

基于关键词注意力的细粒度面试评价方法

陈楚杰¹ 吕建明^{1,2} 沈华伟³

¹(华南理工大学计算机科学与工程学院 广州 510006)
²(大数据与智能机器人教育部重点实验室(华南理工大学) 广州 510006)
³(中国科学院计算技术研究所 北京 100190)
(cscjchen@mail.scut.edu.cn)

Fine-Grained Interview Evaluation Method Based on Keyword Attention

Chen Chujie¹, Lü Jianming^{1,2}, and Shen Huawei³

¹(School of Computer Science and Engineering, South China University of Technology, Guangzhou 510006)
²(Key Laboratory of Big Data and Intelligent Robot (South China University of Technology), Ministry of Education, Guangzhou 510006)
³(Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190)

Abstract Massive online interview video data provides an important data basis for intelligent interview evaluation. With the spread of the current global epidemic, the demand for online interviews has increased, as well as the intelligent interview evaluation tools. In a structured interview, the interviewer needs to observe the interviewee's answers based on the evaluation criteria, and form a profile evaluation of the interviewee's personality traits, communication skills, and leadership, so as to judge whether the interviewee's characteristics match the position. Among them, personality evaluation is a widely accepted evaluation method among companies. Because personality traits affect people's language expression, interpersonal communication and other aspects, it is an important reference to assist the interviewer to decide whether an interviewee meets their job requirements. Based on this, a fine-grained interview evaluation method based on the long short term memory (LSTM) and the hierarchical keyword-question attention mechanism (HKQA-LSTM) is proposed, which aims to score the different personality dimensions of the interviewees and obtain a comprehensive interview score based on this. First, we effectively filter out important words and sentences that are closely related to personality traits in the interview dialogue by introducing a keyword attention mechanism. Then, we use keyword-question level attention mechanism and two-stage model learning mechanism on this basis, and fully combine the multi-scale contextual features of the texts expressed by interviewees to accurately predict personality traits. Finally, through the fusion of personality traits, a comprehensive interpretive evaluation result of the interview is obtained. The experimental results based on real interview scene data show that this method can effectively evaluate the interviewees' different personality traits scores and accurately predict the interviewees' overall scores.

收稿日期:2020-08-18;修回日期:2020-12-07
基金项目:国家自然科学基金项目(61876065);广东省自然科学基金项目(2018A0303130022);广州市科技计划项目(201904010200);中央高校基本科研业务费专项资金项目(D2182480, D2200150)
This work was supported by the National Natural Science Foundation of China (61876065), the Natural Science Foundation of Guangdong Province (2018A0303130022), the Science and Technology Program of Guangzhou (201904010200), and the Fundamental Research Funds for the Central Universities (D2182480, D2200150).
通信作者:吕建明(jmlv@scut.edu.cn)

Key words keyword attention; fine-grained scoring; interview evaluation; two-stage learning; natural language processing

摘 要 海量的在线面试视频数据为智能面试评价提供了重要的数据基础.随着目前全球疫情的蔓延,网络在线面试的需求程度上升,对智能面试评价工具的需求也随之上升.结构化面试中,面试官需要依据评价标准,观察面试者所做的回答,并形成面试者人格特性、沟通技能以及领导力等方面的画像评估,以此判断面试者的特质是否与应聘职位相匹配.其中人格特性评估是公司间广泛使用的一种评估方法,因为人格特性影响着人们的语言表达、人际交往等多个方面,是辅助面试官决策该面试者是否符合其应聘岗位需求的重要参考.基于此,提出了基于循环神经网络长短期记忆(long short term memory, LSTM)以及关键词-问题注意力机制的多层次(hierarchical keyword-question attention LSTM, HKQA-LSTM)细粒度面试评价方法,旨在针对面试者的不同人格特性维度进行打分,并据此得到综合面试得分.首先,通过引入关键词注意力机制有效筛选出面试对话中与人格特性密切相关的重要词句;然后,在此基础上采用了关键词-问题层次注意力机制和 2 阶段的模型学习机制,充分结合面试者表述文本的多尺度上下文特征,对人格特性进行准确预测;最后通过融合人格特性得到具有较高解释性的面试综合评价结果.基于真实面试场景数据的实验结果表明,该方法能有效地评价面试者的不同人格特性得分,并准确地预测面试者总体得分.

关键词 关键词注意力;细粒度评分;面试评价;2 阶段学习;自然语言处理

中图法分类号 TP391

随着互联网技术不断发展,异步视频面试技术越来越成熟^[1].面试者只需登录到特定的平台,回答出平台上事先设定好的问题.之后面试官们在平台上评估面试者的面试表现,从而判断面试者是否能进入下一轮的面试.这种方式和线下面试相比,大大缩短了招聘周期,面试者在任何地点和任何时间都可以回答问题^[2].

然而,在招聘需求比较旺盛的季节,如针对大学校园招聘的春招和秋招季,异步视频的数量会急剧增加,并且由于全球疫情的蔓延,越来越多的公司采用了这种异步视频的面试形式,使得面试官需要耗费大量的精力处理视频.有研究表明,面试招聘标准越客观,该公司的绩效相对就越好^[3-4],而在实际面试中,面试官通常是在一个连续、集中的时间内,根据一系列的标准对应聘者们的表现做出评价,这些评价会受到周围环境的影响.例如前面应聘者的评分结果可能会影响到面试官对后面应聘者的评分决策,这是面试官的记忆容量瓶颈所导致的.由于面试过程中,面试官需要在短时间内对应聘者的表现做出分析,因此面试官会重点关注与评分标准相关的信息,以此形成应聘者人格特性、沟通技能以及领导力等方面的画像评估,从而做出评分决策.在此过程中,面试官对前面应聘者的记忆会短暂存储下来,当应聘者数量过多时,面试官的工作记忆便会超负荷,

因此长时间处理视频的状态下难以保证面试官评价的客观性^[5-6].

作为缓解面试官人力瓶颈的自动化工具,自动面试评价旨在于针对面试者所做出的回答,自动分析出面试者的人格特性、沟通技能以及领导力等方面的特质,从而辅助面试官判断面试者的特质是否符合其岗位胜任要求,避免了面试官可能存在的主观性.其中的人格特性评估是公司间最广泛使用的一种评估方法^[7],因为人格特性影响着人们的语言表达、人际交往等多个方面,比如人际敏感度过高的人在人际交往中会更加在意他人的评价,导致与他人相处时存在社会焦虑^[8].因此本文将重点研究面试评价中的人格特性评价.然而,为了训练精准可用的自动化评价模型,往往需要大量专业人士的精标数据,构建代价昂贵.与此同时,已有的自动化评价模型可解释性较低,无法结合人力资源部门的先验专业知识.

