软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software, 2023,34(3):1087−1108 [doi: 10.13328/j.cnki.jos.006794]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

预训练语言模型实体匹配的可解释性*

梁 峥, 王宏志, 戴加佳, 邵心玥, 丁小欧, 穆添愉

(哈尔滨工业大学 计算学部, 黑龙江 哈尔滨 150001)

通信作者: 王宏志, E-mail: wangzh@hit.edu.cn

摘 要: 实体匹配可以判断两个数据集中的记录是否指向同一现实世界实体, 对于大数据集成、社交网络分析、网络语义数据管理等任务不可或缺. 作为在自然语言处理、计算机视觉中取得大量成功的深度学习技术, 预训练语言模型在实体识别任务上也取得了优于传统方法的效果, 引起了大量研究人员的关注. 然而, 基于预训练语言模型的实体匹配技术效果不稳定、匹配结果不可解释, 给这一技术在大数据集成中的应用带来了很大的不确定性. 同时, 现有的实体匹配模型解释方法主要面向机器学习方法进行模型无关的解释, 在预训练语言模型上的适用性存在缺陷. 因此, 以 Ditto、JointBERT 等 BERT 类实体匹配模型为例, 提出 3 种面向预训练语言模型实体匹配技术的模型解释方法来解决这个问题: (1) 针对序列化操作中关系数据属性序的敏感性, 对于错分样本, 利用数据集元特征和属性相似度实现属性序反事实生成; (2) 作为传统属性重要性衡量的补充, 通过预训练语言模型注意力机制权重来衡量并可视化模型处理数据时的关联性; (3) 基于序列化后的句子向量, 使用 k 近邻搜索技术召回与错分样本相似的可解释性优良的样本, 增强低置信度的预训练语言模型预测结果. 在真实公开数据集上的实验结果表明, 通过增强方法提升了模型效果, 同时, 在属性序搜索空间中能够达到保真度上限的 68.8%, 为针对预训练语言实体匹配模型的决策解释提供了属性序反事实、属性关联理解等新角度.

关键词: 实体匹配; 预训练语言模型; 可解释性

中图法分类号: TP18

中文引用格式: 梁峥, 王宏志, 戴加佳, 邵心玥, 丁小欧, 穆添愉. 预训练语言模型实体匹配的可解释性. 软件学报, 2023, 34(3): 1087–1108. http://www.jos.org.cn/1000-9825/6794.htm

英文引用格式: Liang Z, Wang HZ, Dai JJ, Shao XY, Ding XO, Mu TY. Interpretability of Entity Matching Based on Pre-trained Language Model. Ruan Jian Xue Bao/Journal of Software, 2023, 34(3): 1087−1108 (in Chinese). http://www.jos.org.cn/1000-9825/6794.htm

Interpretability of Entity Matching Based on Pre-trained Language Model

LIANG Zheng, WANG Hong-Zhi, DAI Jia-Jia, SHAO Xin-Yue, DING Xiao-Ou, MU Tian-Yu

(Faculty of Computing, Harbin Institute of Technology, Harbin 150001, China)

Abstract: Entity matching can determine whether records in two datasets point to the same real-world entity, and is indispensable for tasks such as big data integration, social network analysis, and web semantic data management. As a deep learning technology that has achieved a lot of success in natural language processing and computer vision, pre-trained language models have also achieved better results than traditional methods in entity matching tasks, which have attracted the attention of a large number of researchers. However, the performance of entity matching based on pre-trained language model is unstable and the matching results cannot be explained, which brings great uncertainty to the application of this technology in big data integration. At the same time, the existing entity matching model interpretation methods are mainly oriented to machine learning methods as model-agnostic interpretation, and there are shortcomings in

* 基金项目: 国家重点研发计划(2021YFB3300502); 国家自然科学基金(62232005, 62202126); CCF-华为胡杨林基金数据库专项(CCF-HuaweiDB202204); 黑龙江省博士后资助项目(LBH-Z21137)

本文由“大数据治理的理论与技术”专题特约编辑杜小勇教授、杨晓春教授和童咏昕教授推荐.

收稿时间: 2022-05-16; 修改时间: 2022-07-29; 采用时间: 2022-09-23; jos 在线出版时间: 2022-10-27

their applicability on pre-trained language models. Therefore, this study takes BERT entity matching models such as Ditto and JointBERT as examples, and proposes three model interpretation methods for pre-training language model entity matching technology to solve this problem. (1) In the serialization operation, the order of relational data attributes is sensitive. Dataset meta-features and attribute similarity are used to generate attribute ranking counterfactuals for misclassified samples; (2) As a supplement to traditional attribute importance measurement, the pre-trained language model attention weights are used to measure and visualize model processing; (3) Based on the serialized sentence vector, the *k*-nearest neighbor search technique is used to recall the samples with good interpretability similar to the misclassified samples to enhance the low-confidence prediction results of pre-trained language model. Experiments on real public datasets show that while improving the model effect through the enhancing method, the proposed method can reach 68.8% of the upper limit of fidelity in the attribute order search space, which provides a decision explanation for the pre-trained language entity matching model. New perspectives such as attribute order counterfactual and attribute association understanding are also introduced.

Key words: entity matching; pre-trained language model; interpretability

大数据治理确保以正确的方式对数据和信息进行管理，为大数据的有效应用保驾护航，使得数据成为一个有机整体而不是各自为政. 数据集成是大数据治理的重要基石，其目标是为一组自治和异构的数据源提供一个统一的访问入口，以满足多个应用程序和组织组成的信息孤岛之间数据的管理、分析与共享需求[1]. 然而在不同数据源中，人、书籍、组织等同一实体，即现实世界中的同一个独立客体可能有多种表述方式，同一表述也可能对应着不同的实体，因此，如何识别表述之间存在差异的同一实体，成为大数据集成的一个重要课题，这一任务通常被称为实体识别.

主流的实体识别方法分为 3 步[2]: (1) 将可能描述某一真实世界实体的记录初步形成记录对; (2) 判断记录对是否为同一实体; (3) 对于匹配后的记录对进行聚类，每一类的记录描述同一个真实世界实体. 其中，一个核心任务是实体匹配，如图 1 所示，实体匹配可以通过识别不同源中的相同实体来解决语义歧义问题.

产品名称	价格	厂商地址
苹果	5.00	山东济南
华为P40	4500.00	广东东莞
高数	30.99	黑龙江哈尔滨

产品名称	价格	厂商地址
苹果	5199.00	NULL
华为P50	5200.00	NULL
高等数学	30.99	NULL

图 1　实体匹配——匹配不同数据源中的相同实体对

- 实体匹配

现有的实体匹配方法主要分为传统的基于规则的方法[3]、众包等基于特定优化目标的方法[4,5]、机器学习方法[6–8]. 其中，深度学习方法的嵌入表征技术可以处理实体对中的长文本、劣质数据，该领域最引人瞩目的技术是以 BERT 为代表的预训练语言模型[7,9]. 这类模型具有强大的上下文关联表征学习能力，带给了实体识别领域显著准确率和召回率提升，其效果已经达到了工业级可用的程度[2].

- 可解释性

基于预训练的实体匹配方案与传统的基于规则、基于众包的方案不同，虽然在模型表现上相比于其他深度学习和机器学习方法有很大的提升，但这种方案有着数据需求大、效果不稳定、决策难以理解的缺陷. 尽管已有学者对各类实体匹配问题提出了可解释性分析方案[9,10]，但由于这些模型无关的方案针对通用的机器学习模型或者深度模型设计，在预训练语言模型上的适用情况尚不理想.

- 难点与挑战

面向预训练语言模型实体匹配决策解释的难点包括:

(1) 序列敏感性: 从属性角度，现有模型大多对每个元组进行序列化处理，再将得到的序列化文本数据嵌入到向量空间作为模型的输入. 然而，序列化的合理性在现有方法中分析不足，且实验结果表明，尽管如 Ditto 等方法使用了交换属性序等方法试图解决序列化造成的不稳定性，但由于模型本身的上下文关联训练任务影响，效果对于属性序敏感性依旧很高.

(2) 数据关联性: 从样本角度, 以 LIME[11]为代表的现有可解释性方法多衡量样本各部分对决策结果的影响, 这种方法通过对样本的某一属性值施加较小的扰动得到属性值的重要性. 但注意力机制主要捕捉数据上下文关联, 如何将数据各部分的关联性合理分析、筛选, 形成更清晰的样本决策过程可视化描述, 是在样本角度可解释性的难点.

(3) 样本模糊性: 从数据集角度, 一些可解释性方法对错分样本使用同一数据集中的近似样本进行解释, 但是这些方法仅考虑了样本级的近似性, 近似样本解释与错分样本解释之间的近似性没有很好地度量, 近似样本难以清晰给出错分样本决策解释的提示. 此外, 用输入数据归一化距离度量样本相似度, 并不能很好地利用预训练模型强大的表征能力, 且在数据标注不可见的情况下, 如何将近似样本与模型预测结果结合, 提升可解释性的同时提升模型效果, 是可解释性增强模型的一个机遇和挑战.

如上所述, 基于预训练语言模型的实体匹配技术在数据的序列、关联、样本等角度有着鲜明的难解释特点, 一方面, 尽管在时间序列[12]、自然语言处理[13]等其他应用领域的可解释性研究中有与上述难点有关的利用序列关系[12]、关联约束[14,15]、近似样本搜索[16]等研究成果, 但这些模型相关的工作并不能适用于实体匹配问题; 另一方面, 在本文关注的实体匹配领域, 现有的模型无关的可解释性方法未考虑序列化、将算法视作黑盒、未考虑高维嵌入表征, 因而其应用在基于预训练语言模型的实体匹配方法时, 在反事实空间、样本关联、表征解释等角度存在信息缺失, 给出的模型决策解释难以助力用户深入理解模型决策的影响因素. 可见, 本文所总结的 3 个可解释性方面的难点和挑战是预训练语言模型这一自然语言处理领域模型应用到实体匹配这一任务上造成的. 而其后果是, 如本文第 2.1 节和第 2.2 节的预实验所示, 模型对于属性序敏感、在低置信度样本上表现较差, 且错分样本解释结果的针对性较差, 这些缺陷的原因仍然不为用户所知. 此外, 随着近年来自然语言处理领域新型、大型预训练语言模型的涌现, 该类模型在工业界场景下语义复杂、数据质量问题容忍的实体匹配任务上有着巨大的性能潜力. 因此, 亟需针对该类模型的特点提出针对性决策解释策略, 以保证数据治理流水线中该类模型部署的可信度和稳定性.

本文旨在解决上述预训练语言模型实体匹配任务中的可解释性研究中的挑战, 针对给定的实体匹配数据集, 本文提供样本的列级属性序反事实、列级数据关联解释、行级近似样本解释提示, 可以用于错分样本子集或用户关注样本点的兼顾行列的决策解释, 并为低置信度样本提供了可解释效果增强方案. 本文的主要贡献包括:

(1) 基于元学习的列级属性序反事实生成: 针对序列敏感性的挑战, 本文在测试属性变换对数据集匹配结果的影响后, 通过元特征的选取与属性对齐, 将经验映射到新任务上, 构建属性-偏序图, 并求解属性-偏序图上的最大权拓扑排序以生成反事实样本, 搜索错分样本的最优属性序. 本文提出的 MLARC (meta-learned attribute ranking counterfactual)算法生成属性序反事实, 保真度平均达到属性序空间保真度上限的 68.8%.

(2) 基于注意力机制的列级数据关联性度量: 针对数据关联性的挑战, 通过模型的隐层输出计算注意力矩阵, 进而通过注意力矩阵中的高权重热力图, 以挖掘样本各部分上下文表征的关联性度量, 以此作为属性重要性度量的补充, 以生成深入理解错分样本决策依据的样本解释结果. 通过高注意力权重样本筛选, 生成实体对之间属性关联对决策结果影响的可解释分析.

(3) 基于一致显著度的行级近邻搜索与模型增强: 针对样本模糊性的挑战, 对于错分样本, 本文提出的 CSKNN (consistent-saliency k-nearest neighbor)方法搜索具有相近关联表征的近邻训练样本, 以提供易于理解的提示给复杂样本的决策解释; 同时, 使用基于一致显著度评分的 k 近邻搜索方法 KNNE 增强低置信度样本的模型决策, 以兼顾模型效果的可解释性和稳定性. 一致显著度兼顾样本表征近似度和解释显著性, 通过简单的加权 k 近邻搜索实现了解释提示与低置信度模型决策增强.

