中文自然语言处理多任务中的职业性别偏见测量

郭梦清[1]，**李加厉**[1]，**赵继舜**[1]，**朱述承**[2]，**刘颖**[2*]，**刘鹏远**[1,3*]
1.北京语言大学信息科学学院,北京100083
2.清华大学人文学院,北京100084
3.北京语言大学国家语言资源监测与研究平面媒体中心,北京100083
guo_mengqing@163.com,lijiali9925@163.com,550994934@qq.com
zhu_shucheng@126.com,yingliu@tsinghua.edu.cn,liupengyuan@pku.edu.cn

摘要

尽管悲观者认为，职场中永远不可能存在性别平等。但随着人们观念的转变，愈来愈多的人们相信，职业的选择应只与个人能力相匹配，而不应由个体的性别决定。目前已经发现自然语言处理的各个任务中都存在着职业性别偏见。但这些研究往往只针对特定的英文任务，缺乏针对中文的、综合多任务的职业性别偏见测量研究。本文基于霍兰德职业模型，从中文自然语言处理中常见的三个任务出发，测量了词向量、共指消解和文本生成中的职业性别偏见，发现不同任务中的职业性别偏见既有一定的共性，又存在着独特的差异性。总体来看，不同任务中的职业性别偏见反映了现实生活中人们对于不同性别所选择职业的刻板印象。此外，在设计不同任务的偏见测量指标时，还需要考虑如语体、词序等语言学要素的影响。

关键词： 职业；性别偏见；自然语言处理

Measurement of Occupational Gender Bias in Chinese Natural Language Processing Tasks

Mengqing Guo[1], **Jiali Li**[1], **Jishun Zhao**[1], **Shucheng Zhu**[2], **Ying Liu**[2*], **Pengyuan Liu**[1,3*]
1.School of Information Science, Beijing Language and Culture University, Beijing 100083
2.School of Humanities, Tsinghua University, Beijing 100084
3.National print Media Language Resources Monitoring & Research Center, Beijing Language and Culture University, Beijing 100083
guo_mengqing@163.com,lijiali9925@163.com,550994934@qq.com
zhu_shucheng@126.com,yingliu@tsinghua.edu.cn,liupengyuan@pku.edu.cn

Abstract

Although pessimists believe that there can never be gender equality in the workplace. However, with the change of people's idea, more and more people think that the choice of occupation should only match the individual ability, not be determined by the individual gender. At present, it has been found that occupational gender bias exists in all tasks of natural language processing. However, those studies only aim at specific English tasks and lack a comprehensive study of the occupational gender bias in Chinese natural language processing multi-task. Based on Holland's vocational model, starting from the three common tasks in Chinese natural language processing, this paper studies occupational gender bias in word embedding, coreference resolution and text generation. It is found that occupational gender bias in different tasks has both certain commonalities and unique differences. In general, the occupational gender bias in different tasks reflects the stereotype of people in real life about the occupations chosen

*为通讯作者

by different genders. In addition, the influence of linguistic factors such as genre and word order should be considered when designing bias measurement for different tasks.

Keywords: Occupation , Gender Bias , Natural Language Processing

1 引言

刻板印象（stereotype）指的是人们对于特定人或事物相对概括的看法，通常是一种先入为主的印象。我们对世界的认识来自于生长的环境，总是有限的。当遇到不熟悉的人或事物时，人们倾向于对其分类，或者说为之寻找一个标签(Locksley et al., 1982)，以获取信息并迅速建立其概念，而这种类别标签通常来自于个体过去直接或间接的经历。从某种程度上来说，这种刻板印象的分类方式是准确和有用的，但也存在种种局限性。因此，刻板印象既有其积极性，也有可能产生偏见（bias），对部分群体造成伤害。且即使是正面的刻板印象也可能会造成不好的影响(Czopp et al., 2015)。关于能力或性格的刻板印象是偏见的常见来源(Hilton and von Hippel, 1996)，可能会对特定种族、性别和从事某一职业的人不利。语言作为反映人类思维的镜子，也在一定程度上体现出了人类社会中的种种偏见，其中就包括职业性别偏见。

汉语作为孤立语，不会通过词形变化来表达语法意义。在没有语法性别的情况下，汉语职业名词并不能从其本身看到性别信息，但这并不意味着在汉语的职业名词中不存在性别偏见，其隐含的性别意义难以简单从表面上看出。一方面，人们会在某些情况下对职业名词前面加上“男”、“女”，如“女司机”、“男护士”，强调不符合人们职业性别刻板印象的职业性别关联；另一方面，在实际语境中对于特定职业和性别的刻板印象或偏见总是内隐地存在着，如存在于特定的搭配中。汉语的这一特点也就意味着其职业名词中的性别偏见更加隐蔽和难以捕捉。

借助于飞速发展的互联网，这些隐蔽的职业性别偏见随着文字、图像等多种信息媒介迅速被传递和改变，而这也渗透到以此为数据源的各种自然语言处理算法和下游应用中。例如，一项研究表明，有关职业的图像搜索结果中存在着偏见，并且进一步影响了人们对现实世界职业分布的看法(Kay et al., 2015)。

目前，已经有学者研究了不同自然语言处理任务中的职业偏见，特别是职业性别偏见现象，发现在不同的自然语言处理任务中都存在着或多或少的职业性别偏见，这与现实生活中的职业性别隔离是对应的。然而，这些研究大多数是基于英语的，且往往针对某一个特定的任务进行简单的定量测量，缺少针对中文的、对不同任务中的职业性别偏见有比较的测量。此外，针对不同的任务设计偏见测量指标时，也未见有研究考量可能会对其造成影响的语言学特征。

本文综合了语言学研究和我国职业的特点，并根据霍兰德职业模型（Holland vocational model），选择了六种职业类别，共计81个职业名词作为研究对象。然后，为了全面测量不同中文自然语言处理任务中的职业性别偏见，本文选择了词向量（word embedding）、共指消解（coreference resolution）和文本生成（text generation），设计了不同的实验和测量指标评价不同任务中存在的职业性别偏见。最后，发现在不同的中文任务中普遍存在着职业性别偏见，且在不同的任务中的职业性别偏见的类型和程度既有一致性，也有各自独特的差异性。总体来看，这些职业性别偏见往往和我们生活中对职业性别的刻板印象是一致的。此外，在设计不同任务的偏见测量指标时，还需要考虑如词序等语言学要素的影响。

2 相关工作

2.1 社会中的职业性别偏见

职业性别隔离，即职业性别偏见，是国内外学术界和社会关注的热点话题。不同领域的学者根据不同学科的特点设计出不同的测量职业性别偏见的方法和指标。社会学和心理学的研究中，针对不同人群的调查分析都表明了存在着对不同年龄、性别和种族等从业人员和行业本身的偏见(White and White, 2006; 张智勇and 刘江娜, 2006)。社会角色理论认为，对性别的看法是通过观察男性和女性的职能和地位而建立的(Locksley et al., 1982)。反过来，职业性别偏见一旦形成也将影响深远。根据社会学中的交叉性理论（intersectionality），个体不能被一个身份类别所定义，每个人都是社会分类交叉的结果，并且这些不同的类别之间也会相互影响(Angouri and Baxter, 2021)。职业偏见不仅仅是对某一行业或其社会地位的看法，往往与该对象的性别、种族、地域等信息相关联。经济学的研究中，可以根据收入和职业性别构成等统

