计算机科学
COMPUTER SCIENCE

CCSD:面向话题的讽刺识别方法

刘其龙, 李弼程, 黄志勇

引用本文

刘其龙, 李弼程, 黄志勇. CCSD:面向话题的讽刺识别方法[J]. 计算机科学, 2024, 51(9): 310-318.

LIU Qilong, LI Bicheng, HUANG Zhiyong. CCSD:Topic-oriented Sarcasm Detection[J]. Computer Science, 2024, 51(9): 310-318.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230600217

CCSD:面向话题的讽刺识别方法

刘其龙　李弼程　黄志勇
华侨大学计算机科学与技术学院　福建 厦门 361000
(21014083073@stu.hqu.edu.cn)

摘　要　随着社交媒体的发展,越来越多的人在社交平台上发表对热点话题的看法,其中讽刺手法的运用严重影响了社交媒体中情感分析的精度。目前面向话题的讽刺识别研究未同时考虑上下文和常识知识的作用,也忽略了在同一个话题下进行讽刺识别的场景。为此,提出了基于上下文和常识的讽刺识别模型(Sarcasm Detection with Context and Common Sense,CCSD)。首先,模型使用 C^3KG 常识库生成常识文本,并将目标句、话题上下文和常识文本作为预训练 BERT 模型的输入。其次,使用注意力机制来关注目标句和常识中重要的信息。最后,通过门控机制和特征融合,实现讽刺识别。文中构建了一个面向话题的讽刺识别数据集,以验证模型在特定话题中的有效性。实验结果表明,相比基线模型,新模型的性能更优。

关键词:讽刺识别;面向话题的讽刺识别;上下文;常识知识;注意力机制
中图分类号　TP391

CCSD:Topic-oriented Sarcasm Detection

LIU Qilong,LI Bicheng and HUANG Zhiyong
School of Computer Science and Technology,Huaqiao University,Xiamen,Fujian 361000,China

Abstract　With the development of social media,an increasing number of people express their opinions about hot topics on social platforms,and the utilization of sarcastic expression has severely affected the accuracy of sentiment analysis in social media. Currently,topic-oriented sarcasm detection research does not consider the role of context and common sense knowledge simultaneously,and also ignores the scene of sarcasm recognition under the same topic. This paper proposes a sarcasm detection with context and commonsense(CCSD) approach. Firstly,the model uses the C^3KG commonsense knowledge base to generate commonsense text. Then,the target sentence,topic context,and commonsense text are concatenated as the input to the pre-training BERT model. In addition,an attention mechanism is used to focus on important information in the target sentence and commonsense text. Finally,sarcasm detections are realized through gating mechanism and feature fusion. A topic-oriented sarcasm detection dataset is constructed to verify the effectiveness of the proposed model in specific topics. Experimental results show that the proposed model achieves better performance compared to baseline models.

Keywords　Sarcasm detection,Topic-oriented sarcasm detection,Context,Common sense knowledge,Attention mechanism

1　引言

讽刺是互联网上常用的表现手法,使用轻蔑或调侃的语气表达出与所描述的事物相反的意思或效果,具有隐喻性、真实性的特点,往往难以识别。随着社交媒体的发展,Twitter、Instagram、微博等软件已成为人们日常生活中的一部分。其中讽刺手法的运用严重影响了社交媒体中情感分析的精度,提高了舆情监测的难度[1]。因此,针对社交网络中讽刺识别的研究,对情感分析、意见挖掘、舆情监测等任务具有重要意义。

目前,讽刺识别的研究主要分为上下文无关的讽刺识别[2-3]、上下文有关的讽刺识别[4-6],以及多模态讽刺识别[7-8],大多都集中于句子级别的研究。然而在社交网络中,用户发表的评论往往是针对某一事件的观点和看法。对帖文评论的讽刺识别并不同于传统的句子级别的讽刺识别,帖文本身是对事件的描述或者作者的观点,是作者想法的主观体现,与传统讽刺识别中客观的上下文有所区别,且评论与帖文呈现多对一关系。针对这一现象,Liang 等[9]将此任务定义为面向话题的讽刺识别,并提出了一种面向话题的讽刺表达提示学习(Topic-Oriented Sarcasm Prompt Learning,TOSPrompt)模型。该模型针对话题设置提示模板,使用提示学习的方法来挖掘预训练语言模型中的事实和常识知识,更好地从大规模预训练语言模型中理解评论对于话题表达的讽刺信息,从而判断该句子是否为讽刺句子。尽管该模型效果优于基线方法,但预训练模型的语料多且杂,训练出的模型捕获的知识

到稿日期:2023-06-29　返修日期:2023-10-22
基金项目:装备预研教育部联合基目(8091B022150)
This work was supported by the Joint Fund of Ministry of Education for Equipment Pre-research(8091B022150).
通信作者:李弼程(lbclm@163.com)

包含大量噪声，不加区分直接使用其中的知识会影响模型最终的效果。

短文本作为各大社交媒体平台发布信息的主流方式，使得用户在发表自己言论时省略了许多常识和背景知识，因而模型仅凭借帖文和评论往往难以判断评论是否具有讽刺信息。例如在“环保少女对日本排放核废水不做表态”这一事件的热点帖文评论“环保少女先干了这杯氚牌矿泉水再说”中，若要判定此评论具有讽刺信息，则需要具备核废水中具有氚这种放射性物质的常识，本文称其为常识依赖。然而，大多数社交媒体的评论并不包含这些潜在的常识信息。随着自然语言处理的发展，常识知识的重要性越来越明显，但在讽刺识别中却鲜有研究考虑常识知识。

上下文信息在讽刺识别中也至关重要，因为相同的文本在不同的上下文中可能会表现出截然不同的含义。以图 1 为例，相同的目标句在上下文 1 中表达正面信息，而在上下文 2 中却包含明显的讽刺信息，本文称其为上下文依赖。因此，如何在讽刺识别中同时考虑上下文以及引入外部常识是一个值得研究的问题。

