CoreValue：面向价值观计算的中文核心价值-行为体系及知识库

刘鹏远　　张三乐　　于东　　薄琳
北京语言大学信息科学学院
国家语言资源监测与研究平面媒体中心
北京市海淀区学院路15号, 100083
liupengyuan@pku.edu.cn　sanle0409@163.com　yudong@blcu.edu.cn　bolin_blcu@163.com

摘要

由主体行为推断其价值观是人工智能理解并具有人类价值观的前提之一。在NLP相关领域，研究主要集中在对文本价值观或道德的是非判断上，鲜见由主体行为推断其价值观的工作，也缺乏相应的数据资源。该文首先构建了中文核心价值-行为体系。该体系以社会主义核心价值观为基础，分为两部分：1）类别体系。共包含8大类核心价值，进一步细分为19小类双方向价值并对应38类行为；2）要素体系。划分为核心与非核心要素共7种。随后，抽取语料中含有主体行为的文本句，依据该体系进行人工标注，构建了一个包含6994个行为句及其对应的细粒度价值与方向，34965个要素的细粒度中文价值-行为知识库。最后，该文提出了价值观类别判别、方向判别及联合判别任务并进行了实验。结果表明，基于预训练语言模型的方法在价值观方向判别上表现优异，在细粒度价值类别判别以及价值类别多标签判别上，有较大提升空间。

关键词： 价值观计算；人工智能伦理；价值-行为体系；价值-行为知识库

CoreValue: Chinese Core Value-Behavior Frame and Knowledge Base for Value Computing

Pengyuan Liu　　Sanle Zhang　　Dong Yu　　Lin bo
Beijing Language and Culture University
National language resources monitoring and research print media center
15 Xueyuan Road, Haidian District, Beijing, 100083
liupengyuan@pku.edu.cn　sanle0409@163.com　yudong@blcu.edu.cn　bolin_blcu@163.com

Abstract

It is one of the prerequisites for artificial intelligence to understand and possess human values to infer their values from their behavior. However, in NLP related fields, the current research mainly focuses on the judgment of the values or morality of the text, rarely inferring their values from the subject's behavior, and also lacks corresponding data resources. This paper first constructs the Chinese core value-behavior frame. It is based on China's socialist core values and is divided into two parts: 1) category system. There are 8 categories of core values, which are further subdivided into 19 categories of bi-directional values and corresponding to 38 types of behaviors; 2) Factor system. There are 7 types of factors. Then, text sentences containing subject behavior are extracted from the corpus and manually labeled according to the system. Then, a fine-grained Chinese value-behavior knowledge base containing 6994 behavior sentences and their corresponding fine-grained values and directions, and 34965

©2022 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

第二十一届中国计算语言学大会论文集，第417页-第430页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

elements is constructed. Finally, this paper puts forward the tasks of value category classification, direction detection and joint discrimination. Experimental results show that the method based on the pretraining language model performs well in judging the direction of values, and has great room for improvement in fine-grained value category classification and multi-label value category classification.

Keywords: Values Computing , Artificial Intelligence Ethics , Value-Behavior Frame , Value-Behavior Knowledge Base

1 引言

人工智能正在对世界产生重大且深远的影响。一些基于人工智能的算法可代替人自动执行决策，或者说人授权这些算法可以自动执行决策，如自动驾驶、简历筛选(Dastin, 2018; Weed, 2021)、基于法律判决预测(Feng et al., 2022)进行自动判案、甚至是执行武器射击(Vynck, 2021)，而这些行为如果不加某种人类的伦理道德约束，可能会产生巨大风险。因此，人工智能治理(Munoz et al., 2016; Smuha, 2019; 中国国家新一代人工智能治理专业委员会, 2019)正日益受到重视，使人工智能或者机器具有人类伦理道德价值观意义与价值凸显。

近年来，机器伦理、机器道德领域的相关研究主要集中在：1）道德判断即判断某事件或行为是否道德(Prabhumoye et al., 2020; Zhou et al., 2021; Botzer et al., 2022; 彭诗雅 等, 2021)；2）基于社会规范的行为与后果推理(Forbes et al., 2020; Emelin et al., 2020; Lourie et al., 2020)；3）伦理道德的描述与建模(Prabhumoye et al., 2020; Schramowski et al., 2021, 2020)。此外，对社会偏见的检测、分类与消除(Sap et al., 2019; Blodgett et al., 2020)的相关研究也可纳入到机器伦理研究范畴。

在人工智能嵌入人类价值观方面的研究较少。价值观本身抽象、多样、难以具体描述，这一点可从价值观定义[0]上管窥一斑：价值观是一种外显的或内隐的，有关什么是“值得的”的看法，它是个人和群体的特征，影响着人们对行为方式、手段及目的的选择(Kluckhohn et al., 1948)。在科幻小说中，“机器人三定律”的情节说明简单的规则难以编码人类复杂的价值观(Asimov, 2004)。迄今为止，尚无方法对机器是否具有普遍的人类价值观进行测量(Müller, 2020)。虽然Hendrycks et al. (2020)尝试将人类共享的价值观与人工智能对齐，并对某行为做价值观是非判断，但该研究尚存在一些问题，即所基于的人工标注数据并没有预先确定标注者的价值观。这一问题与Talat et al. (2021)发现Jiang et al. (2021)研究中对道德进行判断所存在的问题类似。此外，虽然机器对各种具体行为的价值观或道德是非判断非常重要，但在做出是非判断之前，机器需要理解人类的具体行为是基于什么价值观做出的，这是机器真正理解进而具有人类价值观或者能够进行是非判断的前提。但目前鲜见将价值与行为统一起来建设的语义体系与数据资源，导致难以对机器是否理解、如何理解及如何判别主体行为的价值观进行深入研究。资源的匮乏已经成为制约该方向进一步发展的瓶颈与挑战。

本文首先对价值观体系进行分析梳理，在此基础上结合中国社会环境的特点与实际，确定以社会主义核心价值观作为中文价值-行为体系的基础。随后，从社会主义核心价值观中选取了八类核心价值，通过对大量新闻语料文本中含有主体[1]和具体行为实例的观察，发现价值-行为具有方向性，然后根据具体行为将价值进行进一步细分，将具体行为与价值及方向相对应，建立了中文核心价值-行为类别体系并进行了覆盖度验证；通过对具体行为的分析，得到价值-行为关联的要素，建立了价值-行为的要素体系，类别与要素体系共同构成了中文核心价值-行为体系。基于该体系，对价值-行为实例进行人工标注，每条实例均标注了细分价值与行为、价值-行为关联的所有要素，初步建立了一个中文核心价值-行为知识库。最后，还提出了价值观类别判别、方向判别以及联合判别三个任务并进行了初步实验与分析。本文贡献可总结如下：

