字里行间的道德：中文文本道德句识别研究

彭诗雅[+] 刘畅[+] 邓雅月 于东[*]
北京语言大学/ 信息科学学院
北京市海淀区学院路15号, 100083
pengshiya_blcu@163.com liuchang2014@gmail.com
1420048330@qq.com yudong_blcu@126.com

摘要

随着人工智能的发展，越来越多的研究开始关注人工智能伦理。在NLP领域，道德自动识别作为研究分析文本中的道德的一项重要任务，近年来开始受到研究者的关注。该任务旨在识别文本中的道德片段，其对自然语言处理的道德相关的下游任务如偏见识别消除、判定模型隐形歧视等具有重要意义。与英文相比，目前面向中文的道德识别研究开展缓慢，其主要原因是至今还没有较大型的道德中文数据集为研究提供数据。为解决上述问题，本文在中文语料上进行了中文道德句的标注工作，并初步对识别中文文本道德句进行探索。我们首先构建了国内首个10万级别的中文道德句数据集，然后本文提出了利用流行的几种机器学习方法探究识别中文道德句任务的效果。此外，我们还探索了利用额外知识辅助的方法，对中文道德句的识别任务进行了进一步的探究。

关键词： 人工智能伦理 ； 文本道德

Morality Between the Lines: Research on Identification of Chinese Moral Sentence

Shiya Peng Chang Liu Yayue Deng Dong Yu
Beijing language and Culture University
pengshiya_blcu@163.com liuchang2014@gmail.com
1420048330@qq.com yudong_blcu@126.com

Abstract

With the development of Artificial Intelligence, AI ethics has captured a great deal of public attention. Automatic moral recognition, as an important task in textual morality study, has attracted a lot of interest of NLP researchers in recent years. This task aims to identify fragments of text that involves morality, which is of great significance to moral-related downstream NLP tasks, such as model bias recognition and elimination. Compared with English, the study on textual moral identification for Chinese is slow, a main reason is that there is no large Chinese moral dataset to support research. Aiming to tackle these issues, we proposed a moral sentence labeling work in Chinese, and conducted a pilot study on the Chinese moral sentence recognition task. In this paper, we first constructed Chinese MOral Sentence dataset (CMOS) which consists of over

[*]为通讯作者
[+]为共同一作
基金项目：国家社会科学基金(17ZDA305);北京语言大学校级项目资助(中央高校基本科研业务费专项资金)(17PT05)

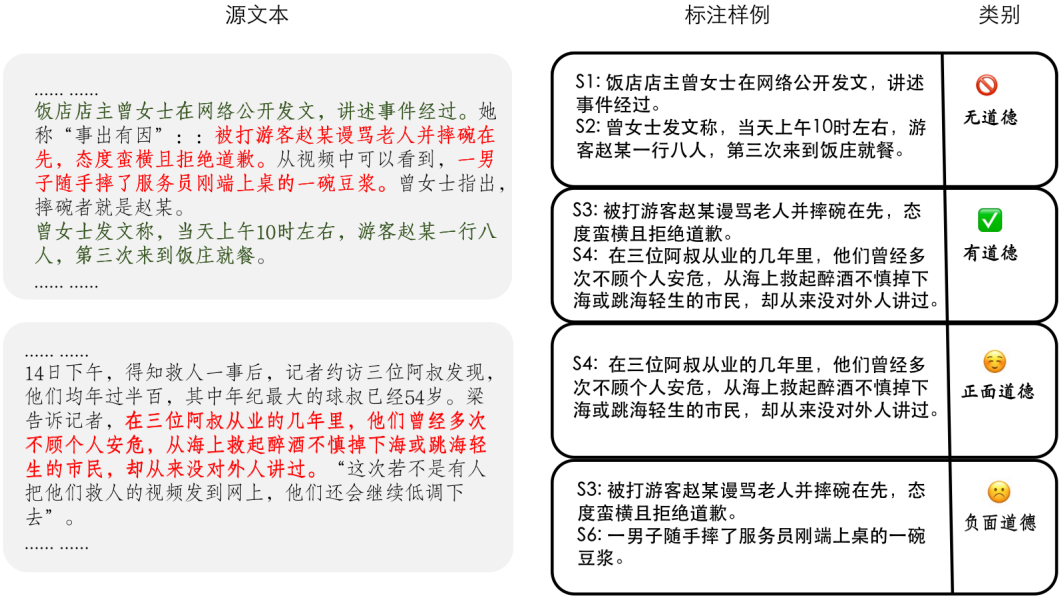

Figure 1: 本文的中文道德句的样例。左边为源文本，右边为数据集中的样例。

100k sentences with moral labels. Then we proposed to carry out the identification task using multiple popular machine learning methods. And we also further explored the identification task with knowledge-aided method.