基于上述挑战,文本提出了一种基于循环神经网络长短期记忆(long short term memory, LSTM)以及关键词-问题注意力机制的多层次(hierarchical keyword-question attention LSTM, HKQA-LSTM)细粒度面试评价方法.该模型是一种 2 阶段学习的分层注意力网络:第 1 阶段中,通过结合人力资源顾问的专业知识,我们将外部具有指导意义的关键词、

问题作为输入,根据面试者在面试中表现出的社交信号^[9-10],评价其不同人格特性的得分,例如针对面试题“请分享一个实例说明您是如何带领团队成功完成某一个任务的?”,如果面试者的回答表述中蕴含沟通、合作、共赢、资源共享等类似含义,则面试官会倾向于认为该面试者的人格是较为宜人合群的;其次,考虑到面试官在面试中会对面试者的综合表现进行评估,即分析面试者的综合素质是否符合应聘岗位,从而决定是否录取^[11].基于此,模型在第2阶段的训练中,通过融合不同人格特性特征向量,对面试者回答的问题进行了一个总体打分.实验结果表明,本文提出的方法能有效地评价面试者的不同人格特性得分与总体得分,在性能上明显超过基线方法.

1 相关工作

1.1 细粒度面试评价

结构化面试(structured interview)是面试中经常采用的一种面试形式,它根据特定职位的胜任能力要求,采用固定的一套设计题库,是一种标准化、公平性强的面试形式.心理学研究组织证明,结构化面试比非结构化面试更加有效地反映面试者能力^[12-13].面试者要在有限时间内证明自己的就业资格.面试官根据面试者的回答,对其人格特性、沟通技能、领导力等方面进行评价,在对面试者有一个总体印象的基础上,面试官会进一步考虑目标岗位的实际需求,判断该面试者的特质是否适合其应聘的职位^[14].其中人格特性评价是公司间广泛使用的面试评价方法^[7],因为它有着持久、稳定的优点,并潜在影响着人们的社交表达,比如社交自信型的人往往会使用更多积极的情感用语,并且对他人会表现出更多的认同和称赞^[15].

细粒度面试评价(特别是自动评估面试者的细粒度人格特性)是社交计算中一个有趣的问题.随着大数据热潮的到来以及计算机算力的提高,人工智能在图像识别与自然语言处理领域均得到了成功的应用,因此也被研究者用来解决面试评价中的人格特性分析问题.

过去的细粒度人格特性分析研究主要依赖于较为繁琐的特征工程,即通过人工定义的一些规则进行特征提取,再利用机器学习方法进行预测.Aydin等人^[16]使用随机森林对视频中的人物性格进行预测;Sinha等人^[17]收集企业的社交媒体帖子,利用词

之间的相似度等统计特征分析员工的6种人格特质(诚实-谦恭性、情绪性、外向性、宜人性、尽职性、经验的开放性);Nambiar等人^[18]利用音频和词汇特征,结合线性回归、隐含狄利克雷分布(latent Dirichlet allocation, LDA)主题生成模型以及支持向量机方法,对面试者的说话流畅度、自信心等特质做出预测;Muralidhar等人^[19]通过场景模拟,收集了169个结构化面试数据集,使用支持向量机对面试者的积极性、热情性等人格特性进行预测;考虑到人们交流过程中经常会使用非正式文本,钟毓等人^[20]利用基于主成分分析的方法,探究非正式词语与说话者人格特性之间的关系.

近年来,神经网络理论与应用不断发展,其中特定的网络结构,比如循环神经网络,可以很好地对文本进行表征建模^[21-22].因此,针对自动面试评价的问题也涌现了不少使用神经网络的方法.Chen等人^[23]应用神经网络为基础的Doc2Vec模型,对面试者的人格特质进行预测;Hemamou等人^[24]专门收集了销售职业的结构化面试视频,使用循环神经网络模型对面试者的语言以及非语言行为(面部表情、语音)进行建模,预测面试者的可雇佣性;Suen等人^[25]使用卷积神经网络,提取120位面试者的面部特征,预测他们的5种人格特性.

1.2 注意力机制

Bahdanau等人^[26]首次在自然语言处理领域(natural language processing, NLP)中引入了注意力机制,并用在了以编码器-解码器为框架的机器翻译任务中,提高了翻译的准确率.经过实验证明,注意力机制能有效地筛选出重要的信息,并提高神经网络的可解释性^[27].因此,在自动面试评价中引入注意力机制,可以使神经网络更多地关注面试者文本中包含的有效信息,忽略不重要的信息.

不少研究者针对不同NLP任务的具体特点,提出了不同的注意力计算方式.Lin等人^[28]提出了一种基于自我注意力(self-attention)机制来提取句子的可解释性嵌入;Yang等人^[29]根据文档中存在词语、句子、篇章这种层次结构特点,提出了一种基于层次注意力(hierarchical attention)机制的模型,用来捕获输入的层次结构信息.

总体上,基于注意力机制的循环神经网络能有效地对文本序列进行建模,但是仅仅依靠端到端的训练而没有融入人力资源的专业知识的方法往往依赖大量的数据训练,需要消耗人力资源顾问大量的时间进行样本标注,人力成本极高.此外,以往面试评价中注意力向量通常是由端到端的学习得来,面试官

无法得知注意力权重所表达的含义.基于此,本文结合人力资源顾问的专业知识,将关键词与问题2种外部信息嵌入到模型中,提高了模型的可解释性,并且能在小样本上取得较好的预测效果.

2 基于关键词注意力的细粒度面试评价方法

2.1 任务定义及方法概述

总体任务结构如图1所示.本文将自动面试评价任务形式化描述为一个回归任务:给定一个面试

者表述的文本内容 $W=(w_1,w_2,\cdots,w_n)$ 和关键词词组 $K_i=(k_{i1},k_{i2},\cdots,k_{im})$,预测对应的面试者不同人格特性得分 y_i 以及对应的总体评分 z .其中 n 是文本 W 的长度, i 是人格特性项的下标, K_i 是不同人格特性下的关键词词组, m 是词组 K_i 的长度.任务定义中,模型除了在第1阶段对面试者的不同人格特性进行打分,还会在第2阶段根据第1阶段抽取出的人格特征向量,对面试者进行小题总分 z 的预测,人格特征向量的融合分为先融合与后融合2种模型.