1）建立了中文核心价值-行为体系。该体系分为两大部分：a）类别体系。包含8类核心价值，细分为19小类双向价值并对应38类行为；b）要素体系。为核心与非核心要素共7种；

2）依据上述体系进行人工标注，最终构建了一个包含6994个行为句及其对应的细粒度价值与方向，34965个要素的细粒度中文价值-行为知识库[2]并进行了初步分析，该知识库为进一步研

[0]研究者们对于“价值观是什么”一直众说纷纭，本文采用了大多数研究者的共识经典表达。
[1]本文研究中的主体，如无特别说明，均指个人或个体，这也是国内外价值观研究的重点。
[2]已开源在：https://gitee.com/NLUSOCO/CoreValue.git

(c) 2022 中国中文信息学会计算语言学专业委员会

究中文核心价值与行为之间的关系、将中文核心价值嵌入到人工智能等价值观计算等相关研究提供了数据基础；

3）为考察机器对文本中主体行为的价值类别与价值方向进行判别的能力，提出了价值观类别判别、方向判别及联合判别三个任务，并进行了基线模型实验和分析。

2 相关工作

2.1 资源

当前规模最大且最有影响的资源为SOCIAL CHEMISTRY 101(Forbes et al., 2020)，该资源面向社交与道德准则推理，包含十二个不同维度的人类关于社交、道德、预期文化压力以及责任承担等的判断，共包含450万个人工标注的类别标签以及上下文描述。基于这个资源，学者们又构建了一些相关资源如：Moral Stories(Emelin et al., 2020)，该资源共包含1.2万个短文本，主要目的是考察机器在社交情况下面向目标的道德推理与生成能力；ValueNet(Kim et al., 2022)，针对价值观约束下的对话进行研究，包含21374个文本场景的人类价值观。该数据集按Schwartz et al. (2012)的十个价值维度进行组织。

SCRUPLES(Lourie et al., 2020)也是一个针对道德判断的数据集。该数据集的标注对象是包括一个标题以及文本正文的描述真实生活的场景，对3.2万个场景共标注了62.5万个道德判断。在SOCIAL CHEMISTRY 101(Forbes et al., 2020)、Moral Stories(Emelin et al., 2020)、SCRUPLES(Lourie et al., 2020) 等的基础上，Jiang et al. (2021)构建了DELPHI数据集，共包含170万个在日常生活中的人类道德判断，其目标是使机器理解具有道德和社会准则，并具有相应判断的能力。CMOS(彭诗雅 等, 2021)，是中文道德句资源，共包含10万个人工判断的是否包含道德以及是否道德的文本句。

其他有影响的还有ETHICS(Hendrycks et al., 2020)和myPersonality(Kosinski et al., 2013)。ETHICS数据集通过人工撰写的方式共获得了超过13万个样例，每个样例由多个文本句构成。数据集包含公正、美德、义务论、利己主义、常识道德共五类。myPersonality是一个包含约15万Facebook用户状态与信息的数据集。从用户中提取了多个样本子集进行并完成了数十种问卷调查，这些问卷涉及人格评估、人口统计学信息和价值观。

2.2 体系

Perry (1926)最早将价值观分为六大类，即认知、道德、经济、政治、审美和宗教。Allport et al. (1960)将价值观也分为六大类：经济的、理论的、审美的、社会性的、政治的和宗教，该分类具有较大影响。Kluckhohn et al. (1948)从价值取向进行划分，总结为：人与自然的关系；理想人格类型；人与他人的关系的形态；时间评价和组织；人的本性。Rokeach (1973)的分类突破了上述类别的框架，认为价值观有终极性和工具性两个维度，并将价值观分为工具性价值观和终极性价值观两类。这样的分类将价值观更有层次和顺序的体现出来。

目前应用最为广泛的是Schwartz et al. (1987)的分类体系，他将价值观分为自我提高—自我超越，保守主义—开放性两个垂直维度，根据维度分为权利、成就、享乐、自主、刺激、博爱、慈善、顺从、保守和安全10类价值观。

国内的研究主要有：杨中芳 (2005)将文化价值体系划分为世界观、社会观和个人观三个层面，每类下再继续细分；黄希庭 等 (2005)将价值观分为人生、政治、道德、职业、人际关系、审美、婚恋、宗教、自我价值和幸福价值观；张进辅 (1998)认为价值观由价值目标、价值手段和价值评价维度组成，把价值观分为人生、政治、道德、职业、婚恋、消费、审美、人际、宗教、知识、教育价值观等。

综上，现有价值观体系主要是根据维度和类型进行分类，分类粒度较粗，对价值观类别细分类的研究较少。鲜见同时从价值观与主体行为两个角度出发建立两者统一的体系。此外，在价值观所体现的具体行为上，鲜见细粒度描述框架的研究。

3 中文核心价值-行为体系设计

人类有着复杂的思想，由于社会环境、所受教育和宗教信仰等的不同，人们的价值观也有所差异。不同文化和社会政治制度中可能存在不同的价值选择(Aizenberg et al., 2020)，比如我国孔子强调以仁、义、礼治天下，而西方的苏格拉底则注重幸福、正义与勇气。

(c) 2022 中国中文信息学会计算语言学专业委员会

党的十八大提出了社会主义核心价值观，即“三个倡导”，这全面概括了全党全社会的价值共识。其中国家层面的价值目标是：富强、民主、文明、和谐；社会层面的价值取向是：自由、平等、公正、法治；个人层面的价值准则：爱国、敬业、诚信、友善[3]。在价值观体系内涵层面，社会主义核心价值观不仅涵盖了广大人民群众的普遍愿望，更凸显了当今中国社会主流意识形态的核心价值理念，容纳了历史文化传统、鲜明时代精神和未来价值追求。社会主义核心价值观就是我国社会做出的价值选择。

3.1 数据来源

中文核心价值-行为体系的设计不但需对社会主义核心价值观内涵的正确理解，还需要找到对应核心价值观主体的具体行为并对其进行分析，因此需要在真实数据中进行考察。该数据需要包含人们日常生活中发生的各种事件。同时，本小节的目标是针对社会主义核心价值观进行价值-行为体系构建，因此希望数据中这些日常生活中发生的各类事件能与社会主义核心价值观有较强的相关性。本文选择的数据的来源为：