Keywords: Artificial intelligence ethics , Text morality

1 引言

随着人工智能的应用逐渐进入人们的生活的方方面面，伦理问题逐渐成为需要关注的焦点，针对人工智能伦理的研究在当前的时代背景下具有深远且重要的意义。目前业界的学者已发现在一些领域中，由于训练学习的数据偏差，模型在判断时会出现某些隐形歧视的问题 (Zhang et al., 2020)。越来越多的AI学者开始关注如何让机器了解道德以及如何让机器具有人类道德 (Cervantes et al., 2013; Schramowski et al., 2019; Lourie et al., 2020)。

然而道德观念是无形的，如何才能让机器学习到人类的道德观念呢？其中一个重要的途径，就是通过筛选出的道德文本学习 (Xie et al., 2020; Shahid et al., 2020)。作为人类思维的最重要的载体之一，文本里包含了人类丰富的道德价值观。但目前针对文本道德的相关研究仍处于发展阶段，其原因之一是目前针对文本道德的数据较少。正确识别文本中的道德能促进对文本中的道德数据的收集，其对文本道德研究的相关任务具有极大帮助，从而促进对机器伦理学习的研究。因而文本道德识别任务是文本中关于机器伦理研究的不可或缺的基础性任务。

对道德文本的研究可从文本颗粒度角度，分为词级别，句子级别和篇章级别 (Araque et al., 2020; Johnson and Goldwasser, 2018; Shahid et al., 2020)。相比于含有知识较少的词，和含有复杂混合知识的篇章，句子具有相对适宜的信息载量。所以本文研究选择以句子级的文本为焦点，来探究识别自然语言中的道德。

面向中文的道德句识别任务的研究面临着很多难点。其中一个主要难点是数据短缺问题，目前在国内还不存在较大型的道德句数据集。且由于理论基础和思维方式等诸多差异，使得英语中的道德识别研究较难以直接迁移到中文里，对中文道德句的相关研究也因此难以发展。

为解决上述问题，本文首先构建了国内首个10万级别的大型中文道德句数据集CMOS(Chinese MOral Sentence dataset)。数据集样例如图1所示，该数据集为后续针对

数据集	来源	语言类别	标注类别及理论
(Johnson and Goldwasser, 2018)	Twitter	英文	MFTC五分类，MFTC
(Hoover et al., 2020)	Twitter	英文	MFTC五分类，MFTC
(Shahid et al., 2020)	在线新闻	英文	MFTC五分类，MFTC
(Lourie et al., 2020)	Reddit	英文	对涉及的人判断，描述伦理
(Forbes et al., 2020)	场景短句	英文	混合，多种分类规则
(Hendrycks et al., 2020)	场景短句	英文	混合，多种分类规则
CMOS	新闻，传记	中文	有/无，正/负，结果主义

Table 1: 本文数据集和以往数据集的对比。来源代表标注数据的领域来源，标注类别及理论表示数据集包含的标签和理论。

文本道德句的相关研究提供基础，并为促进文本道德的研究提供数据支撑。为了加大本任务的挑战难度，我们其次还利用机器辅助的方法对得到的中文道德句数据集进行进一步筛选，以一定的规则抽选出另一个难度较高的数据集。最终我们得到了两个道德句数据集，一个为全范围的CMOS数据集，一个为限定难度的CMOS-select-hard数据集。

作为国内对于文本道德句的早期研究，我们还为日后针对中文文本道德识别的研究提供了一组基线。基于本文构建的道德句数据集，我们利用几种当前流行的机器学习方法，对中文道德句识别任务的实验表现进行了探索，并得到了对应的基线结果，为日后针对中文道德句的识别方法提供了相应参考。

此外，我们还利用增加外部知识辅助的方法，对中文道德句识别的任务进行了进一步的探究。我们选取已有的中文道德词典作为外部知识源，探索其引入对识别效果的改进。在如何引入词典知识方面，我们尝试了两种方法：第一种是基于特征融合的方法，我们利用词典识别出相关的特征，然后将该特征和已有的道德句特征进行融合；第二种方法是基于Attention的方法，在此方法里，我们主要利用词典的知识在Attention上进行改进。本文在基于TextRNN+Attention的模型方法上对这两种方法进行了实验探索。实验表明，增加外部知识辅助能不同程度的提高模型对中文道德句识别任务的能力。本文的主要贡献有以下三个方面：

- 本文首先构建了一个10万级别的大型中文道德句数据集CMOS，其次还利用机器辅助的方法抽取构建了另一个难度较高的CMOS-select-hard数据集。该系列数据集为后续针对文本道德句的识别研究提供了数据基础，也为日后面向文本道德的相关研究提供了数据支撑。

- 基于文本构建的数据集，我们利用目前流行的机器学习方法对中文道德句识别任务进行了初步实验探究，为日后相关研究提供了参考。

- 此外，本文还利用增加外部知识辅助的方法对中文道德句识别任务进行了进一步探究。我们选取中文道德词典作为外部知识源，对引入其在识别中的效果进行了探索。实验结果表明，引入词典知识可以有效提升模型的识别性能。

2 相关工作

由于道德是无形的，因此语言作为载体，成为人们表达自己的道德价值观的重要途径。分析文本中含有的道德价值观对洞察人类道德具有重要意义，文本道德相关任务因而逐渐被越来越多的学者所关注 (Xie et al., 2020; Shahid et al., 2020; Lourie et al., 2020)。目前，国外针对文本道德的研究逐渐进入发展阶段，而相比之下国内对于文本道德的相关研究仍处于起步阶段，进展相对较慢。