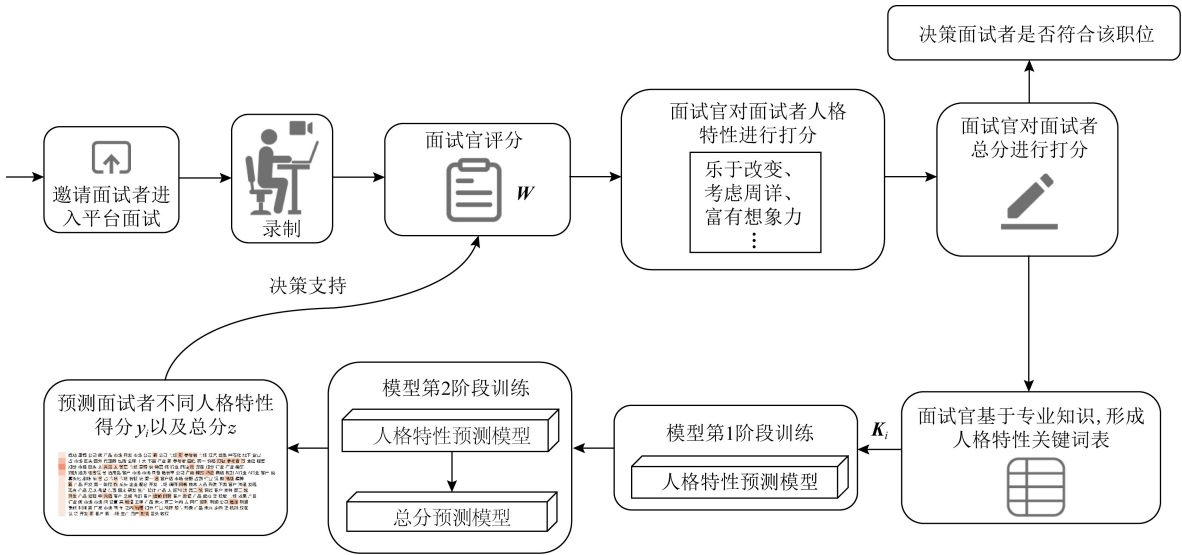


Fig. 1 The overall task structure

图1 总体任务结构

鉴于注意力机制在神经网络模型中有着出色的表现以及可解释性,本文在自动面试评价中引入关键词以及面试问题2种外部信息作为注意力导向.其中关键词是通过对面面试者的回答进行词频统计,接着人力资源顾问依据专业经验筛选出最能代表其人格特性的关键词,不同的人格特性关键词最终形成关键词表,本文实验部分给出了关键词表示例.由于模型训练过程中有了人的参与,该方法能有效提升模型的可解释性.

此外,考虑到面试者在面试中往往会分层次地表达自己的观点,本文使用句子-文档的层次网络来模拟面试者表述文本中的层次结构关系.在句子层次中,我们使用关键词注意力来对句子进行建模,筛选出句子中的重要单词.在文档层次中,我们使用问题注意力来对句子层次的输出进行建模,筛选出面试者表述中更贴切问题的回答,下面将对我们的方法做出详细介绍.

2.2 基于关键词注意力机制的双向 LSTM 模型

双向 LSTM 网络能有效编码上下文信息^[30],本文为了使双向 LSTM 模型能关注到面试者表述中的不同重要信息,从而分析出面试者的不同人格特性,提出了一种基于关键词注意力机制的双向 LSTM 模型.不同于以往的方法,我们考虑到面试官往往会根据面试者表述中的文本内容来推断面试者的教育背景和人格特征,因此提出了关键词注意力以及关键词匹配2种机制,重现面试官在视频面试中采用的评价方法.

如图2所示,模型包含词向量层、双向 LSTM 层、关键词注意力层、关键词匹配层、多层感知机 (multi-layer perceptron, MLP) 层与 sigmoid 层,接下来对基于关键词注意力机制的双向 LSTM 模型进行介绍.

1) 词向量层.给定面试者表述的文本内容 $W=(w_1,w_2,\cdots,w_n)$ 与人格特性关键词词组 $K_i=(k_{i1},$

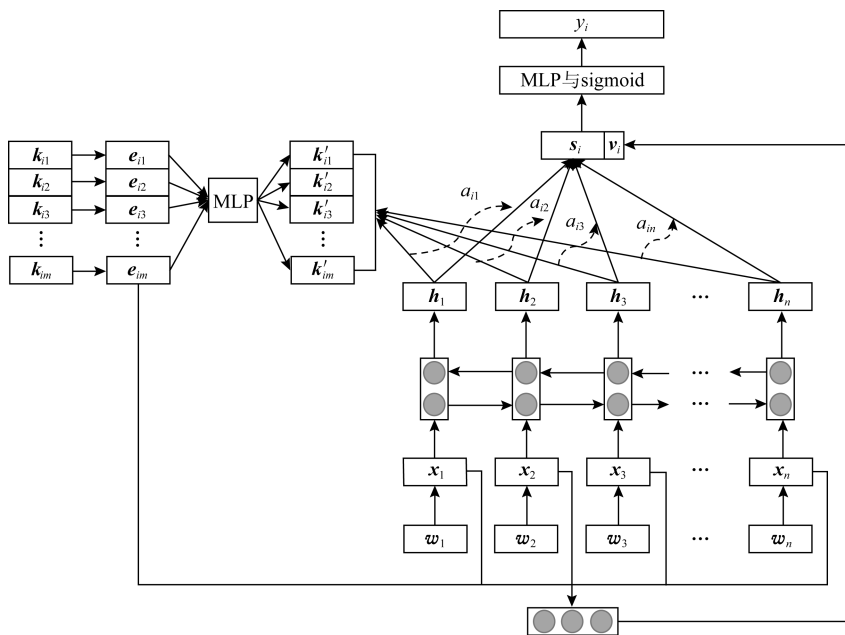


Fig. 2 Keyword attention model

图2 关键词注意力模型

k_{i2}, \dots, k_{im}), 该模型首先使用由百度百科语料预训练的中文词向量^[31], 将基于词袋模型的 one-hot 向量转换成嵌入向量表示, 我们将文本内容的嵌入矩阵记为 $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)$, 作为双向 LSTM 层的输入. 将关键词词组的嵌入矩阵记为 $\mathbf{E}_i = (\mathbf{e}_{i1}, \mathbf{e}_{i2}, \dots, \mathbf{e}_{im})$, 作为关键词注意力层以及关键词匹配层的输入, 设 j 为输入元素的下标, i 为人格特性项的下标, $\mathbf{W}_e \in \mathbb{R}^{d \times N}$ 为预训练词向量, 其中 d 为词向量的长度, N 为词表的大小, 则

$$\mathbf{x}_j = \mathbf{W}_e \mathbf{w}_j, \quad (1)$$

$$\mathbf{e}_{ij} = \mathbf{W}_e \mathbf{k}_{ij}. \quad (2)$$

2) 双向 LSTM 层. 将嵌入矩阵 $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)$ 输入到 LSTM 层, 在获得前向 LSTM 与后向 LSTM 的编码输出后, 对其进行拼接, 得到双向 LSTM 层的输出 $\mathbf{H} = (\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_n)$, 设 Γ 为 LSTM 的基本单元, 则

$$\vec{\mathbf{h}}_j = \vec{\Gamma}(\mathbf{x}_j), \quad (3)$$

$$\overleftarrow{\mathbf{h}}_j = \overleftarrow{\Gamma}(\mathbf{x}_j), \quad (4)$$

$$\mathbf{h}_j = [\vec{\mathbf{h}}_j; \overleftarrow{\mathbf{h}}_j]. \quad (5)$$

3) 关键词注意力层. 将关键词嵌入矩阵 $\mathbf{E}_i = (\mathbf{e}_{i1}, \mathbf{e}_{i2}, \dots, \mathbf{e}_{im})$ 输入到 MLP 层做映射, 形式为

$$\mathbf{k}'_{ij} = \mathbf{W}_k \mathbf{e}_{ij} + \mathbf{b}_k, \quad (6)$$

其中 $\mathbf{W}_k \in \mathbb{R}^{d \times d}$, $\mathbf{b}_k \in \mathbb{R}^d$, 记映射后的矩阵为 $\mathbf{K}'_i = (\mathbf{k}'_{i1}, \mathbf{k}'_{i2}, \dots, \mathbf{k}'_{im})$, 将矩阵 \mathbf{K}'_i 作为注意力机制的导向, 此时, 不同的矩阵 \mathbf{K}'_i 将会分配不同的注意力权

