结合规则蒸馏的情感原因发现方法*

鲍建竹，蓝恭强，巫继鹏，徐睿峰

(哈尔滨工业大学（深圳） 计算机科学与技术学院，广东 深圳 518055)

摘要：现有的基于深度学习的情感原因发现方法往往缺乏对文本子句之间关系的建模，且存在学习过程不易控制、可解释性差和对高质量标注数据依赖的不足。针对以上问题，本文提出了一种结合规则蒸馏的情感原因发现方法。该方法使用层次结构的双向门限循环单元(Bi-GRU)捕获词级的序列特征和子句之间的潜层语义关系，并应用注意力机制学习子句与情感关键词之间的相互联系，同时结合相对位置信息和残差结构得到子句的最终表示。在此基础上，通过知识蒸馏技术引入逻辑规则，从而使该模型具有一定的可控性，最终实现结合逻辑规则的情感原因发现。在中文情感原因发现数据集上的实验结果显示，该方法达到了目前已知的最优结果，F1 值提升约 2 个百分点。

关键词：情感原因发现；注意力机制；门限循环单元；知识蒸馏

中图分类号：TP391 **文献标识码：**A

Emotion Cause Extraction Combined with Regular Distillation

BAO Jianzhu, LAN Gongqiang, WU Jipeng, XU Ruifeng

(School of Computer Science and Technology, Harbin Institute of Technology (Shenzhen), Shenzhen, Guangdong 518055, China)

Abstract: Existing deep learning-based emotion cause extraction methods often lack the ability of modeling the latent semantic relationship between clauses. Besides, these methods are totally uncontrollable, uninterpretable and highly dependent on high-quality annotation data. To solve the above problems, this paper proposes an emotion cause extraction method based on the rule distillation. In this method, a hierarchical bidirectional gated recurrent unit (Bi-GRU) is used to capture the sequence features of the word level and the latent semantic relationship between the clauses. The final representation of clauses is obtained by combining the position information and the residual structure. Besides, the attention mechanism is applied to learn the latent semantic relationship between the clause and the emotional expression. Furthermore, an adversarial-based knowledge distillation architecture is introduced to combine the logic rules and the deep neural network. The experimental results on the Chinese emotion cause extraction dataset show that proposed method outperforms the state-of-the-art by 0.02 in F1.

Key words: emotion cause extraction；hierarchical attention network；domain knowledge；knowledge distillation

* **收稿日期：** **定稿日期：**

基金项目：国家自然科学基金(U1636103，61632011，61876053)；深圳市基础研究项目(JCYJ20180507183527919，JCYJ20180507183608379)；深圳市技术攻关项目(JSGG20170817140856618)；深圳证券信息联合研究计划

作者简介：鲍建竹 (1996—)，男，硕士研究生，主要研究领域为情感分析；蓝恭强(1996—)，男，硕士研究生，主要研究领域为自然语言处理；巫继鹏(1993—)，男，博士研究生，主要研究领域为情感分析；徐睿峰(1973—)，男，教授，主要研究领域为自然语言处理，文本情绪计算，认知计算。

0 引言

随着互联网的飞速发展，Twitter、Facebook 和微博等社交网络被广泛使用，因而产生了大量蕴含丰富情感的文本内容，这为文本情感原因发现工作提供了大量的研究资源。

文本情感分析任务的主要目标是识别出文本中蕴含的情感倾向性，但是在实际应用中，往往需要挖掘更加丰富的情感信息。以舆情监督为例，只知道舆论的倾向性并不足以提供疏导民意的方案，如果能找到舆情的来源或原因，就能提出更有针对性的解决方案。情感原因发现这一任务的提出正是旨在解决对情感“追根溯源”的问题，即找到文本中触发情感表达的源头。下面给出一个情感原因发现的例子。

例：〈1〉上午 10 时，〈2〉李世铭一家闻讯后赶到侯马市公安局，〈3〉**门口看到失而复得的儿子**，〈4〉一家人忍不住泪流满面。

该段文本包括四个子句，子句以逗号、句号、分号等标点符号划分，其中粗体部分为情感原因，下划线部分为情感关键词，情感关键词所在句为情感中心句，情感原因发现任务的目标是在给定情感关键词的情况下识别包含情感原因的子句。例 1 中情感关键词为子句<4>中的“泪流满面”，情感原因为子句<3>，在其他样例中情感原因可能为多个子句。

目前，情感原因发现方法主要包括基于规则和统计机器学习的方法以及基于深度学习的方法。基于规则和统计机器学习的情感原因发现方法[1-3]的主要思路是利用一些语言学特征构建情感原因发现规则集，随后对文本中的原因事件进行定位，或者将规则匹配结果作为特征使用机器学习分类器分类。此类方法的缺点是需要人工构造大量规则集且制定的规则并不能完全覆盖所有的语言现象，而且规则的指定、特征的提取和筛选需要经验指导。同时，该方法还缺乏对序列关系的建模，忽略了上下文语义信息。基于深度学习的情感原因发现方法[4,5]使用深度神经网络学习词语、句子和篇章的向量化表示，并建模文本序列信息以及情感词和文本之间的关系。然而，深度神经网络的训练往往依赖于高质量标注数据，缺乏对人工规则集等资源的有效利用。从以往的研究工作中看到，仅在深度模型的基础上采用规则后处理的方法不仅增加了预测过程的计算复杂性，也很难取得理想的效果。然而，深度神经网络有着可解释性弱和可控性差等缺点。

