

**研究报告主题**：深度学习

**主题方向：**基于深度学习的Blocking模型

**关键词：**深度学习、实体匹配、阻塞（Blocking）

**报告概述：**

报告围绕基于深度学习的Blocking模型为主题，分别介绍了DeepER、AutoBlock和DeepBlocker三大典型模型的框架、所蕴含的科学问题、所用的理论方法以及研究成果，对前两者的研究工作的不足也进行一定的指出和评述。在介绍各模型之前，我首先对相关概念进行了一定的学习并呈现，包括深度学习、自然语言处理、实体匹配以及阻塞（Blocking）

**相关概念介绍：**

**1、深度学习**

**1.1解读**：深度学习 (Deep learning) 通过建立深层神经网络, 模拟人脑的机制进行解释并分析学习图像、语音及文本等数据, 是目前机器学习研究中的一个热点领域。

**1.2本质**：一种基于原始特征（或者是未经过人类思维分析的数据）输入，通过多层非线性处理来学习复杂特征表示的方法。

**1.3任务：**自动完成数据表示和特征提取工作，以及提取不同水平、不同维度的有效表示提高不同抽象层次上对数据的解释能力。

**1.4主要思想：**摆脱人选择的局限性，试图从大量可观测到的浅层感官数据中识别或解释关键特征。

**1.5存在的研究问题：**一个特征表示优于另一个表示的主要因素是什么?给定一个表示对象, 我们如何学习好的特征表示?

**1.6涉及的算法：**无监督学习算法预训练得到下一层的输入集；有监督学习方法调整所有层的网络参数降低误差。

**2、自然语言处理**

**2.1定义：**自然语言处理（Natural Language Processing，NLP）是AI的一个领域，旨在让计算机理解和使用人类语言，从而执行有用的任务。主要划分为两个部分：自然语言理解和自然语言生成。

**2.2目前常应用领域**：主要应用于机器翻译、舆情监测、自动摘要、观点提取、文本分类、问题回答、文本语义对比、语音识别、中文OCR等方面。

**2.3涉及的科学问题：**语言模型、词向量化（数据表示）以及学习模型

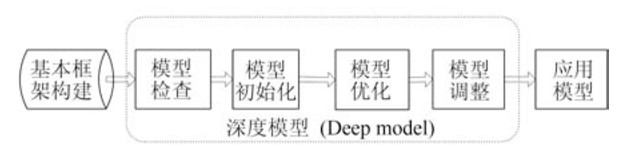
**2.4常用的方法：**

**2.4.1数据表示方法**：one-hot representation、词向量、词袋模型、向量空间模型

**2.4.2学习模型**：神经网络于n元语法模型、Log-bilinear语言模型、Hierarchical log-bilinear语言模型、循环神经网络语言模型等等

**2.5深度学习在自然语言处理领域应用满足的需求**：特征表示学习的需要、无监督特征和权重学习的需要以及学习多层分类表示的需求

**2.6面向自然语言处理的深度学习应用策略框架**



**3、实体匹配**

**3.1定义：**又名实体解析、记录链接以及重复检测或引用协调。指确定两个不同的表示是否引用相同的真实世界实体任务

**3.2对象：**关系表、半结构化或非结构化文本文档

**3.3作用：**识别和删除重复键，创建实体整体视图，构建主数据库管理和知识库构建

**3.4介绍：**大多数的实体匹配解决方案包含两个步骤：blocking和matcing。Blocking阶段会使用启发式的方法快速过滤掉不可能匹配的实体对；matching阶段则会使用一个匹配器来预测保留下来的实体对是否匹配

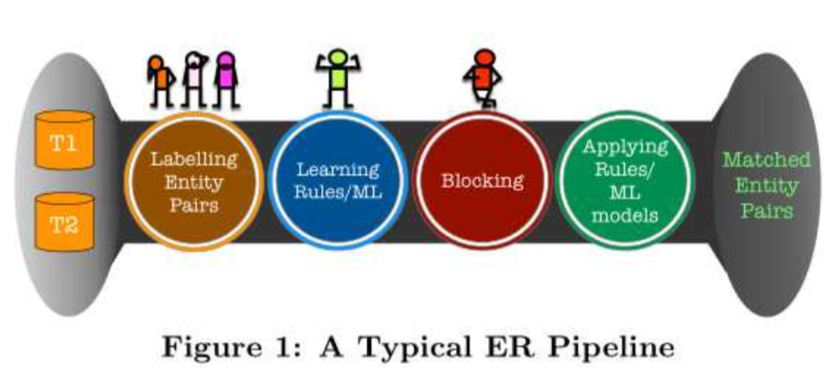