重到双向 LSTM 的输出中. 因此模型能够根据不同的人格特性, 关注到面试者文本信息中的不同方面. 注意力计算方式为

$$u_{it} = \text{att}(\mathbf{h}_t, \mathbf{K}'_i), \quad (7)$$

$$a_{it} = \frac{\exp(u_{it})}{\sum_{k=1}^n \exp(u_{ik})}, \quad (8)$$

$$\mathbf{s}_i = \sum_{t=1}^n a_{it} \mathbf{h}_t, \quad (9)$$

$$\text{att}(\mathbf{h}_t, \mathbf{K}'_i) = \max_{j \in \{1, 2, \dots, m\}} \{ \mathbf{k}'_{ij}^T \tanh(\mathbf{W}_s \mathbf{h}_t + \mathbf{b}_s) \}, \quad (10)$$

其中 att 为注意力函数. 根据注意力函数计算 \mathbf{h}_t 的权重 u_{it} 后, 对 u_{it} 进行归一化处理得到 a_{it} , 最后根据 a_{it} 对矩阵 $\mathbf{H} = (\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_n)$ 中的元素进行加权求和, 得到特征向量 \mathbf{s}_i .

4) 关键词匹配层. 除了考虑经过上下文编码的隐含向量外, 我们认为原始的嵌入向量表示同样重要, 因此引入了关键词匹配机制. 人力资源顾问形成的关键词表中, 共包含卓越关键词词表、优秀关键词词表、普通关键词词表, 我们将其嵌入后的矩阵分别记为 $\mathbf{E}_i, \mathbf{G}_i, \mathbf{O}_i$, 其中 i 是关键词词组的下标. 关键词匹配的计算方法为

$$\mathbf{v}_i = [\text{match}(\mathbf{E}_i, \mathbf{X}); \text{match}(\mathbf{G}_i, \mathbf{X}); \text{match}(\mathbf{O}_i, \mathbf{X})], \quad (11)$$

$$\text{match}(\mathbf{E}_i, \mathbf{X}) = \max_{z \in \{1, 2, \dots, n\}} \{ \max_{j \in \{1, 2, \dots, m\}} \{ \mathbf{e}_{ij}^T \mathbf{x}_z \} \}, \quad (12)$$

$$\mathbf{s}'_i = [\mathbf{s}_i; \mathbf{v}_i], \quad (13)$$

其中 \mathbf{X} 为嵌入后的文本信息矩阵. 将匹配向量 \mathbf{v}_i 与

关键词注意力层输出的特征向量 s_i 作拼接,得到最终的文本特征向量 s'_i .

5) MLP层与sigmoid层.我们将文本特征向量 s'_i 输入到 MLP 层后,通过 sigmoid 层归一化得到最终的预测得分 y_i

$$y_i = F \times \text{sigmoid}(\mathbf{W}_o \mathbf{s}'_i + \mathbf{b}_o), \tag{14}$$

其中 F 为缩放因子,即实验中设定的总得分.

2.3 关键词-问题层次注意力网络

考虑到面试过程中面试者表述的内容是具有层次关系的长文本,且不同的句子表述的内容相对问题来说有侧重点,受 Yang 等人^[29]工作的启发,我们将长文本分为句子层次以及文本层次,在引入关键词注意力的基础上,采用了一种关键词-问题的层次注意力网络,用来捕获文本信息中的层次结构信息.此外,模型使用 2 阶段的学习机制,在第 1 阶段中,模型针对文本信息输入,预测不同的人格特性得分;在第 2 阶段中,模型将对第 1 阶段输出的人格特征向量进行融合,得到面试者小题总得分.

2.3.1 第 1 阶段训练:单项人格特性项训练

如图 3 所示,首先我们将句子层次的文本信息

按式(1)~(13)编码,得到加权后的向量 s'_i 作为句子层次的代表向量,以此类推,在输入不同的句子进行编码后,我们将得到句子表征矩阵 $\mathbf{S}_i = (\mathbf{s}'_{i1}, \mathbf{s}'_{i2}, \dots, \mathbf{s}'_{il})$,其中 l 为句子的数量, i 为关键词词组的下标.同样地,我们将 \mathbf{S}_i 输入双向 LSTM 网络中,并对输出进行拼接,得到隐向量矩阵 $\mathbf{H}'_i = (\mathbf{h}'_{i1}, \mathbf{h}'_{i2}, \dots, \mathbf{h}'_{il})$,其中:

$$\overrightarrow{\mathbf{h}}'_{ij} = \overrightarrow{\Gamma}(\mathbf{s}'_{ij}), \tag{15}$$

$$\overleftarrow{\mathbf{h}}'_{ij} = \overleftarrow{\Gamma}(\mathbf{s}'_{ij}), \tag{16}$$

$$\mathbf{h}'_{ij} = [\overrightarrow{\mathbf{h}}'_{ij}; \overleftarrow{\mathbf{h}}'_{ij}]. \tag{17}$$

在结构化面试过程中,面试者需要围绕问题内容进行回答,因此我们在文档层次上融入了问题注意力机制,该机制会将面试者回答文本中更加贴合问题的语句赋予更高的权重,设问题为 $\mathbf{Q} = (\mathbf{q}_1, \mathbf{q}_2, \dots, \mathbf{q}_v)$,其中 v 为问题长度,将问题输入到 LSTM 中进行编码,形式为