现有的道德文本研究主要包括基于词级别，句子级别以及篇章的研究。这些方法主要通过对文本中进行词，句子或篇章层面的分析判别，来对其进行道德属性的分析。

针对英文文本的道德研究发展迅速，最具规模。首先在词级别的研究上，Graham等人(2009)提出的道德基础词典(MFD)已在道德词研究上颇具影响力，该词典包含151个正面词和168个负面词，带有人工标注的标签。MoralStrength (Araque et al., 2020)则是在道德基础词典分类的基础上，利用WordNet词汇的数据库对MFD进行了扩展，最终获得了含有520个正面词和476个负面词的数据集。而在句子级别的研究上，Hoover等人(2020)构建了一个道德基

础Twitter语料库MFTC，该文集包含35,108条推文，这些推文来自七个不同的领域，并已用于探究识别道德情绪的预测。Johnson等(2018)基于MFT理论的分类，提供了标注准则的描述和2,050条推特的标注数据集，并提出了PSL模型用于对推特上的政治言论中表达的道德类型进行分类。此外，Garten等人(2016)则收集了社交媒体数据语句，用以分析其中的道德修辞，他们还探究使用较长时间的演讲来探讨随着时间的推移检测该修辞的变化。另外在篇章级别的研究上，Clifford和Jerit(2013)使用MFD来执行手动文本分析，分析12年来的纽约时报关于干细胞研究的篇章报道中的道德修辞。Dehghani等人(2014)则使用基于LDA的方法研究了博客语料库中自由主义者和保守主义者的道德价值体系之间的差异。总的来说，基于句子的研究是文本道德目前流行的主要方向。句子作为语言运用的基本单位，能较好地表达一个完整的意思。相比于所含信息单一的道德词，以及所含信息较为混合复杂的篇章，道德句具有相对适宜的信息含量。因而目前针对文本道德的研究较为集中于句子级别。

与此同时，国内人工智能伦理研究仍大多集中于从哲学领域的理论层面进行的探讨(王东浩, 2014; 李伦and 孙保学, 2018)，针对中文文本道德的自然处理领域研究尚在初探过程中。王弘睿等(2020)对词的道德倾向性进行研究分析，提出面向的中文道德词典构建任务。他们将词典词分为四类标签和四种类型，通过词向量扩展和人工标注构建中文道德词典资源。该词典包含25,012个词，其中正向道德词7,912个，负向道德词7,647个，中性词8,963个，被动词490个。但他们的关注中心主要在词级别文本，未涉及到对大规模文本道德句子识别的研究。

目前，尚未有面对中文道德句的识别研究，其原因主要在于缺乏相关的数据。而由于理论基础和思维习惯等方面的诸多差异，使得英语中的道德识别研究较难直接迁移到中文上。在表1中，我们选取了几个典型的国外文本道德句数据集与本文提出的数据集CMOS进行了对比，说明了针对中文构建道德句数据集的必要性。

3 中文道德句识别数据集

3.1 数据选择和预处理

道德句可能出现在任何类型的自然语言文本中。由于不同来源的语料资源具有不同的特点，本文选取并对比分析了不同语料。我们着重依照两个原则考虑：来源语料表述的内容应为客观清晰的行为事件，以及来源语料应具有适宜的难度和较大的数量，使得数据集具有扩展性，为未来扩充做好准备。

统计项	数量	占比（%）
无道德	63,414	62.39
有道德	38,089	37.61
正面道德	11,605	30.46
负面道德	26,484	69.54
数据集合计	101,503	

(a) CMOS基础信息

统计项	数量	占比（%）
无道德	21,980	61.99
有道德	13,477	38.01
正面道德	6898	56.91
负面道德	5223	43.09
数据集合计	35,457	

(b) CMOS-select-hard基础信息

Table 2: 中文道德句CMOS系列数据集统计信息。上面为全范围的CMOS数据集统计信息，下面为限定难度的CMOS-select-hard数据集统计信息。

根据以上原则，本文选取了新闻文本和传记文本作为语料来源。新闻文本，尤其是社会新闻文本，对于事件的表述较为清晰，且道德含量比较高，是标注道德句的合适来源；传记文本

对于事件的描述也较为清楚，特定的传记文本里也会包含一定的道德表述事件。通过比对并考虑到语料的平衡性，我们最终选定本次标注的语料来源为新闻和传记。

我们选取新闻来源为两个。第一个为网络爬取的新闻，筛选后的网络爬取新闻文本规模为10657篇，内容全部为社会类新闻。第二个为BDCI数据集[0]，该数据集为互联网新闻情感数据集。考虑到道德和情感具有一定的相关性，我们选择该数据集部分数据作为新闻来源，为日后探究道德情感联合任务做准备。我们选择的传记文本来源为中国文明网登载的《中国好人传》[1]，该书包含各类好人故事，我们利用其中的10000篇作为标注的传记来源。两类原始语料经过清洗去重，作为本次标注源语料。

3.2 人工数据标注规则和流程

3.2.1 数据标注规则

中文道德句是一段以句为单位的连续的文本，本文中的中文道德句数据集的标注方法参考目前道德理论中的结果主义理论(Thiroux and Krasemann, 1980)。由于道德句标注涉及到对文本的理解，标注具有一定的难度，为此本文预先设定了一组标注准则。具体规则如下：