图 1 依赖上下文的讽刺识别案例

Fig. 1 Context-dependent sarcasm detection cases

为解决以上问题，本文提出了一种基于上下文和常识的讽刺识别(CCSD)模型，为了同时考虑目标句的上下文和常识知识的作用，该模型引入了注意力机制模块和门控机制模块，使得模型可以动态融合多种特征以实现针对特定话题的讽刺识别。在外部常识的选择上，本文使用 C^3KG[10] 作为外部常识库。C^3KG 是融合了社会常识知识和对话流信息的中文常识知识图谱，该知识图谱并不仅仅由一个个实体组成，还包含丰富的日常生活中的事件，满足本文对常识库的要求。

由于不同语言表现讽刺的方式各不相同，国内目前面向话题的讽刺识别的公开数据集较少。因此，本文构建了一个面向微博话题的讽刺识别数据集。该数据集包含 92 个各种类型的微博热点帖文，以及帖文对应的根评论，以讽刺和非讽刺进行标注。

本文的主要贡献如下：

1)提出了一个基于上下文和常识的讽刺识别模型(CCSD)，旨在解决现有讽刺识别方法未同时考虑上下文与常识知识的问题。

2)构建了一个新的面向微博话题的讽刺识别数据集，实现特定话题下的讽刺识别。

3)提出的模型能够有效判断评论中是否包含讽刺信息，取得了比基线方法更优的性能；通过对数据案例进行分析，得到了未来可能的研究方向。

2 相关工作

相关工作主要包含两部分：文本讽刺识别以及在自然语言处理任务中引入外部知识。

2.1 讽刺识别

讽刺识别指给定一个目标句，判断目标句是否具有讽刺性。面向话题的讽刺识别与讽刺识别类似，是针对特定事件的评论分析其是否包含讽刺信息。根据方法是否关注上下文信息，可将文本讽刺识别方法分为上下文无关的讽刺识别和上下文有关的讽刺识别。

2.1.1 上下文无关的讽刺识别

上下文无关的讽刺识别指不结合上下文信息，凭借目标句本身来判断是否为讽刺句。Ghosh 等[11] 提出的讽刺识别模型包含 CNN，LSTM 和 DNN 3 部分，其中 CNN 用于提取局部语义特征，LSTM 用于提取全局语义特征，DNN 用于融合两种特征实现讽刺识别。Van 等[12] 首次提出细粒度讽刺识别，将讽刺类型划分为 4 个类别，该团队使用各种神经网络方法和手工制作的特征，结果表明，细粒度讽刺识别比二元讽刺识别更具挑战性。Joshi 等[13] 使用单词嵌入之间的语义相似性和不一致性完成讽刺识别。同样，Tay 等[14] 发现前后情感矛盾式的讽刺句占比较大，他们利用这种发现设计了句内注意力网络从而实现讽刺识别。

2.1.2 上下文有关的讽刺识别

事实上，讽刺识别严重依赖上下文信息，相同的句子在不同的上下文中会产生不一样的理解。近年来，越来越多学者着手研究上下文有关的讽刺识别任务。Hazarika 等[15] 利用用户的风格特征和个性特征进行编码得到用户嵌入，并记录用户每一条推文的论坛信息得到主题嵌入，将用户嵌入和主题嵌入作为上下文特征。Kolchinski 等[16] 通过设计作者与文本间的密集嵌入作为上下文向量，辅助完成讽刺识别任务。Babanejad 等[17] 使用情感特征和上下文特征扩展了 BERT 的体系结构，实现讽刺识别。面向话题的讽刺识别也是上下文有关的讽刺识别，其话题信息即可作为特殊的上下文信息。现有研究结果表明，结合上下文特征的讽刺识别模型的效果通常优于上下文无关的讽刺识别模型，这说明上下文特征可以有效提高模型判定讽刺信息的准确率，但外部常识知识在讽刺识别中的作用却被忽略。

2.2 外部知识整合

在自然语言处理任务中，整合外部知识的研究有很多。Bi 等[18] 在生成式问答的机器阅读理解中使用问题、文章、词汇表和外部知识来生成答案，通过事实抽取和事实选择来保留有用的外部知识以提高答案质量。Liu 等[19] 将知识图谱作为外部知识库引入搜索系统中，不仅使用了 query 和 Document 的实体嵌入，还使用了实体描述的嵌入信息和实体种类的嵌入信息，从而让模型更加理解实体的具体含义，而不是停留在对字面意思的理解。Wang 等[20] 在自然语言推理任务中引入外部知识，通过搜索句子中主谓宾对应的实体得到知识图谱中实体间的路径，然后输入 BiLSTM 中得到知识表示，最后结合原文完成推理。

越来越多学者尝试在各种自然语言处理任务中引入外部知识，但关于讽刺识别的研究较少。在讽刺识别模型中考虑上下文的同时引入外部常识知识能有效满足讽刺识别的上下文依赖和常识依赖，可以提高面向话题的讽刺识别准确性。

3 基于上下文和常识的讽刺识别模型

针对讽刺识别需要满足的上下文依赖和常识依赖，本文提出了基于上下文和常识的讽刺识别模型 CCSD。该模型可以有效对目标句与上下文之间的关系进行建模，并且引入了外部常识，从而提高了讽刺识别的准确率。CCSD 主要由 4 个模块组成：语义编码模块、注意力机制模块、门控机制模块和特征融合模块。语义编码模块负责将文本转化为特征嵌入，注意力模块关注特征中更加重要的信息，门控机制模块可以为目标句选择其所需要的特征信息，最后通过特征融合模块完成特征融合，实现讽刺识别。接下来将详细介绍问题定义以及模型的各个组成部分，模型结构如图 2 所示。

