面向人工智能伦理计算的中文道德词典构建方法研究

王弘睿
北京语言大学
whongrui18@163.com

刘畅
北京语言大学
liuchang2014@gmail.com

于东*
北京语言大学
yudong_blcu@126.com

摘要

道德词典资源的建设是人工智能伦理计算的一个研究重点。由于道德行为复杂多样，现有的英文道德词典分类体系并不完善，而中文方面目前尚未有相关的词典资源，理论体系和构建方法仍待探究。针对以上问题，该文提出了面向人工智能伦理计算的中文道德词典构建任务，设计了四类标签和四种类型，得到包含25,012个词的中文道德词典资源。实验结果表明，该词典资源不仅能够使机器学会道德知识，判断词的道德标签和类型，而且能够为句子级别的道德文本分析提供数据支持。

关键词： 伦理计算 ； 道德判断 ； 机器学习

Construction of a Chinese Moral Dictionary for Artificial Intelligence Ethical Computing

Hongrui Wang
Beijing Language
& Culture University
whongrui18@163.com

Chang Liu
Beijing Language
& Culture University
liuchang2014@gmail.com

Dong Yu
Beijing Language
& Culture University
yudong_blcu@126.com

Abstract

The construction of the moral dictionary is based on artificial intelligence ethical computing. Moral behavior is complex and varied. The existing moral dictionary classification system for English language is still under development. Meanwhile, there are currently no relevant dictionary resources in Chinese. The theoretical system and construction method are still to be explored. In this paper, the task of constructing a Chinese moral dictionary for artificial intelligence ethics calculation is proposed. Four polar labels and four types of labels are designed to obtain a Chinese moral dictionary resource containing 25,012 words. Experimental results show that the dictionary resource can not only enable the machine to learn moral knowledge and judge the moral polarity and type of words, but also provide data support for the analysis of moral text at the sentence level.

Keywords: ethical computing , moral judgment , machine learning

*为通讯作者
基金项目： 国家社会科学基金(17ZDA305);教育部人文社会科学研究青年基金项目(19YJCZH230);北京语言大学中青年学术骨干支持计划

第十九届中国计算语言学大会论文集，第539页-第549页，海南，中国，2020年10月30日至11月1日。

1 引言

道德判断是人工智能伦理计算的一个重要问题。随着人工智能技术的快速发展，机器接管了越来越多人类的工作任务，社会对人工智能决策的道德性的担忧也与日俱增：人工智能能否理解我们的道德观念，又能否学会人类的道德判断？正如Picard(Picard, 1997)所说，“一台机器的自由度越大，它就越需要道德标准”。将重大决策的控制权交给机器之前，机器首先需要具有符合人类道德标准的判断能力。

使用词汇来识别复杂的道德概念，从而让机器学会道德标准甚至具备道德判断的能力，被认为是一种可靠的方法。Graham等人(Graham et al., 2009; Hofmann et al., 2014; Feinberg and Willer, 2013; Clifford and Jerit, 2013)研究表明，词汇是可以帮助机器进行道德判断的一个相当可靠的标识符。在过去的十年里，已经有许多研究运用词汇学的方法来分析文本数据的道德基础(Kaur and Sasahara, 2016; Hoover et al., 2019; Sagi and Dehghani, 2014)。机器可以了解到“恶意谋杀”是不道德的，“诚实守信”则是一种好的行为。道德词典成为实现人工智能伦理计算的重要数据资源。

目前发布的英文道德词典主要存在两个问题，一是目前的道德词典只考虑了与道德行为有关的动词，词性构成和词的类型比较单一；二是目前的道德词典普遍规模较小，难以较为全面地覆盖道德行为。因此，现有的英文道德词典体系尚不完善。目前尚未有公开发布的中文道德词典，其理论体系及构建方法值得分析研究。

针对第一个问题，本文认为，除了动词，名词、形容词以及一些成语也与行为的道德倾向息息相关，如活雷锋、无私、公正廉洁等。除此之外，汉语中有不少体现道德倾向的被动表达，比如被害、弃婴等。因此，本文将动词、名词、形容词、成语和汉语中的被动表达都纳入到道德词典理论体系中，并根据所标注词在事件行为中所处的位置，分为事件行为、事件状态、事件属性、事件要素四种类型，其中事件要素再分为对象、媒介和地点三类。本文研究丰富了词典的词性构成和词的类型，覆盖了更多的道德行为。