**4、阻塞（Blocking）：**与macthing共同组成实体匹配两大主要内容

**4.1目标**：是从进一步考虑中快速消除已知不匹配的高确定性记录对，从而减少在随后的EM步骤需要完成的匹配工作量

**4.2大多数Blocking方法的不足：**基于key且不需要学习的，成功建立在干净数据上的人工努力

**4.3常见的Blocker：**基于属性等价、哈希和相邻排序的

**三大模型的介绍：**

早期的实体匹配是通过两个记录之间相似度方法将句法相似性和两个记录相似属性进行比较然后进行配对。随着自然语言处理的发展和深度学习的发展，前者使基于文本中出现在属性值中的标记的语义来采用相似性度量成为可能，后者使模型不需要事先制定两个记录的对应属性来做出匹配决策，使实体匹配工作得到飞快的发展。而今天，我所要谈论的便是实体匹配的一项重要的工作——blocking环节。下面，我将要分别介绍三大基于深度学习的Blocking典型模型：DeepER、AutoBlock、DeepBlocker  


图为一种典型的ER管道（也为DeepER的框架图）：

1）对于输入的实体数据，先将实体标记为匹配或非匹配对；

2）使用标记数据学习规则或者是训练机器模型；

3）再通过Blocking（分块，阻塞）分块以减少比较次数；

4）最后应用到学习的规则或机器学习模型中，对需要匹配的实体进行配对。

**1、DeepER：**通过对单个单词的向量进行未加权聚合为tuple向量

**1.1模型的方法论：**

1）通过考虑被匹配值的先验知识而需要更少的被标记的数据

2）在没有特征工程和参数调整的情况下捕获句子和语义相似性

3）提供一种自动化和可定制的blocking方法

4）采用所用属性的整体试图

**1.2涉及的技术：**

**1）Glove**：对单词通过Glove获得分布式表示（向量化）

**2）Adam优化算法**：采用Adam算法进行模型的优化

**3）LSTM-RNN：**使用具有长短记忆（LSTM）隐藏单元的单向和双向递归神经网络（RNN）从单词的分布式表达获取元组的分布式表达

**4）LSH**：应用基于元组分布式和局部敏感哈希（LSH）的Blocking方法进行近似分块

**5）余弦相似度：**使用余弦相似度计算元组间的相似度

**6）K折验证法：**使用重复与非重复比1：100的K折验证，对模型性能进行验证

**1.3研究成果：**

1）介绍了DeepER，一种基于深度学习的ER方法

2） 提出了将元组转换为元组分布式的算法，构建了元组分布式感知分类器和基于LSH的有效blocking策略

**1.4研究不足及评述：**

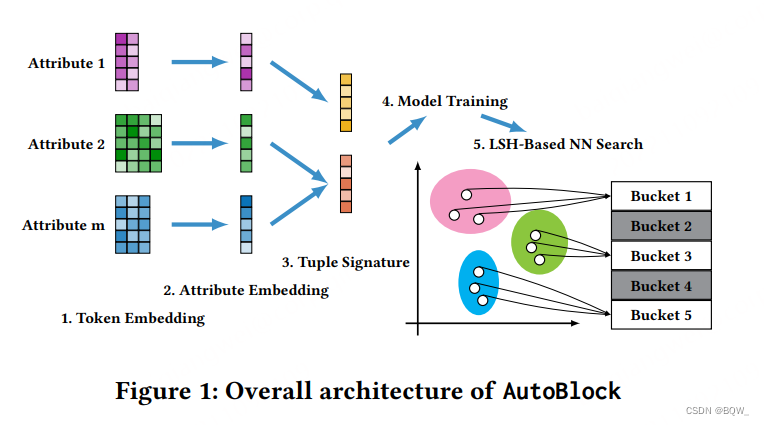
**1）数据表示缺少对上下文的理解。**由于DeepER是早期将深度学习应用于实体匹配的的探索成果，当时流行的三种分布式模型为Glove、Word2Vec和FastText，都是单独对单个对象进行向量化，再综合得到数据表示，而缺少相关理论技术以整体视图对数据进行表示，比如Transformer框架。因此，对于此问题，可以借助Transformer进行解决；