结果主义原则 标注员在标注道德句时，应优先考虑以结果主义原则确定道德属性。如例2.1所示，“拿到项目”并没有对他人产生正面或负面的结果，因此该句应被标注为无道德。而在例2.2中，则反馈出对他人的正面的结果，据此，标注者应将该句标注为正面道德。

例1 卫小军以80万元的价格拿到了某项目部25号、26号楼的基建工程项目。

例2 他不仅将小孩从冰冷的水中救出，并在危急时刻运用医学基础知识和临床技能将小孩救活。

事件优先原则 关于道德的描述有不同的表述方式，本文目前主要标注带有具体行为事件的道德句。标注者在标注道德句时，被标注的道德文本表述中应描绘具体事件。如例1.1所示，句中虽然存在正面词诚信，但该句并无具体行动。而例1.2中则包含关于道德事件的具体行动描述，该句子应标注为正面道德。

例1 她用朴素的方式，告诉我们诚信的含义。

例1 孩子们获救后，李军再次游到河中间，分别将小孩母亲和驾驶员从车窗内拉出来。

独立优先原则 部分涉及道德的句子依赖于上下文环境。在标注时，标注者应尽量选取独立于环境的语句，或选取具有最短上下文的句段，保证其具有独立性质。如例3.1，单看这句并不能判断其道德性质，所以该句并非道德句。而例3.2满足独立性质，应被标注为负面道德句。

例1 男子往后撤了一下，可随着车厢晃动，男子又往前贴近。

例2 男子大概有两分钟的时间，一直贴着红裙子姑娘的臀部。

3.2.2 标注流程

中文道德句标注准备过程分为两个阶段。首先，经过调研现有道德理论和英文数据集以及适合语料后，制定初步规则并收集处理选定的源语料。在标注前，我们对源语料进行了一定的预处理，如切段分行，分离出易于标注的语段。其次，我们从不同来源的语料中各抽取一部分进行试标注，根据试标注结果选定最终决定采用的语料来源，并针对标注出的样本结果进行讨论并修改标注规则。

在明确语料和规则后，我们招募了两批次共13位具有语言学背景的本科生和硕士生作为标注人员进行标注。标注分为三个阶段，分别是线下培训阶段、试标注阶段和正式标注阶段。在线下培训阶段，我们向标注人员介绍本次标注的背景知识和具体的标注规则，并提供相关的标注范例。在试标注阶段，我们向每一名标注人员分发200条标注语料，在标注人员将其标注后由审核人员进行核查反馈。标注人员的试标注结果符合标注规则的情况下，才可以进入正式标注阶段，对我们抽取的待标注语料进行标注。

3.3 机器辅助标注的流程和方法

由于人工标注的数据难度不一，为了增加数据集的挑战性，我们利用机器辅助标注的方法对人工标注的数据进行了进一步的筛选，从中筛选出对于模型辨识难度较高的数据集部分组成数据集，作为CMOS-select-hard数据集。挑选CMOS-select-hard数据集是为了筛选出人工标注的数据中对人类来说可以轻易判断，但是对模型来说容易混淆的数据。

[0]https://www.datafountain.cn/competitions/350

[1]http://www.wenming.cn/book/wmws/

3.3.1 机器辅助的标注流程

机器辅助构建的CMOS-select-hard数据集流程如图2所示，从人工构建的道德句标注数据集开始，经过切分处理数据、训练模型、模型预测输出概率、挑选概率，经过图中的四个标注流程，我们最终得到机器辅助构建的CMOS-select-hard数据集。

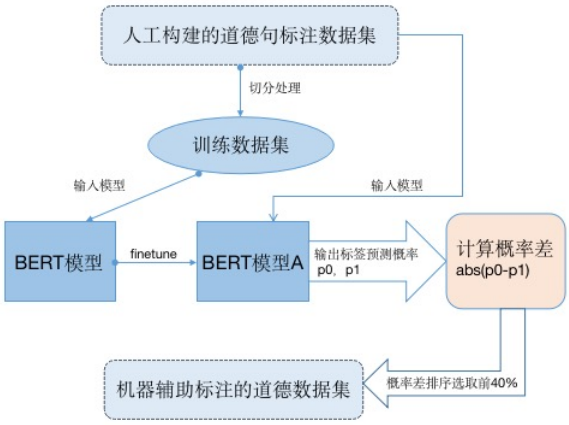

Figure 2: 机器辅助标注资源建设流程图

3.3.2 机器辅助的标注方法

在进行机器辅助标注之前，我们先将整个人工标注完成的数据集D按照3:1:1的比例划分为训练集、验证集和测试集，随后我们选用预训练语言模型BERT作为辅助模型，在上述数据上finetune该模型，得到模型A，并记录D的测试集在A上的结果；然后我们将D的所有数据作为A的输入，记录A输出的最后一层全连接层的概率预测结果，作为判定数据难度的依据。