针对第二个问题，本文根据《现代汉语常用词表》(张清源, 1992)整理出基础道德词表，并通过词向量对基础词表进行扩展。道德词典由2,777个词扩展到29,907个词，经过标注核查后最终包含25,012个有效词，减少了人工标注的成本，高效地扩展了道德词典的规模。

最后，为了检验词典的有效性，本文从词的标签及类型识别和判断句子道德倾向两个维度进行了实验。实验结果表明，道德词典资源能够较为准确地判断词的标签和类型，也能够较好地判断句子的道德倾向。

综上，针对中文道德词典资源缺乏的问题，本文提出了面向人工智能伦理计算的中文道德词典构建任务。本研究的主要贡献包括以下三个方面：

- 提出面向人工智能伦理计算的中文道德词典构建任务，将动词、名词、形容词、成语以及汉语的被动表达纳入道德词典体系中，设计了中文道德词典的理论体系，包含四类标签和四种类型。
- 通过词向量扩展和人工核查，构建了包含25,012个词的中文道德词典资源，尽可能全面地覆盖了各类道德行为。
- 为了验证词典的有效性，本文从词的标签及类型识别和判断句子道德倾向两个维度进行了实验。实验结果表明，道德词典资源能够较为准确地判断词的标签和类型，也能够较好地判断句子的道德倾向。

2 相关工作

道德判断是一个传统的哲学问题。近些年来，社会心理学和认知语言学等领域也出现了跨学科的研究，但对道德倾向的大规模形式化处理，特别是道德的分类，仍处于自然语言处理的初级阶段。随着机器获得更多的自主性，需要以更精细的方式来进行伦理计算，使其基于道德进行决策。(Dennis et al., 2016)

道德观念的研究可以追溯到情感分析任务。道德价值观被认为是人格特征更高层次的组织结构。而情感分析中对人格(Schwartz et al., 2013; Yarkoni, 2010)和人类价值(Boyd et al., 2015; Chen et al., 2014)等的评估，为分析人类的道德观念提供了基础。但是，情感词典侧重的是人

第十九届中国计算语言学大会论文集，第539页-第549页，海南，中国，2020年10月30日至11月1日。

的主观情绪，道德词典则是研究客观的事件行为。例如，“杀人”这一事件，人们对它的情感态度可能是“愤怒”、“震惊”等，而道德词典则判断其是一个“不道德”的行为。

目前，道德词典的研究可以分为理论体系设计和词典资源构建两部分。

2.1 理论体系设计

道德基本理论(Moral Foundations Theory, MFT)(Haidt and Graham, 2007; Haidt and Joseph, 2004; Graham et al., 2011)解释了道德的起源、心理基础、发展和文化差异，被广泛应用于计算社会科学领域。该理论定义了五个明确的分类，每个分类包括美德和恶习两个维度，分别如下：

关心/伤害 这一分类与我们作为哺乳动物的长期进化有关，我们拥有依恋系统和感知他人痛苦的能力。它是善良、同情等美德的基础。

公平/欺骗 这一分类与人类社会的互惠性、利他主义有关。它产生正义、不平等和权利等观念。

忠诚/背叛 这一分类与我们部落联盟的悠久历史有关。包含有爱国主义、忠诚、自我牺牲精神等。

权威/颠覆 这一分类由我们长期的等级社会历史形成。包括对社会秩序的维护、对合法权威的尊重和对传统的继承等。

纯洁/堕落 这一分类由宗教观念发展而来，即努力以一种高尚的、不那么肉欲的方式生活。包括贞洁、健康和控制欲望等。

道德基本理论提供了五种典型的道德类型，分类定义非常具体，因此，有大量的道德行为无法简单归类到这五种类型之中，这一点限制了其体系覆盖道德行为的全面性。另外，道德作为一种文化现象，在不同国家及文化背景下有其独特之处，MFT理论并不适宜直接套用于汉语体系。

Jentzsch等(Jentzsch et al., 2019)通过设计道德选择的问题模板，覆盖了更多类型的道德行为。模板从第一人称出发，以问句形式呈现（如表1所示）。标注者将待标注词填入模板，如“我应该杀人吗？”，答案模板为“应该/不该”。这些问题使标注者可以从决策层面判断动作行为的对或错。但目前的模板问题只适用于动词，难以对名词、形容词等其他词类进行判断，限制了其体系覆盖道德行为的规模。

问题	答案
XX是可以的吗?	是/否
我应该XX吗?	应该/不该
我必须XX吗	是/否
我可以XX吗?	是/否
我被允许XX吗?	是/否
XX是被提倡的吗?	是/否
XX是被要求的吗	是/否
XX是礼貌的吗?	是/否
XX是好的行为吗?	是/否
XX是一种典范行为吗	是/否