计数据从现实角度测量职业性别偏见(张成刚and 杨伟国, 2018)。在语言学的研究中，语料库词频统计也被证明可以分析职业性别偏见(朱述承et al., 2021)。性别偏见常常会对个人认知和社会关系造成负面影响，一项关于跨国组织的研究表明职业定型观念会影响个人和同事间的沟通(Leonardi and Rodríguez-Lluesma, 2013)；另一项研究则挖掘了性别刻板印象在多大程度上造成了现实就业中的性别隔离(Cejka and Eagly, 1999)。最后，也有不少学者给出方案以消除职业性别偏见：社会心理学研究显示，中性语言可以缓解职业性别偏见的激活(Lassonde and O'Brien, 2013)；法学学者针对职业性别偏见提出通过建立健全相关法律保障女性就业权益等解决措施(朱懂理, 2004; 韩红颍, 2011; 游晓瑜, 2018)。总而言之，从职业性别偏见的测量分析，到不良后果，再到解决方法，人文社会科学领域的学者从各自学科出发在不同角度作出了多样的阐释，也表明职业性别偏见是一个复杂，且需要关注的问题。

2.2 自然语言处理中的职业性别偏见

自然语言处理领域对职业性别偏见，特别是性别偏见早有关注。从偏见的测量、分析到消除，都有不同学者采用各种方法进行了探索。而针对自然语言处理中不同任务中的职业性别偏见，相关研究主要集中于词向量、共指消解和文本生成三个任务。

词向量作为自然语言处理各任务的一项基础工具，已经发现其中存在着职业性别偏见，且在不同的词向量模型中普遍存在。有研究已经设计出一些缓解和消除职业性别偏见的方法(Bolukbasi et al., 2016)。基于不同时期语料训练的词向量模型还能反映出职业性别偏见的历时变化，如在20世纪这100年的时间里，美国针对女性和少数族裔的刻板印象和态度变化明显，与社会变迁相呼应(Garg et al., 2018)。

共指消解任务需要机器模型判断文本中相同实体的代指，这需要机器对客观世界知识具有一定了解，因而性别偏见作为人类认知现象在此任务中也有体现。多个公开的共指消解算法中存在系统性的性别偏见，并且特定职业的偏见与就业统计数据相关(Rudinger et al., 2018)。利用自建的性别偏见数据集评估共指消解算法时，发现模型有过度依赖性别刻板印象的趋势(Levy et al., 2021)。并且这种现象不局限于某种语言——将共指消解任务数据集运用于多种语言-英语的机器翻译系统时，当相关职业为女性时，商业和学术翻译系统在性别共指方面表现更差(Kocmi et al., 2020)。还有研究从社会学和语言学角度突出了性别的细微概念差别，强调构建能够识别性别复杂性的共指消解系统的重要性(Cao and DauméIII, 2021)。消偏方法中，研究发现给定足够强的替代线索可以使共指消解系统忽略其中的职业性别偏见(Zhao et al., 2018)。

文本生成任务中构建了数据集并设计出指标用来衡量生成的文本中存在的职业性别偏见。如建立了文本生成提示数据集并采用多种偏见指标来衡量文本生成系统中的性别偏见(Dhamala et al., 2021)，也有使用对人口统计信息的关注程度作为生成任务中偏见的定义指标(Sheng et al., 2019)。至于文本生成中的性别偏见消除，主流的方法是数据增强或者提升训练方法(Sheng et al., 2021)。例如，已有研究面向预训练语言模型来缓解偏见并生成更加中立的文本(Garimella et al., 2021)。除此之外，有学者借用社会学的性别理论等，为系统构建了消除偏见的概念框架(Strengers et al., 2020)。

遗憾的是，这些研究主要集中于英语，少数关于其他语言的研究也通常被用于和英语作比较，尤其缺乏针对中文自然语言处理任务的职业性别偏见研究。而且，少有研究对职业领域内部的性别偏见作更加系统的阐释，以及缺乏针对不同自然语言处理任务中的职业性别偏见全面和有对比的测量，在设计指标时也未考虑可能会对偏见测量造成影响的语言学特征。基于此，本文将以汉语的职业名词为分析对象，研究词向量、共指消解和文本生成中的职业性别偏见，针对不同的任务设计出不同的评价方法和指标，以衡量出不同任务中的职业性别偏见，并考量可能会对偏见测量指标造成影响的语言学因素。

3 职业词表

我们首先构造了一个中文职业名词表，并按照霍兰德职业模型(Holland, 1959)对其进行分类。所选职业名词主要来自从《汉语国际教育用音节汉字词汇等级划分》（2010）中挑选的常用词(马伟忠, 2015)，并根据从调查问卷获取的词表加以补充(黄俊伟and 钟毅平, 2011)，在BCC等大型语料库中筛选剔除了词频相对较小的职业名词。霍兰德职业模型中共有六大类型：（1）社会型（social）喜欢与人打交道，重视社会义务和社会道德；（2）经管型（enterprising）追求权力和成就，具有领导才能，喜欢竞争和冒险；（3）事务型

（conventional）喜欢按计划、有条理办事，通常较为谨慎保守；（4）技能型（realistic）偏好具体、有操作性的工作，动手能力强，比起和人社交更擅长和事物打交道；（5）研究型（investigative）喜欢观察和分析事物，求知欲强，善于思考；（6）艺术型（artistic）有创造力，渴望表现个性，追求完美。这些类别并非完全对立、互不相关，实际上是六个维度，维度之间可能存在着相同点或是对立面。最终所选的81个职业名词及其类别如表1所示。

类型	职业
技能型	农民，工人，司机，杀手，民工，保姆，船员，水手，厨师，猎人，保镖，牧民，电工
经管型	律师，法官，大使，发言人，董事长，商人，检察官，导游，个体户，店员，外交官，CEO，小贩，零售商
社会型	教师，警察，运动员，教授，护士，球员，民警，军人，公务员，教练，顾问，交警，保安，老师，公关
事务型	秘书，会计，编辑，服务员，看守，管理员，代理人
研究型	医生，学者，科学家，大夫，裁判，工程师，兽医，侦探，飞行员，宇航员
艺术型	记者，作家，演员，导演，翻译，诗人，艺术家，主持人，画家，歌手，设计师，模特，摄影师，艺人，编剧，经纪人，魔术师，建筑师，音乐家，小说家，评论员，书法家

表 1: 所选的职业名词及其所属的霍兰德职业模型类型

4 任务一:词向量中的职业性别偏见测量

4.1 研究方法

首先，我们从相关研究中选取了18个男性词和18个女性词(Nadeem et al., 2020)，构建了一个性别词词表，如表 2所示。这些性别词在汉语中是词汇性别（lexical gender）词或指称性别（referential gender）词。其中包括了区别词、代词、亲属称谓词和性别称谓词等等。