1）网络爬取的新闻语料。爬取的网站为中国文明网、青少年爱国主义网和搜狐新闻网；

2）中文道德句子库(彭诗雅 等, 2021)。以新闻文本及传记文本作为语料来源，约10万句。

以上数据源均为新闻语料，语体为书面语，语言使用客观、准确。从内容看，数据中包含大量的、多样的、真实的社会事件及生活事件，且所报道的事件多与当今社会主流或核心价值观相关。将上述语料进行分段、清洗和去重后，作为本小节设计中文核心价值-行为体系以及后续知识库构建的数据来源。

3.2 类别体系设计

3.2.1 核心价值选取

在价值观应用层面，社会主义核心价值观采取了一种开放性的表述方式，即对核心价值观进行直接列举，因此对每种具体相关的行为，其所对应的价值目标、取向、准则相对明晰，易于识别与标注。在十二个社会主义核心价值中，“富强、民主、和谐”体现在经济富强和政治民主、社会和谐，较为宏观，“自由”这一核心价值是指人们的意志自由、存在和发展的自由。在观察语料中我们发现：

1）体现“富强、民主、和谐”价值行为的实例通常以国家、集体为主体，而本文针对个体；

2）体现“自由”价值行为的实例过于宽泛，绝大多数行为都基于人们的“自由意志”。

最终，本文选取文明、公正、平等、法治、爱国、敬业、诚信、友善作为价值-行为类别体系的核心价值，共八种；同时，将从个人层面进行价值-行为类别体系及后续知识库的构建。

3.2.2 基于个人层面的社会主义核心价值内涵

8类核心价值的表述主要基于季明 (2013)对核心价值观的解读，本文根据个人层面行为所体现的价值观，略有调整：

文明。是个人素养、教养的重要体现，与社会个体在文化和道德品行上的素质紧密相关。

公正。就是有着不偏私、以公为首的思维方式，对待事物公平正直，没有偏私。

平等。指个人在社会关系、社会生活中处于同等的地位，具有相同的发展机会，享受着平等的权利和义务。

法治。知法、懂法、守法就是公民法治观念的体现。

爱国。是中华民族精神最稳定的文化基因，体现了人们对自己祖国最深厚的感情。是基于个人对自己祖国依赖关系的深厚情感，要求人们以振兴中华为己任，自觉促进民族团结、维护祖国统一、报效祖国。

敬业。对公民职业行为准则的价值评价，是一个人对自己职业的基本尊敬和负责的态度。对于每一个公民来说，敬业精神的内涵表现在三个方面：热爱自己的工作和所投身的事业；勤勉努力、付出劳动；克制自己恣意享乐、纵情狂欢的欲望。

诚信。诚实守信，包括诚和信两方面。“诚”的内容又包括两方面：一是为人真实，不有意歪曲客观事物的本来面貌，实事求是；二是信守承诺，指人说话要算数、讲信用，对自己的承诺负责，要言而有复，诺而有行。

友善。公民的核心价值规范之一，推动和谐社会的构建。友善指人与人之间和睦、友好、亲近,需要公民做到待人如己、宽厚、助人为乐，努力形成社会主义的新型人际关系。

[3]价值目标、价值取向、价值准则三者内容的具体体现均为价值，后续本文将用“价值”表述，不做深层次区分。

(c) 2022 中国中文信息学会计算语言学专业委员会

3.2.3 价值方向

具体行为在其所体现的价值上具有“方向性”，如“扶老奶奶过马路”的行为，与其关联的价值为“友善”；而“对儿童的求助视而不见”的行为，与其关联的价值也应认为是“友善”，但两者方向相反，即：前者行为与“友善”价值相符，后者行为与“友善”价值相悖。此外，不能就此推断行为主体是在有/没有（或富有/缺乏、认可/不认可）“友善”这一价值时分别做出的行为，也不应推断行为主体是在认可“冷漠”或“恶”等“负”价值情况下做出的行为。实际上，“文明、公正、平等、法治、爱国、敬业、诚信、友善”均为人们或者说行为主体共同享有的价值观，这一点并不会因为行为主体做出某些与价值相悖的行为而改变或在行为主体观念中消失，也不能简单推断行为主体认可“负”价值。

行为主体在基于某价值作出某种行为时，采取何种价值方向是主体的一种选择性注意(Treisman, 1964)，也就是说，行为主体有意识或无意识的选择注意/不注意该价值。更进一步，即该价值在行为主体脑中得到激活/抑制。如果某种价值观被激活，则行为主体会做出符合某种价值观的行为，反之则反，即有两种价值方向：激活（↑）与抑制（↓）。如：“一位年过六旬的老教师依然奋战在教学一线，坚持手写教案”。结合价值观相应的内涵，我们可以推断出“一位年过六旬的老人”是在激活了敬业这一价值的情况下，做出了“坚持手写教案”这一行为。反之，如：“这位银行员工在上班时间玩斗地主”。则由于“这位银行员工” 抑制了敬业这一价值观，做出了“在上班时间玩斗地主”的行为。

此外，存在价值互相影响的情况，如：“为了践行承诺，他带头清理垃圾”。从行为“带头清理垃圾”来看，该行为与“文明”这一价值相关，但该行为是在行为主体为了“诚信”（“践行承诺”）这一价值而做出的。此类现象仍然可以用价值“激活”/“抑制”解释，即：行为主体激活了“诚信”这一价值，做出了“带头清理垃圾”的行为。对这种情况，由于行为主要是由“诚信”价值导致的，因此本文暂不做“文明”这一价值是否被激活/抑制的判断。

3.2.4 基于行为实例的价值-行为类别细分

在明确了8类价值观内涵的基础上，本文通过观察语料中包含主体及其行为的实例，尝试将其按照核心价值归类。在归类过程中发现，单独从每类价值观指导的具体行为来看，主体所持有的价值观仍有明显差异，且能够进一步细分为几个类别。如在“文明”价值观下，还可以进一步细分为：公共文明（爱护公物、环境等相关行为）、仪表文明（个人卫生、仪表等相关行为）、言语文明（礼貌用语等相关行为）和思想文明（宣传文明理念等相关行为）。