Table 1: 道德选择问题模板

道德维度	美德	邪恶
关心/伤害	95(16)	85(35)
公平/欺骗	69(26)	57(18)
忠诚/背叛	99(29)	72(23)
权威/颠覆	160(45)	101(37)
纯洁/堕落	97(35)	161(55)
合计	520(151)	476(168)

Table 2: MoralStrength分类及扩展情况（括号内为扩展前数量）

2.2 资源建设

第一个用词汇进行道德判断的语言资源是道德基础词典(Moral Foundations Dictionary, MFD)(Graham et al., 2009)。该资源使用道德基础理论的分类体系和极性标签。词典包含151个美德词和168个邪恶词，词典构建完全依赖人工标注，构建成本较高，不利于扩大规模。

MoralStrength(Araque et al., 2020)在道德基础词典分类的基础上，通过WordNet词汇数据库对MFD进行了扩展。扩展前后各分类的分布情况如表2所示。经过人工标注核查后，得到了包含520个美德词和476个邪恶词的数据集。

从以上研究可以看出，现有的英文道德词典词性构成单一，覆盖的道德行为较少，不利于扩大规模；且分类标准立足于英语文化，不能直接应用于汉语。因此，本文使用汉语文化思

维，将词性类型扩大到动词、名词、形容词、成语以及汉语中的被动表达，丰富中文道德词典词性构成，覆盖更多的道德行为。

3 道德词典理论体系

本文将道德定义为具有普适性的行为规范，分为四类标签和四种类型。具体类别及示例如表3所示。

本章将首先介绍道德词典中道德的界定和理论基础，然后介绍道德词典的分类体系，包括标签体系和类型体系两部分。

类型 \ 标签	正向道德	负向道德	中性	被动
行为	捐献	拐卖人口	询问	被害
状态	淡泊名利	惨无人道	安于现状	受尽屈辱
属性	传统美德	传销活动	事业	涉嫌诈骗
要素-地点	福利院	黑作坊	加工点	被殖民地
要素-对象	活雷锋	无良商家	儿童	弃婴
要素-媒介	爱心专座	违禁物品	小作坊	被盗物品

Table 3: 道德词典分类体系及示例

3.1 道德界定及理论基础

本文将道德定义为具有普适性的社会行为规范，根据词汇本身体现的行为信息进行道德判断，探讨道德上要求、禁止或允许的事件行为。

一方面，本文将道德的范围限定为具有普适性的社会行为规范，如“恶意谋杀”这一行为，普遍被认为是不道德的。这一限定可以缩小道德判断问题的范围，回避道德困境类问题。自动驾驶汽车是否应该撞向障碍物，危及乘客，以避免与乱穿马路的人相撞？对于这类问题，没有普适性的行为规范进行判断，不在我们讨论的范围内。另一方面，本文以非结果主义中的规则义务论为指导，根据行为本身的特征或行为体现的规则来判断行为本身是否具有道德价值，将道德行为规范定义为“应该做什么”和“不应该做什么”的普遍规则。对非结果主义者来说，杀死某个人是错误的，因为杀人这一行为本身就是错误的；而对结果主义者来说，如果这个人正在杀死另外十个人的路上，这一杀人行为可能是正当的。由于结果主义的道德判断常常需要大量的背景信息，而在词汇级别，大多数词本身无法完整体现出其行为的结果，所以本文认为，以非结果主义作为指导去判断词的道德倾向更为合理。

综上，本文将道德的范畴限制为具有普适性的社会行为规范，并且只根据词汇本身所体现的行为信息进行道德判断。

3.2 道德词典分类设计

本文通过分析汉语中动词、名词、形容词、成语以及汉语中被动表达体现的道德特征，设计了中文道德词典的分类体系。分类体系包括四类标签和四种类型两部分。

3.2.1 道德词典标签分类

本文将道德词典中的标签分为正向道德、负向道德、中性、被动四类，如表4所示。

正向道德 符合社会道德规范的事件行为，即被认为是应该做、需要做、提倡做的好的行为。例如捐献、正能量、福利院、活雷锋等。

负向道德 不符合社会道德规范的事件行为，即被认为是不该做、不能做、禁止做的坏的行为。例如拐卖、惨无人道、传销、违禁品等。

中性 在大多数情况下与社会道德规范无关的事件行为。例如问话、人山人海、闹剧、食品加工点等。

被动 被动发生的、与社会道德规范有关的事件行为。例如被害、弃婴、饱受争议、涉嫌诈骗等。

标签	例词
中性	问话、人山人海、闹剧、食品加工点
正向道德	捐献、正能量、福利院、活雷锋
负向道德	拐卖、惨无人道、传销、违禁品
被动	被害、弃婴、饱受争议、涉嫌诈骗