本文第 1 节介绍深度实体匹配和其上可解释性研究的相关工作. 第 2 节介绍基于预训练语言模型的实体匹配原理和预实验. 第 3 节介绍本文提出的多层次模型解释方法. 第 4 节给出实验评估验证本文方法的有效

性. 最后, 在第 5 节总结本文主要内容并展望未来工作.

1 深度实体匹配相关工作

1.1 实体匹配

主流实体匹配方法包括基于启发式规则的方法、基于众包的方法和基于机器学习的方法等.

- 基于规则的实体匹配. 基于规则的实体匹配通常有 3 种产生方式产生规则: 自动生成、基于概率模型以及专家提供规则. 文献[17]提出了 Swoosh 实体解析算法框架, 当匹配操作和合并操作满足 ICAR 特性: 幂等性、交换性、结合性、可被代表性, 就能使用 F-Swoosh 或 R-Swoosh 降低匹配操作的次数、提高实体解析的效率. 文献[3]在依赖推导的基础上提出了 3 种算法: SiFi-Greedy、SiFi-Gradient、SiFi-Hill, 为匹配规则和候选规则选择相似性函数和阈值. 文献[18]提出运算符树(operator tree)结构, 把匹配规则模型转化为树: 根节点是并集操作符, 中间节点是相似性函数, 叶子节点是数据. 最后, 通过训练集来产生结果.
- 基于众包的实体匹配. 众包实体匹配方法将实体对候选集合发布到众包平台上, 由大量人员对实体对匹配与否给出判断来得到结果. 传统众包方法代价高且存在错误回答的风险, 因此研究人员提出了许多优化方法. 文献[19]首先将未标记的实体对训练集分成两部分: 一部分利用众包进行标记, 另一部分利用记录之间的传递关系进行标记, 以最小化众包对数量、降低成本. 文献[20]同样着力于利用传递性来减少众包操作中询问的数量, 证明了文献[19]中提出的寻找最优询问策略是 NP 难问题, 并提出两种方案: 随机顺序询问和基于优先级的询问. 文献[4]提出了一个基于偏序的众包框架 POWER, 通过构造一个基于偏序的询问图进行提问和推断. 该框架降低了众包成本, 且可以容忍众包和偏序引入的错误.
- 基于机器学习的实体匹配. 实体匹配可以视作二分类问题, 该类方法把传统的分类、特征提取等思想应用在实体对上以得到二分决策器的方法, 通常有更高的精确率. 文献[21]提出了一种 Magellan 的实体匹配系统, 该系统提供了完整的操作指南以及多步骤的工具供选择, 已被应用于许多工业场景中. 文献[22]着力于解决模型训练需要大量标签数据的问题, 提出了无监督学习方法 ZeroER. 该方法基于高斯混合生成训练集, 提出了特定正则化技术改善过拟合, 最后利用传递性提高准确率, 达到与监督学习相当的性能. 文献[23]则提出了基于主动学习的实体匹配框架, 允许用户对组件进行选择, 对于某些数据集, 甚至可以超过监督学习的性能.

1.2 深度实体匹配

- 深度实体匹配

随着深度学习快速发展, 这一技术开始被应用于实体匹配领域. 文献[6]提出了一种准确、高效且易于使用的基于元组分布式表示的新型实体匹配系统——DeepER, 元组表示的学习采用了具有长短时记忆隐藏单元的双向递归神经网络, 并通过端到端的全局方法进行调整. 与最先进的实体匹配解决方案以及在 citations, products 和 proteomics 的多个基准数据集上公布的方法相比, DeepER 显示出优越的性能. 文献[24]探索了深度学习技术在实体匹配问题上的应用, 对不同深度学习技术, 如 word2vec、GloVe、fastText、RNN、Attention 等在不同实体匹配问题上的表现进行了对比、评估.

- 基于预训练语言模型的实体匹配

预训练语言模型能够捕捉文本中的上下文关联形成表征嵌入, 其很明显的优势就是直接提高了模型性能, 在许多数据集上都取得了最先进的结果, 尤其擅长处理长文本属性值、脏数据等复杂场景下的情况[7,9,25]. 2017 年出现的 Transformer 架构[13]引发了基于预训练模型的研究浪潮, 随后出现的 Bert 模型更是进一步提高了许多任务的性能. 文献[26]分析了 4 种最新的基于注意力的 Transformer 结构——BERT、XLNet4、RoBERTa 和 DistilBERT——在实体匹配任务中的表现, Transformer 架构在 EM 中的表现优于经典的深度学习方法. 文献

[7]提出了一种基于预训练语言模型 Bert 的实体匹配系统 Ditto, 首先将实体记录对建模成序列对进行预训练, 并加入了领域知识注入、概括过长序列、数据增强的模块, 性能大大提升. 文献[9]提出了双目标 BERT 训练方法——JointBERT, 使得模型在预测匹配或不匹配之外, 预测训练对中每个实体的实体标识符(图 2(a)中, ①代表 DITTO[7]模型特有部分, ②代表 JointBERT[9]模型特有部分, 其余部分为二者公共组件).

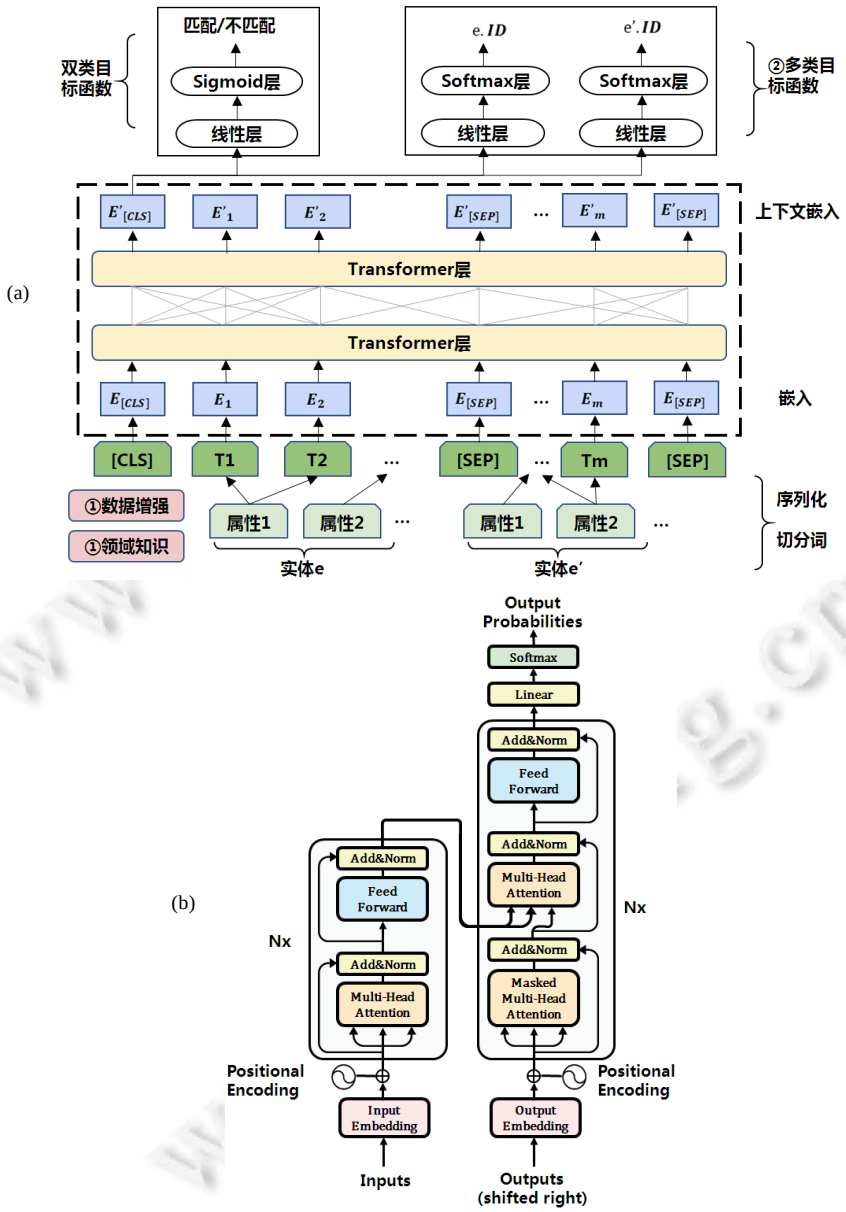


图 2 基于预训练语言模型的实体匹配模型[7,9]和其中 Transformer 层的架构[13]

1.3 实体匹配可解释性

尽管深度实体匹配模型提供了优秀的结果, 但大多数模型是由数据驱动的黑盒模型, 预测过程是无法理解的, 这一方面会降低模型的可信度, 难以将模型部署; 另一方面, 也会带来难以解决的安全问题. 对于专家和开发人员而言, 不透明性会对模型的调试、比较带来困难, 从而模型的改进受到了极大的限制.

因此, 随着深度实体匹配的研究, 实体匹配模型的可解释性成为研究人员关注的课题. 文献[10]提出了一个解释不同粒度实体解析分类器的工具 EXPLAINER, 通过不同场景展示了 EXPLAINER 的全局解释功能、模型分析功能及差异分析功能. 文献[27]提出了一种面向机器学习实体匹配模型预测错误风险的可解释排序方法, 基于风险特征自动化设计和风险评估模型构建与训练, 给出了可解释的预测错误风险分析. 针对预训练模型, JointBERT[9]结合 LIME[11]解释器和领域特定词分析了不同深度学习模型的匹配决策, 最终得出基于 BERT 的模型比基于 RNN 的模型更能专注于相关词类的结论.

然而, 现有可解释性方法少有关注任务的属性关联性, 以及预训练模型在数据预处理、表征学习过程中的特点. 因此, 本文以现有的预训练语言模型实体匹配方法为基础, 整合模型在不同数据粒度和模型角度的信息, 从而生成关联分析和模型反事实, 以理解模型的预测结果. 利用一致显著度搜索加权 k 近邻样本, 提高模型的性能、降低解释的模糊性.

2 预备知识

本文所提方法主要针对预训练语言模型实体匹配的不稳定性和不可解释性, 下面就相关概念和基本知识予以介绍.

2.1 基于预训练语言模型的实体匹配

在实体匹配的应用中, 其最具代表性的预训练语言模型是 BERT 类模型, 包括 BERT、distilBERT、RoBerta 等. 这类模型以 Transformer[13]为基本单元, 构建复杂的深度神经网络结构. 相较于传统的 RNN 和词向量方法, 它们采用多个如图 2(b)所示的 Transformer 层进行预训练, 以生成深层双向语言表征的 BERT 模型, 能够更充分地挖掘文本中蕴含的上下文语义信息.

得益于维基百科等大规模数据集上进行 MLM (masked language model)和 NSP (next sentence prediction)任务[16]的无监督预训练, BERT 类模型一般将[CLS]标志[7]放置在句首, 该标志本身没有语义, 但其嵌入向量可以生成结合上下文多尺度语义信息的精准句子表征[加引用], 以用于后续文本分类任务; 此外, 训练 BERT 时通常会输入 2 个句子, 因此在 BERT 模型的设置中, 这两个句子间用[SEP]来进行分隔. Ditto 和 JointBERT 等基于预训练语言模型的实体匹配方法沿用了这一设计, 但这类方法在处理实体匹配时, 还面对着两个关键需求.

(1) 预处理需求: 如何将实体记录对转化为“句子”输入给模型?

(2) 精调需求: 如何设计微调过程的目标函数?

2.1.1 序列化

针对预处理需求, 首先需要将原始数据集中元组对表示成序列形式输入给预训练模型. 本文假设所处理的数据集中, 实体对的属性是完全相同的. 对于元组对 $e,e'=\{attr_i,val_i\}_{1\leqslant i\leqslant k},\{attr'_j,val'_j\}_{1\leqslant j\leqslant k}$, JointBERT 将其序列化为:

$$[CLS]val_1...val_p[SEP]val'_1...val'_q[SEP].$$

Ditto 将其序列化为:

$$[CLS]serialize(e)[SEP]serialize(e')[SEP],$$

其中, $serialize(e)$=$[COL]attr_1[VAL]val_1\ldots[COL]attr_k[VAL]val_k$.

此外, Ditto 还基于领域知识标注关键的字符串(如在电话号码的最后 4 位前后添加[*LAST*]与[/*LAST*])避免错配、统一化数据格式(如 5%与 5.00%被改写为 5.0%), 且使用属性、字符串的删除、打乱、交换等操作进行数据增强.