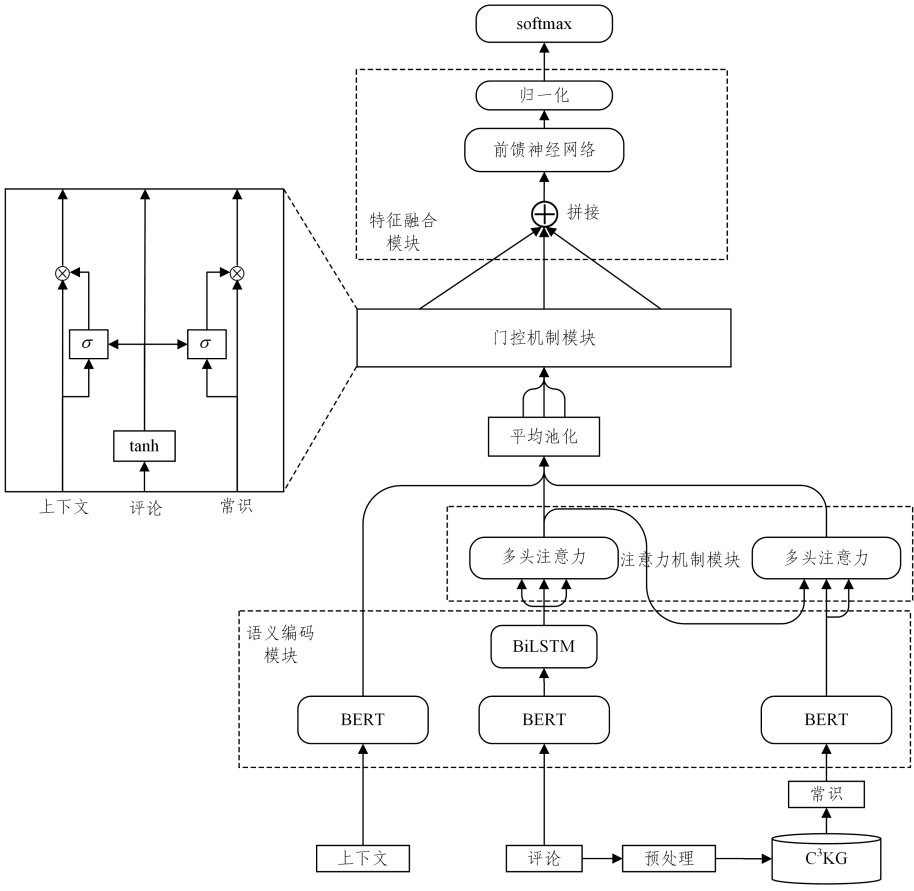

图 2 基于上下文和常识的讽刺识别模型结构

Fig. 2 Structure of sarcasm detection model based on context and common sense

3.1 问题定义

面向话题的讽刺识别旨在识别特定话题下的帖文评论是否具有讽刺意义。形式上，给定一条数据 $D=\langle C,T\rangle$，其中，T 为目标句，即帖文评论；C 为话题上下文，即评论所属的帖文。面向话题的讽刺识别的目标为根据话题上下文 C 和外部常识，准确判断目标句 T 为讽刺或非讽刺。

3.2 语义编码模块

首先，将评论进行停用词过滤、分句和删除标点符号等预处理操作；然后，使用 Sentence-bert[21] 将处理后的目标句和 C^3KG 常识库中的头实体进行语义相似度计算，得到语义最相似的头实体。通过头实体查询尾实体，得到常识知识，再按照常识库中的权重从高到低将尾实体使用逗号拼接得到常识文本。图 3 展示了对该过程的可视化。

在语义编码模块，将得到的上下文文本、评论文本和常识文本分别输入预训练好的 BERT[22] 模型中，得到 3 种文本的语义嵌入向量。将得到的评论语义嵌入再输入双向长短期记忆网络中进一步融合上下文语义，使得每个字符的向量都融合了上下文信息。

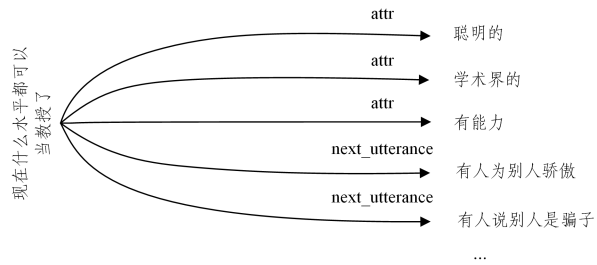

图 3 常识文本生成

Fig. 3 Generation of common sense texts

3.3 注意力机制模块

注意力机制模块中包含两次注意力机制的使用，分别是评论向量的多头自注意力机制和常识向量的多头注意力机制。

以下列举了一些互联网上人们常用的讽刺表达手法：

1）使用带有讽刺意味的词语，如“沽名钓誉”“蛇鼠一窝”“呵呵”；

2）使用特殊标点符号，如双引号和书名号；

3）使用谐音梗和网络流行语，如“真刑啊”代替“真行啊”和“真可铐啊”代替“真可靠啊”。

当出现这些讽刺的表达手法时，往往不需要上下文的参与便能判断该评论包含讽刺信息。因此，在评论向量中引入自注意力机制有助于模型更加关注评论中有用的词语信息，忽略其他无关紧要的表达。而使用多头注意力机制是希望模型可以基于相同的注意力机制学习到不同的信息，然后将不同的信息作为知识组合起来，捕获序列内各种范围的依赖关系。

常识文本是由常识库中多个尾实体拼接而成的文本，但并非每个常识都一定能帮助模型进行讽刺识别。因此作用于常识向量的多头注意力机制旨在让模型更关注有用的常识信息而忽略不相关的常识信息。多头注意力机制实现如下：

$$h_i=\alpha(QW_i^Q,KW_i^K,VW_i^V) \tag{1}$$

$$\alpha(q,k,v)=\mathrm{softmax}\left(\frac{qk^{\mathrm{T}}}{d_k}\right)v \tag{2}$$

$$F=[h_1\oplus h_2\oplus\cdots\oplus h_n]W^O \tag{3}$$

其中，$\boldsymbol{Q}\in R^{lq*d}$ 代表 query 向量，lq 为 query 的长度，d 为字向量维度；$\boldsymbol{K}\in R^{lk*d}$ 代表 key 向量，lk 为 key 的长度；$\boldsymbol{V}$ 代表 value 向量，lv 为 value 的长度；$\boldsymbol{W}_i^Q\in R^{d*\frac{d}{n}}$，$\boldsymbol{W}_i^K\in R^{d*\frac{d}{n}}$，$\boldsymbol{W}_i^V\in R^{d*\frac{d}{n}}$ 分别为 query，key 和 value 的可学习参数矩阵；h_i 代表第 i 个头输出的特征，n 为多头注意力的头数；$\alpha(q,k,v)$ 为注意力打分函数；$\oplus$ 为向量拼接操作，$\boldsymbol{W}^O\in R^{d*d}$ 为输出的可学习参数矩阵。