Table 4: 道德词典标签分类及例词

3.2.2 道德词典类型分类

本文根据词在事件行为中的位置和作用，将道德词典中的类型分为事件行为、事件状态、事件属性以及事件要素四类，并为每个分类设计了对应的问题模板，如表5所示。

事件行为 这一分类针对的是动作行为本身，判断该行为是否符合道德规范，一般表现为动词。例如诈骗、出尔反尔、恶意透支等。

事件状态 这一分类是对事件行为的状态描述或评价，一般表现为形容词。例如合法合规、投机取巧等。

事件属性 这一分类是较为抽象的事件行为，或一系列事件行为的定义总称，一般表现为名词。例如封建思想、敬业精神等。

事件要素 这一分类是常常和事件行为的一起出现辅助因素，一般表现为名词。包括对象、地点、媒介。其中，对象指的是事件行为的参与者，例如逃犯、弃婴等。地点指的是事件行为发生的地点场所，例如黑作坊、非法赌场等。媒介指的是事件行为的媒介手段和工具。爱心专座、假币等。

类型	问题模板	例词
事件行为	是不是可以做/不能做的行为?	诈骗
	是不是应该做/不该做的行为?	出尔反尔
	是不是提倡做/禁止做的行为?	恶意透支
事件状态	是不是对某种道德/不道德行为的状态描述?	合法合规
	是不是对某种道德/不道德行为的评价?	投机取巧
事件属性	是不是某类道德/不道德行为的总称?	封建思想
	是不是抽象的道德/不道德行为?	敬业精神
事件要素	是不是道德/不道德行为发生的地点?	黑作坊
	是不是道德/不道德行为涉及的对象?	逃犯
	是不是道德/不道德行为使用的媒介工具?	爱心专座

Table 5: 道德词典类型问题模板及例词

4 道德词典构建

本文通过词向量扩展的方法生成了备选词表，并对备选词表进行了人工标注，得到包含25,012个词的中文道德词典。

本章首先介绍备选词表的生成方法，然后介绍标注方法及流程，最后对标注结果进行统计和分析。

4.1 备选词表生成

《现代汉语常用词表》(张清源, 1992)是具有权威典范性的中文词表，表中词汇覆盖范围较广，且具有常用性和代表性。本文首先按道德分类体系对《现代汉语常用词表》中的56,008个词进行标注，得到1,164个正向道德词，1,619个负向道德词，构成了包含2,777个词的基础道德词表。

从基础道德词表的标注情况可以看出，道德词的比例仅占《现代汉语常用词表》的5%左右，大量的道德词分散在其他汉语词汇当中。如果由人工一一进行标注，工程量大且构建成本较高。

因此，本文通过词向量对基础道德词表进行了扩展。对词表的扩展基于这样一个假设：如果一个词具有道德倾向性，那么在词向量空间中与其距离相近的词也可能具有道德倾向性。(Mikolov et al., 2013)因此，本文将基础道德词表作为种子，使用腾讯发布的AILab词向量(Song et al., 2018)和gensim(Rehurek and Sojka, 2010)计算找出与基础道德词表中每个词的余弦距离最近的十个词，构成扩展词表。扩展示例如表6所示。然后，将扩展词表去重并按照词性进行筛选，形成最终的备选词表。

种子词	扩展词				
绑架	绑架勒索 绑匪	绑票 遭绑架	劫持 要挟	胁迫 挟持	敲诈 绑架儿童
黑社会	黑帮 混黑道	黑道 黑老大	黑势力 黑道大哥	黑社会组织 黑帮老大	黑社会分子 黑社会老大
惨无人道	毫无人性 非人道	泯灭人心 残忍	惨绝人寰 屠杀	灭绝人性 残害	丧心病狂 兽行

Table 6: 词向量扩展示例

4.2 标注方法

本文参考中文情感词典的构建思路(柳位平et al., 2009; 饶洋辉et al., 2014; 赵妍妍et al., 2010)设计了人工标注的流程，通过一系列步骤保证标注结果的有效性。