2.1.2 目标函数

针对精调需求, 需要构造特定的预训练任务训练模型参数, 训练后的模型在下游任务中只需进行参数微调就能取得很好的结果. 对于元组对 $e,e'=\{attr_i,val_i\}_{1\leqslant i\leqslant k},\{attr'_j,val'_j\}_{1\leqslant j\leqslant k}$, 二者 id 分别为 $e.ID$ 和 $e'.ID$, 二者相等的 one-hot 标签为 y; 模型的预测结果分别为 $\hat{e}.ID$, $\hat{e}'.ID$, $\hat{y}$. Ditto 将损失函数设置为:

$$L_{Ditto} = BCEL(y, \hat{y}).$$

JointBERT 将损失函数设置为:

$$L_{JBERT} = BCEL(y, \hat{y}) + CEL(e.ID, \hat{e}.ID) + CEL(e'.ID, \hat{e}'.ID),$$

其中, CEL 代表交叉熵, BCEL 代表二元交叉熵. 精调时, 均将 CLS 的 embedding 输入给线性层和 Sigmoid(二元目标)/Softmax(多元目标)激活函数层.

2.2 模型的不稳定性与不可解释性

尽管基于预训练语言模型的实体匹配有着优秀的效果, 其针对序列数据的黑盒深度神经网络设计导致其稳定性和可解释性上存在缺陷, 包括以下几个方面.

(1) 不稳定的属性序: 交换实体记录的属性序生成不同数据序列输入对于 JointBERT 和 Ditto 模型的效果有着很大的稳定性影响. 即便 Ditto 模型使用属性换序操作 *attr_swap* 来进行数据增强试图解决序列顺序的问题, 但效果并不显著. 如表 1 所示, 在实体匹配公开数据集上, Ditto 模型的 *F*1-*score* 在两个数据集上模型表现的波动($\Delta F1/F1_{avg}$)平均为 37.52%, JointBERT 模型的 *F*1-*score* 在两个数据集上模型表现的波动($\Delta F1/F1_{avg}$)平均为 27.27%.

(2) 不稳定的低置信度预测结果: JointBERT 和 Ditto 模型在部分数据上的预测置信度偏低, 其结果是在实体匹配公开数据集上, 这部分低置信度数据上的模型表现显著弱于高置信度数据上的性能. 如表 2 所示, 10 个数据集上, 随着置信度的升高, 在大部分数据集上, 其 *F*1-*score* 呈现逐渐上升的趋势.

(3) 难以解释的样本关联: 机器学习任务的庞大数据量往往带来解释信息冗余, 一种常见的可解释方法是使用有代表性的样本[10]或对于给定错分样本的近似样本[28]来解释模型的效果. 然而, 针对这些代表性或近似样本本身的属性级解释在实体匹配的场景下却存在缺陷: 一方面, 实体匹配主要关注实体对两个同义属性值之间的关联, 而现有 LIME 等主流样本解释方法的主流思路是尝试将微小扰动施加在属性值上以衡量单个属性值的重要性; 另一方面, 如图 2(b)所示的 Transformer 结构图中, 核心部分是多头注意力机制(multi-head attention), 但其主要目的是利用模型集中于不同位置的上下文关联能力, 而非单个属性的重要性.

表 1 属性序对模型稳定性的影响

模型	数据集	属性数	$F1_{max}$	$F1_{min}$	$F1_{avg}$	$\Delta F1$	$\Delta F1/F1_{avg}$ (%)
Ditto	Beer	4	0.846 2	0.434 8	0.726 9	0.411 3	56.58
	Amazon-Google	3	0.744 3	0.725 7	0.735 1	0.018 6	2.53
	DBLP-ACM	4	0.977 6	0.756 2	0.908 2	0.221 4	24.38
	DBLP-GoogleScholar	4	0.949 0	0.782 6	0.871 0	0.166 4	19.10
	Walmart-Amazon	5	0.830 2	0.513 8	0.710 3	0.316 4	44.54
	Fodors-Zagats	6	1.0	0.565 2	0.882 2	0.434 8	49.29
JointBERT	Cameras	3	0.832 6	0.690 6	0.761 6	0.142 0	18.64
	Computers	3	0.925 1	0.799 4	0.861 9	0.125 7	14.58
	Shoes	3	0.800 8	0.580 8	0.690 3	0.220 0	31.87
	Watches	3	0.861 3	0.548 3	0.711 8	0.313 0	43.97

表 2 置信度模型预测结果的影响(*F*1-*score*)

模型	数据集	<50%	50%–60%	60%–70%	70%–80%	80%–90%	90%–100%
Ditto	Beer	0.956 5	1	None	None	None	None
	Amazon-Google	0.384 6	0.56	0.565 2	0.730 8	0.8	0.832 6
	DBLP-ACM	0.555 6	0.666 7	0.8	0	0.75	0.977 0
	DBLP-GoogleScholar	0	0	0	0	0	0.974 3
	Walmart-Amazon	0.666 7	0.75	0.812 5	0.777 8	0.769 2	0.693 6
	iTunes-Amazon	0.956 5	1	None	None	None	0
JointBERT	Cameras	0.705 8	0.758 6	0.882 3	0.739 1	0.826 6	0.924 5
	Computers	0	0.5	0.5	0.777 8	0.839 4	0.950 5
	Shoes	None	0.533 3	0.536 6	0.363 6	0.845 1	0.923 1
	Watches	0.593 8	0.9	1	0.8	0.835 8	0.868 1

以图 3 中的 JointBERT 中所给出的重要性评估热力图为例, 其中, 橙色表示向匹配结果正确方向的贡献,

蓝色表示向错误方向推动. 图中“L0|chromebook”和“R0|chromebook”的属性重要性评估显然并不具备单独的解释能力, 具备解释能力的是二者之间的关联. 因此, 如何挖掘、度量并评估预训练语言模型中多头注意力机制的重要性, 是一个十分重要的问题.

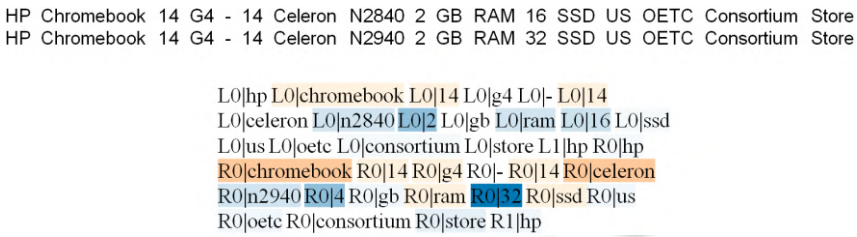


图 3 JointBERT[9]中, 基于 Mojito 框架 LIME 方法的实体对各部分的重要性评估(saliency score)

2.3 关键概念

最后, 我们给出本文的关键概念定义: 反事实解释、属性值关联与近似样本. 本文解释预训练语言模型实体匹配决策的方法主要围绕着这 3 个定义展开.

首先, 我们给出反事实解释的概念. 反事实从已确认的某一历史事实的反面提出疑问或对立的假设, 再从这样的疑问或假设出发, 寻找和收集有关论据. 反事实解释要求优化目标要求生成的反事实解释与原事实数据点相近的同时, 生成的反事实解释能够得到想要的结果[29], 其形式化定义如下.

定义 1(反事实解释). 对于分类模型 f 和样本 x, 样本 x 的反事实解释为 x', 满足:

$$x'=\text{argmin}\ d(x,x');$$

$$f(x')=y',$$

其中, d 为距离度量函数, y'为要求反事实解释能够得到的分类结果.

如表 1 所示, 模型属性换序可以造成分类结果的改变, 因此在 f 为预训练实体匹配模型 M 的场景下, 本文用属性序反事实代指定义 1 中 $x'=x$、但二者输入给模型的属性序不同的情况. 本文中, 我们主要研究匹配错误, 需要反事实修正的情况, 即规定 y'为 x 的分类标注, $f(x)\neq y'$. 有关属性序反事实生成的概念, 我们将在第 3.1 节的定义 5 中进一步介绍.

接着, 我们介绍属性值关联的概念. 实体匹配不同于常规二分类任务之处在于, 其属性值之间的关联对于决策影响较大. 实体对属性值及其子序列之间的关联尤为重要, 其形式化定义如下.

定义 2(属性值关联). 对于模型 M, 给定元组对 $e,e'=\{attr_i,val_i\}_{1\leqslant i\leqslant k},\{attr'_j,val'_j\}_{1\leqslant j\leqslant k}$, 属性值关联定义为三元组:

$$full\_corr(e,e')=\{\langle v_1,v_2,w\rangle\},\ v_1,v_2\in\{\{val_i\}_{1\leqslant i\leqslant k}\cup\{val'_i\}_{1\leqslant i\leqslant k}\}_{subseq},$$

其中, $\{\{val_i\}_{1\leqslant i\leqslant k}\cup\{val'_j\}_{1\leqslant j\leqslant k}\}_{subseq}$ 表示 e, e'所有属性值的子序列集合, 权重 w 用于度量三元组对模型决策的贡献.

显然, 试图衡量单个样本对的全部属性值关联是一个计算代价巨大的任务. 给定模型 M, 其分词策略不同会缩减属性值关联空间的大小, 但仍然为指数级别. 给定一个元组对, 如何得到可解释性最优的属性值关联子集, 是本文要解决的问题.

最后, 我们简要介绍样本近似度的概念. 衡量两个样本近似度有着大量的衡量指标, 如余弦距离、欧式距离、编辑距离等, 基于近似度, 我们可以给出近似样本的定义如下.

定义 3(近似样本). 给定样本 $x\in D$, 定义其上的一个近似度函数 $sim(x,x')$, 则 x 的近似样本定义为:

$$\{x'|sim(x,x')\leqslant\delta\}.$$

即二者距离小于阈值δ.

上述定义中, 近似度的度量是近似样本是否满足可解释需求的关键, 本文主要围绕 top-k 的情况展开讨

论. 搜索历史数据集或训练集中的近似样本, 对于模糊解释的提示、模型性能的增强起着重要的作用.

基于以上定义, 给出本文针对预训练语言模型特点定义的模型解释问题如下.

定义 4(兼顾行列的实体匹配解释问题). 给定一个实体匹配模型 M, 其在实体匹配数据集 D 上的错分集合 E_0, 兼顾行列的实体匹配解释问题解决以下 3 个目标.

① 对于 E_0, 生成属性序反事实 E_0'.

② 对于$\forall e_w \in D$, 计算可解释性最优的属性值关联子集 $e_w.corr \subseteq full\_corr(e_w)$.

③ 对于$\forall e_w \in D$, 搜索有效增强 $e_w.corr$ 和 M 的近似样本集合$\{x'|sim(x,x') \leqslant \delta\}$.

上述 3 个目标中, 目标①可以视作数据集 D 的可解释性最优的列排序, 目标②可以视作 D 上可解释性最优的列组合, 目标③可以视作 D 的可解释性最优的行组合. 而由于关系数据的交换不变性, D 的行排列并不具有意义. 因此, 定义 4 可以看作在反事实、关联分析、解释提示、模型增强的需求下, 在预训练语言模型实体匹配的特定样本集上, 搜索数据集 D 的行列可解释性最优排列组合的过程.

在第 3 节中, 我们将逐一介绍上述问题中 3 个目标的求解方法.

3 面向预训练语言模型实体匹配的可解释性分析方法

基于预训练语言模型的实体匹配技术对属性序敏感、低置信度预测结果不稳定、样本关联难以解释, 给模型应用部署造成了可信度缺失. 针对以上挑战, 本文使用输入属性序反事实以体现序列敏感性对决策的影响, 通过关联分析以解释样本内部的属性影响机制、基于一致性评分搜索近似样本并解决模型不稳定的问题. 通过收集模型的序列化顺序、注意力权重、元组对表征的多层次兼顾行列的信息, 能够提供多角度、针对性、易理解的模型决策解释与可信赖的模型决策增强. 本文提出的兼顾行列的预训练语言模型实体匹配解释方法由以下 3 部分组成.