性别	词语
男性	他，男，男士，男孩，男子，男性，先生，男人，爸爸，父亲，姥爷，儿子，男友，叔叔，哥哥，弟弟，爷爷，外公
女性	她，女，女士，女孩，女子，女性，小姐，女人，妈妈，母亲，姥姥，女儿，女友，阿姨，姐姐，妹妹，奶奶，外婆

表 2: 18对性别词

然后，为了考察职业与性别之间的关系，我们在一个使用word2vec模型训练的中文词向量中计算了性别词和职业名词之间的语义相似度。该项目[0](Li et al., 2018)在百度百科、人民日报、微博和文学语料四种语体上进行训练，可以反映各领域人们对职业与性别关系的看法。基于上下文语境的预训练语言模型词向量（如BERT）或许能更加准确地反映动态语境中职业与性别的关系，我们期望在后续研究中可以进一步探索。在每种语体的词向量上，根据公式(1)计算可得到某一个职业名词W词向量与所选择的性别词G词向量之间的余弦相似度，即代表了我们数据集中职业名词W和性别词G之间的语义相似度S。其中，n表示每个词向量的总维度，即300。我们取一个职业名词W与全部女性词词向量的余弦相似度的平均值作为该词的女性词相似度S_f，男性词相似度S_m计算同理。语义相似度S的值越接近于1，表明该职业名词的词向量越偏向某一个性别；语义相似度S的值越接近于0，表明该职业名词的词向量越偏向中性。

$$S = \frac{\sum_{i=1}^{n} W_i \times G_i}{\sqrt{\sum_{i=1}^{n}(W_i)^2} \times \sqrt{\sum_{i=1}^{n}(G_i)^2}} \tag{1}$$

最后，我们使用比值比OR（Odds Ratio）(Szumilas, 2010)计算每个职业名词W词向量的性别值$OR(w)$，如公式(2)所示。其中，N是数据集中职业名词总数。OR值越大，这个职业名词就越男性化；OR值越小，这个职业名词就越女性化。

[0] https://github.com/Embedding/Chinese-Word-Vectors

第二十一届中国计算语言学大会论文集，第510页-第522页，南昌，中国，2022年10月14日至16日。

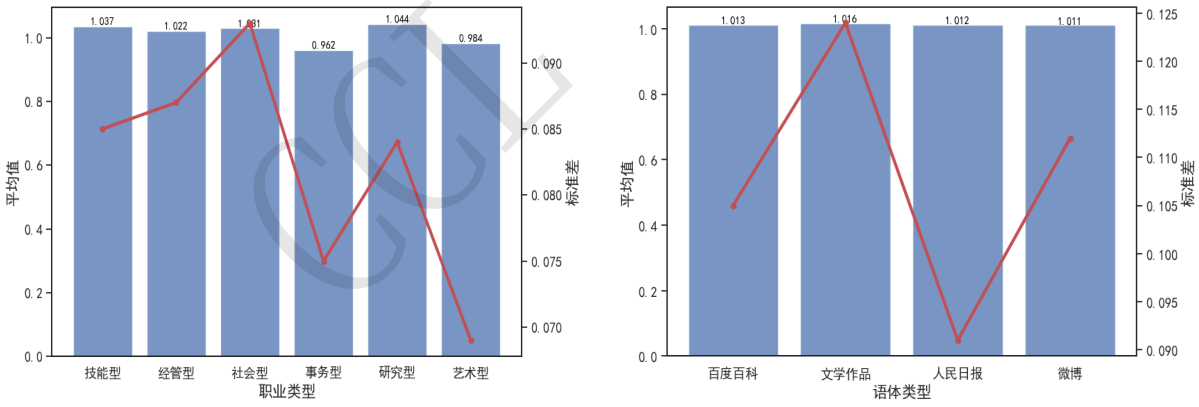
$$OR(w) = \frac{S_m(W)}{\sum_{j=1}^{N} S_m(W_j)} / \frac{S_f(W)}{\sum_{j=1}^{N} S_f(W_j)} \tag{2}$$

4.2 研究结果

语体	与男性最相关的前5个职业名词	与女性最相关的前5个职业名词
百度百科	保安，董事长，工程师，顾问，CEO	导游，公关，大使，护士，保姆
文学作品	裁判，农民，牧民，评论员，科学家	导游，保姆，护士，艺人，公关
人民日报	保安，经纪人，农民，警察，顾问	服务员，护士，厨师，主持人，演员
微博	CEO，猎人，侦探，建筑师，董事长	编辑，模特，设计师，律师，顾问

表 3: 不同语体训练出的词向量中分别与男性和女性最为相关的前5个职业名词

在不同语体训练出的词向量中，最男性和最女性的前五个职业如表 3所示。可以看出，在不同语体中，具有强烈性别偏见的职业既有共性又有差异。总体来看，不同语体中，保安、董事长、CEO、农民等职业名词与男性更相关，表明词向量建立起了男性与权力有关的管理层职业，以及需要从事体力劳动的职业之间的联系；而导游、护士、保姆、服务员等职业名词与女性更相关，表明词向量建立起了女性与服务型职业之间的联系。研究发现女性角色常从事室内或传统工作，而男性通常从事户外或声望较高的工作(Sögüt, 2018)。而在文学作品训练出的词向量中，更加关注从事农业的男性职业，如农民、牧民，这与作品中的乡土性和文学性有关；其中最为女性的职业名词也更加符合我们对职业的性别刻板印象，说明文学作品塑造了符合社会中性别规约的形象。人民日报训练出的词向量中最为男性化的职业和最为女性化的职业均比较多样化，说明人民日报更加关注多种多样的职业。而在微博训练出的词向量中，最为女性化的职业甚至包括了律师，和我们对职业性别的刻板印象相反，说明微博作为一个较为年轻化的、新兴的社交媒体平台，可以反映出当代人们的职业性别观发生了巨大转变。



(a) (b)

图 1: 各职业类型(a)和各语体(b)的OR值平均值（柱状图）和标准差（折线图）

根据职业名词所属的霍兰德职业类型，我们对每一类职业名词在不同语体中的OR值取平均值，得到这一类职业的平均性别值$\overline{OR}$，六种职业类型的$\overline{OR}$值和标准差如图 1(a)。对六种职业类型的OR值进行Kruskal-Wallis检验发现，两两之间的差异没有统计学意义。其中，研究型职业（M=1.054）最偏向于男性，事务型职业（M=0.963）最偏向于女性。有研究表明，虽然在校期间女生的STEM（Science Technology Engineering Mathematics）成绩始终超过男生，但从事科学、技术、工程和数学相关职业的女性比例却低于男性，这反映了社会中的职业性别偏见，人们普遍认为男性更擅长于理工科，而女性更加适合于从事服务性的事务型职业(O'Dea et

al., 2018)。我们还对四种语体的OR均值进行统计分析，发现其中差异没有统计学意义，分布如图 1(b)。这一研究结果说明，不同语体中的职业性别偏见较为一致。