对以上缺点，本文提出一种结合规则蒸馏的情感原因发现方法(Rule Distilled Hierarchical Attention Network, RD-HAN)。通过基于 Bi-GRU 的层次注意力网络模型对文本序列建模，既能从正向词序和逆向词序分别提取句子内部的语义信息，又能结合上下文句子的语义特征，得到包含丰富语义信息的文本编码。同时，通过注意力机制，在文本建模的过程中根据各部分与情感关键词的相关度赋予不同的注意力权重，从而使得重要的语义特征能够被有效利用。同时，在神经网络中引入情感原因发现早期研究中人工提取的专家规则，利用规则指导神经网络的学习，并利用知识蒸馏技术将逻辑规则迁移到网络参数中。实验结果显示，相对过去的基线模型，结合了位置向量和残差结构的层次注意力网络模型取得了明显的性能提升。在此基础上，引入规则蒸馏后的 RD-HAN 方法表现出了更好的性能。

本文将在第 1 节介绍情感原因发现和知识蒸馏的相关工作，在第 2 节介绍结合规则蒸馏的情感原因发现方法的模型结构和原理，第 3 节对实验数据集进行介绍，并分析实验结果，最后一节对本文的工作进行总结并展望。

1 相关工作

现有的情感原因发现方法包括基于规则集和机器学习的方法以及基于深度学习的方法。Lee 等[1]提出了一个情感原因发现人工规则集，并标注了第一个情感原因发现语料库。利用该规则集，Lee 等[2]还提出了

一种基于规则匹配的情感原因发现方法，并在上述语料上做了情感原因发现实验。以上基于规则集的情感原因发现方法具有很好的可解释性和可控性，但是需要人工构造规则集。Gui 等[3]将规则集的匹配结果作为部分特征，加上其他语言学特征使用多核 SVM 对情感原因子句进行分类，取得了不错效果。Gui 等[4]首次将深度学习方法引入情感原因发现问题，提出了一种新的基于卷积操作的深度记忆网络，对以后研究工作有很大启发。Ding 等[5]认为除了文本内容之外，相对位置和全局标签也是非常重要的信息，因此引入相对位置嵌入学习算法，很好的建模了各个子句之间的位置关系，在 Gui 等人提出的中文情感原因发现数据集[6]上达到了目前最优的性能。余传明等[7]提出了一种基于多任务学习的情感原因分析模型，将词性标注作为辅助任务引入情感原因识别任务，实验结果表明，采用多任务学习的策略可以有效地缓解数据不平衡的问题，表现出更好的性能。Hinton 等[8]在 2015 年提出了知识蒸馏(Knowledge Distillation)，这是一种将复杂网络模型学习到的参数迁移到简单模型的框架。Hu 等[9]将一阶逻辑作为后验正则化约束神经网络中文本的表示，并采用知识蒸馏技术将包含逻辑规则的结构信息迁移到网络权重中，并在情感分析和命名实体识别任务上进行了验证。Liu 等[10]提出了一种基于对抗模仿学习的知识蒸馏方法，首先构建一个教师网络从真实标注中学习知识表示，再利用对抗学习的方式训练一个学生网络去模仿教师网络的输出，从而得到一个只接受原始文本的事件抽取特征编码器。

现有的情感原因发现规则集中包含着专家通过观察总结得到的领域知识，而基于深度学习的情感原因发现方法可以从标注文本中学习语义表示。为了将人工规则集中的领域知识结合到深度学习模型中，使得模型具有一定的可解释性，提出了一种结合规则蒸馏的情感原因发现方法。该方法利用层次注意力网络捕获文本子句内部的语义信息以及子句-子句和子句-情感关键词之间的潜层语义关系，在此基础上借助知识蒸馏技术引入逻辑规则进行情感原因发现。