**2）要先验知识代替人工打标签，即仍需要人类专家的大量参与**。以往更多的是人为的参与对两个实体进行标签（是否匹配），来获取标签化的实体对，而该论文提出了通过人类专家的参与成比例的减少人工打标签。目前，ChatGpt大模型的出现，通过有效的指引工程和示例的输入，能够在没有人类过多的参与下，快速对数据进行标签化或者领域参与。

**3）对脏数据集的处理不当**。一般的基于深度学习的实体匹配对于结构型数据集的处理都是比较突出的，即超过非基于深度学习的模型。而对于非结构型和脏数据集的处理，增强效果并不显得突出。在未来，探索一种新的模型框架，解决脏数据集和非结构数据集的问题也是一大探索方向。

**2、AutoBlock:** 标注数据来学习聚合词的权重

**2.1框架**



**2.2涉及的技术**

**1）fastText：**通过对tokens的字符学习一个嵌入，再通过字符求和计算单词的词向量

**2）注意力机制：**在获取词向量后，通过权重的分配得到属性嵌入向量

**3）元组的多次签名**（Signature）：通过对每个元组的所有属性的多次随机选择，对每个元组形成多个签名（Signatures）形成元组级别的表示

**4）cosine相似度（余弦相似度）：**计算元组之间的相似度

**5）快速最近零搜索和LSH：**设置阈值，依据相似度找到阈值内的最近邻，再通过局部敏感哈希获取候选对

**2.3涉及的理论：**

**1）最近邻搜索：**输入元组集合，获取相似度度量，设置阈值，输出得到候选对

**2）哈希冲突（Hash Collision）**：通过相同的哈希值将元组进行分块

**2.3研究成果（贡献）：**

**1）Automation：**AutoBlock 从大量人工清洗数据和blocking key调整中解放了用户；

**2） Scalability：**AutoBlock 具有次均分时间复杂度，并能够轻易应用在上百万records；

**3） Effectiveness：**AutoBlock 在多个大规模、现实数据集上优于各类baselines，特别是在dirty和unstructured数据集

**2.4研究不足以及未来研究方向**

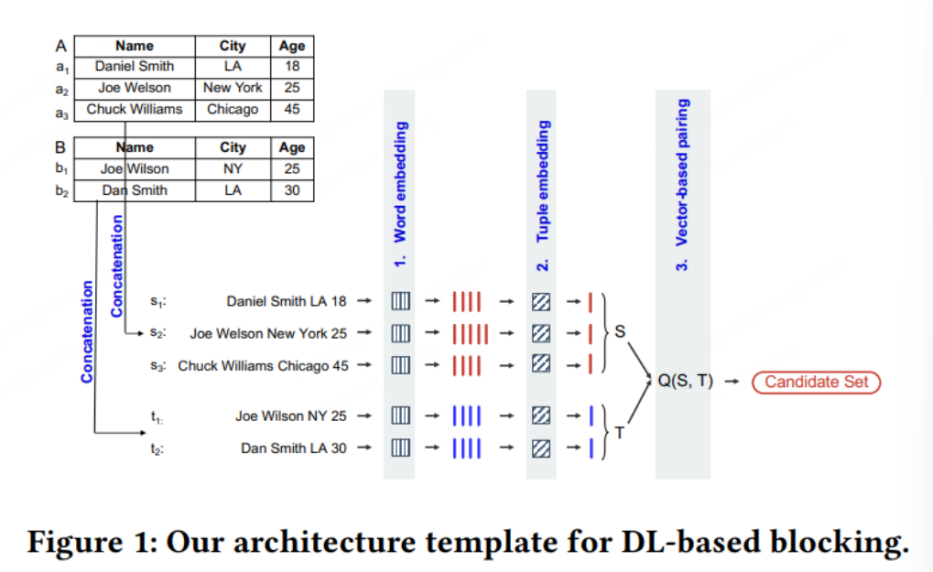
**1）仍具有数据表示不合理问题**：采用的是fastText进行词向量化，不能完全符合满足上下文理解。依旧是Transformer的引入