3.4 门控机制模块

大多数情况下，仅凭评论本身不足以判别讽刺信息，根据讽刺识别的上下文依赖和常识依赖可知，往往需要结合评论、上下文和常识 3 个特征才能进行讽刺判定。但对于不同的文本，每种特征对预测的贡献大小往往不同。例如同一句话在不同上下文中出现，但标签不一致的情况。因此，模型引入了门控机制模块，让模型动态地从 3 种特征中选择所需要的信息。

首先，将语义编码模块输出的上下文特征和注意力机制模块输出的评论特征和常识特征执行平均池化的操作，将字符向量转化为句向量。然后，将评论特征输入 tanh 激活函数中激活。最后，将激活后的评论特征与上下文特征和常识特征分别输入门控单元中，使用 sigmoid 激活函数计算上下文和常识的权重系数，将得到的系数与特征相乘，再经过 tanh 激活函数得到最终上下文和常识特征。具体操作如下：

$$H_{\mathrm{comment}}=\tanh(F_{\mathrm{comment}}) \tag{4}$$

$$b_{\mathrm{context}}=\mathrm{sigmoid}([F_{\mathrm{context}}+H_{\mathrm{comment}}]W_{g1}) \tag{5}$$

$$b_{\mathrm{common}}=\mathrm{sigmoid}([F_{\mathrm{common}}+H_{\mathrm{comment}}]W_{g2}) \tag{6}$$

$$H_{\mathrm{comtext}}=\tanh(b_{\mathrm{context}}\cdot F_{\mathrm{context}}) \tag{7}$$

$$H_{\mathrm{common}}=\tanh(b_{\mathrm{common}}\cdot F_{\mathrm{common}}) \tag{8}$$

其中，$H_{\mathrm{comment}}\in R^d$ 为激活的评论特征；$b\in(0,1)$ 为经过 sigmoid 函数计算的特征权重，$W_g\in R^{d*d}$ 为门控单元的可学习参数，$H_{\mathrm{context}}\in R^d$ 为激活后的上下文特征；$H_{\mathrm{common}}\in R^d$ 为激活后的常识特征。

3.5 特征融合模块

特征融合模块旨在将评论、上下文和常识特征进行拼接融合，并且进一步提取深层次语义特征。

首先，将门控机制模块输出的 3 个特征进行拼接，得到初步融合的特征；其次，将融合的特征输入线性层中进行维度转换；

然后，将转换后的特征输入前馈神经网络。前馈神经网络由两个线性层和一个 ReLU 激活函数组成。

$$FFN(x)=\mathrm{ReLU}(xW_1+b_1)W_2+b_2 \tag{9}$$

其中，$\boldsymbol{W}_1\in R^{d*d}$ 和 $b_1\in R^d$ 分别为第一个线性层的权重矩阵和偏置，$\boldsymbol{W}_2\in R^{d*d}$ 和 $b_2\in R^d$ 分别为第二个线性层的权重矩阵和偏置。

将前馈神经网络输出的结果进行层归一化操作，保证数据分布的稳定性，再和前馈神经网络的输入进行残差连接，进一步确保模型性能。

预测部分由线性层和 softmax 函数组成，先使用线性层将特征维度映射为 2 维，再使用 softmax 函数进行分类，此过程可以描述为：

$$\hat{y}=\mathrm{softmax}(HW_p+b_p) \tag{10}$$

其中，$H\in R^d$ 为融合后的特征，$\hat{y}$ 为预测为讽刺文本的概率。

本文采用交叉熵损失函数对模型进行优化：

$$Loss=-\sum_{i=1}^{N}[y_i\log\hat{y}_i+(1-y_i)\log(1-\hat{y}_i)] \tag{11}$$

其中，N 为训练数据大小，y_i 为样本 i 的真实标签。

4 实验

本章首先介绍面向话题的讽刺识别数据集、评估指标以及实验参数设置；然后介绍基线模型，并比较本文模型和基线模型的性能；再通过一系列消融实验来验证模型各个模块的有效性；最后对实验进行分析，包括模型分析以及错误案例分析。

4.1 面向话题的讽刺识别数据集

目前，国外有较多规模大、质量高的公开讽刺识别数据集，但权威的中文讽刺识别公开数据集较少，且国内面向话题的讽刺识别公开数据集更为稀缺。为此，本文参照 gong 等[22]构建数据集的规范，基于中文社交媒体平台，构建一个面向微博话题的讽刺识别数据集。微博作为国内最大的用户关系社交媒体平台之一，集成了社会热点话题、用户实时评论和用户间传播互动等功能。本文选取微博作为数据来源，为了保证语料库的全面性以及讽刺数据的占比，首先筛选出近年来较为讽刺的热点事件，包括社会、政治、娱乐、农业、医疗、经济、环保、体育和教育等领域的事件；然后采集这些热点话题帖文以及帖文的一级评论作为初始语料库。其中，帖文作为面向话题的讽刺识别中的话题信息，评论作为讽刺识别的目标句。接下来重点介绍数据处理和数据标注部分。