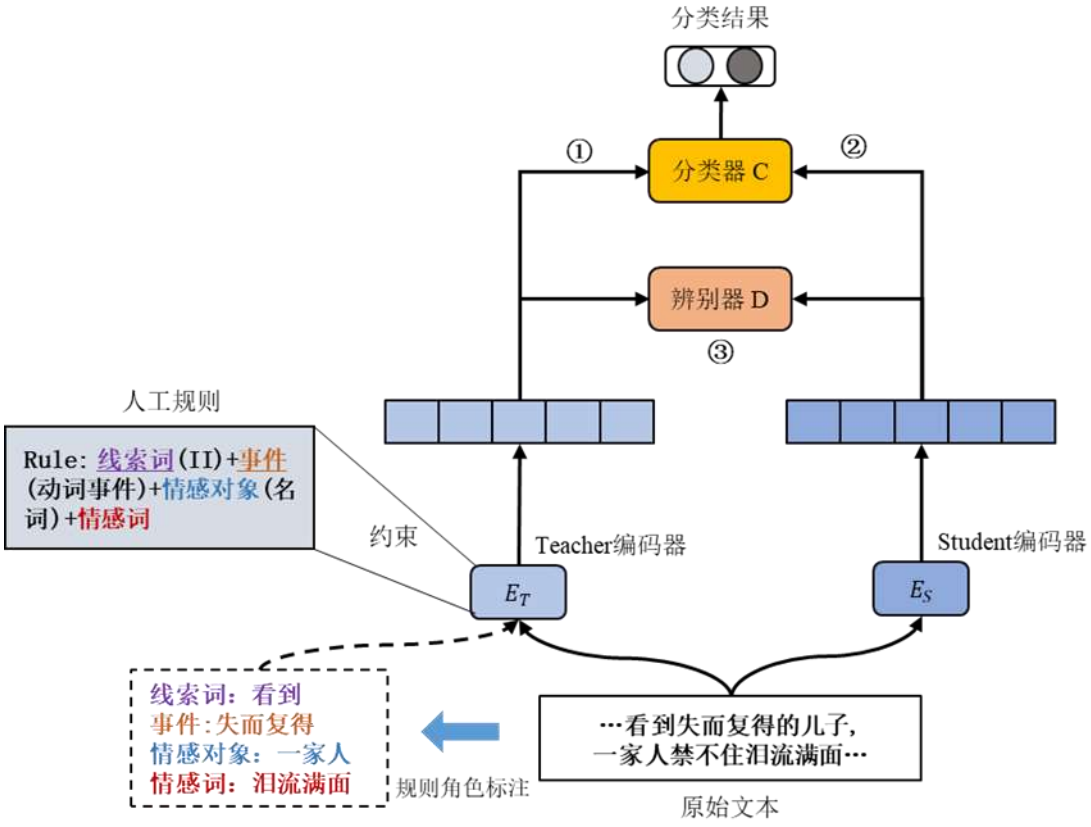

图 1 RD-HAN 情感原因发现方法框架图

2 模型框架

引言中说明了情感原因发现是一个子句级别的分类问题，现有的情感原因发现方法通常将子句看作独立个体，通过学习子句的语义表示及子句-情感关键词之间的语义关系对子句进行分类。此种方法忽视了子句级别的上下文语义信息，此外，基于深度神经网络的编码器存在依赖标注数据和可控性差的缺点。因此，本文提出了一种结合规则蒸馏的情感原因发现方法(RD-HAN)，该方法整体框架如图 1 所示。

该方法分为三个步骤：

1) 采用结合位置信息和残差结构的层次注意力编码器对输入文档进行语义编码和规则角色编码，并将语义编码与逻辑规则编码拼接，得到文本的最终向量表示。通过将文本的最终向量表示送入分类器可以得到预测概率分布。此外，为了得到教师编码器E_T和分类器 C，本文使用规则标签将概率分布映射到规则约束下的概率子空间，并使用真实标签和规则约束后的概率来训练模型。

2) 固定分类器，使用真实标签训练一个只接受文本输入的学生编码器E_S。

3) 交替训练辨别器 D 和学生编码器，辨别器的训练目标是分辨教师编码器和学生编码器的输出，而学生编码器的训练目标是尽量“骗过”辨别器，即让自己的输出尽量接近教师编码器的输出。

2.1 结合位置信息和残差结构的层次注意力编码器

图2给出结合位置信息和残差结构的层次注意力编码器的整体框架图。为了综合考虑子句与情感表达的语义信息以及上下文关系，本文使用层次 Bi-GRU 捕获词级的序列特征和子句之间的潜层语义关系。为了更好地建模情感关键词与其他词之间的关系，本文使用注意力机制学习子句与情感关键词之间的相互联系。在情感原因发现任务中，子句与情感中心句之间的相对位置是一个重要的信息，因此本文引入相对位置编码向量来获取子句与中心句之间的相对位置信息。此外，为了有效地结合句子本身的语义信息和上下文信息，本文采用残差结构来得到子句的最终表示。

词序列编码器：情感原因发现任务中，将情感关键词文本记作$u \in R^{q\times d}$，其中q为文本长度，d为词向量维度。情感原因子句记作$c_i = [w_{i,1}, w_{i,2}, \dots, w_{i,p}] \in R^{p\times d}$，其中i表示第i个子句，$p$为子句长度，$w_{i,j}$代表维度为$d$的词向量。词向量使用 Word2Vec 的 skip-gram 方法在中文情感原因发现数据集上进行预训练。然后，使用 Bi-GRU 网络建模，得到词序列的上下文表示h_i，如公式(1)所示。

$$\begin{aligned} h_i &= [h_{i,1}, h_{i,2}, \dots, h_{i,p}] \\ &= BiGRU([w_{i,1}, w_{i,2}, \dots, w_{i,p}]) \end{aligned} \tag{1}$$

