

总方向：深度学习+实体匹配

1、深度学习：一篇中文论文。视频

2、实体匹配：一篇论文。视频

3、深度学习与实体匹配的结合：三篇英文论文、视频

4、总结：一篇论文。视频

报告结构：

1、深度学习是什么（涉及的科学、所用的理论以及研究结果）、带来了什么现实问题（现实不足，以及评述）

2、实体匹配是什么（涉及的科学、所用的理论以及研究结果）、带来了什么现实问题（现实的不足以及对其评述）

3、深度学习与实体匹配的结合应用（涉及的技术问题、方法以及研究成果）、仍然存在的不足

4、总结

# 篇1：深度学习：

[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.

## 一、深度学习的概述

1、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

2、深度学习的任务：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。在认知科学角度上思路与人类学习机理非常吻合【不免引发一个思考：深度学习是否能够逐渐使机器拥有自己的意识】

3、深度学习的发展历程

3.1BP神经网络：模仿人的大脑分层次结构式处理信息（在减少数据量的同时保存关键的结构信息）——BP神经网络：初始值随机选定以及梯度下降优化策略。缺点：容易导致局部收敛最小而非全局最优

3.2浅层学习结构：将单个原始输入信号直接映射到特定问题空间的简单特征结构，CRF\HMM\SVM（支持向量机）\MLP（多层感知机）\ME（最大熵模型）。缺点：对复杂函数表示能力有限、对复杂问题泛化处理能力不足的局限性

3.3深度可信网络结构（DBN）：在每一层局部进行无监督的训练算法，引导完成特征中间表示层的训练目标

3.4深度神经网络

### （一）深度结构

4、深度结构：3层以上的隐层节点（完成复杂函数的逼近）、强调特征学习（特征工程）的重要性（非监督预训练算法）、避免过拟合问题

### （二）应用动机

5、深度学习应用的动机：解决自动从数据中抽取和组织信息的能力

6、深度学习的主要思想：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。因此，深度学习也称为无监督特征学习

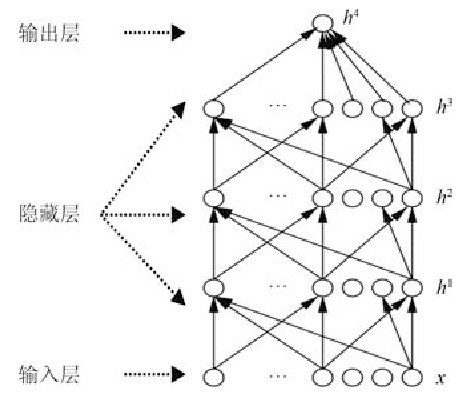
### （三）首要任务

7、深度学习的首要任务：尽可能采用一种简单的算法来实现所求解问题的分层特征表示。强调特征学习或表示学习的重要性

8、深度学习的存在的研究问题：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?

### （四）基本框架

9、深度学习的本质：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



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

无监督构建自动编码器：设置Encoder—Decoder解码编码器，调整编码器从而降低损失，实现每一层输入集的表达特征的自动获取

有监督训练分类器：为了实现分类, 我们需要在自动编码器最高层的编码层添加分类器 (Classiﬁer) , 结合标签 (Label) 样本, 基于标准神经网络的有监督训练方法调整参数.

## 二、面向自然语言处理的深度学习研究及研究

10、面向自然语言处理的深度学习研究以及应用

语言模型是最早采用神经网络开展研究的自然语言处理问题

基于词向量方法以及多层一推卷积神经网络实现了实现了一个同时处理词性标注、语块切分、命名实体识别、语义角色标注四个典型自然语言处理任务的SENNA系统

基于深度模型的特征学习还在词义消歧、情感分析等自然语言处理任务中超越当时最优系统

### （一）深度学习在自然语言处理领域应用的可行性分析

11、深度学习在自然语言处理领域应用的可行性分析

11.1特征表示学习的需要：对处理对象的表示形式，深度学习中的特征提取可以自动从数据中学习获取特征.

