2021)等工作提出的基于提示学习的模型，以前缀提示（prefix prompt）模板的形式针对输入样本构造面向话题的讽刺表达模板。因此，本文基于预训练语言模型的掩码语言模型（masked language model）来对“$[MASK]$”标记位置进行词语填补。使用掩码语言模型的优势在于，通过掩码语言模型，可以基于大规模的预训练语料，利用非掩码区域的特征来为掩码位置“$[MASK]$”预测出合适的词语，从而预测出合适的类别标签。所构造的面向话题的讽刺表达模板定义如下：

$$\mathbf{x}_{prompt} = s\text{是对}t\text{的讽刺吗? } [MASK] \tag{1}$$

基于此，可以得到输入样例的面向话题的讽刺表达模板。接下来，需要借助预训练语言模型（PLM）对“$[MASK]$”位置进行类别标签词预测，从而识别该样例是否为讽刺表达。

5.2 标签词预测

针对给定输入(t, s)的模板“$\mathbf{x}_{prompt} = s$是对t的讽刺吗? $[MASK]$”，本文使用BERT中文预训练语言模型（BERT-base Chinese）(Devlin et al., 2019)作为预训练语言模型对输入样例进行建模。模型的输入表示为：

$$\boldsymbol{x} = [CLS]\mathbf{x}_{prompt}[SEP] \tag{2}$$

随后，将输入表示输入到BERT预训练语言模型$\mathcal{M}$中，以掩码语言模型的方式通过语言模型$\mathcal{M}$预测“$[MASK]$”位置的类别标签词分布：

$$P_{\mathcal{M}}([MASK]|\boldsymbol{x}) = \mathcal{M}(\boldsymbol{x}) \tag{3}$$

这里，为了使提出的TOSPrompt模型简单且具有更强的通用性，本将词语“是”和“否”作为模型的类别标签词。即，标签词集合$\mathcal{V} = \{\text{是}, \text{否}\}$，分别对应“讽刺”类别和“非讽刺”类别。

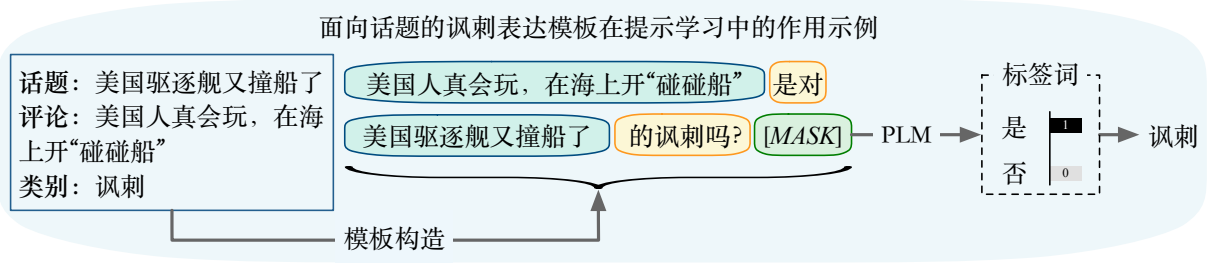

Figure 2: 面向话题的讽刺表达模板在提示学习中的作用示例

图 2给出了一个面向话题的讽刺表达模板在提示学习中的作用示例。如图2所示，借助面向话题的讽刺表达模板，模型针对该输入样本，可以更好地从预训练语言模型中学习到“$[MASK]$”位置的标签词为“是”，对应“讽刺”类别。

5.3 模型训练

基于上述的预训练语言模型$\mathcal{M}$，可以得到$[MASK]$位置预测为标签词集合中每一个标签词v的概率分布。这里，为了将单词的概率映射到标签的概率上，本文定义了一个映射器（verbalizer）将标签词集合$\mathcal{V}$中的词语映射到类别分布空间$\mathcal{Y}$中，即$f: \mathcal{V} \mapsto \mathcal{Y}$。因此，对于输入样本$\boldsymbol{x}$预测出标签词$v \mapsto y$的类别分布$P(y|\boldsymbol{x})$计算如下：

$$P(y|\boldsymbol{x}) = g\big(P_{\mathcal{M}}([MASK] = v|\boldsymbol{x})\big) \tag{4}$$