- 列级属性序反事实生成. 从输入和样本层面, 首先根据历史数据集中的属性序和模型效果, 结合元学习策略与属性相似度来生成错分样本列级属性序反事实. 反事实生成搜索对样本可施加的最小属性值改变, 以实现模型决策的修改. 然而, 尚未有可解释性方法从反事实这一预训练语言模型的特殊角度探究过属性序列化的语言模型会受到属性序的影响. 因此, 本文的反事实方法改变属性的顺序, 可以达到完全不改变数据点即更正模型决策错误的目的.
- 列级属性关联性分析. 从训练和属性层面, 采用基于多头注意力机制中样本的注意力权重进行筛选, 将注意力权重分布中的异常值对应的词语对进行筛选、评分和热力图可视化. 挖掘样本各部分对于决策的影响是重要的, 但现有的属性值重要性衡量不适用于关注实体对关联的预训练语言模型实体匹配. 为了合理解释模型各层捕捉上下文关系的能力如何影响实体匹配决策, 本文使用列级属性关联性对样本进行可解释度评分, 这一评分在错分样本的近邻解释法、模型的代表性样本子集解释法等可解释方法中起着关键的作用.
- 行级解释与决策增强. 从表征和数据集层面, 将[CLS]项的嵌入表征向量用于模型预测的低置信度样本 k 近邻搜索, 实现模型“不擅长处理”部分样本的行级预测结果可解释增强. 预训练语言模型虽然在大部分场景下有着优秀的效果, 但在部分样本上还是呈现了预测置信度下降, 其预测效果较差. 因此, 我们基于精调后的[CLS]嵌入以在表征空间中表示实体对, 结合 k 近邻搜索结果来增强这一部分低置信度样本的弱预测.

3.1 基于元学习的属性序反事实生成

本节介绍一种对数据集生成列级属性序反事实的方法. 该方法从历史数据中学习排序评价经验, 为新数据集的错分样本子集生成属性序, 该属性序下的数据输入给模型后, 能够将大量错分样本正确分类.

如定义1所介绍, 属性序搜索空间大小为全排列的指数级, 且现有的反事实方法难以在该空间上使用. 因此, 为了实现计算有效的属性序反事实生成, 本节提出一种基于元学习的属性序反事实生成方法(MLARC). 其思想是: 从历史数据中学习排序的经验, 搜索属性-偏序图上的最大权拓扑排序, 以高效得到接近最优解的

反事实属性序.

首先, 我们对于属性序反事实生成问题定义如下.

定义 5(属性序反事实生成问题). 对于一个给定的实体匹配模型 M、并给定数据集 D 上分类错误的元组对集合 $E_0=\{e,e'\}=\{\{attr_i,val_i\}_{1\leqslant i\leqslant k},\{attr'_j,val'_j\}_{1\leqslant j\leqslant k}\}$, 属性序反事实生成问题的目标定义为搜索一个元组对属性集合 $A=\{\{attr_i\}_{1\leqslant i\leqslant k},\{attr'_j\}_{1\leqslant j\leqslant k}\}$ 的排列 $Rank_A$, 满足:

$$Rank_A=\max_{Rank_A\in P(|V|,|V|)} fidelity(Rank_A)=\max_{Rank_A\in P(|V|,|V|)}\frac{|\{e,e'\mid M(Rank_A,e,e')=y\}|}{|E_0|},$$

其中, $M(Rank_A,E_0)$为反事实预测结果, $fidelity(Rank_A)$指错分样本分类正确的比例.

根据反事实的一般形式[30], 我们得到的反事实为: “模型 M 对于元组 e, e'的预测结果是〈原实体匹配结果〉. 如果特征集合的排列是不同的〈新排列而不是原排列〉, 那么预测结果就会变成〈反事实实体匹配结果〉”, 从而在不改变属性值、仅改变属性序的情况下, 使得错分样本跨越决策边界, 解释了属性序关系对预训练语言模型 M 决策的影响.

本文中, M 为 BERT 类预训练模型, 我们需要考察属性集合的所有排列, 才能得到性能 $fidelity(Rank_A)$最高的精确结果 $Rank_A^*$, 其时间复杂度为 O(k!). 该计算代价在实体识别模型解释的场景下, 显然是不可接受的.

- MLARC 算法

元学习技术使用相似机器学习任务的经验加速新任务学习过程中的模型选择、参数调优等过程[23]. 基于元学习思想, 我们整合历史数据中不同排列对于结果的影响, 将排列分解为评分不同的偏序, 最大化总偏序评分, 以生成近似最优的属性序反事实. 具体来说, 我们假设同一任务上有着更高经验评分的候选解有更高的概率能够在相似任务上表现优秀, 即使用相似任务的经验和元特征大幅度减小搜索反事实的时间代价, 求近似解 $Rank'_A$.

值得注意的是, 上述“偏序”不是严格满足传递性的偏序. 由于考虑将单个属性排列断成多个偏序, 这些偏序之间可能存在冲突.

本节构造了一种离线-在线的方法 MLARC, 其中, 离线阶段, 利用历史经验构造元数据库; 在线阶段, 基于属性-偏序有向图和最大权拓扑排序的方法来消解冲突.

I. 离线阶段

对于每个历史数据集 $d'\in D$, 计算所有候选排列 $Rank_A$ 的性能 $fidelity(Rank_A)$, 并提取这些数据集的元特征向量 V'_d, 形成大量 $tri(Rank_A)=\langle V_d,Rank_A,fidelity(Rank_A)\rangle$经验三元组. 构建经验三元组库需要约 O($|D'|\cdot k!$)次预测, 虽然数据集属性数量普遍较少、预测操作时间代价小且此操作在离线阶段进行, 这仍是一笔较大的时间开销. 可以通过抽样, 在容忍损失一定效果的情况下减少计算.

II. 在线阶段

首先, 需要构建属性-偏序图. 对于离线 $tri(Rank_A)=\langle V_d,Rank_A,fidelity(Rank_A)\rangle$经验三元组, 首先需要将历史数据集 D'的属性 $\hat{A}$ 上的经验迁移到新数据集 D 中的属性 A 上, 上述属性经验迁移任务可以使用数据集成中经典的模式对齐(schema alignment)算法来解决. 为了兼顾经验迁移的鲁棒性和有效性, 本文采用在关系数据上普适性良好、融合属性值信息、效果优秀的 EMBDI 算法[31]计算嵌入向量, 根据其余弦值实现模式对齐并计算属性相似度 $sim(A,\hat{A})$. 该算法的时间代价较高, 用户可以通过专家和领域知识, 减少 EMBDI 的代表样本数量到常数级别, 或选用编辑距离、正则表达式相似度、属性值集合相似度、属性值分布相似度等朴素的模式对齐方法进行属性匹配和相似度计算.

上述经验三元组中, $Rank_A$ 不是最小的偏序单位, 三元组 $Rank_A$ 可以拆分 $C_{|A|}^2$ 条偏序, 每条偏序表示经验三元组库对两个属性序关系的评分, 则拆分后的偏序评分可以用属性-偏序图来表示.

定义 6(属性-偏序图). 数据集 D 的属性-偏序图是一个带有边权的有向图 $G=(V,E)$, 其中, V 为 d 的所有属性集合 A, 其起点为 v_i、终点为 v_j 的有向边 $e_{ij}=\langle v_i,v_j\rangle$上的边权为:

$$w(e_{ij})=\sum_{d'\in D'}\mathbf{1}_{\hat{A}_i>\hat{A}_j}\times\cos(V_{d'},V_d)\times sim(\hat{A}_i,A_i)\times sim(\hat{A}_j,A_j),$$

其中, $\mathbf{1}_{\hat{A}_i>\hat{A}_j}$ 代表历史数据 $d'\in D'$中, $\hat{A}_i$ 是否出现在 $\hat{A}_j$ 前的 0-1 指示函数.

我们将偏序视作有向边, 其权重视作重要性, 则最优性能的反事实可以近似视作最多偏序“投票”的属性序. 因此, 在上述属性-偏序图上求解最大权拓扑排序问题, 即可得到生成属性序反事实的近似解.

定义 7(最大权拓扑排序问题). 对于一个属性-偏序图 $G=(V,E)$, 最大权拓扑排序指求解其某一子图 $G'=(V,E')$的拓扑排序 $Rank_A$, 使得:

$$Rank_A=\underset{Rank_A\in P(|V|,|V|)}{\arg\max}\sum_{e_{ij}\in Rank_A}w(e_{ij}),$$

其中, $P(|V|,|V|)$表示的全排列, $e_{ij}\in Rank_A$ 表示 $e_{ij}>0$ 且 $Rank_A$ 中第 i 个属性在第 j 个属性前.

上述有向图可能存在有向环, 即偏序集的冲突. 为了解决这一问题, 我们需要首先破除有向图上的圈, 再求解拓扑排序.

算法 1 可以求得最大权拓扑排序的近似解, 具体来说, 首先使用 Kahn 算法(第 1 行)判断有向图无有向环, 若存在有向环, 使用 Chu-Liu Edmonds 算法(第 5 行)求解以入度最小顶点为根的有向图上最大树形图, 以破除有向环, 再使用 Kahn 算法输出最大树形图的拓扑排序即可.

算法 1. 基于元学习的属性序反事实生成算法.

输入: 属性-偏序图 $G=(V,E)$.

输出: 最大权拓扑排序 $Rank_A$.

1. **if** *len*(*kahn*(*G*))=*V*: //根据 kahn 算法输出长度判断有无环
2. **return** *kahn*(*G*) //返回上一步的拓扑排序
3. **else**:
4. *root*←*minInDegree*(*V*) //将入度最小的点作为 *root*
5. *T*←*Chu-LiuEdmonds*(*G*) //求最大树形图
6. **return** *kahn*(*T*) //返回最大树形图的拓扑排序

时间代价方面, Chu-Liu Edmonds 算法的时间复杂度为 O(Elog$V$), Kahn 算法的时间复杂度为 O(V+E), 因此总时间复杂度为 O(Elog$V$).

3.2 基于注意力权重的属性值关联性

本节介绍如何对于一个给定的实体对进行列级属性关联解释, 通过计算、筛选并可视化注意力机制权重, 定位影响模型对该实体对分类决策的关键属性值对, 其结果可用对于任何一个用户关注的实体对(如第 3.1 节的错分实体对、第 3.3 节的低置信度实体对、用户自定义的高敏感实体对等)提供决策解释分析.

- 属性重要性的不足

在模型决策中, 一个重要的影响因素是属性关联, 在实体匹配的场景下尤为如此. 然而, 现有的主流可解释性方法往往关注属性值重要性评估, 这类方法以 LIME 算法[11]为代表, 用线性分类器等简单模型去拟合复杂分类器在样本附近的局部决策. 以图 4 中的实体匹配为例(其中, 黄色部分表示模型参与解释的部分), 算法在原始实体对的周围进行扰动, 得到一些采样样本后分类, 同时根据采样样本与原样本的距离赋予分类器决策的权重, 以衡量属性值对决策的正负向贡献程度, 即属性重要性.

然而, 仅衡量重要性并不适用于实体匹配的场景. 图 4 中, 我们显然还需要关注同一位置的两个样本之间的属性值关联.

以易于理解的主键一致性为例, 图 4 中, “L0|chromebook”与“R0|chromebook”之间的一致性对“匹配”决策有着较大的正向贡献; 图 5 中, “Double Dragon Imperial”与“Scuttlebutt Mateo Loco”之间的不一致性对“不匹配”决策有着较大的正向贡献. 而 LIME 的结果对这些属性值单独评价, 无法体现这样的关联性.