对核心价值进行更细粒度的划分有助于更深层次的理解核心价值的外延与内涵。同时，对价值-行为做更细粒度的刻画，能够对行为主体对应价值进行细化区分，有助于深刻理解和把握价值引导下的主体行为差异，并为分析价值-行为的关系提供了更精细的视角。在已有的核心价值观相关研究中，鲜见对核心价值进行进一步细分的系统性研究。于是本文尝试从语料中的具体行为入手，通过对语料库中主体的真实行为实例进行分析，将其与核心价值关联，进一步考察各个实例间价值的差别，细分价值-行为，并将相似价值-行为进行比较、整理归纳与合并，自底向上进行核心价值体系的构建。尽量避免自顶向下划分类别时类别划分粒度不易把控，特别是一些过细的类别难以在真实语料中找到对应具体实例的问题。归类与细分主要依据：1）行为所依据的具体价值内涵、外延与程度；2）行为主体自身的特点；3）行为所作用的对象。

具体而言，如“见义勇为”和“互相帮助”两类均是体现“友善”价值的行为，虽然“见义勇为”也是“互相帮助”的一种体现，但“见义勇为”包含了更多的价值，或者说比一般的“互相帮助”的行为“价值”更高，事件主体甚至有“舍己救人”的可能，且这类的行为一般发生在紧急、危险的时刻，因此本文将“见义勇为”这一行为类别单独分类。如“关心慰问”体现友善价值的行为，本文认为这一行为类别行为的价值主要体现在看望、慰问、关心，其作用对象的范围很广，与“孝顺长辈”等行为所体现的价值内涵不同。又如对“平等”中包含“人格平等”，本文将行为主体的性别、地域、种族特点独立出来，细分为“性别平等”、“地域平等”和“种族平等”这三类价值。最终，核心价值次分类共包含19小类，对应38类行为。文明价值包括思想文明、公共文明、言语文明和仪表文明；公正价值包括思想公正、机会公正；平等价值包括思想平等、人格平等；法治价值观包括知法懂法、守法用法；爱国价值包括思想爱国、以身作则；敬业价值包括热爱岗位、忠于职守；诚信价值包括传播诚信、诚实待人、信守诺言；友善价值包括乐于助人、宽厚待人。表 1是“文明”这一核心价值的类别体系，包含了价值-行为类别细分，以及激活/抑制该价值对应的行为实例。包含全部核心价值的完整类别体系详见附录A。

(c) 2022 中国中文信息学会计算语言学专业委员会

价值次类	行为类	价值激活行为示例	价值抑制行为示例
思想文明	宣传学习	宣传塑料危害	封建迷信
公共文明	爱护公物	保护古城墙	破坏共享单车
	爱护环境	清理垃圾	乱扔垃圾
	遵守秩序	按序排序	霸占座位
	积极参与	参加水资源保护活动	
言语文明	用语礼貌	礼貌询问	骂脏话
仪表文明	个人卫生	饭前洗手	饭桌抠脚
	穿着服饰	着装合适	坦胸露乳

Table 1: “文明”价值类别体系

3.2.5 类别体系覆盖度验证

为验证类别体系对现实行为的覆盖程度，本文在3.1小节中的第一类数据来源即网络爬取的新闻语料中，随机抽取并人工筛选得到1000条行为主体为个人且其行为是基于8类核心价值观的句子，然后由标注员（两名语言学与应用语言学的硕士生）将其归类到类别体系，无法归类的单独列出。最终结果由另一名标注员进行统计。统计结果表明，本文的类别体系可覆盖96%新闻事件主体的行为，未覆盖的行为，表现为低频长尾分布。未来将考虑增设“其他”类别，对体系进行进一步完善。较高的覆盖率表明本文分类体系能够较好的覆盖真实的新闻语料，能够对当下中国社会环境的主流价值观和相应的行为进行较为全面的归类。

3.3 要素体系设计

主体的具体行为与信息抽取中的事件类似，都表示动作的发生或状态的变化，需要进行分解与表示才能准确刻画。ACE2005(Walker et al.)中事件的表示为：事件触发词、事件类型、事件论元和其他论元角色。如对例句：“英美轰炸伊拉克”，其事件类型为：攻击；“轰炸”是“攻击”事件的触发词；“英美”与“伊拉克”均为论元角色。但对于价值-行为，由于日常生活中的行为及影响因素众多，基于新闻事件的表示并不完全适用。如对例句：“陈某在正在行驶中的公交车上强行踩下刹车...”，按照事件表示分析，可认为事件类型是“法制”，“踩下”或“踩下刹车”是触发“法制”事件的触发词，但这些触发词所能触发的事件类型非常广泛，如果没有例句提示，很难将“踩下”或“踩下刹车”与“法治”这一价值观相联系。类似的，日常生活中的各类常见行为如：“走，拿，拉，上，喜欢，奔跑”，单从这些特定动词来看，基本都无法与特定价值观相联系。此外，行为的表示需要比事件表示更为具体，很多时候需要事件表示以外的元素才能与特定价值观相详细，如上例中的行为：“强行踩下刹车”，在识别这个具体行为依据的价值观时，“强行”这个修饰语起到很大作用；此外，行为发生的地点“正在行驶中的公交车上”，也是帮助价值观识别与判断的重要因素，修饰语“正在行驶中的”同样不能忽视或省略。

考察行为句中各个因素与价值观的关联，目的是帮助推断与考察价值观与行为间的关联。本文参考ACE2005对事件的表示方法，没有设定价值观触发词，而是直接将行为分解为7个要素，如表 2所示。其中，主体与行为是组成一个具体行为必不可少的两个要素，前提（动机、目的、意图、条件）是主体价值选择的基础，很大程度上决定主体的行为。因此，本文将主体、行为、前提这三个要素设置为**核心要素**，其余四个要素虽然对行为主体的价值推断有所影响，但相对次要，设为**非核心要素**。

4 中文核心价值-行为知识库构建

4.1 价值-行为句筛选

价值-行为句根据以下三个条件进行筛选：1）需要同时包含主体及其主观行为，但不包括描述感觉、感受或情感的行为。去除如“立即将执法车停在左侧道路”—缺少主体、“某人...感觉很冷”及“某人...初尝到骗保的甜头”—是对感觉感受的描述；2）句中主体严格限定为“某一个人”。需要去除如“他们开始在超市疯狂购物...”—行为主体为两或多人。3）句中主体的行为所基于的价值观需包含在八种核心价值观中，且价值观可明确推断。去除如“某人手持警务通、通过系统查询获悉了女孩身份信息”—难以推断出主体行为是在何种价值观下做出的、“为了省