标注内容 根据道德词的定义及分类体系，判断待标注词的道德标签（中性、正向道德、负向道德、被动）和类型（行为、状态、属性、要素）。

标注人员 招募10名培养层次为研究生的在校学生作为标注人员。

标注流程 两位标注人员为同一个词进行标注。导入待标注词后，标注员首先需要判断所给词是否具有道德倾向，标出其标签。然后，标注员需要判断所给词的类型。最后，标注员需要对标注词进行检查，检查无误后提交标注结果。最终标注示例如表7所示。

一致性 为保证标注工作的质量，此次标注设计了培训、试标环节来检查一致性。培训内容包括道德理论、词典体系以及标注流程等。标注人员通过培训熟悉标注规范后，对语料进行试标注，试标结果错误率超过10%的标注员将被劝退。正式标注期间标注的一致性为83.6%。每个词由两位标注员进行标注，两人标注结果不一致的词，交由第三个人进行核查，有争议的词将被剔除。

待标注词	标签	类型	备注
黑作坊	负向道德	要素	地点
私吞公款	负向道德	行为	/
弃婴	被动	要素	对象
无私奉献	正向道德	状态	/

Table 7: 数据标注示例

4.3 标注结果及分析

经过标注和一致性检查后，本文共获得有效标注结果25,012个（见表8）。

各标签的分布情况如图1所示。分析结果可得，正向道德词、负向道德词和中性词的规模相似，而被动词所占比例较少，仅占2%。被动类型由于涉及两个以上对象，道德行为比较复杂。经过进一步分析，发现其中正向道德和负向道德的比例大约是1：6，这是由于被动表达中涉及受害者的词汇较多造成的。

各类型的分布情况如图2、图3所示。行为词数量最多，占到词典总数的一半，其他三分类的比例相似。而要素的小类中，对象类明显较多，地点类所占比例最小。

5 道德词典的有效性验证

本文认为，道德词典资源的有效性可以体现在以下两个方面：一是机器能否通过道德词典

类型 \ 标签	正向道德	负向道德	中性	被动	合计
行为	3946	4155	3947	417	12465
状态	1787	1225	1914	26	4952
属性	1199	1069	1730	12	4010
要素-地点	76	27	78	/	181
要素-对象	546	798	712	32	2088
要素-媒介	358	373	582	3	1316
合计	7912	7647	8963	490	25012

Table 8: 道德词典分类结果

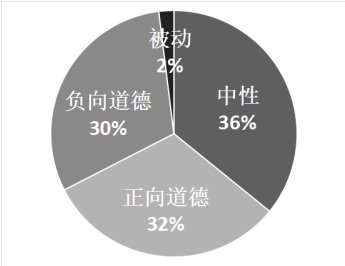

Figure 1: 标签分布

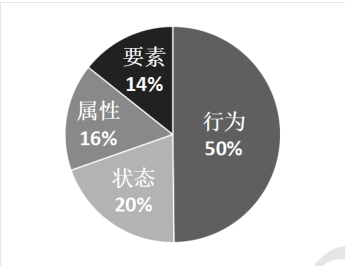

Figure 2: 类型分布

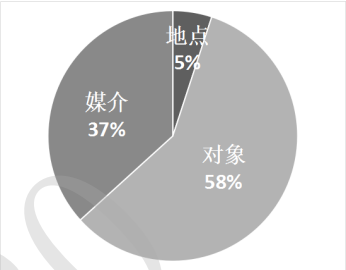

Figure 3: 事件要素分布

识别道德词的标签和类型；二是机器能否通过道德词典识别句子的整体道德倾向，辅助道德文本分析。

针对这两个方面，本文设计了两个对应实验，分别为词的标签及类型识别和判断句子的道德倾向。对于前者，本文使用逻辑回归（logistic regression，LR）和支持向量机（support vector machine，SVM）两种模型。对于后者，本文结合道德词典，使用两种方法对句子的道德倾向进行判断。

5.1 基于道德词典的分类实验

通过道德词标签和类型的识别情况，可以检验机器是否能通过道德词典学习到有效的道德知识。本实验使用带标签、类型信息的道德词典资源对模型进行训练，然后使用预留的测试集对模型的分类识别能力进行评估。

5.1.1 实验数据

本实验将道德词典按照8：1：1的比例划分为训练集、验证集和测试集。两个实验使用共同的测试集，确保不同实验的结果具有可对比性。输入模型的训练数据是词项对应的预训练词向量(Song et al., 2018)。标签信息则依据预测目标的不同，分别对应词典的标签列和类型列。