词注意力：该部分的目的是使用注意力机制建模情感关键词与其他词之间的关系，以此来提取文档中每个子句的语义信息。在

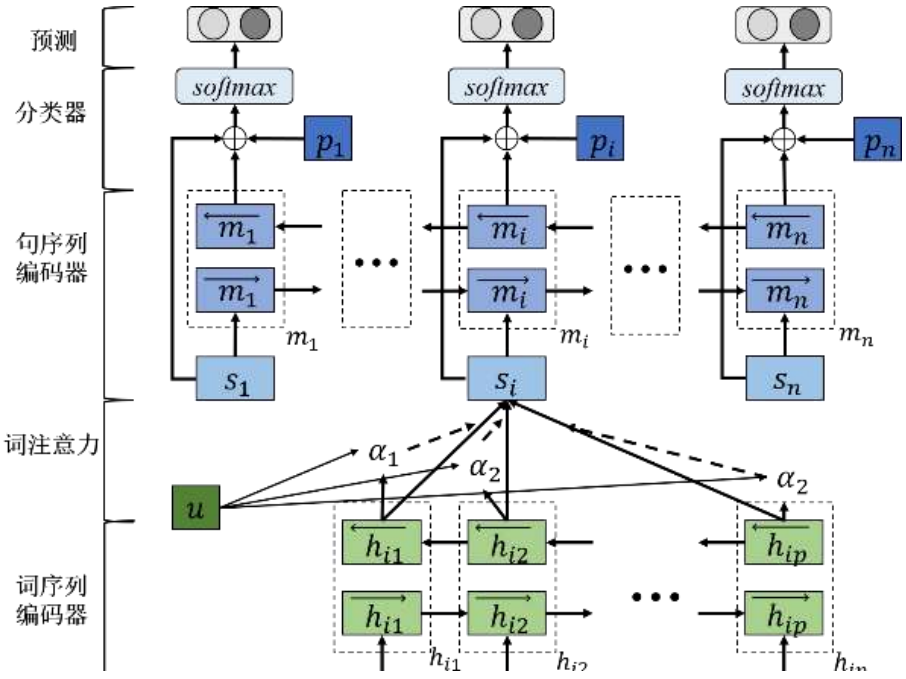

图 2 结合位置向量和残差结构的层次注意力编码器

Q-K-V 注意力机制中，情感表达文本为 Query，情感原因子句是 Key 和 Value，分别由 Bi-GRU 对原始输入编码得到。

$$Q = BiGRU(u) \quad (2)$$

$$K_{c_i} = V_{c_i} = h_i \quad (3)$$

然后，根据公式(4)得到子句中每个词语对情感表达的注意力权重α。

$$\alpha = Attention(Q, K_{c_i}, V_{c_i}) = \frac{QK_{c_i}^{T}}{\sqrt{d}} V_{c_i} \quad (4)$$

其中d是词向量的维度。最后通过公式(5)得到具有情感注意力倾向的子句语义表达s_i。

$$s_i = \alpha^T h_i = \sum_j \alpha_j h_{ij} \quad (5)$$

层得到最终的输出概率y_i，见公式(8)，其中W是权重矩阵。

$$y_i = softmax(W \cdot r_i) \quad (8)$$

最后，具有最高概率的标签作为最终的分类结果。

2.2 结合规则角色编码的文本表示方法

情感原因发现的早期研究方法主要基于人工提取的规则，通常一条包含情感表达和情感原因的文本包含以下几个重要元素：1)情感词 K，2)线索词 I，3)情感对象 E，4)情感原因事件 C。这些元素也称作规则角色。基于规则的情感原因发现方法就是通过情感词和线索词来定位情感对象并最终发现情感原因。

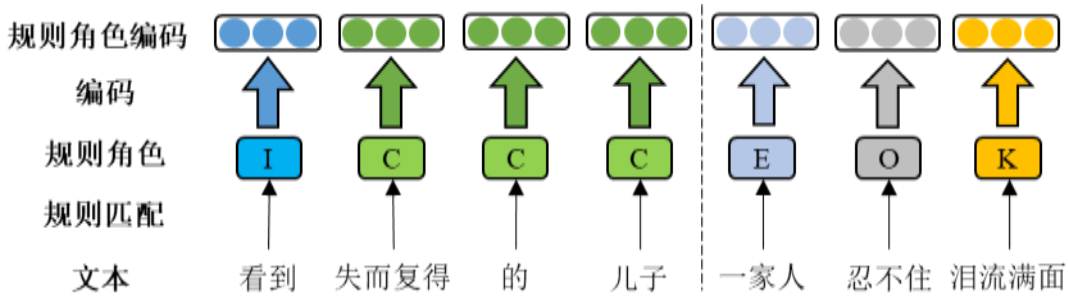

图 3 规则角色编码示例

句序列编码器：该层的目的是捕捉子句之间的上下文信息。本文使用 Bi-GRU 对句子向量序列建模，由公式(6)得到每个子句的上下文向量表示$[m_1, m_2, \dots, m_n]$，其中n为情感原因子句的数目。

$$[m_1, m_i, \dots, m_n] == BiGRU([s_1, s_2, \dots, s_n]) \quad (6)$$

由于上下文向量m_i同时包含了子句的语义信息和上下文信息，当文本序列较长时可能出现梯度消失或梯度爆炸的情况，所以将上下文向量m_i与子句的原始向量s_i拼接作为最终的语义向量。此外，在情感原因发现的相关研究中发现子句与情感关键词所在句的相对位置是非常重要的一个特征，因此将相对位置编码为p_i作为子句位置的向量表示。最终，将语义向量与位置向量拼接得到子句的表示r_i，见公式(7)。

$$r_i = m_i \oplus s_i \oplus p_i \quad (7)$$

分类器：文本情感原因发现本质上是一个分类任务，所以模型最后一层是分类器。情感原因发现任务中分类的目标是子句，所以直接将子句的表示经由分类器和 softmax

以图 3 为例，如果在情感词(泪流满面)的前面出现了线索词(看到)，在情感词与线索词之间存在名词情感对象(一家人)，则线索词与情感对象之间的事件(失而复得的儿子)为情感原因事件，包含该情感原因事件的子句为情感原因句。

基于规则的方法通过用各种规则去匹配文本来确定情感原因句，本文采用 Lee 等[1]提出的线索词和规则。以往的将神经网络与规则结合的方法一般是用规则对神经网络的预测结果进行后处理，以此来修正部分预测错误的标签。但是，由于规则匹配本身的准确率有限，这种方法可能会使模型总体性能下降。因此，本文不采用这种后处理的方式。

为了将规则嵌入到神经网络模型中，可以将规则角色进行编码。具体来说，对于某个文本，首先对其进行规则匹配，确定每个词对应的规则角色，然后对每种规则角色指定不同的编码，以此得到每个词的规则角色编码。