11.2无监督特征和权重学习的需要：无监督方法解决从未标注数据中挖掘信息。深度神经网络采用无监督方式完成预训练过程

11.3学习多层分类表示的需求：需要利用好的学习模型，更多地抽取出有用的中间表示形式。深度学习能够较好抽取处理任务的多层分类表示

11.4当前可用的技术及硬件平台支撑：高性能计算的硬件平台、良好的算法支持技术；自然语言处理的语言模型/算法优化

### （二）面向自然语言处理的深度学习研究模型

12、面向领域任务的深度学习研究及应用需要解决的普适问题

12.1 应用领域的原始特征表示——数据的表示问题

12.2选择合适的深度学习算法——深度学习结构问题（深度学习模型）

13、面向自然语言处理的深度学习研究考虑的两个普适问题

13.1典型的有基于词向量空间、词袋模型、向量空间模型等表示方式

13.2需要根据自然语言递归特性的特点，选择合适的深度学习模型——循环神经网络、递归神经网络、卷积神经网络等

14、自然语言处理中应用深度学习的方式，两类：

14.1在深度学习模型中，直接使用原始特征，构建一类端到端系统，完成处理任务。基于词向量方法及多层一维卷积神经网络完成了词性标注、语块切分、命名实体识别等系列任务;基于递归神经网络实现情感分析、句法分析等多项任务

14.2在现有模型中，将训练后的原始特征作为辅助特征扩充使用。词向量作为额外的特征加入到现有最优系统中[52], 进一步提高了命名实体识别和短语识别的效果.

15、数据表示：

15.1 one-hot representation：每个词表示为一个很长的向量;其中只有一个维度的值为1, 代表了当前的词;其他绝大多数元素都为0;向量的维度是词表的大小。稀疏方式储存

