2018 年以来自然语言处理前沿论文最新进展

2018.11.03 方建勇

提示：采用手机 safari 微软翻译技术

1. 第 1811: 347[pdf,其他] Cs。Cl

   how2hop: 一种用于多模态语言理解的大型数据集

   作者:ramon sanabria, ozan caglayan , shruti palaskar, desmondelliott , loic barrault, lucia 卑鄙, florian metze

   摘要: 本文介绍了 how2，这是一个多式联运教学视频集合，带有英文字幕和众包葡萄牙语翻译。我们还提供了机器翻译、自动语音识别、口语翻译和多模总结的综合序列到序列基线。通过为几种多模式**自然语言**任务提供数据和代码，我们希望促进对这些挑战和类似挑战的更多研究，从而更深入地了解**语言中的多模态处理**。少

   2018 年 11 月 1 日提交;最初宣布 2018 年 11 月。

2. 第 1811.00228[pdf,其他] Cs。简历

   一种注重图像字幕的顺序导引网络

   作者:daouda sow, zengchangqin , mouhamed niasse, tao wan

   文摘:计算机视觉 (cv) 和**自然**学习的最新进展。更多

   2018 年 11 月 1 日提交;最初宣布 2018 年 11 月。

   评论:5 页, 2 个数字, 1 个表, icassp 2019 年

3. 第 1811: 0025[pdf,其他] Cs。Cl

   用 svcca 理解语言模型的学习动态

   作者:naomi saphra, adam lopez

   摘要: 最近的研究表明，神经**语言**模型以多种方式隐式地编码语言结构。然而，现有的研究并没有揭示在培训期间获得这一结构的**过程**。我们使用 svcca 作为一种工具，用于理解**语言**模型是如何隐式预测各种单词群集标记的。我们提出的实验表明，**语言**模型的单个递归层分阶段学习语言结构。例如，我们发现,**语言**模型**比**学习语义和主题信息更早地稳定其对部分语音的表示。少

   2018 年 11 月 1 日提交;最初宣布 2018 年 11 月。

4. 第 1811.00196[pdf,其他] Cs。Cl

   实现可探索的 nlp: 文本分类的生成解释框架

   作者:刘辉,尹庆宇,王威廉

   摘要: 构建可解释的系统是**自然语言处理**(nlp) 领域的一个关键问题，因为大多数机器学习模型没有为预测提供任何解释。现有的可解释机器学习系统的方法往往侧重于解释输出或输入和输出之间的连接。然而，细粒度的信息往往被忽略，系统并没有显式生成人类可读的解释。为了更好地缓解这一问题，我们提出了一个新的生成解释框架，该框架学习进行分类决策，同时生成细粒度解释。更具体地说，我们介绍了可解释的因素和最低风险培训方法，学习生成更合理的解释。我们构建了两个新的数据集，其中包含摘要、评级分数和细粒度

原因。我们在两个数据集上进行实验，并与几个强神经网络基线系统进行比较。实验结果表明，该方法超过了两个数据集上的所有基线，能够同时生成简洁的解释。少

2018 年 10 月 31 日提交;最初宣布 2018 年 11 月。

5. 第 1810.12836[pdf,其他] Cs。CI

基于多任务二编码器模型的跨语言句子表示学习

作者:muthuraman chidambaram, yyfei yang, daniel cer, steve yuan, yun-hsuansung, brian st 横,ray kurzweil

摘要: 神经**语言**模型已被证明在许多**语言处理**任务上实现了令人印象深刻的性能水平。然而，由于其他**语言**提供的标记数据数量有限，这些模型大多仅限于对英文文本进行预测。克服这一问题的一个潜在方法是学习跨语言文本表示，这些表示可以用来将英语任务的培训转移到非英语任务，尽管很少或根本没有特定于任务的非英语数据。在本文中，我们探索了学习跨语言句子表示的**自然**设置：双编码器。我们对一些单语、跨语言和零尝试学习任务的跨语言表达进行全面评估，并对不同的学习跨语言嵌入空间进行分析。少

2018 年 10 月 30 日提交;最初宣布 2018 年 10 月。

6. 建议: 1810.12754[pdf,其他] Cs。Lg

经常性关注股

作者:钟国强,岳国华,肖玲

文摘: 递归神经网络 (rnn) 已成功地应用于许多序列学习问题中。如手写识别、图像描述、**自然语言处理**和视频运动分析。经过多年的发展，研究人员改进了 rnn 的内部结构，并引入了许多变种。除其他外，门控递推单元 (gru) 是使用最广泛的 rnn 模型之一。但是，gru 缺乏自适应关注某些区域或位置的能力，因此在倾斜过程中可能会导致信息冗余或丢失。本文提出了一种名为 "复发式注意单元 (rau)" 的 rnn 模型，该模型通过添加注意门，将注意机制无缝集成到 gru 内部。注意门可以提高 gru 记忆长期记忆的能力，帮助记忆细胞迅速丢弃不重要的内容。rau 能够通过自适应地选择一系列区域或位置来从序列数据中提取信息，并在学习过程中更加关注选定的区域。在图像分类、情感分类和**语言**建模方面的大量实验表明，rau 的性能一直优于 gru 和其他基线方法。少

2018 年 10 月 30 日提交;最初宣布 2018 年 10 月。

7. 建议: 1810.12738[pdf,其他] Cs。简历

文本发现中自然语言理解的视觉重新排序

作者:ahmed sabir, francesc Moreno-Noguer, lluís padró

摘要: 许多场景文本识别方法都是基于纯粹的视觉信息，而忽略了场景与文本之间的语义关系。本文从**自然语言处理**的角度来解决这一问题，以填补**语言**与视觉之间的空白。提出了一种**利用**词语的出现概率 (unigram **语言**模型) 以及场景与文本之间的语义相关性来提高场景文本识别精度的后处理方法。为此，我们最初依靠现成的深层神经网络，它已经接受了大量数据的训练，它为每个输入图像提供了一系列文本假设。然后使用单词频率和与图像中的对象或场景的语义相关性重新排列这些假设。由于这种组合，原始网络的性能得到了提升，几乎没有额外的成本。我们验证了我们对 icdare17 数据集的方法。少

2018 年 10 月 29 日提交;最初宣布 2018 年 10 月。

评论:accv 2018. arxiv 管理说明: 与 arxiv:1810.09776 有实质性的文本重叠

8. 第 1810.12446[pdf,其他] Cs。CI

用加减法双控循环网络简化神经机器翻译

作者:张彪,熊德义,苏金松,钱林, 张慧吉

摘要: 在本文中, 我们提出了一个加法减法双门递归网络 (atr), 以简化神经机器翻译。atr 的复发单元被大大简化, 在所有现有的门控 rnn 单元中具有最小数量的权重矩阵。通过简单的加减法运算, 我们引入了一种双门机制来建立高度相关的输入和忘记门。尽管进行了这种简化, 但仍保留了对远程依赖关系进行建模的基本非线性和能力。此外, 由于简化, 所提出的 atr 比 lstmsgu 更透明。在 atr 中可以很容易地建立正向自我关注, 这使得所提出的网络可以解释。对 wmt14 翻译任务的实验表明, 基于 atrr 的神经机器翻译可以在翻译质量和速度两方面对英语-德语和英语**-法语**进行竞争。nist 汉英翻译、**自然语言**推理和汉语分词的进一步实验验证了 atr 在不同**自然语言中的通用性和适用性处理**任务。少

2018 年 10 月 30 日提交;最初宣布 2018 年 10 月。

评论:emnlp 2018, 长纸, 源代码发布

9. **建议: 1810.12065[pdf, ps,其他] Cs。Lg**

训练递归神经网络收敛率的研究

作者:泽源·天珠, 李元志,赵松

摘要: 尽管深度学习取得了巨大的成功, 但我们对如何训练非凸神经网络的理解仍然相当有限。现有的大部分理论工作只处理具有一个隐藏层的神经网络, 而对于多层神经网络来说, 这种方法鲜为人知。递归神经网络 (rnn) 是一种特殊的多层神经网络, 广泛应用于**自然语言处理**应用中。与前馈网络相比, 它们特别难以分析, 因为权重参数在整个时间范围内被重用。可以说, 我们对训练 rnn 的收敛速度进行了初步的理论理解. 具体而言, 当神经元数量足够大的时候---即训练数据大小和时间范围内的多项式---以及权重为时随机初始化,

我们表明梯度下降和随机梯度下降都最大限度地减少训练损失在线性收敛速度, 即,大

1-1 1e- ω (t). 少

2018 年 10 月 29 日提交;最初宣布 2018 年 10 月。

10. **建议: 181011:04[pdf,其他] Cs。Cl**

机器人学习说 ' 不 ': 语言否定习得的禁止与拒绝机制

作者:frank förster, joe saunders, hagen lehmann, chrystopher l. nehaniv

摘要: "不" 属于儿童使用的前十个词, 体现了语言否定的第一个积极形式。尽管它的早期发生, 但其收购**过程**的细节在很大程度上仍然是未知的。"否" 不能被解释为可感知对象或事件的标签的情况使其超出了大多数现代**语言**习得帐户的范围。此外, 由于这个词的非参照性质, 大多数符号接地架构都很难接地。在一项涉及儿童类人形机器人 icub 的实验研究中, 该机器人被设计用来照亮否定词的获取**过程**, 该机器人被部署在几轮说话时不受约束的与天真的相互作用中参与者充当**其语言**教师。研究结果证实了影响或意志在社会分布**的获取过程**中起着关键作用的假设。否定词在禁止性话语和否定意图解释中是荒谬的突出, 因此它们很容易与老师的言语信号隔离开来。这些词随后可能以消极的情感状态为基础。然而, 对禁止行为的**性质**及其语言和语言外成分之间的时间关系的观察, 对希伯来语类型的算法是否适合**语言**提出了严重的问题接地。少

2018 年 10 月 28 日提交;最初宣布 2018 年 10 月。

评论:提交期刊的文章. 21 页的主要论文加 28 页补充资料/附录。

类:I.2.6;I.2.7;I.5.5;H.5。2

11. **第 1810.11227[pdf,其他] cs.PL**

合成对称镜头

作者:ansders miltner, solomon maina, kathleen fisher, benjamin c.皮尔斯, davidwalker, steve zdancewic

摘要: 镜头是既可以运行 "前后" 又可以运行的程序，可以在两个方向上转换数据和更新。自引进以来，镜片得到了广泛的研究和应用。最近的研究还展示了如何利用类型定向程序合成中的技术来有效地合成一个非常简单的透镜–所谓的超弦数据上的双透镜–给定一对类型 (正则表达式) 和一个小的的例子的数量。我们展示了如何将这种合成算法扩展到更宽的透镜类，我们称之为简单的对称透镜，包括双透镜和更广泛使用的 "非对称" 透镜，以及成熟的 "对称透镜" 的丰富子集。简单对称透镜具有独立的理论意义，是不依赖于持久内部状态的最大一类对称透镜。合成简单对称透镜比合成双透镜具有更大的挑战性: 由于两侧的某些信息可以与另一侧 "断开连接"，因此通常会有许多镜头与给定的示例一致。为了指导搜索**过程**，我们使用信息论中的随机正则表达式和思想来估计候选镜头传播的信息量。我们描述了简单对称透镜的实现和我们的合成过程作为 boomerang **语言**的扩展。我们在 48 个基准示例中评估其性能，这些基准示例来自 flash fill、augeus、双向编程文献和电子文件格式同步任务。经过适度的调整，这是由工具的交互**性鼓励**，我们的实现可以合成所有这些镜头在不到 30 秒。少

2018 年 10 月 26 日提交;最初宣布 2018 年 10 月。

12. **第十四条: 1810.11498[pdf, ps,其他] Cs。红外**

社交媒体宏观社区应急辅助设备的自动识别与排名

作者:bhaskar gautam, annappa basava

摘要: 包括推特在内的在线社交微博平台越来越多地用于协助灾害事件期间的救济行动。在大多数可能是**自然灾害**甚至武装袭击的灾难中，非政府组织寻找关于资源的关键信息，以支持受灾人民。尽管最近在使用深层神经网络的**自然语言处理**方面取得了进展，但短文本的检索和排名却成为一项具有挑战性的任务，因为大量的对话和同情内容与关键信息。在本文中，我们讨论了分类信息检索和大多数相关性信息的排名问题，同时考虑到在这些事件中出现的短文和多语言**的**存在。我们提出的模型是基于在文本和统计预处理的帮助下形成嵌入向量的基础上，最后利用前馈神经网络对整个训练的 2, 100, 000 个向量进行了归一化，以满足需求和可用性推特。本文的另一个重要贡献在于基于前五个一般术语的新加权排序键算法，该算法对分类推特进行了与分类最相关的排序。最后，我们在尼泊尔地震数据集上测试了我们的模型 (包含简短的文本和多语言的推特)，并在 525 万未标记的救灾推特嵌入向量上实现了 6.81 的平均平均精度。少

2018 年 10 月 26 日提交;最初宣布 2018 年 10 月。

评论:irmidis track 2017-信息检索评估论坛的工作说明，印度班加罗尔, 2017 年 12 月 8 日至 10 日, http://ceur-ws.org/Vol-2036

13. **第 1810.11190[pdf,其他] Cs。CI**

规模: 一种快速、高效的通用矢量嵌入实用程序包

作者:ajay patel,亚历山大 sands, chris callison-burch, marianna apidianaki

摘要: 矢量空间嵌入模型，如 word2vec、环球、快速文本和 elmo，是**自然语言处理**(nlp) 应用中非常流行的表现形式。我们介绍了 "规模"，这是一种快速、轻量级的工具，用于利用和**处理**嵌入。重要性是一个开源的 python 包，具有紧凑的矢量存储文件格式，允许高效地操作大量的嵌入。规模执行常见操作的速度比 gensim 快 60 到 6, 000 倍。重要性引入了几个新功能，以提高鲁棒性，如词汇外查找。少

2018 年 10 月 26 日提交;最初宣布 2018 年 10 月。

14. **建议: 1810.11066[pdf,其他] Cs。Lg**

自动化低精度深度学习运算符的生成

作者:meghan cowan, thierry moreau, tiqi chen, luis cze

摘要: 最先进的深度学习模型在计算机视觉和**自然语言处理**领域取得了稳步进展，牺牲了不断增长的模型大小和计算复杂性。在低功耗和移动设备上部署这些模型是一个挑战，因为它们的计算能力有限，能源预算严格。一个引起了重大研究兴趣的解决方案是部署高度量化的模型，这些模型在低精度输入上运行，重量小于 8 位，从而降低了准确性的性能。这些模型显著减少了内存占用 (最多减少 32x)，并且可以在计算密集型卷积和完全连接的层期间用按位运算取代多累积。大多数深度学习框架依赖于高度工程化的线性代数库，如 atlas 或英特尔的 mkl 来实现高效的深度学习运算。到目前为止，流行的深度学习都不直接支持低精度运算符，部分原因是缺乏优化的低精度库。本文介绍了一种工作流程，以快速生成高性能低精度深度学习运算符，实现任意精度，针对多个 cpu 体系结构，并包括内存平铺和矢量化等优化。我们提供了一个关于低功耗 arm cortex-a53 cpu 的广泛案例研究，并展示了如何在优化的 16 位整数基线上生成 1 位、2 位卷积，速度高达 16x，比手写实现好 2.3 倍。少

2018 年 10 月 25 日提交;最初宣布 2018 年 10 月。

评论:10 页, 11 位数字

15. **建议: 1810.10927[pdf,其他] Cs。Cl**

自然语言处理中的贝叶斯压缩

作者:nadezhda chirkova, ekaterina lobacheva, dmitry vetrov

文摘: 在**自然语言处理**中，很多任务都是用递归神经网络成功解决的，但这样的模型有大量的参数。这些参数中的大多数通常集中在嵌入层，其大小会随着词汇长度的增长而增长。我们提出了一种用于 rnn 的贝叶斯稀疏技术，该技术允许压缩 rnn 数十次或数百次，而无需耗时的超参数调整。我们还推广了词汇稀疏模型，以过滤掉不必要的单词，并进一步压缩 rnn。我们表明，保留词的选择是可以解释的。少

2018 年 10 月 25 日提交;最初宣布 2018 年 10 月。

评论:发布于 emnlp 2018

16. **建议: 1810.1002[pdf,其他] Cs。Cl**

用神经网络处理序列映射问题的序列处理

作者:雷宇

摘要: 在**自然语言处理**(nlp) 中，检测两个序列之间的关系或生成给定另一个观察序列的令牌序列非常重要。我们将建模序列对上的问题类型称为序列到序列 (seq2seq) 映射问题。已经进行了大量的研究，以找到解决这些问题的方法，传统的方法依赖于手工制作的功能、对齐模型、分段启发式和外部语言资源的组合。虽然取得了很大的进展，但这些传统的方法还存在着管道复杂、特征工程费力、域适应困难等诸多弊端。近年来，神经网络成为解决 nlp、语音识别和计算机视觉等诸多问题的一个很有希望的办法。神经模型是强大的，因为它们可以端到端训练，很好地概括到看不见的例子，同样的框架可以很容易地适应一个新的领域。本论文的目的是利用神经网络推进 seq2seq 映射问题的最先进。我们从三个主要方面探索解决方案: 调查用于表示序列的神经模型、模拟序列之间的交互以及使用未配

对的数据来提高神经模型的性能。对于各个方面，我们提出了新的模型，并评估了它们在 seq2seq 映射的各种任务中的有效性。少

2018 年 10 月 25 日提交;最初宣布 2018 年 10 月。

评论:博士论文

17. 第 1810.752[pdf,其他] Cs。CI

基于 word 嵌入的编辑距离

作者:牛一林,赵桥,李航,黄敏丽

摘要: 文本相似度计算是**自然语言处理**和相关领域的一个基本问题。近年来，为了完成这项任务，开发了深度神经网络，取得了很高的性能。神经网络通常在监督学习中使用标记数据进行训练，创建标记数据通常成本很高。在本文中，我们讨论了文本相似度计算的无监督学习问题。我们提出了一种新的方法，称为基于 word 嵌入的编辑距离 (wed)，它将嵌入字集成到编辑距离中。对三个基准数据集的实验表明, wed 的性能优于最先进的无监督方法，包括编辑距离、基于 tf-idf 的余弦、基于余弦的词嵌入、jacard 索引等。

2018 年 10 月 25 日提交;最初宣布 2018 年 10 月。

18. 建议: 1810.10641[pdf,其他] Cs。CI

预测与暹罗 cnn 和 lstm 的语义文本相似性

作者:elvys linhares pontes, stphane huet, andréa carneiro linhares, juan-manuel torres-moreno

摘要: 语义文本相似性 (sts) 是**自然语言处理**(nlp) 中许多应用的基础。我们的系统结合卷积和递归神经网络来测量句子的语义相似性。它使用卷积网络来考虑单词的局部背景，使用 lstm 来考虑句子的全局上下文。这种网络的组合有助于保存句子的相关信息，并改进句子之间相似性的计算。我们的型号取得了良好的效果，并与最好的最先进的系统竞争。少

2018 年 10 月 24 日提交;最初宣布 2018 年 10 月。

19. 建议: 1810.10401[pdf,其他] Cs。CI

基于图像的基于自然语言的三维构想神经网络

作者:erinc merdivan , anastasios vafeiadis, dimitrios kalatzis, sten henke, johes kropf, konstantinos votis, dimitrios giakoumis, dimitrios tzovaras, l 葛 ingchen, raouf hamzaoui, matthieu geist

文摘: 我们提出了一种新的 **自然语言**理解方法，我们将输入文本视为图像，并应用二维卷积神经网络从视觉的变化中学习句子的局部和全局语义单词的模式。我们的方法证明，在不使用光学字符识别和顺序**处理**管道的情况下，可以从带有文本的图像中获取语义上有意义的特征，而不使用传统**的自然技术。语言**理解算法需要。为了验证我们的方法，我们提供了两个应用程序的结果: 文本分类和对话框建模。利用二维卷积神经网络，我们能够超越非拉丁字母文本分类的最新精度结果，并在 8 个文本分类数据集中取得了很有希望的结果。此外，我们的方法在使用 babi 对话框数据集的任务 4 中的词汇实体时的性能优于内存网络。少

2018 年 10 月 24 日提交;最初宣布 2018 年 10 月。

评论:自然语言处理 (nlp)，情感分析，对话建模

20. 建议: 1810.09849[pdf,其他] Cs。简历

下拉滤波器：用于卷积的下拉器

作者:陈建伟田志素

文摘: 利用大量的参数，深度神经网络在计算机粘滞和**自然语言处理**任务上取得了显著的性能。然而，网络通常会因为使用过多的参数而受到过度拟合的影响。辍学是一种广泛使用的处理超拟合的方法。尽管在神经网络中，辍学率可以显著地规范密集连接的层，但在使用卷积层时，会导致次优结果。为了跟踪这个问题，我们提出了一种新的卷积层的辍学方法–dropfilter。dropfilter 随机抑制某些过滤器的输出。因为人们观察到，共同适应更有可能发生在内部滤波器中，而不是在卷积层中发生内部滤波器。利用 dropfilter，显著提高了 cifar 和 imagenet 上卷积网络的性能。少

2018 年 10 月 23 日提交;最初宣布 2018 年 10 月。

21. **建议: 1810.0945[pdf, ps,其他] cs. cy**

学生教育过程的可扩展性、灵活性

作者:bhairav mehta, Adithya ramanathan

摘要: 我们提出了一个新的智能辅导系统，它建立在教育心理学的既定假设基础上，并将其集成到一个可扩展的软件体系结构中。具体而言，我们基于知识发声、并行学习和学生学习背景下的即时反馈的已知好处。我们表明，开源数据与深度学习和**自然语言处理**中最先进的技术相结合，可以在规模上应用这三个因素的好处，同时仍然以个人学生的需求和建议。此外，我们允许教师保持对算法输出的完全控制，并提供学生统计数据，以帮助更好地指导课堂讨论，探讨将受益于更多面对面审查和覆盖的主题。我们的实验和试点项目显示了有希望的结果，并巩固了我们的假设，即系统足够灵活，可以在课堂和课堂环境中满足各种用途。少

2018 年 10 月 17 日提交;最初宣布 2018 年 10 月。

评论:提交给 2018 年 NIPS ai 社会公益讲习班

22. **建议: 1810.09717[pdf,其他] Cs。Lg**

没有人有时间进行编码：从自然语言综合结构感知程序

作者:jakub bednarek, karol piaskowski, krzysztof kwwiec

文摘: 从**自然语言**(nl) 进行程序合成对人类是实用的，一旦技术上可行，将显著促进软件开发并使最终用户编程发生革命性的变化。我们提出了 saps，一个端到端神经网络，能够映射相对复杂的多句子 nl 规范到可执行代码的片段。该体系结构完全依赖于神经组件，并建立在在抽象语法树训练的树 2 树自动编码器上，并结合预先训练的单词嵌入和用于 nl **处理**的双向多层 lstm。该解码器具有一种新的信号传播方案和软注意机制的双循环 lstm。当应用于以前研究中提出的大量问题数据集时，saps 的性能与在那里提出的方法相当或更好，在 90% 以上的情况下生成正确的程序。与其他方法不同的是，它不涉及任何非神经组件来**处理**生成的程序，并使用固定维的潜在表示作为 nl 分析器和源代码生成器之间的唯一链接。少

2018 年 10 月 23 日提交;最初宣布 2018 年 10 月。

23. **第: 1810.9683[pdf,其他] Cs。Lg**

基于图嵌入神经网络的二元相似性无监督特征提取

作者:roberto baldoni, giuseppe antonio di luna, luca masrelli, fabio petroni, leonardo querzoni

摘要: 本文考虑二进制相似问题，该问题包括确定两个二进制函数是否相似，只考虑它们的编译形式。在多个应用程序场景中，例如版权争议、恶意软件分析、漏洞检测等，此问题是至关重要的。目前该领域最先进的解决方案通过创建一个嵌入模型来工作，该模型将二进制函数映射到向量中。R^n.这种嵌入模型捕获二进制文件之间的句法和语义相似性，即类似的二进制函数映射到向量空间中接近的点。此策略有许多优点，其中之一是可以预先计算多个二进制函数的嵌入，然后将它们与简单的几何运算 (例如，点积) 进行比较。在 [32] 中，函数首先在由手动设计的特征构成的注释控制流图 (acfg) 中转换，然后使用深层神经网络体系结构将图形嵌入到向量中。在本文中，我们提出并测试了几种计算带注释的控制流图的方法，这些方法使用无监督的方法进行特征学习，而不会产生人为的偏差。我们的方法是**在自然语言处理社区中**使用的技术之后产生启发的 (例如，我们使用 word2vec 来编码程序集指令)。我们表明，我们的方法确实是成功的，它带来了比以前最先进的解决方案更好的性能。此外，我们还报告了对功能嵌入的定性分析。我们发现了一些有趣的情况，在这些案例中，嵌入是根据原始二进制函数的语义聚集在一起的。少

2018 年 10 月 23 日提交;最初宣布 2018 年 10 月。

24. **第: 1810.9648[pdf,其他] Cs。艾**

ai 可以为我做什么: 在合作游戏中评估机器学习解释

作者:石峰,乔丹·博伊德

摘要: 机器学习是决策的重要工具，但其合乎道德和负责任的应用需要严格审查其可解释性和实用性: 这是一个研究不足的问题，特别是**在自然语言方面处理**模型。我们为回答问题的任务设计了一个特定于任务的评估，并评估模型解释在人机协作环境中如何提高人类的性能。我们在一个接地气、逼真的环境中评估解释方法: 作为一个团队进行一场琐事游戏。我们还为**自然语言处理**人环设置提供设计指导。少

2018 年 10 月 24 日提交;v1 于 2018 年 10 月 22 日提交;最初宣布 2018 年 10 月。

25. **建议: 1810.09 580[pdf,其他] Cs。CI**

多伊10.1109/IJCNN.2018.8489656

一种完全基于关注的信息检索

作者:alvaro henrique chaim correia,豪尔赫·路易斯·莫雷拉·席尔瓦, thiago de castro martins, fabio gagliardi cozman

摘要: 递归神经网络现在是**自然语言处理**中最先进的，因为它们可以构建丰富的上下文表示和**任意**长度的处理文本。然而，最近在注意机制方面的发展为前馈网络提供了类似的能力，因此，由于可以并行化的操作数量增加，因此能够更快地计算。我们在回答问题的领域中探索了这种新型的体系结构，并提出了一种新的方法，称为基于完全注意的信息检索 (fair)。我们表明, fabir 在斯坦福问答数据集 (squad) 中取得了有竞争力的结果，同时参数较少，并且在学习和推理方面比竞争对手的方法更快。少

2018 年 10 月 22 日提交;最初宣布 2018 年 10 月。

评论:可在 2018 年神经网络问题国际联席会议上发言

日记本参考:a. h. c. correia, j. l. m. silva, t. d. c. martins and f. g. cozman, "完全关注的信息检索", 2018 年神经网络问题国际联席会议，巴西里约热内卢, 2018 年,第 27-2806 页

26. **建议: 1810.09506[pdf] Cs。CI**

在微博上自动检测自报告的出生缺陷结果，进行大规模流行病学研究

作者:ari z. klein, abed sarker, davy weissenbacher, Graciela gonzalez-hernandez

摘要: 在最近的工作中，我们确定并研究了一小群推特用户，他们的怀孕有出生缺陷的结果可以通过他们公开发布的推特观察到。利用社交媒体的大规模潜力来补充研究出生缺陷的有限方法，而出生缺陷是婴儿死亡的主要原因，取决于自动方法的进一步发展。本研究的主要目的是采取第一步，扩大使用社交媒体观察有出生缺陷结局的妊娠的范围，即开发报告其出生缺陷的用户自动检测推特的方法结果。我们对大约 23，000 条提及先天缺陷的推特进行了注释和预处理，以便培训和评估受监督的机器学习算法，包括功能工程和基于深度学习的分类器。我们还试验了各种采样不足和过采样的方法，以解决类不平衡问题。在原始不平衡数据集上训练的支持向量机 (svm) 分类器 (n 克、词簇和结构特征) 实现了正类的最佳基线性能: "缺陷" 类的 f1-分数为 0.65 分，"缺陷" 类的 f1 分数为 0.65 分"可能的缺陷" 类。我们的贡献包括 (i)**自然语言处理**(nlp) 和受监督的机器学习方法，用于自动检测报告其出生缺陷结果的用户的推特; (ii)对不平衡、采样不足和过度采样数据进行了功能设计和基于深度学习的分类器培训，以及 (iii) 错误分析，可使用我们公开的语料库为分类改进提供信息。今后的工作将侧重于自动化用户级分析，以实现队列包含。少

2018 年 10 月 22 日提交;最初宣布 2018 年 10 月。

27. **建议: 1810.09431**[**pdf**, **ps**,**其他**] **Cs。CI**

主动安全: 用于暴力和滥用语音识别的嵌入式 ai 解决方案

作者:christopher dane shulby, leonardo pombal, vitor j 若尔多, guilherme ziolle, bruno martho, antnio 邮局, thiago prochnow

摘要: 暴力在巴西是一种流行病，在全世界都在上升。移动设备提供的通信技术可用于监测和警惕暴力局势。然而，目前的解决方案，如恐慌按钮或安全话语，可能会增加暴力情况下的生命损失。我们提出了一个嵌入式人工智能解决方案，使用**自然语言**和语音**处理**技术，默默提醒在这种情况下可以帮助的人。使用的语料库包含 400 个阳性短语和 800 个否定短语，总共 1200 句，使用两种众所周知的提取方法进行自然 **语言处理**任务: 单词袋和单词袋嵌入并使用支持向量机进行分类。我们描述了正在开发的概念验证产品，并有很好的效果，指出了通往商业产品的道路。更重要的是，我们表明，通过单词嵌入和数据增强技术进行模型改进提供了一个本质上可靠的模型。最终的嵌入式解决方案的占用空间也小于 10 mb。少

2018 年 10 月 22 日提交;最初宣布 2018 年 10 月。

评论:6 页, bracis 2018 预印

28. **建议: 1810.09302**[**pdf**, **ps**,**其他**] **Cs。CI**

生物 senmvec: 为生物医学文本创建句子嵌入

作者:陈庆宇,彭一凡,陆志勇

摘要: 句子嵌入已成为当今**自然语言处理**(nlp) 系统的重要组成部分，尤其是结合先进的深度学习方法。尽管预先训练的句子编码器在一般领域可用，但到目前为止，生物医学文本中还没有任何编码器。在这项工作中，我们介绍了 bisennvec: 第一套开放的句子嵌入培训超过 3，000 万份文件，从两个学术文章在 pubmed 和临床笔记在 mimic-iii 临床数据库。我们评估了 bisentvec 嵌入在两个句子对相似任务在不同的文本类型。我们的基准测试结果表明，与其他竞争替代方案相比，biosentvec 嵌入可以更好地捕获句子语义，并在这两项任务中实现最先进的性能。我们期望 biosentvec 能促进生物医学文本挖掘的研究和开发，并补充生物医学词嵌入的现有资源。生物传感器以
https://github.com/ncbi-nlp/BioSentVec 的价格公开提供

2018 年 10 月 26 日提交;v1 于 2018 年 10 月 22 日提交;最初宣布 2018 年 10 月。

评论:4 页, 1 张表格

29. 第 1810.09230[pdf,其他] cse

多伊10.114/3243434.32784996

基于 asp 的恶意电源检测深度学习

作者:gili rusak, abdullah al-dujaili, una-may o ' reilly

摘要: 随着深度学习的著名成功，一些尝试开发有效的方法来检测恶意 powershell 程序使用神经网络在传统**的自然语言处理**设置，而另一些使用卷积神经网络，用于在字符级别检测模糊的恶意命令。虽然这些表示可以表达 powershell 的突出属性，但我们的假设是，来自静态程序分析的工具会更有效。我们提出了一种混合方法，结合了传统的程序分析 (以抽象语法树的形式) 和深度学习。这张海报介绍了我们方法中一个基本步骤的初步结果: 学习 powershell ast 节点的嵌入。我们按家族类型对恶意脚本进行分类，并探索嵌入式程序矢量表示。少

2018 年 10 月 3 日提交;最初宣布 2018 年 10 月。

评论:出席含石棉 ccs 2018 年海报会议

30. 建议: 1810.09154[pdf,其他] Cs。Cl

一种用于对话行为分类的双注意层次递归神经网络

作者:李瑞哲、林成华、马修·科林森、小李、陈冠义

摘要: 识别对话行为 (da) 对于许多**自然语言处理**任务 (如对话生成和意图识别) 都很重要。本文提出了一种用于对话行为分类的双关注分层递归神经网络。我们的模式部分是受这样一种观察的启发，即谈话话语通常既与对话行为有关，也与主题有关，前者抓住了社会行为，后者描述了主题。然而，大多数现有的发展议程分类系统都没有利用对话行为与专题之间的这种依赖关系。有了新的针对任务的双重关注机制，我们的模型就能在话语中捕捉有关对话行为和话题的信息，以及关于它们之间互动的信息。我们评估模型在两个公开可用的数据集 (即 "交换机" 和 "每日对话") 上的性能。实验结果表明，通过将建模主题作为辅助任务，我们的模型可以显著提高 da 分类。少

2018 年 10 月 22 日提交;最初宣布 2018 年 10 月。

评论:8 页, 5 个数字

31. 建议: 1810.08899[pdf,其他] Cs。Lg

压缩，或不压缩: 为嵌入式推理描述深度学习模型压缩

作者:秦青,任杰,余家龙, 高玲,王海, 郑洁,冯燕松, 方建斌,王正

摘要: 深度神经网络 (dnn) 的最新进展使其对嵌入式系统具有吸引力。但是, dnn 可能需要很长时间才能在资源受限的计算设备上进行推理。模型压缩技术可以解决嵌入式设备深度推理的计算问题。此技术非常有吸引力，因为它不依赖于专门的硬件，也不依赖于由于隐私问题或高延迟而通常不可行的计算卸载。然而，目前仍不清楚模型压缩技术如何在广泛的 dnn 中执行。要设计高效的嵌入式深度学习解决方案，我们需要了解他们的行为。本文开发了一种定量方法来描述具有代表性的嵌入式深度学习体系结构 nvidia jetson tx2 上的模型压缩技术。我们从图像分类和**自然语言处理**领域考虑了 11 个有影响的神经网络架构，进行了广泛的实验。我们实验证明，数据量化和修剪这两种主流压缩技术是如何在这些网络体系结构上进行的，以及压缩技术对模型存储大小、推理时间、能耗和性能指标。我们证明了在嵌入式系统上实现快速深度推理的机会，但必须仔细选择压缩设置。我们的研究结果为何时以及如何应用模型压缩技术以及设计高效嵌入式深度学习系统的指南提供了见解。少

2018 年 10 月 21 日提交;最初宣布 2018 年 10 月。

评论:8 页，将出现在 ispa 2018

32. 建议: 1810.08838[pdf,其他] Cs。CI

应用注意神经技术的摘要

作者:jacob krantz, jugal kalita

摘要: 在一个数据激增的世界里，迅速总结文本的能力越来越重要。文本的自动总结可以看作是序列问题的序列。**自然语言处理**的另一个领域是机器翻译，由于基于注意的编码器解码器网络的发展，机器翻译正在迅速发展。这项工作将这些现代技术应用于抽象摘要。我们对各种关注机制进行了分析，以总结，目的是开发一种旨在提高艺术状态的方法和架构。特别是，我们修改和优化了一个自我关注的翻译模型，以生成抽象的句子摘要。在标准化评估集和测试指标的背景下，对该基本模型的有效性以及关注变量进行了比较和分析。然而，我们表明，这些指标在有效地对抽象摘要进行评分的能力方面是有限的，并根据抽象模型需要抽象评价的直觉提出了一种新的方法。少

2018 年 10 月 20 日提交;最初宣布 2018 年 10 月。

评论:可在第十五届自然语言处理国际会议 (icon 2018) 上进行口头陈述

33. 建议: 1810.0802[pdf,其他] Cs。CI

使用大纲生成分层文本

作者:mehdi drissi, olivia watkins, jugal kalita

文摘: 自然语言处理中的许多挑战需要生成文本，包括**语言**翻译、对话生成和语音识别。对于所有这些问题，随着文本变得更长，文本生成变得更加困难。目前的**语言**模型往往难以跟踪长一段文字的一致性。在这里，我们尝试建立模型构造，并使用它生成的文本的大纲来保持其焦点。我们发现大纲的使用改善了困惑。我们没有发现使用大纲可以改善人类对更简单基线的评价，从而揭示出困惑和人类感知的差异。同样，没有发现分层生成可以提高人类评价分数。少

2018 年 10 月 20 日提交;最初宣布 2018 年 10 月。

评论:8 页，参加自然语言处理国际会议

34. 第 1810.08436[pdf,其他] Cs。CI

高效的依赖引导命名实体识别

作者:张明杰, aldrian obaja muis, wei lu

摘要: 命名实体识别 (ner) 侧重于从文本中提取语义上有意义的命名实体及其语义类，是几种下游**自然语言不可或缺的组成部分处理**(nlp) 任务，如关系提取和事件提取。另一方面，依赖树也传递关键的语义级别信息。此前已经证明，这些信息可用于提高 ner 的性能 (sasano 和 kurohashi, 2008 年; ling and weld, 2012 年)。在本工作中，我们研究如何更好地利用依赖树传递的结构化信息来提高 ner 的性能。具体而言，与仅利用依赖关系信息来设计本地功能的现有方法不同，我们表明，在构建 ner 模型时，可以利用依赖树的某些全局结构化信息，在这些模型中，依赖关系树可以利用这些信息。提供有指导的学习和推理。通过大量的实验，我们证明了我们提出的新的依赖引导的 ner 模型与基于传统半马尔可夫条件随机场的模型具有竞争力，同时要求的运行时间明显较低。少

2018 年 10 月 22 日提交;v1 于 2018 年 10 月 19 日提交;最初宣布 2018 年 10 月。

评论:8 + 1 页，补充 9 页。发表于第 31 届阿拉伯人工智能会议 (阿拉伯人工智能 2017)。此版本修复了两个公式中的错误。arxiv 管理说明: 文本重叠与 arxiv:1711.07010 由其他作者

35. 第 1810.08305[pdf,其他] Cs。Lg

具有图形结构化缓存的源代码开放词汇学习

作者:milan cvitkovic, badal singh, anima anandkumar

摘要: 以计算机程序源代码为输入的机器学习模型通常使用**自然语言处理**(nlp) 技术。然而，一个主要的挑战是，代码是使用开放的、快速变化的词汇来编写的，例如，由于新变量和方法名称的造币。对这样的词汇进行推理并不是大多数 nlp 方法都是为之设计的。我们引入了图形结构缓存来解决此问题;此缓存包含模型遇到的每个新单词的节点，这些单词的边缘将每个单词连接到代码中出现的单词。我们发现，将这种图形结构化缓存策略与最近基于图形神经网络的代码模型相结合，可在代码上进行监督学习，从而提高模型在代码完成任务和变量命名任务上的性能----具有 100% 以上的相对性对后者的改进----以计算时间适度增加为代价。少

2018 年 10 月 18 日提交;最初宣布 2018 年 10 月。

36. 第 1810.0136[pdf,其他] Cs。铬

基于对流神经网络的语义空间文本动力学分析

作者:杨忠良,南伟，军义生,黄永峰,张玉进

文摘 隐身分析一直是网络安全领域的一个重要研究课题，有助于识别公共网络中的隐蔽攻击。近两年来，**随着自然语言处理**技术的快速发展，无覆盖隐写技术得到了极大的发展。以前的文本隐身分析方法在这种新的隐写技术上显示出不令人满意的结果，仍然是一个尚未解决的挑战。与以往的文本隐身分析方法不同，本文提出了一种基于语义分析的文本隐身分析方法 (ts-cnn)，该方法利用卷积神经网络 (cnn) 提取文本的高级语义特征，并发现其微妙之处。嵌入秘密信息前后语义空间的分布差异。为了训练和测试所提出的模型，我们收集并发布了一个大型文本隐身分析 (ct-steg) 数据集，其中包含总共超过 216,000, 000 个具有不同长度和不同嵌入速率的文本。实验结果表明，该模型能实现近 100% 的精度和召回率，优于以往的所有方法。此外，该模型甚至可以估计内部隐藏信息的容量。这些结果有力地支持了利用秘密信息嵌入前后语义空间的细微变化进行文本隐身分析是可行和有效的。少

2018 年 10 月 18 日提交;最初宣布 2018 年 10 月。

评论:提交给 aaai2019

37. 建议: 1810.07411[pdf,其他] cs. ne

基于局部对齐分布表示的重复神经体系结构的在线学习

作者:亚历山大·奥罗比亚,安库尔马里, c. lee giles, daniel kifer

摘要: 基于递归神经网络的时间模型已被证明在语言建模和语音**处理**等各种应用中相当强大。然而，要训练这些模型，就需要依靠时间的反向传播，这就需要在许多时间步骤中展开网络，从而使进行信贷分配**的过程**更具挑战性。此外，反向传播本身的**性质**不允许使用不可微分的激活函数，并且本质上是顺序的，这使得基础训练**过程**的并行化非常困难。在这项工作中，我们提出了平行时间神经编码网络，这是一个由本地学习算法训练的生物启发模型，称为局部表示对齐，旨在解决困扰重复网络的困难和问题通过时间的反向传播训练。最值得注意的是，这种体系结构既不需要展开，也不需要其内