具体来说，假设用层次注意力模型得到的每个词的语义向量为x_{s_i}，用规则角色编码得到的规则向量为x_{r_i}，那么该词的最终向量表示为：

$$x_i = x_{s_i} \oplus x_{r_i} \tag{9}$$

其中，$\oplus$指对两个向量进行拼接，这两个向量都将在训练过程中更新。

2.3 基于对抗学习的知识蒸馏训练方法

假设有输入变量$x \in \mathcal{X}$和目标变量$y \in \mathcal{Y}$，由于情感原因发现是一个二分类问题，故$\mathcal{Y} = \triangle^2$是二维的概率空间，而$y \in \{0,1\} \in \mathcal{Y}$则是类别标签。在得到训练样本集合$D = \{(x_n, y_n)\}_{n=1}^{N}$的情况下，利用 2.1 节中基于层次注意力机制的文本编码器，通过有监督学习方法，可以得到一个分类模型$p_\theta(y|x)$。

假设有一组逻辑规则，定义为：$R = \{(R_l, \lambda_l)\}_{l=1}^{L}$，这里$R_l$是输入-目标空间$(\mathcal{X}, \mathcal{Y})$中的第$l$个规则，而$\lambda_l \in [0, \infty]$是第$l$个规则的置信度，$\lambda_l = \infty$表明这是一个强规则。假设给定一组样本$(X, Y) \subset (\mathcal{X}, \mathcal{Y})$，在一个神经网络中，通过 Softmax 层可以得到预测概率向量，记为$\sigma_\theta(x)$，其中θ为需要训练的参数。通常，神经网络训练过程是通过样本预测概率和真实标签迭代更新参数θ。为了引入规则标签信息。在训练过程中对网络输出施加规则标签约束，这使得原本的概率分布p_θ被映射到了受规则约束的子空间q。

为了编码器在学习的过程中自动学到规则知识，即将规则“消化”在模型参数中，借助知识蒸馏技术，将规则约束下训练得到的网络当做教师网络(Teacher Network)，用$f_\theta(x)$来表示其输出的概率预测向量，其中θ为训练参数。为了隐式地学到规则信息，需要训练一个学生网络(Student Network)，学生网络只接受语义向量输入而不需要额外的规则信息。传统的知识蒸馏方法使用 KL 散度来更新学生网络，但是在实际应用中，以 KL 散度约束的知识蒸馏更新方法往往难以取得期望的效果[11]，所以本文采用 Liu 等[10]提出的基于对抗的知识蒸馏训练方法，如图 1 所示。这一过程分为三个阶段：教师编码器构建阶段、学生编码器构建阶段和对抗学习阶段，其中前两个阶段为预训练阶段。

在预训练阶段，首先训练编码器$\mathrm{E_T}$和分类器C，损失函数如下：

$$\begin{aligned} L_T(\Theta) &= -\frac{1}{N}\sum_{n=1}^{N}(L_1 + L_2) \\ L_1 &= (1-\mu)\ell(y_n, f_{\Theta_T}(x_n)) \\ L_2 &= \mu\ell\left(q_n, f_{\Theta_T}(x)\right) \end{aligned} \tag{10}$$

其中Θ包含编码器E_T和分类器C的所有参数；ℓ为损失函数，本文中使用交叉熵损失函数；y_n是样本真实标签；q_n是规则约束后的概率子空间；μ是调和系数，用来平衡两个目标的重要度。这一步是对E_T和 C 同时训练。

接下来，将冻结分类器 C，然后连接编码器E_S和分类器 C，构建一个纯文本输入的情感原因发现模型。损失函数如下：

$$L_S(\Theta_S) = -\frac{1}{N}\sum_{n=1}^{N}\ell(y_n, f_{\Theta_S}(x_n)) \tag{11}$$

其中，Θ_S指学生网络编码器的参数，$f_S(x_n)$是学生网络 S 对样本的预测概率。

最后冻结编码器E_T和编码器E_S，将教师编码器E_S的输出当做正例，将学生编码器E_S的输出当做负例来训练辨别器 D。此时的交叉熵损失函数如下：

$$\begin{aligned} L_D(\Theta_D) = -\frac{1}{N}\sum_{n=1}^{N}[&\log(D(f_{\Theta_T}(x_n))) \\ &+ \log(1 - D(f_{\Theta_S}(x_n)))] \end{aligned} \tag{12}$$

其中，Θ_D是辨别器 D 的参数。在对抗训练的阶段，学生编码器E_S将要尽量“骗过”辨别器 D，即让自己的输出尽量接近教师编码器E_T的输出，而辨别器 D 则要尽量去辨别这两个编码器的输出。公式(11)为第二个环节的损失函数，而第一个环节的损失函数如下：

$$\begin{aligned} L_{adv}(\Theta_S) &= \frac{1}{N}\sum_{n=1}^{N}\log(1 - D(f_S(x_n))) \\ &= -\frac{1}{N}\sum_{n=1}^{N}\log(D(f_S(x_n))) \end{aligned} \tag{13}$$

而后将公式(11)与公式(13)联合，得到学生编码器E_S的损失函数为：

$$L_{S-adv}(\Theta_S) = L_S(\Theta_S) + \pi * L_{adv}(\Theta_S) \quad (14)$$

其中，π是超参数，用来平衡两个损失函数。

实际训练中，以往的实验表明过去常用的交替更新策略会产生不均衡的效果，比如辨别器性能太强，所以学生编码器E_S很难“骗过”它，导致训练无法进行。于是本文将采用动态适应的学习策略，学生编码器E_S正常迭代更新，同时预先设定一个判别器阈值，判别器 D 的参数仅在其判别准确率较低时更新。这样可以保证学生编码器不会因为判别器性能过高而无法学习参数。具体算法如下所示：