**2）可解释性低：**整个实现过程属于“黑箱”，因此实验过程可解释性很低，尤其在于协调人机学习过程中难以进行衡量。因此，未来可以在模型的可解释性方面进行探索，以求更好实现人机学习的协调

**3）匹配检索的速度不够。**近年来，快速发展的Faiss技术较为快速的解决此问题。’

**3、DeeoBlocker框架：**仍然会执行基于关键词的聚合，但是会使用词向量计算关键词的语义相似度来进行优化

**3.1框架的模板**



**3.2所用的理论或方法**

**1）自编码和LSTM：**采用encoder-decoder（编码-解码架构）。在编（解）码器中都使用LSTM-RNNS模型

**2）FastText：**在word embedding环节中主要采用fastText字符级向量嵌入

**3）Aggregation或者Self-Supervised：**前者包括不加权平均和加权平均，后者包括Self-reproduction、cross-tuple training、Triplet loss minimization和Hybrid

**4）向量配对：**哈希、排序、相似度。主要采用top-k cosine相似度配对

**5）模型：**SIF、LSTM、Feed-Forward network、BERT、Siamese network、Transformer

**6）评估方法：**使用Recall（召回率）和CCSR评估模型的精确度，CCSR，RunTime评估模型的运行速度

**3.3研究成果（贡献）：**

1）定义用于blocking的深度学习解决方案的解空间，包含不同复杂性的解决方案和大量先前的工作

2）解空间中开发8个具有代表性的解决方案：解决方案不需要标注数据，并利用了深度学习近期进展

3）最优方案在dirty和textual数据上超越现有解决方案，并在结构化数据上具有更好的竞争力

**参考文献：**

[1]奚雪峰,周国栋.面向自然语言处理的深度学习研究[J].自动化学报,2016,42(10):1445-1465.DOI:10.16383/j.aas.2016.c150682.

[2]唐晓彬,沈童.深度学习框架发展综述[J].调研世界,2023(04):83-88.DOI:10.13778/j.cnki.11-3705/c.2023.04.009.

[3]马郅斌.自然语言处理中的深度学习:方法及应用[J].科技传播,2020,12(21):128-130.DOI:10.16607/j.cnki.1674-6708.2020.21.041.

[4]Wei Zhang, Hao Wei, Bunyamin Sisman, Xin Luna Dong, Christos Faloutsos,

and Davd Page. 2020. AutoBlock: A hands-off blocking framework for entity matching. In WSDM. 744–752.

[5]Thirumuruganathan, S.; Li, H.; Tang, N.; Ouzzani, M.; Govind, Y.; Paulsen, D.; Fung, G.; Doan, A. Deep Learning for Blocking in Entity Matching: A Design Space Exploration. Proc. VLDB Endow. 2021, 14, 2459–2472

[6] Ebraheem, M ; Thirumuruganathan, S ; Joty, S; Ouzzani, M ; Tang, N .Distributed Representations of Tuples for Entity Resolution.2018,11,1454-1467

[7] Gurajada, S., Popa, L., Qian, K., and Sen, P. (2019). Learning-based methods with human-in-the-loop for entity resolution. In Proceedings of the 28th ACM International Conference on Information and Knowledge Management, Beijing, China, November 3–7, 2019. pp. 2969–70

[8] Li, Y., Li, J., Suhara, Y., Wang, J., Hirota, W.: Deep entity matching: challenges and opportunities. ACM J. Data Inf. Qual. 13(1), 1:1-1:17 (2021)