部激活函数的导数。我们将我们的模型和学习过程与其他在线反向传播时间替代方案进行比较 (这也往往是计算成本高昂的), 包括实时经常性学习、回声状态网络和无偏见的在线重复优化, 并表明它在序列建模基准上的表现优于它们, 例如弹跳 mnist, 我们称之为 "弹跳 notist" 和 penn treebank 的新基准。值得注意的是, 在某些情况下, 我们的方法甚至可以通过时间本身以及稀疏专心回溯等变体超越完全反向传播。此外, 我们提出了有希望的实验结果, 证明了我们的模型进行零镜头适应的能力。少

2018 年 10 月 17 日提交;最初宣布 2018 年 10 月。

评论:即将取得一些其他成果

38. **第: 1810.07091[pdf,其他] Cs。Cl**

信息: 学习文本表示的开源框架

作者:ahmad tae, raphael rubino , josef van genabith

摘要: 表示学习方法的出现使各种**语言**任务的性能大大提高, 从而减少了对手动功能工程的需求。虽然工程表示通常基于某些语言理解, 因此更容易解释, 而博学的表示法则更难解释。从经验上研究这两种方法的互补性可以提供更多的语言见解, 有助于在可解释性和性能之间达成更好的妥协。我们提出了一个框架, 研究学习和工程的文本表示在文本分类任务的背景下。它旨在简化功能工程的任务, 并为提取学习的特征和结合这两种方法提供基础。inodens 具有灵活性、可扩展性, 学习曲线较短, 并且易于与许多可用且广泛使用的**自然语言处理**工具集成。少

2018 年 10 月 16 日提交;最初宣布 2018 年 10 月。

39. **修订: 1810.0825[pdf,其他] Cs。Lg**

用于稀疏数据的快速随机 pca

作者:徐峰,谢玉阳,宋明业,于文健,唐杰

文摘: 主成分分析 (pca) 广泛应用于社会网络分析、信息检索和**自然语言处理**等实际数据的维数约简和嵌入。在本文中, 我们提出了一种快速随机 pca 算法来**处理**大量稀疏数据。该算法与基本随机 svd (rpca) 算法 (halko 等人, 2011 年) 具有相似的准确性, 但在很大程度上针对稀疏数据进行了优化。它还具有很好的灵活性, 可以将运行时与精度进行权衡, 以实现实际使用。实际数据实验表明, 该算法比基本的 rpca 算法快 9.1X, 精度损失大, 比 matlab 中的 svds 快 20x, 误差小。该算法在 24 核计算机上不到 400 秒的时间内计算具有 12, 889, 521 人的大型信息检索数据的前 100 个主要组件, 而所有常规方法都由于内存不足问题而失败。少

2018 年 10 月 16 日提交;最初宣布 2018 年 10 月。

评论:16 页, 接受 acml2018

40. **建议: 1810.06640[pdf,其他] Cs。Cl**

不强化学习的对抗性文本生成

作者:david donahue, anna rumshisky

摘要: 生成对抗性网络 (gans) 最近经历了人气的激增, 在各种任务中具有竞争力, 尤其是在计算机视觉方面。然而, gan 培训在**自然语言处理**方面的成功有限。这主要是因为文本序列是离散的, 因此渐变不能从鉴别器传播到生成器。最近的解决方案使用强化学习将近似梯度传播到发电机, 但这是低效的训练。我们建议使用自动编码器来学习句子的低维表示。然后训练 gan 在这个空间中生成自己的向量, 从而解码为现实的话语。我们报告了来自生成器的随机样本和插值样本。句子向量的可视化表明我们的模型正

确地学习了自动编码器的潜在空间。人的评分和 BLEU 分数都表明，我们的模型根据竞争基线生成逼真的文本。少

2018 年 10 月 11 日提交;最初宣布 2018 年 10 月。

评论:四页，不带引用。acl 乳胶风格。四位数字

41. **建议: 1810.06638[pdf,其他] Cs。Cl**

u-net: 无答案问题的机器阅读理解

作者:孙富阳,李林阳,邱锡鹏,刘阳

摘要: 无答案问题的机器阅读理解是自然**语言处理中一**项新的具有挑战性的任务。一个关键的子任务是可靠地预测问题是否无法回答。在本文中，我们提出了一个称为 u-net 的统一模型，它有三个重要的组成部分: 应答指针、无应答指针和答案验证器。我们引入一个通用节点，从而将问题及其上下文段落作为一个连续的令牌序列**进行处理**。通用节点对问题和段落中的融合信息进行编码，对预测问题是否可回答起重要作用，也大大提高了 u-net 的简洁性。与最先进的管道模型不同，u-net 可以以端到端的方式学习。在 squad 2.0 数据集上的实验结果表明，u-net 能够有效地预测问题的不可回答性，并在 squad 2.0 上达到 f1 的评分 71.7。少

2018 年 10 月 12 日提交;最初宣布 2018 年 10 月。

评论:9 页

42. **建议: 1810.06599[pdf, ps,其他] cse**

多伊 10.114/3283812.3283822

使用 cmg 从源代码生成注释

作者:sergey matskevich, colin s. gordon

摘要: 好的注释可以帮助开发人员更快地理解软件并提供更好的维护。然而，评论往往缺失，一般不准确，或过时。其中许多问题可以通过自动生成注释来避免。本文提出了一种利用**自然语言处理**中的通用技术直接从源代码生成信息性注释的方法。我们使用现有**的自然语言**模型生成注释，该模型将单词与其各自的逻辑含义和语法规则结合起来，允许通过从程序文本的声明性描述中进行注释生成。我们在 python 中实现的几个经典算法上评估我们的算法。少

2018 年 10 月 15 日提交;最初宣布 2018 年 10 月。

评论:nl4se 2018 预印

43. **第 1810.06 339[pdf,其他] Cs。Lg**

深层强化学习

作者:李玉熙

摘要: 我们讨论了深度强化学习的概述风格。我们画了一幅大图，充满了细节。我们讨论了六个核心要素、六个重要机制和 12 个应用，重点是当代工作和历史背景。我们从人工智能、机器学习、深度学习和强化学习 (rl) 的背景开始，从资源开始。接下来我们讨论 rl 的核心要素，包括价值函数、策略、奖励、模型、勘探与开发和表示。然后讨论了 rl 的重要机制，包括注意力和记忆、无监督学习、分层 rl、多代理 rl、关系 rl 和学习学习。之后，我们将讨论 rl 应用，包括游戏、机器人技术、**自然语言处理**(nlp)、计算机视觉、金融、商业管理、医疗保健、教育、能源、交通、计算机系统、科学、工程和艺术。最后简要总结，讨论挑战和机遇，最后结语。少

2018 年 10 月 15 日提交;最初宣布 2018 年 10 月。

评论:摩根大通和克莱浦: 人工智能和机器学习综合讲座

44. 第 1810.06245[pdf,其他] Cs。CI

将简单和轻盈带回神经图像字幕

作者:jean-benoit delbrouck, stphane dupont

文摘: 神经图像字幕 (nic) 或神经字幕的生成在过去几年中引起了人们的广泛关注。用**自然语言**描述图像一直是计算机视觉和**语言处理**领域新出现的挑战。因此，很多研究都集中在用新的创意推动这项任务向前发展上。到目前为止，目标是最大限度地提高自动度量的分数，要做到这一点，就必须拿出多个新的模块和技术。一旦这些模型加起来，模型就变得复杂而急需资源。在本文中，我们采取了一个小的倒退，以研究一个在性能和计算复杂性之间进行有趣权衡的体系结构。为此，我们处理神经字幕模型的每个组件，并提出一个或多个解决方案，使模型总体上更轻松。我们的想法灵感来自两个相关的任务: 多式联运和单体神经网络翻译。少

2018 年 10 月 15 日提交;最初宣布 2018 年 10 月。

45. 第 1810.05788[pdf,其他] Cs。CI

专家网络的混合: 可扩展的半监督学习框架

作者:shun kiyono, jun suzuki , kentaro inui

摘要: 目前深度神经网络 (dnn) 在越来越广泛的人工智能任务范围内的成功在很大程度上取决于标记训练数据的质量和数量。一般来说，标记数据的稀缺是需要解决的最重要问题之一，这在许多**自然语言处理**任务中经常被观察到。半监督学习 (ssl) 是一种很有前途的方法，可以通过合并大量未标记的数据来克服这一问题。本文提出了一种新的文本分类任务的可扩展 ssl 方法。我们的方法 "expert/imitator 网络的混合" 的独特特性是，模拟网络学会在未标记的数据上 "模拟" 专家网络的估计标签分布，这可能有助于成为分类。我们的实验表明，该方法持续提高了几种类型的基线 dnn 的性能。我们还证明了我们的方法具有更多的数据，更好的性能属性，并有希望的可伸缩性的未标记的数据。少

2018 年 10 月 12 日提交;最初宣布 2018 年 10 月。

46. 第 1810.0320[pdf,其他] Cs。CI

知识图中的重要属性识别

作者:孙胜杰,杨东,张洪春,陈延旭,赵伟,孟晓南, 胡毅

摘要: 由具有描述和属性以及实体之间关系的实体组成的知识图 (kg) 在各种**自然语言处理**任务中发现了越来越多的应用程序场景。在维基数据这样的典型知识图中，实体通常有大量的属性，但很难知道哪些属性很重要。属性的重要性可以成为从信息检索到**自然语言**生成的各种应用中的一条有价值的信息。在本文中，我们提出了一种使用外部用户生成的文本数据来评估实体属性的相对重要性的通用方法。更具体地说，我们使用单词/子词嵌入技术将外部文本数据与实体的属性名称和值相匹配，并根据它们的匹配一致性对属性进行排序。据我们所知，这是将基于向量的语义匹配应用于重要属性识别的第一项工作，其性能优于以往的传统方法。我们还将检测到的重要属性的结果应用到**语言**生成任务中;与以前生成的文本相比，新方法会生成更多自定义和信息丰富的消息。少

2018 年 10 月 11 日提交;最初宣布 2018 年 10 月。

47. 特别报告: 1810.04805[pdf,其他] Cs。CI

bert: 用于语言理解的深双向变压器的预培训

作者:jacob devlin,张明伟, k 据说 lee , kolina Toutanova

文摘: 我们引入了一种新的**语言**表示模型，称为 bert，它代表来自变压器的双向编码器表示。与最近**的语言**表示模型不同, bert 设计用于通过在所有图层的左右上下文上共同调节来预训练深双向表示。因此，可以使用一个额外的输出图层对预先培训的 bert 表示进行微调，以便为各种任务 (如问题回答和**语言**推断) 创建最先进的模型，而无需大量内容特定于任务的体系结构修改。bert 在概念上很简单，经验也很强大。它在 11 **项自然语言处理**任务上获得了新的最先进的结果，包括将 glue 基准提高到 80.4 (绝对提高 7.6%)、多 nli 精度提高到 18.7 (绝对提高 5.6%) 和 squad v1.1 回答测试 f1 到 93.2 (1.5% 的绝对改善)，比人类的表现高出 2.0%。少

2018 年 10 月 10 日提交;最初宣布 2018 年 10 月。

评论:13 页

48. **第 1810.04700[pdf,其他] Cs。Cl**

数据到文本生成的端到端内容和计划选择

作者:sebastian gehrmann, falcon z. dai , henry elder, 亚历山大 m.拉什

摘要: 学习利用神经网络从结构化数据中生成流畅的**自然语言**已成为 nlg 的常用方法。当结构化数据的形式在不同的示例中发生变化时，此问题可能具有挑战性。本文对序列到序列模型的几个扩展进行了综述，以考虑潜在的内容选择过程，特别是复制注意和覆盖解码**的**变体。我们进一步提出了一种基于不同组合的训练方法，以鼓励模型在训练过程中学习不同的句子模板。对这些技术的实证评价显示，在五个自动化指标以及人工评估中生成的文本的质量有所提高。少

2018 年 10 月 10 日提交;最初宣布 2018 年 10 月。

评论:inlg 2018

49. **建议: 1810.04440[pdf] Cs。Cl**

研究 bhartrhari 的新 vista: 认知 nlp

作者:jyashree gajjam, Diptesh kanojia, malhar kulkarni

摘要: 梵语语法传统主要从帕尼尼的阿斯塔希亚伊开始，以帕达萨斯特拉的身份，最终成为一个位于 bhartrhari 手中的瓦基萨索斯特拉。自 ashok aklujkar 在哈佛大学提交博士论文以来，语法哲学家 bhartrhari 和他的权威著作 "vakyapadiya" 至少 50 多年来一直是现代学者研究的问题。句子和单词作为**语言**中一个有意义的语言单位的概念一直是后来许多作品讨论的主题。虽然一些学者运用语言学技术批判地建立了 bhartrhari 作品的文本，但也有一些学者致力于从这些技巧中探索哲学见解。还有一些人从现代语言学和心理学的角度对他的作品进行了研究。很少有人试图通过逻辑讨论来证明这些观点是合理的。在本文中，我们提出了一个新的观点，研究 bhartrhari，和他的作品，特别是 ' 瓦卡帕迪亚 '。这种观点是从**自然语言处理**(nlp) 领域，更具体地说，什么被称为认知 nlp。我们研究了 bhartrhari 在 "vakyapadiya" 第二章开头给出的句子的定义。我们研究了其中的一个定义，进行了实验，并遵循了沉默阅读梵文段落的方法。我们收集参与者的 gaze 行为数据并对其进行分析，以了解人类头脑中的基本理解过程，并展示我们的结果。我们使用 t 检验来评估我们的结果的统计意义，并讨论我们工作的注意事项。我们还对这一实验和这种方法的有用性提出了一些一般性意见，以便在 bhartrhari 的工作中获得更多的见解。少

2018 年 10 月 10 日提交;最初宣布 2018 年 10 月。

评论:19 页

50. 建议: 1810.04437[pdf,其他] Cs。Lg

坚持是有回报的: 关注 lstm 的起动机制所存在的问题

作者:giancarlo d. salton , john d. kelleher

摘要: **使用语言**模型 (lm) 是几种**自然语言处理**系统中的重要组成部分。由 lstm 单元组成的复发性神经网络 lm, 特别是那些增强外部存储器的单元, 取得了最先进的结果。然而, 由于信息衰落和对最新信息的偏见, 这些模型仍然难以**处理**更有可能包含远程依赖关系的长序列。本文根据 lstm 门控机制持久化信息的时间数, 在对内存中信息进行处理的基础上, 演示了一种在内存增强型 lstm lm 中检索信息的有效机制。少

2018 年 10 月 10 日提交;最初宣布 2018 年 10 月。

51. 第 1810.0101[pdf,其他] Cs。简历

作为 sockeye 神经机器翻译任务的图像字幕

作者:loris bazzani, tobias domhan, felix hieber

摘要: 图像字幕是一个介于计算机视觉和**自然语言处理**之间的跨学科研究问题。任务是生成图像内容的文本描述。图像字幕的典型模型是编码器解码器深层网络, 编码器捕获图像的本质, 而解码器负责生成描述图像的句子。注意机制可以用来自动将解码器集中在与预测下一个单词有关的图像部分。本文探讨了神经机器翻译中常用的不同解码器和注意模型, 即注意力递归神经网络、自注意变压器和完全卷积网络, 它们代表了神经机器翻译的当前状态。神经机器翻译的艺术。图像字幕模块可作为 sockeye 的一部分在 https://github.com/awslabs/sockeye 可以在 https://awslabs.github.io/sockeye/image_captioning.html 找到教程。少

2018 年 10 月 15 日提交;v1 于 2018 年 10 月 9 日提交;最初宣布 2018 年 10 月。

52. 第: 1810.03996[pdf, ps,其他] Cs。Cl

基于序列神经网络的学习名词案例

作者:新浪艾哈迈迪

摘要: 形态衰退是**自然语言处理**中的一项重要任务, 其目的是反映名词的数量、案例和性别。本研究提案旨在解决递归神经网络 (rnn) 在多大程度上能够有效地学习减少名词的情况。考虑到数据稀疏在**处理**形态丰富**的语言**方面的挑战, 以及这些**语言**中句子结构的灵活性, 我们相信, 建模形态依赖关系可以改善神经网络模型的性能。建议进行各种实验, 以了解可解释的特征, 从而更好地推广跨语言任务中的学习模型。少

2018 年 10 月 9 日提交;最初宣布 2018 年 10 月。

评论:3 页研究建议

53. 第: 1810.03993[pdf,其他] Cs。Lg

模型报告的模型卡

作者:margaret mitchell, simone wu, andrew zaldivar, parker barnes, lucy vasserman, ben hutchinson, elena spitzer, inioluwa deborah raji, timnit ge 兄弟

摘要: 受过训练的机器学习模式越来越多地用于在执法、医学、教育和就业等领域执行影响大的任务。为了阐明机器学习模型的预期用例, 并最大限度地减少它们在不太适合的上下文中的使用, 我们建议在发布的模型中附上详细说明其性能特征的文档。在本文

中，我们提出了一个我们称之为模型卡的框架，以鼓励这种透明的模型报告。模型卡是附带训练有素的机器学习模型的简短文档，可在各种条件下提供基准评估，例如跨不同的文化、人口或表型组（例如种族、地理位置、性别、与预期应用领域相关的菲茨帕特里克皮肤类型）和交叉组（例如，年龄和种族，或性别和菲茨帕特里克皮肤类型）。模型卡还披露了打算使用模型的背景、业绩评价程序的细节以及其他相关信息。虽然我们主要关注计算机视觉和**自然语言处理**应用领域中以人为本的机器学习模型，但这个框架可以用来记录任何训练有素的机器学习模型。为了巩固这一概念，我们为两个受监督的模型提供卡片：一个是检测图像中笑脸的培训，另一个是检测文本中有毒评论的培训。我们建议使用模型卡，作为实现机器学习和相关 ai 技术负责任民主化的一个步骤，提高人工智能技术运作情况的透明度。我们希望这项工作鼓励那些发布训练有素的机器学习模型的人在模型发布的同时提供类似的详细评估数字和其他相关文档。少

2018 年 10 月 5 日提交;最初宣布 2018 年 10 月。

54. 第: 1810.03918[pdf,其他] Cs。红外

利用结构特征和依赖原则在答题中提取答案

作者:lokesh kumar sharma, namita mittal

摘要: 问答（qa）研究是**自然语言处理**中一项具有重要挑战性的课题。qa 旨在从相关文本片段或文档中提取准确答案。qa 研究背后的动机是使用最先进搜索引擎的用户的需求。用户需要的是准确的答案，而不是可能包含答案的文档列表。本文从相关文献中提取出成功的答案，需要几个有效的特点和关系来提取。这些特征包括各种词汇、句法、语义和结构特征。建议的结构特征是从问题和支持的文档的依赖特征中提取出来的。实验结果表明，结合基本特征，利用依赖原则设计，结构特征提高了答案提取的准确性。提出的结构特点采用了新的设计原则来提取远程关系。这一增加是整体答案提取精度提高的一个可能原因。少

2018 年 10 月 9 日提交;最初宣布 2018 年 10 月。

评论:12 页, 11 个图, 6 个表, 4 算法和 ieee 格式

55. 建议: 1810.03552[pdf,其他] Cs。Cl

零资源多语言模式转移: 学习共享内容

作者:xilun chen, ahmed hassan awadallah, hani hassan, wei wang, claire cardie

摘要: 利用神经网络模型，现代**自然语言处理**和理解应用得到了极大的提升。然而，对于大多数**语言**，特别是没有足够的附加说明的培训数据的资源不足的语言来说，情况并非如此。跨语言迁移学习方法通过利用来自其他（源）**语言**的标记数据（通常借助并行语料库等跨语言资源）来提高低资源目标**语言**的性能。在这项工作中，我们提出了第一个零资源多语言传输学习模型，它可以利用多种源**语言**的培训数据，同时不需要目标**语言**培训数据，也不需要跨语言监督。与现有的仅依赖**于语言**不变特征进行跨语言传输的方法不同，我们的方法以一致的方式利用语言不变和**特定于语言**的特征。我们的模型利用对抗网络来学习**语言**不变的特征和专家混合模型，动态地利用目标**语言**和每个单独源**语言**之间的关系。这使我们的模型能够有效地了解在多语言设置中在各种**语言**之间共享的内容。与现有技术相比，它显著提高了性能，如在多文本分类和序列标记任务（包括大型实际行业数据集）上进行的一系列广泛实验中所示。少

2018 年 10 月 8 日提交;最初宣布 2018 年 10 月。

56. 建议: 1810.03445[pdf] Cs。Cl

基于词向量组合模型的语言进化树的构建

作者:朱高,姜艳辉,高俊辉

文摘: 本文试图通过案例计算来探讨**语言**的演变。首先，我们选择了 1400 至 2005 年十一位英国作家的小说，找到了相应的作品;然后，利用**自然语言处理**工具构造相应的 11 个语料库，计算 11 个语料库中 100 个高频词的相应词向量;接下来，对于每个语料库，我们将 100 个单词向量从开始到结束连接为一个;最后，利用相似性比较和分层聚类方法生成了组合的 11 个词向量之间的关系树。这棵树代表了十一个语料库之间的关系。我们发现，在聚类生成的树中，语料库与与语料库相对应的年份之间的距离基本相同。这意味着我们发现了一个特定的**语言**进化树。为了验证这种方法的稳定性和多功能性，我们增加了另外三个主题: 狄更斯的八部作品、19 世纪诗人的作品和近 60 年来的艺术批评。对于这四个主题，我们测试了不同的参数，如语料库的时间跨度、语料库之间的时间间隔、词向量的维度以及高频公共词的数量。结果表明，该方法相当稳定，用途广泛。少

2018 年 10 月 4 日提交;最初宣布 2018 年 10 月。

57. 建议: 1810.03430[pdf] Cs。红外

来自维基百科的交叉脚本印地语英语 ner 语料库

作者:mohd zeeshan ansari, tanvir ahmad, md Ansari ali

摘要: 社交媒体平台上生成的文本本质上是混合的语言文本。任何形式的**语言**混合都会给**语言处理**系统带来相当大的困难。此外,**语言处理**研究的进步取决于标准语料库的可用性。混合语言的印度命名实体识别 (ner) 系统的开发正面临着障碍，因为没有标准的评价公司。这种语料库可能是混合的语言**性质**，其中文本是使用多种**语言**编写的，主要是使用单个脚本。我们工作的动机是强调这种公司的自动生成，以鼓励混合语言的印度 ner。本文介绍了维基百科类别页面中的交叉脚本 hindi-nh 语料库的编写过程。使用标准的 conll-2003 类别的 per、loc、org 和 misc 对该公司进行了成功的注释。对各种机器学习算法进行了评价，取得了良好的效果。少

2018 年 10 月 8 日提交;最初宣布 2018 年 10 月。

评论:智能数据通信技术与物联网国际会议 (icici-2018)

58. 建议: 1810.02100[pdf,其他] Cs。Cl

域外依赖关系分析的半监督方法

作者:于俊涛

摘要: 依赖关系分析是将句法树分配给文本的重要**自然语言处理**任务之一。由于依赖语料库的更广泛可用性以及改进的解析和机器学习技术，基于监督学习的系统的解析精度得到了显著提高。然而，由于监督学习的**性质**，这些解析系统高度依赖于手动注释的培训语料库。它们在域内数据上的工作效果相当好，但在域外文本上测试时，性能显著下降。为了弥补域内和域外的性能差距，本文研究了三种域外依赖关系分析的半监督技术，即共同训练、自我训练和依赖**关系语言**模型。我们的方法使用易于获得的未标记数据来提高域外解析精度，而无需昂贵的语料库注释。对几个英语领域和多语言数据的评价表明，在解析精度方面有了相当好的提高。总体而言，这项工作对域外依赖关系分析的半监督方法进行了调查，在统一的框架中，我扩展并比较了一些重要的半监督方法。这些技术之间的比较表明，自我训练与领域外解析的共同训练同样有效，而依赖**语模型**可以提高域内和域外的准确性。少

2018 年 10 月 4 日提交;最初宣布 2018 年 10 月。

评论:博士论文

59. 第 1810.01570[pdf,其他] Cs。CI

一种用于患者笔记的深层学习体系：实现与评价

作者:kaung khin, phip burckhardt, rema padman

摘要: 除身份识别是将 18 种受保护的健康信息 (phi) 从临床笔记中删除的**过程**，以便将文本视为无法单独识别。**自然语言处理**(nlp) 的最新进展使得能够使用深度学习技术来完成去识别任务。在本文中，我们提出了一个深入的学习架构，它建立在最新的 nlp 进步的基础上，结合了深刻的上下文化嵌入和变分退出双 lstm。我们在两个金标准数据集上测试此体系结构，并表明该体系结构在这两个数据集上实现了最先进的性能，同时也比其他系统收敛得更快，而不使用字典或其他知识源。少

2018 年 10 月 2 日提交;最初宣布 2018 年 10 月。

评论:提交第 28 届信息技术和系统讲习班

60. 建议: 1810.01466[pdf,其他] si

开放源俄罗斯推特数据的无监督机器学习揭示了全球范围和运营特征

作者:克里斯托弗·格里芬,布雷迪·比克尔

摘要: 我们开发并使用了一系列统计方法 (无人监督的机器学习)，从推特提供的数据集中提取相关信息，这些数据集由据称试图影响 2016 年美国总统的俄罗斯手推车组成选举。这些无人监督的统计方法可以快速识别 (一) 巨魔人口中的新兴语言社区; (二) 作业的跨国范围和 (三) 可用于以下目的的巨魔**的**业务特点:将来的识别。通过**自然语言处理**、多种学习和傅立叶分析，我们确定了一个不仅包括 2016 年美国大选，还包括法国国家和地方及全国德国选举的行动。我们展示了由此产生的巨魔群体是由具有共同但明确定制的行为特征的用户组成的。少

2018 年 10 月 2 日提交;最初宣布 2018 年 10 月。

评论:19 页, 14 数字

61. 建议: 1810.01127[pdf] Cs。艾

神经系统中的预测学习：发现潜在的生成结构

作者:andrea e. martin, leonidas a. a. doumas

摘要: 人类从他们的环境中学习复杂的潜在结构 (例如，**自然语言**、数学、音乐、社会等级)。在认知科学和认知神经科学中，从感官或一阶表示中推断高阶结构的模型被提出来解释人类行为的复杂性和灵活性。但这些模型调用的结构是如何在首先出现在神经系统中的呢？为了回答这个问题，我们解释了系统如何从完全非结构化数据的经验中学习潜在的表示结构 (即谓词)。在谓词学习**的过程**中，人工神经网络利用跨神经元程序集的**分布式**计算的自然动态特性来学习谓词，同时也将它们结合起来从概念上讲，两个计算方面似乎是必要的人类行为，根据形式理论在多个领域。我们描述了如何使用神经振荡将谓词推广到人工神经网络中，以实现类似人类的外推和组合。从经验中学习谓词、从组合上表示结构以及推断看不见的数据的能力为理解和建模最复杂的人类行为提供了一条前进的道路。少

2018 年 10 月 2 日提交;最初宣布 2018 年 10 月。

62. 第: 1810.00664[pdf, ps,其他] Cs。CI

向量空间模型中的文本相似性：一个比较研究

作者:omid shahmirzadi, adam Lugowski, kenneth younge

摘要: 语义文本相似性的自动测量是**自然语言处理**中的一项重要任务。在本文中，我们评估了不同的向量空间模型的性能来执行此任务。我们解决了对专利与专利相似性建模的实际问题，并比较了 tfidf (和相关扩展)、主题模型 (例如，潜在语义索引) 和神经模型 (例如，段落向量)。与预期相反，只有在: 1) 对目标文本进行压缩时，文本嵌入方法的附加计算成本才是合理的;和 2) 相似度的比较是微不足道的。否则，tfidf 在其他情况下的表现令人惊讶: 特别是在更长和更技术性的文本中，或在最近的邻居之间进行更细粒度的区分。出乎意料的是，对 tfidf 方法的扩展，如增加名词短语或递增计算术语权重，在我们的上下文中没有帮助。少

2018 年 9 月 24 日提交;最初宣布 2018 年 10 月。

评论:17 页

63. 第 1810.00660[pdf,其他] Cs。CI

基于注意的拼写和语法纠错编码网络

作者:新浪艾哈迈迪

摘要: 自动拼写和语法更正系统是**自然语言**应用中使用最广泛的工具之一。在本文中，我们假设错误修正的任务是一种单语言机器翻译，其中源句可能是错误的，目标句子应该是输入的修正形式。我们在这个项目中的主要重点是建立神经网络模型的错误修正任务。特别是，我们研究了序列到序列和基于关注的模型，这些模型最近显示出比许多**语言处理**问题的最新性能更高的性能。证明了神经机器翻译模型可以成功地应用于纠错任务。虽然本研究的实验是在阿拉伯语料库上进行的，但我们的研究方法可以很容易地应用于任何**语言**。少

2018 年 9 月 21 日提交;最初宣布 2018 年 10 月。

评论:75 页, 20 位数字，作为硕士论文提交巴黎笛卡尔大学

64. 建议: 1810.00462[pdf, ps,其他] Cs。Hc

遗憾理论定量测量的人机界面设计

作者:姜龙生,王悦

摘要: 遗憾理论是描述风险下人类决策的理论。获得遗憾理论的量化模型的关键是，当人类在一组选项中做出选择时，要衡量他们心目中的偏好。与物理量不同的是，测量心理偏好不是程序不变的，即读数在方法改变时改变。在这项工作中，我们通过选择与个人做出选择的方式相兼容的程序来缓解这种影响。我们认为，由此产生的模式更接近人类决策的**性质**。偏好激发**过程**被分解为一系列简短的调查，以减少认知工作量，提高反应精度。为了使问题**自然**和熟悉的主题，我们遵循的洞察，人类产生，量化和沟通的偏好在**自然语言**。因此，模糊集理论被用来模拟主体的反应。基于这些想法，图形人机界面 (hci) 被设计出来，以阐明信息，并有效地收集人类的反应。该设计还考虑了人的启发式和偏差，例如范围效应和锚固效应，以提高其可靠性。调查的总体表现是令人满意的，因为测量模型显示预测精度相当于被试的修正效果。少

2018 年 9 月 30 日提交;最初宣布 2018 年 10 月。

评论:6 页, 5 个数字

65. 决议: 1810.00162[pdf,其他] Cs。简历

nice: 神经网络量化中的噪声注入与夹紧估计

作者:chaim baskin, natan liss, yoav chai , evgenii zheltonozhskii, eli schwartz, rajaggiryes, avi mendelson , 亚历山大 m. bronstein

摘要: 卷积神经网络 (cnn) 在计算机视觉、语音识别、**自然语言处理**等许多领域都很受欢迎。尽管深度学习导致了这些领域的突破性性能，但所使用的网络在计算上要求很高，即使在 gpu 上也远远不是实时的，gpu 不高效，因此不适合移动设备等低功耗系统. 为了克服这一挑战，提出了一些量化这些网络的权重和激活的解决方案，这大大加快了运行时的运行。然而，这种加速是以更大的错误为代价的。本文提出的 \ uniqname 方法通过噪声注入和学习夹紧来训练量化神经网络，提高了神经网络的精度。这将导致各种回归和分类任务的最先进的结果，例如，imagenet 分类与体系结构，如 resnet-18\ 34/50 与低的 3 位重量和激活。我们在 fpga 上实现了该解决方案，以证明其在低功耗实时应用中的适用性。本文的实施情况可 https://github.com/Lancer555/NICE

2018 年 10 月 2 日提交;v1 于 2018 年 9 月 29 日提交;最初宣布 2018 年 10 月。

66. 第 189.11086[pdf, ps,其他] Cs。Lg

学习经常性的双元/三元加权性

作者:arash ardakani, zhengyunji, sean c. smythson, brett h. meyer , warren j. gross

摘要: 递归神经网络 (rnn) 在**处理**序列数据方面表现出了优异的性能。但是，由于它们的递归**性质**，它们既复杂又占用大量内存。这些限制使得 rnn 难以嵌入到需要硬件资源有限的实时**处理**的移动设备上。为了解决上述问题，我们引入了一种方法，可以在训练阶段学习二进制和三元权重，以促进 rnn 的硬件实现。因此，使用此方法可以通过简单的积累来替代所有多重累积操作，从而在硅面积和功耗方面为自定义硬件带来显著的好处。在软件方面，我们使用长的短期存储器 (lstm) 在各种序列模型 (包括序列分类和**语言**建模) 上评估我们的方法的性能 (在准确性方面)。我们证明，我们的方法在运行时使用双核/三元权重时，在上述任务上取得了有竞争力的结果。在硬件方面，我们提出了自定义硬件，用于加速具有双元/三元权重的 lstm 的重复计算。最后，我们证明，与 asic 平台上的全精度实现相比，具有 binary/tary 权重的 lstm 可以实现高达 12x 的内存节省和 10 倍的推理加速。少

2018 年 9 月 28 日提交;最初宣布 2018 年 9 月。

67. 建议: 1809.10763[pdf,其他] Cs。Cl

为索拉尼库尔德人建造 lemmatizer 和拼写检查器

作者:shahin salavati, sina ah 迪

摘要: 本文旨在提出索拉尼库尔德语的引引种和词级纠错系统。我们提出了一种基于形态规则和 n-gram **语言**模型的混合方法。据我们所知，我们分别将我们的引引信和纠错系统称为 peyv 和 rénh s，这是为索拉尼库尔德人提供的第一个工具。peyv lemmatizer 的准确率为 18.7%。至于 rén s，使用词典，我们已经获得了 96.4% 的准确性，而没有词典，校正系统有 87% 的精度。作为两个基本的文字**处理**工具，这些工具可以为进一步研究索拉尼库尔德语的更**自然语言处理**应用铺平道路。少

2018 年 9 月 27 日提交;最初宣布 2018 年 9 月。

评论:6 页的文章，发表在《第八届语言与技术会议》上，波兰波兹南

68. 第 1809. 10707[pdf,其他] Cs。简历

交通摄像图像的语义主题分析

作者:jeffreyliu, andrew weinert, Saurabh amin

摘要: 交通摄像头通常部署在道路基础设施网络中监控组件，为运营商提供有关网络关键点状况的视觉信息。然而，人类观察员**处理同时处理**信息来源的能力往往有限。在深度学习方法的推动下，计算机视觉的最新发展使一般对象识别、基于相机的传感机会超越了现有的人类观察者范式。在本文中，我们提出了一个**自然语言处理**(nlp) 启发的方法，名为标签词 (blw)，用于分析图像数据集专用文本标签。llw 模型以传统的矩阵形式表示数据，支持数据压缩和分解技术，同时保留语义可解释性。我们应用潜在的 dirichlet 分配 (lda) 主题模型将标签数据分解为少量语义主题。为了说明我们的做法，我们使用 2017 年 12 月至 2018 年 1 月期间从波士顿地区收集的高速公路摄像机图像。我们分析摄像机对天气事件的敏感性;识别时间流量模式;并分析了冬季假期、"炸弹旋风" 冬季风暴等不常见事件的影响。这项研究展示了我们的方法的灵活性，它允许我们分析天气事件和高速公路交通只使用交通摄像头图像标签。少

2018 年 9 月 27 日提交;最初宣布 2018 年 9 月。

评论:将于 2018 年 11 月 3 日至 7 日

69. 第: 1809.10617[pdf,其他] Cs。CI

通过研究对象促进地球科学的公平研究

作者:andres garcia-silva, jose manuel Gomez-Perez, raul palma, marcin krystek, simone mantovani, f 里分 fomflini, valentina grande, francesco de leo, Stefano salvi, elisa trasati, vito romaniello, mirko alami, cristiano silvagni, rosemarie 塞拉利昂, fulvio marelli, sergio alani, miche Lazzarini, hazel j.napier, helen m.glaves, timothy aldridge, charles meertens, fran boler, henry w.loescher, christine laney, melissaa genazzio 等人 (另有 2 名作者没有出示)

摘要: 数据密集型科学界正在逐步采用 fair 的做法，以提高科学突破的能见度并实现重用。这一运动的核心是以符合 fair 原则并持续发展关键基础设施和工具的方式包含和描述科学信息和资源。本文介绍了围绕几个地球科学学科围绕研究对象采用 fair 所涉及的挑战、经验和解决方案。在这一旅程中，我们的工作是全面的，其成果包括: 适应地球科学家需要的扩展研究对象模型;提供数字对象标识符 (doi)，以便能够持续识别并给予作者应有的信任;通过**自然语言处理**生成基于内容的、语义丰富的研究对象元数据，通过推荐系统和第三方搜索引擎提高可见性和重用性;以及各种类型的检查表，这些检查表提供了研究对象质量的紧凑表示，作为科学重用的关键推动因素。所有这些结果都集成在 rohub 中，该平台为跨不同科学界的大量应用程序和接口提供研究对象管理功能。为了监测和量化社区对研究对象的理解，我们定义了指标，并通过 rohub 获得了措施，本文也对这些指标进行了讨论。少

2018 年 9 月 27 日提交;最初宣布 2018 年 9 月。

70. 建议: 1809.10267[pdf,其他] Cs。CI

解释和文本总结的语义句子嵌入

作者:张志, shagan sah, thang nguyen, dheeraj peri,亚历山大 loui, carl salvaggio, raymond ptucha

文摘: 本文介绍了一种适用于高级**自然语言处理**的矢量编码框架句子。我们的潜在表示被证明是编码句子与共同的语义信息与相似的向量表示。矢量表示是从一个编码器-解码器模型中提取出来的，该模型是在句子释义对上训练的。我们演示了句子表示在两种不同任务中的应用--句子解释和段落摘要，使其对**处理**文本的常用的经常性框架具有吸引力。实验结果有助于深入了解矢量表示如何适合于高级**语言嵌入.**少

2018 年 9 月 26 日提交;最初宣布 2018 年 9 月。

评论:5 页, 4 位数, ieee 全球 2017 年会议

71. 第: 1809.10158[pdf,其他] si

"参议员，我们销售广告": 2016 年俄罗斯脸谱广告活动分析

作者:ritam dutt, ashok deb, emilio ferrara

摘要: 美国民主的关键方面之一是自由和公正的选举，使一任总统能够和平移交权力。2016 年美国总统选举之所以突出，是因为在选举之前、期间和之后都有外国的影响力。这种怀疑影响的很大一部分是通过社交媒体进行的。在本文中，我们特别研究了据称俄罗斯政府购买的 3, 500 个 facebook 广告。这些广告由美国国会众议院情报委员会于 2018 年 5 月 10 日发布。我们使用**自然语言处理**技术对广告进行了分析，以确定与最有效的广告相关的文本和语义特征。随着时间的推移，我们将广告集中到各种广告系列和与之相关的标记方中。我们还在个人、活动和政党的基础上研究了广告的有效性。最有效的广告往往有不太积极的情绪，关注过去的事件，在**本质上**更具体、更个性化。更有效的运动也表现出如此相似的特点。竞选的持续时间和广告的推广表明，人们希望播下分裂的阵营，而不是左右选举。少

2018 年 9 月 26 日提交;最初宣布 2018 年 9 月。

评论:接受: 第三届智能信息技术国际会议 (iciit 2018)，印度钦奈, 2018 年 12 月 11 日至 14 日

72. 第 1809.10025[pdf, ps,其他] cs. cy

大规模的个性化教育

作者:sam saarinen, evan cater, michael littman

摘要: 根据学生个人的需要调整信息呈现方式，可使学生 outcomes~\cite{bloom19842} 获得巨大收益。这一发现可能是由于不同的学生学习方式不同，可能是能力、兴趣或其他 factors~\cite{schiefele1992interest} 不同的结果。使演讲适应个人的教育需求历来是专家的领域，因此在规模上的工作成本很高，在后勤上也很有挑战性，也导致教育成果的不公平。增加的课程规模和大量的 mooc 注册提供了前所未有的学生数据访问权限。我们建议，强化学习 (rl) 以及半监督学习、**自然语言处理**和计算机视觉方面的新兴技术对于利用这些数据提供个性化服务至关重要大规模的教育。少

2018 年 9 月 24 日提交;最初宣布 2018 年 9 月。

73. 第 xiv: 1809.09096[pdf,其他] Cs。红外

文本综述----------

作者:dav 上个月 bacciu, antonio bruno

摘要: 提取压缩是一个具有挑战性的**自然语言处理**问题。这项工作有助于制定神经萃取压缩作为一个解析树转导问题，而不是序列转导任务。在此基础上，我们引入了一个深度神经模型，用于通过扩展标准的长期短期记忆来学习结构到子结构树的转换，同时考虑到结构递归中的父子关系。该模型在准确性和压缩率方面均能达到句子压缩基准的最佳性能。少

2018 年 9 月 24 日提交;最初宣布 2018 年 9 月。

评论:将出现在 ieee scci 深度学习 2018

74. 第 xiv: 180 099.09078[pdf,其他] Cs。CI

通过基于基于关注的图形信息聚合联合提取多个事件

作者:小刘,罗振晨,黄和燕

摘要: 事件提取在**自然语言处理**中具有实用的应用价值。在现实世界中，存在于同一句子中的多个事件是一种常见现象，在这种现象中，提取这些事件比提取单个事件更困难。以前使用顺序建模方法建模事件之间关联的工作在很大程度上受到了捕获非常长的依赖关系的低效率的影响。在本文中，我们提出了一个新的联合多事件提取（jmee）框架，通过引入句法快捷方式弧，以增强信息流和基于注意的卷积网络，共同提取多个事件触发器和参数。模型图信息。实验结果表明，与最先进的方法相比，我们提出的框架取得了较好的效果。少

2018 年 10 月 23 日提交;v1 于 2018 年 9 月 24 日提交;最初宣布 2018 年 9 月。

评论:被 emnlp 2018 所接受

75. 第 1809. 08730[pdf,其他] Cs。Cl

命名实体识别的可变形堆叠结构

作者:曹淑阳,邱锡鹏,黄宣静

摘要: 用于命名实体识别的神经结构在**自然语言处理**领域取得了巨大的成功。目前，主导体系结构包括一个双向递归神经网络（rnn）作为编码器和一个条件随机场（crf）作为解码器。本文提出了一种用于命名实体识别的可变形堆叠结构，在该结构中，动态建立了相邻层之间的连接。我们通过将变形堆叠结构调整到不同的层来评估它。我们的模型在 ontonotes 数据集上实现了最先进的性能。少