对任一条数据，模型A的最终输出是一组两个概率的tuple，分别对应该条数据应该分为0标签或者1标签（针对任务一和任务二，0标签和1标签分别对应无/有道德判断和负面/正面道德）的概率。对模型而言，两个标签之前的区别越是显著，则两个标签对应的概率的差值越大。如果一条数据经过A的预测，有0.9的概率属于标签1，而有0.1的概率属于标签0，则模型能很大程度上肯定该条数据属于标签1；反之，若一条数据属于标签1的概率为0.51，而有属于标签0的概率为0.49，则模型层面上可以认为很难区分这条数据的标签归属。

在记录了D的所有数据的模型预测输出后，我们将D的数据以概率差值的绝对值为key进行排序，并根据对数据的观察和对概率差值绝对值的分布的观察，取D的概率差值绝对值最小的前40%数据作为CMOS-select-hard数据集，记为D2。

为了证明D2的难度大于D的难度，我们又将D2按照与D同样的划分方式划分为训练集、验证集和测试集，并选用同样结构的预训练语言模型BERT为模型，在D2上finetune了一个新的模型A2，并记录D2的测试集在A2上的结果。结果的对比可以参见表3。从表3可以看出，挑选出来的CMOS-select-hard数据集的难度增加了，模型的表现相比在全范围的CMOS数据集上的表现出现了明显的下滑，说明我们的使用机器辅助的标注方法是可行且有效果的。

	Task1			Task2		
Model	**Acc**	**Precision**	**F1**	**Acc**	**Precision**	**F1**
CMOS						
BERT	90.39	89.33	89.46	93.71	92.59	92.94
CMOS-select-hard						
BERT	77.10	75.86	75.43	84.30	84.27	83.91

Table 3: CMOS数据集和CMOS-select-hard数据集上模型的表现对比

Models	Task1			Task2		
	Acc	Precision	F1	Acc	Precision	F1
CMOS						
SVM	69.00	71.00	55.00	78.00	78.00	72.00
LR	69.00	69.00	57.00	75.00	76.00	67.00
FastText	87.06	88.44	84.78	80.17	77.75	77.01
TextCNN	84.05	83.23	81.85	77.59	74.62	74.51
TextRNN	87.64	87.95	85.78	77.20	74.50	72.85
TextRNN+att	87.83	88.69	85.87	80.14	78.61	75.98
BERT	90.39	89.33	89.46	93.71	92.59	92.94
CMOS-select-hard						
SVM	63.00	66.00	42.00	64.00	67.00	57.00
LR	63.00	63.00	45.00	63.00	66.00	57.00
FastText	76.14	75.50	73.49	79.97	80.41	79.08
TextCNN	71.01	69.84	66.94	77.41	77.10	77.19
TextRNN	74.76	73.51	72.42	79.97	79.65	79.51
TextRNN+att	76.14	74.75	74.68	76.50	79.10	74.40
BERT	77.10	75.86	75.43	84.34	84.27	83.91

Table 4: 数据集在机器学习的表现。本文探究了三类主流方法的表现，包括统计学习方法，卷机神经方法和预训练方法。采取的评价指标为正确率，准确度，F1值。

3.4 标注结果

经过人工标注和机器辅助标注，我们最终构建了本文的两个数据集，一是全范围的中文道德句CMOS数据集，二是难度限定的CMOS-select-hard数据集。数据集统计信息如表2所示，本文的道德句数据为相关研究的开展提供了基础。

为了计算标注一致性，我们分别在CMOS的任务一和任务二的数据集中抽取了300条和500条数据，分别找2位标注员进行统一标注，并利用Fleiss' kappa (Cooper, 2003)计算方法进行一致性计算。经过计算，一致性检验结果分别为70.54和86.93，标注数据结果一致性较好，说明标注质量有一定的保证。在日后对数据集的维护和扩充时，我们也会进一步进行复标提高标注质量。

4 中文道德句识别初步探究

中文道德句识别任务本质上可以分为两个，一个是识别出某一句文本是否包含道德内容，另一个则是识别其包含道德内容的正负。我们将这两个任务抽象为两个二分类任务。

4.1 实验方法

为探索和分析机器学习方法在本文构建数据集上的表现，本文选取了目前流行的三类主流分类方法的七个文本分类模型：SVM，LR，TextCNN，TextRNN，TextRNN+attention,FastText，以及Bert。下面将相关的文本分类模型进行介绍:

(1)SVM: 支持向量机(Support Vector Machine,SVM) (Cortes and Vapnik, 1995)在解决小样本、非线性及高维模式识别中表现很好，是应用广泛的一种分类算法。

(2)LR: 逻辑回归(LR,Logistic Regression)(Pedregosa et al., 2011)是传统机器学习中的一种分类模型,简单、高效、易于并行和在线学习(动态扩展),在工业界具有非常广泛的应用。

(3)FastText: (Bojanowski et al., 2017)主要思想基于word2vec中的skip-gram模型，在训练文本分类模型的同时，也将训练出字符级n-gram词向量。该方法专注于文本分类，在许多问题上有很好的表现。