4.1.1 数据处理

考虑到数据集后续研究的可扩展性和可挖掘性，对采集到的数据执行以下输出处理的操作：

1）删除帖文中不影响语义的敏感词汇，以及包含敏感词汇的评论。

2）删除同一帖文下的相同评论。

3)过滤帖文和评论中的特殊符号、网页地址等不包含语义的字符。

4)过滤字符长度小于3的评论。

以上操作可以初步提高语料库质量,并减少数据标注的工作量。每条数据样例都由一条帖文和一条帖文评论组成。

4.1.2 数据标注

在数据标注过程中,为了保证数据的高质量,每一条数据都由3位标注者独立标注,标注者为具备深度学习基础的本科生和研究生。本数据集共有10位标注者,每位标注者负责标注5000条数据样本,历时一个月完成。对于标注不一致的数据,选取票数更高的标签作为最终标签,并在过程中剔除清洗不到位的数据。对于标注者无法根据帖文和评论确定标签的数据,则将该数据丢弃。

为了确保数据集中每个话题的可用性以及数据集的数据量,本文在已标注好的数据中随机抽取,并使每个话题的讽刺样本或非讽刺样本数最多不超过该话题数据量的70%,最后得到一个面向微博话题的讽刺识别(Weibo Topic Oriented Sarcasm Detection, WTO)数据集。该数据集包括WTO-V1和WTO-V2两部分。WTO-V1包含20566条数据,按6∶2∶2的比例划分为训练集、验证集和测试集,其中验证集和测试集的话题上下文会在训练集中出现。WTO-V2包含1516条数据,其中每条数据的话题上下文均不在WTO-V1中,该数据集仅用来测试,以验证模型在面对未知话题的泛化性。WTO数据集的数据分布占比如表1所列。

表1 WTO数据集的数据分布情况

Table 1 Data distribution of WTO dataset

样本数	WTO-V1 样本数	WTO-V1 样本占比/%	WTO-V2 样本数	WTO-V2 样本占比/%
讽刺样本数	13653	66.39	1006	66.36
非讽刺样本数	6913	33.61	510	33.64
总计	20566	100.00	1516	100.00

将最后的数据集按照所属话题进行统计,不同类别话题占比如图4所示。可以发现,社会、娱乐、教育这3类话题的讽刺话题数占比较大。图5和图6分别展示了数据集中评论和帖文长度的分布统计,评论文本长度的中位数位于11～20个字符之间,话题文本长度的中位数位于151～200个字符之间,符合微博文本的贴特点。

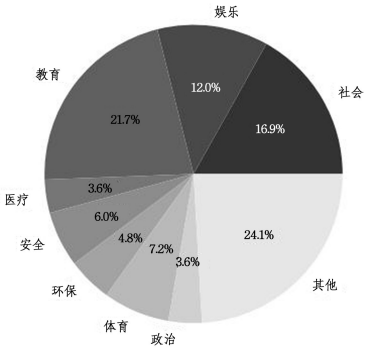

图4 各话题占比

Fig. 4 Proportion of each topic

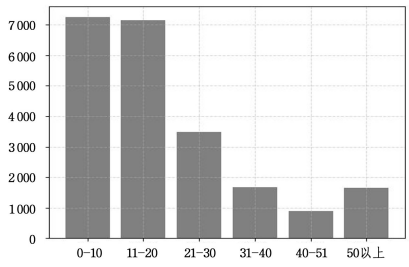

图5 评论字符长度分布

Fig. 5 Distribution of comment length

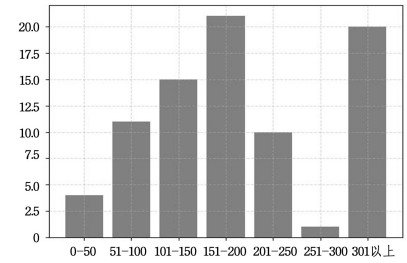

图6 帖文字符长度分布

Fig. 6 Distribution of post length

4.1.3 ToSarcasm数据集[1)]

本文在两个数据集上评估所提出的模型:1)本文构建的WTO数据集;2)ToSarcasm数据集[9]。ToSarcasm数据集包含4871个话题-评论对组成的样本,其中话题707个。数据集在各类别下的样本分布情况如表2所列。

表2 ToSarcasm数据集的数据分布

Table 2 Data distribution of ToSarcasm dataset

	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	100.00

4.2 评估指标

本文通过准确率(Accuracy, Acc)、精确率(Precision, P)、召回率(Recall, R)和F1值(F1-score, $F1$)来评估模型有效性,定义如下:

$$Acc=\frac{TP+TN}{TP+TN+FP+FN} \tag{12}$$

$$P=\frac{TP}{TP+FP} \tag{13}$$

$$R=\frac{TP}{TP+FN} \tag{14}$$

$$F_1=\frac{2\times P\times R}{P+R} \tag{15}$$

其中,TP表示预测类别和真实类别均为"讽刺"的数量;TN表示预测类别和真实类别均为"非讽刺"的数量;FP表示预测类别为"讽刺"、真实类别为"非讽刺"的数量;FN表示预测类别为预测类别为"非讽刺"、真实类别为"讽刺"的数量。

4.3 基线模型

本文将CCSD模型与以下基线模型进行对比:

1)BERT:BERT是Devlin等[23]提出的预训练语言模型,在许多NLP任务中取得了出色的效果。

1) https://github.com/HITSZ-HLT/ToSarcasm

2）KC-ISA：KC-ISA 是 Xu 等[24]提出的一种隐式情感分析模型，其使用共同注意力机制来整合外部常识和上下文特征。其中，外部常识为知识图谱。讽刺是隐式情感的一种，所以该模型也适用于讽刺识别任务。本文将 KC-ISA 模型与 CCSD 模型进行比较，目的是区分共同注意力机制与 CCSD 中的注意力机制模块和门控机制模块对面向话题的讽刺识别的贡献。