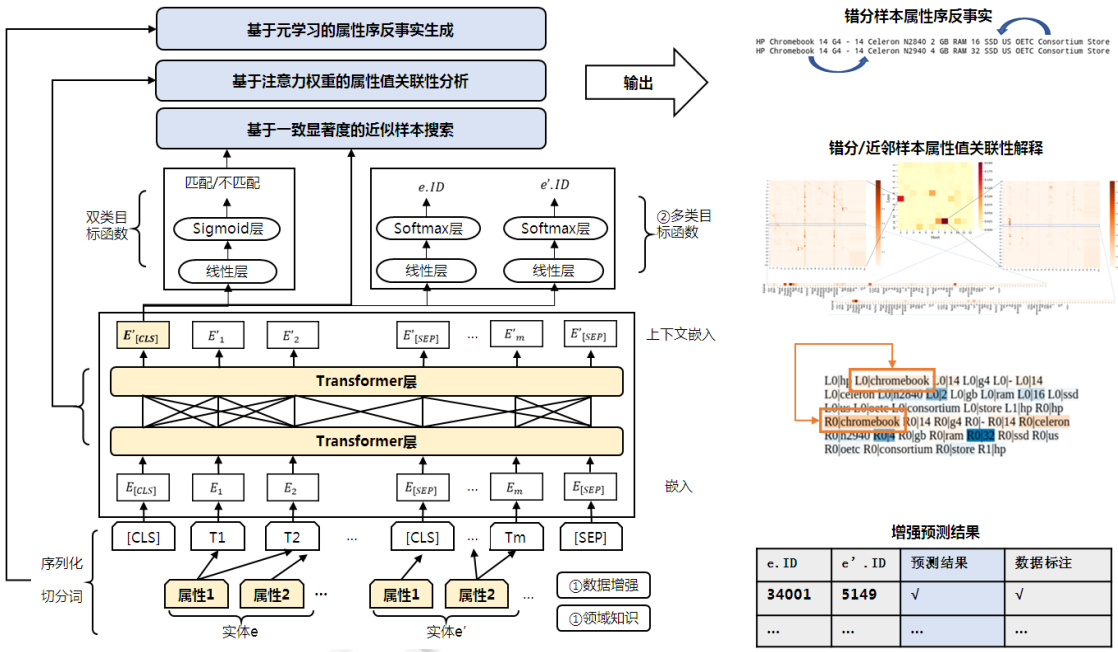


图 4　兼顾行列的预训练语言模型实体匹配解释方法

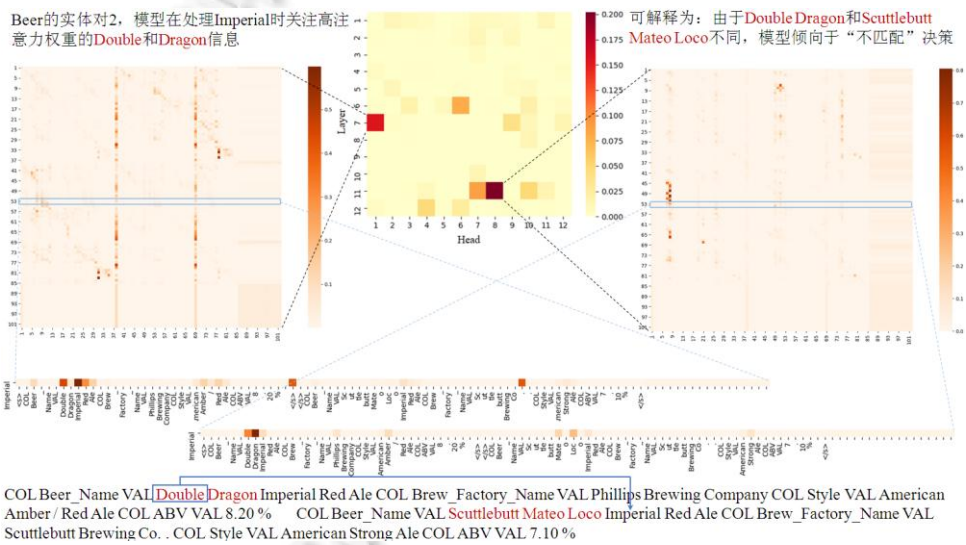


图 5　基于注意力机制的关联分析一例

- 注意力机制与属性关联性

上文所述的属性关联性，是预训练语言模型在实体匹配上取得成功的关键. 这类模型的核心模块正是利用多层 Transformer 中的多头注意力机制，将上下文相关的信息以一定的关联权重融合入每个单词和句子的表征中. 因此，理解注意力机制就理解了深度模型表征的关键环节. 注意力机制的可解释性已经在自然语言处理中得到了广泛应用[32]，体现为高注意力权重词语与数据集中专家预标注关联词语的对应关系. 我们参考这一方法，将高注意力机制的样本属性对作为样本的强关联解释. 首先，我们对注意力机制简要介绍.

注意力机制的一般形式由 *Query*(目标字或待生成的标注的词)、*Value*(上下文各字的原始 *Value* 表示)、*Key*(上下文各字的 *Key* 向量表示)组成，通过计算 *Query* 和 *Key* 的相似度对 *Value* 加成，生成目标字的 *Attention* 值:

$$Attention(Query,Source)=\sum_i Attention(Query,Key_i)\times Value\ .$$

其中, $Source=\langle Key,Value\rangle$, 得到目标字增强的语义表示. 多头自注意力机制将每个字的多个增强语义向量线性组合, 最终获得一个与原始字向量长度相等的增强语义向量. 多头的使用是因为不仅词具有多义性, 有些句子在不同场景下表达的意思也不一样, 可以看作是对文本中每个字分别增强其语义向量表示的黑盒.

模型 M 决策样本 e 的所有权重可以视作四维张量:

$$T[h,l,s,s],$$

其中, h 表示注意力头的数量, l 表示 Transformer 的层数, s 表示样本序列化后的语句长度.

- 属性值关联筛选

下面我们介绍如何筛选高权重的词语对[32]作为模型解释. 由于实体匹配数据集缺乏文本数据集中的关联词作为评价注意力权重的标准, 因此针对某一实体对 x_i 直接计算所有注意力头对应的权重会得到 $h\times l\times s\times s$ 的 ($word_1$,$word_2$,*attention weight*)三元组, 造成单个样本的信息冗余. 但若直接筛选最高权重很可能会得到可解释性较弱的关联, 造成属性值关联可解释性弱的问题. 针对这一问题, 一种方法是增加判据严格筛选词语对, 共计 3 个筛选判据包括:

- J_1(跨样本属性): $word_1$ 与 $word_2$ 属于两个样本, 即在[SEP]的两侧.
- J_2(强语义关联): $word_1$ 与 $word_2$ 为两个单词, 而不是数字、占位符等.
- J_3(注意力权重): 注意力权重 *attention weight* 在该样本所有候选关联中为 top-k.

上述 3 个条件中, J_1 避免了同一样本内部属性关联的出现, J_2 避免了弱语义关联的出现, J_3 保证了关联的相对显著性. 显然, 前两个筛选条件尚不足以完全解决样本模糊性的问题, 我们将在下一节中介绍一种基于近邻样本搜索的解释提示方法, 用训练集中专家预先标注解释信息的样本辅助属性值关联解释.

作为可视化, 我们连接满足 3 个条件的关联词语对, 并给出各层、各注意力头、样本内部的显著性热力图. 以图 5 中的实体对为例, 第 7 层和第 11 层的多头注意力机制在计算 Imperial 表征时关注了 Double、Dragon 两词, 这一部分与 Scuttlebutt Mateo Loco 不同, 使模型倾向于“不匹配”这一正确决策.

3.3 基于一致显著度的近似样本搜索

本节介绍一种针对单个样本点, 融合模型表征和关联显著程度的行级近似样本搜索方法. 相比于传统的 k-近邻样本搜索, 该方法利用预训练语言模型的注意力关联机制和样本对嵌入, 搜索到的结果能够更好地补足实体对粒度上决策结果可解释性弱和在数据集粒度上模型效果不稳定的问题.

第 3.2 节中的关联分析方法虽然能够提供更全面的解释信息, 但相比于重要性度量不够直观, 在部分数据上需要辅以所捕捉关联属性值的专家理解方可为用户所用. 而在大批量数据集上, 可能出现大量的错分样本或用户关注样本, 对这些样本逐个添加专家理解, 代价高、信息繁冗、直观程度不足. 因此, 需要设计支持数据集内部样本间近似度量, 对于实体对粒度的近似样本解释提示等有着重要的作用.

- 近似样本搜索

现有模型无关的解释方法中, 近似度在样本上直接计算, 没有很好地融合模型表征信息的近似性和样本解释部分的近似性. 然而, 前者是模型决策的直接影响因素, 后者是样本解释信息的参与部分, 因此传统的近似度度量搜索到的样本集与待解释样本点的关联较弱. 为了有效地解决这个问题, 我们利用模型生成样本表征, 给出融合一致性得分、显著程度的一致显著度定义, 并给出基于这一近似度量的模型解释(consistent-saliency k-nearest neighbor, CSKNN)和预测增强(k-nearest neighbor enhancing, KNNE)方法.

定义 8(一致显著度). 对于样本 x, x_i, 二者的一致显著度定义为:

$$cosistent\_saliency(x,x_i)=G\{d(E'_{[CLS]}(x_i),E'_{[CLS]}(x))\}\times\frac{maxweight-\mu}{\sigma},$$

其中, $E'_{[CLS]}(x_i)$ 为预训练后 CLS 嵌入向量 $d(E'_{[CLS]}(x_i),E'_{[CLS]}(x))$类似于文献[9]中的一致性得分, 表示 x_i 与样本 x 之间的欧氏距离; G 为标准正态分布函数; μ和σ表示张量 T 中所有元素的均值和方差, $\frac{maxweight-\mu}{\sigma}$ 用于度

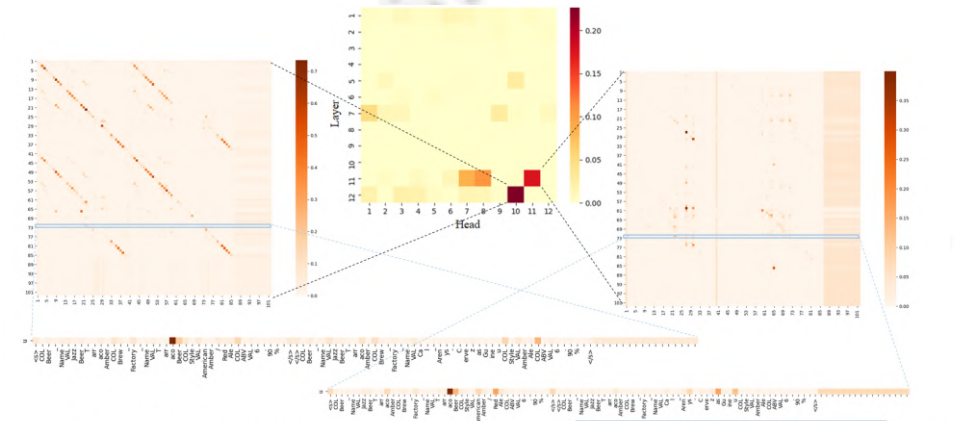
量样本 x_i 中最大注意力权重偏离均值的程度，这一程度可以用来评估关联性解释的显著度.

一致显著度的本质思想是加权的 k 近邻搜索，即在距离度量中引入关联对应注意力矩阵最大权重的离群度，即认为解释性更强的样本对于模型“不确定”的样本分类决策的辅助提示作用更大. 本文从下述两个样本近似度度量的角度，引入一致显著度.

I. 解释提示

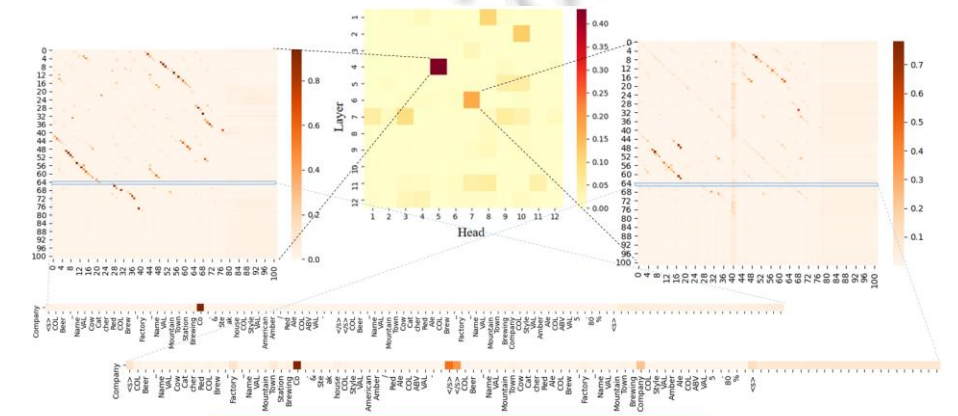
一方面，在本文的错分样本近似样本搜索解释方法 CSKNN 中，我们引入一致显著度用来度量搜索近邻样本和关联性解释与错分样本之间的近似度. 最近邻解释方法将样本归一化后的距离被视为与测试点相关联的权重，被用于检索训练集中与测试点最近的 k 个样例[10]. 由于我们的目标是生成样本解释的启发性提示，直接将基于归一化距离的 k 近邻搜索结果展示给用户，会导致缺少模型相关表征信息和决策解释信息，造成样本解释的模糊性.