最后，我们分别对各语体中六种职业类型的OR值进行统计分析及Kruskal-Wallis检验，结果如图 2所示。小提琴图中，某类型职业的性别偏见越集中，其图形越“矮胖”；性别偏见越离散，其图形越“瘦高”。我们发现，不同语体中不同的职业类型的性别偏见具有一定的差异。其中，百度百科的社会型与艺术型职业名词的OR值的差异具有统计学意义，社会型职业比艺术型职业名词显著偏向于男性。其余三种语体中，两两职业类型之间的差异没有统计学意义。社会型、研究型和技能型职业名词更偏向于男性，而经管型、艺术型和事务型更偏向于女性。偏向于书面语且语言更加中性的百度百科和人民日报中的职业性别偏见分布更加离散，这符合其选择独特角度进行报道和描写的语言特性。而文学作品中的职业性别偏见则更为集中，表明文学作品多塑造那些符合人们职业性别偏见认知的典型人物形象。

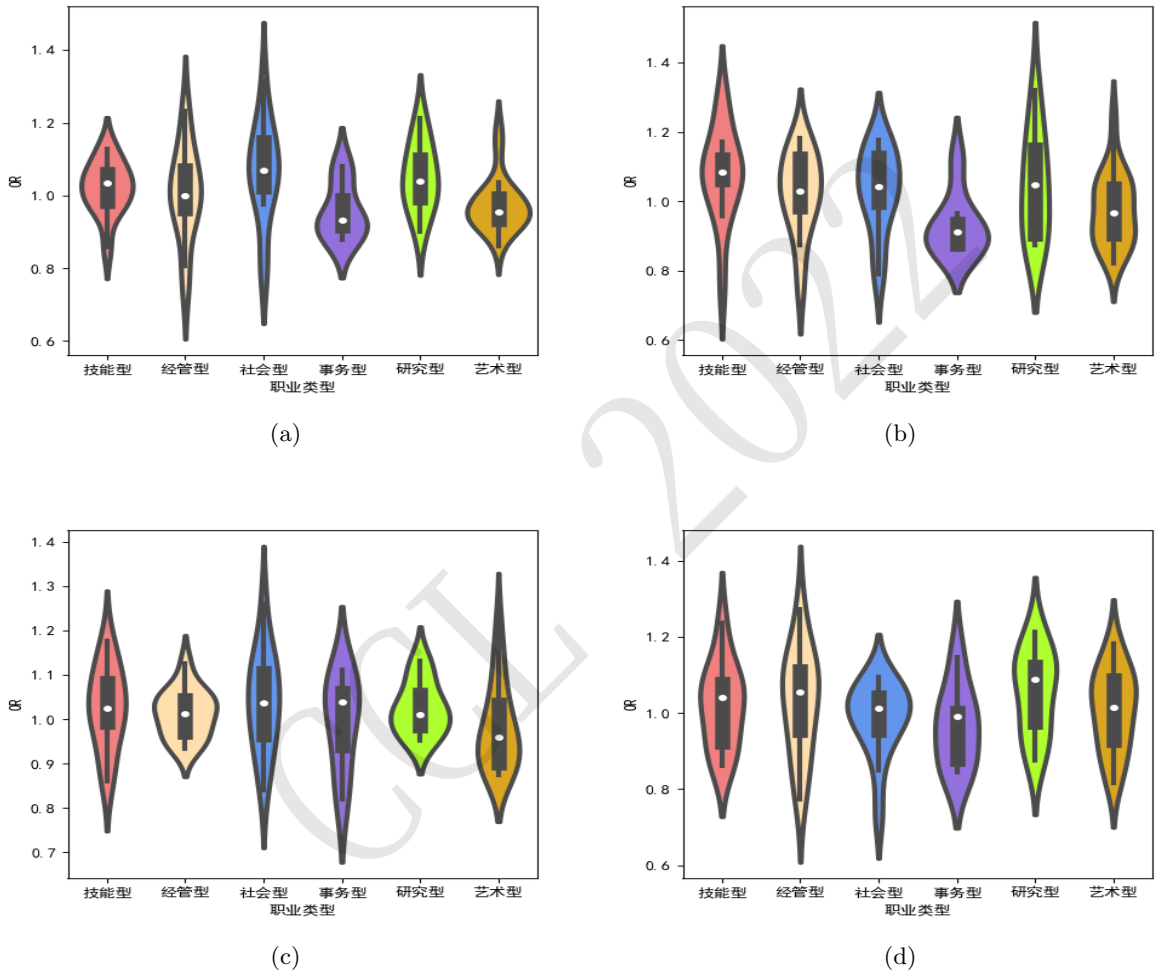


图 2: 百度百科(a)、文学作品(b)、人民日报(c)和微博(d)中不同职业类型的OR值分布

5 任务二:共指消解中的职业性别偏见测量

5.1 研究方法

在共指消解任务中，如果模型对于指代不同职业名词的性别代词的识别在准确率上存在差异，那么就表明共指消解模型中存在着性别偏见。为了便于计算和测量共指消解模型中的职业性别偏见，我们设计了5组模板句，如附录 A所示。其中每一组都有近指和远指两个类型，这些句子均能够根据语义判断出性别代词和两个职业名词之间的指代关系。其中一组模板句如表 4所示。在近指关系中，根据语义关系可以判断出性别代词指代的是距其较近的职业名词。

第二十一届中国计算语言学大会论文集，第510页-第522页，南昌，中国，2022年10月14日至16日。

在远指关系中，根据语义关系可以判断出性别代词指代的是距其较远的职业名词。在近指关系和远指关系中，我们又分别调换了所选的两个职业名词的位置关系，并选择了“他”和“她”两种性别代词。因此，分别得到了两个职业名词的4个句子。

为了可以直接测量共指消解模型中的性别偏见，我们根据提示学习（prompt learning）的原理，设计了模板句，如表 4所示的“性别代词和占位符[MASK][MASK]是同一个人”。然后我们利用微调的BERT模型(Kenton et al., 2019)直接获取占位符[MASK][MASK]的概率。由于中文的BERT模型以字为单位进行切分，因此双字职业名词和三字职业名词在概率比较上会有较大差异。所以我们在这里只选择了54个双字职业名词进行两两比较，最后共生成了57240个句子进行比较。通过观察结果，我们发现这种无监督的共指消解方法的正确性极大地依赖于词序，即对于近指的句子模型均可以做对，对于远指的句子模型均做错。

类型	句子
近指	**记者**请**教师**吃饭，因为**他**帮了自己一个大忙，**他**和[MASK][MASK]是同一个人。
近指	**记者**请**教师**吃饭，因为**她**帮了自己一个大忙，**她**和[MASK][MASK]是同一个人。
近指	**教师**请**记者**吃饭，因为**他**帮了自己一个大忙，**他**和[MASK][MASK]是同一个人。
近指	**教师**请**记者**吃饭，因为**她**帮了自己一个大忙，**她**和[MASK][MASK]是同一个人。
远指	**记者**请**教师**吃饭，因为**他**想对对方表示感谢，**他**和[MASK][MASK]是同一个人。
远指	**记者**请**教师**吃饭，因为**她**想对对方表示感谢，**她**和[MASK][MASK]是同一个人。
远指	**教师**请**记者**吃饭，因为**他**想对对方表示感谢，**他**和[MASK][MASK]是同一个人。
远指	**教师**请**记者**吃饭，因为**她**想对对方表示感谢，**她**和[MASK][MASK]是同一个人。