(c) 2022 中国中文信息学会计算语言学专业委员会

要素	定义	文本实例	要素实例
主体	行为的实施者	...男子将车停在...	男子
行为	主体主观做出的行为	...公交车上强行踩下刹车...	强行踩下刹车
前提	行为主体的动机、目的、意图、条件	...为了践行承诺，他...	践行承诺
对象	行为所针对或作用的客体	...哄骗李某到其家中...	李某
工具	行为主体所用的工具、方式、手段	...用铁锹殴打他...	铁锹
时间	行为发生的时间	...在半夜大声唱歌...	半夜
地点	行为发生的地点	...在马路中央刷抖音...	马路中央

Table 2: 要素定义与示例

	文明	公正	平等	法治	爱国	敬业	诚信	友善	合计
激活（↑）	376	213	167	88	500	1776	425	3006	6551
抑制（↓）	487	167	331	1141	40	447	410	1384	4407
总	863	380	498	1229	540	2223	835	4390	10958

Table 3: 价值-行为粗粒度类别标注统计结果

钱，某人偷偷地在垃圾桶里捡同学们用剩的铅笔..."—虽然可以推断出其价值观是"节约"，但这一价值观并不在八种核心价值观之中。

按上述条件，从本文3.1小节的数据来源中随机抽取文本句13000句左右，经筛选并得到符合要求的句子共7130句作为下一步的标注对象。

4.2 标注原则

价值-行为知识库标注的四项基本原则为：1）要素需要保持完整，包含关联的修饰语；2）对价值观进行推断的主要依据是某个行为直接关联的要素（含行为）而不是根据整句，以避免多个行为的影响；3）为避免仅根据主体行为直接做出价值观判断，必须考察所有要素，并标记除主体/行为之外，哪个（些）要素可推断对主体价值观；4）若存在多个主体行为，则仅标注出体现八类核心价值观的行为，忽略其他行为。

4.3 标注过程与质量控制

为了确保标注质量，标注分为四个阶段：培训、试标注、二次培训以及正式标注。在培训阶段中，标注人员需要熟悉理论背景、标注规则和标注流程。在试标注阶段，标注人员对200条语料进行试标注，试标注正确率低于80%的标注人员将被劝退。试标注阶段后，对试标注中出现的问题对标注员进行进行二次培训。在正式标注阶段，标注工作分批次进行，每批次中的每条句子均由两位标注员进行标注。每批次标注语料由第三名标注员对标注不一致的数据进行详细核查。若标注正确率低于85%语料则需要全部核查，标注正确率低于80%的语料则需要重新标注。部分标注不一致的句子将交由第三名标注员进行讨论确定，有争议的句子将会被删除。由于需要标注要素以及粗细粒度类别，故将整个标注分为两个任务：1）价值-行为粗粒度类别及要素标注；2）价值-行为细粒度类别标注。两个任务的标注流程与质量控制方法类似。

标注质量。第一个任务的总体标注一致率为85.52%，第二个任务中次分类标注一致率为90.8%,行为归类标注一致率为89%。此外，对所有标注不一致的句子，均进行讨论后重新进行了标注。

4.4 构建结果

最终有效标注结果共6994句，含34965个要素。其中1128句与两或多个价值相关联。表 3与表 4是价值-行为粗粒度类别标注与要素标注统计结果，价值-行为细粒度统计结果见附录A。

4.5 初步分析

基于本文所建立的价值-行为知识库，可以从价值-行为粗细粒度分布、价值-要素、行为-要素等多个角度对相关规律进行分析与研究。限于篇幅，本文仅以"文明"为例列出其细粒度价值

(c) 2022 中国中文信息学会计算语言学专业委员会

	主体	行为	对象	原因	工具	地点	时间
激活（↑）	6418	6418	4285	1095	340	1180	2754
抑制（↓）	3592	3592	2505	434	625	740	987
总	10010	10010	6790	1529	965	1920	3741

Table 4: 价值-行为要素标注统计结果

分布与行为比例，并以“文明”与“公正”为例进行最重要的核心要素-行为进行初步考察。分布见图1-2，词云见图3，其中↓与↑分别指抑制与激活。

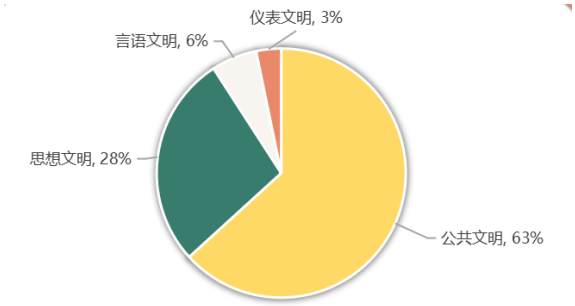

Figure 1: 次类分布（文明-行为类）

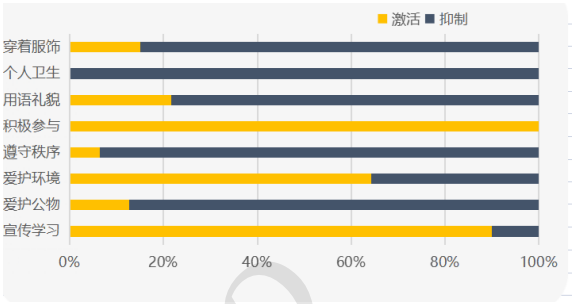

Figure 2: 激活抑制占比（文明-行为类）

从价值-行为次分类分布来看，占比从高到低依次为：公共文明>思想文明>言语文明>仪表文明。体现出人们更关注个体在社会中的文明表现，更关注宣传学习文明的行为。在文明价值激活与抑制方面，除“积极参与”、“爱护环境”与“宣传学习”外，其余类别均是价值抑制多于激活。侧面反映当下社会倾向在前三类文明激活的行为中塑造文明价值观，以“穿着服饰”、“个人卫生”、“用语礼貌”等文明抑制的行为作为文明价值的负向建构。

在要素方面，抑制与激活时两类价值的词汇各具特色。当主体的文明价值抑制时，主要会做出：1）不遵守规则的行为，如“逆行、闯红灯”等；2）个人素养低下的行为如“扔、破坏”等。反之，主体经常会成为文明知识/思想的传播者或学习者。当主体的公正价值抑制时，虽然会有“帮助、接受”等正面词，但是实际此时的主体是置其他人利益于不顾，在办事时多“帮助”自己的亲戚朋友，枉顾公正原则。同样，我们从“拒绝、坚持”这样的词也可得知主体在公正价值激活时会拒绝不公正现象、坚持原则、秉公办事。