15.2 词向量（Distributed representation）：低维实数向量，相似词的词向量距离相近

15.3词袋模型：忽略文本的语法和语序。计算词频

15.4向量空间模型（VSM）：把对文本内容的处理简化为向量空间中的向量运算, 并且它以空间上的相似度来表示语义的相似度

16、学习模型：诸多模型, 本质上非常相似, 差别主要在于隐藏层到输出层的语义定义

16.1神经网络与n元语法模型

16.2Log-bilinear语言模型

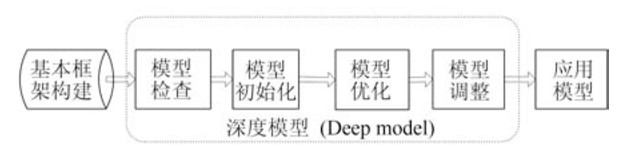
16.3Hierarchical log-bilinear语言模型

16.4循环神经网络语言模型

16.5基于词向量的改进模型

### （三）面向自然语言处理的深度学习应用策略

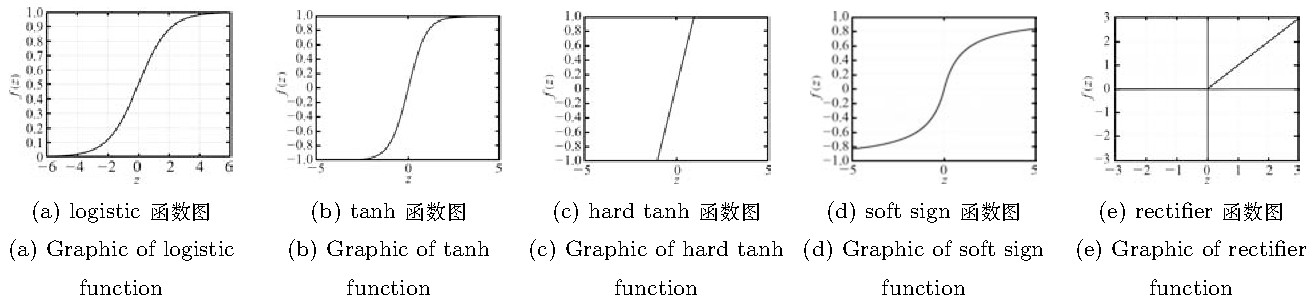
17、面向自然语言处理的深度学习应用策略



17.1步骤一：构建基本模型框架：选择合适的神经网络结构，构建深度学习基本模型框架

17.1.1基本表示结构：词向量化、

17.1.2非线性过程：激活函数：logistic、sigmoid、tanh、hard tanh



17.2步骤二：模型检查：采用梯度下降法检查模型实现是否存在错误

17.2.1梯度下降、损失函数

17.3步骤三：模型初始化：设计神经网络隐藏层偏置量和网络结点权重矩阵的参数初始化

17.3.1隐藏层偏置量为0，权重值都为0

17.4步骤四：模型优化：涉及模型参数调整优化

17.4.1损失函数采用梯度下降而下降

17.5步骤五：模型调整：是否过拟合？采用正则化调整，防止过拟合

17.5.1降低模型规模

17.5.2使用标准的L2、L1的regularity限制调整权重值或者稀疏化方式降低模型复杂度

### （四）面向自然语言处理的深度学习典型应用

18、面向自然语言处理的深度学习典型应用

18.1分词和词性标注：英文分词和词性标注：SENNA系统；中文分词和词性标注：特征发现、数据表示和模型算法三方面

18.2句法分析

18.3词义学习

18.4情感分析、

18.5机器翻译

19、面向自然语言处理的深度学习平台工具

19.1函数库/工具包：Python库：Theano、Kears神经网络库、TensorFlow

19.2数据表征工具：SENNA、Word2vec、

19.3经典神经网络模型：卷积神经网络、循环神经网络和递归神经网络

19.4深度神经网络组件：RBM、GPU、DBN、LSTM、递归自动解码器

19.5循环神经网络等工具

## 三、存在的问题和未来研究方向

20、存在的问题和未来研究方向：

20.1数据表示问题及展望

问题：自然语言是包含了人类的语义解释，并非是原始信息；图像和语音是原始的。

展望：考虑如何将深度学习与自然语言处理任务结合的具体落地应用也是值得研究的重点。即如何将自然语言的数据进行表示，再作为深度学习的输入进行分析

20.2学习模型问题及展望

问题：深度学习模型研究中，难点在于模型构建过程中参数的优化调整方面：网络层数、正则化问题和学习率问题等的调整。最突出的问题是训练速度较慢

展望：提升网络训练速度、选择合适的正则化

20.3其它问题及思考

20.3.1自动学习和人工结合

问题：一直强调学习特征采用自动方法

思考：再针对问题领域，选择合适的模型架构后，仍然需要人类知识的融入

20.3.2自然语言的不确定性

问题：一词多义的存在，可能使不确定性随着层数的增加不断的放大

思考：结合上下文语言情景：突破自然语言字、词、短句等局部表示，面向篇章分析、篇章理解是重点发展方向

## 四、结束语

21、结束语

未来的发展方向：

寻找更好的特征表示方式？

构建具有明显分层机制的模型？

# 篇2：深度学习框架发展概述

[1]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.