$$\mathbf{q}'_j = \mathbf{W}_e \mathbf{q}_j, \tag{18}$$

$$\mathbf{u}'_j = \overrightarrow{\Gamma}(\mathbf{q}'_j). \tag{19}$$

将 LSTM 层中最后一个时间步输出的隐含向量

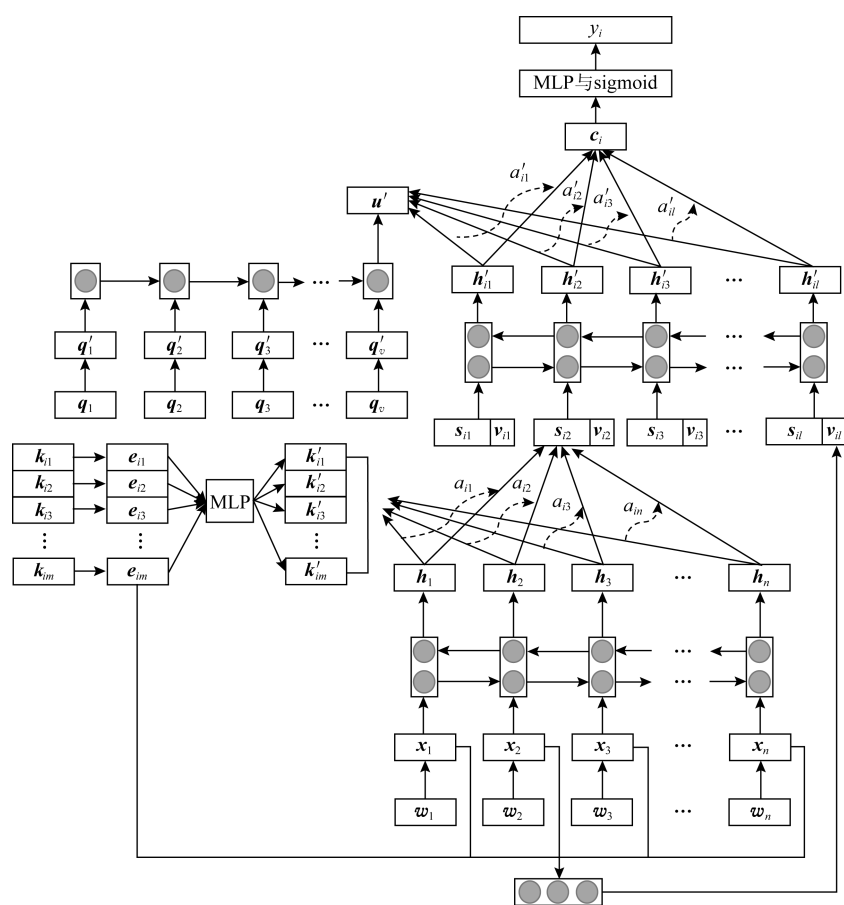


Fig. 3 Keyword-question attention model

图 3 关键词-问题注意力模型

作为问题编码向量 u' , 问题注意力计算方式为

$$u'_{it} = \tanh(W_c[h'_{it}; u'] + b_c), \tag{20}$$

$$a'_{it} = \frac{\exp(u'_{it})}{\sum_{k=1}^l \exp(u'_{ik})}, \tag{21}$$

$$c_i = \sum_{t=1}^l a'_{it} h'_{it}. \tag{22}$$

经过问题注意力机制加权处理后, 获得人格特征向量 c_i , c_i 在通过 MLP 层与 sigmoid 层的计算后获得面试者该项的分数预测:

$$y_i = F \times \text{sigmoid}(W_o c_i + b_o). \tag{23}$$

任务的目标函数为均方误差(mean-square error, MSE), 给定 N 对回答记录-人格特性关键词-问题对 (W_i, K_i, Q_i) , MSE 损失的计算方式为

$$\begin{aligned} loss = \frac{1}{N} \sum_{i=1}^N (p(W_i, K_i, Q_i) - \\ y(W_i, K_i, Q_i))^2 + \lambda \|\theta\|, \end{aligned} \tag{24}$$

其中, $p(W_i, K_i, Q_i)$ 为模型预测得分, $y(W_i, K_i, Q_i)$ 为面试官打的真实得分, λ 为 L_2 正则化惩罚项的因子, θ 为模型参数.

由于在模型训练的过程中, 面试问题和面试者作答共同作为输入信号进行编码和参与训练, 模型所学习到的是面试问题和面试者作答之间的匹配程度, 因此模型不依赖于特定的问题, 具有较好的泛化能力. 在实际的应用中, 可通过提高面试问题的多样性, 进一步提升问题适配能力.

2.3.2 第 2 阶段训练: 人格特性特征融合

将第 1 阶段训练得到的基础模型进行迁移, 设输入的文本信息为 W , 基础模型输出的第 i 项人格特性的特征向量为 $c_i, i \in [1, k], k$ 为人格特性项个数, 如图 4 所示:

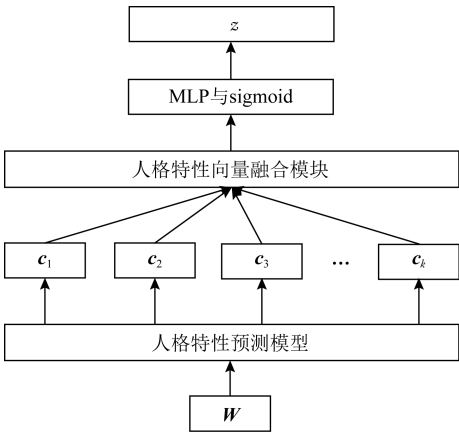


Fig. 4 Personality traits vector fusion

图 4 人格特性特征向量融合

我们将不同人格特性的特征向量进行融合, 分别使用先融合和后融合 2 种模式, 计算方式为

$$z_{\text{first}} = F \times \text{sigmoid}\left(\frac{1}{k} W_u \sum_{i=1}^k c_i + b_u\right), \tag{25}$$

$$z_{\text{last}} = F \times \text{sigmoid}(W_r[c_1; c_2; \dots; c_k] + b_r). \tag{26}$$

先融合模式中, 我们把不同的特征向量取平均; 后融合模式中, 把不同的特征向量进行拼接; 获得融合的特征向量后, 将其输入到 MLP 层与 sigmoid 层, 得出面试者的小题总分预测.

与 2.3.1 节相同, 任务的目标函数为 MSE, 设给定基础模型 M 以及 N 个回答记录 W_i , 设 $p(W_i, M)$ 为模型预测的小题总得分, $z(W_i)$ 为面试官对该问题的真实打分, 则损失的计算方式为

$$loss = \frac{1}{N} \sum_{i=1}^N (p(W_i, M) - z(W_i))^2 + \lambda \|\theta\|. \tag{27}$$

3 实验与分析

3.1 实验数据

我们通过与人力资源公司进行合作, 搭建了异步在线面试平台, 模拟了真实场景下的结构化面试过程. 面试共分为 3 道题, 每道题设定 5min 答题时间. 首先, 我们邀请面试者登录平台, 针对事先设定好的题目进行回答录制, 接着 3 位专业的人力资源顾问依据评分标准, 将对面试者进行人格特性打分(包括乐于说服他人、乐于改变、考虑周详等 18 项人格特性)以及小题总分打分, 打分范围为 0~5 分. 最终 3 位能力资源顾问所打的分值取平均, 作为面试者的最终得分. 打分过程中根据面试者回答记录做出词频统计, 人力资源顾问分别选出最能代表人格特性的关键词形成卓越、优秀、普通关键词表. 我们最终共收集了 96 位面试者的面试数据. 表 1 给出了数据集的详细说明, 表 2 给出了人力资源顾问形成的关键词词表中的其中一项示例.