(4)TextCNN: 本文使用的CNN模型基于(Rakhlin, 2016)描述的TextCNN模型，在句子分类任务上有不错的表现。将卷积神经网络CNN应用到文本分类任务，利用多个不同尺寸的卷积核进行卷积来提取句子中的关键信息。CNN的并行计算能力很强，可以快速实现特征提取。

Models	Task1	Task2		
	无道德	有道德	正面道德	负面道德
CMOS				
SVM	69.00	73.00	79.00	78.00
LR	68.00	70.00	77.00	75.00
FastText	91.37	85.51	71.70	83.81
TextCNN	81.14	85.32	66.21	83.03
TextRNN	88.69	87.20	68.59	80.40
TextRNN+att	90.64	86.75	75.50	81.72
BERT	85.70	92.96	89.05	96.12
CMOS-select-hard				
SVM	62.00	70.00	71.00	62.00
LR	63.00	63.00	69.00	62.00
FastText	73.88	77.12	82.00	78.81
TextCNN	67.34	72.35	71.93	82.27
TextRNN	69.92	77.10	77.71	81.59
TextRNN+att	69.06	80.43	85.04	73.17
BERT	71.52	80.19	84.60	83.94

Table 5: 模型在不同道德属性标签上的分类性能

(5)TextRNN: 在实验中，我们采用基于双向LSTM的textRNN模型(Liu et al., 2016)。循环神经网络善于捕捉更长的序列信息，其在每个时间步上的输入有两部分信息：部分是前一个时间步的保留信息，部分是当前时间步对应的原始信息。

(6)BERT: BERT（Bidirectional Encoder Representations from Transformers）是近年来最为热门的预训练方法的代表(Devlin et al., 2018)。作为一种新的语言模型，它在各种文本任务如问答、命名实体识别、自然语言推理、文本分类等上表现突出。针对道德句识别任务，本文实现了Bert，并探究其在道德识别任务上的效果。

此外，以上方法在文本分类任务中尽管效果显著，但都有一个不足的地方就是不够直观，可解释性不好。而注意力（Attention）(Vaswani et al., 2017)机制是自然语言处理领域一个常用的建模长时间记忆机制，能够很直观的给出每个词对结果的贡献。所以我们除了以上五个模型外，还利用注意力机制构建了TextRNN+Attention方法，用以提高可解释性，辅助后续实验分析。

4.2 实验设置

数据集选取自本文构建的中文道德句数据。我们按照3:1:1的比例将数据打乱后划分为训练集，验证集和测试集。在参数设置上，对SVM模型，使用网格搜索的方法确定最佳的参数为gamma=1，C=10；对lr，使用网格搜索确定最佳参数为tol=1e-4，C=1；对FastText，TextCNN，TextRNN，TextRNN+Attention，都设置学习率为1e-3，并设置0.5的dropout概率。对TextCNN，有卷积核尺寸为(2，3，4)，卷积核数量为256；对TextRNN和TextRNN+Attention，有hidden size=128和lstm层数=2，对FastText，有hidden size=256。

为了更好的分析模型的识别表现，以及探究模型的表现效果。我们另外还在测试集中选取了一些样例,对TextRNN+Attention方法的Attention矩阵进行了可视化。

4.3 实验结果及分析

为探究中文道德句识别的基线效果，我们基于前述的六种主流文本分类方法在CMOS数据集上进行了实验。实验结果如表4所示，我们选取的评价指标为正确率，准确度和F1值。

对于全范围的CMOS数据集，在任务一识别句子道德的有无实验中，表现最好的模型为BERT，其次为TextRNN+Attention，两者在正确率，准确度和F1值三个方面均高于其他方法。在任务二识别句子道德的正负实验中，表现最好的模型为为BERT，其次为FastText。

Figure 3: 模型在识别任务上的Attention示意图。

从整体可以看出，除了BERT外，任务一的识别平均效果在0.85左右，任务二的识别平均效果为0.79左右。在本次实验中，在使用同样的模型的前提下，我们发现其对句子道德的有无更为敏感，对句子道德的正负的识别能力则相对弱一些。我们推测这一方面可能是因为任务二的数据因为是任务一数据的一个子集，其本身数量要少一些。另一方面则是在任务本身的难度方面，对于参与实验的几种模型而言，任务二比任务一要更难一些。

除了任务的平均效果，我们还分析了各模型在两个任务的四个标签上的单独表现。具体实验结果如表5所示，在表中我们展示了六个模型在无道德，有道德，正面道德和负面道德的识别正确率。在任务一的有道德和无道德标签中，有道德标签表现最好的是BERT,无道德标签表现最好的是FastText。从模型的整体表现可以看出，模型对于有无的识别能力并不一致，部分模型如FastText和TextRNN对于无道德标签识别能力更强，而TextCNN和SVM等则对有道德标签识别能力更强。在任务二的正面道德和负面道德标签中，正面道德标签表现最好的BERT，负面道德标签表现最好的为FastText。而从模型的整体识别表现来看，模型对对负面道德的识别表现最佳，而对正面识别效果偏弱。该识别表现差异性原因可能因为在新闻数据里负面道德句中的道德特征更明显更易于识别，而正面道德的表述则不太明显。