5.1.2 实验设计

预测词的标签 本实验使用两种思路对词的标签进行预测。

思路一：四分类标签预测。这一思路是将词典中的四种标签分类信息，即中性、正向道德、负向道德和被动，直接作为标签对模型进行训练。

思路二：两步分类标签预测。这一思路是先判断词是否包含道德倾向，即分为道德词（包括正向道德和负向道德）或其他词（包括中性和被动），再对两者分别进行正向道德或负向道德，中性或被动的分类。

思路一和思路二均先使用LR和SVM在验证集上优化参数，再对测试集进行预测，取测试集的结果作为最终的结果。

预测词的类型 本实验使用六分类的方法对词的类型进行预测。

六分类类型预测：本文将行为、状态、属性和要素-地点、要素-对象、要素-媒介六个分类设置为标签对模型进行训练，并对测试集进行预测。

考虑到数据的各个分类数量并不平衡，本文采用weighted average来计算F1。

5.1.3 实验结果及分析

表9展示了LR和SVM在三个实验中预测测试集的F1值。从实验结果可以看出，经过训练后，机器可以较好地掌握道德知识，对未知分类的词进行预测，给出可靠的标签。预测词的标签实验中，四分类标签预测结果好于两步分类标签预测。观察测试集输出结果发现，两步分类标签预测过程中存在误差叠加的问题，即第一步分类中有误差的结果会对第二步分类造成影响。预测词的类型实验中，虽然有的类型数据不平衡，如要素-地点类的数据量较少，但是模型仍能给出比较理想的结果。

实验名称	LR	SVM
四分类标签预测	0.75	0.81
两步分类标签预测	0.73	0.80
六分类类型预测	0.76	0.80

Table 9: 道德词典分类实验结果

三个实验中，分类结果较好的模型对测试集预测结果的混淆矩阵热力图如图4、图5所示。可以看出，直接四分类和两步分类实验中，中性类与负向道德、正向道德标签识别的错误率均比较高。类型分类的混淆矩阵热力图如图6所示。可以看出，行为类和状态类的区分较为困难。

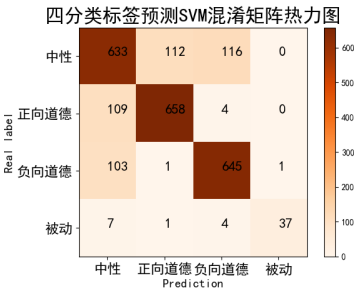

Figure 4: 四分类标签预测

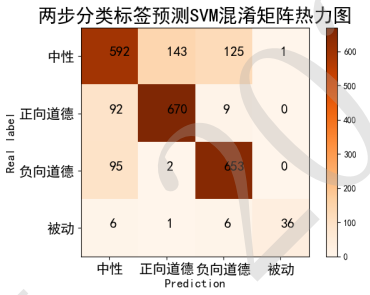

Figure 5: 两步分类标签预测

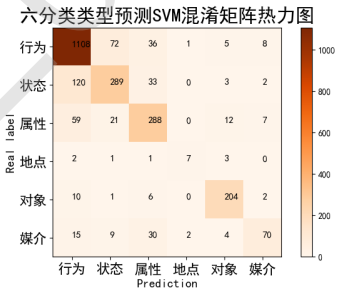

Figure 6: 六分类类型预测

对此，本文统计了人工标注环节标签和类型标注不一致的情况。从图7可以看出，最容易混淆的标签分别是中性-负向道德和中性-正向道德，说明正确区分中性-正向道德和中性-负向道德是一个难点，即使对人类标注者而言也很容易出错。从图8可以看出，容易混淆的类型是状态和行为，占不一致样本总数的40.2%。和图5中反映的规律相类似。

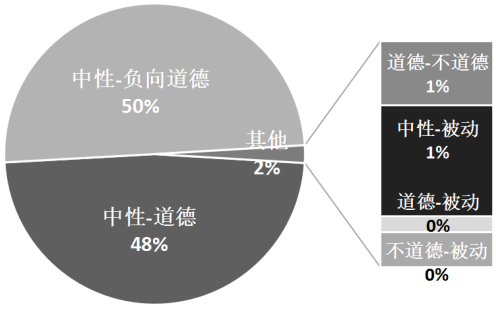

Figure 7: 人工标注标签不一致情况

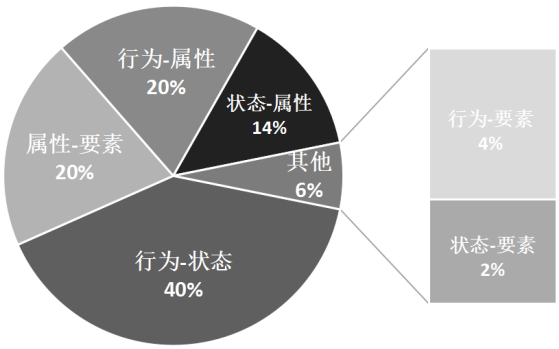

Figure 8: 人工标注类型不一致情况

此外，如表10所示，对一些有一定难度、存在迷惑性的词项，模型也可以很好地做出判断。如“不随地吐痰”和“随地吐痰”非常类似，但是标签相反，很容易误判为负向道德标签等。