一致显著度可以避免样本解释的模糊性，同时保证近似样本提示的近似性. 对于测试集中的某一错分样本，CSKNN 将所有训练集中的样本按照一致显著度排序，基于排序结果，在训练集中搜索一致显著度最高的 k 个样本，将这部分带有专家解释信息标注的样本作为“提示样本集”供给用户参考，作为待解释样本 x 决策结果的解释提示. 这一过程在算法 2 中表示为 *CSKNN*(·)函数. 在图 6、图 7 中，我们展示了一个解释的例子.



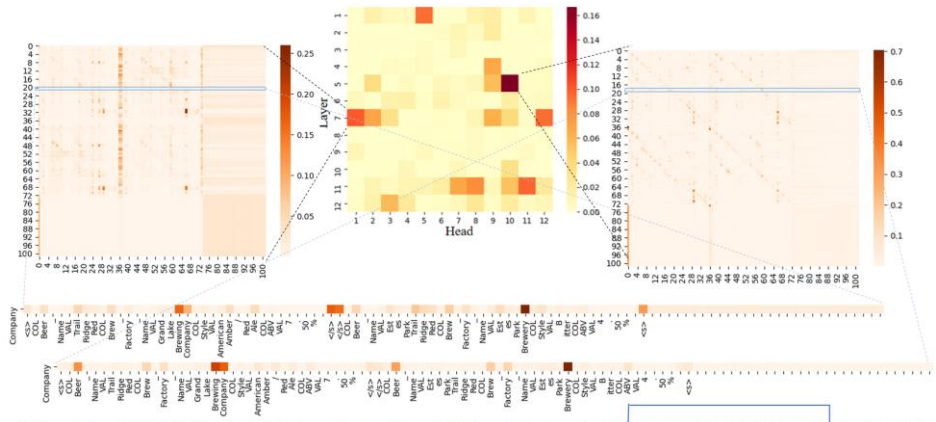
COL Beer_Name VAL Jazz Beer Tarraco Amber COL Brew_Factory_Name VAL Tarraco Beer COL Style VAL American Amber / Red Ale COL ABV VAL 6.90 % COL Beer_Name VAL Jazz Beer Tarraco Amber COL Brew_Factory_Name VAL Ca l_Arenys - Cervezas Guineu COL Style VAL Amber Ale COL ABV VAL 6.90 %

图 6 错分样本案例 x

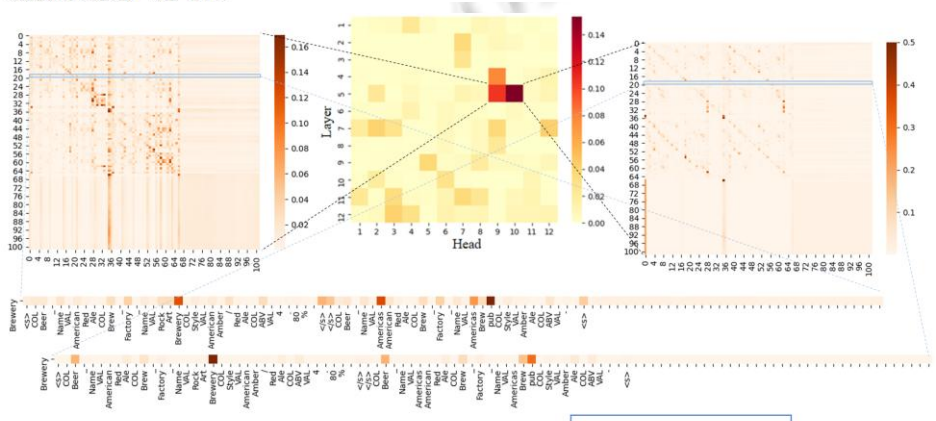


COL Beer_Name VAL Cow Catcher Red COL Brew_Factory_Name VAL Mountain Town Station Brewing Co. & Steakhouse COL Style VAL American Amber / Red Ale COL ABV VAL - COL Beer_Name VAL Mountain Town Cow Catcher Red Ale COL Brew_Factory_Name VAL Mountain Town Brewing Company COL Style VAL Amber Ale COL ABV VAL 5.80 %

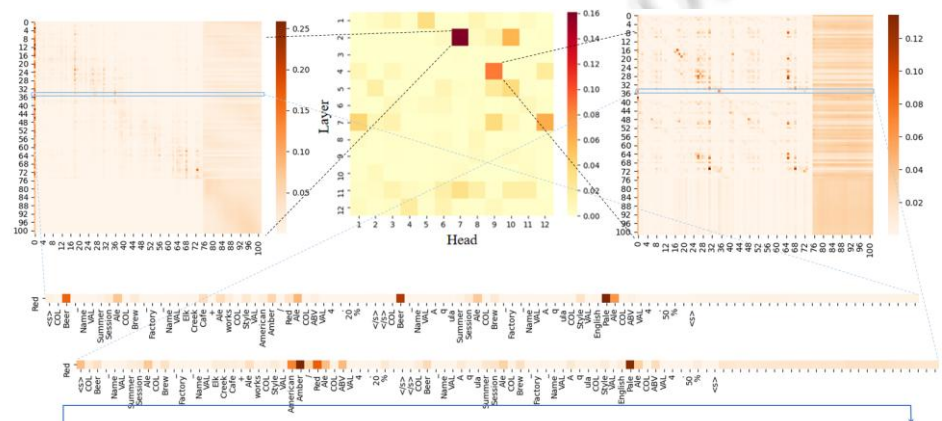
图 7 错分样本在训练集中搜索到的 4 个近邻样本 x_1, x_2, x_3, x_4



COL Beer_Name VAL Trail Ridge Red COL Brew_Factory_Name VAL Grand Lake Brewing Company COL Style VAL American Amber / Red Ale COL ABV VAL 7.50 % COL Beer_Name VAL Estes Park Trail Ridge Red COL Brew_Factory_Name VAL Estes Park Brewery COL Style VAL Bitter COL ABV VAL 4.50 %



COL Beer_Name VAL American Red Ale COL Brew_Factory_Name VAL Rock Art Brewery COL Style VAL American Amber / Red Ale COL ABV VAL 4.80 % COL Beer_Name VAL Americas American Red Ale COL Brew_Factory_Name VAL Americas Brewpub COL Style VAL Amber Ale COL ABV VAL -



COL Beer_Name VAL Summer Session Ale COL Brew_Factory_Name VAL Elk Creek Cafe + Aleworks COL Style VAL American Amber / Red Ale COL ABV VAL 4.20 % COL Beer_Name VAL Aqula Summer Session Ale COL Brew_Factory_Name VAL Aqula COL Style VAL English Pale Ale COL ABV VAL 4.50 %

图 7 错分样本在训练集中搜索到的 4 个近邻样本 x_1, x_2, x_3, x_4 (续)

II. 模型增强

另一方面，在针对低置信度样本的可解释模型增强 KNNE 中，我们引入一致近似度，可以用于将表征信

息引入决策过程. 如表 2 所示, 预训练语言模型低置信度的预测结果往往呈现出较低的性能. 因此, 本文将近似样本搜索的应用范围从单纯的错分样本拓展到低置信度样本上, 作为预训练语言模型实体匹配的决策补充, 以实现模型效果的可解释增强.

一致显著度中的显著度可以作为分类器决策时的权重评分, 一致性评分可以作为对比于归一化距离的样本更精细化的表征. 本文使用的 k 近邻预测增强分类器将模型 M 的预测结果与 $E'_{[CLS]}(x_i)$ 表征空间中以欧式距离为度量的 k 近邻分类器结合, 其分类决策函数如下:

$$y'=\begin{cases}\arg\max\limits_{c_j}\sum_{x_i\in N_k(x')} consitent\_saliency(x,x_i)\times I(y_i=c_j), & conf(M(x'))<\delta\\ M(x'), & conf(M(x'))\geqslant\delta\end{cases},$$

其中, $N_k(x)$为样本 x 的 k 个一致显著度最高样本组成的邻域. 将一致显著度作为权重求取邻域内样本标签的加权平均结果作为目标样本的标签 y', 在算法 2 中, 这一操作用 $KNNE(\cdot)$函数表示. 实验结果表明, KNNE 可以在较短的时间代价内有效提升预训练语言模型 M 的效果, 且受近邻数 k 和置信度阈值δ的影响较小.

算法 2. 一致显著度-k 近邻增强算法.

输入: 模型、训练集 D_{train}, 测试集 D_{test}, 用户给出的查询 q_{user}, 置信度阈值δ.

输出: e_w 关联分析结果 $e_w.corr$, e_w 的近似样本集合 $e_w.csknn$, 低置信度样本的预测增强结果 E'_{low}

1. **for** d in T_D:
2. $d.corr \leftarrow Filter(J_1 \cap J_2 \cap J_3, d)$ //根据 $J_1 \cap J_2 \cap J_3$ 筛选 $d.corr$, 构建候选解释样本集合
3. $D_O \leftarrow q_{user}(D_{test})$ //根据查询从测试集中选择待解释样本集
4. $D_Q.csknn \leftarrow CSKNN(D_Q, T_D.corr, E'_{[CLS]})$ //搜索显著度为权重、 $E'_{[CLS]}$ 为表征的 e_w 加权 k 近邻
5. $E'_{low} \leftarrow KNNE(e_w.corr, e_w.csknn)$ //使用加权 k 近邻算法进行分类增强
6. **return** $e_w.corr$, $e_w.csknn$, E'_{low}

算法 2 简要阐明了第 3.2 节和第 3.3 节的流程, 包括关联性分析可视化(第 1 行)、CSKNN 搜索近似样本集(第 2 行)以及 KNNE 增强模型(第 3 行). 可以看到, KNNE 利用了 CSKNN 的搜索结果; 而 CSKNN 给关联分析可视化提供了解释提示, 也给 KNNE 提供了近邻样本集. 时间代价方面, 上述过程主要时间代价仅为一次加权 KNN 搜索, 其余筛选和分类的时间代价较低, 满足了模型解释的低延时需求.

4 实验分析

4.1 实验数据

为了衡量预训练语言模型处理含有复杂文本实体的能力, 本文选择包含部分文本的多个属性公开实体匹配数据集上进行实验, 包括 Beer、Amazon-Google、DBLP-ACM、DBLP-GoogleScholar、Walmart-Amazon、Fodors-Zagats、Cameras、Computers、Shoes 和 Watches. 表 3 给出了数据集所对应的详细信息.

表 3 实验数据集

模型	数据集	训练集大小	测试集大小	错误实例数
Ditto	Amazon-Google	6 874	2 293	140
	Beer	268	91	3
	DBLP-ACM	7 417	2 473	26
	DBLP-GoogleScholar	17 223	5 742	96
	iTunes-Amazon	321	109	3
	Walmart-Amazon	6 144	2 049	99
	Fordors-Zagats	846	100	110
JointBERT	Cameras	3 154	1 051	67
	Computers	4 858	1 618	62
	Shoes	3 484	1 161	128
	Watches	3 848	1 283	91

数据集中, Ditto 使用的 7 个数据集来自 ER Benchmark 数据集和 Magellan 数据集, 这些数据集包含产品、

出版文献、商业等领域, 在实体匹配文献中应用广泛. 每个数据集包括两个相同模式的关系表的候选实体对, 实体对来自实体分块结果中抽样并由人工标注, 属性个数从 1 到 8 个不等.

JointBERT 使用的 4 个数据集来自用于大规模产品匹配的 WDC 产品数据语料库(WDC LSPC)[33]. 产品数据从 schema.org 电子商店收集, 记录了数据领域、产品描述和其 HTML 页面中的产品 ID. 本文使用 computers, camera, shoes 和 watches 这 4 个类别的训练、验证以及测试集. 测试集中包含的所有实体, 都在训练集中用不同的实体描述出现过.

为了保证数据集与模型的适配性, 避免出现基准模型在特定数据集上表现较差的情况, 我们在 Ditto 和 JointBERT 上使用的所有数据集均为两个模型使用的原始的数据集.