Table 1 Statistics of Interview Datasets

表 1 面试数据集统计信息

属性	值
人格特性项数目	18
〈回答记录/人格特性/问题〉对数目	3 456
平均文本长度	912.7
平均文本分词数	519.4
去停用词后平均文本分词数	177.4
卓越关键词平均个数	14.3
优秀关键词平均个数	16.3
普通关键词平均个数	11.1

Table 2 Examples of Keywords
表 2 关键词示例

人格特性项	关键词		
	卓越	优秀	普通
乐于说服他人		过程	
	沟通	安排	
	成功	最后	运营
	要求	流程	支持
	解决	转型	为中心
	必须	合作	拜访
	结果	指定	交流
	目标	计划	收集问题
	挑战	协调	跟进
	反思	分工	分派
	引领需求	分享	倾听
	化解	尊重	配合
	谈判	分清主次	不计较
	不厌其烦	认可	
		共识	

3.2 评价指标

评价指标除了模型训练阶段使用的 MSE 之外, 本文还采用皮尔逊相关系数 (Pearson correlation coefficient, PCC)、斯皮尔曼秩相关系数 (Spearman's rank correlation coefficient, SCC) 来表示模型预测结果与人力资源顾问打分的相关程度. 此外, 二次加权 Kappa (quadratic weighted Kappa, QWK) 也常作为分数预测类任务的评价指标. QWK 会对预测分数与真实分数差值较大的项做出更大的惩罚, 计算方式为

$$k = 1 - \frac{\sum_{i,j} w_{i,j} O_{i,j}}{\sum_{i,j} w_{i,j} E_{i,j}}, \tag{28}$$

$$w_{i,j} = \frac{(i-j)^2}{(R-1)^2}, \tag{29}$$

其中, \mathbf{O} 是观察分数矩阵, \mathbf{E} 是预期分数矩阵, i 和 j 分别指实际分数和预测分数, R 是分数能取得的最大值. 矩阵 \mathbf{O} 和矩阵 \mathbf{E} 需要进行归一化. 归一化前的 $O_{i,j}$ 的值表示实际标签为 i 分、模型预测为 j 分的数量; 归一化前的矩阵 \mathbf{E} 通过实际标签值、模型预测值 2 个向量做外积得到.

3.3 参数设置

我们对训练集、验证集、测试集按 8:1:1 的比例划分. 使用预训练的中文词向量^[31], 维度设置为 300, 对于词表中不存在的词, 词向量进行随机采样, 采样范围为 $[-0.01, 0.01]$. 优化器选择 Adam-optimizer, 学习率设置为 0.001. LSTM 的隐藏层维度设置为 200. 第 1 阶段训练的 *batch_size* 设置为 128, 第 2 阶段训练的 *batch_size* 设置为 16. 对于最

大句子长度与句子中的最大单词数长度, 我们使用网格搜索, 在句子最大长度设置为 15, 句子中单词数最大长度设置为 20 的条件下实验结果最佳, 因此实验中统一采用此参数. 此外, 为了防止过拟合, 实验中加入 dropout 机制与 L_2 正则化机制, 并在训练过程中采用了 early stopping 技巧. 为了避免实验的随机性, 本文在实验上采取 10 折交叉验证, 以产生不同的训练集、验证集和测试集, 选择在 10 折交叉验证中的测试数据集上结果的平均值作为模型最终表现.

3.4 实验对比与结果分析

本文采用 3 个基线模型用于对文本信息进行编码, 并将编码向量输入 MLP 层与 sigmoid 层进行预测, 用来与我们的模型作比较:

1) BiLSTM. 该模型利用一个双向的 LSTM 来对文本信息进行编码, 考虑了上下文关系. 模型中的记忆门缓解了序列模型容易发生梯度消失、爆炸的问题, 使模型可以学习到长文本中远距离的依赖关系.

2) HA-LSTM (hierarchical attention LSTM)^[29]. 该模型考虑了文本的层次结构并融入了注意力机制, 将文本信息中不同的句子输入到 LSTM 中获得句子向量表示后, 再将不同的句子向量表示输入到另一个 LSTM 中获得输出.

3) KA-LSTM (keyword attention LSTM). 该模型为 2.2 节介绍的基于关键词注意力的双向 LSTM 模型.

表 3 给出了我们的方法 (hierarchical keyword-question attention LSTM, HKQA-LSTM) 与其他不同的基线模型的性能对比, 分别比较了面试数据集中人格特性项以及小题总分的预测情况. 从表 3 可以发现:

1) HA-LSTM 模型在小题总分的预测上效果比 BiLSTM 模型效果有一定的提升, 其中 PCC 高出 0.034, SCC 高出 0.035, QWK 高出 0.023. 说明在面试过程中, 面试官从整体上看更注重面试者表述文本信息的层次结构. 即对于面试官来说, 他会被面试者回答记录中的某几句话着重吸引, 这也符合现实场景, 证明了层次建模的重要性.

2) KA-LSTM 模型在人格特性以及小题总分的预测上效果比 BiLSTM 模型、HA-LSTM 模型有优势, 说明在文本信息编码的过程中, 关键词注意力机制发挥了作用, 我们将关键词作为注意力的外部导向, 可以有效地对重要的信息进行保存, 对不重要

的信息进行过滤,并且提高了模型的可解释性.

3) 从整体上看, HKQA-LSTM 模型取得的效果均高于基线方法,说明关键词注意力机制以及问题注意力机制能有效地关注到更精确的局部信息,从而帮助模型提升预测的准确性.层次结构也使得模型能够捕获更长序列的依赖关系.

4) 对于小题总分的预测, HKQA-LSTM 模型比 KA-LSTM 模型有了较大的提升, PCC 高出 0.239, SCC 高出 0.259, QWK 高出 0.287.这种现象的原因在于 HKQA-LSTM 模型使用了 2 阶段学习的训练方式,模型在第 2 阶段的训练中,对第 1 阶段抽取出来的人格特性向量进行融合,使得文本表征更加丰富.

Table 3 Performance Comparison of Different Models

表 3 不同模型性能对比

模型	人格特性			小题总分		
	PCC	SCC	QWK	PCC	SCC	QWK
BiLSTM	0.465	0.440	0.376	0.485	0.399	0.344
HA-LSTM	0.458	0.430	0.369	0.519	0.434	0.367
KA-LSTM	0.595	0.588	0.546	0.544	0.447	0.392
HKQA-LSTM	0.624	0.619	0.578	0.783	0.706	0.679

注:加粗部分为最优结果.

表 4 列出了我们的方法 (HKQA-LSTM) 与基线方法在测试集上的预测值与真实值的对比,使用 MSE 来评估不同模型的表现.可以看出,在人格特性和小题总分 2 种预测上,本文方法的预测值相比于基线方法都更加接近真值.与 HA-LSTM 相比,人格特性预测值与真实值的 MSE 误差由 1.764 降低到了 1.376,小题总分 MSE 误差由 0.695 降低到了 0.314.

Table 4 MSE Results of Different Models

表 4 不同模型预测结果的均方误差

模型	人格特性	小题总分
BiLSTM	1.732	0.778
HA-LSTM	1.764	0.695
KA-LSTM	1.543	0.643
HKQA-LSTM	1.376	0.314

注:加粗部分为最优结果.

为了验证本文提出的方法对细粒度评分预测的准确程度,我们对 18 项不同人格特性的预测做出比较分析,使用 PCC 来评估不同方法的表现.实验结果如图 5 所示:

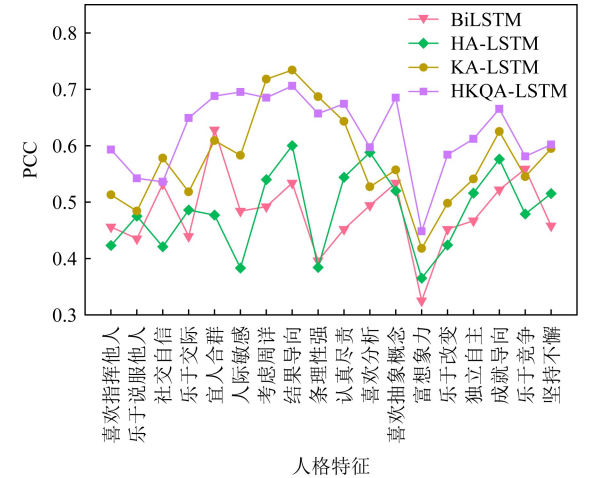


Fig. 5 Comparison of fine-grained ratings of different models

图 5 不同模型细粒度评分对比

从图 5 中可以看出,相比与基线方法,本文提出的 2 种基于关键词注意力的方法在不同人格特性预测的表现上都更为优秀,主要原因在于面试者的口语表述是相对较长的文本,我们的方法拥有关键词注意力机制,可以更有质量地捕捉到不同人格特性下所侧重关注的不同词语和句子,从而使得模型预测效果得到提升.