其中，$g(\cdot)$为将标签词的概率转换为类别标签概率的函数。因此，对于本文设计的标签词集合$\mathcal{V}$，输入样本x的预测类别标签定义为：

$$\text{argmax } \hat{\boldsymbol{y}} = \underset{y \in \mathcal{Y}}{\text{argmax}}\ P_{\mathcal{M}}([MASK] = v_y|\boldsymbol{x}) \tag{5}$$

其中，$\hat{\boldsymbol{y}}$为预测类别分布。v_y为类别标签y对应的标签词。随后，基于训练数据，通过最小化交叉熵损失对提出的TOSPrompt模型进行训练和优化：

$$\mathcal{L} = -\sum_{i=1}^{N}\sum_{j=1}^{L} y_i^j \log \hat{y}_i^j + \lambda ||\Theta||^2 \tag{6}$$

其中，N为训练集大小。L为类别数量。$\boldsymbol{y}_i$和$\hat{\boldsymbol{y}}_i$分别代表训练样本i的真实类别和预测类别分布。Θ为模型中所有的可训练参数。λ为L_2正则化系数。

6 实验

本文在构建的面向话题的讽刺识别新数据集（ToSarcasm数据集）上进行实验和分析。通过与现有的讽刺识别基线模型和大规模预训练语言模型进行对比实验，评估本文提出的TOSPrompt模型在面向话题的讽刺识别任务的性能。同时，通过实验和分析，评估本文提出的面向话题的讽刺识别新任务的研究价值和挑战性。

6.1 实验数据与参数设置

为了评估本文提出的TOSPrompt模型在面向话题的讽刺识别任务中的性能，本文首先对ToSarcasm数据集进行数据划分。对于每一条标注数据$x_i = (s_i, t_i, y_i)$，文本将其随机地分配给训练集、验证集或测试集。其中，训练集、验证集和测试集的比例为：6:2:2。基于此，可以得到ToSarcasm数据集的在训练集、验证集和测试集上的数据集合，数据统计如表5所示。

在实验中，本文采用JIEBA分词工具对文本进行中文分词处理[2]。对于本文提出的TOSPrompt模型，使用BERT中文预训练语言模型（BERT-base Chinese）(Devlin et al.,

[2] https://github.com/fxsjy/jieba

标签	训练集样本数	验证集样本数	测试集样本数
讽刺	1464	486	486
非讽刺	1461	487	487
总计	2925	973	973

Table 5: ToSarcasm数据集的数据统计信息

2019)作为预训练语言模型编码器，将输入样本编码成768维的向量。L_2正则化系数λ设置为0.00001。Dropout设置为0.1。模型采用Adam算法作为参数优化器，学习率设置为0.00002，权重衰减系数设置为0.002。批量大小Mini-batch设置为32。

6.2 评估指标

在模型评估中，本文采用精确率（Precision）、召回率（Recall）、F1值（F1-score）以及准确率（Accuracy）综合评估模型在面向话题的讽刺识别分析任务中的性能。因为面向话题的讽刺识别任务主要关注模型是否能正确识别出样本中的讽刺表达，同时区分非讽刺表达。基于此，本文设定“讽刺”类别为正例（Positive），“非讽刺”类别为反例（Negative）。因此，准确率表示预测正确的样本占所有样本的比例；精确率表示预测为“讽刺”的样本中有多少是真正的“讽刺”样本；召回率表示针对原来的样本而言，样本中的“讽刺”类别样本有多少被预测正确了；F1值则综合精确率和召回率的结果。各评估指标计算公式如下：

$$\text{准确率} = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

$$\text{精确率} = \frac{TP}{TP + FP} \tag{8}$$

$$\text{召回率} = \frac{TP}{TP + FN} \tag{9}$$

$$\text{F1值} = \frac{2 \times \text{精确率} \times \text{召回率}}{\text{精确率} + \text{召回率}} \tag{10}$$

其中，TP（True Positive）表示预测类别和实际类别都为“讽刺”。FP（False Positive）表示预测类别为“讽刺”，而实际类别为“非讽刺”。FN（False Negative）表示预测类别为“非讽刺”，而实际类别为“讽刺”。TN（True Negative）表示预测类别和实际类别都为“非讽刺”。