算法 1 动态适应学习算法
输入：训练数据$D = \{(x_n, y_n)\}_{n=1}^{N}$;预训练后的$E_T, E_S$，$D$，$C$
1 固定E_T和C
2 根据公式(14)更新E_S
3 如果D的准确率低于阈值则根据公式(12)更新D
4 重复 2-3 直到收敛
输出：对抗学习强化后的E_S

3 实验与分析

3.1 实验数据

本文使用的实验数据为中文情感原因发现数据集[6]。该数据集的统计信息如表 1 所示。

表 1 中文情感原因发现数据集总体统计

项目	数量(条)	占比
文档	2,105	-
子句	11,799	-
情感原因	2,167	-
含一条原因的文档	2,046	97.20%
含两条原因的文档	56	2.66%
含三条原因的文档	3	0.14%

该数据集共收集 2105 个文档，包含 11799 条子句，其中有 2167 条子句为情感原因所在句。绝大多数的文档只含有一条情感原因子句。

3.2 实验设置

本章提出模型的超参数设置如下：所有子句的文本截断长度为 41(如果过长则截断，过短则用“<PAD>”标签填充)；词向量由在该语料上使用Word2Vec中的skip-gram方法预训练的 300 维词向量初始化，在训练的初始阶段以 0.0001 的学习率微调；所有 Bi-GRU 的隐层维度均为 300 维。训练时，使用 Adam 优化器，学习率(learning rate, lr)设置为 3×10^{-4}，梯度以及梯度平方的运行平均值的系数(β)设置为(0.9,0.999)，数值计算稳定项(ϵ)为 1×10^{-8}，权重衰减系数(weight decay)为 1×10^{-5}。训练数据的批次大小(batch size)为 16，采用了提前停止策略(early stopping)，容忍度为 5 个训练周期(epoch)。

通过实验验证，在 Lee 等[1]提出的 15 条规则中选取了 8 条有效规则。规则置信度λ均设置为 0.95，在规则约束学习的对抗训练阶段，调和参数$\pi = 1 - 0.9^t$，其中t为训练轮次，初始值为 0.1，随着训练过程逐渐升高。

3.3 对比模型

为了验证模型的有效性，将提出的模型与如多种基准模型进行比较，具体如下：

RB(Rule-based method)：Lee 等在 2010 年工作中提出的基于规则匹配情感原因发现方法[1]。

CB(Commonsense-based method)：此方法为 Russo 等在 2011 年提出的基于情感常识库匹配的方法，Gui 等将情感认知词典引入此方法，并在中文情感原因发现数据集上进行了复现[6]。

RB+CB+ML(Machine Learning)：Gui 等在中文情感原因发现数据集上结合了基于规则匹配以及常识库匹配方法从数据中抽取特征，并结合机器学习算法进行分类[6]。

Multi-kernel：Gui 等使用基于多核(多项式核函数与修改后的卷积核函数以“相乘”方式结合)支持向量机分类器的事件驱动的极限方法[6]。

Memnet：Gui 等提出的基于层级注意力记忆网络进行情感原因发现方法[4]。

ConvMS-M：Gui 等提出的基于卷积的层级注意力记忆网络情感原因发现方法[4]。

PAE-DGL：Ding 等提出的结合相对位置和全局标签的情感原因发现方法[5]。

CANN：Li 等提出的基于协同注意力网络的情感原因发现方法[12]。

HAN: Yang 等提出的层次注意力网络[13]。

HAN(Ours)：本文提出的结合相对位置和残差结构的层次注意力网络模型。

RD-HAN：本文提出的以教师网络的输出和真实标签为学习目标基于对抗训练得到的学生网络模型。

3.4 实验结果及分析

本文实验均取 90%的数据作为训练集，10%的数据作为测试集，取 25 次独立实验结果的平均值作为最终的实验结果，对比实验结果与分析如下：

不同方法性能对比：首先比较不同方法在中文情感原因发现数据集上的实验结果，如表 2 所示。

表 2 不同方法性能对比

模型	P	R	F1
RB	0.6712	0.5247	0.5890
CB	0.2672	**0.7130**	0.3887
RB+CB+ML	0.5921	0.5307	0.5597
Multi-kernel	0.6588	0.6927	0.6752
Memnet	0.5922	0.6354	0.6131
ConvMS-M	0.7076	0.6838	0.6955
PAE-DGL	0.7619	0.6908	0.7242
CANN	**0.7721**	0.6891	**0.7266**
HAN	0.7232	0.6605	0.6904
HAN(Ours)	**0.7771**	0.6897	0.7304
RD-HAN	0.7706	**0.7203**	**0.7446**

可以看到，本文提出的 RD-HAN 模型在准确率、召回率和 F1 值上都明显超过了传统的基于规则和机器学习的方法，也超过了大多数的深度学习方法。相对于基于规则的方法(RB)，该模型的 F1 值高出了约 15 个百分点；相对于多核 SVM(Multi-kernel)的方法，该模型的 F1 值高出将近 7 个百分点。与目前已知的在中文情感原因发现数据集上达到当前最优(state-of-the-art)性能的协同注意力模型 CANN 相比，该模型在准确率、召回率和 F1 值指标上均有提升。这说明，本文提出方法在融合了规则知识和文本表示的情况下，取得了比只基于规则或文本表示学习的方法更优的性能。