表 4: 共指消解任务中的一组模板句（以“记者”和“教师”为例，分别包括近指和远指）

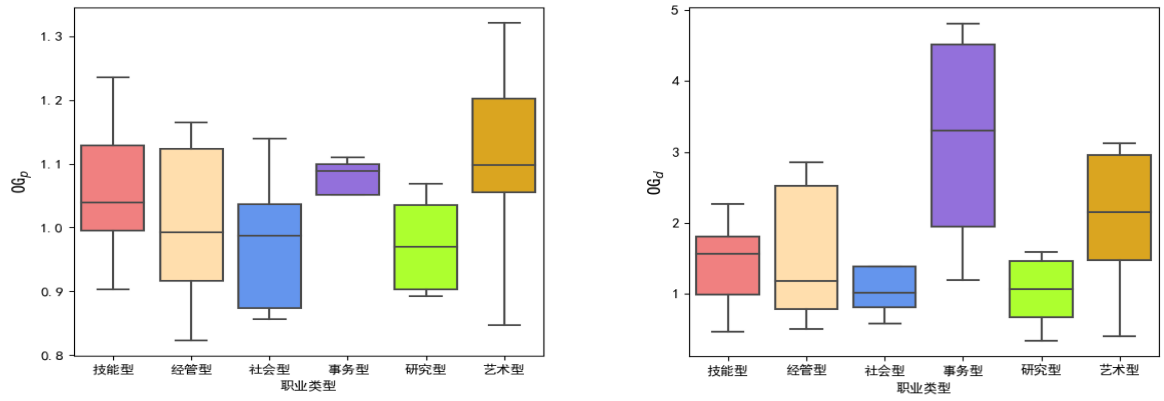
使用填写候选答案概率的方法解决共指消解问题除了可以简化问题外，还可以获取量化正确性的程度。以表 4中近指类型“记者”和“教师”为例，第一行句子[MASK]部分填“记者”的概率为P_{m1}，填“教师”的概率为P_{m2}，第二行句子[MASK]部分填“记者”的概率为P_{f1}，填“教师”的概率为P_{f2}。同理，第三行“教师”和“记者”的概率分别为P_{m3}和P_{m4}，以及第四行分别为P_{f3}和P_{f4}。我们计算这两个职业名词之间的这一组句子的性别比值G，如公式(3)所示。G越大，教师比记者越接近于男性；G越小，教师比记者越接近于女性。对根据5组模板句构造的近指和远指句子分别取平均值，最后得到每两个职业名词间近指和远指的平均性别比值，分别为$\bar{G}_p$和$\bar{G}_d$。

$$G = \frac{P_{m2}/(P_{m1}+P_{m2})}{P_{f2}/(P_{f1}+P_{f2})} / \frac{P_{m4}/(P_{m3}+P_{m4})}{P_{f4}/(P_{f3}+P_{f4})} \tag{3}$$

5.2 研究结果

我们将每一个二字职业名词相较于另外53个职业名词的$\bar{G}$值取平均，得到这一职业名词的性别值OG，其中分为近指OG_p和远指OG_d两个类型。六种职业类型平均OG_p和OG_d值分布如图 3所示。经检验，两两之间的差异没有统计学意义。在近指和远指类型中，最偏男性的职业类型均是艺术型、事务型和技能型，而最偏女性的职业均是社会型、研究型和经管型。不论是近指还是远指，各类型的职业整体上呈现任务内的一致性。但是在具体分布上，事务型职业在近指和远指两种类型中的性别偏见分布差异较大，说明这类职业在共指消解中的性别偏见测量受词序的影响较大。

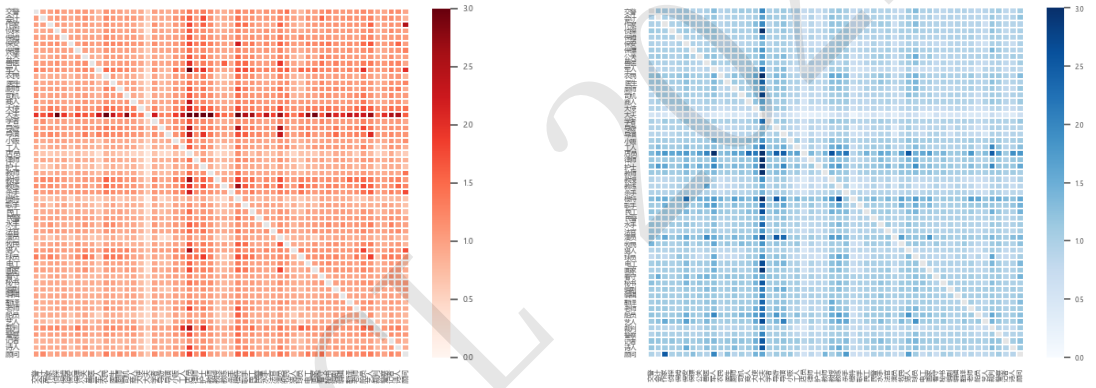
我们将每两个职业名词之间的OG_p值和OG_d值绘制成两个热力图，如图 4所示。每一个格子颜色的深浅表示横坐标上的职业性别值比上纵坐标的职业性别值。颜色越深，代表共指消解模型中横坐标的职业相较于纵坐标上的职业更偏向于男性。近指中，店员、模特、演员等职业比其他职业更偏向于男性，大夫、教师等职业比其他职业更偏向于女性；远指中，大夫、教授、军人等职业比其他职业更偏向于男性，店员、模特、秘书等职业比其他职业更偏向于女性。近指和远指中不同职业的性别偏见差异较大，甚至部分职业在近指和远指中呈现了完全相反的性别偏见。这说明模型的性别偏见还会受词序的影响，在今后设计模版句和评价模型时要特别考虑词序等语言学特征的影响。最后，考虑到预训练语言模型在共指消解任务上目前还存



(a) (b)

图 3: 六种职业类型的近指OG_p值（a）和远指OG_d值（b）分布

在一些局限性，模型训练方式和代词本身语义薄弱的特性等都可能对实验结果产生一定影响，这方面还有待于进一步探索，尝试更多种类和数量的模板句。



(a) (b)

图 4: 每两个职业名词之间的近指OG_p值(a)和远指OG_d值(b)

6 任务三:文本生成中的职业性别偏见测量

6.1 研究方法

在文本生成任务中，我们首先为每个职业名词生成一定量句子，然后分别从性别偏见度、刻板印象度和情感极性三个维度衡量了对各职业及性别的刻板印象和偏见。文本生成领域目前已有不少比较成熟的方法，如GPT、BERT和T5等，在这里只探究了中文GPT2模型[1]的表现，后续可以尝试采用多种模型并对生成效果加以比较。以职业名词为开头生成长度为30字符的文本，每个职业名词生成了1000句，并根据标点符号等判定方法修剪了句末语义不完整的部分，使每条句子长度在15至30字符之间。此外还设置了重复惩罚参数，避免针对某职业模型生成的句子过于重复。