6.3 对比模型

在实验中，本文选取以下基线模型作为本文提出的TOSPrompt模型的对比模型：

- **BiLSTM**：使用一个双向的LSTMs (Hochreiter and Schmidhuber, 1997)分别学习句子和目标的隐藏层表示，最终将隐藏表示拼接作为讽刺识别的分类特征。
- **MIARN** (Tay et al., 2018)：句子级别的讽刺识别模型。使用基于注意的神经网络模型，分别对评论和话题进行讽刺表达的不一致性学习，最终将隐藏层表示拼接作为讽刺识别的分类特征。
- **ADGCN** (Lou et al., 2021)：句子级别的讽刺识别模型。基于外部情感知识的图网络模型，通过学习情感的不一致性挖掘上下文的讽刺表达。本文使用ADGCN分别对话题和评论进行建模，最终将隐藏表示拼接作为讽刺识别的分类特征[3]。
- **BERT** (Devlin et al., 2019)：原始的BERT中文预训练语言模型BERT-base Chinese，该模型使用“$[CLS]s[SEP]t[SEP]$”作为模型输入。
- **ADGCN-BERT** (Lou et al., 2021)：ADGCN的变种，编码器使用BERT-base Chinese。
- **PET** (Schick and Schütze, 2021)：基于Schick和Schutze (2021)的提示学习研究工作，使用“$[CLS]s[SEP]t[SEP]$这是$[MASK]$”构建模板输入到BERT-base Chinese中，预测$[MASK]$位置是“讽刺”或“非讽刺”。

[3]其中中文情感词得分来自BosonNLP：http://bosonnlp.com/

其中，对于BiLSTM、IAN、MIARN和ADGCN这四种非BERT基线模型，本文使用Shen等人 (2018)提出的Chinese Word Vectors中文词向量对词语进行词向量初始化。

6.4 面向话题的讽刺识别实验结果

为了评估本文提出的TOSPrompt模型在面向话题的讽刺识别任务中的有效性，本文在ToSarcasm数据集上与基线模型进行了对比实验，实验结果如表6所示。从表中结果可以看出，本文提出的TOSPrompt模型在所有的评估指标上都取得了最佳性能。这显示了本文提出的TOSPrompt模型在面向话题的讽刺识别任务中的有效性。与句子级讽刺识别对比模型（MIARN、ADGCN和ADGCN-BERT）相比，本文提出的TOSPrompt模型在所有评估指标上都取得了大幅度的提升。这表明，单纯的句子级讽刺识别模型并不能很好地处理面向话题的讽刺识别任务，而本文提出的TOSPrompt模型借助话题信息的建模，可以更好地解决面向话题的讽刺识别任务。另一方面，基于BERT的模型相比基于传统词向量的模型总体上取得更优的性能。这说明在面向话题的讽刺识别任务中，使用更强大的预训练语言模型在学习讽刺表达时能取得更好的学习效果。此外，相比现有基于提示学习的模型（PET），本文提出的TOSPrompt模型在所有评估指标上的性能都取得了提升，其中准确率提升了1.06%，F1值提升了1.76%。这说明，在面向话题的讽刺识别任务中，通过面向话题来设计讽刺表达的提示学习模板，可以更好地针对话题来学习评论上下文中的讽刺表达信息，从而取得更优的性能。

模型	准确率	精确率	召回率	F1值
Bi-LSTM	63.72	61.65	72.55	66.65
MIARN (Tay et al., 2018)	65.32	63.25	74.12	68.25
ADGCN (Lou et al., 2021)	65.90	63.16	76.50	69.19
BERT (Devlin et al., 2019)	69.66	67.26	74.79	70.83
ADGCN-BERT (Lou et al., 2021)	70.57	68.93	75.10	71.88
PET (Schick and Schütze, 2021)	70.70	67.57	75.78	71.44
TOSPrompt	**71.76**	**70.02**	**76.68**	**73.20**

Table 6: 模型在面向话题的讽刺识别任务上的性能（%）

6.5 不同模板对实验性能的影响

为了分析本文提出的TOSPrompt模型在使用不同模板的提示学习对性能带来的影响，本文以非面向话题和面向话题两个角度来构建模板，针对TOSPrompt模型设计了以下变种。

1）非面向话题的模板：

- $\mathbf{x}_{prompt} = s[SEP]t[SEP]$这是$[MASK]$表达：该模板采用完形填空提示（cloze prompt）模板，预测$[MASK]$位置是“讽刺”或“非讽刺”。
- $\mathbf{x}_{prompt} = s[SEP]t[SEP]$这是讽刺吗？$[MASK]$：该模板采用前缀提示模板，预测$[MASK]$位置是“是”或“否”。