而对于CMOS-select-hard数据集，模型方法的表现效果均有明显下降，这说明我们的使用机器辅助的标注方法构建的较难数据集是有一定挑战性的。

此外，我们还从Attention的角度对模型效果进行了分析。对一条数据而言，句中对结果贡献大的部分，Attention矩阵中对应位置的值越大，在图中对应部分的颜色就越浅，样例如图3所示。从图中可以看出，每个部分的Attention权重并不统一。我们分析了一些具体的数据后发现，图中偏亮的部分有与句中道德词的部分重叠的情况，这说明模型在识别时有效利用到了部分道德词的知识信息；但也存在部分浅色部分与句中道德相关的部分无任何联系的情况。

从Attention矩阵可视化的分析中，我们发现模型在一定程度上可以捕捉到词汇层面上的道德相关信息，但某些句子中词汇级别的道德信息并未得到明显关注，所以我们考虑是否可以通过引入外部知识，如道德词典知识的方法，来辅助模型进行道德句识别。为了探究该方法的表现，我们在后续对其进行了进一步尝试探究。

5 基于外部知识辅助的中文道德句识别探究

为了进一步探究中文道德句识别任务，我们在基线实验外，又探索了利用外部知识辅助识别的方法。为了探究加入外部知识如道德词典知识是否能提升模型识别的整体效果，我们提出利用两种方法引入知识，并对比基线模型进行了实验验证。

Setting	Accuracy	Precision	F1
Task1			
TextRNN+Attention	87.83	88.69	85.87
+ dictionary(add feature)	89.08	88.51	**87.77**
+ dictionary(mod Attention)	88.58	88.05	87.17
Task2			
TextRNN+Attention	80.14	78.61	75.98
+ dictionary(add feature)	86.19	86.13	**83.46**
+ dictionary(mod Attention)	79.38	78.01	74.71

Table 6: 引入外部知识Attention方法。我们验证了利用外部知识改进Attention的实验效果，并在表中与Task1和Task2的Baseline效果和添加feature的方法进行了对比。

5.1 基于道德词典的识别方法

我们选取的外部知识为道德词典，该词典由王弘睿等(2020)提出，他们将词典词分为四类标签和四种类型，通过词向量扩展和人工标注构建中文道德词典资源。该词典包含25,012个词，其中正向道德词7,912个，负向道德词7,647个，中性词8,963个，被动词490个。

在本研究中，我们主要利用的为正向道德词和负向道德词两类词知识。在引入词典知识的方法中，我们尝试了两种方法:

第一种方法是基于特征融合的方法。我们利用道德词典的识别出句中正向和负向的道德词的数量作为特征，然后将其与道德句输入模型的向量表示进行拼接。

第二种方法是基于对Attention矩阵进行修改的方法。我们主要利用道德词典识别出句中道德词的位置，并对Attention矩阵中对应位置的权重进行放大。本文在基于TextRNN+Attention的模型方法上对这该方法进行了实验探索。

5.2 实验设计及结果分析

为了验证我们提出的两种引入词典知识方法是否有效，我们在本文构建的CMOS数据集上进行了实验探究。实验利用的训练集，测试集和验证集数据和基线方法一致。两中方法的实验都是在基线方法中的TextRNN+Attention模型的基础上进行的，实验结果表6所示。

可以看出，第一种方法对两个任务都有一定的提升，而第二种方法仅对任务一有提升。我们认为，第一种方法直接引入了道德词的数量和正负信息，是很强但并非决定性的特征，可以在基线系统的基础上取得一定的提升，但是幅度是有限的；第二种方法实质上引入的信息是“哪些部分是道德词”，而没有引入道德词的正负信息。这个信息对区分道德内容的有无有一定的帮助，但是对判断道德的正负而言，这个信息效果有限，且改动了本来有效的Attention矩阵，所以反而会造成模型效果的降低。

6 结论

本文的主要工作是在自然语言处理中对文本道德进行研究，具体的研究任务是面向中文的道德句识别。为探究针对中文文本的道德句识别任务，本文主要从资源建设和识别方法两个方面展开了研究。

目前开源的文本道德数据集仍较少，数据短缺问题导致相关领域的任务难以开展。在现有的中文道德研究领域里，还不存在较大型的中文文本道德句相关的数据集，所以面向中文的文本道德句识别任务较难发展。为了解决这个问题，我们首先建设了一个10万级别的大型中文道德句数据集CMOS。该数据集为后续针对文本道德句的相关研究提供了数据支撑，也为日后面向文本道德的相的研究提供了数据基础。其次，为了加大本任务的挑战难度，我们还利用机器辅助的方法对得到的中文道德句数据集进行进一步筛选，以相应的规则抽选出另一个难度较高的数据集。最终我们得到了两个道德句数据集，一个为全范围的CMOS数据集，一个为限定难度的CMOS-select-hard数据集。

本研究基于CMOS系列道德句数据集，对中文道德句识别任务进行了探索。在道德识别方面，本文分别从基线识别方法和外部知识辅助识别方法两方面进行了探究。在基线方法

中，本文选取了目前流行的三类主流分类方法的七个文本分类模型：SVM，LR，TextCNN，TextRNN，TextRNN+attention,FastText，以及Bert。通过在中文道德句识别任务上运用以上七种模型方法，我们得到了一组基础的基线结果。该组基线结果为日后针对相关研究提供了参考。在外部知识辅助识别方法中，我们尝试利用已有的道德词典知识来辅助道德句识别，对利用外部知识辅助中文道德句识别的方法进行了初步探究。我们探究了两种加入词典知识的方法，特征融合和在Attention上改进。经过实验验证，我们证明加入外部知识的两种方法均能有效提高模型识别效果。