2018 年 9 月 28 日提交;v1 于 2018 年 9 月 23 日提交;最初宣布 2018 年 9 月。

76. 第 1809. 08396[pdf,其他] Cs。铬

gdpr 后的隐私政策格局

作者:thomas linden, hamza harkous, kassem fawaz

摘要: 每一个新的隐私规定都会带来这样的问题：它是否会改善用户的隐私，或者是否会对理解和行使他们的权利造成更多的障碍。欧盟一般数据保护条例（gdpr）是有史以来要求最高、最全面的隐私法规之一。因此，在它生效几个月后,**自然**会研究它对网上隐私政策格局的影响。在这项工作中，我们对 gdpr 前后的隐私政策进行了第一次纵向、深入和规模的评估。我们衡量这些政策的完整消费周期，从第一次用户印象到合规性评估。我们创建了一个由 3 086 英语**语言**隐私政策组成的多样化语位，并为其获取 gpdr 前版本和 gdpr 后版本。通过 amazon mturk 上 530 参与者的用户研究，我们发现，除了欧洲顶级网站外，隐私政策的可视化呈现在有限的数据敏感类别中略有改进。我们还发现，隐私政策的可读性受到 gdpr 的影响，因为尽管为减少对被动判决的依赖做出了努力，但判决和词语增加了近 3 0%。我们在自动化**自然语言处理**技术的基础上，进一步开发了一种新的工作流程，用于自动评估隐私政策中的要求。我们发现了 gdpr 引发积极变化的证据，模糊程度模糊，平均超过 8 个指标，超过 20.5% 的政策有所改善。最后，我们展示了隐私策略涵盖了更多的数据实践，特别是围绕数据保留、用户访问权限和特定受众，平均 15.2% 的策略在 8 个合规性指标中得到了改进。然而，我们的分析揭示了目前的现状与 gdpr 的最终目标之间存在着很大的差距。少

2018 年 9 月 22 日提交;最初宣布 2018 年 9 月。

77. 第: 189.07954[pdf,其他] cse

利用自然语言处理预测堆栈溢出问题和片段的编程语言

作者:kamel alreshedy, dhanush dharmaretnam , daniel m.german, venkatesh srinivasan, t. aarongulliver

摘要: 堆栈溢出是软件开发人员中最受欢迎的问答网站。作为知识共享和获取的平台，堆栈溢出中发布的问题通常包含一个代码段。堆栈溢出依赖于用户正确标记问题的编程**语言**，它只是假定问题中的片段的编程**语言**与问题本身的标记相同。本文提出了一种分类器，利用**自然语言处理**(nlp) 和机器学习 (ml) 来预测堆栈溢出中发布的问题的编程**语言**。该分类器通过结合问题的标题、正文和代码段中的功能，在预测 24 种最流行**的编程语言**时实现了 91.1 的准确性。我们还建议使用一个分类器，该分类器只使用问题的标题和正文，其准确性为 81.1。最后，我们提出了一个仅实现 77.7% 精度的代码段分类器。这些结果表明，在问题的文本和代码段的组合上部署机器学习技术提供了最佳性能。这些结果还证明，可以确定几行源代码的片段的编程**语言**。我们可视化了两种编程**语言** java 和 sql 的功能空间，以确定与这些**语言**相对应的堆栈溢出中的问题中的一些特殊信息属性。少

2018 年 9 月 21 日提交;最初宣布 2018 年 9 月。

78. **第: 189.07945[pdf,其他] cse**

scc: 代码段的自动分类

作者:kamel alreshedy, dhanush dharmaretnam , daniel m.german, venkatesh srinivasan, t. aarongulliver

摘要: 确定源代码文件的编程**语言**已在研究界得到考虑;研究表明，机器学习 (ml) 和**自然语言处理**(nlp) 算法可以有效地识别源代码文件的编程**语言**。但是，确定代码段或几行源代码的编程**语言**仍然是一项具有挑战性的任务。在线论坛 (如堆栈溢出) 和代码存储库 (如 github) 包含大量代码段。在本文中，我们描述了源代码分类 (scc)，它是一种分类器，可以识别用 21 种不同编程**语言**编写的代码段的编程**语言**。采用多项式朴素贝叶斯 (mnb) 分类器，采用堆栈溢出柱进行训练。结果表明，该精度达到 75%，高于编程**语言**识别 (pli 是一个专有的在线分类器的片段)，其精度只有 55.5。该工具的精度、召回和 f1 得分分别为 0.76 分、0.76 分和 0.76 分。此外，它还可以区分代码段与一系列编程**语言**(如 c、c++ 和 c#)，还可以识别编程**语言**版本 (如 c# 3.0、c# 4.0 和 c# 5.0)。少

2018 年 9 月 21 日提交;最初宣布 2018 年 9 月。

日记本参考:2018 年源代码分析和操作工作会议

79. **第: 189.07889[pdf,其他] Cs。CI**

英语前置词短语的预测

作者:njoung kim, kyle rawlins, benjamin van durme, paul smolensky

摘要: 区分动词的核心依赖项和非核心依赖项 (即参数和附加项) 是一个长期存在的、非平凡的问题。在**自然语言处理**中，语篇信息在语义角色标记 (srl) 和前置词 (pp) 依恋消歧等任务中具有重要意义。在理论语言学中，存在许多关于性参数的诊断测试，但它们往往产生相互矛盾和潜在梯度的结果。这对于语法斜项目 (如 pp) 尤其如此。我们提出了两个 pp 参数预测任务分支这两个动机: (1) 二元参数/辅助分类的 pp 在 verbnet, (2) 梯度参数预测使用人类的判断作为黄金标准，并报告结果使用预先训练的单词嵌入和其他语言信息功能的预测模型。我们在每个任务上的最佳结果是 (1)$acc.=0.9\ 55$, $F_1=0.9\ 54$(elmo + bilstm) 和 (2) 皮尔逊的

r=0.624(word2vec+MLP)。此外，当句子编码器预先训练完成我们的任务时，我们还演示了参数预测在通过 srl 上的性能增益来改善句子表示的效用。少

2018 年 9 月 24 日提交;v1 于 2018 年 9 月 20 日提交;最初宣布 2018 年 9 月。

80. **第: 189.07485[pdf, ps,其他] Cs。CI**

问答系统自然语言问题解释的定量评价

作者:takuto asakura, jin-dongkim, yasunoriyamamoto, yuka tateisi , toshihisa takagi

文摘: 系统基准评价在改进问答系统技术的**过程**中发挥着重要作用。虽然目前有一些现有的**自然语言**(nl) qa 系统的评估方法，但大多数方法只考虑最终答案，限制了它们在黑匣子风格评价中的效用。在此，我们提出了一种细分的评估方法，以便能够对 qa 系统进行更细粒度的评估，并提出了一个针对 nl 问题 (nlq) 解释步骤的评估工具，这是 qa 管道的第一步。使用两个公共基准数据集的实验结果表明，我们可以使用该方法更深入地了解 qa 系统的性能，这应该为改进系统提供更好的指导，而不是使用黑匣子式方法。少

2018 年 9 月 20 日提交;最初宣布 2018 年 9 月。

评论:16 页, 6 位数字, jist 2018

81. **第 xiv:1809. 06943[pdf] Cs。红外**

论证挖掘: 利用多源和背景知识

作者:anastasios lytos, thomas lagkas, panagiotis sarigiannidis, kalina bontcheva

摘要: 论证挖掘领域产生于需要从所表达的观点中确定根本原因，以及发展已建立的观点挖掘和情绪分析领域的紧迫性。最近在更广泛的人工智能领域取得的进展，结合通过社交网站获得的数据，为**自然语言过程**的每一个子领域创造了巨大的潜力，包括论证挖掘。少

2018 年 9 月 18 日提交;最初宣布 2018 年 9 月。

评论:第十二届东南欧博士生年会 (dsc2018)，国际标准书号: 978-960-94-94-20-7，66-74 页，希腊塞萨洛尼基, 2018 年 5 月

82. **修订: 1809. 06858[pdf,其他] Cs。CI**

法语: 频繁-不可知语词表示

作者:龚成岳,狄河,徐坦, 秦涛, 王立伟, 刘铁燕

摘要: 连续字表示 (又名词嵌入) 是**自然语言处理**任务中许多基于神经网络的模型的基本组成部分。虽然人们普遍认为，在嵌入空间中，语义相似的单词应该彼此接近，但我们发现，在几个任务中学习到的单词嵌入对单词频率有偏见: 高频和低频单词的嵌入位于嵌入空间的不同次区域，一个罕见的词和一个流行的词的嵌入即使在语义上是相似的，也可以相去甚远。这使得学习过的单词嵌入无效，特别是对于罕见的单词，从而限制了这些神经网络模型的性能。在本文中，我们开发了一种清晰、简单而有效的方法，通过对抗训练来学习 \ 意思 {议定-敏捷字嵌入} (frage)。我们对四个**自然语言处理**任务中的十个数据集进行了全面的研究，包括单词相似性、**语言**建模、机器翻译和文本分类。结果表明，使用 frage，我们在所有任务中都获得了比基线更高的性能。少

2018 年 9 月 18 日提交;最初宣布 2018 年 9 月。

评论:将于 2018 年 NIPS 出现

83. **第 xiv:1809.06639[pdf] Cs。艾**

西班牙临床叙述中的肺癌概念注释

作者:marjan najafabadipour, juan manuel tuñas, alejandro rodríguez-zález, ernestina menasalvas

摘要: 最近临床数据生成的迅速增长和计算科学的快速发展使我们能够从医疗保健行业的海量数据集中提取新的见解。肿瘤学临床笔记正在创建丰富的数据库来记录患者的病史，它们可能包含大量的模式，可以帮助更好地管理疾病。但是，这些模式被锁定在临床文档的自由文本 (非结构化) 部分中，其结果是限制卫生专业人员从这些文档中提取有用的信息，并最终在准确的方式。信息提取 (ie)**过程**需要**自然语言处理**(nlp) 技术来为这些模式分配语义。因此，本文分析了可通过 apache 非结构化信息管理体系结构 (uima) 框架集成的特定肺癌概念注释器的设计。此外，我们还解释了注释结果的生成和存储的详细信息。少

2018 年 9 月 18 日提交;最初宣布 2018 年 9 月。

评论:10 页, 6 个数字

日记本参考:生命科学中的数据集成 (料展 2018)

84. **第: 1809. 06187[pdf] Cs。简历**

利用卷积神经网络研究不同隐藏层和时代手写数字识别精度变化的研究与观察

作者:rezoana bente arif, md. abu bakr siddique, mohammad mahmudur rahman khan, mahjabin rahman oishe

文摘: 如今，深度学习可以应用于医学、工程等多个领域。在深度学习中，卷积神经网络 (cnn) 广泛应用于模式和序列识别、视频分析、**自然语言处理**、垃圾邮件检测、主题分类、回归分析、语音识别、图像分类、目标检测、分割、人脸识别、机器人和控制。与其在大型应用中近乎人的水平准确相关的好处，导致近年来美国有线电视新闻网的接受程度越来越高。本文的主要贡献是分析美国有线电视新闻网隐藏层模式对网络整体性能的影响。为了证明这种影响，我们在修改后的国家标准与技术研究所 (mnist) 数据集上应用了不同层次的神经网络。同时，观察不同数量的隐藏层和时代的网络精度变化，并对它们进行比较和对比。利用随机梯度和反向传播算法对系统进行训练，并采用前馈算法进行测试。少

2018 年 9 月 22 日提交;v1 于 2018 年 9 月 17 日提交;最初宣布 2018 年 9 月。

评论:将在第四届 ieee 电气工程与信息与通信技术国际会议 (iceeieect 2018) 上发表

85. **第 xiv: 1809.05896[pdf,其他] Cs。Lg**

利用递归神经网络对过程实例进行分类

作者:markku hinkka, teemu lehto , keijo heljanko, 亚历山大 jung

摘要: **生产过程**挖掘包括将操作系统创建的日志转换为**过程**模型的技术。在**过程**挖掘工具中，通常需要能够对正在进行的**流程**实例进行分类，例如,**预测流程**仍需多长时间才能完成，或**将流程**实例分类到不同的过程实例类仅基于到目前为止在**流程**实例中发生的活动。经常神经网络及其子类，如门式递归单元 (gru) 和长期记忆 (lstm)，已被证明能够学习随后分类任务的相关时间特征。本文将递归神经网络应用于**过程**实例的分类。使用从事件日志跟踪中提取的标记**流程**实例，以监督方式训练建议的模型。这是我们第一次知道 gru 被用于对业务**流程**实例进行分类。我们的主要实验结果表明, gru 在训练时间上的表现明显优于 lstm，同时获得了与 lstm 模型几乎相同的精度。我们论文

的其他贡献是通过过滤不常见的活动来改进分类模型训练时间，这是一种常用的技术，例如在**自然语言处理**(nlp) 中。少

2018 年 9 月 16 日提交;最初宣布 2018 年 9 月。

评论:bpm 2018 年度研讨会论文集

86. 第 xiv:1809.05889[pdf,其他] Cs。铬

多伊 10.1109/SNPD.2018.8441123

深层学习与经典机器学习算法在恶意软件检测中的比较

作者:mohit sewak, sanjay k. sahay, hemant rathore

摘要: 近年来，深度学习在图像识别、**自然语言处理**、语言建模、**神经**机等各种人工智能应用中取得了可喜的成果翻译等。虽然一般来说，它是计算成本比传统的机器学习技术，他们的结果被认为是更有效的在某些情况下。因此，本文研究并比较了一种称为深层神经网络 (dnn) 的深层学习体系结构，并将其与经典的随机林 (rf) 机器学习算法进行了恶意软件分类。我们研究了具有四种不同特征集的经典 rf 和 dnn 架构的性能，发现无论要素输入如何，经典 rf 精度都优于 dnn。少

2018 年 9 月 16 日提交;最初宣布 2018 年 9 月。

评论:11 页, 1 图

日记本参考:ieee，第 19 页 ieeeeacis 软件工程、人工智能、网络和并行分布式计算 (snpd) 国际会议, 2018

87. 第 xiv:1809.05814[pdf,其他] Cs。CI

开发深度学习算法对糖尿病的自由文本笔记进行分类：卷积神经网络比支持向量机获得更高的精度

作者:杨博义,亚当·赖特

摘要: 卫生专业人员在审查电子健康记录 (ehr) 时可以使用**自然语言处理**(nlp) 技术。机器学习自由文本分类器可以帮助他们识别问题并做出关键的决定。我们的目标是开发深度学习神经网络算法，用于识别与糖尿病相关的 ehr 进度注意事项，并在两个机构中验证这些算法。使用的数据是从糖尿病患者中检索到的 2,000 张 ehr 进度注释，所有笔记都手动注明为糖尿病或非糖尿病患者。开发了几种深度学习分类器，并对其性能进行了对 roc 曲线 (auc) 下区域的评价。卷积神经网络 (cnn) 模型具有可分离卷积层，准确地识别了 brigham 和妇女医院测试中与糖尿病有关的音符，最高的 auc 为 0.975。深度学习分类器可用于识别与糖尿病相关的 ehr 进度注意事项。特别是，基于 cnnna 的分类器可以实现比基于 svm 的分类器更高的 auc。少

2018 年 9 月 16 日提交;最初宣布 2018 年 9 月。

评论:2018 年 9 月 15 日，向美国医学信息学协会 (jamia) 杂志提交了 9 页, 4 个数字

88. 第 xiv:1809.05724[pdf,其他] Cs。艾

利用科学问题领域的外部知识提高自然语言推理

作者:王晓燕, pavan kapanipathi, ryan musa, mo yu, kartik talamadupula, ibrahim abdelaziz, maria chang, acille fokoue, bassem makni, nicolas mattei, michael witbrock

摘要: **天然**的**使用语言**推理 (nli) 是许多**自然语言处理**(nlp) 应用的基础，包括语义搜索和问题回答。由于发布了具有挑战性的大规模数据集, nli 问题得到了极大的关注。目前对这一问题的处理方法主要侧重于基于学习的方法，这些方法只使用文本信息，以

便对给定的前提是否需要、矛盾或对特定假设保持中立进行分类。令人惊讶的是，在 nli 问题上，使用基于结构化知识的方法----人工智能的一个中心议题----并没有得到太多关注。虽然有许多开放的知识库包含各种类型的推理信息，但它们在 nli 中的使用还没有得到很好的探索。为了解决这个问题，我们提出了一个技术的组合，利用知识图，以提高在科学问题领域的 nli 问题的性能。我们介绍了将我们的技术应用于基于文本、图形和文本到图形的模型的结果，并讨论了外部知识在解决 nli 问题中的应用影响。我们的模型通过科学尾科学问题数据集，在 nli 问题上实现了新的最先进的性能。少

2018 年 9 月 15 日提交;最初宣布 2018 年 9 月。

评论:9 页, 3 个数字, 5 个表

89. 第 xiv: 1809.05679[pdf,其他] Cs。CI

文本分类的图卷积网络

作者:梁耀,毛成生,袁罗

摘要: 文本分类是**自然语言处理**中一个重要而经典的问题。已经有许多研究将卷积神经网络 (规则网格上的卷积，例如序列) 应用于分类。然而，只有数量有限的研究探索了更灵活的图形卷积神经网络 (卷积在非网格上，例如任意图) 的任务。在本工作中，我们建议使用图形卷积网络进行文本分类。我们建立了一个基于词共现和文档词关系的语料库的单一文本图，然后学习语料库的文本图卷积网络 (text gcn)。我们的文本 gcn 是用一个热表示的单词和文档初始化的，然后它在文档的已知类标签的监督下，共同学习单词和文档的嵌入。我们在多个基准数据集上的实验结果表明，没有任何外部文字嵌入或知识的香草文本 gcn 优于最先进的文本分类方法。另一方面，文本 gcn 还学习预测词和文档嵌入。此外，实验结果表明，随着训练数据百分比的降低，文本 gcn 的改进比最先进的比较方法更加突出，表明文本 gcn 对文本分类中的训练数据较少具有鲁棒性。少

2018 年 10 月 17 日提交;v1 于 2018 年 9 月 15 日提交;最初宣布 2018 年 9 月。

90. 第 189.05636[pdf,其他] Cs。CI

多伊 10.1111/gec3.12404

地理文本数据与数据驱动的地理空间语义学

作者:胡英杰

摘要: 现在的许多数据集包含地理位置和**自然语言**文本之间的链接。这些链接可以是地理标记，如地理标记推文或地理标记维基百科页面，其中位置坐标显式附加到文本。这些链接也可以是地方提及，如那些在新闻文章，旅游博客，或历史档案，其中文本隐式连接到上述地方。这种类型的数据称为地理文本数据。大量地理文本数据的提供带来了挑战和机遇。一方面，由于某些地方的非结构化文本和复杂的空间足迹，自动**处理**此类数据具有挑战性。另一方面，地理文本数据通过文本中包含的丰富信息以及文本与地理之间的特殊联系提供了独特的研究机会。因此，地理文本数据促进了各种研究，特别是数据驱动的地理空间语义方面的研究。本文讨论了地理文本数据和相关概念。本文以数据驱动研究为重点，系统地回顾了大量从地理文本数据中发现多种类型知识的研究。在文献综述的基础上，提取了广义工作流，并对今后工作面临的关键挑战进行了探讨。少

2018 年 9 月 14 日提交;最初宣布 2018 年 9 月。

评论:地理指南针, 2018

91. 第 1809. 05286[pdf,其他] Cs。简历

从自然语言处理中吸取的经验教训与美国有线电视新闻网框架的深度插值

作者:kian ghodoussi, nihar sheth , zane durante, markie wagner

摘要: 深层学习中的一个主要增长领域是卷积神经网络的研究和实现。在深层学习社区中，对卷积神经网络 (cnn) 在图像识别中的鲁棒性的一般解释是基于 cnn 能够提取局部特征的想法。然而，自然 **语言处理**等领域的最新发展表明，这种范式可能是不正确的。本文分析了美国有线电视新闻网有关领域的现状，提出了一个假设，为美国有线电视新闻网模型的鲁棒性提供了一个新的解释。从那里，我们展示了我们的方法的有效性，通过提出新的深 cnn 帧插值架构，可与最先进的插值模型的复杂性的一小部分相媲美。少

2018 年 9 月 16 日提交;v1 于 2018 年 9 月 14 日提交;最初宣布 2018 年 9 月。

评论:10 页, 5 个数字

92. **第 xiv:180 9.05053[pdf,其他] Cs。CI**

xnli: 评价跨语言句子的表达

作者:亚历克西斯·康纳、纪尧姆·兰普尔、鲁蒂·里诺特、阿迪纳·威廉姆斯、塞缪尔·鲍曼、霍尔杰·施文克、 维塞林·斯托扬诺夫

摘要: 最先进 **的自然语言处理**系统依靠注释数据形式的监督来学习有能力的模型。这些模型一般是对单一**语言**(通常是英语) 的数据进行培训的，不能在该**语言**之外直接使用。由于以每**种语言**收集数据都不现实，人们对跨语言理解 (cross-lingual) 和低资源跨**语言**转移的兴趣越来越大。在这项工作中，我们构建了 xlu **的评估**集，将多种自然**语言**推理语料库 (multinli) 的开发和测试集扩展到 15 **种语言**，包括低资源**语言**如斯瓦希里语和乌尔都语。我们希望我们的数据集，即所谓的 xnli，将通过提供一个信息丰富的标准评估任务，促进跨语言句子理解的研究。此外，我们还为多语言句子理解提供了几个基线，其中两个基于机器翻译系统，两个基线使用并行数据来训练对齐的多语言词袋和 lstm 编码器。我们发现 xnli 代表了一个实用且具有挑战性的评估套件，直接转换测试数据可在可用基线中获得最佳性能。少

2018 年 9 月 13 日提交;最初宣布 2018 年 9 月。

评论:emnlp 2018

93. **第 xiv:1809.04835[pdf] Cs。简历**

基于深层强化学习的图像描述

作者:石海超、李鹏、王波、王振宇

文摘: 近年来，加强学习的策略梯度方法已被用来训练关于自然**语言处理任务的**深度端到**端**系统。此外，随着对图像内容理解的复杂性和以**自然语言**描述图像内容的方式的多样化，图像字幕一直是一个具有挑战性的问题。据我们所知，最先进的方法遵循顺序模型的模式，如递归神经网络 (rnn)。然而，在本文中，我们提出了一种新的图像字幕结构，并进行了深度强化学习，以优化图像字幕任务。我们利用两个称为 "策略网络" 和 "价值网络" 的网络来协作生成图像的字幕。在微软 coco 数据集上进行了实验，实验结果验证了该方法的有效性。少

2018 年 9 月 13 日提交;最初宣布 2018 年 9 月。

94. **第 xiv:1809.04698[pdf,其他] Cs。CI**

学习总结放射学发现

作者:张宇豪,黛西·易丁,钱天培,克里斯托弗·曼宁,柯蒂斯 p. 朗洛茨

摘要: 放射学报告的印象部分总结了**自然语言**中的重要放射学发现，并在向医生传达这些发现方面发挥着核心作用。然而，通过总结调查结果来产生印象的**过程**对放射科医生来说是耗时的，而且容易出错。我们建议自动生成放射学印象与神经序列到序列学习。我们进一步提出了一个定制的神经模型，学习编码研究背景信息，并使用这些信息来指导解码**过程**。在从实际医院研究中收集的大量放射学报告数据集上，我们的模型优于 rouge 指标下现有的非神经和神经基线。在一个盲目的实验中，一位经董事会认证的放射科医生表示，67% 的抽样系统摘要至少与相应的人工摘要一样好，表明临床有显著的有效性。据我们所知，我们的工作是朝着这个方向迈出的第一次尝试。少

2018 年 10 月 8 日提交;v1 于 2018 年 9 月 12 日提交;最初宣布 2018 年 9 月。

评论:emnlp 2018 健康文本挖掘和信息分析讲习班 (emnlp-louhi)。代码和预培训模型可在: https://github.com/yuhaozhang/summarize-radiology-findings

95. 第 xiv:1809.04686[pdf,其他] Cs。CI

基于多语言神经机器翻译的零射击跨语言分类

作者:akiko eriguchi, melvin johnson, orhan firat , hidto kazawa, wolfgang macherey

摘要: 将表示从大型监督任务转移到下游任务在人工智能领域 (如计算机视觉和**自然语言处理**(nlp)) 中显示出很有希望的结果。与此同时，机器翻译 (mt) 的最新进展使人们能够培训多语言神经 mt (nmt) 系统，这些系统可以在多种**语言**之间进行翻译，并且还能够执行零镜头翻译。但是，很少注意利用多语言 nmt 系统所学到的表示，以便在其他 nlp 任务中实现零拍摄多语言性。在本文中，我们演示了一个简单的框架，多语言编码器分类器，通过重用编码器从多语言 nmt 系统，并将其与任务特定的分类器组件拼接，进行跨语言的跨语言迁移学习。我们提出的模型在三个基准任务 (amazon review、sst 和 snli) 的英语设置方面实现了显著改进。此外，我们的系统可以使用在培训期间看不到分类数据的新**语言**进行分类，这表明零镜头分类是可能的，而且竞争非常激烈。为了了解促成这一发现的潜在因素，我们对共享词汇的影响、nmt 的训练数据类型、分类器的复杂性、编码器表示能力和模型泛化进行了一系列分析。零拍摄性能。我们的研究结果提供了有力的证据，证明从多语言 nmt 系统中学到的陈述广泛适用于各种**语言**和任务。少

2018 年 9 月 12 日提交;最初宣布 2018 年 9 月。

96. 第 xiv:180 9.04556[pdf,其他] Cs。CI

无监督的可控制文本形式化

作者:parag jain, abhijit mishra, amar prakash azad,karthik sankaranarayanan

文摘: 我们提出了一个可控**自然语言**转换的新框架。认识到并行语料库的要求对于可控生成任务实际上是不可持续的，提出了一种无监督的培训方案。该框架的关键是一个深度神经编码器解码器，它通过辅助模块 (称为记分员) 通过文本转换知识得到增强。记分员基于现成的**语言处理**工具，根据编码器的操作来决定编码器的学习方案。我们将此框架应用于通过提高输入文本的可读性等级来形式化输入文本的文本转换任务;用户可以在运行时控制所需的形式化程度。在公共数据集上的实验证明了我们的模型在以下方面的有效性: (a) 将给定的文本转换为更正式的样式，以及 (b) 在与输入控件相关的输出文本中引入适当数量的格式。我们的代码和数据集发布供学术使用。少

2018 年 9 月 10 日提交;最初宣布 2018 年 9 月。

97. 第 1809. 04280[pdf,其他] 反渗透委员会

复杂场景中的人的指令安全导航

作者:哲胡,潘佳,范廷祥,杨瑞刚 , 迪内什·马诺查

文摘: 在本文中，我们提出了一个机器人导航算法与**自然语言**接口，使机器人通过遵循人类的指令，如 "去到餐厅，并远离人 "。我们首先将人工指令分为三种类型：目标、约束和不提供信息的短语。接下来，我们在导航过程中，以动态的方式为提取**的**目标和约束项提供接地，以处理距离太远的目标对象，而这些目标对象对于传感器的观察和像人类这样移动的障碍物的出现。特别是，对于目标短语 (例如, "去餐厅")，我们将其接地到预定义语义地图中的某个位置，并将其视为全局运动规划师的目标，该规划师在工作区中规划一个无冲突路径，供机器人遵循。对于约束短语 (例如, "远离人员")，我们通过根据对象检测模块返回的结果调整本地成本映射的值，动态地将相应的约束添加到本地规划师中。然后使用更新的成本图计算用于计算机器人安全导航的局部避碰控制。通过**将自然语言处理**、运动规划和计算机视觉相结合，我们开发的系统被证明能够成功地遵循**自然语言**导航指令。在模拟和真实场景中实现导航任务。视频可在 https://sites.google.com/view/snhi 少

2018 年 9 月 12 日提交;最初宣布 2018 年 9 月。

98. 第 xiv:1809.04047[pdf,其他] Cs。Cl

awe：文本授权的非对称构词

作者:马腾飞,吴嘉民,曹晓,孙继蒙

摘要: 文本包络是**自然语言处理**中的一项基本任务。它指的是文本片段之间的方向性关系，使 "前提" 可以推断 "假设"。近年来，深度学习方法在这一任务中取得了巨大的成功。他们中的许多人考虑了前提假设对之间的句子间词相互作用，然而，很少有人考虑这些相互作用的 "不对称"。与释义识别或句子相似性评价不同，文本包络本质上是确定前提与假设之间的方向性 (不对称) 关系。本文提出了一种简单而有效的利用非对称词嵌入来增强现有文本嵌入算法的方法。在科学尾巴和 snli 数据集上的实验结果表明，学习的非对称单词嵌入可以显著改善基于文字交互的文本包络模型。值得注意的是，拟议的 awe-deiste 模型可以获得 2.1% 的精度提高比以前的最先进的科学尾巴。少

2018 年 9 月 12 日提交;v1 于 2018 年 9 月 11 日提交;最初宣布 2018 年 9 月。

99. 第 xiv:1809.04022[pdf, ps,其他] Cs。Cl

lstm 能否学习捕获协议？巴斯克人的案例

作者:shauli ravfogel, francis m.tyers , yoav goldberg

摘要: 顺序神经网络模型是各种**自然语言处理**(nlp) 任务中的强大工具。这些模型的顺序性质提出了一个问题：这些模型在多大程度上可以含蓄地学习人类**语言**特有的等级结构，以及它们能获得什么样的语法现象？我们关注巴斯克语中的协议预测任务，作为一项需要隐含理解句子结构和获取复杂但一致的形态系统的任务的案例研究。通过分析两个句法预测任务--动词数预测和后缀恢复--的实验结果，我们发现，在巴斯克，顺序模型在协议预测方面的表现比之前的协议预测所预期的要差用英语工作。基于诊断分类器的初步发现表明，网络利用局部启发式作为句子层次结构的代名词。我们提出巴斯克协议预测任务作为具有挑战性的基准模型，试图学习规律的人类**语言**。少

2018 年 9 月 21 日提交;v1 于 2018 年 9 月 11 日提交;最初宣布 2018 年 9 月。

评论:在 2018 年 emnlp 会议上参加 "神经网络分析和解释" 研讨会

100. 第:1809.03485[pdf,其他] Cs。Cl

新闻文章思想政治工作的多视点模型

作者:vivek kulkarni, juntingye , steven skiena, william yang wang

摘要: 新闻文章的标题、内容和链接结构往往揭示其政治意识形态。然而，现有的大多数关于自动政治意识形态检测的作品只利用文本线索。借鉴神经推理的最新进展，我们提出了一个新的基于注意力的多视图模型，利用来自上述所有观点的线索，以确定一个新闻文章所体现的意识形态。我们的模型借鉴**了自然语言处理**和网络科学中表象学习的进步，从新闻文章的文本内容和网络结构中获取线索。我们根据一组基线对我们的模型进行经验评估，并表明我们的模型比最先进的 f1 分数高出 10 个百分点。少

2018 年 9 月 10 日提交;最初宣布 2018 年 9 月。

评论:10 页。emnlp 2018。增加了版权声明陈述这是作者草稿 (也注意和固定的问题与引文 (间距和可读性))

101. **第 1809. 03411[pdf,其他] Cs。Cl**

填写缺失路径: 用于识别词汇语义关系的 word 对的模拟出现和依赖路径

作者:koki washio, tsuneaki kato

摘要: 识别词对之间的词汇语义关系是**自然语言处理**中许多应用的重要任务。这个任务的主流方法之一是利用连接两个目标词的词汇句法路径，这反映了单词对的语义关系。但是，这种方法要求被考虑的单词在句子中同时出现。由于 zipf 定律的规定，这一要求几乎得不到满足，该定律规定，大多数内容词很少出现。在本文中，我们提出了新的方法与神经模型 $P(path|w_1,w_2)$来解决这个问题。我们提出的模型 $P(path|w_1,w_2)$ 可以在无监督的方式学习，并可以概括单词对和依赖路径的共存。该模型可用于增强语料库中不同时出现的词对的路径数据，并从词对中提取捕获关系信息的特征。实验结果表明，我们的方法改进了以往基于依赖路径的神经方法，成功地解决了焦点问题。少

2018 年 9 月 10 日提交;最初宣布 2018 年 9 月。

评论:11 页, naacl2018

102. **新建: 1809. 03401[pdf,其他] Cs。Cl**

神经关联分析在向量空间中捕获词汇语义关系

作者:koki washio, tsuneaki kato

摘要: 捕获向量空间中单词的语义关系有助于执行许多**自然语言处理**任务。一种很有前途的方法是利用词汇句法模式作为单词对的特征。在本文中，我们提出了这种基于模式的方法--神经潜在关系分析 (nlra) 的新模型。njra 可以概括单词对和词汇句法模式的同现现象，并获得不同时出现的单词对的嵌入。这克服了以往基于模式的模型中遇到的关键数据稀疏问题。我们测量关系相似度的实验结果表明, njra 的性能优于以前基于模式的模型。此外，当与矢量偏移模型结合使用时, njr 的性能与利用额外语义关系数据的最先进模型相当。少

2018 年 9 月 10 日提交;最初宣布 2018 年 9 月。

评论:7 页，在 emnlp2018 接受

103. **第: 1809.03385[pdf,其他] Cs。Lg**

基于深层图像字幕网络的科学有源搜索系统

作者:邱迪孔德

摘要: 火星探测器的行星探测任务很复杂，一般需要人类专家详细的任务规划，从路径到图像捕捉。美国宇航局一直在利用这一**过程**从火星上获取超过 2200 万张图像。为了提高这一**过程**的自动化程度和效率，我们提出了一个行星漫游者系统，积极寻找捕获

图像中预先确定的科学特征的突出。科学家可以用**自然语言**预先指定此类搜索任务，并将其上传到漫游者，在漫游者身上，部署的系统会不断用深层图像字幕网络捕获图像，并将自动生成的字幕与漫游者进行比较。通过某些指标预先指定的搜索任务，以便确定传输图像的优先级。作为一个有益的副作用，该系统还可以作为基于内容的搜索引擎部署到地面行星数据系统。少

2018 年 9 月 10 日提交;最初宣布 2018 年 9 月。

评论:9 页, 5 个数字, 1 个表。预打印。进行中的工作

104. **第: 1809. 03348[pdf,其他] Cs。CI**

木感: 学习可探索的语感网络的快速分离的稀疏表示和文本定义

作者:张廷云,大涌,蔡尚基,陈云农

摘要: 尽管在各种**自然语言处理**任务上取得了成功，但由于向量表示的密集，单词嵌入很难解释。本文主要从各个方面来解释嵌入，包括矢量维度中的感觉分离和定义生成。具体而言，给定一个上下文和一个目标词，我们的算法首先将目标词嵌入投影到一个高维稀疏向量，并选择能够通过编码来最好地解释目标词语义含义的特定维度上下文信息，其中目标单词的意义可以间接推断。最后，我们的算法应用 mn 生成了人类可读形式的目标词的文本定义，从而可以直接解释相应的单词嵌入。本文还介绍了一个大型、高质量的上下文定义数据集，该数据集由感官定义和每个多义词的多个示句组成，是定义建模和词感消歧的宝贵资源。实验表明, BLEU 得分和人体评价测试具有较好的性能。少

2018 年 9 月 10 日提交;最初宣布 2018 年 9 月。

105. **第: 1809. 03094[pdf,其他] lo c**

多伊10.4204/EPTCS.277。4

作为并行程序的经典证明

作者:fedico aschieri, agata ciabattoni, francesco antonio genco

摘要: 我们为经典逻辑引入了第一个证明作为并行程序对应。我们定义了一个平行的和更强大的扩展的简单类型的 lambda 演算对应于基于被排除的中间定律的分析**自然**演绎。由此产生的功能**语言**具有自然的高阶进程之间**的通信**机制，这也支持广播. 规范化过程利用减少来实现处理和传输**过程**闭包的新技术。少

2018 年 9 月 9 日提交;最初宣布 2018 年 9 月。

评论:在《2018 年程序 ganalf 》中, arxiv:1809. 02416. arxiv 管理说明: 文本与 arxiv:1607.05120 重叠

日记本参考:ettcs 277, 2018, 43-57 页

106. **第 xiv:1809. 002824[pdf,其他] Cs。红外**

多伊10.1080/13658816.2018.1458986

从地理标记房屋广告中获取当地地名的自然语言处理和地理空间聚类框架

作者:胡英杰,毛惠娜,格兰特·麦肯齐

摘要: 居住在一个地理区域的居民经常使用当地地名。这些地名可能不会记录在现有的地名录中，因为它们是白话性质的, 对覆盖大片地区 (例如整个世界) 的地名录没有意义，最近设立的地名 (例如，新开设的购物中心的名称)或其他原因。虽然并不总是有记录，但地方地名在许多应用中发挥着重要作用，从支持公众参与城市规划到在灾害应对中寻找受害者。在本文中，我们提出了一个计算框架，从地理标记的住房广告中收集

当地地名。我们利用那些发布在面向当地的网站上的广告，比如 craigslist，经常提到当地的地名。拟议的框架包括两个阶段:**自然语言处理**和地理空间集群。nlp 阶段检查住房广告的文本内容，并提取地名候选项。地理空间阶段的重点是与提取的地名候选对象相关的坐标，并执行多尺度地理空间聚类，以过滤掉非地名。我们通过将其性能与六个基线的性能进行比较来评估我们的框架。我们亦会把我们的结果与四个现有的地名录进行比较，以显示我们的架构所发现的本地地名，而这些地名是尚未记录的。少

2018 年 9 月 8 日提交;最初宣布 2018 年 9 月。

评论:2018 年国际地理信息科学杂志

107. **第 xiv:1809.02823[pdf] si**

多伊 10.1080/13658816.2017.1367797

用新闻文章提取和分析城市间的语义相关性

作者:胡英杰,叶新岳,肖世龙

摘要: 新闻文章捕捉到了我们社会的各种话题。它们不仅反映了在我们的物质世界中发生的社会经济活动，也反映了一些只存在于人们认知中的文化、人类利益和公众关切。新闻文章中经常提到城市，同一篇文章中可能有两个或两个以上的城市同时出现。这种现象的发生往往表明上述城市之间存在一定的相关性，根据新闻文章的内容，相关性可能会在不同的主题下。我们认为不同主题下的关联是语义关联。通过阅读新闻文章，可以把握城市之间的一般语义关联，然而，考虑到几十万篇新闻文章，任何人都很难甚至不可能手动阅读。本文提出了一个能够 "阅读" 大量新闻文章并提取城市间语义关联的计算框架。该框架基于**自然语言处理**模型，并采用机器学习**过程**来确定新闻文章的主要主题。我们描述了这个框架的总体结构及其各个模块，然后将其应用到实验数据集，其中有超过 50 万篇新闻文章，涵盖了美国前 100 个城市，跨越了 10 年的时间。我们在不同的主题下和多年内对提取的语义相关性进行了探索性可视化。我们还分析了地理距离对语义关联的影响，发现了不同距离衰变的影响。该框架可用于支持城市网络研究中的大规模内容分析。少

2018 年 9 月 8 日提交;最初宣布 2018 年 9 月。

评论:2017 年国际地理信息科学杂志

108. **第 1809. 002811[pdf,其他] Cs。Lg**

在线新闻中用户反应的多标签分类

作者:zacarias curi, alceu de souza britto, emerson cabrera paraiso

摘要: 互联网用户数量的增加和 web2.0 带来的强有力的互动，使得《观点挖掘》成为**自然语言处理**领域的一项重要任务。尽管有几种方法能够执行此任务，但很少使用多标签分类，其中每个示例都有一组真正的标签。这种类型的分类是有用的情况下，从读者的角度分析意见。最近，深度学习已经注册的最新的几个单一标签的问题。本文讨论了与传统的多标签分类方法相比，长期短期存储器的效率。为此，对用巴西葡萄牙文撰写的两家新闻语料库进行了广泛的测试，并作出了反应。提出了一种新的语料库 bfrc-pt。在所执行的测试中，使用分类链方法和随机林算法获得了最高数量的正确预测。在考虑类分布时，二元关联方法结合 lstm 和随机林算法获得了最佳结果。少

2018 年 9 月 8 日提交;最初宣布 2018 年 9 月。

109. **第 xiv:1809. 002796[pdf,其他] Cs。CI**

带有边界指示器的注意语义角色标注

作者:张卓生,何世霞,李祖超,赵海

摘要: 语义角色标记 (srl) 的目的是发现句子的谓词参数结构，它在**自然语言**的深度**处理**中起着至关重要的作用。本文介绍了基于依赖的 srl 简单而有效的辅助标签，以增强具有多跳自注意功能的语法无关模型。我们的语法不可知论模式通过 conll-2009 中的英语和中文基准上的最先进的模型实现了竞争性能。少

2018 年 9 月 8 日提交;最初宣布 2018 年 9 月。

110. 第 1809. 002701[pdf,其他] Cs。CI

欺骗我，如果你可以：三角挑战问题的对抗性写作

作者:eric wallace, pedrorodriguez, shi feng, jordan boyd-graber

摘要: 现代**自然语言处理**系统被吹捧为接近人类的表现。但是，现有数据集并不完善。示例是在考虑到人而不是计算机的情况下编写的，通常不能正确地暴露模型限制。我们通过开发一个新的流程来解决这个问题，它用于众包注释、对抗性书写，在这个过程中，人类与训练有素的模型进行交互，并试图打破它们。将这一注释**过程**应用于 trivia 问答会产生一个挑战集，尽管人类玩家很容易回答，但系统地阻碍了自动问答系统。在评估数据上诊断模型错误为开发更强大、更通用的问答系统提供了可操作的见解。少