## 一、引言

1、深度学习的作用：深度学习方法特有的深层结构使得模型具有极强的特征表示和概念抽象能力，避免了传统机器学习方法对特征工程的依赖

2、深度学习框架的作用：

2.1在没有框架的情况下，开发者和使用者需要编写大量代码来完成复杂的计算操作，具有很高的学习成本和门槛；

2.2而框架通过将底层算法进行模块化封装极大提高了代码编写的效率并降低了使用门槛，推动了深度学习模型的发展和应用部署

## 二、国外深度学习框架

### 1、早期深度框架：

1.1Theano：现已停止开发

1.1.1特点：Python库；专用于定义、优化、求值数学表达式，适用于多维数组；特点在于：集成了Numpy库、可以使用GPU快速运算、编译优化提升速度

1.1.2缺点：偏向底层、难以调试、编译时间长、代码与其它框架相比不够精简、缺乏可使用的预训练模型

1.2Caffe：现已并入PyTorch框架

1.2.1特点：C++语言。

### 2、成熟的深度学习框架：对分布式提出要求

2.1Tensor Flow：谷歌

2.1.1优点：采用数据流图进行计算，节点代表数学运算，边表示节点间的数据；对不同平台有极强的支持；完善的资源和生态

2.1.2问题：接口变动太频繁且版本之间兼容性差；系统设计复杂；功能实现不统一；调试困难

2.2PyTorch：Facebook

2.2.1优点：使用了动态图框架，框架更容易调试，灵活性更高；设计简洁，具有张量、变量和模块三个抽象层次；具有速度优势‘极强的易用性；具有完善活跃的社区环境

2.2.2不足：运行效率低于静态图；在分布式和并行上存在操作不便的缺点

2.3MXNet：亚马逊，数据流图

2.3.1特点：全采用模块化设计，对编译依赖小，适合快速开发；对分布式提供了极强的支持，在内存和显存优化具有优势

2.3.2不足：文档不完善、推广力度不够

2.4CNTK



## 三、国内深度学习框架

## 四、深度学习框架发展趋势

1、易用性

2、分布式与并行能力：降低训练费用

3、统一性：框架各版本之间的接口与使用方法一致，框架对各类硬件及各类部署统一性

4、扩展性：更多模型与深度学习模型结合

5、计算效率：训练耗时长是一大主要缺点

## 五、简要评述——深度学习发展历程和使用情况的特点

1、深度学习框架处于快速发展阶段

2、新框架后发优势明显

3、深度学习框架的竞争已经超出公司层面

# 篇3：Learning-based Methods with Human-in-the-Loop for Entity Resolution-2019

## 1、现代ER系统的理想要求：

1.1通过（最小的）人类交互获取领域知识

1.2通过机器学习技术提供尽可能多的自动化

1.3实现高可解释性

## 2、实体解析（ER）：

目标：在不同的数据集中链接和融合对同一现实世界实体的不同提及

## 3、现代ER管道预计需要满足的三个重要需求

3.1获取人类领域专业知识：决定两个提及是否指的是同一个现实世界实体，即人工参与打标签

3.2自动化：能够广泛使用ER

3.3可解释性（白盒）：避免模型的黑盒性，防止ER模型存在偏见和衡量人类领域专业知识参与程度

4.1众包：

4.1.1人类智能任务（HIT）：通过向一群工人呈现成对的记录来获取标签想法

4.2积极学习：

4.3深度学习：分布式表达方面的有效性；不同数据集之间转移学习模型的能力

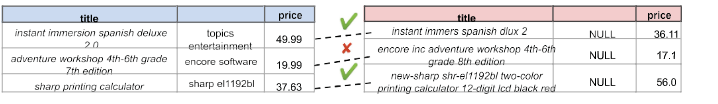
4.4可解释的ER：允许最终用户验证逻辑性，甚至定制和修改

# 篇4：Deep Entity Matching: Challenges and Opportunities—2021

1、实体匹配：确定两种不同的表示是否引用相同的真实世界实体的任务；或者确定不同实体的两个异构表示是否应关联在一起

## 一、引言

1、实体匹配(EM)：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



2、 实体匹配的作用

2.1识别和删除重复项，创建实体整体试图，构建主数据库管理和知识库构建

2.2从匹配子公司到母公司，到匹配求职者的职位描述

3、实体匹配中的发展

3.1早期通过两个记录之间的相似程度方法将句法相似性和两个记录相似属性进行比较

3.2自然语言处理的发展，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已经成为可能

3.3Transformer框架基于输入序列中的所有令牌生成令牌嵌入，高度上下文，更加奥捕捉单词的语义和上下文理解

3.4深度学习的发展，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，但仍能够学会关注正确的组件。形成预先训练的语言模型

## 二、DITTO

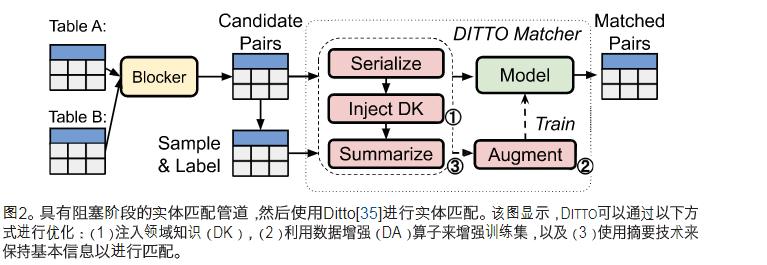
1、DITTO：依赖预训练的，基于Transformer的LMs（预训练的LMs）