3）KL-BERT：KL-BERT 是 Li 等[25]提出的基于 BERT 的讽刺识别模型，该模型在讽刺识别中融合了常识知识。

4）TOSPrompt：TOSPrompt 是基于提示学习和大规模预训练语言模型提出的一种面向话题的讽刺表达提示模型。

4.4 实验参数设置

实验使用 PyTorch 深度学习框架实现，在 RTX 3090 GPU 上运行，使用随机梯度下降优化模型。为了提升优化的效率，采用不同参数设置不同学习率的策略。对预训练的 BERT 模型，学习率设置为 5×10^{-5}；对其他参数，学习率为5×10^{-3}。考虑到数据集短文本的限制，将目标句和常识文本的最大长度设置为 64，将话题文本（上下文）的最大长度设置为 256。基线方法的参数设置均为原论文中默认的最优设置。TOSPrompt 采用的模板为原论文中效果最优的模板：

$$x_{\text{prompt}}=s \text{ 是对 } t \text{ 的讽刺吗？[MASK]} \tag{16}$$

其中，s 代表评论，t 代表话题。最后使用模型生成的标签词是/否来作为最后的结果。

4.5 实验结果

为了评估 CCSD 模型在面向话题的讽刺识别任务中的有效性，本文在 WTO 和 ToSarcasm 数据集上和基线模型进行对比实验，实验结果如表 3 所列，粗体表示最优结果。可以观察到，本文模型在所有数据集上均取得了最佳性能。具体而言，在数据集 WTO-V1，WTO-V2 和 ToSarcasm 上，CCSD 模型的 F1 值较经过微调的 BERT 模型分别提高了 10.57%，2.73%和 5.34%；与引入常识知识的 KL-BERT 模型相比，F1 值分别提高了 2.75%，1.38%和 5.14%；与引入上下文信息的 TOSPrompt 模型相比，F1 值分别提高了 13.99%，1.58%和 2.8%。以上数据说明引入上下文信息和常识知识均可以提高面向话题的讽刺识别的效果，也证明了本文提出的上下文依赖和常识依赖成立。并且与同样引入了上下文信息和常识知识的 KC-ISA 模型相比，CCSD 模型的 F1 值在 3 个数据集上分别提高了 1.66%，9.8%和 5.09%。这可能是因为 KC-ISA 使用 TransE[26]模型训练的嵌入作为常识信息，但 TransE 模型在处理常识库这种复杂的知识图谱方面的能力比较欠缺，也无法表达常识的语义信息。这说明本文提出的 CCSD 模型能有效利用上下文信息和常识知识来完成面向话题的讽刺识别任务。

表 3 模型在面向话题的讽刺识别任务上的性能

Table 3 Performance of CCSD in topic-oriented irony recognition task

（%）

Method	WTO-V1				WTO-V2				ToSarcasm			
	Acc	*P*	*R*	*F*1	*Acc*	*P*	*R*	*F*1	*Acc*	*P*	*R*	*F*1
BERT	78.70	76.63	75.27	75.83	77.21	74.44	74.28	74.36	68.35	65.79	66.02	65.89
KC-ISA	86.53	86.10	83.79	84.74	74.83	74.03	66.20	67.29	70.40	67.63	65.62	66.14
KL-BERT	85.02	83.20	84.20	83.65	79.14	77.12	74.84	75.71	69.17	66.35	65.91	66.09
TOSPrompt	77.30	77.09	71.06	72.41	79.85	**79.11**	74.00	75.51	71.06	68.81	68.51	68.43
CCSD	**87.70**	**86.61**	**86.20**	**86.40**	**80.03**	76.89	**77.62**	**77.09**	**71.94**	**70.92**	**71.55**	**71.23**

为了验证模型的泛化性，使用在 WTO-V1 中训练好的模型在未知话题的 WTO-V2 数据集上实现面向话题的讽刺识别。除了 TOSPrompt 模型外，其他模型的效果在 WTO-V1 数据集上的效果均不如本文方法。这表明 TOSPrompt 模型使用的基于预训练 BERT 模型的提示学习方法及其模板具有较强的泛化性。同样只使用了目标句作为输入的 BERT 模型，因不受上下文和常识知识的影响，F1 值也仅下降 1.47%。受常识知识影响的 KL-BERT 模型，其 F1 值下降 7.94%。引入上下文信息和常识知识的 KC-ISA 模型和 CCSD 模型的 F1 值分别下降 17.45%和 9.31%，这一方面是因为模型参数较多且数据集规模不大，使得模型在 WTO-V1 上训练出现过拟合，另一方面是因为这两个模型均受到上下文信息和常识知识的影响。但 CCSD 模型在 WTO 数据集上始终表现出优于基线模型的性能。综合来看，本文提出的 CCSD 模型不仅能够在已知话题的数据集上达到优秀的效果，并且还能迁移到未知话题的数据，实现面向话题的讽刺识别。

KC-ISA 和 TOSPrompt 模型均直接或间接使用了话题上下文和常识知识。在输入信息相同的情况下，CCSD 模型的效果更优，这表明模型中的注意力机制模块和门控机制模块可以自适应学习到更有用的信息。

4.6 消融实验

为了进一步验证上下文和常识信息对面向话题的讽刺识别的作用，以及 CCSD 模型各个模块的有效性，本节进行了消融实验。首先，将常识特征删除得到 CCSD(-common)，以研究常识知识对讽刺识别的作用；其次，删除上下文特征得到 CCSD(-context)，以研究上下文的有效性；然后分别移除注意力机制模块中的上下文和常识注意力机制模块得到 CCSD(-att1)和 CCSD(-att2)，以研究注意力机制是否能有效帮助模型关注有用信息；再移除常识知识和上下文信息以及常识注意力机制得到 CCSD(att1)，即单个 BERT 和多头注意力机制的组合；最后，删除门控机制模块，以研究其是否能帮助模型动态选择不同信息。消融实验结果如表 4 所列。