Figure 3: 核心要素—行为的词云（文明↓、文明↑、公正↓、公正↑）

5 价值观计算任务

对给定主体行为的文本句，要使机器真正理解人类价值观，需要机器具有判别主体行为所基于价值的能力、判别主体做出行为时所处价值方向（激活/抑制）的能力。为此本文设计了三个任务：1）价值类别判别；2）价值方向判别；3）价值类别方向联合判别。

5.1 任务定义

价值类别判别。多分类任务。输入是一个单价值行为句[4] $X = x_1, x_2...x_i...x_m$，机器需要判别并输出对应的价值观标签 $Y \in \{y_1, y_2...y_i...y_n\}$。其中 x_i 为字或词，m为文本长度，y_i 为价值类别，n为价值类别总数。

[4]仅包含一个主体行为且该行为仅体现某一类价值观的文本句子。

(c) 2022 中国中文信息学会计算语言学专业委员会

	类别判别		方向判别		联合判别	
Baseline	ACC	F_1	ACC	F_1	ACC	F_1
Random	11.7	9.7	52.5	59.5	5.5	4.4
Majority	43.2	7.5	67.4	40.3	33.3	3.1
BERT	89.6	**85.7**	97.0	96.6	85.5	69.6
RoBERTa	89.2	85.3	98.2	**98.0**	88.0	**71.0**

Table 5: 价值类别判别、价值方向判别及价值类别方向联合判别的实验结果（%）

价值方向判别。二分类任务。输入是一个单价值行为句$X = x_1, x_2...x_i...x_m$以及对应的价值$C \in \{c_1, c_2...c_n\}$，机器需要判别并输出对应的方向标签$Y \in \{y_1, y_2\}$。其中x_i为字或词，m为文本长度，c_i为价值类别，n为价值类别总数，y_i为方向：激活（↑）/抑制（↓）。

价值类别方向联合判别。多分类任务。输入是一个单价值行为句$X = x_1, x_2...x_i...x_m$，机器需要同时判别对应的价值与方向并输出价值方向联合标签：$Y \in \{y_1, y_2...y_i...y_n\}$。其中$x_i$为字或词，m为文本长度，$y_i$为价值方向联合标签如：文明激活（↑）、文明抑制（↓）等。

5.2 实验

5.2.1 实验设置

实验数据。通过保留单一行为要素的方式，形成10010条仅包含一个行为的句子，其中有948条与两个价值相关，其余句子为单价值行为句共9062条，按照约8:1:1的比例划分为训练集（7250句）、验证集（906句）和测试集（906句）。

基线模型。1）Random。即随机选取类别标签作为分类结果；2）Majority。即选取频次最高的类别标签作为分类结果。鉴于目前在文本分类任务上基于预训练微调的方法性能较好，因此本文选取了两个基于预训练语言模型微调的方法：3）BERT(Devlin, 2018)、4）RoBERTa(Liu et al., 2019)。

参数设置。BERT为bert-base-chinese[5]默认设置，RoBERTa为chinese-roberta-wwm-ext[6]默认设置。均采取max-len为256，batch-size为64，learning-rate为1e-05。

评价指标。分类问题中常用的评价指标为准确率（（ACC, Accuracy）、精确率（P），召回率（R）与F值（$F-Score$）。本文使用准确率（ACC）与能综合反映分类器性能的宏平均F值（F_1），宏平均F_1值可视为多个二分类F-Score的算数平均值。

5.2.2 实验结果

三个任务的实验结见表5。对价值方向判别任务，BERT与RoBERTa的性能表现均十分优异，F_1值均超过了96%，这说明价值激活和抑制时所做出的的行为在语义上区分明显，预训练语言模型能够根据行为以及给定的价值做出正确的价值方向判断。对价值类别判别任务，BERT与RoBERTa也取得了较好的结果，F_1值均超过了85%，这说明各个价值的行为之间具有较好的语义区分度。表现最差的是价值类别与方向联合判别任务，F_1值在70%左右，RoBERTa模型的性能比BERT略高。联合判别任务需要同时判别价值类别与价值方向，因此这是一个16分类任务，难度低于分别为8分类与2分类的价值判别与价值方向任务，而且类别判别与方向判别可能会互相影响，故模型性能较差。

图4是RoBERTa模型在价值类别方向联合判别任务中各个价值方向上的性能比较柱状图，目的是考察模型对各个价值及不同方向判别的性能。模型对价值激活或抑制的判别上，不同价值方向的性能并不相同，且没有明显规律。值得注意的是，虽然“爱国”的样本句子数较少，约仅占总样本句子的6%，但其类别与方向的性能均为最高。经考察语料发现，“爱国”这一价值关联的行为，语义较为单调，“思想爱国”、在爱国的事情上“以身作则”这两类行为与其他价值关联的行为区分度较高。

5.2.3 讨论

以上三个任务，仅考虑了单价值行为句，而真实语料中确实存在一个行为句中与多（两）

[5]https://github.com/google-research/bert
[6]https://github.com/ymcui/Chinese-BERT-wwm

(c) 2022 中国中文信息学会计算语言学专业委员会

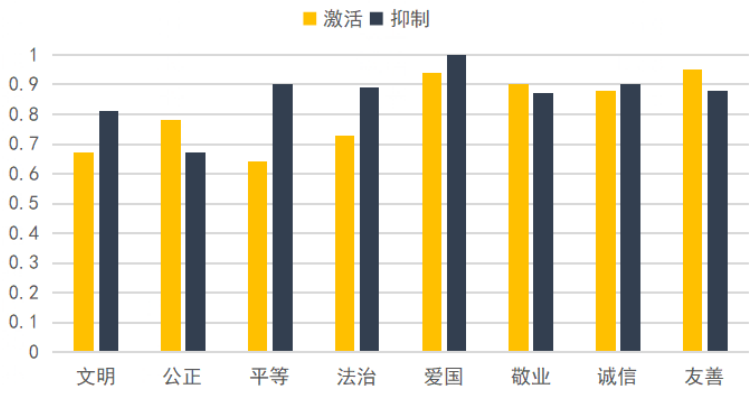

Figure 4: 基于RoBERTa的价值类别方向联合判别个价值方向别性能比较（F_1）

	多价值判别-所有			多价值判别-多标签			细粒度判别	
baseline	P	R	F_1	P	R	F_1	ACC	F_1
BERT	89.6	75.2	81.2	28.2	29.4	28.7	82.8	**54.2**
RoBERTa	89.7	78.6	**83.6**	42.8	32.3	**33.9**	83.5	53.8