2）面向话题的模板：

- $\mathbf{x}_{prompt} =$ 针对t的评论s是$[MASK]$表达：该模板采用完形填空提示模板，预测$[MASK]$位置是“讽刺”或“非讽刺”。
- $\mathbf{x}_{prompt} = s$针对t是$[MASK]$表达：该模板采用完形填空模板，预测$[MASK]$位置是“讽刺”或“非讽刺”。

表 7给出了不同模板在面向话题的讽刺识别任务上的实验性能。从表中结果可以看出，针对话题构建面向话题的讽刺表达提示模板在所有评估指标上都取得了比非面向话题的模板更好的性能。这说明，在面向话题的讽刺识别任务中，针对话题来设计讽刺表达提示模板可以更好地从预训练语言模型中挖掘出关于话题的语言知识，从而能更好地根据评论文本学习面向该话题的讽刺识别特征信息，取得更优的性能。

模板	准确率	精确率	召回率	F1值
$\mathbf{x}_{prompt} = s[SEP]t[SEP]$这是$[MASK]$表达	70.86	68.62	76.43	72.31
$\mathbf{x}_{prompt} = s[SEP]t[SEP]$这是讽刺吗？$[MASK]$	70.92	69.49	75.61	72.42
$\mathbf{x}_{prompt}$ = 针对t的评论s是$[MASK]$表达	71.37	69.93	76.22	72.94
$\mathbf{x}_{prompt} = s$针对t是$[MASK]$表达	71.60	**70.35**	76.13	73.13
TOSPrompt	**71.76**	70.02	**76.68**	**73.20**

Table 7: 不同模板的性能（%）对比

6.6 不同比例训练数据的性能分析

为了评估训练数据样本数量对本文提出的TOSPrompt模型的性能影响，基于BERT、PET和所提出的TOSPrompt模型使用不同比例的训练数据集合进行了对比实验，结果如图3所示。从图中结果可以看出，相比于原始的BERT模型，基于提示学习的PET和本文提出的TOSPrompt模型在各个比例的训练数据下都取得了更优的性能。这也进一步显示了提示学习在面向话题的讽刺识别任务中的有效性。此外，本文提出的TOSPrompt模型在不同大小的训练数据下性能都始终优于BERT和PET，特别是在仅使用20%-60%的训练数据时提升尤为显著。这说明本文提出的TOSPrompt模型由于设计了面向话题的讽刺表达提示模板，能更好地从预训练语言模型中针对话题学习上下文中的的讽刺表达信息，因此在缺少训练数据时也能取得令人满意的性能。

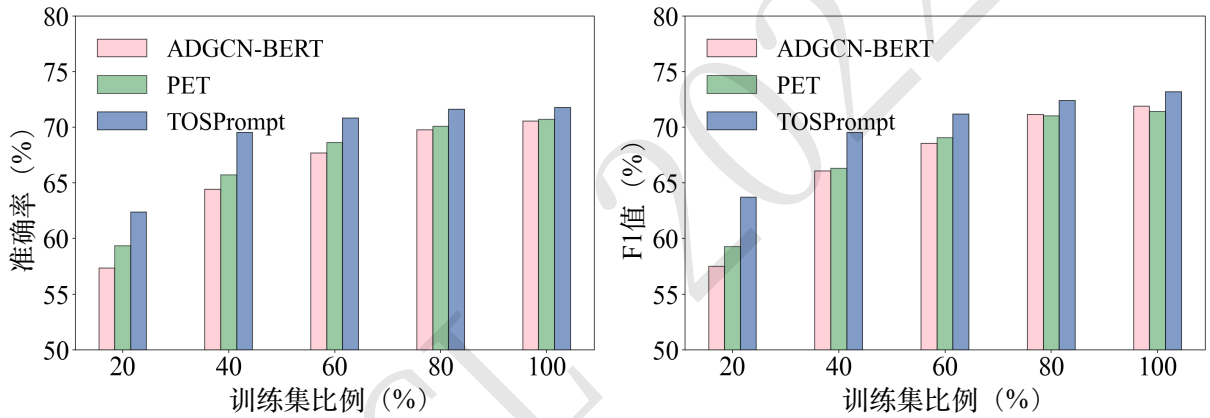

Figure 3: 使用不同比例训练数据的实验结果

样例	话题	评论
1	三位 大妈偷走巨型烧烤炉 ，被抓时正在买食材	中国 大妈们真厉害 ！
2	马克龙访美与特朗普 互动存在分歧	好甜蜜 ，羡慕这样的爱情。
3	特朗普当着一屋子 大学生 的面说： 我超爱贷款 ！	真棒 ， 好潇洒 的美国人！