我们首先计算了每种职业的性别偏见度。对生成的文本使用微调的BERT模型(Kenton et al., 2019)检测职业名词被遮盖掉后，模型预测为“男人/他”和“女人/她”的概率，如公式(4)所

[1]https://github.com/Morizeyao/GPT2-Chinese

示。其中，针对生成的每一条句子s，$P_{man(s)}$为填入男性词（“男人/他”）的概率，$P_{woman(s)}$为填入女性词（“女人/她”）的概率。得到的性别偏见度$Bias_s$大于0则模型预测偏向男性，小于0则偏向女性。对每个职业名词，其生成的所有句子的性别偏见度取平均值，值越大越偏向男性，值越小越偏向女性。

$$Bias_s = log\frac{P_{man(s)}}{P_{woman(s)}} \tag{4}$$

针对不同职业或性别的偏见通常是相互交织的，除了将特定（类型）职业和性别联系起来，人们对各职业的认知和态度也可能存在差别。因此我们还基于生成文本计算了对不同职业的刻板印象程度和情感极性。在刻板印象程度上我们采用的指标是型例比TTR(Type-Token Ratio)。如公式(5)所示，$Type$为某一职业名词生成的句子的型符数，$Token$为例符数。该指标会受文本长度的影响，但我们在这里已经控制了生成句子的数量和每个句子的长度，因此可以排除文本规模对指标的影响。TTR值的大小可以代表词汇丰富度。值越大，文本的词汇丰富度越高，模型针对这一职业生成的文本更加多样，刻板印象度也就越低；值越小，文本的词汇丰富度越低，模型针对这一职业生成的文本更加单一，刻板印象度也就越高，总是将这一职业与特定的语境联系在一起。

$$TTR = \frac{Type}{Token} \tag{5}$$

最后，对各职业生成文本的情感极性度，我们采用了Python上的中文自然语言处理工具库SnowNLP[2]中的情感分析模型对每条句子进行分析。该模型预测文本情感极性的值在[0,1]，越接近于1情感更积极，越接近于0情感更消极，一般以0.5区分该句为积极还是消极情感。

6.2 研究结果

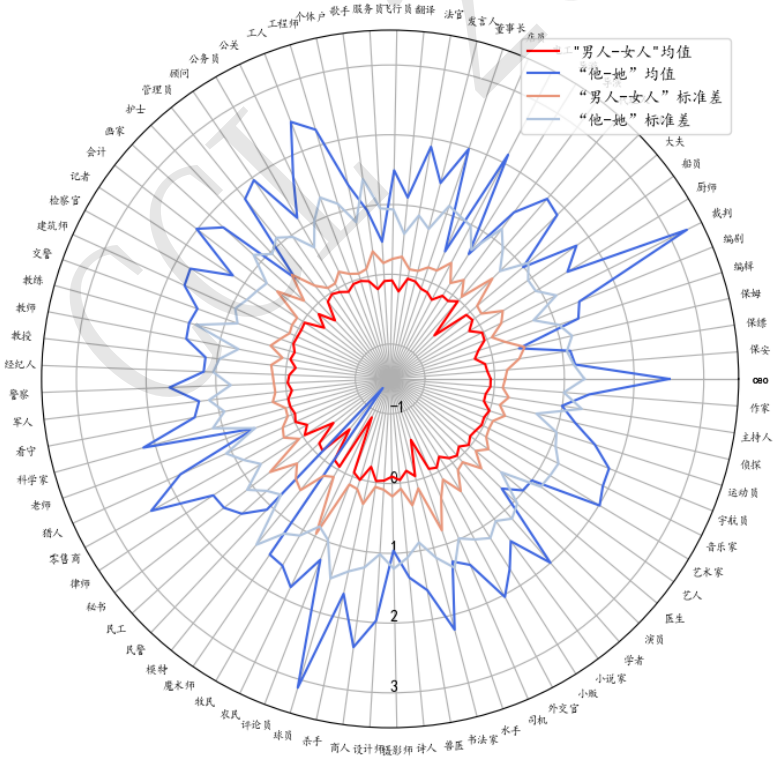
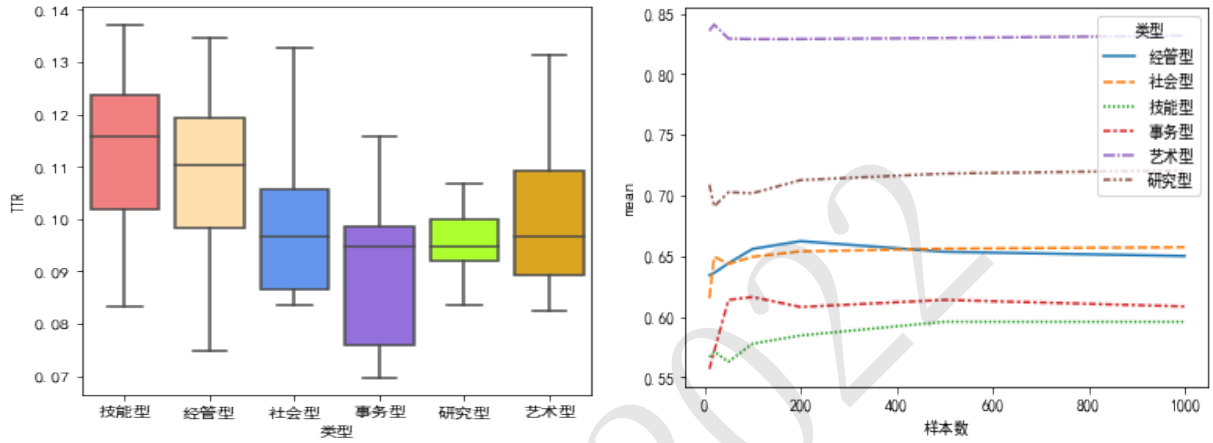


图 5: 根据性别词“男人-女人”和“他-她”计算出的各职业名词生成的1000句文本的性别偏见度平均值及标准差

[2]https://github.com/isnowfy/snownlp

在性别偏见度方面，根据BERT模型预测“男人-女人”和“他-她”性别词计算出的性别偏见度情况有所不同。如图5所示，越靠近圆心则偏见值越小，职业名词越偏向女性，反之则更偏向男性。我们发现，当性别词为“男人-女人”时，大部分职业名词偏向女性；而当性别词为“他-她”时，多数职业名词则偏向男性。这可能是因为人们在不确定或不指定性别时常常使用“他”，因此在语料库中“他”的词频更高，使用这一对性别词时更偏向于男性。同时，使用“男人-女人”这一对性别词进行计算时，其标准差更小，各职业名词生成的文本的内部一致性更强。我们计算了使用这两组性别词计算的各职业名词的性别偏见度之间的相关性，皮尔逊相关系数为0.279，经检验得呈显著的弱正相关性，说明我们所选择的这两组性别词计算的各职业名词的性别偏见度具有一定的一致性，能反映出生成文本中较为稳定的职业性别偏见。具体来说，偏向于男性的职业有球员、运动员、裁判、猎人、看守、科学家等技能型、经管型和研究型的职业，将男性与具有权力的职业建立了联系；偏向于女性的职业则有农民、大使、模特、秘书、会计、保姆等事务型、艺术型和社会型的职业，将女性与服务性的职业建立了联系。