Table 6: 多价值判别与细粒度判别的实验结果（%）。其中**多价值判别-所有**为对该任务所有样本进行性能评价的结果，**多价值判别-多标签**为仅对多标签样本进行性能评价的结果。

个价值观相关联的情况。此外，价值分类判别任务仅进行了8分类的粗粒度价值判别，而实际上本文所构建的知识库，包含更细粒度的价值共19类。考虑以上两点，本小节做了两个补充实验：1）**多价值判别**：含有多个行为多个价值句子的价值分类判别实验，一个句子可以有两或多个价值标签，因此是一个多标签分类任务；2）**细粒度判别**：针对细粒度的单价值行为句，进行19分类的价值分类判别。

多价值判别实验数据：将知识库中的全部文本句共6994条，按照约8:1:1的比例划分为训练集（5595句）、验证集（699句）和测试集（700句）。**细粒度判别实验数据**：知识库中的全部单价值细粒度行为句共9025条[7]，按照约8:1:1的比例划分为训练集（7220句）、验证集（903句）和测试集（902句）。**基线模型**、**参数设置**同前。**评价指标**：细粒度判别任务，同前；对多价值判别任务，采用了多标签分类常用评价指标宏平均P、R、与F_1。

实验结果见表6。对多价值判别任务，BERT与RoBERTa的表现良好，后者略高于前者，但由于其中84%都是单价值标签的样本，因此基于所有样本得到的性能难以说明模型对多标签样本的价值判别能力。表6中的"多价值判别-多标签"是仅针对多标签样本进行单独统计得到的结果。多标签样本数量占总样本的16%左右，均为每个样本2个价值标签。由于样本较少，且一般多标签分类相比单标签分类难度更高，BERT与RoBERTa的表现均不佳，F_1值不足34%。表6最右侧两列是细粒度判别任务的结果，可知BERT与RoBERTa的性能基本类似，且均较差，F_1值均在55%以下，原因可能是：1）由于是19分类，每个细粒度中的样例较少且不平衡；2）细粒度之间行为的语义较难区分。

6 结语

本文基于核心价值观建立了首个面向价值观计算的中文核心价值-行为体系及相应的知识库，该知识库可支持中文价值观计算与分析。基于该知识库，本文提出了3个价值观计算任务并进行了实验，实验结果表明基于预训练语言模型的方法在价值观方向判别上表现优异，在细粒度价值类别判别以及价值类别多标签判别上，有较大提升空间。本文工作尚存在一些局限如：整体规模较少尤其是细粒度价值-行为、语料类别分布不太均衡，个别类别中样本较少等。未来将：1）扩大知识库规模；2）针对个别类别的特点寻找适合的新闻语料来源；3）在扩大知识库规模的过程中对细粒度类别进行适当合并增减等调整，以增加覆盖度，在保证细粒度类别间区分度的同时，保证每一个细粒度类别能够有足够多的样本；4）将研究进一步拓展到语用层面。

[7]存在一些在粗粒度下为单价值行为句但是在细粒度下并非单价值行为句的情况，排除此类句子。

(c) 2022 中国中文信息学会计算语言学专业委员会

参考文献

中国国家新一代人工智能治理专业委员会, 2019. 新一代人工智能治理原则——发展负责任的人工智能[Z].

季明, 2013. 核心价值观概论[M]. 北京: 人民日报出版社.

张进辅, 1998. 我国大学生人生价值观特点的调查研究[J]. 心理发展与教育(2): 26-30.

彭诗雅, 刘畅, 邓雅月, 等, 2021. 字里行间的道德: 中文文本道德句识别研究[C]//第20届中国计算语言学大会. 537-548.

杨中芳, 2005. 中国人真是"集体主义"的吗?——试论文化、价值与个体的关系[J]. 中国社会心理学评论, 1(1): 55-93.

黄希庭, 郑涌, 2005. 当代中国青年价值观研究[M]. 人民教育出版社.

Aizenberg E, Van Den Hoven J, 2020. Designing for human rights in ai[J]. Big Data & Society, 7(2): 2053951720949566.

Allport G W, Vernon P E, Lindzey G, 1960. Study of values.[M]. Houghton Mifflin.

Asimov I, 2004. I, robot: volume 1[M]. Spectra.

Blodgett S L, Barocas S, Daumé III H, et al., 2020. Language (technology) is power: A critical survey of" bias" in nlp[A].

Botzer N, Gu S, Weninger T, 2022. Analysis of moral judgment on reddit[J]. IEEE Transactions on Computational Social Systems.

Dastin J, 2018. Amazon scraps secret ai recruiting tool that showed bias against women[M]// Ethics of Data and Analytics. Auerbach Publications: 296-299.

Devlin C M W L K T K, 2018. Bert: Pre-training of deep bidirectional transformers for language understanding[A].

Emelin D, Bras R L, Hwang J D, et al., 2020. Moral stories: Situated reasoning about norms, intents, actions, and their consequences[A].

Feng Y, Li C, Ng V, 2022. Legal judgment prediction via event extraction with constraints [C/OL]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland: Association for Computational Linguistics: 648-664. https://aclanthology.org/2022.acl-long.48.

Forbes M, Hwang J D, Shwartz V, et al., 2020. Social chemistry 101: Learning to reason about social and moral norms[A].

Hendrycks D, Burns C, Basart S, et al., 2020. Aligning ai with shared human values[A].

Jiang L, Hwang J D, Bhagavatula C, et al., 2021. Delphi: Towards machine ethics and norms [A].

Kim H, Yu Y, Jiang L, et al., 2022. Prosocialdialog: A prosocial backbone for conversational agents[A].

Kluckhohn C E, Murray H A, 1948. Personality in nature, society, and culture.[M]. Knopf.

Kosinski M, Stillwell D, Graepel T, 2013. Private traits and attributes are predictable from digital records of human behavior[J]. Proceedings of the national academy of sciences, 110 (15): 5802-5805.

Liu Y, Ott M, Goyal N, et al., 2019. Roberta: A robustly optimized bert pretraining approach [Z].