结果表明，与 BERT 模型相比，融合上下文信息或者

常识知识均可提高面向话题的讽刺识别的效果，但其效果不如同时融合了上下文与常识的 CCSD 模型。同样，CCSD(-att2)模型效果优于 CCSD(att1)，这也表明了上下文依赖的重要性。单 BERT 和多头注意力的 CCSD(att1)在 3 个数据集上的 F1 值分别降低了 3.83%，3.66%和 4.16%，这表明常识和话题上下文的组合可进一步提高模型对讽刺信息的判别能力。取消上下文的注意力机制，模型在 3 个数据集上的 F1 值分别下降 1.99%，2.88%和 3.06%；取消常识的注意力机制，模型在 3 个数据集上的 F1 值分别下降 2.59%，1.42%和 4.08%。这表明注意力机制模块可以有效保留有用信息，充分利用上下文和常识知识。取消门控机制模块，模型在 3 个数据集上的 F1 值分别下降 2.49%，1.97%和 3.87%，这表明门控机制模块有助于模型动态选择 3 种信息。

表 4 消融实验结果

Table 4 Results of ablation experiment

(%)

Method	WTO-V1		WTO-V2		ToSarcasm	
	Acc	*F1*	*Acc*	*F1*	*Acc*	*F1*
CCSD	**87.70**	**86.40**	**80.03**	**77.09**	71.94	**71.23**
CCSD(-common)	86.25	84.33	77.41	73.00	69.65	67.41
CCSD(-context)	85.30	83.80	76.88	70.87	**72.35**	67.68
CCSD(-att1)	85.98	84.41	78.07	74.21	69.68	68.17
CCSD(-att2)	85.54	83.81	78.67	75.67	71.43	67.15
CCSD(att1)	84.98	82.57	77.94	73.43	69.84	67.07
CCSD(-gate)	85.01	83.91	79.12	75.12	71.53	67.36

CCSD(-context)模型和 KL-BERT 模型均只采用评论本身和常识信息。在 WTO-V1 和 ToSarcasm 数据集上，CCSD(-context)的效果更好，但在 WTO-V2 数据集上 KL-BERT 的效果更优。这可能是因为 WTO-V2 数据集的语料从未在训练集中出现，使得常识特征变得尤其重要，而 KL-BERT 所选择的常识库以及常识知识的选择策略使得模型在未知数据上的泛化性能更强。

4.7 模型分析

尽管面向话题的讽刺识别存在上下文和常识依赖，但表 4 的结果表明，在两个数据集上，仅包含评论注意力机制的模型的效果仍优于 BERT 模型。因此本节对评论注意力机制的结果进行可视化，进一步分析该模块的作用，并对 CCSD 的错误案例进行总结，得到文本讽刺识别的难点以及未来可能的改进方向。

4.7.1 评论注意力模块分析

CCSD 模型的评论注意力模块采用多头自注意力机制实现，不同的注意力头关注目标句的不同方面，以形成目标句的整体语义。图 7 为目标句“吃得像个丧尸一样”其中一个头的权重分布。颜色越深表示模型对该字符的关注度越高，颜色越浅表示关注度越低。权重热度图显示，模型的关注重点在于目标句的“丧尸”二字，这两个字用于形容人时具有强烈的讽刺意味。因此，CCSD 中的评论注意力模块更加关注目标句中具有讽刺含义的词汇，有助于讽刺的识别。

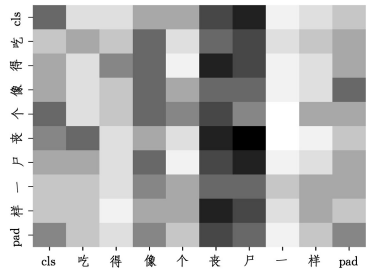

图 7 权重分布

Fig. 7 Weight distribution

4.7.2 错误案例研究

由表 5 中的结果可知，面向话题的讽刺识别难度较高，但其在情感分析和社交网络等方面具有重大研究价值。因此，本小节对错误预测的样例进行研究，并将其分为 3 类，以进一步分析讽刺识别的挑战，分类结果如表 5 所列。

表 5 错误案例分析

Table 5 Error case analysis

类别	错误案例信息
话题上下文过长	目标句：工作太认真了 话题上下文：……小区业主捐 100 万元附近大同小学，校门口保安拦住……
常识或背景知识缺失	目标句：996，福报啊 话题上下文：#女子求职开口问月休被 HR 怼#：不问薪资待遇，一来就问休息？还是在家休息吧…… 常识：聪明的，满怀希望的，救援到达，得到帮助，不幸的，倒霉的……
使用谐音梗、流行语	目标句：她只是犯了所有女孩子都会犯的错误，你们为什么要这么对她 话题上下文：#娄底警方通报撞人拖行事件#9 月 4 日……

对于表中第一个案例，尽管目标句与上下文的情感存在一定差异，但该案例的话题上下文较长，导致得到的上下文嵌入包含的语义不明确，无法保留重要信息；对于表中第二个案例，目标句中的“996”是近几年的时代热词，而目标句本身可以代表某一社会热点事件，但由于常识库中缺乏该类常识和背景知识，使得最后模型辨别失败；对于表中第三个案例，该目标句属于最近网络上表达讽刺的流行句式，模型在预训练和微调阶段不包含该类语料，导致模型分类错误。而谐音梗作为热门的网络表达，同样也加大了讽刺识别的难度。

针对第一个问题，可以通过缩短上下文，必要情况下对较长话题进行人工修改，也可以通过改进目标句与上下文特征的融合方式来解决；针对第二个问题，则需要一个不断更新的知识库来补充常识知识；针对第三个问题，可以在模型预训练或微调过程中加入大量的网络流行语和谐音梗来训练模型，以提升对该类目标句的讽刺识别效果。

结束语 针对面向话题的讽刺识别中未同时考虑上下文和常识的问题，本文提出了 CCSD 模型，该模型能有效整合上下文和常识特征。该模型利用 C^3KG 常识库来生成常识文本，该常识库整合了事件和对话流的常识信息。为了更加关注目标句中的重点信息和常识文本中有用的常识，本文设计了注意力机制模块。在此基础上，为了实现不同特征的合理

利用，设计了门控机制模块，让模型动态地从 3 种特征中选择所需要的信息。最后，结果特征融合模块将 3 种特征进行融合，并进一步提取特征，实现面向话题的讽刺识别任务。为了验证本文方法的有效性，在两个数据集上进行实验。与基线方法相比，本文提出的 CCSD 模型的性能更优。

参 考 文 献

[1] PANG B, LEE L. Opinion mining and sentiment analysis[J]. Foundations and Trends © in information retrieval, 2008, 2(1/2): 1-135.