2018 年 9 月 7 日提交;最初宣布 2018 年 9 月。

111. 建议: 1809. 02665[pdf] Cs。Lg

dreamnlp：一种新的基于计数草图数据流算法的临床报告元数据提取 nlp 系统：初步结果

作者:choi sanghyun, nikita ivkin, vladimir braverman, michael a. jacobs

摘要: 从电子健康记录 (ehr) 中提取信息是一项具有挑战性的任务，因为它需要事先了解报告和一些**自然语言处理**算法 (nlp)。随着 ehr 实施的数量不断增加，以高效的方式获取此类知识的难度越来越大。我们通过提出一种新的方法来分析大量的 ehrs 使用改进的计数草图数据流算法，称为 dreamnlp 来应对这一挑战。通过使用 dreamnlp，我们生成了一个字典，其中使用的是 ehr 中频繁出现的术语或重打手，与其他 nlp 程序使用的传统计数方法相比，计算内存较低。我们展示了从 ehr 中提取最重要的乳房诊断特征的一系列患者接受乳房成像。在分析的基础上，提取这些术语将有助于定义下游任务的重要功能，如精密医学的机器学习。少

2018 年 8 月 25 日提交;最初宣布 2018 年 9 月。

评论:13 页, 3 个数字，美国专利

类:e.1;e.2;F.2.2;I.2。7

112. 第 1809. 02444[pdf,其他] Cs。Lg

基于可差神经计算机对抗攻击的变形关系

作者:陈爱文,马磊,李峰，谢晓飞,刘洋,耀顺

摘要: 深度神经网络 (dnn) 虽然已成为许多新技术的推动力，并在许多尖端应用中取得了巨大的成功，但仍然容易受到对抗攻击。可分辨神经计算机 (dnc) 是一种新型的计算机，以 dnn 为中心控制器，在外部内存模块上运行，用于**数据处理**。dnc 独特的体系结构有助于其在任务中的最先进性能，这需要能够表示变量和数据结构，以及在较长时间内存储数据。然而，仍然缺乏关于对抗性例子在鲁棒性方面如何影响 dnc 的全面研究。本文针对**自然处理语言**领域中描述的一系列任务，提出了基于变形关系的对抗技术。我们表明, dnc 在 babi 逻辑问题回答任务中近乎完美的表现可以通过对抗性

注入的句子来降低。我们进一步深入研究了 dnc 内存大小在其健壮性方面的作用，并分析了导致 dnc 失败的潜在原因。我们的研究表明，目前在构建更强大的 dnc 方面面临的挑战和潜在机遇。

2018 年 9 月 7 日提交;最初宣布 2018 年 9 月。

113. **第 1809. 02428[pdf,其他] Cs。CI**

多伊 10.100/13218-018-0557-5

用于词汇分析的多任务多语言建模

作者:约翰内斯·比耶瓦

摘要: 在**自然语言处理**(nlp) 中，传统上一次考虑单个**语言**(如英语) 的单个任务 (例如词性标记)。然而，最近的工作表明，利用任务之间以及**语言**之间的相关性可能是有益的。在本文中，我研究了关联性的概念，并探讨了如何利用它来构建需要较少手动注释数据的 nlp 模型。对于包含 60 **种语言**的大量**语言**示例，将对大量的 nlp 任务进行调查。研究结果显示了联合多任务和多语言建模的潜力，并暗示了可以从这些模型中获得的语言见解。少

2018 年 9 月 7 日提交;最初宣布 2018 年 9 月。

评论:论文摘要。这是一篇在 ki/künstliche intelligenz 上发表的文章的预印。最终的经过身份验证的版本可在 https://doi.org/10.1007/s13218-018-0557-5 在线查阅:

114. **建议: 1809. 008[pdf,其他] Cs。红外**

用于视觉推理的级联互斥调制

作者:姚一群,徐嘉明,王峰, 徐波

摘要: 视觉推理是一个特殊的视觉问题，本质上是多步骤的、**构成**的，也需要密集的文本视觉交互。我们提出了 cmm: 级联互调制作为一种新的端到端视觉推理模型。cmm 包括一个多步骤的问题和图像理解**过程**。在每个步骤中，我们都使用特征式线性调制 (film) 技术，使 textual/visual t 管能够相互控制。实验表明, cmm 的性能明显优于大多数相关模型，并在两个视觉推理基准上达到了最新水平: clevr 和 nlvr，从合成语言和**自然语言**收集。消融研究证实，我们的多步框架和视觉引导**语言**调制都是这项任务的关键。我们的代码可在 https://github.com/FlamingHorizon/CMM-VR。少

2018 年 9 月 6 日提交;最初宣布 2018 年 9 月。

评论:将出现在 emnlp 2018 中

115. **第 1809. 01500[pdf, ps,其他] Cs。CI**

神经药物网

作者:n 克里斯·nikhil , shivansh mundra

摘要: 在本文中，我们描述了由团队 light 提交的用于 "社交媒体挖掘用于健康应用的社交媒体挖掘" 的共享任务的系统。以前的工作表明, lstm 在**自然语言处理任务中**取得了显著的性能。我们部署了两个 lstm 模型的组合。第一个是附加分类器的**预训练语言**模型，以单词作为输入，第二个是以字符三克为输入的注意单元的 lstm 模型。我们把这两种型号的组合称为神经-药物网。我们的系统在第二个共享任务中排名第二: 描述药物摄入量的职位的自动分类。少

2018 年 8 月 31 日提交;最初宣布 2018 年 9 月。

116. **第十四条: 1809. 01496[pdf,其他] Cs。CI**

学习性别中立的单词嵌入

作者:赵杰玉,周一超,李泽宇, 王伟, 张启伟

摘要: word 嵌入模型已成为各种**自然语言处理**(nlp) 应用中的基本组成部分。然而，在人为语料库上接受培训的嵌入已被证明继承了反映社会结构的强烈的性别陈规定型观念。为了解决这一问题，本文提出了一种新的培训程序，用于学习性别中立的单词嵌入。我们的做法旨在在文字向量的某些维度中保留性别信息，同时迫使其他方面不受性别影响。在该方法的基础上，我们生成了全球的生中性变种 (gn-gllowe)。定量和定性实验表明, GN-GloVe 在不牺牲嵌入模型功能的情况下成功隔离性别信息。少

2018 年 8 月 29 日提交;最初宣布 2018 年 9 月。

评论:emnlp 2018

117. 第十四条: 1809. 01495[pdf,其他] Cs。Cl

一种强化学习驱动的搜索性会话系统翻译模型

作者:wafa aissa, laure soulier, ludovic denoyer

文摘 面向搜索的对话系统依赖于**自然语言**(nl) 所表达的信息需求。我们在这里重点介绍了用于构建基于关键字的查询的 nl 表达式的理解。我们提出了一个增强学习驱动的翻译模型框架，该框架能够 1) 以监督的方式学习从 nl 表达式到查询的翻译, 2) 通过将翻译模型作为单词选择来克服大规模数据集的缺乏方法，并在学习**过程**中注入相关性反馈。在两个 trec 数据集上进行了实验，并概述了该方法的有效性。少

2018 年 8 月 29 日提交;最初宣布 2018 年 9 月。

评论:这是作者的预印版的作品。它张贴在这里供您个人使用，而不是重新分配。请引用将在 2018 年 emnlp 研讨会 scai 会议记录中发表的最终版本: 第二届面向搜索的对话国家 ai 国际研讨会-国际标准书号: 971-948087-75-9

118. 第 xiv: 1809. 01331[pdf,其他] Cs。Cl

对话中表达新人物的神经多语音模型

作者:shereen oraby, lena reed, sharath ts, shubhangi tandon, marilyn walker

摘要: **天然**的面向任务的对话框的**语言**生成器应该能够改变输出话语的样式，同时仍然能够有效地实现系统对话框操作及其关联的语义。虽然使用神经生成来培训会话代理的响应生成组件有望简化在新领域产生高质量响应的**过程**，但据我们所知，几乎没有进行调查神经生成器的任务导向的对话框，可以改变他们的响应风格，我们知道没有实验的模型，可以产生不同的风格不同的风格，而在训练中看到，同时仍然维护语义保真度的输入意思是表示。在这里，我们表明，一个模型，被训练，以实现一个单一的风格人格目标可以产生输出，结合了风格的目标。我们仔细评估多语音输出的语义保真度，以及与原始训练风格所特有的语言特征的异同。我们表明，与我们的预测相反，学到的模型并不总是简单地插值模型参数，而是产生不同的风格，并根据他们所训练的个性而新颖。少

2018 年 9 月 5 日提交;最初宣布 2018 年 9 月。

评论:2018 年互动

119. 第 xiv: 1809. 01316[pdf,其他] Cs。Lg

多伊 10.114/32626266.3271712

学习用户首选项和了解事件计划的日历上下文

作者:kim donghyeon, jjhyuklee, donghee choi, jaeon choi, jaewoo kang

摘要: 随着在线日历服务在全球范围内越来越受欢迎，日历数据已成为了解人类行为的最丰富的上下文来源之一。但是，即使在开发联机日历的情况下，事件调度仍然非常耗时。尽管基于机器学习的事件调度模型在一定程度上具有自动调度**过程**，但它们往往无法理解微妙的用户首选项和复杂的日历上下文，这些内容都是**用自然编写的事件标题语言**。在本文中，我们提出了神经事件调度助手 (neesa)，它直接从原始在线日历中学习用户首选项并了解日历上下文，以便实现完全自动化和高效的事件调度。我们利用超过 593k 日历事件为 neesa 学习调度个人事件，我们进一步利用 nesa 进行多与会者事件调度。nesa 成功地融合了深度神经网络，如双向长期短期存储器、卷积神经网络和公路网，用于了解每个用户的偏好，并了解基于**自然的**日历上下文**语言**。实验结果表明，在个人事件调度任务和多与会者事件调度任务的各种评价指标方面，neesa 的性能明显优于以往的基线模型。我们的定性分析显示了 nesa 中每一层的有效性，并了解了用户偏好。少

2018 年 10 月 17 日提交;v1 于 2018 年 9 月 5 日提交;最初宣布 2018 年 9 月。

评论:cikm 2018

120. **第 09iv:1809. 00794[pdf,其他] Cs。CI**

texar: 用于文本生成的模块化、多功能和可扩展工具包

作者:李婷胡志兰,石浩然，杨子超，谭伯文,赵天成 ,何俊贤, 王文涛，于兴江，李辉,迪王, 马学哲,刘赫克托·梁小丹,朱万荣，德文德拉·辛格·萨昌, 埃里克·p·兴

摘要: 我们介绍了 texar，这是一个开源工具包，旨在支持将任何输入**转换为自然语言**的一系列文本生成任务，如机器翻译、摘要、对话、内容操作等.考虑到模块化、多功能性和可扩展性的设计目标, texar 提取了各种任务和方法背后的常见模式，创建了一个高度可重用模块和功能的库，并允许任意模型体系结构和算法范例。在 texar 中，模型体系结构、损耗和学习**过程**被完全分解。高概念级的模块可以自由组装或插上。这些功能使 texar 特别适合研究人员和从业者进行快速原型设计和实验，以及促进跨不同文本生成任务的技术共享。我们提供案例研究，以证明该工具包的使用和优势。texar 在 https://github.com/asyml/texar 根据 apache 许可证 2.0 发布。少

2018 年 9 月 4 日提交;最初宣布 2018 年 9 月。

评论:14 页;github: https://github.com/asyml/texar

121. **第 09iv:809.00640[pdf,其他] Cs。CI**

认知行为疗法对心理健康概念的语言理解的深度学习

作者:lina rojas-barahona, Bo-Hsiang tseng, yinpeidai, clare mansfield, osman ramadan, stefan ultes, michael crawford, milica gasic

摘要: 近年来，我们看到了对单词和句子的深入学习和分布式表示，对一些**自然语言处理**任务产生了影响，如相似性、包衣和情感分析。在这里，我们将介绍一个新的任务:理解从认知行为治疗 (cbt) 衍生的心理健康概念。我们根据 cbt 原理定义了一个心理健康本体，注释了一个大的语料库，在这个语料库中，这种现象被展示出来，并使用深度学习和分布式表示来进行理解。结果表明，在这一艰巨任务中，深度学习模型与单词嵌入或句子嵌入相结合的性能明显优于非深度学习模型。这一理解模块将是提供治疗的统计对话系统的重要组成部分。少

2018 年 9 月 3 日提交;最初宣布 2018 年 9 月。

评论:接受在 2018 年《卢伊岛：第九次健康文本挖掘和信息分析国际讲习班上出版

122. 第 09iv:1800.00589[pdf,其他] Cs。Cl

基于语义角色标注的移位提取与推理

作者:daniel loureiro , alípio mário 豪尔赫

摘要: 常识推理对于**自然语言处理**的进步越来越重要。虽然 "嵌入" 一词非常成功，但他们无法解释 "咖啡" 和 "茶" 的哪些方面使它们相似，或者它们如何与 "商店" 有关。本文提出了一个明确的词表示，它建立在分布假设的基础上，从语义角色来表示意义，并允许从它们的网格中推断关系，这得到了基于功能的多值假设的支持。我们发现，我们的模型改进了无监督单词相似任务的最新技术，同时允许从相同的向量空间直接推断新关系。少

2018 年 9 月 3 日提交;最初宣布 2018 年 9 月。

评论:在 fever-emnlp 2018 接受

123. 第 1809. 00072[pdf,其他] cs et

rx-caff: 评估和训练电阻式横杆深度神经网络的框架

作者:shubham jain, abhronil sengupta, kaushikroy, anand Raghunathan

文摘: 深神经网络 (dnn) 被广泛用于在语音、图像和**自然语言处理**中执行机器学习任务。dnn 的高计算和存储需求导致了对节能实现的需求。电阻式横杆系统由于能够紧凑、高效地实现原始 dnn 运算，即矢量矩阵乘法，已成为很有前途的候选系统。然而，在实际应用中，电阻横杆的功能可能会与理想的抽象有很大的不同，因为器件和电路级的非理想性，如驱动器电阻、传感电阻、潜行路径和互连寄生。尽管 dnn 对其计算中的错误有些容忍，但仍有必要评估由于横杆非理想性而引入的错误对 dnn 精度的影响。遗憾的是，在具有 2.6-155 亿个突触连接的大型 dnn 的情况下，设备和电路级模型是不可行的。在这项工作中，我们提出了一个快速和准确的仿真框架，使训练和评估大规模的 dnn 基于电阻横杆的硬件织物。我们提出了一种快速横杆模型 (fcm)，该模型能够准确捕获由于非理想性而产生的误差，同时比电路仿真快 5 个数量级。我们开发了 rx-cafe，这是流行的 caffe 机器学习软件框架的增强版本，用于在横梁上培训和评估 dnn。我们使用 rx-cafe 来评估为 imagenet 数据集设计的大规模图像识别 dnn。我们的实验表明，横杆非理想性可显著降低 dnn 精度 9.6%-32%。据我们所知，这项工作是对大型 dnn 在电阻横梁上的准确性的首次评估，并强调需要进一步努力应对非理想性的影响。少

2018 年 8 月 31 日提交;最初宣布 2018 年 9 月。

124. 第 180.00039[pdf,其他] cse

全面召回、语言处理和软件工程

作者:zheyu, tim menzies

文摘: 一类广泛的软件工程问题可以概括为 "总召回问题"。本文认为，识别和探索软件工程中的总召回**语言处理**问题是一项具有广泛适用性的重要任务。为了证明这一点，我们表明，通过应用和调整最先进的主动学习和文本挖掘，解决总召回问题，可以帮助解决两个重要的软件工程任务: (a) 支持大型文献评论和 (b) 识别软件安全漏洞。此外，我们推测 (c) 测试用例优先级和 (d) 静态警告识别也可归类为总召回问题。"完全召回 " 在软件工程中的广泛适用性表明，存在一些基础框架，不仅包括**自然语言处理**，还包括广泛的重要软件工程任务。少

2018 年 8 月 31 日提交;最初宣布 2018 年 9 月。

评论:4 页, 2 个数字。提交给 2018 年 NL4SE@ESEC/FSE

125. 第 1808. 10822[pdf,其他] Cs。简历

查看颜色: 用于分类的学习语义文本编码

作者:shah nawaz, alessandro calefati, muhammad kamran janjua, ignazio gallo

摘要: 我们用这项工作回答的问题是: 我们能否将文本文档转换为图像, 以利用最佳图像分类模型对文档进行分类? 为了回答这个问题, 我们提出了一种新的文本分类方法, 该方法利用嵌入字和卷积神经网络 (cnn) 的功能, 将文本文档转换为编码图像, 并成功地应用于图像分类。我们通过在一些著名的文本分类基准数据集上获得有希望的结果来评估我们的方法。这项工作允许应用许多先进的 cnn 架构开发的计算机视觉自然 **语言处理**。我们在多模态数据集上测试了所提出的方法, 证明可以使用单个深部模型来表示同一要素空间中的文本和图像。少

2018 年 8 月 31 日提交;最初宣布 2018 年 8 月。

评论:9 页。正在 ijdar 审查

126. 第 xiv:1808. 10685[pdf,其他] Cs。CI

从开放式商业调查问题中提取关键词

作者:barbara mcillivray, gard jenset, dominik heil

摘要: 不限成员名额的调查数据是研究和作出商业决定的重要基础。收集和手动分析自由文本调查数据的费用通常高于收集和分析由多项选择问题的答案组成的调查数据。然而, 自由文本数据允许新的内容在预定义的类别之外表达, 是对人们意见有价值的新见解的来源。同时, 调查总是对所研究的实体的**性质**作出本体论假设, 这具有至关重要的道德后果。人类的解释和意见只能用文本数据来源在其丰富的内容中得到适当的确定;如果对这些来源进行适当的分析, 人类和社会实体的基本语言**本质**就得到了保障。**天然**的**使用语言加工过程**(nlp) 通过自动化自然 **语言**分析, 从而允许对人类判断进行有见地的调查, 为应对这一道德商业挑战提供了可能性。我们提出了一个计算管道, 用于分析对调查中的开放式问题的大量回答, 并提取适当代表人们意见的关键字。该管道满足了在商业软件和定制分析范围之外执行此类任务的需要, 在最先进的系统中超越了性能, 并以透明的方式执行这一任务, 以便对其进行审查和公开分析中的潜在偏差。遵循开放数据科学的原则, 我们的代码是开源的, 可推广到其他数据集。少

2018 年 8 月 31 日提交;最初宣布 2018 年 8 月。

评论:1 图

127. 第 1808. 10603[pdf,其他] cs.PL

非自由数据类型的非线性模式匹配与回溯

作者:satoshi egi, yuichi nisiwaki

摘要: 非自由数据类型是其数据没有规范形式的数据类型。例如, 多集是非免费数据类型, 因为多集{a,b,b}有另外两个等价但字面上不同的形式{b,a,b}和{b,b,a}.模式匹配是众所周知的提供了一个方便的工具集, 以处理此类数据类型。尽管到目前为止, 已经提出了许多关于实际编程**语言**模式匹配和实现的研究, 但我们观察到, 这些研究都不符合实际模式匹配的所有标准, 如下所示: i)非线性模式回溯算法的效率, 二) 匹配过程的可扩展**性**, 以及 iii) 模式中的多态性。本文旨在设计一种满足上述三个标准的面向模式匹配的编程**语言**。该**语言**具有干净的类似架构的语法和高效、可扩展的模式匹配语义。这种编程**语言**对于**处理**复杂的非自由数据类型特别有用, 这些类型不仅包括多

集和集，还包括图形和符号数学表达式。我们讨论了我们的实用模式匹配标准的重要性，以及我们的**语言**设计是如何自然地从这些标准**中产生**的。拟议**的语言**已经实现，并作为埃格森编程**语言开源.**少

2018 年 8 月 31 日提交;最初宣布 2018 年 8 月。

评论:20 页，2018 年

128. 第 1808. 10503[pdf,其他] Cs。CI

可解释序列分类的迭代递归注意模型

作者:martin tutek, jan šnajder

文摘: **天然**的**的语言加工**的引入极大地受益于注意机制的引入。但是，对于涉及一系列推理步骤的任务，标准注意模型的可解释性有限。我们描述了一个迭代递归注意模型，该模型通过重用以前计算的查询的结果来构造输入数据的增量表示。我们对情绪分类数据集的模型进行了培训，并展示了其识别和组合输入的不同方面的能力，同时获得接近最新水平的性能。少

2018 年 8 月 30 日提交;最初宣布 2018 年 8 月。

评论:在 emnlp 2018 中举办 nlp 研讨会，共 7 页，5 个数字，分析和解释神经网络

129. 第 xiv:1808. 09935[pdf,其他] Cs。Lg

基于附件的神经文本分割

作者:pinkesh badjatiya, litton j k 鲁西金克尔, manish gupta , vasudeva varma

摘要: 文本分割在各种**自然语言处理**(nlp) 任务 (如摘要、上下文理解、文档索引和文档噪声去除) 中起着重要的作用。以前此任务的方法需要手动功能工程、巨大的内存要求和较大的执行时间。据我们所知，本文是第一个提出一种新的监督神经方法进行文本分割的方法。具体而言，我们提出了一个基于注意的双向 lstm 模型，其中学习句子嵌入使用 cnn 和段预测的基础上的上下文信息。此模型可以自动处理可变大小的上下文信息。与现有的竞争基线相比，该模型显示三个基准数据集的 windiff 分数性能提高了约 7%。少

2018 年 8 月 29 日提交;最初宣布 2018 年 8 月。

130. 第 1808. 09861[pdf,其他] Cs。CI

最小资源的神经跨语言命名实体识别

作者:谢家腾,杨志林,格雷厄姆·诺伊比希,诺亚 a.史密斯，海梅·卡博内尔

文摘: 对于没有附加注释资源的语言，在不受监督的**情况**下从资源丰富的**语言**进行自然**语言处理**模型 (如命名实体识别 (ner)) 将是一个吸引人的问题能力。然而，不同**语言**之间的单词和单词顺序的差异使其成为一个具有挑战性的问题。为了改进**跨语言**词汇项的映射，我们提出了一种基于双语单词嵌入的查找翻译的方法。为了提高词序差异的鲁棒性，我们建议使用自我关注，这允许在词序方面有一定程度的灵活性。我们证明，这些方法在跨语言环境下，在通常测试的**语言**上实现了最先进或有竞争力的 ner 性能，所需资源比过去的方法要低得多。我们还评估了将这些方法应用于低资源语言维吾尔语的挑战. 少

2018 年 9 月 11 日提交;v1 于 2018 年 8 月 29 日提交;最初宣布 2018 年 8 月。

评论:emnlp 2018 长纸

131. 第 1808. 09551[pdf, ps,其他] Cs。CI

解释用于语言级预测的特征感知神经网络：它们是否发现语言规则?

作者:fréderic godin, kris demuynck, juni dambre, wesley de neve, thomas demeester

摘要: 在不同的基于神经网络的**自然语言处理算法中**，目前使用的是特征级特征。然而，对这些模型学习的字符级模式了解甚少。此外，在缺乏定性分析的情况下，模型往往只进行定量比较。在本文中，我们研究了哪些字符级模式神经网络学习，以及这些模式是否与手动定义的单词分段和注释一致。为此，我们将上下文分解技术 (murdoch 等人，2018 年) 扩展到卷积神经网络，使我们能够比较卷积神经网络和双向长短期记忆网络。我们对这些模型进行了评估和比较，以完成三种不同**语言**的形态标记任务，并表明这些模型含蓄地发现了可理解的语言规则。我们的实现可以在 https://github.com/FredericGodin/ContextualDecomposition-NLP 找到。少

2018 年 8 月 28 日提交;最初宣布 2018 年 8 月。

评论:2018 年被接受

132. 第 1808.09500[pdf] Cs。CI

用形态和语音子词表示使词汇嵌入适应新语言

作者:adti chaudhary, chuntingzhou, lori leevin, graham neubig, david r.mortensen , jaime g. carbonell

摘要: **自然语言处理**(nlp) 的大量工作都是针对资源丰富的**语言**，这使得对资源较少的新**语言**的泛化具有挑战性。我们提出了两种方法，通过使用语言动机的子词单位，调整连续的单词表示，提高对低资源**语言**的泛化：音素、语素和图形。我们的方法既不需要并行语料库，也不需要双语词典，与以前依赖这些资源的方法相比，它的性能有了显著的提高。我们展示了我们对四**种语言**(即维吾尔语、土耳其语、孟加拉语和印地语) 的命名实体识别方法的有效性，其中维吾尔语和孟加拉语是资源较低的**语言**，并在机器上进行实验翻译。利用转述和转移学习的子词使我们的维吾尔语增加了 + 15.2 ner f1，为孟加拉语提供了 + 9.7 f1。我们还展示了单语设置的改进，我们在这里实现了 (avg.) + 3 f1 和 (avg.) + 1.35 BLEU。少

2018 年 8 月 28 日提交;最初宣布 2018 年 8 月。

评论:2018 年被接受

133. 第 1808.09419[pdf,其他] Cs。CI

识别格式良好的自然语言问题

作者:manaal faruqui, Dipanjan das

摘要: 了解搜索查询是一个难题，因为它涉及到处理用户普遍发布的 "单词沙拉" 文本。但是，如果查询类似于格式良好的问题,**则自然语言处理管道能够**执行更准确的解释，从而减少下游复合误差。因此，确定查询是否格式正确可以增强对查询的理解。在这里，我们将介绍一个新的任务，即识别一个格式良好的**自然语言**问题。我们构建并发布了 25 100 个可公开的问题数据集，这些问题分为格式良好和非格式良好的类别，并在测试集中报告了 0.7% 的准确性。我们还证明了我们的分类器可以用来提高神经序列到序列模型的性能，以产生阅读理解的问题。少

2018 年 8 月 28 日提交;最初宣布 2018 年 8 月。

日记本参考:2018 年 emnlp 项目

134. 第 1808.09408[pdf,其他] Cs。CI

文本的神经表示

作者:maximin coavoux, shashi narayan, shay b. cohen

文摘: 本文在隐私保护的背景下，讨论了对**自然语言处理**(nlp) 深度学习系统的对抗攻击。我们研究一种特定类型的攻击: 攻击者窃听神经文本分类器的隐藏表示，并尝试恢复有关输入文本的信息。在神经网络的计算在多个设备之间共享的情况下，可能会出现这种情况，例如，某些隐藏的表示形式由用户的设备计算并发送到基于云的模型。我们通过攻击者准确预测其中的特定私人信息的能力来衡量隐藏表示的隐私，并描述隐私与神经表示的效用之间的权衡。最后，我们提出了几种基于改进训练目标的防御方法，并表明它们改善了神经表示的隐私。少

2018 年 8 月 28 日提交;最初宣布 2018 年 8 月。

评论:emnlp 2018

135. 第 1808. 09397[pdf] Cs。红外

医学: 临床语义相似性的一种资源

作者:王燕山,纳维德·阿夫扎尔,孙阳·傅善大 , 王立伟, 沈飞辰, 马吉德·拉斯泰格-莫贾拉德,刘洪芳

摘要: 电子健康记录 (ehr) 的广泛采用使人们能够利用 ehr 数据实现广泛的应用。然而, ehr 数据的有意义的使用在很大程度上取决于我们是否能够有效地提取和整合嵌入在临床文本中的信息，因为**自然语言处理**(nlp) 技术是必不可少的。测量文本片段之间语义相似性的语义文本相似性 (sts) 在许多 nlp 应用中起着重要的作用。在一般的 nlp 域中, sts 共享任务提供了大量的文本代码段对，并在不同的域中进行手动注释。在临床领域, sts 可以使我们检测和消除冗余信息，这些信息可能导致认知负担的减轻和临床决策**过程**的改进。本文阐述了我们为医疗领域的 sts 资源组装所做的努力。它包括总共 174, 629 对句子对收集从一个临床语料库在梅奥诊所。两位语义相似度评分为 0-5 (低至高相似度) 的医学专家对包含 1, 068 对句子对的 medsts (medsts _ ann) 的一个子集进行了注释。我们进一步分析了 medsts 语料库中的医学概念，并在 medsts _ ann 语料库上测试了四个 sts 系统。在未来，我们将通过发布 medsts _ ann 语料库来组织一个共享的任务，以激励社区解决现实世界中的临床问题。少

2018 年 8 月 28 日提交;最初宣布 2018 年 8 月。

136. 建议: 1808. 09357[pdf,其他] Cs。Cl

理性复发

作者:郝鹏, roy schwartz, sam 汤姆森, noah a. smith

摘要: 尽管神经模型在**自然语言处理**中取得了巨大的经验成功，但许多神经模型缺乏伴随着经典机器学习方法的强烈直觉。近年来，卷积神经网络 (cnn) 和加权有限状态自动机 (wfsa) 之间存在联系，从而产生了新的解释和见解。在这项工作中，我们展示了一些递归神经网络也共享与 wfsa 的这种连接。我们将这种连接正式描述，将理性递归定义为可写入有限 wfsa 集的正向计算的递归隐藏状态更新函数。我们表明，最近的几个神经模型使用合理的递归。我们的分析提供了这些模型的全新视图，并有助于设计新的神经架构，从 wfsa 中汲取灵感。我们提出了一个这样的模型，它在**语言**建模和文本分类方面的表现优于最近的两个基线。我们的研究结果表明，从 wfsa 等经典模型中转移直觉可以成为设计和理解神经模型的有效方法。少

2018 年 8 月 28 日提交;最初宣布 2018 年 8 月。

评论:emnlp 2018

137. 第 1808. 09315[pdf,其他] Cs。Cl

具有递归神经滤波器的卷积神经网络

作者:杨毅

摘要: 我们引入了一类利用递归神经网络 (rnn) 作为卷积滤波器的卷积神经网络 (cnn)。卷积滤波器通常是作为线性仿射变换实现的，然后是非线性函数，它不能考虑**语言**组合。因此，它限制了通常用于**自然语言处理**任务的高阶筛选器的使用。在这项工作中，我们用 rnn 对卷积滤波器进行建模,**这些滤波器自然地**捕获语言中的组合性和长期依赖关系。我们展示了配备了递归神经过滤器 (rnf) 的简单 cnn 架构所取得的结果与斯坦福情绪树库上公布的效果相当，并提供了两个答案句子选择数据集。少

2018 年 8 月 28 日提交;最初宣布 2018 年 8 月。

评论:被 emnlp 2018 作为简短文件接受

138. 第 xiv:1808. 09092[pdf,其他] Cs。Cl

基于自动相关神经网络的不流畅性检测

作者:paria jamshid lou, peter anderson, mark johnson

文摘: 近年来,**自然语言处理**社区已经从特定于任务的功能工程，即研究人员发现各种任务的临时特征表示，而倾向于使用通用方法，学习输入表示自己。然而，在自发的语音记录中检测不流利的最先进的方法目前仍然取决于一系列手工制作的特征，以及从语言等预先存在的系统的输出中获得的其他表示形式模型或依赖关系解析器。作为替代方案，本文提出了一种简单而有效的自动不流利检测模型，称为自动相关神经网络 (acnn)。该模型使用卷积神经网络 (cnn)，并在最低层使用一个新的自相关运算符进行扩充，该运算符可以捕获修复语音不稳定的特征的 "粗略复制" 依赖关系的种类。在实验中, acnn 模型在不流利检测任务上优于基线 cnn, f-分数增加了 5%，接近以前在这一任务上的最佳结果。少

2018 年 10 月 27 日提交;v1 于 2018 年 8 月 27 日提交;最初宣布 2018 年 8 月。

139. 第 xiv:1808. 09040[pdf,其他] Cs。Cl

知识图的一次性关系学习

作者:熊文汉, 莫宇,张世宇, 郭晓晓,王威廉

摘要: 知识图 (kg) 是各种**自然语言处理**应用程序的关键组件。为了进一步扩大 kg 的覆盖面，以前关于知识图完成的研究通常需要为每个关系提供大量的培训实例。然而，我们注意到，长尾关系实际上在 kg 中更为常见，而这些新增加的关系往往没有许多已知的培训倍数。在这项工作中，我们的目标是在一个只有一个培训实例的具有挑战性的环境下预测新的事实。我们提出了一个一次性的关系学习框架，它利用嵌入模型所提取的知识，并通过考虑学习的嵌入和单跳图形结构来学习匹配度量。从经验上讲，与现有的嵌入模型相比，我们的模型带来了相当大的性能改进，并且在处理新添加的关系时消除了对嵌入模型进行再培训的需要。少

2018 年 8 月 27 日提交;最初宣布 2018 年 8 月。

评论:emnlp 2018

140. xiv:1808.08850[pdf, ps,其他] Cs。Cl

基于窗口的句子边界评估

作者:carlos-emiliano gonzlez-galcarardo, juan-manuel torrens-moreno

摘要: 句子边界检测 (sbd) 一直是一个主要的研究主题，因为自动语音识别记录已被用于进一步**的自然语言处理**任务，如部分语音标记，问题回答或自动汇总。但评价呢?做标准的评估指标，如精度，召回，f-分数或分类错误;更重要的是，在文字记录的最终应用下，根据独特的参考评估自动系统就足以得出 sbd 系统的性能? 本文提出了一种基于窗口的句子边界评估 (wisebe)，这是一种基于多参考 (dis) 协议的句子边界检测系统的半监督度量方法。我们使用 wisebe 和标准指标在一组 youtube 成绩单上评估和比较不同 sbd 系统的性能。这种双重评估使人们了解 wisebe 如何成为 sbd 任务更可靠的指标。少

2018 年 8 月 27 日提交;最初宣布 2018 年 8 月。

评论:在 2018 年第 17 届墨西哥人工智能国际会议 (micai) 会议上

141. 第 1808. 08703[pdf,其他] Cs。CI

利用浏览思维载体通过对抗性训练生成文本

作者:afroz ahamad

摘要: 在过去几年中，由于生成对抗性网络 (gans) 的形成，生成模型取得了各种进展。在与图像生成和风格转换有关的各种任务中，有机生物已被证明具有极好的性能。在**自然语言处理**领域，单词嵌入 (如单词 2vec 和 glovee) 是在文本数据上应用神经网络模型的最先进的方法。有人试图利用带有单词嵌入的甘醇生成文本。本文提出了一种基于梯度惩罚函数和 f-度量的利用 skip-wuing 句子嵌入与 gans 一起生成文本的方法。使用基于输入信息的文本的 gans 句子嵌入的结果与使用单词嵌入的方法相当。少

2018 年 8 月 27 日提交;最初宣布 2018 年 8 月。

评论:版本 1. 6 页, 1 个数字, 2 个表格

142. 第 1808. 08575[pdf,其他] Cs。CI

用于关键字生成的标题导向编码

作者:王晨,高一凡,张佳妮,欧文国王,刘先生

摘要: 关键字生成 (kg) 旨在生成一组给定文档的关键短语，这是**自然语言处理**(nlp) 中的一项基本任务。以前的大多数方法都以提取的方式解决了这个问题，而最近，在使用深层神经网络的生成设置下进行了几次尝试。但是，最先进的生成方法只是平等地对待文档标题和文档主体，而忽略了标题在整个文档中的主导作用。为了解决这个问题，我们引入了一种新的模型，称为标题引导网络 (tg-net)，用于基于编码器-解码器体系结构的自动关键字生成任务，具有两个新功能: (i) 标题另外用作类似于查询的输入, (ii) a 标题引导编码器收集从标题到文档中每个单词的相关信息。在一系列 kg 数据集上的实验表明，我们的模型优于最先进的模型，具有较大的利润，特别是对于标题长度比率非常低或非常高的文档。少

2018 年 9 月 6 日提交;v1 于 2018 年 8 月 26 日提交;最初宣布 2018 年 8 月。

143. 第 1808. 08357[pdf] Cs。红外

tux 博士: ubuntu 用户的问答系统

作者:bijil abraham phillip, manas jog, aprv milind upasani

摘要: 各种论坛和问答 (问答) 网站可在线使用，使 ubuntu 用户能够找到与他们的查询类似的结果。但是，搜索结果通常非常耗时，因为它要求用户从大量问题中查找与他

/她的查询相关的特定问题实例。在本文中，我们为 ubuntu 用户提供了一个名为 tux 博士的自动问答系统，该系统旨在通过从在线数据库中选择最相似的问题来回答用户的问题。该原型在 python 中实现，并将 nltk 和 corenlp 工具用于**自然语言处理**。原型的数据来自 askubuntu 网站，其中包含大约 150k 的问题。从对原型的人工评价中获得的结果很有希望，同时也提供了一些有趣的改进机会。少

2018 年 8 月 25 日提交;最初宣布 2018 年 8 月。

144. **第 xiv:1808. 08316[pdf,其他] Cs。红外**

动态实体相关性排序的三元神经模型

作者:tu ngoc nguyen, tuan tran , wolfgang nejdl

摘要: 测量实体相关性是许多**自然语言处理**和信息检索应用的一项基本任务。以前的工作经常研究静态环境中的实体相关性和无监督的方式。然而，现实世界中的实体往往涉及许多不同的关系，因此实体关系随着时间的推移非常动态。在本工作中，我们提出了一种基于神经网络的动态实体关联方法，利用集体关注作为监督。我们的模型能够在一个联合框架中学习丰富和不同的实体表示。通过对大型数据集的大量实验，证明了我们的方法比竞争基线取得了更好的结果。少

2018 年 9 月 6 日提交;v1 于 2018 年 8 月 24 日提交;最初宣布 2018 年 8 月。

评论:在 comnll 2018 论文集中

145. **第: 1808. 07724[pdf,其他] Cs。Cl**

使用多感 lstm 将文本映射到知识图形实体

作者:dimitri kartsaklis, mohammad taher pilehvar, nigel collier

文摘: 本文讨论了**将自然语言**文本映射到知识库实体的问题。映射**过程**是作为一个短语或句子的组合，到从知识图获得的多维实体空间中的一个点。组合模型是一种 lstm，它在输入字嵌入 (多感 lstm) 上具有动态消歧机制，解决多义问题。此外，知识库空间是通过从具有文本特征增强的图形中收集随机游走来准备的，而文本特征是文本和知识库实体之间的一组语义桥梁。在大规模文本到实体映射和实体分类任务上展示了这项工作的思想，并取得了最先进的结果。少

2018 年 8 月 23 日提交;最初宣布 2018 年 8 月。

评论:可在 2018 年 emnlp (主要会议) 上发言

146. **特别报告: 1808.07604[pdf, ps,其他] Cs。Cl**

利用风格相关性对多标签音乐风格进行评论驱动分类

作者:赵光祥,徐晶晶,曾琪,任宣城

文摘: 本文探讨了一种新的**自然语言处理**任务--评论驱动的多标签音乐风格分类。这项任务要求系统根据其在网站上的评论来识别多种风格的音乐。最大的挑战在于复杂的音乐风格关系。它给许多多标签分类方法带来了失败。针对这一问题，我们提出了一种新的深度学习方法，以自动学习和利用风格相关性。该方法由两部分组成: 基于标签图的神经网络和基于相关的连续标签表示的软训练机制。实验结果表明，我们的方法对建议数据集上的基线有了很大的改进。特别是微 f1 从 53.9 提高到 64.5，单误差从 30.5 减少到 22.6。此外，可视化分析表明，我们的方法在捕捉风格相关性方面表现良好。少

2018 年 8 月 22 日提交;最初宣布 2018 年 8 月。

147. **第: 1808.07383[pdf,其他] Cs。Lg**

动态自注意：句子嵌入的动态计算单词注意力

作者:yoon deunsol, dunboklee, ssnkeun lee

摘要: 本文提出了一种新的句子嵌入自注意机制--动态自注意 (dsa)。我们通过修改胶囊网络中的动态路由 (sabouretal.2017) 来设计自然**语言处理**的 dsa。dsa 关注具有动态权重向量的信息性单词。我们在斯坦福**自然语言**推理 (snli) 数据集中的句子编码方法中获得了新的最先进的结果，参数最少，同时在斯坦福情绪树库 (sst) 中显示了比较结果数据。少

2018 年 8 月 22 日提交;最初宣布 2018 年 8 月。

评论:7 页, 4 个数字

148. **第: 1808.07235[pdf,其他] Cs。CI**

为文本分类寻找情感的良好表现

作者:吉和公园

摘要: 机器正确地解释人类的情感以促进人机的交流是很重要的，因为情感是人与人之间交流的重要组成部分。情感的一个方面反映在我们使用的**语言**中。如何在文本中表现情感是**自然语言处理**(nlp) 中的一个挑战。尽管连续向量表示 (如 word2vec) 已成为 nlp 问题的新规范，但它们的局限性在于，它们没有考虑到情绪，并且可能无意中包含对某些身份 (如不同性别) 的偏见。本文的重点是通过明确考虑文本内部的情感和模型偏差在训练**过程**中的表达，来改进单词和句子两个层次的现有表现形式。我们改进的表示可以帮助构建更强大的机器学习模型，用于与情感相关的文本分类，如情绪分析和辱骂**性语言**检测。我们首先提出了一种叫做情感词向量 (evec) 的表示形式，它是从一个具有情绪标记语料库的卷积神经网络模型中学习的，该模型是使用哈希标记构造的。其次，我们扩展到学习句子级表示与一个巨大的文本语料库与识别表情符号的伪任务。结果表明，通过对数以百万计的推文进行了演示，并带有弱监管标签 (如哈希标签和表情符号)，我们可以更有效地解决情感分析任务。最后，作为现有方法表示中模型偏差的例子，我们探讨了滥用**语言**自动检测的一个具体问题。我们通过进行实验来测量和减少表示中的这些偏差，以建立更可靠的分类模型，从而解决各种神经网络模型中的性别偏见问题。少