为了判断模型是否能真的分辨出面试中综合表现较为卓越的面试者,我们将模型对小题总分的预测结果分为 2 类,即卓越(小题总分 ≥ 2.5)与一般(小题总分 < 2.5),评价标准采用准确率(Precision)、召回率(Recall)和 Macro-F1.

我们分别尝试了先融合模式 (HKQA-LSTM-Early) 和后融合模式 (HKQA-LSTM-Late),对第 1 阶段抽取出来的人格特征向量进行融合.表 5 展示了 2 种模式的实验对比.

Table 5 Accuracy of Total Score in Different Modes

表 5 不同模式下小题总分的预测精度

方法	小题总分		
	Precision	Recall	Macro-F1
HKQA-LSTM-Early	0.794	0.848	0.816
HKQA-LSTM-Late	0.824	0.872	0.845

注:加粗部分为最优结果.

从表 5 可以看出,后融合模式比先融合模式有更好的效果,其中 Macro-F1 值提高了 2.9%.造成这种现象的原因是,在实际面试过程中,由于题目的不同,面试官所关注的面试者的人格特性侧重点也

不同.而后融合模式的计算方法可以动态地调整不同人格特性的权重,更符合实际的面试情况,因此对于模型的准确率、召回率都有一定的提升.

3.5 案例分析

为了分析关键词注意力机制预测不同人格特性的有效性,我们进行了样例分析.图 6 包含了 2 个可视化的例子,我们对文本信息编码过程中单词、句子注意力的权重进行了高亮显示,其中颜色越深说明该单词或句子的注意力权重越大.

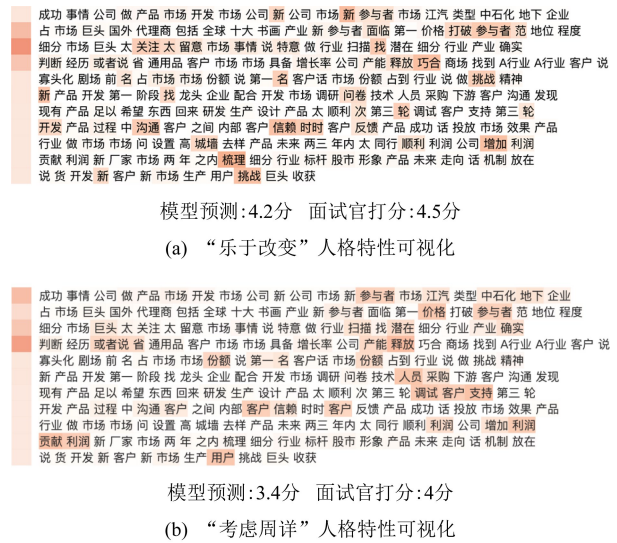
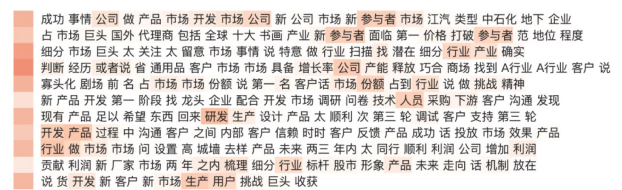


Fig. 6 Attention weight visualization of interview texts
图 6 面试文本注意力权重可视化

从图 6 的可视化例子中可以看出,我们的模型可以有效地关注文本信息中重要的单词,忽略掉不重要的单词.并且针对不同的人格特性,关键词注意力机制使得模型可以关注到同一文本的不同方面.如图 6(a)中,为了判断面试者的“乐于改变”人格特性,模型更加关注“打破”“梳理”“挑战”等词.而图 6(b)中,由于考虑的是面试者的“考虑周详”人格特性,模型则将注意力更多地放在“用户”“客户”“调试”等词,因此该模型可以有效地帮助面试官进行决策.

图 7 展示了模型预测误差较大的一项人格特性项“富想象力”.在该例子中,模型对不同句子的注意力比较平均,并将单词注意力更多地放在了重复率较高的一些名词上.这是由于关键词表中,人力资源顾问给出的该人格特性项的关键词为“提出”“战略布局”“新思路”等,与面试者所表述的出入较大,导致最后模型对各个句子的注意力较为分散.此外,从图 5 的细粒度评分比较中可以看出,模型对于人格特性项“富想象力”的预测与面试官打分的相关程度

较低.一个可能的原因是“富想象力”本身是一项较为抽象的人格特性,因此面试官在打分的过程中会受到面试者的声音、表情动作等其他方面的影响.



模型预测:1.3分 面试官打分:3.5分
“富想象力”人格特性可视化
Fig. 7 Example of negative samples
图 7 负例样例

4 结 语

本文提出了一种基于关键词注意力机制的细粒度面试评价方法,模型利用外部的关键词以及问题作为模型不同层次上的注意力导向,对面试者面试过程中所展现出来的不同人格特性进行预测.模型对第 1 阶段训练得到的人格特性表征尝试了不同的融合方式,得到第 2 阶段的最终模型.在数据集上的实验结果表明,本文提出的模型能有效地对面试者的不同人格特性进行细粒度评价,对于不同的人格特性,模型能筛选出同一文本中不同的重要词句,有效帮助面试官做出决策.

在未来的工作中,我们将考虑进一步探索面试中的语音、视觉模态特征并融合进模型中,并在模型增量学习过程中产生新的人格特性关键词,让人力资源顾问参与完善关键词表,考虑基于图谱的方法挖掘关键词词语之间的同义、上下位等更丰富的语义关系.结合其他半监督的学习方法,通过利用更多面试过程产生的无标签数据来进一步提升细粒度评分的性能.