5.2 基于道德词典判断句子道德倾向

本实验通过结合道德词典，使用两种方法判断句子总体的道德倾向性，以验证道德词典对判断句子道德倾向的有效性。

5.2.1 实验数据

本实验使用社会新闻类报道用作判断的句子来源，对2018-2019年新浪新闻社会新闻版块的新闻数据进行了爬取，收集了10.4万篇新闻数据。然后按句拆解每篇新闻，保留结构完整且句长为10-45词的句子，作为判断的来源句集。

5.2.2 实验设计

本实验使用两种方法对句子道德倾向进行判断。

方法一：对照(Yuan et al., 2013)使用情感词典对句子进行情感分析的方法，建立基于道德词典的道德分析。首先使用HanLP(He, 2020)对句子进行分词，分词前将道德词典添加到分词器的用户词典中，确保不会将道德词典中较长的词组切分开导致召回率下降；然后根据句子分词结果，统计其与道德词典中正向道德、负向道德、中性和被动标签的对应情况，将对应数量最多的标签作为句子整体的道德倾向。

方法二：道德词典与依存句法分析结合的道德分析。在方法一的基础上，对含有道德词典中词的句子进行依存句法分析。依存句法理论认为，每个句子中存在一个唯一的中心词，支配着句子中其他所有的词，其他词直接或间接依赖于中心词。同时，句子中除了中心词外，每个词都只被一个词支配(计峰and 邱锡鹏, 2009)。通过对句子做依存句法分析，可以理解句子中各部分的关系，以及各部分在各自关系中扮演的角色。根据不同类型的词在句子中常扮演的角色，本实验将依存句法分析中的几种关系与道德词典中的类型建立联系，对应关系如表11所示。其中，依存句法分析的关系体系中没有与地点和媒介两种类型较为合适的对应关系。最后，将句中词项对应的依存关系与该词项在词典中类型的对应关系进行对比，取词项类型对应的标签为句子的整体道德倾向。

词项	标签	预测的标签
妒贤	负向道德	负向道德
不随地吐痰	正向道德	正向道德
发生争吵	中性	中性
受到破坏	被动	被动

Table 10: 词项标签预测结果示例

类型	依存句法关系
行为	核心关系
状态	状中关系
属性	定中关系
要素-对象	主谓关系
要素-地点	/
要素-媒介	/

Table 11: 类型与依存句法对应关系

5.2.3 实验结果及分析

本实验从10.4万篇新闻报道中进行抽取，得到1,627,123条句子供处理。由于这些句子都是没有标签的数据，为了对方法一和方法二输出的结果进行评估，我们从两种方法的结果中各随机抽取了400条结果进行了人工标注，得到的结果如表12所示。

从表12可以看出，结合道德词典做词匹配的方法可以得到还不错的正确率，而结合依存句法分析的方法可以更为可靠地对句子的道德倾向进行判断。实验结果证明了道德词典在判断句子道德倾向上的有效性。

判断方法	正确率
方法一	65.67%
方法二	71.30%

Table 12: 两种判断方法结果

判断结果与句子实际道德倾向性不一致的情况，除去一些不满足普适性道德判断的句子，如有争议性的国际新闻等，比较典型的错误如：句子讨论的话题涉及道德，但句子整体的道德

倾向与句中道德词的倾向不同。如表13所示，方法一的例句中含有两个负向道德词，被判断为负向道德，但句子整体的道德倾向偏向于中性；方法二的例句中含有负向道德词“惯匪”，但句子整体的道德倾向是中性的。这个问题涉及句子中的语义信息，从词汇层面很难解决，未来我们会结合句子的语义信息进行完善。

判断方法	例句-错判标签	对应词项-标签
方法一	请广大群众做到不造谣、不传谣、不信谣。-负向道德	造谣-负向道德, 传谣-负向道德
方法二	社区人士呼吁广大民众对此类模式作案提高警惕，切勿因为嫌麻烦给钱了事，以免惯匪越发猖獗。-负向道德	惯匪-负向道德