2018 年 8 月 22 日提交;最初宣布 2018 年 8 月。

评论:香港科技硕士论文, 87 页

日记本参考:香港科技大学硕士论文, 2018

149. **决议: 1808. 07166[pdf, ps,其他] Cs。CI**

多伊 10.1016/j.physa.2018.08.031

使用频率分布来确定有争议音素的状态;西班牙语中的半声纳应用

作者:manuel ortega-rodríguez, hugo solís-sánchez, ricardo gamboa-alfaro

摘要: 利用**自然语言**是复杂的系统这一事实，本文提出了一种基于频率分布的直接方法，在决定有问题的音素的状态时可能会很有用。语言学中的公开问题。主要的概念是，**自然语言**，可以从一个复杂的角度考虑作为**信息处理**机器，并设法设置适当的冗余级别，已经 "作出选择" 是否语言单位是否是音素，这将反映在频率与等级图的更大平滑度上。对于我们选择研究的特殊情况，我们的结论是，将西班牙半声纳/w/w/w/视为与元音对应的音素是合理的，一方面，也可能将半声纳/音素视为与元音的独立音素对

应的//，另一方面。由于**语言**一直是复杂性研究的中心议题，这一讨论还使我们有机会在更广泛的复杂系统辩论中深入了解新出现的属性。少

2018 年 8 月 21 日提交;最初宣布 2018 年 8 月。

评论:24 页, 10 个数字

日记本参考:码 a 503 (2018) 1020-1029

150. **第: 1808. 06359[pdf,其他] cse**

利用需求和源代码之间的历史关联来识别受影响的类

作者:davedde falessi, justinroll, jin guo, jane Cleland-Huang

摘要: 随着软件系统中引入和实现新的要求，开发人员必须确定需要更改的源代码类集。因此，过去的工作重点是预测受需求影响的类集。在本文中，我们引入和评估了一种新的信息类型，基于这样的直觉，即与特定类的历史变化相关的一组需求可能会表现出与影响新需求的语义相似性。类。这种对需求集 (r2rs) 系列度量的新要求捕获了新需求与以前与类关联的现有需求集之间的语义相似性。本文的目的是介绍和评估 r2rs 度量在预测受需求影响的类集合方面的有用性。我们考虑了 18 个不同的 r2rs 指标，结合六**种自然语言处理**技术来测量文本之间的语义相似性 (例如, vsm) 和三个分布分数来计算总体相似性 (例如,相似性分数之间的平均值)。我们评估 r2rs 是否可用于组合预测受影响的类，以及是否与基于更改的时间位置、代码的直接相似性、复杂性度量和代码气味的其他四个度量系列进行预测。我们的评估包括五个分类器和 78 个版本属于四个大型开源项目，这导致超过 700, 000 个候选受影响的类。实验结果表明，利用 r2rs 信息可以在各种分类器和项目中平均提高 60% 以上的预测受影响类的准确性。少

2018 年 8 月 20 日提交;最初宣布 2018 年 8 月。

151. **第 1808. 06305[pdf,其他] Cs。Cl**

基于方差归一化和动态嵌入的词表示后处理

作者:王斌,陈芬晓,王安吉拉,郭建杰

摘要: 尽管嵌入的词的矢量表示在许多**自然语言处理**(nlp) 应用程序中提供了令人印象深刻的性能，但有序输入序列的信息在一定程度上会丢失，即使只是这样培训中使用了基于上下文的示例。为了进一步提高性能，本文提出了两种新**的后处理**技术，即通过方差归一化 (pvn) 进行后**处理**和通过动态嵌入后处理 (pde).pvn 方法归一化字向量主成分的方差，而 pde 方法从有序输入序列中学习正交潜在变量。pvn 和 pde 方法可以集成，以实现更好的性能。我们将这些**后处理**技术应用于两种流行的词嵌入方法(即第 2 vec 和环球词)，以获得它们的**后处理表示.**进行了大量实验，以证明拟议**的后处理**技术的有效性。少

2018 年 9 月 5 日提交;v1 于 2018 年 8 月 20 日提交;最初宣布 2018 年 8 月。

评论:8 页, 2 个数字

152. **第 xiv:1808. 05857[pdf,其他] cse**

elica: 一种动态提取需求相关信息的自动化工具

作者:zahra shakeri hossein abad, vincenzo gervasi, didar zowghi, ken barker

摘要: 需求激发需要对最终系统所在的问题领域有广泛的知识和深刻的理解。然而，在许多软件开发项目中，分析师被要求从一个不熟悉的领域中获取需求，这往往会导致分析师和利益相关者之间的沟通障碍。在本文中，我们提出了一个需求 elic 背诵援助

工具 (elica)，以帮助分析师更好地了解目标应用领域的动态提取和标签的要求相关的知识。为了提取相关术语，我们利用加权有限态传感器 (wfst) 的灵活性和强大功能进行自然 **语言** 处理任务的动态建模。除了通过文本传达的信息外，elica 还捕获和**处理**有关发言者意图的非语言信息，如他们的置信度、分析语气和情绪。提取的信息作为一组带有突出显示的相关术语的标记片段提供给分析师，这些术语也可以作为需求工程 (re)**过程**的工件导出。通过案例研究验证了自译烟子的应用和实用性。这项研究表明，如何捕获有关应用程序域和在启发式会议期间捕获的信息的现有相关信息，如对话和利益相关者的意图，以支持分析师实现其任务。少

2018 年 7 月 20 日提交;最初宣布 2018 年 8 月。

评论:2018 年 ieee 第 26 届国际需求工程研讨会

153. **第 1808.0531[pdf,其他] Cs。Ds**

有效地学习垫模型的混合

作者:alen liu, ankur moitra

摘要: m 允许模型的混合是一种流行的生成模型，用于对来自异构人群的数据进行排名。他们有各种各样的应用，包括社会选择、推荐系统和**自然语言处理**。在这里，我们给出了第一个多项式时间算法，可以证明地学习了具有任何常量分量的 m 允许模型的混合的参数。在我们工作之前，只有两个组成部分的案件得到了解决。我们的分析围绕着 zagier 的确定性同一性，这是在数学物理学的上下文中得到证明的，我们用它来显示多项式可辨识性，并最终构造测试函数来一次剥离一个组件。为了补充我们的上界，我们在样本复杂度上显示了信息理论的下限，并对只进行局部查询的受限算法家族显示了较低的边界。这些结果共同表明，在改善对各组成部分数量的依赖方面存在各种障碍。他们还从超越最坏情况分析的角度推动学习 m 允许模型混合物的研究。在这个方向上，我们表明，当 m 允许模型的缩放参数有分离时，有更快的学习算法。少

2018 年 8 月 16 日提交;最初宣布 2018 年 8 月。

评论:35 页

日记本参考:focs 2018

154. **第 1808.05697[pdf,其他] Cs。Cl**

自然语言处理中的深度贝叶斯主动学习：一个大规模的实证研究结果

作者:aditya sidhant, zachary c. lipton

文摘: 最近的几篇论文探讨了主动学习 (al)，以减轻深层学习对**自然语言处理**的数据依赖。然而，al 对现实世界问题的适用性仍然是一个悬而未决的问题。在监督学习中，从业者可以尝试许多不同的方法，在选择模型之前根据验证集对每个方法进行评估，al 不会提供这样的奢华。在一次 al 运行过程中，代理会对其数据集进行注释，耗尽其标签预算。因此，给定一个新任务，一个活跃的学习者没有机会比较模型和采集功能。本文对深度主动学习进行了大规模的实证研究，解决了多个任务，并针对每个数据集、多个模型和全套采集功能进行了研究。我们发现，在所有设置中，贝叶斯通过不同的方式主动学习，使用 dropout 或 bay-by backprop 提供的不确定性估计显著改善了 i. d. 基线，通常优于经典的不确定性采样。少

2018 年 9 月 24 日提交;v1 于 2018 年 8 月 16 日提交;最初宣布 2018 年 8 月。

评论:将在 2018 年 emnlp 会议上提交

155. **第 1808.05505[pdf,其他] Cs。Cl**

释义思想：句子嵌入模块，模仿人类语言识别

作者:张明君,康必成

文摘 句子嵌入是**自然语言处理**中的一个重要研究课题。为了提高机器翻译等各种自然语言处理任务的性能，必须生成一个充分反映句子语义意义的嵌入向量 和文档分类。到目前为止，已经提出了各种句子嵌入模型，并通过在嵌入后的任务中的良好表现，如情感分析和句子分类，证明了它们的可行性。然而，由于使用简单的句子表示方法可以提高句子分类和情绪分析的性能，因此仅仅声称这些模型充分反映了基于好的句子的含义是不够的此类任务的性能。本文在人类**语言**识别的启发下，提出了语义连贯的以下概念，对于一种好的句子嵌入方法，应满足于这样的概念：在嵌入中，相似的句子应该彼此靠近。空间。然后，提出了路径-思想 (p-思想) 模型，以尽可能多地追求语义连贯。对两个释义识别数据集 (ms coco 和 sts 基准) 的实验结果表明，p-思想模型的性能优于基准句子嵌入方法。少

2018 年 10 月 14 日提交;v1 于 2018 年 8 月 16 日提交;最初宣布 2018 年 8 月。

评论:10 页

156. 第 xiv:1808.05374[pdf, ps,其他] Cs。Cl

使用频谱聚类计算单词类

作者:effi levi, saggy herman, ari rappoport

摘要: 对词汇词汇进行聚类是**自然语言处理**(nlp) 中一个研究得很好的问题。词汇集用于处理统计**语言处理**中的稀疏数据，以及用于解决各种 nlp 任务 (文本分类、问题回答、命名实体识别等) 的功能。频谱聚类技术是图像**处理**和语音识别领域中广泛应用的一种技术。然而，它几乎没有在 nlp 的上下文中探讨;具体而言，在这 (meila 和 shi, 2001) 中使用的方法从来没有被用来聚一个一般的词词典。我们将谱聚类应用于单词的词汇，通过将它们作为解决两个经典 nlp 任务的特征来评估产生的聚类：语义角色标记和依赖关系解析。我们将性能与布朗聚类分析 (一种广泛使用的单词聚类技术) 以及其他聚类方法进行了比较。我们表明，光谱聚类产生类似于布朗聚类的结果，并且优于其他聚类方法。此外，我们量化光谱和棕色集群之间的重叠，显示每个模型捕获一些信息，而这些信息是另一个模型没有捕获的。少

2018 年 8 月 16 日提交;最初宣布 2018 年 8 月。

157. 第 xiv:1808.04865[pdf,其他] Cs。Cl

自上而下的树结构化文本生成

作者:郭启鹏,邱锡鹏,薛向阳,郑章

摘要: 文本生成是**自然语言处理任务中**的一个基本组成部分。现有的序列模型直接对文本序列进行自回归，难以生成复杂结构的长句。本文提倡一种简单的方法，将句子生成视为树生成任务。通过显式建模组成语法树中的语法结构并执行自上而下的宽度-第一树生成，我们的模型适当地修复依赖关系并执行隐式全局规划。这与基于转换的深度第一代**过程**形成鲜明对比，后者在分析时难以处理不完整的文本，也没有将未来的上下文纳入规划。我们在两代任务和一个解析任务上的初步结果表明，这是一个有效的策略。少

2018 年 8 月 14 日提交;最初宣布 2018 年 8 月。

158. 第 xiv:1808.04752[pdf,其他] Cs。Lg

量化神经网络的方法与理论综述

作者:郭云辉

摘要: 深度神经网络是许多现实世界任务的最先进的方法，如计算机视觉、**自然语言处理**和语音识别。深部神经网络尽管很受欢迎，但在训练和推理过程中消耗了大量的内存，消耗了设备的电池寿命，也受到了批评。这使得很难在资源约束严格的移动或嵌入式设备上部署这些模型。量化被认为是满足深度神经网络模型所需的极端内存要求的最有效方法之一。量化表示使用更紧凑的格式 (如整数甚至二进制数字) 存储权重，而不是采用 32 位浮点格式来表示权重。尽管预测性能可能会降低，但量化为大幅降低模型大小和能耗提供了一个潜在的解决方案。在这项调查中，我们对量化神经网络的不同方面进行了深入的回顾。讨论了量化神经网络目前面临的挑战和发展趋势。少

2018 年 8 月 13 日提交;最初宣布 2018 年 8 月。

评论:0.0 版本

159. **第 xiv:1808.07006[pdf,其他] cse**

神经机器翻译超越函数对的二值码相似性比较

作者:赵飞,李小鹏,张志新, 杨志强, 罗兰南,曾强

摘要: 二进制代码分析允许分析二进制代码，而无需访问相应的源代码。二进制文件，在拆卸**后，用汇编语言**表示。这激励我们利用**自然语言处理**(nlp) 中的思想和技术来处理二进制分析，这是一个专注于**处理**各种**自然**语言文本的丰富领域.我们注意到二进制代码分析和 nlp 有很多类似的主题，如语义提取、摘要和分类。这项工作利用这些想法来解决两个重要的代码相似性比较问题。(i) 为不同指令集体系结构 (isa) 提供一对基本块，确定它们的语义是否相似;和 (ii) 给出一段感兴趣的代码，确定它是否包含在不同 isa 的另一段装配代码中。这两个问题的解决方案有许多应用程序，如跨体系结构漏洞发现和代码剽窃检测。我们实现了一个原型系统 innere 业，并进行了全面的评估。我们的方法和问题 i 的现有方法之间的比较表明，我们的系统在准确性、效率和可伸缩性方面的性能优于它们。利用该系统进行的案例研究表明，我们对问题二的解决是有效的。此外，本研究还展示了如何将思想和技术应用于大规模二进制代码分析。少

2018 年 8 月 8 日提交;最初宣布 2018 年 8 月。

160. **第 xiv:1808.04614[pdf,其他] Cs。CI**

向非专家解释 web 表查询

作者:jonathan berant, daniel gerch, amir globerson, tova milo, tomer wolfson

摘要: 为查询表设计可靠的**自然语言**(nl) 接口一直是数据管理和**自然语言处理**(nlp) 社区研究人员的长期目标。这样的接口接收作为输入 nl 问题，将其转换为正式查询，执行查询并返回结果。翻译**过程**中的错误并不少见，用户通常难以理解其查询是否已正确映射。我们通过向非专家用户解释所获得的正式查询来解决此问题。提出了两种查询解释方法: 第一种方法将查询转换为 nl，而第二种方法提供基于查询单元的源的图形表示形式 (在给定表中的执行中)。我们的解决方案通过 web 表增强了最先进的 nl 接口，在培训和部署阶段都增强了该接口。实验，包括在 amazon 机械土耳其人身上进行的用户研究，展示了我们提高 nl 接口正确性和可靠性的解决方案。少

2018 年 8 月 14 日提交;最初宣布 2018 年 8 月。

评论:将出现在 icde 2019 年中的短文版

161. **第 xiv:1808.04459[pdf] Cs。CI**

基于深度学习的自然语言处理在端到端语音翻译中的实现

作者:sarvesh patil

摘要: 深度学习方法使用多个**处理**层来学习数据的层次结构表示。它们已经部署在大量的应用程序中，并产生了最先进的结果。近年来，随着计算机处理能力的**提高**，能够进行高维张量计算,**自然语言处理**(nlp) 的应用在效率和准确性。本文将介绍各种信号**处理技术,**然后应用它们，利用深度递归神经网络生成语音到文本系统。少

2018 年 8 月 9 日提交;最初宣布 2018 年 8 月。

评论:4 页, 6 个数字

162. **第 xiv:1808. 04446[pdf,其他] Cs。简历**

多跳功能调制的可视推理

作者:florian strub, mathieu seurin, ethan perez,纱. de vris, jérémiemary, philippe preux, aaron courville, olivier pietquin

文摘: 最近在计算机视觉和**自然语言处理**方面的突破激发了人们对具有挑战性的多式联运任务的兴趣，如视觉问答和视觉对话。对于此类任务，一种成功的方法是通过特征线性调制 (film) 层 (即每个通道缩放和移位) 对**语言**进行基于图像的卷积网络计算。我们建议以多跳的方式，而不是像以前的工作那样，同时生成卷积网络层次结构上的 film 层的参数。通过在处理**语言输入和**生成 film 层参数之间交替使用，此方法能够更好地扩展到具有较长输入序列 (如对话) 的设置。我们展示了多跳 film 一代实现了最先进的短输入序列任务恐怕--与单跳 film 生成相当---同时也显著优于以前最先进的和单跳 film 生成猜猜什么？!视觉对话任务。少

2018 年 10 月 12 日提交;v1 于 2018 年 8 月 3 日提交;最初宣布 2018 年 8 月。

评论:在 eccv 的 pec 2018

163. **建议: 1808. 04343[pdf,其他] Cs。Cl**

regmapr-文本匹配变得简单

作者:悉达多·婆罗门

摘要: 文本匹配是自然**语言处理**中**的**一个基本问题。使用双向 lstm 进行句子编码和句子间注意机制的神经模型在多个基准数据集上表现得非常好。我们建议 regmap 来说--一个简单而通用的文本匹配体系结构，不使用句子间的注意。从暹罗建筑开始，我们根据两个句子中单词之间的精确和短语匹配，以两个特征来增强单词的嵌入。我们利用三种类型的数据集正则化方法对模型进行了训练，以实现文本包络、转译检测和语义关联。与使用大量手工制作的特征的更复杂的神经模型或模型相比, regmapr 的性能是可比较的或更好的。regmapr 在 sick 数据集上的转述检测和 snli 数据集上不使用句子间注意的模型中的文本包包实现了最先进的结果。少

2018 年 9 月 10 日提交;v1 于 2018 年 8 月 13 日提交;最初宣布 2018 年 8 月。

164. **建议: 1808. 04334[pdf,其他] Cs。Cl**

基于角的词汇元认知学习

作者:james o ' neill, d 努 shka bollegala

文摘: 近年来，在自然语言处理任务中，为**改进**分布式**文字表示**而使用单词嵌入已显示出良好的成功。这些方法要么在一组向量上执行简单的数学运算，要么使用无监督的学习来查找低维表示。本工作比较了针对不同损失训练的元嵌入，即考虑重建嵌入和目标之间角度距离的损失函数，以及那些根据矢量长度考虑归一化距离的函数。我们认为，元嵌入更好地对待在无监督学习中的集合集，因为在元嵌入之前，上游任务的每个嵌

入的相应质量是未知的。我们展示了解释这一点的规范化方法，如余弦和 kl 发散目标，比在标准上训练的元嵌入更有效果我 1 和我 2 在具有相同性和相关性的数据集上丢失，并发现它优于现有的元学习策略。少

2018 年 8 月 13 日提交;最初宣布 2018 年 8 月。

评论:5 页, 2 个数字

165. 第: 1808. 04217[pdf, ps,其他] Cs。CI

使用序列一致性的句子表示的无监督学习

作者:悉达多·婆罗门

摘要: 计算句子的通用分布表示是自然**语言处理**中**的**一项基本任务。我们提出了一种简单但功能强大的无监督方法，通过对令牌序列强制实施一致性约束来学习此类表示。我们考虑两类这样的约束--形成句子的序列和合并时形成句子的两个序列之间的序列。我们通过训练来区分一致和不一致的例子来学习句子编码器。对几个转移学习和语言探测任务的广泛评估表明，在强大的无监督和监督基线的情况下，性能有所提高，在一些情况下大大超过了这些基线。少

2018 年 9 月 29 日提交;v1 于 2018 年 8 月 10 日提交;最初宣布 2018 年 8 月。

评论:提交给国际航天中心 2019 年

166. 第 xiv: 1808. 04124[pdf,其他] Cs。红外

多伊 10.3166/dn.2017.00011

科学语料库中研究地点的识别方法

作者:eric kergosien, marie-noëlle bessagnet, Chaudiron teisseire, joachimschöpfel ,mohammad amin farvardin , stphane chaudiron, bernard jjjemin, annig le parc-lakayrele, mathieu roche, christian sallaberry, jean-菲利普·tonneau, marie-noelle bessagnet, amin farvardin, annig lacayrelle

摘要: terre-istex 项目旨在根据科学语料库中的异质数字内容，确定研究工作关系在研究领域、学科交叉口和具体研究方法方面的演变。该项目分为三个主要行动: (1) 确定已成为实证研究主题的时期和地点，并反映所分析的语料库所产生的出版物; (2) 确定这些作品中涉及的主题;3) 开发基于网络的地理信息检索工具 (gir)。前两个操作涉及**将自然语言处理**模式与文本挖掘方法相结合的方法。通过跨越全球关系引擎中的三个维度 (空间、主题和时间)，将有可能了解对哪些领土和在什么时间进行了哪些研究。在该项目中，实验是在一个异构语料库上进行的，包括电子论文和科学文章从 istex 数字图书馆和 cirad 研究中心。少

2018 年 8 月 13 日提交;最初宣布 2018 年 8 月。

日记本参考:revue des 科学和技术公司 de l ' infors-série 文档 Série, lavoisier, 2017, 20 (2-3), pp.11-30

167. 第 1808. 04000[pdf,其他] Cs。简历

语言引导的时尚形象操纵与功能的转变

作者:mehmet günel, erkut erdem, Erkut erdem

文摘: 开发通过**自然**句子编辑服装图像的技术，并相应地产生新的服装，在艺术、时尚和设计方面有着广阔的应用前景。然而，这被认为是一项肯定具有挑战性的任务，因为图像处理只能在图像的相关部分进行，同时保持其余部分不变。此外，这个操作**过程**应

该产生一个尽可能逼真的图像。在这项工作中，我们提出了 firemedgan，它利用特征的线性调制 (film)，在不使用额外空间信息的情况下，将视觉特征与**自然语言**表示关联并转换。我们的实验表明，这种方法与跳过连接和总变异正则化相结合，产生的结果比基线工作更合理，并且在生成新的装备时具有更好的定位能力。与目标描述。少

2018 年 8 月 12 日提交;最初宣布 2018 年 8 月。

评论:接受 eccv 2018，第一个计算机视觉为时尚，艺术和设计研讨会 (扩展版)

168. **第 xiv:1808.03920[pdf,其他] Cs。Lg**

具有递归多级融合的多模态语言分析

作者:paul pu liang, ziyin liu, amir zadeh, louis-philiu mor 露 z

文摘: 人多模态**语言**的计算建模是**自然语言处理**中一个新兴的研究领域，涉及**语言**、视觉和声学模式。理解多模**态语言**不仅需要建模每个模态中的交互 (模内交互)，更重要的是模式之间的交互 (交叉模态交互)。本文提出了将融合问题分解为多个阶段的递归多级融合网络 (rmfn)，每个阶段都集中在多模态信号的子集上，用于专门的、有效的融合。多态融合方法基于前几个阶段的中间表示，采用这种多阶段融合方法对多模态交互进行建模。通过将我们提出的融合方法与递归神经网络系统相结合，建立了时间和模内相互作用的模型。rmfn 在与多模态情绪分析、情感识别和说话人特征识别相关的三个公共数据集中，在模拟人类多模态**语言**方面表现出最先进的性能。我们提供可视化，以表明融合的每个阶段集中在不同的多模态信号子集，学习越来越有鉴别性的多模态表示。少

2018 年 8 月 12 日提交;最初宣布 2018 年 8 月。

评论:emnlp 2018

169. **第 xiv:1808.03915[pdf,其他] Cs。Cl**

多语言对话的收件人和响应选择

作者:元木佐藤,广木大户,汤田津尾

摘要: 开发能够用多种**语言**进行对话的会话系统是**自然语言处理**的一个有趣的挑战。在本文中，我们介绍了多语言收件人和响应选择。在此任务中，会话系统预测多种**语言**的输入消息的适当收件人和响应。开发这种多语言响应系统的一个关键是如何利用高资源语言数据来补偿低资源的**语言**数据。我们提出了几种会话系统的知识转移方法。为了评估我们的方法，我们创建了一个新的多语言对话数据集。对数据集的实验证明了我们方法的有效性。少

2018 年 8 月 12 日提交;最初宣布 2018 年 8 月。

评论:2018 年

170. **第 1808.083806[pdf] Cs。Cl**

自动预注释在临床笔记数据元素提取中的影响--清洁工具

作者:郭宗婷,胡继娜,金继勋, robert el-kreh, Siddharthsingh, stephanie feudjio feupe, vincent kri, gordon lin, micree e. day,许春南, lucila ohno-machado

摘要: 目的。注释是昂贵的，但对于临床笔记回顾和临床**自然语言 处理**(cnlp) 至关重要。然而，计算机生成的预注释在多大程度上有利于人类的注释仍然是一个悬而未决的问题。我们的研究介绍了 clean (clinci-notreas 和意记)，这是一个基于注释的 cnlp 注释系统，用于改进数据元素的临床注释注释，并将 clean 与广泛使用的注释系统 brat rapid 进行了全面的比较注释工具 (brat)。材料和方法。clean 包括一个集成管道

(clean-ep) 和一个新开发的注释工具 (clean-at)。一位领域专家和一位新手用户 (注释者) 通过在 84 个公示中标记与充血性心力衰竭 (chf) 和川崎病 (kd) 队列有关的 87 个数据元素，参与了比较可用性测试。结果。clean 比 brat (0.82) 获得了较高的音符级 f1-分数 (0.896)，其正确性差异显著 (p 值 < 0.001)，主要相关因素是系统软件 (p 值 < 0.001)。在注释时间上没有明显差异 (p 值 0.188)，克莱安 (7.262 分钟注) 和布拉特 (8.286 分钟注) 之间没有明显差异。差异主要与音符长度 (p 值 < 0.001) 和系统软件 (p 值 0.013) 有关。专家报告说，clean 是有用的-满意，而新手报告略有改善。讨论。clean 提高了注释的正确性，并以相同的效率水平提高了有用的满意度。限制包括注释前正确性率的未经测试的影响、样本小、用户小、金标准得到严格验证等。结论。带有预注释的 clean 对于专家处理涉及众多和不同目标数据元素的复杂注释任务是有益的。少

2018 年 8 月 11 日提交;最初宣布 2018 年 8 月。

171. **第 xiv:1808.03728[pdf,其他] Cs。CI**

分层关注：在各种 nlp 任务中真正重要的是什么

作者:杜泽豪,张志华

摘要: 在句子嵌入、文本生成、机器翻译、句子转换、句子转换、句子 转换、句子转换、句子转换、句子转换、句子转换、句子转换、句子转换、句子转换、句子转换、句子转换、句子转换、句子转换、句子转换、句子转换、句子转换、句子翻译、句子化、句子化、句子化、句子化、句子化、句子化、句子化、句子化、句子化、句子化、句子化、句子化、句子化、句子化、句子化、句子化、句子化、句子化机器阅读理解等。不幸的是，现有的关注机制只学习高级或低级功能。本文认为，缺乏层次机制是提高注意机制性能的瓶颈，提出了一种基于多层次不同层次加权和的分层注意机制 (ham)。注意。与现有的机器阅读理解模型 (如 bidaf 和 match-lstm) 相比，ham 在中国诗歌生成任务上实现了最先进的 0.26 分 BLEU 分数，平均提高了近 6.5%。此外，我们的实验和定理表明，火腿具有更强的泛化和表示能力比现有的关注机制。少

2018 年 8 月 10 日提交;最初宣布 2018 年 8 月。

172. **建议:1808.03507[pdf,其他] Cs。CI**

利用数据到文本技术有效利用医疗数据

作者:steffen pauws, albert gatt, emiel krahmer, ehud reiter

摘要: 保健组织正在不断努力改善健康结果，降低成本，并提高患者的护理体验。数据对于衡量和帮助实现这些保健服务方面的改进至关重要。因此，来自各种临床、财政和业务来源的数据流入现在超过了保健组织及其患者。然而，有效利用这些数据是一项重大挑战。显然，文本是使数据可访问的重要媒介。编制财务报告是为了评估医疗保健组织的一些关键绩效指标，以指导其提供医疗保健服务。同样，在临床层面，关于病人状况的数据通过文本描述进行传递，以促进病人的审查、轮班移交和护理过渡。同样，患者通过短信、报告的形式或医生的电子健康平台了解自己的健康状况和治疗数据。不幸的是，如果是由保健专业人员完成的,这种案文就是劳动密集程度很高的过程的结果。它还容易不完整、主观性强，难以扩展到不同的领域、更广泛的受众和不同的传播目的。数据到文本是人工智能领域最新的一项突破性技术，它从数据中自动以文本或语音的形式生成**自然语言**。本章提供了数据到文本技术的调查，重点介绍了如何在医疗保健环境中部署数据到文本技术。它将 (1) 提供数据到文本方法的最新综合，(2) 提供医疗服

务用例的分类概述; (3) 寻求为在医疗保健环境中评估和实施数据到文本提供有力的理由, (4) 突出最近的研究挑战。少

2018 年 8 月 10 日提交;最初宣布 2018 年 8 月。

评论:27 页, 2 个数字, 书章

173. 第: 1808. 03353[pdf,其他] Cs。Cl

通过信息瓶颈原则实现高效的人形语义表示

作者:noga zaslavsky, charles kemp, terry regier , naftali tihby

摘要: 保持环境的高效语义表示是人类和机器面临的一项重大挑战。虽然人类**语言**是这个问题的有用解决方案, 但目前尚不清楚什么计算原理能在机器中产生类似的解决方案。在这项工作中, 我们提出了对这一未决问题的答案。我们建议**语言**通过优化信息瓶颈 (ib) 在其词典的复杂性和准确性之间进行权衡, 将感知压缩为单词。我们通过探索人类语言如何对颜色进行分类, 提出了这一原则可能会导致人类**样**语义表示的经验证据。我们证明, 在 ib 意义上, 跨**语言**的颜色命名系统几乎是最优的, 而且这些**自然**系统类似于人工 ib 颜色命名系统, 具有控制跨语言的单一权衡参数可变性。此外, ib 系统通过一系列结构相变演变, 展示了一个可能的适应**过程**。因此, 这项工作确定了一个计算原则, 该原则是人类语义系统的特征, 可以为机器中的语义表示提供有益的信息。少

2018 年 8 月 9 日提交;最初宣布 2018 年 8 月。

日记本参考:2017 年 NIPS 的认知智能人工智能研讨会

174. 建议: 1808.03 137[pdf,其他] Cs。Cl

文学研究中的情绪分析研究

作者:evgeny kim, roman klinger

摘要: 情感往往是令人信服的叙述的关键部分: 文学讲述的是有目标、欲望、激情和意图的人。在过去, 古典文学研究通常在解释学的框架内审视文学的情感维度。然而, 随着被称为数字人文 (dh) 的研究领域的出现, 一些关于文学语境中情感的研究也发生了计算上的转变。鉴于□生署仍在成立, 这项研究方向是比较新的。同时, 情绪分析的研究始于近二十年前的计算语言, 是当今一个在主要计算语言学会议上有专门的研讨会和轨道的既定领域。这就引出了一个问题, 即计算语言学中的情感分析研究与数字人文学科之间的共性和差异是什么? 在这项调查中, 我们概述了应用于文学的情绪和情绪分析的现有研究体系。在调查的主要部分之前, 我们简要介绍**了自然语言的处理**和机器学习、情感的心理模型, 并概述了现有的情绪和情绪方法计算语言学中的分析。本次调查中提出的论文要么直接来自 dh, 要么来自计算语言学场所, 仅限于应用于文学文本的情感和情感分析。少

2018 年 8 月 9 日提交;最初宣布 2018 年 8 月。

评论:提交 dhq 审查 (http://www.digitalhumanities.org/dhq/)

175. 第 xiv: 1808. 02961[pdf,其他] Cs。Cl

利用多通道卷积神经网络为汉语情感分析寻找有效的表现

作者:刘鹏飞,张吉,梁永基,赵河,托马斯·I·格里菲斯

摘要: 文本的有效表示对于各种**自然语言处理**任务至关重要。对于汉语情感分析的特殊任务, 重要的是要从单词、文字、拼音等不同形式的汉语表现中理解和选择文本的有效表示。本文提出了一种多通道卷积神经网络 (mcncn), 其中各通道对应表示的多通道

卷积神经网络 (mcncn)，系统地研究了这些表示在汉语情绪分析中的作用。实验结果表明: (1) word 在低 oov 速率的数据集上获胜，而字符则以其他方式获胜;(2) 将这些表示组合使用通常会提高性能;(3) 基于 mcncn 的表现优于传统的支持向量机特征;(4) 提出的 mcnnn 模型实现了与最先进的中国情绪分析机型快速文本相比的竞争性能。少

2018 年 8 月 8 日提交;最初宣布 2018 年 8 月。

176. 第 xiv:1808. 02861[pdf,其他] Cs。简历

选择您的神经元: 通过神经元重要性整合领域知识

作者:ramprasath r. selvaraju, prithvijit chattopadhyay, mohamed elhoseiny, tilak sharma, dhruv batra, devi parkh , stefan lee

摘要: 卷积神经网络中的单个神经元被证明是由图像级分类任务监督的，它已经被证明隐式地学习了从简单的纹理和形状到整体或部分物体的语义上有意义的概念--形成一个 "通过学习**过程**获得的概念。在本工作中，我们介绍了一种基于此观察的简单、高效的零镜头学习方法。我们的方法，我们称之为神经元重要性-aware 方差转移 (niwt)，学习将有关新的 "看不见的" 类的领域知识映射到这个学习的概念字典，然后优化网络参数，可以有效地结合这些概念--本质上是通过发现和构成深层网络中的学习语义概念来学习分类器。与以往的 cu 鸟类和 awa2 通用零镜头学习基准相比，我们的方法有了改进。我们在一组不同的语义输入上展示了我们的方法，作为外部领域的知识，包括属性和**自然语言**说明。此外，通过学习逆映射, niwt 可以为新学习的分类器的预测提供视觉和文本解释，并提供神经元名称。我们的代码可在 https://github.com/ramprs/neuron-importance-zsl。少

2018 年 8 月 8 日提交;最初宣布 2018 年 8 月。

评论:在 2018 年 eccv 论文集中

177. 第 xiv:1808. 02374[pdf,其他] Cs。Cl

神经时间关系分类中的字级损失扩展

作者:artuur leeuwenberg, marie-francine moens

摘要: 在**自然语言处理**中的许多任务中，无监督的预先训练的单词嵌入被有效地用于利用未标记的文本数据。通常，这些嵌入可以用作初始化，也可以用作特定于任务的分类模型的固定单词表示形式。在本文中，我们将分类模型的任务丢失扩展到模型的文字嵌入级别上，并给出了一个无监督的辅助损失。这是为了确保学习的单词表示既包含特定于任务的功能，从监督丢失组件中学习，也包含从无监督的损失组件中学习的更一般的功能。我们评估了我们在时间关系提取任务上的方法，特别是从临床记录中提取叙事遏制关系，并表明在非监督目标上继续训练嵌入与任务目标提供了更好的特定于任务的嵌入，并改进了 thyme 数据集上的最新状态，只使用通用域词性标记作为语言资源。少

2018 年 8 月 7 日提交;最初宣布 2018 年 8 月。

评论:参加第 27 届计算语言学国际会议 (coling 2018)

178. 建议: 1808. 02291[pdf, ps,其他] Cs。艾

基于规则的流推理中的窗口有效性问题

作者:alessandro ronca, mark kaminski, bernardo cuenca gru, ian horrocs

摘要: 基于规则的时间查询**语言**提供了所需的表达能力和灵活性，可以**自然**地捕获流数据上复杂的分析任务。但是，流**处理**应用程序通常需要使用有限的资源进行近乎实时的响应。特别是，基本查询**语言**必须具有有利的计算属性，并且流**处理**算法只能在任何内存中保留少量以前接收到的事实。在不牺牲正确性的情况下。本文提出了一个具有可跟踪数据复杂度的时间数据数据采集器递归片段，并研究了该片段的通用流推理算法的性质。我们关注窗口有效性问题，以最大限度地减少流推理算法需要在任何时间点将数据保留在内存中的时间点的数量。少

2018 年 8 月 7 日提交;最初宣布 2018 年 8 月。

179. 第 xiv:1808.02229[pdf,其他] Cs。Lg

草学：浅层和深层学习中的几何意识

作者:张嘉耀,朱光旭,小罗伯特·希思,黄开斌

文摘: 现代机器学习算法已被广泛应用于计算机视觉、**自然语言处理**和人工智能等一系列信号**处理**应用中。许多相关的问题涉及子空间结构特征、正交约束或低阶约束目标函数或子空间距离。这些数学特征是用格拉斯曼流形**自然**表达的。不幸的是，这一事实尚未在许多传统的学习算法中探讨。在过去的几年里，人们对研究格拉斯曼多方面来解决新的学习问题越来越感兴趣。使用深度神经网络在经典学习和学习方面的显著改进使这种尝试得到了保证。我们把前者称为浅薄，后者称为深度格拉斯曼学习。本文的目的是通过对常见数学问题和初步解决方法的调查，并对各种应用进行综述，介绍格拉斯曼年学习的新兴领域。我们希望在研究中启发不同领域的从业者采用格拉斯曼学习的有力工具。少

2018 年 8 月 12 日提交;v1 于 2018 年 8 月 7 日提交;最初宣布 2018 年 8 月。

评论:提交给 ieee 信号处理杂志

180. 第 xiv:1808.02171[pdf,其他] Cs。Cl

对话-上下文感知端到端语音识别

作者:kim suyoun , florian metze

摘要: 现有的语音识别系统通常是在句子级别构建的，尽管众所周知，对话上下文 (例如跨越句子或扬声器的更高级别的知识) 可以帮助**处理**长对话。最近端到端语音识别系统的进展有望将所有可用信息 (如声学、**语言**资源) 整合到一个模型中，然后对其进行联合优化。因此，这种对话上下文信息也应该集成到端到端模型中，以进一步提高识别精度，这似乎是很**自然**的。在这项工作中，我们提出了一个对话-上下文感知语音识别模型，它以端到端的方式明确使用句子级信息以外的上下文信息。我们的对话上下文模型捕获了句子级上下文的历史记录，以便整个系统可以以端到端的方式使用对话上下文信息进行培训。我们评估了我们在交换机会话语音语料库上提出的方法，并表明我们的系统优于类似的句子级端到端语音识别系统。少

2018 年 8 月 6 日提交;最初宣布 2018 年 8 月。

评论:提交给 slt

181. 第 xiv:1808.01729[pdf,其他] cse

可执行触发器操作注释

作者:penyu ie, rai rai, junyi jsy li, sarfraz khurshid, raymond j . mooney, milosgligoric

摘要: **天然**的**语言**元素 (例如, todo 注释) 经常用于在开发人员之间进行通信，并在代码存储库 (触发器) 中包含特定条件时描述需要执行的任务 (操作)。随着项目的发展、

开发过程的变化和开发团队的重组，这些评论由于其非正式性质，往往变得无关紧要或被遗忘。我们提出了第一种称为 trigit 的技术，将触发器操作到执行注释指定为可执行语句。因此，当触发器计算为 true 时，将自动执行操作。规范是用宿主语言(如 java)编写的，并作为生成过程的一部分进行评估。这些触发器被指定为抽象语法树上的查询语句和生成配置脚本的抽象表示形式，并且这些操作被指定为代码转换步骤。我们为 java 编程语言实现了 trigit，并从 8 个流行的开源项目中迁移了 20 个现有的触发器操作注释。我们根据可执行注释中的令牌数量和生成过程中引入的时间开销来评估使用 triit 的成本. 少

2018 年 8 月 6 日提交;最初宣布 2018 年 8 月。

182. 第 xiv:1808.01591[pdf,其他] Cs。CI

lisa: 通过分层语义积累解释递归神经网络的判断和模式转换的实例

作者:pankaj gupta, hinrich schütze

摘要: 递归神经网络 (rnn) 是一种时间网络和累积性网络，在各种自然语言处理任务中显示出了良好的效果。尽管他们取得了成功，但了解他们隐藏的行为仍然是一个挑战。在这项工作中，我们通过一种名为 layer-wise-semic-cat-cat-cat-cat-cat- 累计 (lisa)的拟议技术来分析和解释 rnn 的累积性质，该技术用于解释决策和检测网络所依赖的最可能的 (即显著性) 模式同时做决定。我们演示 (1) lisa: "rnn 如何在给定文本示例和预期响应的顺序处理过程中积累或构建语义" (2) 示例模式: "数据中每个类别的显著性模式是如何根据网络在决策中 "。我们分析了 rnn 对不同输入的敏感性，以检查预测分数的增加或减少，并进一步提取网络学到的显著性模式。我们采用了两个关系分类数据集: 第 10 学期任务 8 和 tac kbp 插槽填充，通过 lisa 和示例模式来解释 rnn 预测。少

2018 年 8 月 5 日提交;最初宣布 2018 年 8 月。

评论:2018 年自然语言处理经验方法会议 (emnlp2018) nlp 神经网络分析与解释研讨会 (blackboxnlp)

183. 第 xiv:1808.01175[pdf,其他] Cs。CI

通过多尺度图形分区对新闻文章进行内容驱动、无监督的聚类分析

作者:m. tarik altuncu, sophia n.yaliraki , mauricio barahona

摘要: 全球新闻和新闻内容数量激增，加上通过网络媒体广泛和即时地获取信息，使监测新闻发展和意见变得困难和耗时实时形成。越来越需要能够对原始文本进行预处理、分析和分类的工具，以提取可解释的内容;具体而言，确定主题和内容驱动的文章分组。我们在这里介绍了这样一种方法，它将自然语言处理中强大的矢量嵌入与图论中的工具结合在一起，利用图形上的扩散动力学来揭示自然跨比例分区。我们的框架使用最近的深度神经网络文本分析方法 (doc2vec) 以矢量形式表示文本，然后应用多尺度社区检测方法 (马尔可夫稳定性) 对文档向量的相似图进行划分。该方法允许我们以不受监督的方式以不同的分辨率获取内容相似的文档集群。我们通过对 vox media 在一年内发表的 9, 000 篇新闻文章的分析来展示我们的方法。我们的结果显示了根据内容对文档进行一致分组，而没有对要找到的群集的数量或类型进行先验的假设。与外部分类服务和标准主题检测方法相比，多级聚类显示了主题和子主题的准层次结构，具有更高的清晰度和更好的主题一致性。少