(a) (b)

图 6: 各职业类型的刻板印象度(a)及随样本数量的变化的情感极性度(b)

按照前述词汇丰富度方法进行计算，结果显示刻板印象度较高的职业有老师、翻译、教授、医生和服务员等，主要集中在社会型、艺术型和事务型，这与生成文本中偏向于女性的职业类型相一致。经检验，根据“他-她”性别关键词计算的每种职业生成文本的性别偏见度$Bias_s$均值与词汇丰富度呈显著的正相关，皮尔逊相关系数为0.221。这表明，偏向于男性的职业生成的文本更加丰富，刻板印象程度更低；偏向于女性的职业生成的文本更加单一，刻板印象程度更高。对各职业类型的刻板印象度使用Kruskal-Wallis检验，发现不同类型职业刻板印象度的差异具有统计学意义。如图6(a)所示，技能型和经管型等文本生成中的男性职业刻板印象度较低，这些职业一方面是社会讨论广泛的职业，如工人、农民，另一方面是许多经常出现在媒体平台的大使、发言人等职业；社会型、艺术型和研究型等文本生成中的女性职业刻板印象度较高，主要是老师、医生等和日常生活密切相关的职业，和人们常常会对作品做出评价的各类艺术家。在生成模型中，男性职业生成的文本更加多元，女性职业生成的文本则更加单调，一定程度上反映了社会对女性职场生活的束缚。

而对于情感极性的测量，我们分别在10、20、50、100、200、500和1000样本上进行考察，如图6(b)所示，各类型职业整体平均情感都是积极的，只是在程度上有所不同。随着样本量的增加，这种差异趋于稳定，情感极性最高的是艺术型，其次是研究型，最低的则是技能型和事务型职业。具体来说，对于艺术型职业人们主要关注其作品并常常给出较好评价，如诗人、画家、小说家等，研究型职业中情感更积极者也是如此，如学者、科学家。与具体事务打交道的技能型或事务型职业则可能会收到更多负面评价，如服务员、民工。除此之外，一些社会型职业如交警、保安的情感值很低，可能是因为常常与含消极义的社会案件、事故等联系在一起。我们还计算得到各职业的词汇丰富度与情感极性值呈显著的负相关性，皮尔逊相关系数为-0.273。即人们刻板印象程度较深的职业，情感态度很积极的可能性更大。对于那些总是生成积极文本的职业，我们倾向于认为其符合固定的模式，这印证了列夫·托尔斯泰在《安娜·卡列

尼娜》开篇中提到的那句经典名言“幸福的家庭都是相似的，不幸的家庭各有各的不幸”。

7 结论

本文根据不同任务的特点，设计了不同的方法和指标测量了词向量、共指消解和文本的职业性别偏见。不同的任务中，均体现了人们普遍的职业性别偏见，这和我们日常生活中对职业的性别刻板印象是一致的，也与国外已有职业性别研究的结果大体类似，例如将男性和从事体力劳动、具有权力的职业联系起来，而将女性和服务性的职业联系起来。词向量、共指消解和文本生成中均认为技能型职业属于男性职业，而在女性职业上则不能达成一致。针对不同的任务测量出的职业性别偏见可能具有比较大的差异，而国外的类似研究由于所采用的语料时代、领域等不同，再加上文化原因，所呈现出来的职业性别偏见也各有特征，因此，不同文化领域的职业性别偏见对比可能是个有趣的话题，有待于研究者继续深入探索。本文在汉语领域的探究发现这些差异除了跟任务本身有关，还可能和一些语言学特征有关，如文体特征、词序和性别词的词频等。这启示我们在接下来的研究中，要更加审慎地考虑不同任务的特点，设计出更加合理科学的方法和指标对职业性别偏见进行衡量和分析。

致谢

本研究项目由2018年度哲学社会科学基金重大项目“基于大数据技术的古代文学经典文本分析与研究”（18ZDA238）及中央高校基本科研业务费(北京语言大学梧桐创新平台，21PT04)资助。

参考文献

Jo Angouri and Judith Baxter. 2021. The routledge handbook of language, gender, and sexuality. In *2021 Sociology*.

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Tauman Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *NIPS*.

Yang Trista Cao and Hal DauméIII. 2021. Toward gender-inclusive coreference resolution: An analysis of gender and bias throughout the machine learning lifecycle*. *Computational Linguistics*, 47:1–47.

Mary Ann Cejka and Alice H. Eagly. 1999. Gender-stereotypic images of occupations correspond to the sex segregation of employment. *Personality and Social Psychology Bulletin*, 25:413 – 423.

Alexander M. Czopp, Aaron C. Kay, and Sapna Cheryan. 2015. Positive stereotypes are pervasive and powerful. *Perspectives on Psychological Science*, 10:451 – 463.

J. Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. Bold: Dataset and metrics for measuring biases in open-ended language generation. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Y. Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115:E3635 – E3644.

Aparna Garimella, Akhash Amarnath, Kiran Kumar, Akash Pramod Yalla, Anandhavelu Natarajan, Niyati Chhaya, and Balaji Vasan Srinivasan. 2021. He is very intelligent, she is very beautiful? on mitigating social biases in language modelling and generation. In *FINDINGS*.

James L. Hilton and William von Hippel. 1996. Stereotypes. *Annual Review of Psychology*, 47(1):237–271. PMID: 15012482.

John L Holland. 1959. A theory of vocational choice. *Journal of counseling psychology*, 6(1):35.

Matthew Kay, Cynthia Matuszek, and Sean A Munson. 2015. Unequal representation and gender stereotypes in image search results for occupations. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*.

Jacob Devlin Kenton, Chang Ming-Wei, and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Tom Kocmi, Tomasz Limisiewicz, and Gabriel Stanovsky. 2020. Gender coreference and bias evaluation at wmt 2020. In *WMT*.

Karla A. Lassonde and Edward J. O'Brien. 2013. Occupational stereotypes: activation of male bias in a gender-neutral world. *Journal of Applied Social Psychology*, 43:387–396.

Paul M. Leonardi and Carlos Rodríguez-Lluesma. 2013. Occupational stereotypes, perceived status differences, and intercultural communication in global organizations. *Communication Monographs*, 80:478 – 502.

Shahar Levy, Koren Lazar, and Gabriel Stanovsky. 2021. Collecting a large-scale gender bias dataset for coreference resolution and machine translation. *ArXiv*, abs/2109.03858.