4.2 实验方法

- 评价指标

在本文中, 我们采用 *F*1 增量、保真度(fidelity)、近似比(approximation ratio, APM)和时间来评价反事实的性能. 并使用常用的评价指标准确率(precision)、召回率(recall)和 *F*1-*score* 和时间来评估模型增强的性能. 上述指标计算如下:

$$fidelity(Rank_A)=\frac{|\{e,e'\,|\,M(Rank_A,e,e')=y\}|}{|M(Rank_A,e,e')|},\ APM(Rank_A)=\frac{fidelity(Rank_A)}{fidelity(Rank_A^*)},$$

$$Precision=\frac{TP}{TP+FP},\ Recall=\frac{TP}{TP+FN},\ F1\text{-}score=\frac{2\times precision\times recall}{precision+recall}.$$

值得注意的是, 可解释性是一个较新的研究领域, 对于其评价指标尚未达成共识. 从样本决策可解释性评价的角度, 常用的评价指标包括所生成反事实的保真度和距离、以 LIME 为代表的模型局部拟合残差等. 由于排序空间的距离度量困难, 本文选择保真度和近似比用于评价样本决策的可解释性; 从样本关联解释的角度, 尚未有成熟的可解释性评价指标被提出, 现有的可解释性分析工作主要关注在关联约束层面, 由于关联样本在实体匹配数据集上的标注缺失, 我们使用案例研究来对其进行评价; 从模型可解释性评价的角度, 线性模型、决策树、*k*-近邻搜索、朴素贝叶斯, 被认为是具有自可解释性的模型[34], 而这一角度显然无法通过指标进行评价. 因此, 我们使用加权 *k*-近邻搜索局部替代神经网络模型, 提升了模型的自可解释性, 并使用准确率、召回率和 *F*1-*score* 评价局部替代后模型的效果增强.

- 基线模型

Ditto 和 JointBERT 模型为本文的基线模型, 介绍见第 2.1 节.

- 实验方法

在实验中, 模型由神经网络构成, 实现基于 Python 以及 Pytorch 框架, 硬件平台为 2×Tesla M40 计算卡. 在实验结果中汇报的值是 5 次随机拆分训练集、测试集和验证集的实验验证后的平均值. 其中, 每次验证的结果是在最优的参数配置下模型收敛时的平均性能.

4.3 实验结果与分析

为了评估基于变分自编码器的异构缺陷预测方法的有效性, 本节通过实验研究了以下问题.

- 问题 1: MLARC 生成的反事实效果如何?
- 问题 2: CSKNN 搜索的近邻样本能否得到与错分样本近似的关联关系?
- 问题 3: KNNE 提升模型的效果和时间受参数的影响如何?

4.3.1 实验 1: MLARC 生成的反事实的效果

为了验证这个问题的结果, 我们在 Ditto 这一使用属性换序数据增强的预训练语言模型方法上开展了实验. 首先, 我们对于 6 个数据集提取元特征, 包括数值属性的数量、类别属性的数量、数值属性的比例、属性的总数、数据集中的元组数、类别属性中类别数最少的类别数量、类别属性中类别数最少的类别信息熵、类别属性中类别属性最少的单个类别的最大比例、类别属性中类别属性最少的单个类别的最小比例、数值属性

的最小平均值、数值属性的最小方差.

- 性能测试

我们测试了反事实带来的 *F*1-*score* 提升, 即 MLARC 得到的反事实排序中, 将跨越决策边界的样本对按照样本对类别计作真阳性样本对和真阴性样本对后, 重新计算 *F*1-*score* 和 *F*1-*score* 的增量. 注意: 反事实属性序的生成仅影响模型的预测, 预训练和精调是不变的.

从表 4 中可以看出, 对于错分样本对集合反事实换序后重新预测, 所提升 *F*1-*score* 约为 5.4%, 其中表现最为突出的是 DBLP-ACM 数据集, 仅仅交换了 *A*3 和 *A*4 的顺序, 就得到了 14.43%的 *F*1-*score* 效果提升. 值得注意的是, 表 4 中第 3 列的反事实 *F*1-*score* 是一种特殊的度量, 其目的是在体现反事实效果的同时, 兼顾错分样本在整个样本中的比例.

表 4 MLARC 生成反事实的 *F*1-*score* 提升、最优排序与最优排序权重和

数据集	原 *F*1-*score*	反事实 *F*1-*score*	Δ*F*1	反事实排序
Beer	0.823 5	0.848 5	0.025 0	*A*1-*A*3-*A*2-*A*4
Amazon-Google	0.681 5	0.744 3	0.062 8	*A*2-*A*1-*A*3
DBLP-ACM	0.827 1	0.971 4	0.144 3	*A*1-*A*2-*A*4-*A*3
DBLP-GoogleScholar	0.880 0	0.949 0	0.069 0	*A*2-*A*3-*A*4-*A*1
Walmart-Amazon	0.696 3	0.697 2	0.000 9	*A*1-*A*2-*A*3-*A*5-*A*4
Fodors-Zagats	0.977 8	1.0	0.022 2	*A*1-*A*6-*A*5-*A*4-*A*2-*A*3

反事实的常用评价指标包括样本距离和保真度, 但是本文的样本距离为 0, 这也导致现有反事实生成方法均不符合本文问题的定义. 因此, 这些方法不适合与 MLARC 对比, 我们需要设计一种对比方法.

- 对比方法

我们将其与朴素遍历(simple traversal, ST)得到的最优解对比. 具体来说, 朴素遍历策略选择元数据构建过程中 *F*1-*score* 最高的属性序作为反事实, 其结果是全局最优解. 在表 5 中, 我们对比了 MLARC 和 ST 的保真度和近似比.

表 5 MLARC 生成反事实的保真度与近似比

数据集	MLARC 保真度	ST 保真度	近似比
Beer	0.4	0.4	1.0
Amazon-Google	0.152	0.176	0.863 6
DBLP-ACM	0.088 2	0.470 6	0.187 4
DBLP-GoogleScholar	0.462 3	0.509 4	0.907 5
Walmart-Amazon	0.102 8	0.607 5	0.169 2
Fodors-Zagats	1.0	1.0	1.0

- 保真度测试

需要强调的是, 两种方法保真度虽然显著低于现有反事实方法, 但其原因是我们严苛地限制了其搜索空间: 仅允许改变属性序不允许改变属性值. ST 表明了遍历本文搜索空间的情况下反事实的最高保真度上限. 表 5 的近似比说明, MLARC 方法在大幅度降低搜索时间的情况下, 得到了平均为上限 68.8%的保真度, 是一种计算有效的属性序反事实计算方法.

4.3.2 实验 2: CSKNN 错分样本近邻关联性的案例研究

CSKNN 方法是实体匹配领域少有的关注属性关联度的模型解释方法, 由于实体匹配数据集缺少指标衡量效果需要的大量对属性关联和的领域专家标注[28], 我们通过展示一个案例研究(case study), 以展示 CSKNN 配合属性关联如何通过关联性信息来让用户获取对预训练语言模型实体匹配决策和近似样本的理解.

- 案例研究

考虑如图 6 所示 Beer 数据集在 Ditto 上的错分样本 *x* 的注意力机制权重关联解释结果. 首先, 由于字典中没有 Tarraco, Ditto 的分词结果将 Tarraco 这一西班牙地名错误地分解为 T、arr、aco 这 3 个单词, 导致其没有合理地结合 Guineu (理解为一种西班牙产的啤酒或者理解为几日亚国家). 而 CSKNN 搜索到的 4 个样本 x_1, x_2, x_3, x_4 展示了一些前/后缀相同的同义缩写(前缀相同的 Co.与 Company)、同义词(前缀相同的 Brewing Company

和 Brewery)、近义词(前缀相同的 Brewery 和 Brewpub)、同领域词(后缀相同的 Red Ale 与 Pale Ale).

从上述实验结果可以看出, 4 个近邻样本在提示, 模型在处理目标样本时可能受到了 aco 和 Guineu 的同义/近义关系的影响(由于没有切分正确, 无法判断它们是同领域词), 且由于前/后缀的问题(实际上体现为 Tarraco 切分错误), 造成了错误决策.

图 6、图 7 中, 显然, 4 样本的解释方法(同义、近义词)与图 5(主键区别)不同, 体现了基于注意力权重的关联分析挖掘不同属性关联的能力, 且搜索到的训练集中最高注意力权重都捕捉到了模型决策依据的重要属性关联部分. 搜索到的历史样本配合专家标注的训练集解释能够提供提示给用户, 辅助对于模型决策的理解.

4.3.3 实验 3: KNNE 提升模型的效果与参数影响

- 性能提升

对于 KNNE, 这一方法作用于模型预测结果. 因此, 为了保证其增强验证的全面性, 我们同时测试了该决策增强方法作用在两个最新 BERT 类模型: Ditto 和 JointBERT 上的准确率、召回率、*F*1-*score* 和时间结果. 实验结果见表 6, 可以看到, 绝大部分情况下, 增强后的 *F*1-*score* 高于增强前, 且主要趋势是在增强后, 模型准确率提升、召回率降低. 注意: 此处增加时间为占预测时间的比例, 预测时间是远远小于预训练和精调时间的. 因此, 增加时间相比于模型处理数据的整个过程几乎可以忽略不计.

表 6 原模型与 KNNE 增强模型性能比较

原模型	数据集	原始准确率	增强准确率	原始召回率	增强召回率	原始 *F*1	增强 *F*1	增加时间(%)
Ditto	Amazon-Google	0.701 1	**0.731 2**	**0.812 0**	0.790 6	0.752 5	**0.759 8**	16.35
	Beer	0.923 1	**0.982 3**	0.857 1	**1.0**	0.888 9	**0.903 2**	26.22
	Cameras	0.803 4	**0.876 2**	**0.846 8**	0.797 3	0.824 6	**0.834 9**	16.41
	Computers	0.932 9	**0.941 0**	**0.909 1**	0.906 2	0.920 9	**0.923 3**	15.79
	DBLP-ACM	0.943 7	**0.966 2**	**0.982 0**	0.966 2	0.962 5	**0.966 2**	15.87
	ITunes-Amazon	0.960 0	**0.960 0**	**0.888 9**	0.888 9	0.923 1	**0.923 1**	16.56
	Watches	0.848 6	**0.887 2**	**0.848 6**	0.831 0	0.848 6	**0.858 2**	16.93
	DBLP-GoogleScholar	**0.962 6**	0.948 1	0.937 4	**0.973 8**	0.949 8	**0.960 8**	17.00
	Shoes	**0.837 8**	0.852 8	**0.893 0**	0.810 7	**0.864 5**	0.831 2	16.06
	Walmart-Amazon	**0.714 3**	0.747 1	**0.729 2**	0.661 5	**0.721 6**	0.701 7	16.33
JointBERT	Cameras	0.844 4	**0.896 4**	**0.855 9**	0.779 3	**0.850 1**	0.833 7	3.8
	Computers	0.800 0	**0.845 1**	**0.897 7**	0.852 3	0.846 1	**0.848 7**	4.54
	Shoes	0.794 9	**0.822 9**	**0.637 6**	0.592 6	**0.707 8**	0.689 0	3.84
	Watches	0.856 1	**0.920 2**	**0.816 9**	0.771 1	0.836 0	**0.839 1**	3.81

- 参数测试

最后, 我们进行参数影响测试, 图 8 中为阈值δ和 *k* 变化时, *F*1-*score* 变化的部分结果. 可以看到, 在大部分数据集上, 选择较小的 *k* (*k*=5)和 70% (紫色折线图)作为阈值, 即可得到较好的结果. Fodors-Zagats 数据集在 Ditto 和 JointBERT 上已经有足够优秀的效果, 我们在实验中观察到, 其错分样本不超过 2 个. 因此我们认为, 该数据集的增强空间较小且实验数据不具有参考意义, 不在表 6 和图 8 中展示.