2018 年 8 月 3 日提交;最初宣布 2018 年 8 月。

评论:8 页; 5 个数字;参加 kdd 2018: 数据科学、新闻和媒体研讨会

184. 第 1808. 0114[pdf,其他] Cs。Lg

深度学习中的泛化错误

作者:daniel jakubovitz, raja giryes, miguel r. d. rodrigues

摘要: 深度学习模型最近在计算机视觉、语音识别、语音翻译、**自然语言处理**等各个领域都表现出了出色的表现。然而，除了他们最先进的性能外，一般还不清楚他们泛化能力的来源是什么。因此，一个重要的问题是，是什么使深度神经网络能够很好地从训练集推广到新的数据。本文结合经典和最新的理论和实证结果，对深部神经网络泛化误差的表征提供了现有理论和边界的概述。少

2018 年 8 月 3 日提交;最初宣布 2018 年 8 月。

185. 第 1807. 1181838[pdf] 反渗透委员会

机器人指令语音的可扩展接地

作者:乔纳森·康奈尔

摘要: 口语**是**指挥移动机器人的方便界面。然而，要做到这一点，一些基本术语必须以感知和运动技能为基础。我们详细介绍了机器人 eli 上使用的**语言处理**，并解释了这种接地是如何执行的，它如何与用户手势交互，以及它如何处理诸如回指等现象。然而，更重要的是，有些概念是机器人无法预先编程的，比如家庭中各种物体的名称或可能要求它执行的具体任务的**性质**。在这些情况下，至关重要的是要有一种扩大接地的方法，本质上是 "通过被告知来学习"。我们描述了如何成功地实现这一方法，以便在桌面设置中学习新的名词和动词。创建这种**语言**学习内核可能是机器人所需要的最后一个显式编程--核心机制最终可以用来传授大量的知识，就像孩子从父母和老师那里学习一样。少

2018 年 7 月 31 日提交;最初宣布 2018 年 7 月。

报告编号: "说话和倾听的机器人" 章节草稿, j. markowitz (编), de grutyer 2014

186. 第 1807. 1171714[pdf,其他] Cs。Cl

神经自然语言处理中的性别偏差

作者:k 子杰, piotr mardziel, f 预科 jing wu, preetam amamcharla, anupam datta

摘要: 我们研究**神经自然语言处理**(nlp) 系统是否反映了训练数据中的历史偏差。我们定义了一个通用基准来量化各种神经 nlp 任务中的性别偏差。我们的经验评估采用了最先进的神经相关分辨率和基于 rnn 的**教科书**语言模型，在基准数据集上进行了训练，发现模型在如何看待职业方面存在显著的性别偏见。然后，我们减轻了对 cda 的偏见:一种通过因果干预来强化语料库的通用方法，以打破性别和性别中立词之间的联系。我们的经验表明, cda 有效地减少性别偏见，同时保持准确性。我们还探索了缓解策略的空间与 cda，一个预先的方法，嵌入词去偏置 (wed)，和他们的组成。我们表明, cda 的表现优于 wed，当单词嵌入训练时，效果非常好。对于预先训练的嵌入，可以有效地组合这两种方法。我们还发现，随着对原始数据集的训练进行梯度下降，性别偏见随着损失的减少而增加，这表明优化鼓励偏差;cda 缓解了此行为。少

2018 年 7 月 31 日提交;最初宣布 2018 年 7 月。

187. 第 1807. 11057[pdf,其他] Cs。Cl

基于远程约束训练的跨语言文档向量神经机器翻译框架

作者:李伟,麦先生

摘要: 文档的通用跨语言表示对于许多**自然语言处理**任务非常重要。本文提出了一种利用神经机器翻译 (nmt) 框架，通过自注意机制有效地创建文档向量的文档矢量化方法。我们的方法所使用的模型可以用与手头任务无关的并行语料库进行训练。在测试过程中，我们的方法将采用单语文档，并将其转换为 "基于神经机翻译框架并进行距离约束训练的跨语言文档向量" (cntdv)。cntdv 是我们以前对基于神经机器翻译框架的文档向量的研究的后续研究。cntdv 可以以极快的速度从编码器的正向传递生成文档矢量。此外，它还具有距离约束，因此从不同**语言**对获得的文档向量始终是一致的。在跨语言文档分类任务中，我们的 cntdv 嵌入超过了在英语到德语分类测试中发布的最先进的性能，据我们所知，它还在德语对英语的分类测试。与以往的研究相比，在测试**过程**中不需要翻译，这使得模型更快、更方便。少

2018 年 9 月 4 日提交;v1 于 2018 年 7 月 29 日提交;最初宣布 2018 年 7 月。

188. 第 xiv:1807.10854[pdf,其他] Cs。Cl

深层学习在自然语言处理中的应用综述

作者:daniel w. otter , julian r. medina , jugal k. kalita

摘要: 在过去几年里，由于深度学习模型的使用激增,**自然语言处理**领域向前推进。本调查简要介绍了该领域，并简要概述了深度学习架构和方法。然后，它筛选了大量的最近的研究，并总结了大量的相关贡献。分析的研究领域包括几个核心语言**处理**问题，以及计算语言学的一些应用。然后讨论了目前的最新情况，并为今后在这一领域的研究提出了建议。少

2018 年 7 月 27 日提交;最初宣布 2018 年 7 月。

189. 特别报告:1807.1080:05[pdf,其他] Cs。Cl

利用附加语言信息改进神经序列标签

作者:mahtab ahmed, muhammad rifayat samee, robert e. mercer

摘要: 序列标记是为数据序列分配分类标签的任务。在**自然语言处理**中，序列标记可应用于各种基本问题，如 "词性部分 (pos)" 标记、命名实体识别 (ner) 和 "块"。在本研究中，我们提出了一种在神经序列框架中添加各种语言特征的方法，以改进序列标记。除了词级知识外，还添加了感觉嵌入来提供语义信息。此外，还添加了字符嵌入的选择性读数，以捕获句子中每个单词的上下文和形态特征。与以前的方法相比，这些增加的语言特征使我们能够设计一个更简洁的模型，并执行更有效的培训。我们提出的架构在 pos、ner 和分块的基准数据集上实现了最先进的结果。此外，我们模型的收敛速度明显优于以前的最先进的模型。少

2018 年 7 月 27 日提交;最初宣布 2018 年 7 月。

评论:9 页, 1 个图，正在审查中

190. 第 1807.10215[pdf,其他] Cs。简历

深 spine: 自动腰椎分割，点状指定，脊柱狭窄分级使用深度学习

作者:jen-tang lu, stedfano pedemonte, bernardo bizzo, sean doyle, katherine p.andriole, mark h.michalski, r. gilberto gonzalez, stuart r . pomerantz

摘要: 脊柱狭窄的高流行率导致大量的 mri 成像，但解释可能是耗时的，即使在最专业的放射科医生中也具有较高的读取器间变异性。在本文中，我们开发了一种有效的方法，利用存储在大规模档案报告和图像数据中的主题专家专业知识，为实现全自动腰椎管狭窄分级的深层学习方法。具体而言，我们引入了三个主要贡献: (1) 一种**自然**语

言处理方案，从不同类型和等级脊柱的自由文本放射学报告中逐级提取地面真相标签狭窄 (2) 精确的椎体分割和圆盘级定位使用 u-net 架构结合脊柱曲线拟合方法，(3) 多输入、多任务、多类卷积神经网络来执行中央管和在轴向和矢状成像系列输入上进行椎管狭窄分级，提取的报告派生标签应用于相应的成像电平段。这项研究使用了从 4075 名患者身上提取的 22796 椎间盘水平的大数据集。我们在腰椎管狭窄分类方面达到了最先进的性能，并期望该技术将提高放射学工作流程的效率，并提高放射学报告对转诊临床医生和患者的感知价值。少

2018 年 7 月 26 日提交;最初宣布 2018 年 7 月。

评论:2018 年被评为医疗保健机器学习 (mlhc) 的焦点演讲。补充视频: https://bit.ly/DeepSPINE

191. 第: 1807. 09986[pdf,其他] Cs。简历

用于图像字幕的递归融合网络

作者:姜文浩,马琳,姜玉刚, 刘伟,张通

摘要: 近年来，图像字幕技术取得了很大进展，所有最先进的型号都采用了编码器解码器框架。在此框架下，输入图像由卷积神经网络 (cnn) 编码，然后用递归神经网络 (rnn) 翻译成**自然语言**。基于此框架的现有模型仅采用一种 cnn，例如 resnet 或感知 x，它仅从一个特定的角度描述图像内容。因此，不能全面理解输入图像的语义含义，这限制了字幕的表现。为了利用多个编码器的互补信息，提出了一种新的递归融合网络 (rfnet) 来处理图像字幕。我们模型中的融合**过程**可以利用图像编码器输出之间的交互，然后为解码器生成新的紧凑而信息化的表示形式。mscoco 数据集上的实验证明了我们提出的 rfnet 的有效性，它为图像字幕提供了新的最新技术。少

2018 年 7 月 30 日提交;v1 于 2018 年 7 月 26 日提交;最初宣布 2018 年 7 月。

评论:eccv-18

192. 第 1807. 09844[pdf,其他] Cs。Cl

模块化力学网络: 自然语言处理中用深神经网络桥接力学和现象学模型

作者:simon dbnik, john d. kelleher

摘要: **天然的语言加工过程**(nlp) 可以使用自上而下 (理论驱动) 和自下而上 (数据驱动) 的方法来完成，我们分别称之为机械方法和现象学方法。这些办法经常被认为是相互对立的。通过研究深度学习中的一些最新方法，我们认为深度神经网络包含了这两个观点，此外，利用这一方面的深度学习可能有助于解决**语言**技术中的复杂问题，例如作为空间认知领域的建模**语言**和感知。少

2018 年 7 月 21 日提交;最初宣布 2018 年 7 月。

评论:11 页, 1 个图，出现在 clasp 论文中的计算语言学第 1 卷: 自然语言中的逻辑和机器学习会议记录 (laml 2017)，第 1-11 页

日记本参考:casp 论文在计算语言学第 1 卷: 自然语言中的逻辑和机器学习会议论文集 (laml 2017)。issn: 2002-9764。uri: http://hdl.handle.net/2077/54911

193. 第 1807. 08792[pdf,其他] cs et

光子卷积神经网络加速器

作者:armin Mehrabian, yousra al-kabani , volker j sorger, tarek el-ghazawi

摘要: 卷积神经网络 (cnn) 一直是许多应用的核心，包括但不限于计算机视觉、语音**处理**和**自然语言处理**(nlp)。然而，计算成本高昂的卷积操作对 cnn 的性能和可伸缩性提

出了许多挑战。同时，传统上用于数据通信的光子系统由于其高带宽、低功耗和可重构性，在**数据处理**方面受到了新的欢迎。在这里，我们提出了一个光子卷积神经网络加速器 (pcnna) 作为概念设计的证明，以加快 cnn 的卷积操作。我们的设计基于最近推出的硅光子微称量库，该协议使用广播和重量协议执行乘法和累积 (mac) 操作，并通过神经网络的各层移动数据。在这里，我们的目标是利用光子学固有的平行性 (波分复用 (wdm)) 与 kernels 中输入要素映射和内核之间的连接稀疏性之间的协同作用。虽然我们的完整系统设计在执行时间内可提供多达 3 个数量级的加速，但与最先进的电子核心相比，其光学核心可能会提供 5 个数量级以上的加速。少

2018 年 7 月 23 日提交;最初宣布 2018 年 7 月。

评论:5 页, 6 位, ieee socc 2018

194. **第 xiv:1807.08077[pdf,其他] Cs。CI**

一种用于创意视觉故事讲述的管道

作者:stephanie m. lukin, reginald hobbs, clare r. voss

摘要: 计算视觉讲故事产生了一个文本描述的事件和解释描绘在一系列的图像。这些文本是通过**自然语言处理**、生成和计算机视觉方面的进步和跨学科方法而成为可能的。我们将计算创意的视觉故事定义为能够从三个方面改变故事讲述的能力：谈论不同的环境，根据叙述目标产生变体，以及使叙述适应观众。创意讲故事的这些方面及其对叙事的影响还有待在视觉故事中探索。本文介绍了一个任务模块、对象识别、单图像推断和多图像叙事的管道，为构建一个富有创意的视觉讲故事者提供了初步的设计。我们在注释任务中对一系列图像进行了这种设计。我们提出和分析收集到的语料库，并描述实现自动化的计划。少

2018 年 7 月 20 日提交;最初宣布 2018 年 7 月。

评论:最初发表于 2018 年北美计算语言学协会 (naacl) 第一次讲故事研讨会论文集 (讲故事)

195. **第 xiv:180.7.07752[pdf] Cs。CI**

多伊 10.5220/jca20188917319

推特情绪分析系统

作者:shaunak joshi, deepali deshpande

摘要: 社交媒体越来越多地被人类用来以短信的形式表达自己的感受和意见。在文本中检测情绪有广泛的应用，包括识别个人的焦虑或抑郁，以及衡量社区的幸福感或情绪。情感可以用许多可以看到的方式来表达，比如面部表情和手势、言语和书面文本。文本文档中的情感分析本质上是一个基于内容的分类问题，涉及**自然语言处理**和机器学习领域的概念。本文讨论了基于文本数据的情感识别和情绪分析中使用的技术。少

2018 年 7 月 20 日提交;最初宣布 2018 年 7 月。

评论:5 页

日记本参考:国际计算机应用杂志 (2018)

196. **第:1807.07425[pdf,其他] Cs。CI**

基于规则的特征和知识引导的卷积神经网络的临床文本分类

作者:梁耀,毛成生,袁罗

文摘: 临床文本分类是医学**自然语言处理**中的一个重要问题。现有的研究传统上侧重于基于规则或知识源的特征工程，但只有少数研究利用了深度学习方法的有效特征学习

能力。在本研究中，我们提出了一种新的方法，结合基于规则的特点和知识引导的深度学习技术，有效的疾病分类。我们的方法的关键步骤包括识别触发短语，用很少的例子预测类使用触发短语，并训练一个卷积神经网络与单词嵌入和统一医疗**语言**系统(umls) 实体嵌入。我们评估了我们的方法在 2008 年集成信息学与生物学和床头 (i2b2) 肥胖挑战。结果表明，我们的方法优于最先进的方法。少

2018 年 7 月 20 日提交;v1 于 2018 年 7 月 17 日提交;最初宣布 2018 年 7 月。

评论:arxiv 管理说明: 文本重叠与 arxiv:1806. 04820 由其他作者

197. 第: 1807: 279[pdf,其他] Cs。CI

将解释性传授给单词嵌入

作者:aykut koic, uhsan utlu, lutfi kerem senel , haldun m. ozaktas

摘要: 词嵌入作为**自然语言处理**中普遍存在的一种方法，被广泛地用来将词的语义属性映射到密集的矢量表示中。它们捕获单词之间的语义和句法关系，但与单词相对应的向量相对于彼此来说只是有意义的。矢量和它的维度都没有任何绝对的、可解释的含义。我们对嵌入学习算法的目标函数进行了加性修改，鼓励语义上与预定义概念相关的单词的嵌入向量沿指定维度获取较大的值，同时离开原来的语义学习机制大多不受影响。换句话说，我们将已经确定为相关的单词与预定义的概念对齐。因此，我们通过将意义赋予嵌入的向量维度，将其传递给嵌入一词的可解释性。预定义的概念来自于外部词汇资源，在本文中选择了一个外部词汇资源作为 roget 的术语词库。我们观察到，沿所选概念的对齐并不限于术语词库中的单词，还扩展到其他相关单词。我们从实验结果中量化了可解释性和意义分配的程度。我们还通过词类类比和相似度测试来证明所产生的向量空间的语义一致性的保持。这些测试表明，所提出的框架所获得的解释性的单词嵌入不会牺牲共同基准测试中的性能。少

2018 年 7 月 19 日提交;最初宣布 2018 年 7 月。

评论:8 页, 2 个数字

198. 建议: 1807. 07108[pdf,其他] Cs。CI

语义分析: 利用 lstm 编码器 cfg-decoder 对目标句子进行句法保证

作者:fabiano ferreira luz, marcelo finger

摘要: 语义解析可以定义**为将自然语言**句子映射到机器可解释的**过程**，它的意义的正式表示。利用 lstm 编码器解码器神经网络进行语义解析已成为一种很有前途的方法。然而,**自然语言**的人工自动翻译并不能为生成这种保证的句子提供语法保证，对于数据库查询可能导致关键的实际情况尤其重要错误，如果句子是不语法的。在这项工作中，我们提出了一个神经架构称为编码器 cfg-decoder，其输出符合给定的无上下文语法。结果显示了此类体系结构的任何实现，显示了其正确性，并提供了比文献更好的基准精度级别。少

2018 年 7 月 18 日提交;最初宣布 2018 年 7 月。

199. 第: 1807. 06978[pdf,其他] Cs。红外

通过综合审查改进可解释的建议

作者:sixun ouyang, aonghus lawlor, felipe costa, peter dolog

摘要: 推荐系统的一项重要任务是为用户提供可解释的解释。这对该系统的信誉很重要。目前可解释的推荐系统倾向于关注已知对用户很重要的某些功能，并以结构化的形式提供他们的解释。众所周知，用户生成的审阅和审阅者的反馈对用户的决策具有很强的

影响力。另一方面，最近的文本生成作品已被证明生成的文本质量与人类的书面文本相似，我们的目的是表明生成的文本可以成功地用于解释建议。在本文中，我们提出了一个由流行的评论导向生成模型组成的框架，旨在为建议创建个性化的解释。解释是在字符和单词级别上生成的。我们构建一个数据集，其中包含来自亚马逊图书评审数据集的审阅者的反馈。我们的跨域实验旨在将自然**语言处理**与推荐系统领域之间的联系起来。除了**语言**模型评价方法外，我们还采用了一种新型的基于深度神经网络的面向评审的推荐系统 deepcon，通过根均方误差 (rmse) 来评价生成的评审的推荐性能。我们证明，综合个性化的评论比人类的书面评论有更好的推荐性能。据我们所知，这提供了第一个机器生成的**自然选择解释** 评级预测。少

2018 年 7 月 18 日提交;最初宣布 2018 年 7 月。

评论:recsys drrs 2018

200. **xiv:1807. 06414[pdf,其他] Cs。红外**

将上下文感知神经网络与去噪自动编码器相结合测量字符串相似性

作者:mehdi ben lazreg, morten goodwin

摘要: 测量字符串之间的相似性是许多成熟和快速增长的研究领域的核心，包括信息检索、生物学和**自然语言处理**。字符串相似度测量的传统方法是在单词空间上定义一个度量，用于量化和汇总两个字符串中字符之间的差异。令人惊讶的是，在过去几十年中，该地区的最先进技术变化不大。大多数指标都是基于字符和字符分布之间的简单比较，而不考虑单词的上下文。本文提出了一个字符串度量，它包含了字符串之间的相似性，基于 (1) 包括单词之间的字符相似性。相同词的非标准拼写和 (2) 词的上下文。我们的建议是一个神经网络，由去噪自动编码器和我们所说的上下文编码器专门设计，以找到基于上下文的单词之间的相似性。实验结果表明，在超过 854% 的情况下，所得到的指标成功地找到了最接近的单词之间非标准拼写的正确版本，而在已建立的 normalised-levenshtein 距离中，这一比例为 63.2%--。此外，我们还表明，在类似上下文中使用的单词与计算为与不同上下文的单词相似的方法，这是在已建立的字符串度量中缺少的理想属性。少

2018 年 7 月 16 日提交;最初宣布 2018 年 7 月。

201. **第: 1807. 06151[pdf,其他] Cs。CI**

注意侵略性检测的 lstm

作者:n 克里斯·尼希尔, ramit Nikhil, Nikhil kumar nirala, rohan kirnani

摘要: 在本文中，我们描述了在脸谱网帖中提交的侵略识别共享任务的系统和 nusnik 团队的评论。以前的工作表明, lstm 在**自然语言处理任务中**取得了显著的性能。我们部署了一个 lstm 模型，上面有一个关注单元。我们的系统在印地语子任务中分别排名第六和第 4 位，用于 facebook 评论和通用社交媒体数据的子任务。在相应的英语子任务中排名第 17 位和第十位。少

2018 年 7 月 16 日提交;最初宣布 2018 年 7 月。

评论:在第 27 届国际计算语言学会议 (coling 2018) 上参加第一次滚动、侵略和网络欺凌研讨会

202. **第 xiv:180. 7.05962[pdf,其他] Cs。CI**

使用短语嵌入的文本文档中关键字的主题加权排名

作者:debanjan mahata, john kuriakose, rajiv rahnshah, roger zimmermann, john r. talburt

摘要: 关键字提取是**自然语言处理**中的一项基本任务，有助于将文档映射到一组具有代表性的单词和多词短语。文本文档中的关键字主要使用监督和非监督方法提取。本文提出了一种非监督技术，该技术采用主题加权个性化的 paragerank 算法和神经短语嵌入相结合的方法提取和排序关键字。我们还介绍了一种使用现有技术处理文本文档和训练短语嵌入的有效方法。我们共享从现有数据集中派生的评估数据集，该数据集用于选择基础嵌入模型。对排名关键字提取的评估是在两个基准数据集上进行的，其中包括短摘要 (成套) 和长科学论文 (semeval 2010)，并显示比最先进的系统产生更好的结果。少

2018 年 7 月 16 日提交;最初宣布 2018 年 7 月。

评论:第一届 ieee 多媒体信息处理与检索国际会议论文集中接受的论文预印

203. 第 1807. 05642[pdf,其他] Cs。Cl

晚不是厄利: 一个更快的平行厄利解析器

作者:peter ahrens, john feser, robin hui

摘要: 我们提出了 ate 算法，这是 earley 算法的一个异步变体，用于分析上下文无关语法。earley 算法**是自然**基于任务的，但由于任务之间的依赖关系，很难并行化。我们提出 ate 算法，它使用额外的数据结构来维护有关分析状态的信息，以便可以按任意顺序**处理**工作项。此属性允许使用任务并行性加快 ate 算法的工作。我们证明了 ate 算法在**自然语言**任务上可以实现比 earley 算法的 120x 加速。少

2018 年 7 月 15 日提交;最初宣布 2018 年 7 月。

204. 第 1807. 05195[pdf,其他] Cs。Cl

使用领域对抗学习的低资源文本分类

作者:daniel Grießhaber, ngoc thang vu, jones maucher

摘要: 深度学习技术最近已证明在许多**自然语言处理**任务中是成功的，这些任务构成了最先进的系统。然而，它们需要大量经常缺失的附加注释的数据。本文探讨了在新的目标域或**语言**的低资源和零资源设置中训练深度、复杂神经网络的域不变特征时，利用域对抗学习作为调节器，避免过度拟合。在新**语言**的情况下，我们表明，单语词向量可以直接用于训练，而无需预对齐。它们投射到一个公共空间可以在训练时临时学习，达到预先训练的多语言文字向量的最终性能。少

2018 年 7 月 13 日提交;最初宣布 2018 年 7 月。

评论:将在 2018 年第六届统计语言和语音处理国际会议上发表

205. 第 xiv: 180 7.04723[pdf, ps,其他] Cs。Lg

瓶颈模拟器: 一种基于模型的深层强化学习方法

作者:iulian vlad serban, chinnadhuraisankar, michael pi 片断 per, joelle pineau, yyokhua bengio

摘要: 深度强化学习最近取得了许多令人印象深刻的成功。然而，将这些方法应用于现实问题的一个主要障碍是它们缺乏数据效率。为此，我们提出了瓶颈模拟器: 一种基于模型的增强学习方法，该方法将学习的环境过渡模型与推出模拟相结合，从几个示例中学习有效的策略。学习的过渡模型采用抽象的、离散的 (瓶颈) 状态，通过减少模型参数的数量和利用环境的结构特性来提高样本效率。我们根据学习策略的固定点对瓶

颈模拟器进行了数学分析，揭示了性能是如何受到四个不同的错误来源的影响的：与抽象空间结构相关的错误、与过渡模型估计方差，一个与过渡模型估计偏差有关的误差，一个与过渡模型类偏差有关的误差。最后，我们评估了两种**自然语言处理**任务的瓶颈模拟器：文本冒险游戏和现实世界中复杂的对话响应选择任务。在这两项任务中，瓶颈模拟器都能产生出色的性能，击败竞争的方法。少

2018 年 7 月 12 日提交;最初宣布 2018 年 7 月。

评论:26 页, 2 个数字, 4 个表

类:I.5.1;I.2。7

206. 第 xiv:1807.03625[pdf, ps,其他] Cs。Sd

从语音模式中调节外语教学

作者:fedor kitashov, Elizaveta svitanko, debojyoti dutta

摘要：最先进的自动语音识别 (asr) 系统与罕见口音数据的缺乏作斗争。对于足够大的数据集，神经引擎在大多数**自然语言处理**问题中往往优于统计模型。然而，语音重音仍然是这两种方法的挑战。音师们手动创建描述说话人口音的一般规则，但他们的结果仍然没有得到充分利用。在本文中，我们提出了一个模型，自动检索从一个小数据集的语音概括。这种方法利用了特定方言和普通美式英语 (gae) 之间的发音差异，并创建了新的重音单词样本。拟议的模型能够学习以前由音系学家手动获得的所有概括。我们使用这种统计方法从 cmu 发音词典生成了 100 万个语音变体，并训练了序列到序列 rnn，以 59% 的准确性识别重音单词。少

2018 年 7 月 9 日提交;最初宣布 2018 年 7 月。

207. 第 xiv:1807.02911[pdf,其他] Cs。Cl

多伊 10.107/978-3-319-99740-7 _ 12

阿拉伯情感分析的 cnn 和 lstm 联合模型

作者:abdulaziz m. alayba, vasile palade, matthew england, raat iqbal

摘要：深度神经网络在处理来自广泛应用领域的具有挑战性的大型数据集时，表现出了良好的数据建模能力。卷积神经网络 (cnn) 在选择良好的功能方面具有优势，长期短期存储器 (lstm) 网络已被证明具有良好的学习序列数据的能力。据报告，这两种方法都能在图像**处理**、语音识别、**语言**翻译和其他**自然语言处理**(nlp) 任务等领域取得更好的成果。twitter 短信的情感分类是一项具有挑战性的任务，阿拉伯语情感分类任务的复杂性增加，因为阿拉伯语是一种丰富的**形态语言**。此外，提供准确的阿拉伯文**预处理**工具是目前的另一个限制，同时在这一领域的研究有限。在本文中，我们研究了整合 cnn 和 lstm 的好处，并在不同数据集上的阿拉伯语情绪分析获得了更好的准确性。此外，我们还试图通过使用不同的情绪分类水平来考虑特定阿拉伯语词的形态多样性。少

2018 年 7 月 21 日提交;v1 于 2018 年 7 月 8 日提交;最初宣布 2018 年 7 月。

评论:作者接受的 ccd-make 提交版本

类:I.2.7;I.2。6

日记本参考:专业进修机器学习和知识提取国际跨领域会议。cd-make 2018。《计算机科学讲座笔记》，第 1115 卷, 179-191 页。springer, cham

208. 第 xiv:1807.02903[pdf,其他] Cs。Cl

通过词汇嵌入预测语言内和语言之间的单词的控制和可感性

作者:nikola ljubešić, darja fišer , anita peti-stantić

摘要: 在心理语言学中，具体性和可象似性的概念历来很重要，在面向语义的**自然语言处理**任务中具有重要意义。本文以词法嵌入为解释变量，通过监督学习研究了这两个概念的可预测性。我们通过利用与单个向量空间对齐的跨语言嵌入集合，在**语言**内部和语言之间执行预测。我们表明，具体性和可成像性的概念在**语言**内部和语言之间都是高度可预测的，在跨**语言**预测时，在相关性中的适度损失高达 20%。我们进一步表明，通过单词嵌入跨语言的迁移比通过双语词典进行简单的跨语言迁移更有效。少

2018 年 7 月 8 日提交;最初宣布 2018 年 7 月。

209. 第 xiv:1807.02857[pdf,其他] Cs。Lg

利用递归神经网络学习序列时间信息

作者:普什帕拉贾·穆鲁根

摘要: 递归网络是**处理自然语言**、声音、时间序列数据等序列数据的最强大、最有前途的人工神经网络算法之一。与传统的前馈网络不同，经常网络具有固有的反馈循环，它允许存储时间上下文信息，并将信息状态传递给事件的整个序列。这有助于在**语言**建模、股市预测、图像字幕、语音识别、机器翻译和目标跟踪等许多重要任务中实现最先进的性能，然而，培训完全连接的人 mn 和梯度流的管理是一个复杂**的过程**。为解决上述限制，进行了许多研究。本文的目的是提供有关复发神经元，其差异和行程技巧的简要细节，以训练完全复发的神经网络。本审查工作是作为我们的 ipo 工作室软件模块 " 多个目标跟踪" 的一部分进行的。少

2018 年 7 月 8 日提交;最初宣布 2018 年 7 月。

评论:17 页

210. 第 1807.02471[pdf, ps,其他] Cs。红外

深层学习对情感分类的不同词语嵌入研究综述

作者:德巴德里·达塔

摘要: 网络上充满了文本内容,**自然语言处理**是机器学习中最重要的领域之一。但是，当数据巨大时，简单的机器学习算法就无法处理它，而当基于神经网络的深度学习发挥作用时。然而，由于神经网络无法**处理**原始文本，我们必须通过一些不同的词嵌入策略来改变它们。本文演示了在 amazon 评论数据集上实现的那些独特的词嵌入策略，它有两种需要分类的情绪: 基于众多客户评论的快乐和不快乐。此外，我们证明了在准确性的区别与关于嵌入哪个词应用的话语什么时候。少

2018 年 7 月 5 日提交;最初宣布 2018 年 7 月。

211. 第 1807.02458[pdf,其他] Cs。铬

安全相关文件自动分类的一种实用方法

作者:安东尼诺·萨贝塔,米歇尔·贝齐

摘要: 缺乏关于开放源码软件组件脆弱性的可靠详细资料来源是维持安全软件供应链和有效的脆弱性管理**进程**的主要障碍。众所周知，国家脆弱性数据库等标准咨询和脆弱性数据来源覆盖面差，质量不一致。为了减少对这些源的依赖，我们提出了一种方法，该方法使用机器学习来分析源代码存储库，并自动识别与安全相关的提交 (即可能修复漏洞的提交)。我们将提交引入的源代码更改视为用**自然语言**编写的文档，并使用标准文档分类方法对其进行分类。结合使用来自提交的不同方面的信息的独立分类器，我们的方法可以产生高精度 (80%)，同时确保可接受的召回 (43%)。特别是，使用从源代

码更改中提取的信息，可以与最流行的方法相比有很大的改进，同时需要的培训数据量要小得多，并采用更简单的体系结构。少

2018 年 7 月 6 日提交;最初宣布 2018 年 7 月。

评论:发表于第 34 届 ieee 软件维护与进化国际会议 (icsme) 2018 年课程

212. 第 xiv:1807.02383[pdf,其他] Cs。CL

信息提取的自然语言处理

作者:索尼特·辛格

摘要: 随着数字时代的兴起，新闻、文章、社交媒体等形式的信息爆发。这些数据大多以非结构化的形式存在，手动管理和有效地利用这些数据是繁琐、繁琐和耗费人力的。这种信息的爆炸式增长和对更复杂和更有效的信息处理工具的需求产生了信息提取 (ie) 和信息检索 (ir) 技术。信息提取系统将**自然语言**文本作为输入，并生成由某些标准指定的结构化信息，这些信息与特定应用程序相关。ie 的各种子任务，如命名实体识别、核心解析、命名实体链接、关系提取、知识库推理，构成了各种高端**自然语言处理的基石**(nlp) 任务，如机器翻译，问答系统,**自然语言**理解，文本摘要和数字助理，如 siri, cortana 和谷歌现在。本文介绍了信息提取技术及其各种子任务，重点介绍了各种 ie 子任务的最新研究、当前的挑战和未来的研究方向。少

2018 年 7 月 6 日提交;最初宣布 2018 年 7 月。

评论:24 页, 1 个图

213. 特别报告: 1807.02257[pdf,其他] Cs。简历

基于自然语言查询的动态多模式实例分割

作者:edgar Margffoy-Tuay, juan c. pérez, emilio botero, pablo arbeláez

摘要: 我们解决了在给定描述对象的**自然语言**表达式的情况下对对象进行分段的问题。目前的技术通过直接或递归合并通道维度中的语言和视觉信息，然后执行卷积来处理这一任务;或通过 (\ texitii) 将表达式映射到可将其视为筛选器的空间，该空间的响应与图像中给定空间坐标处的对象的存在直接相关，以便可以应用卷积来查找对象。为了充分利用语言的递归性质，我们提出了一种将**这**两个见解结合起来的新方法。此外，在上采样**过程**中，我们还利用了在下采样图像时产生的中间信息，从而获得了详细的分割。我们将我们的方法与四个标准数据集中的最先进方法进行比较，在这些方法中，它在此任务的八个拆分中，它超过了以前的所有方法。少

2018 年 7 月 22 日提交;v1 于 2018 年 7 月 6 日提交;最初宣布 2018 年 7 月。

214. 第 xiv:180.7.02250[pdf,其他] Cs。简历

面部帽: 使用面部表情分析的图像字幕

作者:omid mohamad nezami, mark dras, peter anderson, len hamey

摘要: 图像字幕是生成图像的**自然语言**描述的**过程**。然而，目前的大多数图像字幕模型都没有考虑到形象的情感方面，这与其中所代表的活动和人际关系非常相关。为了开发一个模型，可以产生人类样的标题与这些，我们使用从图像中提取的面部表情特征，包括人脸，目的是提高模型的描述能力。在本工作中，我们提出了两个变种的脸谱帽模型，嵌入面部表情特征以不同的方式，以生成图像标题。使用所有标准评估指标，我们的 face-cap 模型在应用于从标准 flickr 30k 数据集中提取的图像标题数据集 (由大约 11k 图像组成) 时，其性能优于生成图像标题的最先进的基线模型包含的面。对字幕

的分析发现，也许令人惊讶的是，字幕质量的提高似乎不是来自与图像的情感方面有关的形容词的增加，而是来自字幕中描述的动作的更多变化。少

2018 年 7 月 6 日提交;最初宣布 2018 年 7 月。

215. 第 1807. 02221 [pdf] Cs。Cl

好莱坞的数据科学：利用电影的情感弧线推动娱乐业商业模式创新

作者:marco del vecchio,亚历山大 kharlamov, glennparry, ganna pogrebna

摘要: 许多商业文献都涉及以消费者为中心的设计问题：企业如何设计能够准确反映消费者偏好的定制服务和产品？本文运用数据科学**的自然语言处理**方法，探讨情绪是否以及在多大程度上塑造了消费者对媒体和娱乐内容的偏好。使用由 6, 174 个电影脚本组成的独特过滤数据集，我们生成屏幕内容的映射，以捕捉每个电影的情感轨迹。然后，我们将获得的映射组合到表示消费者情感旅程分组的集群中。这些集群用于预测电影的整体成功参数，包括票房收入、观众满意度 (由 imdb 评分)、奖项以及观众和评论家评论的数量。我们发现，和书一样，所有的电影故事都被 6 个基本的形状所主导。最高的票房与洞形的人有关，其特点是情感下降，然后是情感的上升。这种形状导致财务上成功的电影，而不考虑类型和制作预算。然而，《人在洞里》成功并不是因为它产生了最 "喜欢" 的电影，而是因为它产生了最 "谈论" 的电影。有趣的是，精心挑选的制作预算和体裁组合可能会产生一个财务上成功的电影与任何情感的形状。讨论了这一分析对产生点播内容和推动娱乐业商业模式创新的意义。少

2018 年 7 月 10 日提交;v1 于 2018 年 7 月 5 日提交;最初宣布 2018 年 7 月。

216. 特别报告：1807. 02200 [pdf,其他] Cs。Cl

多伊 10.1080/09298215.2018.1488878

音乐知识发现的自然语言处理

作者:sergio oramas, luis Espinosa-Anke, francisco gómez, xavier serra

摘要: 今天，大量的音乐知识以书面形式储存，证词可以追溯到几个世纪前。在这项工作中，我们提出了不同**的自然语言处理**(nlp) 方法，以利用这些文本集合的潜力，自动发现音乐知识，涵盖原型 nlp 中的不同阶段管道，即语料库汇编、文本挖掘、信息提取、知识图生成和情感分析。每一种方法都与处理大量文件的不同用例 (即弗拉门戈、文艺复兴和流行音乐) 一起介绍，并提出和讨论数据驱动分析得出的结论.少

2018 年 7 月 5 日提交;最初宣布 2018 年 7 月。

日记本参考:新音乐研究杂志 (2018)

217. 第 xiv:1807. 02164 [pdf, ps,其他] Cs。Hc

当卷积神经网络遇到数据可视化时

作者:毛阳,李波,冯冠雄,严忠江

摘要: 近年来，深度学习几乎在各个领域都构成了深刻的技术革命，引起了产业界和学术界的高度关注。特别是卷积神经网络 (cnn) 作为深度学习的一种代表性模型，在计算机视觉和**自然语言处理**方面取得了巨大的成功。然而，简单或盲目地将美国有线电视新闻网应用于其他领域会导致较低的训练效果，或使调整模型参数变得相当困难。在这张海报中，我们提出了一个名为 v-inn 的一般方法，介绍了美国有线电视新闻网的数据可视化。v-nnn 在美国有线电视新闻网建模之前引入了一个数据可视化模型，以确保**处理**后的数据适合图像的特征以及 cnn 建模。我们将 v-inn 应用于基于著名实用

数据集 awid 的网络入侵检测问题。仿真结果表明, v-nnc 明显优于其他研究，每个入侵类别的召回率均在 99.8% 以上。少

2018 年 6 月 12 日提交;最初宣布 2018 年 7 月。

评论:向 2018 年向 sigcomm 提交 2 页, 2 个数字

218. **第 xiv:1807. 01855[pdf] Cs。CI**

zipf 的 50 种语言定律: 结构模式、语言解释和认知动机

作者:余水元,徐春山,刘海涛

摘要: zipf 定律在包括语言在内的许多与**人类**有关的领域都有发现，在这些领域，一个词的频率一直被发现为其频率等级的权力法律功能，即 zipf 定律。然而，无论是普遍的法律还是统计伪影，都存在很大的争议，对可能形成它的机制了解甚少。为了回答这些问题，本研究对齐普夫定律进行了大规模的**跨语言**调查。统计结果表明, zipf 的 50 **种语言**的定律都有 3 段的结构模式，每个片段都表现出独特的语言性质，而下一段总是向下弯曲，偏离理论期望。这一发现表明，这种偏差是**自然语言**中词频分布的一个基本和普遍的特征，而不是低频词的统计误差。基于双**过程**理论的计算机模拟得出了具有相同结构模式的 zipf 定律，表明 zipf 的**自然语言**定律是由共同的认知机制驱动的。这些结果表明, zipf 的**语言**定律是由控制人类语言行为的双重**处理**等认知机制所驱动的。少

2018 年 7 月 5 日提交;最初宣布 2018 年 7 月。

评论:18 页, 3 个数字

219. **xiv:180. 07. 01704[pdf] Cs。CI**

一种用于视距情感分类的卷积神经网络

作者:邢永平,肖创白,吴一飞,丁子明

摘要: 随着互联网的发展,**以**情感分析为重要任务的自然**语言处理**(nlp) 在信息处理中变得至关重要。情感分析包括方面情感分类。方面情绪可以提供完整和深入的结果，增加了对方面层面的关注。句子中不同的语境词会不同地影响句子的情感极性，而极性则会根据句子中的不同方面而变化。拿一句话来说, ' 我买了一台新相机。图片质量惊人，但电池寿命太短。如果方面是图片质量，那么预期的情绪极性是 ' 积极的 '，如果考虑电池寿命方面，那么情绪极性应该是 ' 负 ';因此，在探索句子中的方面情绪时，方面是很重要的考虑因素。递归神经网络 (rnn) 被认为是处理**自然语言处理**的良好模型，而 rnn 在方面情感分类方面取得了良好的性能，包括目标依赖 lstm (td-lstm)。目标连接 lstm (tc-lstm) (tang, 2015a, b), ae-lstm, at-lstm, aeat-lstm (wang 等人, 2016 年)。关于利用卷积神经网络进行情感分类的文献也越来越多，但关于利用卷积神经网络进行方面情感分类的文献却很少。在本文中，我们开发了基于注意的输入层，其中信息是由输入层考虑的。然后，我们将基于注意的输入层合并到卷积神经网络 (cnn) 中，以引入上下文单词信息。在我们的实验中，将方面信息融入美国有线电视新闻网可以提高后者的方面情感分类性能，而不使用 twitter 基准数据集中的句法分析器或外部情感词汇，但却能获得更好的性能与其他型号相比。少

2018 年 7 月 4 日提交;最初宣布 2018 年 7 月。

220. **第 1807. 01670[pdf,其他] Cs。CI**

从自然语言编码空间关系

作者:tiago ramalho, tomaškočiskü, fredic besse, s . m. ali esami, gábor melis, fabio viola, phil blunsom, karl moritz hermann