Table 8: 典型的面向话题的讽刺样本示例

6.7 话题样例分析

本文提出的面向话题的讽刺识别分析任务相比传统的句子级讽刺识别任务，引入了讽刺表达中的话题对象。因此，本文进一步通过典型的讽刺样例来分析话题对讽刺识别的作用。

表 8给出三个典型的面向话题的讽刺表达样例，可以看出，仅从评论的内容都难以判断这三个样例是否为讽刺表达，因为评论中的上下文只表达了单向的积极情感，没有体现出讽刺表达中的情感不一致描述。而当结合话题信息后，可以看出，这三个样例都是讽刺表达样例。对

于样例1，话题中的“大妈偷走巨型烧烤炉”是负面的。因此，结合评论中的“大妈们真厉害”可以判断出该样例是讽刺样本。同样地，对于样例2和样例3，话题中的阴影部分内容都是带有负面情绪的表达，通过结合评论中的描述，可以推断出样例2和样例3也是讽刺样本。可以看出在面向话题的讽刺识别任务中，话题的内容对于判断样本是否为讽刺表达是至关重要的。这充分显示了本文提出的面向话题的讽刺识别任务的合理性和研究价值。此外，从典型样例也可以看出，解决面向话题的讽刺识别任务不仅仅需要针对话题和评论挖掘上下文的语义信息，还需要对话题和评论的上下文信息进行匹配，挖掘话题和评论中重要内容的联系（图8中对应背景颜色的文字描述）。这也意味着面向话题的讽刺识别任务相比句子级的讽刺识别任务具有更强的挑战性。

6.8 错误样例分析

从表 6中结果可以看出，所有的模型在面向话题的讽刺识别任务中的性能指标都没能超过80%，这也侧面反映了面向话题的讽刺识别任务的挑战性和研究价值。为了进一步分析任务数据中的挑战性，本文对提出的TOSPrompt模型的错误样本进行了分析，并将错误分类的样本大致归类为以下类型：

1）需要一定的背景知识才能了解话题和评论所表达的内容。例如以下例子，其正确类别标签为“讽刺”：

话题：又一国际巨头将撤离深圳！留下超10万平米土地谁接盘?

评论：你说的很有道理，深圳只是个小县城

该例子中“深圳”是一个大都市，而不是“小县城”，因此模型需要加入额外的背景知识才能更好地对其内容进行学习，得出正确的分类结果；

2）评论中带有缩写或非正规用词。例如以下例子，其正确类别标签为“讽刺”：

话题：萨尔瓦多与台湾“断交”

评论：恭喜菜菜，真的快“独”了！

该例子中的“菜菜”指的是“蔡英文”。因此，模型需要将这些词语映射为正规用词才能准确理解评论中的上下文信息表达，得出正确的分类结果；

3）评论中带有隐喻表达的词语。例如以下例子，其正确类别标签为“讽刺”：

话题：因缺少备件，德国海军潜艇全部趴窝

评论：“工匠”们累了，要休息啦！

该例子中评论内容将“德国海军”比喻成“工匠”，因此需要将隐喻表达跟话题中的事物对应起来才能识别样本的讽刺表达信息，得出正确的分类结果。

因此，针对上述的错误样例分析，在未来的研究中，可以考虑探索在模型中融入话题和评论中所讨论事物的背景知识、对评论文本中涉及的实体进行识别和共指消解、对样本中的隐喻表达进行识别和对齐等技术，以进一步提升面向话题的讽刺识别任务的性能。

7 结论

针对现有的讽刺识别研究通常只针对句子级别来挖掘上下文的讽刺表达信息，但忽略了讽刺表达的话题背景或讽刺对象的不足，本文提出一个新的面向话题的讽刺识别任务。该任务通过引入话题信息作为讽刺表达的对象，使讽刺识别的研究更贴切真实场景且更具挑战性。为此，本文构建了一个面向话题的讽刺标注数据集，以推动这一研究的开展。此外，为了解决面向话题的讽刺识别任务，本文基于提示学习，提出了一种面向话题的讽刺表达提示学习（TOSPrompt）模型。与一系列基线模型的对比实验结果表明，本文提出的TOSPrompt模型在面向话题的讽刺识别任务中取得了最佳性能。

参考文献

Nastaran Babanejad, Heidar Davoudi, Aijun An, and Manos Papagelis. 2020. Affective and contextual embedding for sarcasm detection. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 225–243, Barcelona, Spain (Online). International Committee on Computational Linguistics.