Lourie N, Le Bras R, Choi Y, 2020. Scruples: A corpus of community ethical judgments on 32, 000 real-life anecdotes[A].

Müller V C, 2020. Ethics of artificial intelligence and robotics[Z].

(c) 2022 中国中文信息学会计算语言学专业委员会

Munoz, Executive Office of the President C, Director D P C, et al., 2016. Big data: A report on algorithmic systems, opportunity, and civil rights[M]. Executive Office of the President.

Perry R B, 1926. General theory of value: Its meaning and basic principles construed in terms of interest[M]. [etc] Longmans, Green.

Prabhumoye S, Boldt B, Salakhutdinov R, et al., 2020. Case study: Deontological ethics in nlp [A].

Rokeach M, 1973. The nature of human values.[M]. Free press.

Sap M, Gabriel S, Qin L, et al., 2019. Social bias frames: Reasoning about social and power implications of language[A].

Schramowski P, Turan C, Jentzsch S, et al., 2020. The moral choice machine[J]. Frontiers in artificial intelligence, 3: 36.

Schramowski P, Turan C, Andersen N, et al., 2021. Language models have a moral dimension [A].

Schwartz S H, Bilsky W, 1987. Toward a universal psychological structure of human values.[J]. Journal of personality and social psychology, 53(3): 550.

Schwartz S H, Cieciuch J, Vecchione M, et al., 2012. Refining the theory of basic individual values.[J]. Journal of personality and social psychology, 103(4): 663.

Smuha N, 2019. Ethics guidelines for trustworthy ai[C]//AI & Ethics, Date: 2019/05/28-2019/05/28, Location: Brussels (Digityser), Belgium.

Talat Z, Blix H, Valvoda J, et al., 2021. A word on machine ethics: A response to jiang et al.(2021)[A].

Treisman A M, 1964. Selective attention in man[J]. British medical bulletin, 20(1): 12-16.

Vynck G D, 2021. The us says humans will always be in control of ai weapons. but the age of autonomous war is already here[J]. The Washington Post.

Walker C, Strassel S, Medero J, et al. Ace 2005 multilingual training corpus ldc2006t06, 2006 [J]. URL https://catalog. ldc. upenn. edu/LDC2006T06.

Weed J, 2021. Résumé-writing tips to help you get past the ai gatekeepers[J]. New York Times.

Zhou K, Smith A, Lee L, 2021. Assessing cognitive linguistic influences in the assignment of blame[C]//Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media. 61-69.

A 附录

CCL 2022

价值观类别	次分类	行为类	价值观激活时关联的行为	价值观抑制时关联的行为
文明	思想文明	宣传学习(知识/精神)	宣传塑料危害	封建迷信
	公共文明	爱护公物	保护古城墙	破坏共享单车
		爱护环境	清理垃圾	乱扔垃圾
		遵守秩序	按序排队	霸占座位
		积极参与(社会治理活动)	参加水资源保护活动	
	言语文明	用语礼貌	礼貌询问	骂脏话
	仪表文明	个人卫生	饭前洗手	饭桌抠脚
		穿着服饰	着装合适	坦胸露乳
公正	思想公正	宣传公正言论	打抱不平	发表歧视言论
	机会公正	办事公正	公正审判	找关系
		廉洁奉公	不占公家便宜	以权谋私
平等	思想平等	宣传平等言论	发表反歧视文章	传播歧视思想
	人格平等	性别平等	男女平等	重男轻女
		种族平等	尊重种族	歧视黑人
		地域平等	各省平等	地域黑
		其他（人人平等）	平等对话	富人高贵
法治	知法懂法	宣传/学习 法律条款/内容	阅读普法书籍	意识不到犯法
	守法用法	遵守法律	主动报警	违法犯罪
		配合民警执行公务	主动配合调查	拒不配合调查
爱国	思想爱国	宣传/学习 爱国言论/知识	参观革命史馆	散布分裂言论
		心系祖国	关注国家大事	崇洋媚外
	以身作则	维护祖国统一	捍卫领土	支持台独
		投入祖国建设	卫星研制	
敬业	热爱岗位	热爱本职工作	热爱教书	消极工作
	忠于职守	做好本职工作	付出辛勤劳动	偷懒耍滑
		克制欲望	淡泊名利	贪污受贿
诚信	传播诚信	宣传诚信理念	传播诚信文化	
	诚实待人	真实诚恳	直言真相	虚构事实
		拾金不昧	物归原主	隐瞒私吞
	信守诺言	兑现承诺	说话算话	欠债不还
友善	乐于助人	见义勇为	跳水救人	漠视求助
		捐款捐物	捐献衣物	
		互相帮助 （朋友、陌生人）	帮助邻居	
	宽厚待人	以和为贵	劝架	故意伤害
		关心慰问	看望战友	漠不关心
		孝顺长辈	赡养父母	虐待父母
		爱护幼小	收养孤儿	虐待儿童
		爱护动物	救助流浪狗	毒害流浪狗

Table 1: 价值-行为类别体系

第二十一届中国计算语言学大会论文集，第417页-第430页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

次分类	行为类	数量
思想文明	宣传学习（知识/精神）	239
公共文明	爱护公物	71
	爱护环境	112
	遵守秩序	316
	积极参与（社会治理活动）	46
言语文明	用语礼貌	51
仪表文明	个人卫生	8
	穿着服饰	20
思想公正	宣传公正言论	16
机会公正	办事公正	327
	廉洁奉公	37
思想平等	宣传平等言论	32
人格平等	性别平等	141
	种族平等	72
	地域平等	45
	其他（人人平等）	224
知法懂法	宣传/学习法律条款/内容	23
守法用法	遵守法律	1082
	配合民警执行公务	127
思想爱国	宣传/学习爱国言论/知识	126
	心系祖国	136
以身作则	维护祖国统一	197
	投入祖国建设	81
热爱岗位	热爱本职工作	52
忠于职守	做好本职工作	2029
	克制欲望	147
诚实待人	真实诚恳	496
	拾金不昧	116
信守诺言	兑现承诺	228
传播诚信	宣传诚信	6
乐于助人	见义勇为	522
	捐款捐物	502
	互相帮助（朋友、陌生人）	1256
宽厚待人	以和为贵	1395
	关心慰问	314
	孝顺长辈	194
	爱护幼小	204
	爱护动物	4

Table 2: 价值-行为句子标注结果

CCL 2022

第二十一届中国计算语言学大会论文集，第417页-第430页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会