摘要: **天然**的**的语言处理**在通过分布方法学习词汇语义方面取得了显著的进展，但通过这些方法学到的表示无法捕捉到现实世界中隐含的某些种类的信息。特别是，空间关系的编码方式与人类的空间推理不一致，对视点变化缺乏不变性。我们提出了一个系统，能够从**自然语言**中捕获空间关系的语义，如背后、左边等。我们的主要贡献是一个新的多模态目标，其基础是从文本描述生成场景图像，以及一个新的数据集来训练它。我们证明了内部表示对保留描述转换的意义是鲁棒性的 (转述不变性)，而视点不变性是系统的一个紧急属性。少

2018 年 7 月 5 日提交;v1 于 2018 年 7 月 4 日提交;最初宣布 2018 年 7 月。

221. **第 1807. 01337[pdf,其他] Cs。Lg**

多伊 10.11145/3219819851

cota: 通过排名和深度网络提高客户支持的速度和准确性

作者:piero molino,怀秀郑州, yi-chia wang

摘要: 对于一家希望提供令人愉快的用户体验的公司来说，处理任何客户问题都是至关重要的。本文提出了 cota 系统，通过自动机票分类和支持代表的答案选择，提高最终用户客户支持的速度和可靠性。演示了两种机器学习和**自然语言处理**技术: 一种是依靠特征工程 (cota v1)，另一种是通过深度学习架构 (cota v1) 利用原始信号。cota v1 采用了一种新的方法，将多分类任务转换为排名问题，在数千个类的情况下表现出更好的性能。对于 cota v2，我们提出了一个编码器-组合解码器，这是一种新颖的深度学习体系结构，它允许异构输入和输出特征类型，并通过网络体系结构选择注入先前的知识。本文将这些模型及其变种与门票分类和答案选择的任务进行了比较，显示了 cota v2 模型优于 cota v2 的模型，并分析了它们的内部工作原理和缺点。最后，在生产环境中进行了 aeb 测试，验证 cota 在不降低客户满意度的情况下将问题解决时间缩短 10% 的实际影响。少

2018 年 7 月 3 日提交;最初宣布 2018 年 7 月。

222. **第 xiv:1807. 00914[pdf,其他] Cs。Cl**

语言变异与通用建模: 自然语言处理中的类型学语言学研究综述

作者:edoardo maria ponti, helen o ' horan, yevgeni berzak, ivan vulić, roi reichart, thierry poibeau, ekaterina shutova, anna korhonen

摘要: 解决语法结构和意义分类的跨语言变异是多语言自然 **语言处理**的关键挑战。世界上大多数**语言**缺乏资源，使监督学习不可行。此外，大多数算法的性能受到特定语言偏见和对信息多语言数据的忽视的阻碍。语言类型学的学科提供了一个原则框架，可以系统地和经验地比较**语言**，并记录其在公开数据库中的差异。这些文件包含了设计独立**于语言**的算法和改进旨在缓解上述问题的技术的关键信息，包括具有类型学特征的跨语言迁移和多语言联合模型。在本调查中，我们证明了类型学对多个 nlp 应用程序都有好处，这些应用程序涉及语义任务和句法任务。此外，我们还概述了从数据库中提取要素或自动获取这些功能的几种技术: 这些功能随后可以集成到多语言模型中，以便交叉地将参数连接在一起，或者将模型调整为特定的**语言**。最后，我们提倡一种新的类型学，它考虑到单个示例中的模式，而不是整个**语言**中的模式，并考虑到分级的类别而不是离散的类别，以弥补与上下文和连续的**差距**机器学习算法的性质。少

2018 年 7 月 2 日提交;最初宣布 2018 年 7 月。

223. 第 xiv:1807.00847[pdf,其他] Cs。Lg

使 (近) 每一个神经网络更好: 通过权重参数重采样生成神经网络

作者:刘嘉义, samarth tripathi, unmesh k 第 3 款, mohak shah

文摘: 深神经网络 (dnn) 在计算机视觉、**自然语言处理**和其他领域越来越流行。但是,对深度学习模型进行培训和微调是计算密集型和耗时的。我们提出了一种新的方法来提高几乎所有模型的性能,包括预先培训的模型。该方法采用集成方法,通过根据这些参数的概率分布重新分配模型参数值来构造集成中的网络,该参数分布在训练**过程**快结束时计算。对于预先训练的模型,此方法会导致额外的训练步骤 (通常少于一个时代)。我们使用 mnist 数据集执行各种分析,并使用 imagenet 数据集上的预先培训模型,使用多个 dnn 模型验证该方法。少

2018 年 7 月 2 日提交;最初宣布 2018 年 7 月。

评论:接受 uai 深度学习不确定性研讨会

224. 第:1807.00791[pdf] Cs。CI

基于自然语言界面的结构化数据查询的语用方法

作者:aliaksei vertsel, mikhail Rumiantsau

摘要: 随着技术使用的增加和数据分析在许多企业中的组成部分,快速访问和解释数据的能力变得比以往任何时候都更加重要。各组织和公司正在利用信息检索技术来管理其信息系统和**流程**。尽管在关系数据库中组织了大量高效数据的信息检索,但用户仍然需要掌握数据库**语言**/模式来完全制定查询。这给组织和公司带来了雇用精通 db **语言**/架构的员工来制定查询的负担。为了减轻已经捉襟见肘的数据团队的一些负担,许多组织都在寻找允许非开发人员查询其数据库的工具。不幸的是,编写一个有效的 sql 查询来回答用户试图提出的问题并不总是那么容易的。即使是看似简单的问题,比如 " 哪些创业公司获得了超过2亿美元的资金? " 实际上也很难回答,更不用说转换为 sql 查询了。您如何定义初创公司? 按大小、位置、时间长短分列,它们的成立时间? 如果用户正在使用他们已经熟悉的数据库,这可能是可以的,但如果用户不熟悉数据库,这将是正常的。现在需要的是一个集中的系统,它可以有效地将**自然语言**查询转换为不同客户数据库类型的特定数据库查询。有许多因素会显著影响系统体系结构和用于将 nl 查询转换为结构化查询表示的算法集。少

2018 年 7 月 2 日提交;最初宣布 2018 年 7 月。

评论:8 页, 1 个图

225. 第:1807.00735[pdf,其他] Cs。CI

赫赫-日

作者:何阳辉, vishnu jejjala, brent d . nelson

文摘: 我们在**自然语言处理**、计算语言学和机器学习方面的技术在 arxiv 的六和四个相关部分进行论文的研究: hi-ph 、hep-lat 、gr-qc 和 math ph。从 arxiv 的开始到 2017 年底,每个部分的论文标题都被提取出来,并被视为我们用来训练神经网络 2wordec 的语料库。对常见的 n 克、线性句法身份、字云和词的相似性进行了比较研究。我们发现这些领域之间存在着显著的科学和社会学差异。结合支持向量机,我们还表明,在高能和数学物理的不同子领域的标题的句法结构是足够不同的,神经网络可以执行形式的二元分类与现象学部分相比,精度为 87.1%,可在所有部分执行更精细的 5 倍分类,准确率为 65.1%。少

2018 年 6 月 27 日提交;最初宣布 2018 年 7 月。

评论:50 页, 6 个数字

226. 第 xiv:1807. 00543[pdf,其他] Cs。CI

会话语音的标点符号预测模型

作者:piotr al 戈尔 ko, piotrSzymański, jan mizgajski, adrian szymczak, yishay carmiel, najim dehak

摘要: asr 系统通常不预测任何标点符号或大小写。缺少标点符号会导致结果呈现中出现问题，并使人类读者和现成的**自然语言处理**算法变得混乱。为了克服这些限制，我们训练了两个变体的深神经网络 (dnn) 序列标记模型-双向长期短期存储器 (blstm) 和卷积神经网络 (cnn)，以预测标点符号。模型是在费舍尔语料库上训练的，其中包括标点符号注释。在我们的实验中，我们使用序列对齐算法结合时间对齐和标点符号费舍尔语料库的记录。神经网络是在常见的网络爬网格洛维嵌入的单词在费舍尔的文字记录与对话侧指标和单词时间信息一致。cnns 的精度更高, blstm 往往有更好的召回。虽然 blstm 总体上犯的错误较少，但美国有线电视新闻网预测的标点符号更准确--特别是在问号的情况下。我们的研究结果证明了单词在时间上的分布，以及预先训练的嵌入，在标点符号预测任务中是有用的。少

2018 年 7 月 2 日提交;最初宣布 2018 年 7 月。

评论:接受 2018 年互讲会议

227. 建议: 1807. 00267[pdf,其他] Cs。CI

一种理解口语的编码上下文的有效方法

作者:Raghav gupta, abhinav Rastogi, dilek hakani-tur

摘要: 在面向任务的对话系统中，口语**理解**(slu) 是指将**自然语言**使用者的话语解析为语义框架的任务。利用先前对话历史的背景是更有效的 slu 的关键。slu 的最新方法是使用内存网络在每个回合处理对话中的多个话语来对上下文进行编码，从而在准确性和计算效率之间进行重大权衡。另一方面，像对话状态跟踪器 (dst) 这样的下游组件已经跟踪对话状态，这可以作为对话历史的总结。在这项工作中，我们提出了一种有效的方法来编码上下文从先前的话语 slu。更具体地说，我们的体系结构包括一个单独的基于重复神经网络 (rnn) 的编码模块，该模块积累对话上下文，以指导帧解析子任务，并可在 slu 和 dst 之间共享。在我们的实验中，我们从两个领域展示了我们的对话方法的有效性。少

2018 年 7 月 1 日提交;最初宣布 2018 年 7 月。

评论:提交给国际语音

228. 第 xiv:1806. 098009[pdf,其他] Cs。简历

用自然语言生成反事实解释

作者:lisa anne hendricks, rang hu, trevor darrell, zeynep akata

摘要: **天然**的对深度神经网络决策的**语言**解释为 ai 代理阐明推理**过程**提供了一种直观的方法。目前的文本解释学习讨论图像中的阶级判别特征。但是，如果图像中存在，了解哪些属性可能会更改分类决策也很有帮助 (例如, "这不是 scarlet tanager，因为它没有黑色翅膀")。我们将这种文本解释称为反事实解释，并提出一种直观的方法，通过检查输入中缺少哪些证据来生成反事实解释，但如果存在，可能会导致不同的分类决定在图像中。为了演示我们的方法，我们考虑了一个细粒度的图像分类任务，在这个

任务中，我们将图像作为输入，并将反事实类和输出文本作为输入，这解释了为什么图像不属于反事实类。然后，我们使用建议的自动度量标准，对生成的反事实解释进行定性和定量分析。少

2018 年 6 月 26 日提交;最初宣布 2018 年 6 月。

评论:在 2018 年计算机学习中的人类可解释性 icml 研讨会 (whi 2018) 上，瑞典斯德哥尔摩

229. 第 xiv:1806.0953[pdf, ps,其他] Cs。Cl

在新闻标题上使用 nlp 预测指数趋势

作者:marc velay, fabrice daniel

文摘:本文试图利用新闻标题提供一种趋势预测的最新技术。我们介绍了使用**自然语言处理**预测 djia 趋势的研究。我们将解释我们使用的不同算法以及各种嵌入技术...

2018 年 6 月 22 日提交;最初宣布 2018 年 6 月。

230. 第 xiv:1806.09504[pdf,其他] Cs。艾

解释知识库的嵌入模型：一种教学方法

作者:arthur colcollini 古斯芒, alvaro henrique chaim correia, glauber de bona, fabio gagliardi cozman

摘要: 知识库被用于**从自然语言处理**到语义 web 搜索的各种应用;唉，在实践中，他们的有用性是由他们的不完整伤害。嵌入模型在知识库完成过程中获得了最先进的精度，但众所周知，它们的预测很难解释。本文采用 "教学方法" (从神经网络文献中) 来解释嵌入模型，从中提取加权霍恩规则。我们展示了如何调整教学方法，以适应知识库的大规模关系方面，并通过实验显示它们的优缺点。少

2018 年 6 月 20 日提交;最初宣布 2018 年 6 月。

评论:在 2018 年计算机学习中的人类可解释性 icml 研讨会 (whi 2018) 上，瑞典斯德哥尔摩

231. 第 6.6:1806.09464[pdf,其他] Cs。Lg

学习用于紧凑型嵌入表示的 k 路 d 维离散代码

作者:陈婷,陈仁强敏,孙一洲

摘要: 传统的嵌入方法将每个符号直接与一个连续嵌入向量相关联，这相当于应用基于离散符号的 "一热" 编码的线性变换。尽管这种方法很简单，但它产生的参数数量随词汇量的大小呈线性增长，并可能导致过度拟合。在这项工作中，我们提出了一个更紧凑的 k 路 d 维离散编码方案，以取代 "一热" 编码。在建议的 "kd 编码" 中，每个符号都由 D-维度代码，基数为 K 通过构成代码嵌入向量，生成最终的符号嵌入向量。对于具有语义意义的端到端学习代码，我们推导出一种基于随机梯度下降的宽松离散优化方法，该方法一般可应用于任何具有嵌入层的可微化计算图。在我们对从**自然语言处理**到图形卷积网络的各种应用进行的实验中，嵌入层的总尺寸可减少 98%，同时实现相似或更好的性能。少

2018 年 6 月 21 日提交;最初宣布 2018 年 6 月。

评论:icml 2018. arxiv 管理说明: 文本与 arxiv:1711.03067 重叠

232. 第 xiv:1806.09089[pdf,其他] Cs。Cl

密集连接的网络特征级特征提取

作者:李昌熙,金英邦,李东宇,林海硕

摘要: 生成字符级功能是在各种**自然语言处理任务中**取得良好效果的重要步骤。为了减轻人工生成人工特征的需要，提出了利用神经结构 (如卷积神经网络 (cnn) 或递归神经网络 (rnn) 自动提取此类特征的方法，并已显示了很好的结果。但是，cnn 生成与位置无关的功能，而 rnn 速度较慢，因为它需要按顺序**处理**字符。本文提出了一种利用密集连接网络自动提取字符级特征的新方法。拟议的方法不需要任何**特定于语言**或任务的假设，并显示出稳健性和有效性，同时比基于 cnn 或 rnn 的方法更快。在三个序列标记任务 (插槽标记、语音部分 (pos) 标记和命名实体识别 (ner)) 上评估此方法，我们获得了最先进的性能，插槽标记和 pos 标记的精度为 96.62 f1 和 97.73%，和相当的性能，最先进的 91.13 f1 分数上的净入学率。少

2018 年 7 月 26 日提交;v1 于 2018 年 6 月 24 日提交;最初宣布 2018 年 6 月。

评论:12 页，4 位数字，在 coling 2018 被接受为会议文件

233. 第 xiv:1806.08894[pdf,其他] Cs。Lg

多伊 10.1007/978-3-319-56991-8 _ 32

深度强化学习：概述

作者:seyed sajad mousavi, michael schukat, enda howley

摘要: 近年来，一种名为深度学习的特定机器学习方法获得了巨大的吸引力，因为它在模式识别、语音识别、计算机视觉和**自然等广泛应用中取得了惊人的效果语言处理**。最近的研究还表明，深度学习技术可以与强化学习方法相结合，以学习高维原始数据输入问题的有用表示。本章回顾了深度强化学习的最新进展，重点介绍了最常用的深层体系结构，如自动编码器、卷积神经网络和递归神经网络，这些体系结构已成功地与强化学习框架。少

2018 年 6 月 22 日提交;最初宣布 2018 年 6 月。

评论:2016 年 sai 智能系统会议论文集

234. 第 1806.08730[pdf,其他] Cs。Cl

自然语言十项全能：多任务学习作为问题回答

作者:bryan mccann, nitish shirish keskar, c 林区 xong, richard socher

文摘: 深度学习提高了许多**自然语言处理**(nlp) 任务的单独性能。但是，一般的 nlp 模型不能出现在一个侧重于单个指标、数据集和任务的特殊性的范式中。我们介绍**了自然语言**十项全能 (十项全能)，这是一个跨越十个任务的挑战：回答问题、机器翻译、总结、**自然语言**推理、情绪分析、语义角色标记、零镜头关系提取、目标导向对话、语义解析和常识代词解析。我们将所有任务都转换为上下文回答问题。此外，我们提出了一个新的多任务问题回答网络 (mqan) 共同学习 decaNLP 中的所有任务，而无需在多任务设置中的任何特定任务的模块或参数。mqan 在机器翻译的转移学习和命名实体识别、情绪分析和**自然语言**推理的领域适应以及文本分类的零拍摄能力方面都有改进。我们证明，mqan 的多指针-发电机解码器是这一成功的关键，并且通过反课程培训策略进一步提高了性能。虽然 mqan 是为 decaNLP 设计的，但在单任务设置中，它也在 wiksql 语义分析任务上实现了最先进的结果。我们还发布了用于采购和**处理**数据、培训和评估模型以及复制 decaNLP 的所有实验的代码。少

2018 年 6 月 20 日提交;最初宣布 2018 年 6 月。

235. 第 1806.08724[pdf,其他] Cs。Sd

声调和谐的语言模型评价

作者:david r. w. sears , Filip korzeniowski , gerhard widmer

文摘: 本研究借鉴并扩展了自然 **语言处理**中的概率**语言**模型，以发现声调和谐的句法特征。**语言**模型有许多形状和大小，但它们的中心目的总是相同的：以字母、单词、音符或和弦的序列预测下一个事件。然而，很少有研究使用这种模型来评估最先进的建筑使用大规模的西方色调音乐语料库，而不是更愿意使用相对较小的数据集，其中包含当代类型的和弦注释比如爵士乐、流行音乐和摇滚。利用共同实践时期突出的器乐体裁的符号表示，本研究应用一种灵活的数据驱动编码方案，(1) 评估有限上下文 (或 n-gram) 模型和和弦中的递归神经网络 (rnn) 预测任务;(2) 比较每个选定数据集中和弦集的最佳表现模型的预测精度;(3) 在回归分析中解释两种模型体系结构之间的差异。我们发现，使用部分匹配 (ppm) 算法预测的有限上下文模型的性能优于 rnn，特别是对于钢琴数据集，回归模型表明 rnn 与特别罕见的和弦类型作斗争。少

2018 年 6 月 22 日提交;最初宣布 2018 年 6 月。

评论:7 页，4 个数字，3 个表。发表于第 19 届国际音乐学会信息检索大会 (ismir) 论文集，法国巴黎

236. 第 xiv:1806.08077[pdf,其他] Cs。CL

用于释义生成的字典引导编辑网络

作者:黄少汉,吴宇,傅伟,周明

摘要: 一个直观的方法，为人类写释义句子是取代原来的句子中的单词或短语与其相应的同义词，并作出必要的改变，以确保新的句子是流利的和语法正确的。我们提出了一种用句子引导编辑网络**对过程**进行建模的新方法，该网络有效地对源句进行重写，生成释义句子。它结合在原始句子的上下文中，从现成的字典中学习适当的单词级和短语水平的释义对，以及生成流畅的**自然语言**句子。具体而言，系统检索从原句子的回传数据库 (ppdb) 派生的一组单词级别和短语级参照对，用于指导可以删除或插入单词的决定，并注意这两个词序列到序列框架下的机制。我们对两个用于转述生成的基准数据集 (mscoco 和 quora 数据集) 进行了实验。评估结果表明，我们的字典引导编辑网络优于基线方法。少

2018 年 6 月 21 日提交;最初宣布 2018 年 6 月。

237. 第 xiv:1806.07978[pdf,其他] Cs。Lg

语料库复制任务

作者:托比亚斯·艾金格

文摘: 在**自然语言处理**(nlp) 领域，我们重新审视了众所周知的词嵌入算法 word2vec。单词嵌入按向量标识单词，以便捕获单词的分布相似性。出乎意料的是，除了语义相似性外，在单词2vec 生成的单词嵌入中甚至关系相似性也被证明是被捕获的，由此产生了两个问题。首先，哪种关系可以在连续空间中表示，其次，如何建立关系。为了解决这些问题，我们提出了自下而上的观点。我们称生成输入文本，其中 word2vec 输出目标关系解决语料库复制任务。我们认为这种方法对任何可能的关系集都是一种概括，因此我们希望语料库复制任务的解决能够为这些问题提供部分答案。少

2018 年 6 月 20 日提交;最初宣布 2018 年 6 月。

评论:引用可能无法适当呈现. 请与作者联系以了解详细信息

238. 第 1806.07977[pdf,其他] Cs。CL

多伊10323/jifs-16984

txpi-u：大学生人格认同的资源

作者:gabriela Ramírez-de-la-Rosa, esaquu villatoro-tello, héctor jiménez-salazar

摘要: 在 标记语料库等资源是在**自然语言处理**(nlp) 领域中训练自动模型所必需的。从历史上看，关于大量问题的大量资源大多以英文提供。其中一个问题被称为人格识别，在这种情况下，基于心理模型 (例如五大模型)，目标是找到一个主体的个性特征，例如，由同一主题编写的文本。本文介绍了一种新的西班牙语料库，称为人格识别文本 (txpi)。此语料库将有助于开发模型，自动将个性特征分配给文本文档的作者。我们的语料库 txpi-u 包含 416 墨西哥本科生的信息，其中包含一些人口统计信息，如年龄、性别和他们注册的学术课程。最后，作为额外的贡献，我们提出了一套基线，为进一步研究提供了一个比较方案。少

2018 年 6 月 20 日提交;最初宣布 2018 年 6 月。

日记本参考:智能与模糊系统杂志，第 34 卷，第 5 期,第 291-3001 页, 2018 年

239. 第 xiv:1806. 07721[pdf,其他] Cs。CI

语义关系分类：任务形式化与细化

作者:vivian s. silva, manuela hürliman, brian davis, siegfried handschuh, andréfreitas

摘要: 在自然 **语言** 处理中，识别文本中术语之间的语义关系是一项基本任务，它可以支持需要轻量级语义解释模型的应用。目前，语义关系分类主要集中在开放域数据中评估的关系。这部作品对用于语义关系分类的抽象关系集进行了批判，这些关系涉及它们表达在领域特定语料库中发现的术语之间的关系的能力。在此基础上，提出了一种基于重用和扩展 d 变性本体中存在的抽象关系集的替代语义关系模型。由此产生的关系集基础充分，可以捕获广泛的关系，因此可以作为语义关系自动分类的基础。少

2018 年 6 月 20 日提交;最初宣布 2018 年 6 月。

评论:10 页，在《 2016 年 cogalex-v 》上提交

日记本参考:第五次词汇认知方面研讨会论文集，日本大阪, 2016 年，第 30-39 页

240. 第 xiv:1806. 07711[pdf,其他] Cs。CI

字典定义的语义角色的分类

作者:vivian s. silva, siegfried handschuh, andréfreitas

摘要: 理解术语之间的语义关系是**自然语言处理**应用程序中的一项基本任务。虽然能够以正式方式 (如本体) 表达这些关系的结构化资源仍然很少，但收集字典定义的大量语言资源正在变得可用，但对**自然**的**语言**定义是使它们在语义解释任务中有用的基础。在分析 wordnet 光泽度的一个子集的基础上，我们提出了一套语义角色，构成了字典定义的语义结构，并说明了它们与定义的句法结构的关系，确定了可以使用的模式在信息提取框架和语义模型的发展中。少

2018 年 6 月 20 日提交;最初宣布 2018 年 6 月。

评论:9 页, 2 个数字，在 cogalex-v 2016 上提交

日记本参考:第五次词汇认知方面研讨会论文集 (cogalex v)，日本大阪, 2016 年，176-184 页

241. 第 1806. 0787[pdf, ps,其他] Cs。CI

自动事实调查：任务配方、方法和未来方向

作者:james thorne, and 列 as vlachos

摘要: 最近对错误信息的日益关注刺激了事实检查方面的研究，即评估索赔真实性的任务。在自然 **语言处理**、机器学习、知识表示、数据库和新闻等多个学科进行了自动化研究。虽然取得了重大进展，但在往往不了解对方和使用术语不一致的研究界发表了相关论文和文章，从而阻碍了理解和取得进一步进展。在本文中，我们调查了来自**自然语言处理**和相关学科的自动事实检查研究，统一了论文和作者之间的任务公式和方法。此外，我们强调，证据的使用是贯穿任务表述和方法的一个重要区别因素。最后，我们提出了未来 nlp 研究自动事实检查的途径。少

2018 年 9 月 5 日提交;v1 于 2018 年 6 月 20 日提交;最初宣布 2018 年 6 月。

评论:在第 27 届计算语言学国际会议 (coling 2018) 上发表

242. 第 1806. 07336[pdf,其他] Cs。Lg

神经密码理解：代码语义的一种学习性表示

作者:tal ben-nun, alice shoshana jakobov》, torsten hoefler

文摘: 随着**自然语言处理**中嵌入的成功，研究将类似的方法应用于代码分析。大多数作品试图直接**处理**代码或使用句法树表示，将其视为**用自然语言**编写的句子。但是，由于函数调用、分支和语句的可互换顺序等结构特征，现有的方法都不足以能够有力地理解程序语义。本文提出了一种新的**处理**技术来学习代码语义，并将其应用于各种程序分析任务。特别是，我们规定，一个鲁棒的代码分布假说适用于人和机器生成的程序。根据这一假设，我们定义了一个嵌入空间, inst2vec，基于独立于源编程**语言**的代码的中间表示 (ir)。我们利用程序的基础数据和控制流，为该 ir 提供了上下文流的新定义。然后，我们使用类比和聚类分析对嵌入进行定性分析，并评估三个不同高级任务的学习表示形式。我们表明，使用单一的 rnn 体系结构和预先训练的固定嵌入, inst2vec 优于专门的性能预测方法 (计算设备映射、最佳线程粗化);和算法分类从原始代码 (104 个类)，在那里我们设置了一个新的最先进的。少

2018 年 7 月 31 日提交;v1 于 2018 年 6 月 19 日提交;最初宣布 2018 年 6 月。

243. 第 1806. 07243[pdf,其他] Cs。简历

用于可解释视觉问题回答的学习条件图结构

作者:will Norcliffe-Brown, efstathios vafeias, sarah parisot

摘要: 视觉问题回答是一个具有挑战性的问题，需要从计算机视觉和**自然语言处理**的概念的组合。大多数现有的方法使用两个流策略，计算映像和问题功能，因此使用各种技术进行合并。尽管如此，很少有人依赖于更高级别的图像表示，这可以捕获语义和空间关系。在本文中，我们提出了一种新的基于图形的视觉问题回答方法。我们的方法结合了图形学习者模块，该模块学习输入图像的特定问题图形表示，以及最近的图形卷积概念，旨在学习捕获问题特定交互的图像表示。我们使用所建议的图形学习者模块增强的简单基线体系结构测试我们在 v2 v2 数据集上的方法。我们以 66.18% 的精度获得了很有希望的结果，并证明了该方法的可解释性。代码可在 github-comnemainmain/vqa 项目中找到。少

2018 年 11 月 1 日提交;v1 于 2018 年 6 月 19 日提交;最初宣布 2018 年 6 月。

评论:2018 年 NIPS (13 页, 7 位数字)

244. 第 1806. 07139[pdf,其他] Cs。CI

在优化 nlp 模型时，利用 j-k 折叠交叉验证来减小方差

作者:henry b. moss, david s. leslie , paul rayson

摘要: k-折叠交叉验证 (cv) 是一种常用的估计机器学习模型真实性能的方法，允许模型选择和参数调整。然而, cv 的**过程**本身就需要随机划分数据，因此我们的性能估计实际上是随机的，其可变性对于**自然语言处理任务来说**可能是巨大的。我们证明了这些不稳定的估计不能用来进行有效的参数调整。由此产生的调整参数对数据的分区方式高度敏感，这意味着我们通常会选择次优参数选择，并存在严重的重现性问题。相反，我们建议使用较少变量 j-k-折叠 cv，其中 j 独立 k 折叠交叉验证用于评估性能。我们的主要贡献是将 j-k-折叠 cv 从性能估计扩展到参数调整，并研究如何选择 j 和 k。我们认为，可变性比偏见更重要的有效调谐，所以主张更低的选择 k 比通常在 nlp 文献中看到，而不是使用保存的计算来增加 j。为了证明我们的建议的普遍性，我们调查了广泛的案例研究：情绪分类 (一般和特定目标)、词性标记和文档分类。少

2018 年 6 月 19 日提交;最初宣布 2018 年 6 月。

评论:coling 2018。代码可在 https://github.com/henrymoss/COLING2018

245. 第 xiv:1806. 06874[pdf] Cs。Cl

将词特征向量法与卷积神经网络相结合，在口语理解中进行插槽填充

作者:林瑞熙

摘要: 插槽填充是口语理解 (slu) 和 **自然语言处理**(nlp) 中的一个重要问题，它涉及识别用户的意图并为英语中的每个单词分配语义概念。句子。本文提出了一种词特征向量法，并将其结合到卷积神经网络 (cnn) 中。我们考虑 18 个单词的特征，每个单词的特征都是通过合并类似的单词标签来构造的。通过引入外部库的概念，提出了一种有利于从训练数据集中建立一个词与特征之间关系的特征集方法。利用 atis 数据集对计算结果进行了报告，并与传统的 cnn 和双向顺序 cnn 进行了比较。少

2018 年 6 月 18 日提交;最初宣布 2018 年 6 月。

246. 第 xiv:1806. 076571[pdf, ps,其他] Cs。Cl

子格: 使用子字符串扩展 skip-克单词表示

作者:tom kocmi, ond 热埃杰·博贾尔

摘要: 略读图 (word2vec) 是一种利用神经网络创建单词向量表示 ("分布式词表示") 的最新方法。这种表现在**自然语言处理**的各个领域都很受欢迎，因为它似乎在没有任何明确监督的情况下捕捉了关于单词的句法和语义信息。我们建议 subgram，一个改进的 skipgram 模型，也考虑在训练**过程**中的单词结构，实现了大量的收益，在 skipgram 原来的测试集。少

2018 年 6 月 18 日提交;最初宣布 2018 年 6 月。

评论:发布于 tsd 2016

日记本参考:文本、演讲和对话：第 19 届国际会议, tsd 2016

247. 第 xiv:1806 06301[pdf, ps,其他] Cs。Cl

从野生数据中嵌入：测量、理解和删除

作者:adam sutton, thomas lansdall-灵 spyalyalyalyalyalyalyalyalds , nello Cristianini

摘要: 许多现代人工智能 (ai) 系统利用数据嵌入，特别是在**自然语言处理**(nlp) 领域。这些嵌入是从 "从野外" 收集到的数据中收集的，被发现含有不必要的偏见。本文为衡量、理解和消除这一问题做出了三项贡献。我们提出了一个严格的方法来衡量其中的一些偏见，基于使用为社会心理学应用创建的单词列表;我们观察到职业中的性别偏见如何反映现实世界中同一职业中的实际性别偏见;最后，我们演示了一个简单的投影如何可以显著减少嵌入偏置的影响。所有这些都是了解如何将信任构建到 ai 系统中的持续努力的一部分。少

2018 年 6 月 16 日提交;最初宣布 2018 年 6 月。

评论:作者的原始版本

248. 第 1806.06183[pdf,其他] Cs。简历

神经绘制器: 多轮图像生成

作者:ryan y. benmalek, claire cardie , serge belongie, xiadong he, jan feng gao

摘要: 在这项工作中，我们结合了两个研究线程从 vision/ima 素和**自然语言处理**，以制定图像生成任务的条件下，在多圈设置。通过多圈，我们指的是图像是在用户指定的条件信息的一系列步骤中生成的。我们提出的方法非常有用，并提供了对神经可解释性的见解。我们介绍了一个框架，其中包括一个新的训练算法，以及为多轮设置构建的模型改进。我们证明了这个框架生成一系列与给定的条件信息相匹配的图像，这项任务对于更详细地测试和分析条件图像生成方法很有用。少

2018 年 6 月 16 日提交;最初宣布 2018 年 6 月。

249. 第 xiv:1806. 05559[pdf,其他] Cs。CI

利用双向递归神经网络提取平行句提高机器翻译水平

作者:francis grégoire,菲利普·朗莱斯

摘要: 并行句子提取是解决多语言**自然语言处理**应用程序中发现的数据稀疏问题的一项任务。我们提出了一种基于双向递归神经网络的方法，从多语言文本的集合中提取平行句子。我们对嘈杂的并行语料库的实验表明，我们可以通过消除对特定功能工程或额外外部资源的需求，在竞争基线的情况下获得有希望的结果。为了证明我们的方法的效用，我们从维基百科文章中提取句子对，以训练机器翻译系统，并显示翻译性能的显著改进。少

2018 年 8 月 24 日提交;v1 于 2018 年 6 月 13 日提交;最初宣布 2018 年 6 月。

评论:12 页, 7 个数字, coling 2018. arxiv 管理说明: 文本与 arxiv:1709.09783 重叠

250. xiv:1806.05480[pdf,其他] Cs。CI

多伊 10.1109/SYNASC.2015.45

使用停止词和对话语的浪漫语言语言自动识别

作者:Ciprian-Octavian truicáu, jien velcin, 亚历山德鲁·博伊切 a

文摘: 自动**语言**识别是一个**自然语言处理**问题，试图确定给定内容的**自然语言**。本文提出了一种利用含有停止词和变音词的词典**自动识别**书面文本的统计方法。我们提出了不同的方法，结合这两个字典，以准确地确定文本语料库的**语言**。之所以选择这种方法，是因为停止词和变音词是一种**语言**非常具体的，尽管有些**语言**有一些相似的单词和特殊字符，但它们并不都是常见的。考虑到的**语言**是浪漫**语言**，因为它们非常相似，通常很难从计算的角度区分它们。我们使用推特语料库和新闻文章语料库测试了我们的方法。这两个语料库都由 utf-8 编码的文本组成，因此可以考虑变音，在文本没有变音的

情况下，只有停止词用于确定文本的**语言**。实验结果表明，该方法对小文本的准确率在 90% 以上，对更低的文本的准确率在 99.8% 以上。

2018 年 6 月 14 日提交;最初宣布 2018 年 6 月。

251. 第 1806. 05432[pdf] Cs。Cl

使用条件随机字段 (crf) 的乌尔都语分割

作者:haris bin zia, agha ali raza, awais athar

摘要: 最先进 **的自然语言处理**算法在很大程度上依赖于高效的分词。乌尔都语是其中的**一种语言**，其中的分词是一项复杂的任务，因为它表现出空间遗漏以及空间插入问题。这在一定程度上是由于阿拉伯语脚本虽然在**性质**上是草书，但由具有固有的联接和非联接属性的字符组成，而不考虑单词边界。本文提出了一种利用具有正交、语言和形态学特征的条件随机场序列建模器的乌尔都语分词系统。我们提出的模型自动学习将空白作为单词边界以及零宽度非 joiner (zwnj) 作为子词边界进行预测。利用人工注释语料库，我们的模型实现了单词边界识别的 f1 分数为 0.97，子词边界识别任务的 f1 分数为 0.97。我们已经公开了我们的代码和语料库，以使我们的结果可重现。少

2018 年 6 月 14 日提交;最初宣布 2018 年 6 月。

评论:8 页，涂装 2018

252. 第 xiv: 1806. 05219[pdf,其他] Cs。Cl

将复制与复制结合在一起: 用于目标依赖情感分析的三项复制研究

作者:andrew moore, paul rayson

摘要: 缺乏可重复性和通用性是**自然语言处理**中科学持续发展的两个重要威胁。**语言**模型和学习方法非常复杂，科学会议文件不再包含足够的空间，无法提供复制或复制所需的技术深度。以目标依赖情绪分析为案例研究，我们展示了最近在该领域的工作如何没有一致地发布代码，或描述了足够详细的学习方法设置，缺乏可比性和通用性的训练，测试或验证数据。为了研究通用性和使最先进的比较评价，我们对三组互补方法进行了首次复制研究，并对六个不同的英语数据集进行了首次大规模的大规模评价。考虑到我们的经验，我们建议未来的复制或复制实验应始终考虑各种数据集，同时记录和发布其方法和已发布的代码，以最大限度地减少对这两个数据集的障碍。可重复性和通用性。我们已经发布了我们的代码与一个模型动物园在 github 与木星笔记本，以帮助理解和完整的文件，我们建议其他人做同样的与他们的论文在提交时通过匿名 github 帐户。少

2018 年 8 月 6 日提交;v1 于 2018 年 6 月 13 日提交;最初宣布 2018 年 6 月。

评论:coling 2018。代码可在 https://github.com/apmoore1/Bella

253. 特别报告: 1806. 05009[pdf, ps,其他] Cs。Lg

通过自适应符号嵌入进行树编辑距离学习

作者:benjamin pašen, claudio gallicchio , alessio micmicri , barbara hammer

摘要: 公制学习的目的是通过学习距离测量来提高分类的准确性，这种距离测量使同一类的数据点更加紧密地结合在一起，并将不同类的数据点进一步分开。最近的研究表明，通过学习的成本函数，度量学习方法也可以应用于树，如分子结构、计算机程序的抽象语法树或**自然语言**的句法树。编辑距离，即替换、删除或插入树中节点的成本。然而，直接学习这样的成本可能会产生一个违反度量公理的编辑距离，难以解释，并且可能无法很好地概括。在这个贡献中，我们提出了一种新的树度量学习方法，我们称之

为嵌入编辑远程学习 (bedl)，它通过嵌入树节点作为向量间接地学习编辑距离，从而使这些树之间的欧几里得距离向量支持类区分。我们通过减少来自同一类的原型树的距离和增加来自不同类的原型树的距离来学习这种嵌入。在我们的实验中，我们发现 bedl 改进了树木公制学习的最先进技术，从计算机科学到生物医学数据，再到**自然**语言 **处理**数据集包含超过 300，000 个节点。少

2018 年 7 月 16 日提交;v1 于 2018 年 6 月 13 日提交;最初宣布 2018 年 6 月。

评论:在瑞典斯德哥尔摩举行的国际机器学习会议 (2018 年) 上发表论文

日记本参考:机器学习研究论文集 80 (2018) 3973-3982

254. **第 xiv:1806.04822[pdf,其他] Cs。CI**

sgm：多标签分类的序列生成模型

作者:杨彭成,孙旭,李伟，马树明,吴伟,王厚峰

摘要: 多标签分类是**自然语言处理**中一项重要而又具有挑战性的任务。它比单一标签分类更为复杂，因为标签往往是相关的。现有方法倾向于忽略标签之间的相关性。此外，文本的不同部分对于预测不同的标签可能有不同的贡献，而现有模型没有考虑到这一点。本文提出将多标签分类任务视为序列生成问题，并应用一种具有新解码器结构的序列生成模型来求解。大量的实验结果表明，我们提出的方法比以前的工作有相当大的优势。对实验结果的进一步分析表明，该方法不仅能捕捉标签之间的相关性，而且在预测不同标签时自动选择信息最丰富的词语。少

2018 年 6 月 15 日提交;v1 于 2018 年 6 月 12 日提交;最初宣布 2018 年 6 月。

评论:被涂装 2018 年接受

255. **特别报告：1806.04820[pdf] Cs。CI**

基于 ehr 的计算表型的自然语言处理

作者:曾泽贤,邓宇，李晓宇,特里斯坦·娜诺曼，袁罗

文摘: 本文综述了**将自然语言处理**(nlp) 应用于电子健康记录 (ehr) 进行计算表型的最新进展。基于 nlp 的计算表型有许多应用，包括诊断分类、新表型发现、临床试验筛选、药物基因组学、药物相互作用 (ddi) 和不良药物事件 (ade) 检测，以及全基因组和全表关联研究。计算表型的算法开发和资源建设取得了重大进展。在被调查的方法中，设计良好的关键字搜索和基于规则的系统往往取得良好的性能。但是，关键字和规则列表的构造需要大量的手动工作，这很难扩展。监督机器学习模型受到青睐，因为它们能够从数据中获取分类模式和结构。近年来，深度学习和无监督学习越来越受到人们的关注，前者的表现受到青睐，后者因其寻找新表型的能力而受到青睐。集成异构数据源变得越来越重要，并在提高模型性能方面显示出希望。通常，通过将多种信息模式结合起来，就能取得更好的业绩。尽管取得了许多进展，但基于 nlp 的计算表型仍然存在挑战和机遇，包括更好的模型可解释性和泛化性，以及临床叙事中特征关系的适当表征

2018 年 6 月 14 日提交;v1 于 2018 年 6 月 12 日提交;最初宣布 2018 年 6 月。

256. **第 xiv:1806.04284[pdf,其他] Cs。CI**

释义：通过图像提取粘度粉碎释义

作者:朱晨辉,马尤大谷,中岛玉田

摘要: 释义是对换句话说的文本含义的重述。为了提高许多**自然语言处理**任务的性能，对释义进行了研究。在本文中，我们提出了一个新的任务 i 描述，以提取视觉基础释义 (vgp)，这是不同的短语表达式描述相同的视觉概念在图像中。这些提取的 vgp 有可能

改进**语言**和图像多模态任务，如视觉问题回答和图像字幕。如何模拟 vgp 之间的相似性是解释的关键。我们应用了现有的各种方法，提出了一种新的基于神经网络的图像注意方法，并报告了首次尝试对 i 自述的结果。少