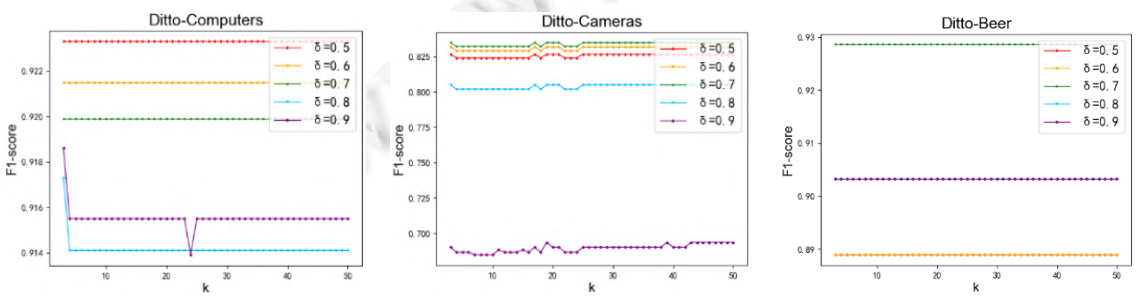


图 8 阈值δ和 *k* 对 *F*1-*score* 的影响

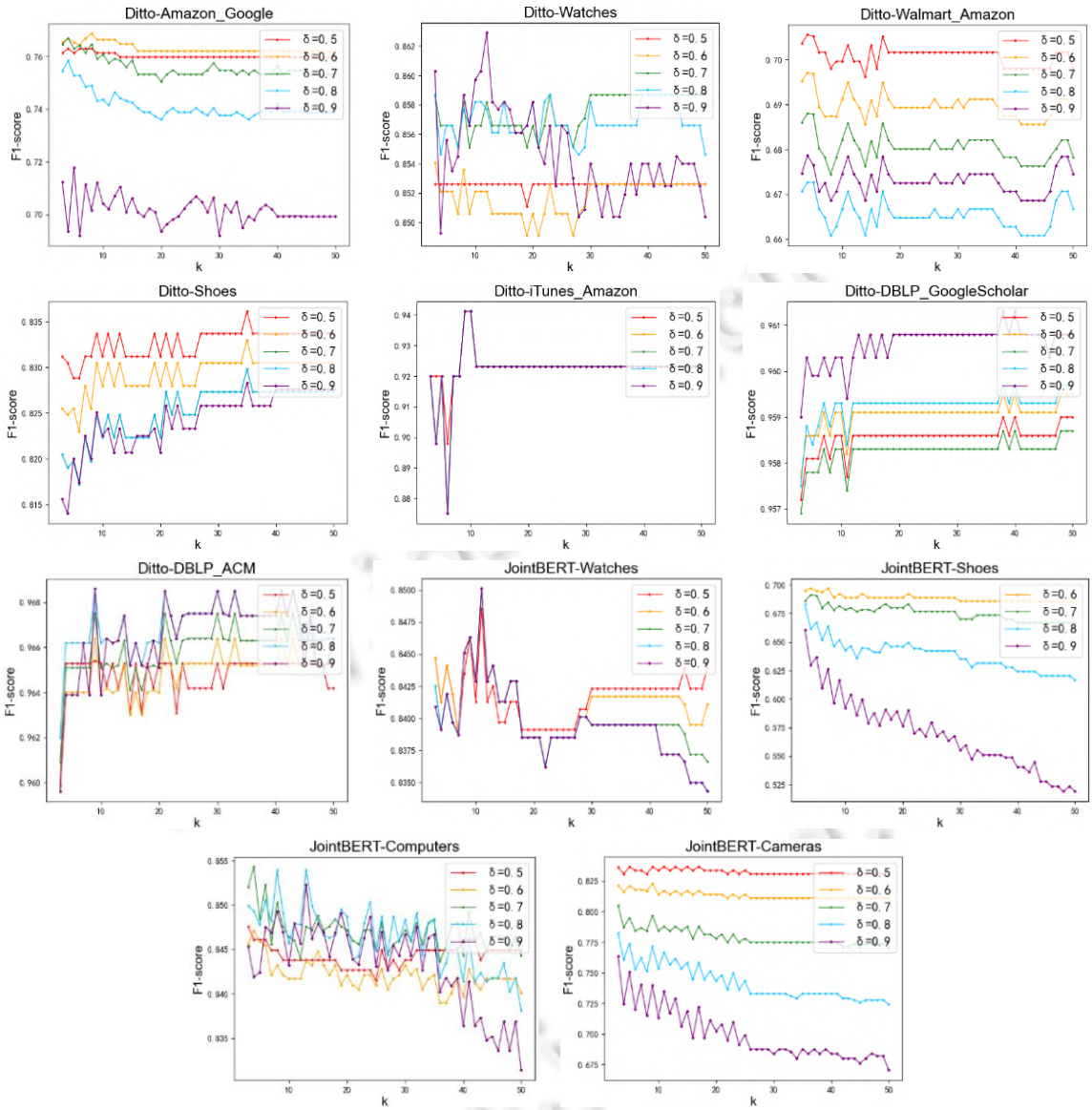


图 8 阈值δ和 k 对 *F1-score* 的影响(续)

5 总 结

基于预训练语言模型的实体匹配方法可以有效地利用实体对之间的语义信息，有着良好的性能和效果. 该类方法的应用，需要定制化的模型解释方法支撑. 针对此需求，本文提出了一种兼顾行列的预训练语言模型实体匹配解释方法. 针对现有实体匹配模型解释方法不能很好地挖掘属性关联、利用属性序信息和表征等问题，本文基于元学习和属性关联度，进行属性-偏序图上的最大权拓扑排序生成属性序反事实，并结合预训练语言模型的注意力机制分布对属性关联进行了挖掘. 通过进一步引入生成的元组对嵌入表示下的一致显著度 k 近邻搜索，可以有效地搜索近似样本给予提示，并提升模型在低置信度样本上的决策能力. 通过在大量实体匹配数据集的实验，验证了本文所提出的可解释性方法不仅可以从属性关联、属性序等新的角度解释模型决策，还可以提升模型在低置信度样本上的决策性能. 我们的未来工作包括面向实体识别模型调优的可解释

性方法设计、针对属性关联可解释性的评估、标注近似样本并构建评价指标等.

致谢 我们首先感谢本文的审稿人, 他们耐心以及极富价值的建议促成了本文的定稿. 此外, 我们特别要感谢李扬名, 感谢他对我们的科研工作一直以来的鼓励支持, 以及对论文涉及原理的宝贵建议.

References:

[1] Doan AH, Halevy AY, Ives ZG. Principles of Data Integration. Morgan Kaufmann Publishers, 2012.

[2] Dong XL, Rekatsinas T. Data integration and machine learning: A natural synergy. In: Proc. of the Int'l Conf. on Management of Data (SIGMOD 2018). ACM, 2018. 1645−1650.

[3] Wang J, Li G, Yu JX, Feng J. Entity matching: How similar is similar. Proc. of the VLDB Endowment, 2011, 4(10): 622−633.

[4] Chai C, Li G, Li J, Deng D, Feng J. A partial-order-based framework for cost-effective crowdsourced entity resolution. VLDB Journal, 2018, 27(6): 745−770.

[5] Das S, *et al*. Falcon: Scaling Up Hands-off Crowdsourced Entity Matching to Build Cloud Services. In: Proc. of the ACM Int'l Conf. on Management of Data (SIGMOD 2017). Chicago, 2017. 1431−1446.

[6] Ebraheem M, Thirumuruganathan S, Joty SR, Ouzzani M, Tang N. Distributed representations of tuples for entity resolution. Proc. of the VLDB Endowment, 2018, 11(11): 1454−1467.

[7] Li Y, Li J, Suhara Y, *et al*. Deep entity matching with pre-trained language models. Proc. of the VLDB Endowment, 2020, 14(1): 50−60.

[8] Tu JH, Fan J, Tang N, Wang P, Chai CL, Li GL, Fan RX, Du XY. Domain adaptation for deep entity resolution. In: Proc. of the Int'l Conf. on Management of Data (SIGMOD 2022). Philadelphia: ACM, 2022. 443−457.

[9] Peeters R, Bizer C. Dual-objective fine-tuning of BERT for entity matching. Proc. of the VLDB Endowment, 2021, 14(10): 1913−1921.

[10] Ebaid A, Thirumuruganathan S, Aref WG, *et al*. EXPLAINER: Entity resolution explanations. In: Proc. of the ICDE. 2019. 2000−2003.

[11] Ribeiro MT, Singh S, Guestrin C. Why should *i* trust you? Explaining the predictions of any classifier. In: Proc. of the 22nd ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. 2016. 1135−1144.

[12] Sood A, Craven M. Feature importance explanations for temporal black-box models. In: Proc. of the AAAI. 2022. 8351−8360.

[13] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I. Attention is all you need. In: Proc. of the NIPS. 2017. 5998−6008.

[14] Mahajan D, Tan CH, Sharma A. Preserving causal constraints in counterfactual explanations for machine learning classifiers. arXiv:1912.03277, 2019.

[15] Karimi AH, Barthe G, Schölkopf B, Valera I. A survey of algorithmic recourse: definitions, formulations, solutions, and prospects. arXiv:2010.04050, 2020.

[16] Rajani NF, Krause B, Yin WP, Niu T, Socher R, Xiong CM. Explaining and improving model behavior with *k* nearest neighbor representations. arXiv:2010.09030, 2020.

[17] Benjelloun O, Garcia-Molina H, Menestrina D, *et al*. Swoosh: A generic approach to entity resolution. VLDB Journal, 2009, 18(1): 255−276.

[18] Chaudhuri S, Chen BC, Ganti V, Kaushik R. Example-driven design of efficient record matching queries. In: Proc. of the VLDB. 2007. 327−338.

[19] Wang J, Li G, Kraska T, Franklin MJ, Feng J. Leveraging transitive relations for crowdsourced joins. In: Proc. of the SIGMOD Conf. 2013. 229−240.

[20] Vesdapunt N, Bellare K, Dalvi N. Crowdsourcing algorithms for entity resolution. Proc. of the VLDB Endowment, 2014, 7(12): 1071−1082.

[21] Konda P, *et al*. Magellan: Toward building entity matching management systems. Proc. of the VLDB Endowment, 2016, 9(12): 1197−1208.

[22] Wu R, Chaba S, Sawlani S, *et al*. ZeroER: Entity resolution using zero labeled examples. In: Proc. of the SIGMOD Conf. 2020. 1149−1164.

[23] Meduri V, Popa L. Sen P, Sarwat M, *et al*. A comprehensive benchmark framework for active learning methods in entity matching. In: Proc. of the SIGMOD. 2020. 1133−1147.

[24] Mudgal S, Li H, Rekatsinas T, *et al*. Deep learning for entity matching: A design space exploration. In: Proc. of the SIGMOD Conf. 2018. 19−34.

[25] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional Transformers for language understanding. In: Proc. of the NAACL-HLT, Vol.1. 2019. 4171−4186

[26] Brunner U, Stockinger K. Entity matching with Transformer architectures—A step forward in data integration. In: Proc. of the EDBT. 2020. 463−473.

[27] Chen ZQ, Chen Q, Hou BY, Li ZH, Li GL. Towards interpretable and learnable risk analysis for entity resolution. In: Proc. of the ACM SIGMOD Int'l Conf. on Management of Data (SIGMOD 2020). Portland: ACM, 2020. 1165−1180.

[28] Wallace E, Feng S, Boyd-Graber J. Interpreting neural networks with nearest neighbors. arXiv:1809.02847, 2018.

[29] Wachter S, Mittelstadt BD, Russell C. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. arXiv:1711.00399, 2017.

[30] Sokol K, Flach PA. Counterfactual explanations of machine learning predictions: Opportunities and challenges for AI safety. In: Proc. of the SafeAI@AAAI, Vol.2301. 2019. Paper 20.

[31] Cappuzzo R, Papotti P, Thirumuruganathan S. Creating embeddings of heterogeneous relational datasets for data integration tasks. In: Proc. of the SIGMOD Conf. 2020. 1335−1349.

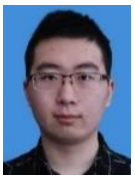
[32] Wang Y, Huang M, Zhu X, Zhao L. Attention-based LSTM for aspect-level sentiment classification. In: Proc. of the Conf. on Empirical Methods in Natural Language Processing. 2016. 606−615.

[33] Primpeli A, Peeters R, Bizer C. The WDC training dataset and gold standard for large-scale product matching. In: Companion Proc. of the World Wide Web Conf. 2019. 381−386.

[34] Yang Q, Fan LX, Zhu J. Introduction to Interpretable Artificial Intelligence. Beijing: Electronic Industry Press, 2022 (in Chinese with English abstract).

附中文参考文献:

[34] 杨强, 范力欣, 朱军. 可解释人工智能导论. 北京: 电子工业出版社, 2022.



梁峥(1998－), 男, 博士生, CCF 学生会员, 主要研究领域为数据集成, 实体识别, 异常检测.

邵心玥(1996－), 女, 博士生, 主要研究领域为黑盒算法可解释性, 反事实解释.

王宏志(1978－), 男, 博士, 教授, 博士生导师, CCF 杰出会员, 主要研究领域为数据库管理系统, 大数据分析与治理.

丁小欧(1993－), 女, 博士, 助理教授, CCF 专业会员, 主要研究领域为数据质量, 数据清洗, 时序数据管理.

戴加佳(2000－), 女, 硕士生, 主要研究领域为数据质量, 实体识别.

穆添愉(1998－), 男, 博士生, CCF 学生会员, 主要研究领域为自动机器学习, 模型自动选择, 超参数优化.