与当前最优的 CANN 模型相比，本文提出的基模型 HAN(Ours)在准确率、召回率和 F1 值上均有提升，在此基础上，结合规则蒸馏的 RD-HAN 模型由于规则标签的引入使得准确率有所下降(不到 1 个百分点)，但是同时将召回率提升了约 3 个百分点，最终在 F1 值上提升了约 2 个百分点。这说明了本文提出的结合位置向量和残差结构的层次注意力网络在情感原因发现上有不错的表现，而基于对抗的规则蒸馏方法又通过引入逻辑规则带来了进一步的性能提升。

不同规则引入方式的对比：为了评估本文提出的规则蒸馏方法的有效性，在基模型 HAN(Ours)的基础上对不同规则的引入方式进行对比实验，实验结果如表 3 所示。

表 3 不同规则引入方式对比

模型	P	R	F1
HAN(Ours)	**0.7771**	0.6897	0.7304
HAN(Ours)+P	0.7276	**0.7208**	0.7242
RD-HAN(T)	0.7665	0.7158	0.7403
RD-HAN(S)	0.7706	0.7203	**0.7446**

HAN(Ours)+P 是指在 HAN(Ours)基础上加入规则后处理的方法，即利用规则标签修改模型预测标签的简单处理方式。RD-HAN(T)指直接引入规则嵌入的教师网络模型。可以看到，如果直接对模型进行规则后处理，即 HAN(Ours)+P 方法，在提高召回率的同时也损失了更多的准确率，因为规则匹配本身的准确率有限，会带来大量的错误正样本(False Positive)现象。换而言之，使用简单的后处理方法反而拖累了模型的整体性能。而采用规则嵌入的 RD-HAN(T)方法以更低的准确率下降代价提升了召回率，使得最终 F1 值相较于 HAN(Ours)上升了约 1 个百分点。基于对抗学习的知识蒸馏方法使得学生网络可以直接学习规则嵌入后的编码结果而不需要引入规则约束，即 RD-HAN(S)方法，能够在降低模型复杂度，

提高可控性的同时，也带来了性能上的提升。

不同蒸馏方式的对比：为了验证对抗训练的作用，本文引入知识蒸馏技术中常用的基于 K-L 散度的训练方法作为对比，将基于 K-L 散度训练得到的教师网络和学生网络分别记为 RD-HAN(T)-KL 和 RD-HAN(S)-KL，将基于对抗训练得到的教师网络和学生网络记作 RD-HAN(T)-Adv 和 RD-HAN(S)-Adv，对比结果见表 4。

表 4 不同蒸馏方式的对比

模型	P	R	F1
RD-HAN(T)-Adv	0.7665	0.7158	0.7403
RD-HAN(S)-Adv	**0.7706**	**0.7203**	**0.7446**
RD-HAN(T)-KL	0.7661	0.7137	0.7389
RD-HAN(S)-KL	0.7658	0.7123	0.7381

结果表明，使用基于 KL 散度的知识蒸馏策略训练得到的学生网络并不能有效地学习到教师网络的信息，其在准确率、召回率和 F1 值上相比教师网络均有下降。因此，基于对抗的知识蒸馏训练方法是有效的。

模型结构消融实验：为了评估本文提出的基础模型 HAN(Ours)中各部分的贡献，本文对该模型的不同结构组成进行了消融实验。结果如表 5 所示。

表 5 HAN(Ours)结构消融实验对比

模型	P	R	F1
HAN	0.7232	0.6605	0.6904
HAN(Ours)	**0.7771**	0.6897	**0.7304**
-word2vev	0.7405	0.6859	0.7117
-attention	0.7641	0.6829	0.7207
-hierarchy	0.6899	0.6833	0.6837
-position	0.7377	0.6544	0.6932
-highway	0.7626	**0.6958**	0.7269

其中 HAN 指原始的 HAN 模型，与 HAN(Ours)相比缺少了位置向量和残差结构；-word2vev 指不用预训练的词向量而是随机生成词向量；-attention 指在模型第一层不引入注意力机制；-hierarchy 指不用层级结构的 HAN(Ours)模型，即用第一层的 GRU 输出的句子向量加上注意力机制直接进行分类；-position 指在第二层句向量不拼接位置向量；-highway 指在模型得到第二层输出上下文向量送入分类器之前不拼接句向量。

结果显示，与原始的 HAN(Ours)相比，在情感原因发现任务中引入位置向量和残差结构给模型性能带来了大约 4 个百分点的提升。若不使用预训练词向量，模型的性能会受到较大的影响，F1 值下降了 2 个百分点。注意力机制的引入可以建模情感原因子句和情感表达句之间的关系，这为模型的分类性能带来了约 1 个百分点的提升，这说明该机制的引入是有必要的。如果忽略子句的上下文信息，模型的性能将大幅下降，F1 值下降了约 5 个百分点。如果不引入位置嵌入向量(Position Embedding)，模型的 F1 值将会下降 5 个百分点之多，这表明相对位置是一个重要的特征。在句子编码层引入残差结构对模型的性能提升有一定的帮助(约 0.4 个百分点)。