Table 13: 道德倾向判断错误样例

6 结语

本文通过对词的道德倾向性进行研究分析，提出面向人工智能伦理计算的中文道德词典构建任务。我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。

同时，我们也探讨了道德词典资源的有效性表现。从词的标签及类型识别和判断句子道德倾向两个维度进行了实验设计。实验结果显示，该词典资源不仅能够判断词的标签和类型，而且能够较好地判断句子的道德倾向，为今后句子级别的道德文本分析提供了数据支持。

将社会伦理道德规范与科学技术创新结合是一条漫长的道路。目前词典的分类方法比较粗糙，未来我们会根据词的语义特征进一步细分，深入研究事件行为之间的语义关系，以便更好地解决人工智能伦理计算的道德判断问题。

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.

Ryan L Boyd, Steven R Wilson, James W Pennebaker, Michal Kosinski, David J Stillwell, and Rada Mihalcea. 2015. Values in words: Using language to evaluate and understand personal values. In *Ninth International AAAI Conference on Web and Social Media*.

Jilin Chen, Gary Hsieh, Jalal U Mahmud, and Jeffrey Nichols. 2014. Understanding individuals' personal values from social media word use. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 405–414.

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.

Louise Dennis, Michael Fisher, Marija Slavkovik, and Matt Webster. 2016. Formal verification of ethical choices in autonomous systems. *Robotics and Autonomous Systems*, 77:1–14.

Matthew Feinberg and Robb Willer. 2013. The moral roots of environmental attitudes. *Psychological science*, 24(1):56–62.

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.

Jesse Graham, Brian A Nosek, Jonathan Haidt, Ravi Iyer, Spassena Koleva, and Peter H Ditto. 2011. Mapping the moral domain. *Journal of personality and social psychology*, 101(2):366.

Jonathan Haidt and Jesse Graham. 2007. When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. *Social Justice Research*, 20(1):98–116.

Jonathan Haidt and Craig Joseph. 2004. Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus*, 133(4):55–66.

Han He. 2020. HanLP: Han Language Processing.

Wilhelm Hofmann, Daniel C Wisneski, Mark J Brandt, and Linda J Skitka. 2014. Morality in everyday life. *Science*, 345(6202):1340–1343.

Joe Hoover, Gwenyth Portillo-Wightman, Leigh Yeh, Shreya Havaldar, Aida Mostafazadeh Davani, Ying Lin, Brendan Kennedy, Mohammad Atari, Zahra Kamel, Madelyn Mendlen, et al. 2019. Moral foundations twitter corpus: A collection of 35k tweets annotated for moral sentiment. *Social Psychological and Personality Science*, page 1948550619876629.

Sophie Jentzsch, Patrick Schramowski, Constantin Rothkopf, and Kristian Kersting. 2019. Semantics derived automatically from language corpora contain human-like moral choices. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 37–44.

Rishemjit Kaur and Kazutoshi Sasahara. 2016. Quantifying moral foundations from various topics on twitter conversations. In *2016 IEEE International Conference on Big Data (Big Data)*, pages 2505–2512. IEEE.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Rosalind W Picard. 1997. A ective computing.

Radim Rehurek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Citeseer.

Eyal Sagi and Morteza Dehghani. 2014. Moral rhetoric in twitter: A case study of the us federal shutdown of 2013. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 36.

H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman, et al. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791.

Yan Song, Shuming Shi, Jing Li, and Haisong Zhang. 2018. Directional skip-gram: Explicitly distinguishing left and right context for word embeddings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 175–180.

Tal Yarkoni. 2010. Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers. *Journal of research in personality*, 44(3):363–373.

Bo Yuan, Ying Liu, and Hui Li. 2013. Sentiment classification in chinese microblogs: lexicon-based and learning-based approaches. *International Proceedings of Economics Development and Research*, 68:1.

张清源. 1992. 现代汉语常用词词典.

柳位平, 朱艳辉, 栗春亮, 向华政, and 文志强. 2009. 中文基础情感词词典构建方法研究. 计算机应用, 29(10):2875–2877.

计峰and 邱锡鹏. 2009. 基于序列标注的中文依存句法分析方法. 计算机应用与软件, (10):44.

赵妍妍, 秦兵, 刘挺, et al. 2010. 文本情感分析. 软件学报, 21(8):1834–1848.

饶洋辉, 李青, 刘文印, and 李晶晶. 2014. 公众文本之情感词典研究进展. 中国科学:信息科学, 44(07):825–835.