目前我们的道德句数据集包含的道德句的标签还比较单一。未来我们将继续完善标注类别，增加如强度属性等相关道德标签，同时也扩充道德强度低的数据，为道德句研究提供更多帮助，也为文本道德研究提供更多数据支撑。

参考文献

Oscar Araque, Lorenzo Gatti, and Kyriaki Kalimeri. 2020. Moralstrength: Exploiting a moral lexicon and embedding similarity for moral foundations prediction. *Knowledge-based systems*, 191:105184.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

José-Antonio Cervantes, Luis-Felipe Rodríguez, Sonia López, and Félix Ramos. 2013. A biologically inspired computational model of moral decision making for autonomous agents. In *2013 IEEE 12th International Conference on Cognitive Informatics and Cognitive Computing*, pages 111–117. IEEE.

Scott Clifford and Jennifer Jerit. 2013. How words do the work of politics: Moral foundations theory and the debate over stem cell research. *The Journal of Politics*, 75(3):659–671.

Harris Cooper. 2003. Psychological bulletin.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.

Morteza Dehghani, Kenji Sagae, Sonya Sachdeva, and Jonathan Gratch. 2014. Analyzing political rhetoric in conservative and liberal weblogs related to the construction of the "ground zero mosque". *Journal of Information Technology & Politics*, 11(1):1–14.

J. Devlin, M. W. Chang, K. Lee, and K. Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Maxwell Forbes, Jena D Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020. Social chemistry 101: Learning to reason about social and moral norms. *arXiv preprint arXiv:2011.00620*.

Justin Garten, Reihane Boghrati, Joe Hoover, Kate M Johnson, and Morteza Dehghani. 2016. Morality between the lines: Detecting moral sentiment in text. In *Proceedings of IJCAI 2016 workshop on Computational Modeling of Attitudes*.

Jesse Graham, Jonathan Haidt, and Brian A Nosek. 2009. Liberals and conservatives rely on different sets of moral foundations. *Journal of personality and social psychology*, 96(5):1029.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2020. Aligning ai with shared human values. *arXiv preprint arXiv:2008.02275*.

Joe Hoover, Gwenyth Portillo-Wightman, Leigh Yeh, Shreya Havaldar, Aida Mostafazadeh Davani, Ying Lin, Brendan Kennedy, Mohammad Atari, Zahra Kamel, Madelyn Mendlen, et al. 2020. Moral foundations twitter corpus: A collection of 35k tweets annotated for moral sentiment. *Social Psychological and Personality Science*, 11(8):1057–1071.

Kristen Johnson and Dan Goldwasser. 2018. Classification of moral foundations in microblog political discourse. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 720–730.

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2016. Recurrent neural network for text classification with multi-task learning. *arXiv preprint arXiv:1605.05101*.

Nicholas Lourie, Ronan Le Bras, and Yejin Choi. 2020. Scruples: A corpus of community ethical judgments on 32,000 real-life anecdotes. *arXiv preprint arXiv:2008.09094*.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

A Rakhlin. 2016. Convolutional neural networks for sentence classification. *GitHub*.

Patrick Schramowski, Cigdem Turan, Sophie Jentzsch, Constantin Rothkopf, and Kristian Kersting. 2019. Bert has a moral compass: Improvements of ethical and moral values of machines. *arXiv preprint arXiv:1912.05238*.

Usman Shahid, Barbara Di Eugenio, Andrew Rojecki, and Elena Zheleva. 2020. Detecting and understanding moral biases in news. In *Proceedings of the First Joint Workshop on Narrative Understanding, Storylines, and Events*, pages 120–125.

Jacques P Thiroux and Keith W Krasemann. 1980. *Ethics: Theory and practice*. Glencoe Publishing Company.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Jing Yi Xie, Renato Ferreira Pinto Jr, Graeme Hirst, and Yang Xu. 2020. Text-based inference of moral sentiment change. *arXiv preprint arXiv:2001.07209*.

Guanhua Zhang, Bing Bai, Junqi Zhang, Kun Bai, Conghui Zhu, and Tiejun Zhao. 2020. Demographics should not be the reason of toxicity: Mitigating discrimination in text classifications with instance weighting. *arXiv preprint arXiv:2004.14088*.

李伦and 孙保学. 2018. 给人工智能一颗“良芯(良心)”——人工智能伦理研究的四个维度. 教学与研究, (2018 年08):72–79.

王东浩. 2014. 人工智能体引发的道德冲突和困境初探. 伦理学研究, 2:68–73.

王弘睿, 刘畅, and 于东. 2020. 面向人工智能伦理计算的中文道德词典构建方法研究. In 第十九届中国计算语学大会*(The Nineteenth China National Conference on Computational Linguistics, CCL 2020)*, pages 539–549.