David Bamman and Noah A Smith. 2015. Contextualized sarcasm detection on twitter. In *Ninth international AAAI conference on web and social media*, pages 574–577.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris 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. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Semi-supervised recognition of sarcasm in Twitter and Amazon. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, pages 107–116, Uppsala, Sweden. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Raymond W Gibbs. 1986. On the psycholinguistics of sarcasm. *Journal of experimental psychology: general*, 115(1):3.

Raymond W Gibbs. 2007. On the psycholinguistics of sarcasm. *Irony in language and thougt: A cognitive science reader*, pages 173–200.

Roberto González-Ibáñez, Smaranda Muresan, and Nina Wacholder. 2011. Identifying sarcasm in Twitter: A closer look. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 581–586, Portland, Oregon, USA. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Aditya Joshi, Vinita Sharma, and Pushpak Bhattacharyya. 2015a. Harnessing context incongruity for sarcasm detection. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 757–762, Beijing, China. Association for Computational Linguistics.

Aditya Joshi, Vinita Sharma, and Pushpak Bhattacharyya. 2015b. Harnessing context incongruity for sarcasm detection. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 757–762, Beijing, China. Association for Computational Linguistics.

Roger Kreuz and Gina Caucci. 2007. Lexical influences on the perception of sarcasm. In *Proceedings of the Workshop on Computational Approaches to Figurative Language*, pages 1–4, Rochester, New York. Association for Computational Linguistics.

Roger J Kreuz and Sam Glucksberg. 1989. How to be sarcastic: The echoic reminder theory of verbal irony. *Journal of experimental psychology: General*, 118(4):374.

Amit Kumar Jena, Aman Sinha, and Rohit Agarwal. 2020. C-net: Contextual network for sarcasm detection. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 61–66, Online. Association for Computational Linguistics.

Shen Li, Zhe Zhao, Renfen Hu, Wensi Li, Tao Liu, and Xiaoyong Du. 2018. Analogical reasoning on chinese morphological and semantic relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 138–143. Association for Computational Linguistics.

Chenwei Lou, Bin Liang, Lin Gui, Yulan He, Yixue Dang, and Ruifeng Xu. 2021. Affective dependency graph for sarcasm detection. In *the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21)*, pages 1844–1849.

Edwin Lunando and Ayu Purwarianti. 2013. Indonesian social media sentiment analysis with sarcasm detection. In *2013 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, pages 195–198. IEEE.

Navonil Majumder, Soujanya Poria, Haiyun Peng, Niyati Chhaya, Erik Cambria, and Alexander Gelbukh. 2019. Sentiment and sarcasm classification with multitask learning. *IEEE Intelligent Systems*, 34(3):38–43.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. 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 2227–2237, New Orleans, Louisiana, June. Association for Computational Linguistics.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, and Prateek Vij. 2016. A deeper look into sarcastic tweets using deep convolutional neural networks. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1601–1612, Osaka, Japan. The COLING 2016 Organizing Committee.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert, and Ruihong Huang. 2013. Sarcasm as contrast between a positive sentiment and negative situation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 704–714, Seattle, Washington, USA. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online, April. Association for Computational Linguistics.

Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. 2020. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235.

Yi Tay, Anh Tuan Luu, Siu Cheung Hui, and Jian Su. 2018. Reasoning with sarcasm by reading in-between. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1010–1020, Melbourne, Australia. Association for Computational Linguistics.

Tao Xiong, Peiran Zhang, Hongbo Zhu, and Yihui Yang. 2019. Sarcasm detection with self-matching networks and low-rank bilinear pooling. In Ling Liu, Ryen W. White, Amin Mantrach, Fabrizio Silvestri, Julian J. McAuley, Ricardo Baeza-Yates, and Leila Zia, editors, *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, pages 2115–2124. ACM.

Meishan Zhang, Yue Zhang, and Guohong Fu. 2016. Tweet sarcasm detection using deep neural network. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2449–2460, Osaka, Japan. The COLING 2016 Organizing Committee.

孙晓, 何家劲, and 任福继. 2016. 基于多特征融合的混合神经网络模型讽刺语用判别. 中文信息学报, 30(6):215–223.

樊小超, 杨亮, 林鸿飞, 刁宇峰, 申晨, and 楚永贺. 2021. 基于多语义融合的反讽识别. 中文信息学报, 35(6):103–111.