2、DITTO的优化

2.1允许专家通过突出在做出匹配决策时值得关注的重要输入片段，将领域知识明确注入到训练实例中

2.2用（困难）例子扩充了训练数据，减少了训练DITTO所需的标记训练数据量

2.3DITTO可以对长字符进行汇总，以便只关注为为EM保留的最重要的信息



### （一）利用预先训练的语言模型

1、使用预先训练的LMs的单词嵌入的高度上下文的

2、上下文单词嵌入可以区分在不同上下文中使用的相同标记，也可以区分相反、在相似上下文中具有相同含义的不同标记

3、LMs：三种代表性的预训练的LMs：BERT、RBERT、DistilleBERT

### （二）优化模型训练

1、注入领域知识：允许工人通过跨度类型化和跨度规范化来突出重要令牌的跨度，从而注入此类领域知识

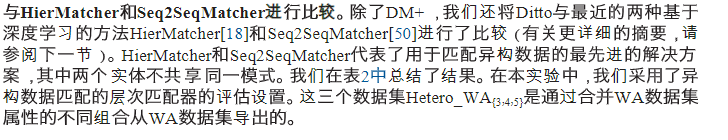
2、扩充训练数据：通过跨度级别、属性级别和实体级别的扩充操作符从现有示例中派生出新的训练示例

3、总结长字符串：TF-IDF衡量代币的重要性，从而将重要的令牌保存在长字符串中

### （三）DITTO的性能

1、基准数据：WDC、ER Magellan

2、LMs：RoBERT



## 三、实体匹配的深度学习模型

1、基于非transformer的方法：

1.1DeepER：基于LSTM的RNN和暹罗架构，即具有一堆完全相同架构和共享参数的神经网路

1.2DeepMatcher：探索了EM的深度学习模型的不同架构。DeepMatcher基于注意力的模型明确考虑了属性相似性

1.3问题：都将属性1视为令牌序列，并利用单词嵌入模型

2、预先培训的基于Transformer的方法：

1、预训练的LM模型能够通过微调应用在其它领域

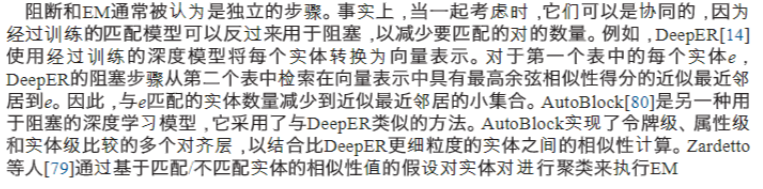
2、常用的预训练LMs：BERT及其变体

## 四、讨论：超越实体匹配

### （一）阻塞和实体匹配在一起

1、阻塞（blocking）是EM之前发生的一个重要步骤

2、阻塞的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



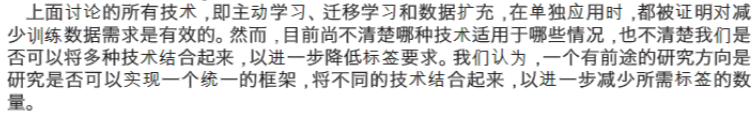
### （二）缓解培训数据需求

1、为了降低获得训练数据的成本，采用：主动学习、迁移学习和数据增强技术

1.1主动学习：一种选择可以改进当前模型的未标记示例的技术，因此人工注释器为所选示例分配标签来更新模型

1.2转移学习：一种将从源领域学到的知识转移到目标领域的技术，因为模型可以利用从源领域学到的知识。关键挑战是可迁移性

1.3数据扩充：一种利用显眼知识基于给定的训练数据集综合生成新训练示例的技术。关键挑战是如何找到并结合领域知识来创建有用的训练实例



### （三）匹配不同的实体

1、匹配不同数据格式描述的实体：关系表、半结构化或非结构化文本文档；匹配不一定相同的实体或不同类型的实体

2、实现更通用的EM解决方案的挑战

2.1构建用于评估一般EM任务的基准数据集：没有用于匹配不同实体或匹配不同数据格式的实体的基准数据集

2.2设计和开发一种有效的EM通道，支持一般的匹配条件：可以通过提供更多的训练示例，使用有效的预处理操作来仅保留用于匹配的基本信息，提高模型的整体性能

2.3开发新的模型来共同学习数据的结构和语义：对于结构化数据上的EM任务，可以将结构信息合并到令牌嵌入中；而对于半结构化数据更复杂

### （四）解释实体匹配的结果