[2] YI T, LUU A T, SIU C H, et al. Reasoning with Sarcasm by Reading In-between[C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Australia: Association for Computational Linguistics, 2018: 1010-1020.

[3] KUMAR A, NARAPAREDDY V T, SRIKANTH V A, et al. Sarcasm detection using multi-head attention based bidirectional LSTM[J]. IEEE Access, 2020, 8: 6388-6397.

[4] BAMMAN D, SMITH N. Contextualized sarcasm detection on twitter[C]// Proceedings of the International AAAI Conference on Web and Social Media. California: AAAI Press, 2015, 9(1): 574-577.

[5] RAJADESINGAN A, ZAFARANII R, LIU H. Sarcasm detection on twitter: A behavioral modeling approach[C]// Proceedings of the Eighth ACM International Conference on Web Search and Data Mining. New York: Association for Computing Machinery, 2015: 97-106.

[6] JOSHI A, SHARMA V, BHATTACHARYYA P. Harnessing context incongruity for sarcasm detection[C]// Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. Beijing: Association for Computational Linguistics, 2015: 757-762.

[7] SANTIAGO C, DEVAMANYU H, VERÓNICA P, et al. Towards Multimodal Sarcasm Detection (An _Obviously_ Perfect Paper)[C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: Association for Computational Linguistics, 2019: 4619-4629.

[8] PAN H, LIN Z, FU P, et al. Modeling intra and inter-modality incongruity for multi-modal sarcasm detection[C]// Findings of the Association for Computational Linguistics: EMNLP 2020. Online: Association for Computational Linguistics, 2020: 1383-1392.

[9] LIANG B, LIN Z, QIN B, et al. Topic-Oriented Sarcasm Detection: New Task, New Dataset and New Method[C]// Proceedings of the 21st Chinese National Conference on Computational Linguistics. Nanchang: Chinese Information Processing Society of China, 2022: 557-568.

[10] LI D, LI Y, ZHANG J, et al. C3KG: A Chinese Commonsense Conversation Knowledge Graph[C]// Findings of the Association for Computational Linguistics: ACL 2022. 2022: 1369-1383.

[11] GHOSH A, VEALE T. Fracking sarcasm using neural network[C]// Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. San Diego: Association for Computational Linguistics, 2016: 161-169.

[12] VAN HEE C, LEFEVER E, HOSTE V. Semeval-2018 task 3: Irony detection in english tweets[C]// Proceedings of The 12th International Workshop on Semantic Evaluation. New Orleans: Association for Computational Linguistics, 2018: 39-50.

[13] JOSHI A, TRIPATHI V, PATEL K, et al. Are Word Embedding-based Features Useful for Sarcasm Detection? [C]// Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin: Association for Computational Linguistics, 2016: 1006-1011.

[14] TAY Y, TUAN L A, HUI S C, et al. Reasoning with sarcasm by reading in-between[C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne: Association for Computational Linguistics, 2018: 1010-1020.

[15] HAZARIKA D, PORIA S, GORANTLA S. Cascade: Contextual sarcasm detection in online discussion forums[C]// Proceedings of the 27th International Conference on Computational Linguistics. Santa Fe: Association for Computational Linguistics, 2018: 1837-1848.

[16] KOLCHINSKI Y A, POTTS C. Representing social media users for sarcasm detection[C]// Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels: Association for Computational Linguistics, 2018: 1115-1121.

[17] BABANEJAD N, DAVOUDI H, AN A, et al. Affective and contextual embedding for sarcasm detection[C]// Proceedings of the 28th International Conference on Computational Linguistics. Barcelona (Online): International Committee on Computational Linguistics, 2020: 225-243.

[18] BI B, WU C, YAN M, et al. Incorporating External Knowledge into Machine Reading for Generative Question Answering[C]// Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Hong Kong: Association for Computational Linguistics, 2019: 2521-2530.

[19] LIU Z, XIONG C, SUN M, et al. Entity-duet neural ranking: Understanding the Role of Knowledge Graph Semantics in Neural Information Retrieval[C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne: Association for Computational Linguistics, 2018: 2395-1405.

[20] WANG Z, LI L, ZENG D. Knowledge-enhanced natural language inference based on knowledge graphs[C]// Proceedings of the 28th International Conference on Computational Linguistics. Barcelona: International Committee on Computational Linguistics, 2020: 6498-6508.

[21] REIMERS N, GUREVYCHI. Sentence-bert: Sentence embeddings using siamese bert-networks[C]// Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Hong Kong: Association for Computational Linguistics, 2019: 3982-3992.

[22] GONG X, ZHAO Q, ZHANG J, et al. The design and construction of a Chinese sarcasm dataset[C]// Proceedings of the Twelfth Language Resources and Evaluation Conference. 2020: 5034-5039.

[23] 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. Minneapolis: Association for Computational Linguistics, 2019: 4171-4186.

[24] XU M, WANG D, FENG S, et al. KC-ISA: An Implicit Sentiment Analysis Model Combining Knowledge Enhancement and Context Features[C]// Proceedings of the 29th International Conference on Computational Linguistics. Gyeongju: International Committee on Computational Linguistics, 2022: 6906-6915.

[25] LI J, PAN H, LIN Z, et al. Sarcasm detection with commonsense knowledge[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 3192-3201

[26] BORDES A, USUNIER N, GARCIA-DURAN A, et al. Translating embeddings for modeling multi-relational data[C]// Proceedings of the 26th International Conference on Neural Information Processing Systems. 2013: 2787-2795.

LIU Qilong, Born in 1999, postgraduate. His main research interests include natural language processing, network public opinion knowledge graph and graph neural network.

LI Bicheng, born in 1970, Ph.D, professor, Ph.D supervisor. His main research interests include intelligent information processing, network ideological security, network public opinion monitoring and guidance, and big data analysis and mining.

（责任编辑：何杨）