参 考 文 献

[1] Nikolaou I, Oostrom J K. Employee Recruitment, Selection, and Assessment: Contemporary Issues for Theory and Practice [M]. Oxford, UK: Psychology Press, 2015
[2] Rasipuram S, Rao P, Jayagopi D B. Asynchronous video interviews vs. face-to-face interviews for communication skill measurement: A systematic study [C] //Proc of the 18th ACM Int Conf on Multimodal Interaction (ICMI). New York: ACM, 2016: 370-377

- [3] Ekwoaba J O, Ikeije U U, Ufoma N. The impact of recruitment and selection criteria on organizational performance [J]. *Global Journal of Human Resource Management*, 2015, 3(2): 22-33
- [4] Klepić I. Correlation of recruitment and selection of human resources and the performance of small and medium enterprises [J]. *Naše Gospodarstvo/Our Economy*, 2019, 65(4): 14-26
- [5] Hinson J M, Jameson T L, Whitney P. Impulsive decision making and working memory [J]. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 2003, 29(2): 298-306
- [6] Biehal G, Chakravarti D. Consumers' use of memory and external information in choice: Macro and micro perspectives [J]. *Journal of Consumer Research*, 1986, 12(4): 382-405
- [7] Ryan A M, Reeder M C, Golubovich J, et al. Culture and testing practices: Is the world flat? [J]. *Applied Psychology*, 2017, 66(3): 434-467
- [8] Joo H J, Yeon B, Lee K U. The impact of personality traits on emotional responses to interpersonal stress [J]. *Clinical Psychopharmacology Neuroscience*, 2012, 10(1): 54-58
- [9] Gatica-Perez D. Signal processing in the workplace [J]. *IEEE Signal Processing Magazine*, 2015, 32(1): 121-125
- [10] Knapp M L, Daly J A. *Handbook of Interpersonal Communication* [M]. Thousand Oaks, CA: Sage, 2002
- [11] Xu Jianping, Zhou Han, Li Wenya, et al. Interviewer's rating and influencing factors in structural interview [J]. *Advances in Psychological Science*, 2014, 22(2): 357-368 (in Chinese)
(徐建平, 周瀚, 李文雅, 等. 结构化面试中面试官的评分及影响因素[J]. *心理科学进展*, 2014, 22(2): 357-368)
- [12] Wang Pei, Feng Lijuan. Applicant impression management [J]. *Advances in Psychological Science*, 2006, 14(5): 743-748 (in Chinese)
(王沛, 冯丽娟. 应聘者印象管理研究述评[J]. *心理科学进展*, 2006, 14(5): 743-748)
- [13] DeGroot T, Gooty J. Can nonverbal cues be used to make meaningful personality attributions in employment interviews? [J]. *Journal of Business and Psychology*, 2009, 24(2): 179-192
- [14] Laserna C M, Seih Y T, Pennebaker J W. Um... who like says you know: Filler word use as a function of age, gender, and personality [J]. *Journal of Language and Social Psychology*, 2014, 33(3): 328-338
- [15] Mairesse F, Walker M A, Mehl M R, et al. Using linguistic cues for the automatic recognition of personality in conversation and text [J]. *Journal of Artificial Intelligence Research*, 2007, 30(1): 457-500
- [16] Aydin B, Kindiroglu A A, Aran O, et al. Automatic personality prediction from audiovisual data using random forest regression [C] //Proc of the 23rd Int Conf on Pattern Recognition (ICPR). Piscataway, NJ: IEEE, 2016: 37-42
- [17] Sinha P, Dey L, Mitra P, et al. Mining HEXACO personality traits from enterprise social media [C] //Proc of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA). Stroudsburg, PA: ACL, 2015: 140-147
- [18] Nambiar S K, Das R, Rasipuram S, et al. Automatic generation of actionable feedback towards improving social competency in job interviews [C] //Proc of the 1st ACM SIGCHI Int Workshop on Multimodal Interaction for Education (MIE). New York: ACM, 2017: 53-59
- [19] Muralidhar S, Gatica-Perez D. Examining linguistic content and skill impression structure for job interview analytics in hospitality [C] //Proc of the 16th Int Conf on Mobile and Ubiquitous Multimedia (MUM). New York: ACM, 2017: 339-343
- [20] Zhong Yu, Fei Dingzhou. Judging personality by informal words: A sparse PCA approach [J]. *Journal of Chinese Information Processing*, 2017, 31(1): 192-204 (in Chinese)
(钟毓, 费定舟. 基于稀疏主成分分析的非正式语词的心理-人格特征研究[J]. *中文信息学报*, 2017, 31(1): 192-204)
- [21] Yu Kai, Jia Lei, Chen Yuqiang, et al. Deep learning: Yesterday, today, and tomorrow [J]. *Journal of Computer Research and Development*, 2013, 50(9): 1799-1804 (in Chinese)
(余凯, 贾磊, 陈雨强, 等. 深度学习的昨天、今天和明天[J]. *计算机研究与发展*, 2013, 50(9): 1799-1804)
- [22] Mikolov T, Karafiát M, Burget L, et al. Recurrent neural network based language model [C] //Proc of the 11th Annual Conf of the Int Speech Communication Association (INTERSPEECH). Grenoble, France: ISCA, 2010: 1045-1048
- [23] Chen Lei, Feng G, Leong C W, et al. Automated scoring of interview videos using Doc2Vec multimodal feature extraction paradigm [C] //Proc of the 18th ACM Int Conf on Multimodal Interaction (ICMI). New York: ACM, 2016: 161-168
- [24] Hemamou L, Felhi G, Vandenbussche V, et al. HireNet: A hierarchical attention model for the automatic analysis of asynchronous video job interviews [C] //Proc of the 33rd AAAI Conf on Artificial Intelligence. Menlo Park, CA: AAAI, 2019: 573-581
- [25] Suen H Y, Hung K E, Lin C L. TensorFlow-Based automatic personality recognition used in asynchronous video interviews [J]. *IEEE Access*, 2019, 7: 61018-61023
- [26] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate [C/OL] //Proc of ICLR 2015. [2020-03-05]. <https://arxiv.org/pdf/1409.0473.pdf>

[27] Chaudhari S, Polatkan G, Ramanath R, et al. An attentive survey of attention models [OL]. [2020-03-06]. <https://arxiv.org/pdf/1904.02874.pdf>

[28] Lin Zhouhan, Feng Minwei, Santos C N, et al. A structured self-attentive sentence embedding [C/OL] //Proc of ICLR 2017. [2020-03-07]. <https://arxiv.org/pdf/1703.03130.pdf>

[29] Yang Zichao, Yang Diyi, Dyer C, et al. Hierarchical attention networks for document classification [C] //Proc of the 2016 Conf of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA: ACL, 2016: 1480-1489

[30] Zhou Peng, Qi Zhenyu, Zheng Suncong, et al. Text classification improved by integrating bidirectional LSTM with two dimensional max pooling [C] //Proc of the 26th Int Conf on Computational Linguistics (COLING). Stroudsburg, PA: ACL, 2016: 3485-3495

[31] Li Shen, Zhao Zhe, Hu Renfen, et al. Analogical reasoning on Chinese morphological and semantic relations [C] //Proc of the 56th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2018: 138-143



Chen Chujie, born in 1997. Master candidate. His main research interests include natural language processing and deep learning.
陈楚杰,1997 年生.硕士研究生.主要研究方向为自然语言处理和深度学习.



Lü Jianming, born in 1980. PhD, professor. Member of CCF. His main research interests include data mining and artificial intelligence.
吕建明,1980 年生.博士,教授,CCF 会员.主要研究方向为数据挖掘和人工智能.



Shen Huawei, born in 1982. PhD, professor. Senior member of CCF. His main research interests include social computing and data mining.
沈华伟,1982 年生.博士,研究员,CCF 高级会员.主要研究方向为社会计算和数据挖掘.