注意力机制作用分析：为了检验注意力机制的作用，采用了热力图工具将模型训练过程中的注意力权重矩阵可视化，如图 4 所示，对其进行定性的研究。

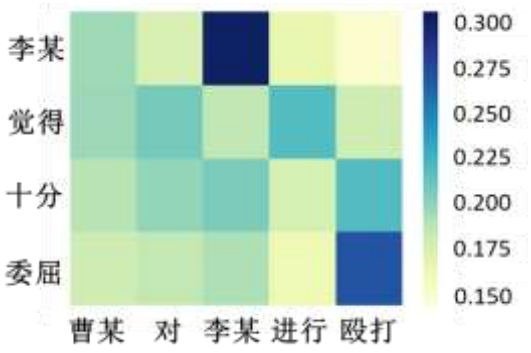

图 4 注意力权重矩阵可视化

该样例中，情感表达句为“李某觉得十分委屈”，分词后为“李某”，“觉得”，“十分”，“委屈”；情感原因句为“曹某对李某进行殴打”，分词后为“曹某”，“对”，“李某”，“进行”，“殴打”。图 4 中纵轴代表前者，横轴代表后者，各矩阵块的颜色代表其横纵坐标轴上的两个词之间的相关性，颜色越深表示越相关。可以看到，图中绝大多数词之间并无明显的相关性，两句中的“李某”由于是同一个词，且模型使用的词向量是上下文无关的，即词向量相同，所以该模块颜色最深。另外情感原因句中的“殴打”与情感表达句中的“委屈”具有较高的相关性，

所以在该

模型建模的过程中，能够更加“关注”这个词的语义，从而正确地将该句分类为情感原因句。这一可视化分析结果支持了实验中关于注意力机制的消融实验结论，显示引入注意力机制是合理且有效的。

错例分析：为了进一步分析模型预测错误的原因，从实验结果中选取了部分错例进行分析，如表 6 所示。

表 6 错例分析

序号	错例（**粗体为情感原因句**）
1	…令人高兴的是，高雁去南京的大医院复查时，被确诊为癌症早期，**治疗后不会有生命危险**…
2	…挺实惠的，**但老去同一家吃就烦了**…
3	…**小雪走了**，留给 20 岁的小娟无尽的悔恨和伤痛…

错例 1 的预测错误是由于情感原因句与情感关键词所在句“令人高兴的是”距离过远，模型不能够很好地建模二者之间的关系。同时，规则也无法匹配距离情感句过远的句子。在错例 2 中，情感原因句省略了主语，但语言学规则通常只能覆盖语法结构完整的语句。错例 3 中情感原因句太短，并且“走了”一词字面含义并无明显情感倾向性，模型难以判断其与情感表达之间的关系。

4 总结与展望

本文提出了一种结合规则蒸馏的情感原因发现方法，利用知识蒸馏技术将情感原因发现专家规则嵌入到基于层次注意力机制的深度神经网络模型，在保证模型性能的前提下提高了模型的可控性。在中文情感原因发现数据集上的实验结果表明，本文提出的方法达到了在该数据集上目前已知的最优结果。

在研究过程中发现，情感原因发现问题的标注语料比较匮乏，可能对方法设计带来限制，并且现有的语料缺乏个性化信息，不具备研究群体情感原因发现问题的条件，在后续的工作中将针对这样的问题展开研究。

参考文献

[1] Lee S Y M, Chen Y, Huang C R. A Text-Driven Rule-Based System for Emotion Cause Detection[C]//Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), 2010: 45-53.

[2] Lee S Y M, Chen Y, Huang C R, et al. Detecting Emotion Causes with a Linguistic Rule-Based Approach[J]. Computational Intelligence, 2013, 29(3): 2-28.

[3] Gui L, Yuan L, Xu R, et al.: Emotion Cause Detection with Linguistic Construction in Chinese Weibo Text[M], Natural Language Processing and Chinese Computing, Heidelberg, Berlin: Springer, 2014: 457-464.

[4] Gui L, Hu J, He Y, et al. A Question Answering Approach to Emotion Cause Extraction[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2017: 1593-1602.

[5] Ding Z, He H, Zhang M, et al. From Independent Prediction to Recorded Prediction: Integrating Relative Position and Global Label Information to Emotion Cause Identification[C]//Proceedings of the National Conference on the Association for the Advance of Artificial Intelligence (AAAI), 2019.

[6] Gui L, Wu D, Xu R, et al. Event-Driven Emotion Cause Extraction with Corpus Constructions[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2016: 1639-1649.

[7] 余传明 李, 安璐. 基于多任务深度学习的文本情感原因分析[J]. 《广西师范大学学报》(自然科学版), 2019, 37(1): 50-61.

[8] Hinton G, Vinyals O, Dean J. Distilling the Knowledge in a Neural

Network[C]//Proceedings of the Conference and Workshop on Neural Information Processing Systems (NIPS), 2015: 9.

[9] Hu Z, Ma X, Liu Z, et al. Harnessing Deep Neural Networks with Logic Rules[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL), 2016: 2410-2420.

[10] Liu J, Chen Y, Liu K. Exploiting the Ground-Truth: An Adversarial Imitation Based Knowledge Distillation Approach for Event Detection[C]//Proceedings of the National Conference on the Association for the Advance of Artificial Intelligence (AAAI), 2019.

[11] Krishna K, Jyothi P, Iyyer M. Revisiting the Importance of Encoding Logic Rules in Sentiment Classification[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018: 4743-4751.

[12] Li X, Song K, Feng S, et al. A Co-Attention Neural Network Model for Emotion Cause Analysis with Emotion Context Awareness[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2018: 4752-4757.

[13] Yang Z, Yang D, Dyer C, et al. Hierarchical Attention Networks for Document Classification[C], 2016: 1480-1489.