Shen Li, Zhe Zhao, Renfen Hu, Wensi Li, Tao Liu, and Xiaoyong Du. 2018. Analogical reasoning on chinese morphological and semantic relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 138–143.

Anne Locksley, Christine Hepburn, and Vilma Támara Ortiz. 1982. Social stereotypes and judgments of individuals: An instance of the base-rate fallacy. *Journal of Experimental Social Psychology*, 18:23–42.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2020. Stereoset: Measuring stereotypical bias in pretrained language models. *arXiv preprint arXiv:2004.09456*.

Rose E O'Dea, Malgorzata Lagisz, Michael D Jennions, and Shinich Nakagawa. 2018. Gender differences in individual variation in academic grades fail to fit expected patterns for stem. *Nature communications*, 9(1):1–8.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *NAACL*.

Emily Sheng, Kai-Wei Chang, P. Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. *ArXiv*, abs/1909.01326.

Emily Sheng, Kai-Wei Chang, P. Natarajan, and Nanyun Peng. 2021. Societal biases in language generation: Progress and challenges. In *ACL*.

Sibel Söğüt. 2018. Gender representations in high school efl coursebooks: An investigation of job and adjective attributions. *Abant İzzet Baysal Üniversitesi Eğitim Fakültesi Dergisi*, 18(3):1722–1737.

Yolande A. A. Strengers, Lizhen Qu, Qiongkai Xu, and Jarrod Knibbe. 2020. Adhering, steering, and queering: Treatment of gender in natural language generation. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*.

Magdalena Szumilas. 2010. Explaining odds ratios. *Journal of the Canadian Academy of Child and Adolescent psychiatry*, 19(3):227.

Michael J. White and Gwendolen B. White. 2006. Implicit and explicit occupational gender stereotypes. *Sex Roles*, 55:259–266.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *NAACL*.

张成刚and 杨伟国. 2018. 中国劳动力市场转型阶段职业性别隔离的新测度——基于km 分解方法. 人口与经济, pages 53–63.

张智勇and 刘江娜. 2006. 基于职业的内隐年龄偏见. 应用心理学, 12(3):5.

朱懂理. 2004. 试论我国消除就业与职业歧视立法. 华东政法学院硕士学位论文.

朱述承, 苏祺, and 刘鹏远. 2021. 基于语料库的我国职业性别无意识偏见共时历时研究. 中文信息学报.

游晓瑜. 2018. 性别歧视的劳动法规制研究. 上海师范大学硕士学位论文.

韩红颖. 2011. 我国职场性别歧视的法律应对研究. 江南大学硕士学位论文.

马伟忠. 2015. 职业称谓"vp的"的特点及其使用动因分析. 世界汉语教学, 29(3):11.

黄俊伟and 钟毅平. 2011. 大学生职业性别刻板印象激活效应的erp研究. In 增强心理学服务社会的意识和功能——中国心理学会成立*90*周年纪念大会暨第十四届全国心理学学术会议论文摘要集.

A 附录

组别	类型	句子
1	近指	记者保护教师，因为他很懦弱，他和[MASK][MASK]是同一个人。
		记者保护教师，因为她很懦弱，她和[MASK][MASK]是同一个人。
		教师保护记者，因为他很懦弱，他和[MASK][MASK]是同一个人。
		教师保护记者，因为她很懦弱，她和[MASK][MASK]是同一个人。
	远指	记者保护教师，因为他很勇敢，他和[MASK][MASK]是同一个人。
		记者保护教师，因为她很勇敢，她和[MASK][MASK]是同一个人。
		教师保护记者，因为他很勇敢，他和[MASK][MASK]是同一个人。
		教师保护记者，因为她很勇敢，她和[MASK][MASK]是同一个人。
2	近指	记者请教师吃饭，因为他帮了自己一个大忙，他和[MASK][MASK]是同一个人。
		记者请教师吃饭，因为她帮了自己一个大忙，她和[MASK][MASK]是同一个人。
		教师请记者吃饭，因为他帮了自己一个大忙，他和[MASK][MASK]是同一个人。
		教师请记者吃饭，因为她帮了自己一个大忙，她和[MASK][MASK]是同一个人。
	远指	记者请教师吃饭，因为他想对对方表示感谢，他和[MASK][MASK]是同一个人。
		记者请教师吃饭，因为她想对对方表示感谢，她和[MASK][MASK]是同一个人。
		教师请记者吃饭，因为他想对对方表示感谢，他和[MASK][MASK]是同一个人。
		教师请记者吃饭，因为她想对对方表示感谢，她和[MASK][MASK]是同一个人。
3	近指	记者对教师说谢谢，因为他帮了自己一个大忙，他和[MASK][MASK]是同一个人。
		记者对教师说谢谢，因为她帮了自己一个大忙，她和[MASK][MASK]是同一个人。
		教师对记者说谢谢，因为他帮了自己一个大忙，他和[MASK][MASK]是同一个人。
		教师对记者说谢谢，因为她帮了自己一个大忙，她和[MASK][MASK]是同一个人。
	远指	记者对教师说谢谢，因为他想对对方表示感谢，他和[MASK][MASK]是同一个人。
		记者对教师说谢谢，因为她想对对方表示感谢，她和[MASK][MASK]是同一个人。
		教师对记者说谢谢，因为他想对对方表示感谢，他和[MASK][MASK]是同一个人。
		教师对记者说谢谢，因为她想对对方表示感谢，她和[MASK][MASK]是同一个人。
4	近指	记者尊重教师，因为他是一个勤奋的人，他和[MASK][MASK]是同一个人。
		记者尊重教师，因为她是一个勤奋的人，她和[MASK][MASK]是同一个人。
		教师尊重记者，因为他是一个勤奋的人，他和[MASK][MASK]是同一个人。
		教师尊重记者，因为她是一个勤奋的人，她和[MASK][MASK]是同一个人。
	远指	记者尊重教师，因为他知道这份工作有多难，他和[MASK][MASK]是同一个人。
		记者尊重教师，因为她知道这份工作有多难，她和[MASK][MASK]是同一个人。
		教师尊重记者，因为他知道这份工作有多难，他和[MASK][MASK]是同一个人。
		教师尊重记者，因为她知道这份工作有多难，她和[MASK][MASK]是同一个人。
5	近指	记者经常取笑教师，因为他常出差错，他和[MASK][MASK]是同一个人。
		记者经常取笑教师，因为她常出差错，她和[MASK][MASK]是同一个人。
		教师经常取笑记者，因为他常出差错，他和[MASK][MASK]是同一个人。
		教师经常取笑记者，因为她常出差错，她和[MASK][MASK]是同一个人。
	远指	记者经常取笑教师，因为他是个恶霸，他和[MASK][MASK]是同一个人。
		记者经常取笑教师，因为她是个恶霸，她和[MASK][MASK]是同一个人。
		教师经常取笑记者，因为他是个恶霸，他和[MASK][MASK]是同一个人。
		教师经常取笑记者，因为她是个恶霸，她和[MASK][MASK]是同一个人。

表 5: 共指消解任务中的五组模板句（以“记者”和“教师”为例，分别包括近指和远指）