2018 年 6 月 11 日提交;最初宣布 2018 年 6 月。

评论:coling 2018

257. 第 xiv:1806.04189[pdf,其他] Cs。CL

使用图形表示导航，以实现神经语言模型的快速、可扩展解码

作者:张敏佳,刘晓东,王文汉,高建峰,何玉雄

摘要: 神经**语言**模型 (nlm) 最近通过在许多**自然语言处理**(nlp) 任务中实现最先进的性能而获得了新的兴趣。然而，nlm 在计算上要求很高，这主要是由于软最大层在大量词汇表上的计算成本。我们观察到，在许多 nlp 任务的解码中，只需要精确计算前 k 假设的概率，而 k 通常比词汇量小很多。本文提出了一种新的软最大层逼近算法，称为快速图解码器 (fgd)，该算法在给定的上下文中快速识别一组最有可能根据 nlm 出现的 k 词。我们证明，fgd 将解码时间缩短了一个数量级，同时在神经机器翻译和**语言**建模任务上达到了接近完全软最大基线精度的水平。并证明了软最大近似质量的理论保证。少

2018 年 6 月 11 日提交;最初宣布 2018 年 6 月。

258. 第 xiv:1806.03688[pdf,其他] Cs。CL

lexnlp：法律和法规文本的自然语言处理和信息提取

作者:michael j bommarito ii, daniel martin katz, eric m detterman

摘要: lexnlp 是一个开源 python 包，专注于**自然语言处理**和机器学习，用于法律和监管文本。该软件包包括以下功能: (一) 分部文件，(二) 确定标题和章节标题等关键文本，(三) 提取超过 18 种类型的结构化信息，如距离和日期; (四) 提取指定实体，如公司和地缘政治实体，(v) 将文本转换为模型培训的功能，(vi) 建立无监督和监督的模型，如单词嵌入或标记模型。lexnlp 包括基于从 sec edgar 数据库提供的真实文档以及各种司法和监管程序中提取的数千个单元测试的预先培训模型。lexnlp 设计用于学术研究和工业应用，并以 https://github.com/LexPredict/lexpredict-lexnlp 分发。少

2018 年 6 月 10 日提交;最初宣布 2018 年 6 月。

评论:9 页，0 个数字; 另见 https://github.com/LexPredict/lexpredict-lexnlp

类:I.2.7;F.2.2;H.3.1;H.3.3;i。7

259. 第 xiv:1806.03537[pdf, ps,其他] Cs。CL

义和诗词嵌入与语义转换：一项调查

作者:and 类 kutuzov, lilja Øvrelid, terrence szymanski, erik velldal

摘要: 近年来，旨在使用分布方法，特别是基于预测的词嵌入模型追踪词汇语义中的时间变化的出版物激增。然而，这种研究缺乏更成熟的**自然语言处理**领域的凝聚力、共同术语和共同做法。本文综述了历时词嵌入和语义移位检测相关的学术研究现状。我们首先讨论语义转换的概念，然后继续概述使用单词嵌入模型跟踪此类时间转换的现有方法。我们提出了可以比较这些方法的几个轴，并概述了 nlp 这个新出现的子领域面临的主要挑战，以及前景和可能的应用。少

2018 年 6 月 13 日提交;v1 于 2018 年 6 月 9 日提交;最初宣布 2018 年 6 月。

评论:coling 2018 会议记录

260. 第 xiv:1806.03379[pdf,其他] Cs。简历

cs-vqa: 用压缩感图像回答视觉问题

作者:li-chihuang , kuldeep kulkarni, anikjha , suhas lohit, suren jayasuriya, pavan turaga

摘要: 视觉问题回答 (vqa) 是一项复杂的语义任务，需要**自然语言处理**和视觉识别。在本文中，我们探讨了当图像在亚尼奎斯特压缩范式中捕获时，vqa 是否可以解决。我们开发了一系列深网络架构，利用可用的压缩数据提高精度，并表明 vqa 在压缩域中确实是可以解决的。我们的结果表明，在使用压缩测量时，vqa 性能会显著下降，但当 vqa 管道与最先进的深部神经网络结合使用时，可以恢复其准确性，从而实现 cs 重建。研究结果对资源受限的 vqa 应用产生了重要影响。少

2018 年 6 月 8 日提交;最初宣布 2018 年 6 月。

评论:5 页, 2 个数字，被接受到 icip 2018

msc 类: 68

261. xiv:1806.03255[pdf,其他] cs. cy

自动生成中国的大型区域性块列表

作者:austin hounsel, prateek mittal, nick feamster

摘要: 互联网审查的措施依赖于要测试的网站列表，或由第三方管理的 "阻止列表"。不幸的是，其中许多列表并不公开，而那些倾向于关注一小部分主题的列表，使其他类型的网站和服务得不到测试。为了增加现有网站列表上的网站集并使其多样化，我们使用**自然语言处理**和搜索引擎自动发现在中国被审查的网站范围更广。使用这些技术，我们创建了一个 1125 网站的列表，其中包括亚历克莎 1000 强之外的中国政治、少数族裔人权组织、受压迫宗教等。重要我们发现的网站都不存在于当前最大的阻止列表中.我们开发的列表不仅极大地扩展了当前互联网测量工具可以测试的网站集，而且加深了我们对中国审查内容性质的理解 。我们已经发布了这个新的块列表和生成它的代码。少

2018 年 7 月 19 日提交;v1 于 2018 年 6 月 4 日提交;最初宣布 2018 年 6 月。

262. 第 1806.03144[pdf, ps,其他] Cs。红外

科技论文研究领域的自动识别

作者:eric kergosien, amin farvardin, Chaudiron teisseire, marie-nole bessagnet, joachim schöpfel, stphane chaudiron, bernard jjemin , annig le parc-lacayle , mathieu roche, christian sallaberry , jean-菲利普·tonneau

摘要: terre-istex 项目旨在根据科学论文中提供的异质数字内容，确定涉及特定地理区域的科学研究。该项目分为三个主要工作包: (1) 确定实证研究的时期和地点，并反映所分析的文本样本所产生的出版物; (2) 确定这些文件中出现的主题,(3) 开发基于网络的地理信息检索工具。前两个操作将**自然语言处理**模式与文本挖掘方法结合起来。将空间、专题和时间层面纳入全球关系有助于更好地了解开展了何种研究、专题及其地理和历史覆盖面。terre-istex 项目的另一个独创性是语料库的异质性，包括 istex 数字图书馆和 cirad 研究中心的博士论文和科学文章。少

2018 年 6 月 8 日提交;最初宣布 2018 年 6 月。

日记本参考:第十一届语文资源与评价国际会议会议记录，第 1902 至 1907 页, 2018 年, http://lrec2018.lrec-conf.org

263. 第 xiv:1806. 02814[pdf,其他] Cs。Cl

低资源医学命名实体识别的嵌入转移--以患者移动性为例

作者:denis newman-griffis, ayah zirikly

文摘: 功能作为全球健康的一个重要指标正在得到人们的认可，但在医学**自然语言处理**研究中仍未得到充分研究。我们首先分析了使用最近开发的自由文本电子健康记录数据集自动提取患者移动性描述的方法。我们将任务定义为命名实体识别 (ner) 问题，并研究 ner 技术在移动性提取中的适用性。由于专注于患者功能的文本语料库很少，我们探索了在复发神经网络 ner 系统中使用的词嵌入的域适应。我们发现，在一个小的领域内语料库上训练的嵌入的性能几乎和从大型领域外语料库中学到的效果一样好，而领域适应技术在精度和召回方面都有额外的改进。我们的分析确定了在提取患者移动性描述时的几个重大挑战，包括注释实体的长度和复杂性以及移动性描述中的高语言变异性。少

2018 年 6 月 7 日提交;最初宣布 2018 年 6 月。

评论:接受 2018 年 binlp. 11 页

264. xiv:1806. 02724[pdf,其他] Cs。简历

用于视觉和语言导航的语音折叠式

作者:daniel fred, ronghang hu, volan cirik, anna rohrbach, jacob Cirik, louis-phipimornyas , taylor berg-kirkpatrick, kate saenko, dan k 贩运,trevor darrell

摘要: **自然语言**指令指导下的导航为指令追随者带来了一个具有挑战性的推理问题。**天然**的**语言**指令通常只识别几个高级决策和地标，而不是完整的低级运动行为;许多缺失的信息必须根据感知语境来推断。在机器学习设置中，这具有双重挑战性: 很难收集足够的带注释的数据，以便能够从零开始学习**此推理过程**，也很难使用通用方法实现推理**过程**序列模型。在这里，我们描述了一种视觉和**语言**导航的方法，该方法使用嵌入式扬声器模型解决这两个问题。我们使用这个扬声器模型来 (1) 合成新的数据扩充指令，(2) 实现语用推理，评估候选操作序列解释指令的程度。这两个步骤都由反映人为指令粒度的全景动作空间支持。实验表明，这种方法的所有三个组成部分--扬声器驱动的数据增强、语用推理和全景行动空间--极大地提高了基线指令追随者的性能，成功率提高了一倍多在标准基准上的现有最佳方法。少

2018 年 10 月 26 日提交;v1 于 2018 年 6 月 7 日提交;最初宣布 2018 年 6 月。

评论:NIPS

265. 第 xiv:1806. 02366[pdf] cs et

lstm 体系结构中 CMOS-memristor 电路的设计

作者:kamilya smagulova, kazybek adam, olga krestinskaya, alex pappachen james

摘要: 长期短期存储器 (lstm) 体系结构是一种众所周知的方法，用于构建**可用于自然语言处理**中的数据顺序**处理**的递归神经网络 (rnn)。由于并行性和复杂性大, lstm 的近传感器硬件实现面临挑战。我们提出了一个 0.18 m cmos, gst 记忆电阻 lstm 硬件架构近传感器**处理**。在一个基于 keras 模型的预测问题中，验证了该系统的有效性。少

2018 年 6 月 6 日提交;最初宣布 2018 年 6 月。

日记本参考:ieee 关于电子器件和固态电路的国际会议, 2018

266. 第 xiv:1806-04[pdf,其他] Cs。简历

挖掘意义：从愿景到语言，通过多个网络共识

作者:iulia duta, andari liviu nicoicioiu, simion-vlad bogolin, marius leordeanu

摘要: 在计算机视觉、**自然语言处理**和机器学习的交汇点上，将视觉数据描述为自然 **语言**是一项非常具有挑战性的任务。**语言**远远超出了对物理对象及其相互作用的描述，可以通过多种方式传达相同的抽象概念。它既是关于最高语义级别的内容，也是关于流畅的形式。在这里，我们提出了一种方法来描述视频的自然**语言**，通过达成共识之间的多个编码器解码器网络。找到这样一个共同的语言描述，与一个更大的群体共享共同的属性，有更好的机会传达正确的含义。我们提出并培训了几种网络架构，并使用不同类型的图像、音频和视频功能。每个模型都会生成自己对输入视频的描述，最好的描述是通过一个高效的两阶段共识**过程**来选择的。我们通过在具有挑战性的 msr-vtt 数据集上获得最先进的结果来展示我们方法的强度。少

2018 年 9 月 18 日提交;v1 于 2018 年 6 月 5 日提交;最初宣布 2018 年 6 月。

评论:2018 年接受 bmvc

267. 第 xiv:1806. 01873[pdf,其他] Cs。简历

焦点可视文本注意视觉问题回答

作者:梁俊伟,蒋鲁江,曹良良, 李丽佳,亚历山大·豪普特曼

摘要: 利用神经网络对**语言**和视觉的最新见解已成功地应用于简单的单图像视觉问题回答。然而，为了解决个人照片等多媒体收藏的现实生活中的问答问题，我们必须查看带有照片或视频序列的整个收藏。在回答大量集合中的问题时,**一个自然**的问题是确定支持答案的片段。本文描述了一种新的神经网络--焦点视觉-文本注意网络 (fvta)，用于视觉问题回答中的集体推理，其中提供了图像和文本元数据等视觉序列信息和文本序列信息。fvta 引入了一种端到端方法，该方法利用分层**过程**动态确定哪些媒体以及在顺序数据中关注哪些时间来回答问题。fvta 不仅能很好地回答问题，还为系统结果提供了得到答案的依据。fvta 在 memexqa 数据集上实现了最先进的性能，并在电影 qa 数据集上获得了竞争结果。少

2018 年 6 月 5 日提交;最初宣布 2018 年 6 月。

268. 第 xiv:1806. 01515[pdf,其他] Cs。Cl

源端单语嵌入如何影响神经机器翻译?

作者:丁淑阳,杜建华

摘要: 在许多**自然语言处理**(nlp) 任务中，使用预先训练的单词嵌入作为输入层是一种常见的做法，但在神经机器翻译 (nmt) 中却在很大程度上被忽略。本文对 nmt 中使用预训练源侧单语嵌入词的效果进行了系统分析。我们比较了几种策略，例如在不同数量的数据的 nmt 培训中修复或更新嵌入，我们还提出了一种新的策略，称为双嵌入，将固定和更新策略混合在一起。我们的研究结果表明，如果适当地将预培训的嵌入纳入 nmt，特别是在并行数据有限或有其他域内单语言数据随时可用的情况下，这种嵌入可能会很有帮助。少

2018 年 6 月 14 日提交;v1 于 2018 年 6 月 5 日提交;最初宣布 2018 年 6 月。

评论:10 页, 4 个数字

269. 第 xiv:1906.00802[pdf,其他] 反渗透委员会

maestrob: 用于低层控制和高级推理的集成编排的机器人框架

作者:asim munawar, giovanni de r 秋护酒店, Tu-Hoapham, daiki kim ura, michiaki tatsubori, takao moriyama, ryuki tachibana, grady booch

文摘: 本文介绍了一个叫做 maestrob 的框架。它的设计是通过**自然语言**或演示给出的简单的高级指令，使机器人能够高精度地完成复杂的任务。为了实现这一点，它通过使用存储在本体形式和规则中的知识来处理层次结构，以便在不同级别的指令之间进行桥接。因此，该框架有多层**处理**组件;低级别的感知和驱动控制，用于认知能力和语义理解的符号规划器和 watson api，以及由一个名为 "intu 项目" 的新开源机器人中间件对这些组件进行编排。我们展示了如何在多个参与者 (人、通信机器人和工业机器人) 协作执行共同工业任务的复杂场景中使用此框架。人类用**自然语言**对话和演示向 pepper (软银机器人的人形机器人) 教授组装任务。我们的框架帮助 pepper 感知人类演示，并为 ur5 (来自通用机器人的协作机器人手臂) 生成一系列动作，最终执行组装 (例如插入) 任务。少

2018 年 6 月 3 日提交;最初宣布 2018 年 6 月。

评论:ieee 机器人与自动化国际会议 (icra) 2018。视频: https://www.youtube.com/watch? v=19JsdZi0TWU

270. 第 xiv:866.00793[pdf,其他] Cs。CI

基于域名特定知识库的转移主题标签: 对英国下议院 1935-2014 演讲的分析

作者:亚历山大·赫尔佐格,彼得·约翰 , 斯拉夫·扬金·米哈伊洛夫

文摘: 主题模型被广泛用于**自然语言处理**，使研究人员能够估计文档集合中的潜在主题。大多数主题模型使用无人监督的方法，因此需要额外的步骤，将有意义的标签附加到估计的主题。这种手动标记过程是不可扩展的，并且存在人为的偏差。我们提出了一种半自动转移主题标记方法，旨在解决这些问题。特定于域的代码库构成了自动主题标记的知识库。我们通过对 195-2014 年英国下议院完整语料句的动态主题模型分析来演示我们的方法，使用比较议程项目的编码说明来标记主题。我们表明，我们的方法很好地适用于我们估计的大多数主题;但我们也发现，具体机构的议题，特别是关于国家以下各级治理的专题，需要人工投入。我们使用人工专家编码验证我们的结果。少

2018 年 8 月 27 日提交;v1 于 2018 年 6 月 3 日提交;最初宣布 2018 年 6 月。

271. 第 xiv:066.00780[pdf,其他] Cs。CI

为面向目标的对话系统构建高级对话管理器

作者:弗拉基米尔·伊利耶夫斯基

摘要: 面向目标 (go) 对话系统，俗称目标导向聊天机器人，可帮助用户在封闭的域内实现预定义的目标 (例如预订电影票)。第一步是通过使用**自然语言**理解技术来理解用户的目标。一旦知道了这个目标，机器人就必须管理对话来实现这个目标，这个目标是根据学习到的政策进行的。对话系统的成功取决于政策的质量，而这又取决于政策学习方法是否有高质量的培训数据，例如深度强化学习。由于领域的特殊性，可用数据的数量通常太低，无法培训良好的对话政策。在本硕士论文中，我们引入了一种转移学习方法来缓解域内数据可用性低的影响。我们基于转移学习的方法提高了机器人的成功率:20%相对于遥远的领域，我们把它增加一倍以上，对于接近的领域，相比模型没有转移学习。此外，转学学习聊天机器人学习政策的速度高达 5 到 10 倍。最后，由于转移学习方法是对其他**处理**(如预热启动) 的补充，我们表明它们的联合应用能产生最佳的结果。少

2018 年 6 月 3 日提交;最初宣布 2018 年 6 月。

评论:硕士论文

272. 第 xiv:006. 727[pdf,其他] 反渗透委员会

用于协作人自主目标搜索的闭环贝叶斯语义数据融合

作者:luke burks, ian loefgren, luke barbier, jeremy muesing, jamison mckinley, sousheel vunnam , nisar ahmed

文摘: 在搜索应用中，自主无人飞行器必须能够有效地重新获取移动目标并将其本地化，这些目标可以在大空间中长时间不在视野中。因此，必须积极利用所有可用的信息来源----包括人类提供的不准确但随时可用的语义观察。为此，本工作开发并验证了一种用于动态目标搜索的新型协作人机传感解决方案。我们的方法使用连续的部分可观察**马尔可夫决策过程**(cpomdp) 规划来生成车辆轨迹，以最佳方式利用机载传感器中不完美的检测数据以及语义**自然** 可以从人类传感器中特别要求的语言观察。关键的创新是一个可扩展的分层高斯混合模型公式，以有效地解决 comdp 与语义观测在连续动态空间。该方法通过一个真正的人机团队在自定义测试台上进行动态室内目标搜索和捕获场景的演示和验证。少

2018 年 6 月 2 日提交;最初宣布 2018 年 6 月。

评论:最终版本被接受并提交给 2018 年融合大会 (2018 年 7 月，英国剑桥)

273. 第 xiv:606.00696[pdf] cse

nlp 辅助软件测试: 系统审查

作者:vahid garousi, sara bauer, michael felderer

摘要: 语境: 为了减少从**自然语言**需求中提取测试用例的人工工作，文献中提出了许多基于**自然语言处理**(nlp) 的方法。鉴于这一领域的方法很多，而且许多从业人员渴望利用这些技术，因此必须综合并概述这一领域的最新情况。目的: 总结 nlp 辅助软件测试的最新技术，这将有利于从业者潜在地利用这些基于 nlp 的技术，使研究人员能够提供研究场景的概述。方法: 针对上述需求，采用系统文献映射 (分类) 和系统文献综述的形式进行了调查。在汇编了 57 篇论文的初步库之后，我们进行了系统投票，最后的一批包括 50 篇技术论文。结果: 本文综述了论文中的贡献类型、用于帮助软件测试的 nlp 方法类型、所需输入要求的类型以及该领域的工具支持的回顾。我们的研究结果如下: (1) 在论文中提供的 28 个工具中，只有 2 个 (7%) 可供下载;(2) 较大比例的论文 (23 份 (50 份) 提供了对 nlp 方面的浅度接触 (几乎没有细节)。结论: 我们相信，本文作为这一领域知识体系的 "指数"，对从业者和研究人员都有好处。这些结果可以帮助从业人员，使他们能够利用任何现有的基于 nlp 的技术，降低测试用例设计的成本，减少用于测试活动的人力资源数量。初步的见解，在与我们的一些工业合作者分享了这一审查之后，表明这次审查确实对从业人员有用和有益。少

2018 年 6 月 2 日提交;最初宣布 2018 年 6 月。

评论:软件测试;自然语言处理 (nlp);系统文献映射;系统的文献综述。arxiv 管理说明: 文本与 arxiv:1801.0. 02201 重叠

274. 第 xiv:866.00186[pdf,其他] Cs。简历

视频描述: 方法、数据集和评估指标的调查

作者:nayyer aafaq, syed zulqarnain gilani, wei liu, ajmal mian

摘要: 自动视频描述可用于帮助视障人士、人机交互、机器人技术和视频索引。在过去的几年里，由于计算机视觉和**自然语言处理**方面的深度学习取得了前所未有的成功，这一领域的研究兴趣大增。文献中提出了许多方法、数据集和评价措施，呼吁进行全面调查，以便更好地将研究工作的重点放在这一蓬勃发展的方向上。本文通过对包括深度学习模式在内的最先进方法的调查，准确地满足了这一需求;比较基准数据集的域、类的数量和存储库大小;并确定各种评价指标的利弊，如 BLEU、rouge、meteor、cider、spice 和大规模毁灭性武器。我们的调查显示，视频描述研究要与人类的表现相匹配，还有很长的路要走，造成这种不足的主要原因有两个。首先，现有数据集不能充分代表开放域视频和复杂语言结构的多样性。其次，目前的评价措施与人的判断不一致。例如，同一个视频可以有非常不同但正确的描述。我们的结论是，在规模、多样性和注释准确性方面，需要改进评价措施和数据集，因为它们直接影响到更好的视频描述模型的开发。从算法的角度来看，对描述质量的诊断是很有挑战性的，因为**与语言**模型自然产生的偏差相比，很难评估视觉特征的贡献程度采用。少

2018 年 6 月 1 日提交;最初宣布 2018 年 6 月。

评论:提交到期刊

275. 第 1805 5.12291 [pdf,其他] Cs。CI

印尼会话文本中基于特征的神经命名实体识别模型的实证评价

作者:kemal k 响文, samuel louvan

摘要: 尽管**在自然语言处理**社区中命名实体识别 (ner) 任务的历史很长，但以前的工作很少研究会话文本的任务。这样的文本是具有挑战性的，因为它们包含了大量的单词变体，增加了词汇外 (oov) 单词的数量。大项的 oov 词数量之多给基于源的神经模型带来了困难。同时，基于字符的神经模型在缓解此 oov 问题方面的有效性也有足够的证据。我们报告了一个经验性的评估神经序列标记模型与字符嵌入，以解决 ner 任务在印尼语对话文本。我们的实验表明, (1) 字符模型的性能优于文字嵌入模型，最多 4F_1 点, (2) 字符模型在 oov 案例中表现更好，提高高达 15F_1 点，并且 (3) 字符模型对非常高的 oov 速率是鲁棒性的。少

2018 年 9 月 19 日提交;v1 于 2018 年 5 月 30 日提交;最初宣布 2018 年 5 月。

评论:参加 emnlp 2018 用户生成的文本 (w-nut) 研讨会

276. 第 1805 5.1124 [pdf,其他] Cs。CI

回指和核心解析: 回顾

作者:rhea Ramkumar, soujanya poria , erik cambria , ramkumar thirunavukarasu

摘要: 实体解析旨在解析文档中对实体的重复引用，并构成**自然语言处理**(nlp) 研究的核心组成部分。该领域在提高机器翻译、情感分析、释义检测、摘要等其他 nlp 领域的性能方面具有巨大的潜力。nlp 中的实体分辨率领域出现了两个单独的子领域的研究激增，即回指决议和相关性解决。通过这篇文章，我们的目标是澄清实体解决中这两个任务的范围。我们还对解决这一 nlp 问题所采用的数据集、评估指标和研究方法进行了详细分析。这项调查的目的是让读者清楚地了解什么是这一 nlp 问题以及需要注意的问题。少

2018 年 5 月 30 日提交;最初宣布 2018 年 5 月。

277. 第 1805 5.1118 [pdf,其他] Cs。CI

视觉参照表达式识别: 系统实际上学到了什么?

作者:volkan cirik, louis-phili-pholan mornyan, taylor berg-kirkpatrick

摘要: 我们对引用表达式识别的最新系统进行了实证分析----在**自然语言**表达所指的图像中识别对象的任务----目的是深入了解如何识别这些系统对**语言**和视觉的推理。令人惊讶的是，我们发现有力的证据表明，即使是复杂的、有语言动机的这项任务模型也可能忽视语言结构，而是依靠数据选择和注释中意外偏见所带来的浅薄相关性**程**。例如，

我们展示了一个系统在输入图像上进行训练和测试没有输入引用表达式在前 2 名

预测中，精度可达到 71.2。此外，只预测给定输入的对象类别的系统可以在前 2 名预测中达到 84.2 的精度。这些令人惊讶的积极结果对于应该是不足的预测情景表明，仔细分析我们的模型正在学习什么----以及进一步分析我们的数据是如何构建的----对于我们寻求在基础**上取得实质性进展至关重要语言**任务。少

2018 年 5 月 30 日提交;最初宣布 2018 年 5 月。

评论:naacl2018 短

278. 第 1805 5.11653[pdf,其他] Cs。CI

lstm 利用数据的语言属性

作者:nelson f. liu, omer levy, roy schwartz, chahao tan, noah a. smith

摘要: 虽然递归神经网络在各种**自然语言处理**应用中取得了成功，但它们是序列数据的一般模型。我们研究**自然语言**数据的属性如何影响 lstm 学习非语言任务的能力：从其输入中回忆元素。我们发现，在**自然语言**数据上训练的模型能够从更长的序列中召回令牌，而不是在**非语言**序列数据上训练的模型。此外，我们还表明，lstm 通过显式地使用其神经元的子集来计算输入中的时间步长来学习解决记忆任务。我们假设，**自然语言**数据中的模式和结构使 lstm 能够通过提供减少损失的近似方法来学习，但了解不同的培训数据对 lstm 学习能力的影响仍然是悬而未决的问题。少

2018 年 5 月 29 日提交;最初宣布 2018 年 5 月。

评论:7 页，4 个数字；接受 acl 2018 repl4nlp 研讨会

279. 第 1805 5.11224[pdf,其他] Cs。CI

基于搜索的结构化预测的提取知识

作者:刘一佳,车万祥,赵怀鹏,秦冰，刘婷

摘要: 许多**自然语言处理**任务可以建模为结构化预测，并作为搜索问题来解决。在本文中，我们将具有不同初始化训练的多个模型的集合提取为一个模型。除了学习匹配合奏在参考状态上的概率输出外，我们还利用合奏来探索搜索空间，并从探索中遇到的状态中学习。两种典型的基于搜索的结构化预测任务--基于过渡的依赖分析和神经机器翻译的实验结果表明，蒸馏可以有效地提高单一模型的性能，最终模型达到了在 las 中的 1.32 和 BLEU 在这两个任务上的得分分别优于强基线，并且优于以往文献中贪婪的结构化预测模型。少

2018 年 5 月 28 日提交;最初宣布 2018 年 5 月。

评论:将出现在 acl 2018

280. 第 1805 5.11222[pdf,其他] Cs。Lg

在无人监督的情况下与沃瑟斯坦教徒对着住的路线

作者:edouard grave, armand joulin, quentin berthet

文摘: 我们考虑的任务是在高维中对齐两组点，这在**自然语言处理**和计算机视觉中有许多应用。例如，最近有证据表明，通过对在单语数据上训练的单词嵌入进行对齐，可以在没有监督数据的情况下推断双语词典。最近的这些进展是建立在对抗训练的基础上的，以学习这两种嵌入之间的映射。在本文中，我们提出使用一个替代公式，基于正交矩阵和置换矩阵的联合估计。虽然这个问题不是凸的，但我们建议使用凸松弛来初始化我们的优化算法，传统上考虑的是图同构问题。我们提出了一个随机算法，以最大限度地减少我们的成本函数在大规模的问题。最后，我们通过对在单语数据上训练的单词嵌入来评估我们对无监督词翻译问题的方法。在这个任务上，我们的方法获得了最先进的结果，同时比竞争方法所需的计算资源更少。少

2018 年 5 月 28 日提交;最初宣布 2018 年 5 月。

281. 第 1805 5.981[pdf,其他] Cs。Lg

基于神经磁测量生成模型的卷积网结构的稳健、适应性强的脑机接口

作者:ivan zubarev, rasmus zetter, Hanna-Leena halme, lauri parkkonen

文摘: 深部神经网络在图像识别和**自然语言处理**中得到了非常成功的应用。近年来，这些强大的方法也受到了脑机接口 (bci) 社区的关注。在这里，我们介绍了一个卷积神经网络 (cnn) 架构优化的大脑状态分类从非侵入性磁脑电图 (meg) 测量。模型结构是由一个最先进的生成模型的 meg 信号的动机，因此很容易用神经生理术语解释。我们证明，该模型是高度准确的解码事件相关的响应，以及调节振荡大脑活动，并对各自的差异是稳健的。重要的是，该模型可以很好地在用户中推广：在对以前用户收集的数据进行培训时，它可以成功地在新用户上执行。因此，可以省略耗时的 bci 校准。此外，该模型可以增量更新，从而使离线实验的平均精度提高 + 8.9%，在实时 bci 中提高 + 17.0。我们认为，这种模型可以用于实际的 bci 和基础神经科学研究。少

2018 年 5 月 28 日提交;最初宣布 2018 年 5 月。

评论:8 页, 2 个数字, 4 个表格. 关键词: meg、bci、实时、卷积神经网络

282. 第 1805 5.796[pdf,其他] Cs。CI

自然语言处理中的卷积神经网络压缩

作者:krzysztof wróbel, mar 拉姆皮奥伦, maciej wielgosz, michal karwatowski, kazimierz wiatr

摘要: 卷积神经网络是现代模型，在许多分类任务中非常有效。它们最初是为**图像处理**目的而创建的。然后进行了一些试验，以便在**自然语言处理**等不同领域使用它们。人工智能系统 (如人形机器人) 通常是基于对内存、功耗等有限制的嵌入式系统。因此卷积神经网络由于其内存容量的存在，应该减少映射到给定的硬件。本文给出了压缩有效卷积神经网络进行情绪分析的结果。主要步骤是量化和修剪**过程**。给出了将压缩网络映射到 fpga 的方法和实现结果。所描述的模拟结果表明, 5 位宽度足以从网络的浮点版本中获得精度下降。此外，还显著减少了内存占用 (从 85% 减少到 93%)。少

2018 年 5 月 28 日提交;最初宣布 2018 年 5 月。

评论:7 页, 4 个数字, 6 个表

283. 第 xiv:1805 5.10685[pdf,其他] Cs。红外

基于文献向量嵌入和深度学习的法律文献检索

作者:keet sugathadasa, budhi ayesha , nisansa de silva, amal shehan perera, vindula jayawardana, dimuthu lakmal, madhavi perera

文摘: 领域特定信息检索**过程**一直是**自然语言处理**领域的一项突出而持续的研究。许多研究人员采用了不同的技术来克服技术和领域的特殊性，并为不同的领域提供了一个成熟的模型。这些研究的主要瓶颈是领域专家的大量耦合，这使得整个**过程**既耗时又繁琐。在这项研究中，我们开发了三种新的模型，与通过提供的在线存储库生成的金标准进行比较，特别是针对法律领域。这三种不同的模型结合了法律领域的向量空间表示，其中文档矢量生成是在两个不同的机制中完成的，并且是上述两个机制的集合。本研究包含了将法律案例文件表示到不同向量空间的**过程**中正在进行的研究，同时纳入了语义词汇测量和**自然语言处理**技术。本研究建立的集成模型具有较高的精度水平，这确实证明了在信息检索**过程**中引入域特定语义相似度度量的必要性。本研究还表明，词相似度度量的不同分布对不同文档矢量维度的影响，可以导致法律信息检索**过程**的改进。少

2018 年 5 月 27 日提交;最初宣布 2018 年 5 月。

284. **第 1805 5.10393[pdf,其他] Cs。CI**

利用深部神经网络对隐私政策中的语言模糊性进行建模

作者:刘飞,李妮可,廖可欣

摘要: 网站隐私政策太长，难以理解。过于复杂的**语言**使得隐私通知的有效性不如应有的效果。当人们认为隐私政策模糊时，他们就更不愿意分享他们的个人信息。本文从**自然语言处理**的角度对模糊进行解码。虽然彻底识别模糊的术语及其语言范围仍然是一个难以捉摸的挑战，但在这项工作中，我们寻求使用深层神经网络来学习隐私策略中单词的矢量表示。矢量表示被输入到交互式可视化工具 (lstmvis)，以测试它们发现与语法和语义相关的模糊术语的能力。该方法为建模和理解**语言**模糊提供了希望。少

2018 年 5 月 25 日提交;最初宣布 2018 年 5 月。

评论:7 页

285. **第 1805 5.09959[pdf,其他] Cs。CI**

乳腺癌治疗经验的情绪分析及微博上的保健意识

作者:eric m. clark, ted james, chris a.jones, amulya alapati,承诺 ukandu, christopher m.Ukandu , peter sheridan dods

摘要: 背景: 社交媒体有能力为医疗行业提供有价值的反馈，让他们透露和表达自己的医疗决策**过程**，并在期间和之后自我报告的生活质量指标治疗。在之前的工作中，[crannell 等人]，我们在微博上研究了一个活跃的癌症患者群体，并汇编了一套推特，描述他们在这种疾病中的经历。我们将这些在线公开证词称为 "无形病人报告的结果" (ipro)，因为它们带有相关指标，但难以通过传统的自我报告手段获取。方法: 本研究旨在确定与患者体验有关的推特，作为监测公共卫生的补充信息工具。我们使用 twitter 的公共流媒体 api，在 2016 年 9 月至 2017 年 12 月中旬的时间里汇编了超过 530 万条与 "乳腺癌" 相关的推文。我们将监督机器学习方法与**自然语言处理**结合起来，筛选与乳腺癌患者经历相关的推特。我们分析了 845 名乳腺癌患者和幸存者账户的样本，负责 48, 000 多个职位。我们通过对冲情绪分析对推特内容进行了调查，以定量提取充满情绪的话题。结果: 我们发现，在患者治疗、提高支持、传播意识等方面，都有积极的经验。与保健有关的进一步讨论十分普遍，而且基本上是负面的，重点是担心可能导致覆盖面丧失的政治立法。结论: 社交媒体可以为患者提供一个积极的渠道，讨论他们的需求以及对其医疗报道和治疗需求的担忧。从在线通信中获取 ipro 可以帮助 it 专业人员了解情况，并导致更多的关联和个性化的治疗方案。少

2018 年 10 月 12 日提交;v1 于 2018 年 5 月 24 日提交;最初宣布 2018 年 5 月。

286. 第 1805.09906[pdf,其他] Cs。CI

文本网络嵌入的扩散映射

作者:张新元,李一通,沈鼎汉,李一通, 李新华

摘要: 文本网络嵌入利用与网络关联的富文本信息来学习顶点的低维矢量表示。最近的研究没有使用典型的**自然语言处理**(nlp) 方法, 而是利用同一边缘上的文本关系以图形方式嵌入文本。但是, 这些模型忽略了测量图形中任何两个文本之间的完整连接级别。我们提出了文本网络嵌入 (dmte) 的扩散映射, 将图形的全局结构信息集成起来, 捕获文本之间的语义相关性, 并在文本输入上应用扩散卷积操作。此外, 还设计了一个新的目标函数, 利用图形扩散有效地保持高阶接近。实验结果表明, 该方法在顶点分类和链路预测任务上优于最先进的方法。少

2018 年 5 月 24 日提交;最初宣布 2018 年 5 月。

287. 第 1805.0. 09644[pdf,其他] Cs。红外

多伊 10.114/2766462.27 67870

dinfra: 计算多语言语义相关性的一站式商店

作者:siamak Barzegar juliano efsonsales, andre freitas, siegfried handschuh, brian davis

文摘: 本演示提供了一种利用三种分布语义模型 (dsm) 计算 12 种**自然语言**的多语言语义关联和相关性的基础结构。我们的恶魔-diinfra (分布式基础架构) 为研究人员和开发人员提供了一个非常有用的平台, 用于**处理**大型语料库和进行分布式语义实验。我们在 web 服务中集成了多个多语言 dsm, 以便最终用户可以获得结果, 而无需担心构建 dsm 所涉及的复杂性。我们的网络服务使用户能够方便地访问不同参数的 dsm 的广泛比较。此外, 用户还可以使用易于使用的 api 配置和访问 dsm 参数。少

2018 年 5 月 16 日提交;最初宣布 2018 年 5 月。

评论:2 页, 2 个数字, 信号会议

288. 第 1805 5.09120[pdf] Cs。红外

多伊 10.1007/10772-017-9411-7

阿拉伯问题向从 web 中自动提取的逻辑表示

作者:patrice bellot, widbakari, mahmoud neji

摘要: 随着网络上阿拉伯语电子数据的不断增长, 提取信息基本上用于构建文档语料库, 而信息提取实际上是问答的主要挑战之一。事实上, 构建语料库是目前在**自然语言处理**(nlp) 中的其他一些主要主题中提到的一个研究主题, 例如信息检索 (ir)、问答 (qa)、自动摘要 (as) 等一般来说, 问答系统提供了回答用户问题的各种段落。为了使这些段落真正提供信息, 该系统需要访问基础知识库;这就需要构建一个语料库。我们研究的目的是建立一个阿拉伯语问答系统。此外, 分析问题必须是第一步。接下来, 从网络中检索一段可以作为适当答案的段落是非常必要的。本文提出了一种用阿拉伯语分析问题和检索段落答案的方法. 为了进行问题分析,**处理**了五种事实问题类型。此外, 我们的目的是尝试从每个问题的声明形式生成逻辑表示。在回答问题的其他**语言**中, 还讨论了一些研究, 涉及回答问题的逻辑方法. 这种表述是非常有希望的, 因为它有助于我们以后选择一个合理的答案。正确分析并转换为逻辑形式的问题的准确性达到 64%。然后, 自动生成的文本段落的结果获得了 87% 的准确性和 98% 的 c@1 评分。少

2018 年 5 月 23 日提交;最初宣布 2018 年 5 月。

日记本参考:《国际语音技术杂志》, springer verlag, 2017, 20 (2), pp.339-353

289. 第 1805 5.09119[pdf, ps,其他] Cs。Cl

选择机器翻译数据进行自然语言理解系统的快速引导

作者:judith gaspers, penny karanasou, rajen chatterjee

文摘: 本文研究了利用机器翻译 (mt) 引导**自然语言**理解 (nlu) 系统的新**语言**, 用于大型语音控制设备的用例。目标是减少为新**语言**获取附加注释的语料库所需的成本和时间, 同时仍有足够大的用户请求覆盖范围。研究了为保持提高 nlu 性能的话语而对 mt 数据进行滤波的不同方法和**语言**专用**后处理**方法。这些方法在一个大规模的 nlu 任务中进行了测试, 将大约 1000 万个训练话语从英语翻译成德语。结果显示, mt 数据的使用比基于语法的基线和内部数据收集基线有了很大改进, 同时大大减少了人工工作。过滤和**后处理**方法都进一步提高了结果。少

2018 年 5 月 23 日提交;最初宣布 2018 年 5 月。

290. 第 xiv: 180 5.09019[pdf,其他] Cs。简历

美国有线电视新闻网 + 美国有线电视新闻网: 卷积解码器图像字幕

作者:王庆忠,陈庆子

摘要: 图像字幕是一项结合计算机视觉和**自然语言处理**领域的具有挑战性的任务。为了实现图像自动描述的目的, 提出了多种方法, 基于递归神经网络 (rnn) 或长期记忆 (lstm) 模型在这一领域占主导地位。但是, rnn 或 lstm 不能并行计算, 而忽略句子的基本层次结构。在本文中, 我们提出了一个框架, 只使用卷积神经网络 (cnn) 来生成字幕。由于采用了并行计算, 我们的基本模型在训练期间的速度是 nic (基于 lstm 的模型) 的 3 倍左右, 同时也提供了更好的结果。我们对 mscoco 进行了广泛的实验, 并研究了模型宽度和深度的影响。与基于 lstm 的应用类似关注机制的模型相比, 我们提出的模型实现了 bleu-1、2、3、4 和 meteor 的可比分数和较高的 cider 分数。我们还在段落注释数据集上测试我们的模型, 并获得更高的 cider 分数比分层 lstem 少

2018 年 5 月 23 日提交;最初宣布 2018 年 5 月。

291. 第: 1805.0866[pdf, ps,其他] Cs。铬

作者使用广义差分隐私进行模糊处理

作者:娜塔莎·费尔南德斯,马克·德拉斯,安娜贝尔·麦基弗

摘要: 模糊文本文档作者身份的问题迄今在文献中很少受到关注。目前的方法是临时性的, 依赖于对手辅助知识的假设, 这使得很难对这些方法的隐私属性进行推理。差异隐私是一种众所周知且强大的隐私方法, 但它依赖于数据集之间邻接的概念, 使其无法应用于文本文档隐私。但是, 广义差分隐私允许将差分隐私应用于具有度量值的任意数据集, 并已在涉及单个数据点释放的问题上得到证明。本文介绍了如何利用文体测量和**自然语言处理**文献中的现有工具和方法, 将广义微分隐私应用于作者的模糊处理。少

2018 年 5 月 22 日提交;最初宣布 2018 年 5 月。

292. 第 1805.0430[pdf,其他] Cs。直流

rpc 考虑有害的问题: rdma 上的快速分布式深度学习

作者:薛玉龙,苗友山,陈成明, 吴明 , 张林涛,周丽东

摘要: 深度学习作为一项重要的新的资源密集型工作, 已成功地应用于计算机视觉、语音、**自然语言处理**等领域。分布式深度学习已成为应对不断增长的数据和模型大小的必要条件。它的计算通常具有一个简单的张量数据抽象来建模多维矩阵, 一个数据流图

建模计算，迭代执行与相对频繁的同步，从而使它在很大程度上减少风格分布式大数据计算。rpc 作为通信的原始，已被流行的深度学习框架所采用，如使用 grpc 的 tensorflow。研究表明，对于分布式深度学习计算，特别是在支持 rdma 的网络上，rpc 是次优的。张量抽象和数据流图，再加上 rdma 网络，提供了在不牺牲可编程性和通用性的情况下减少不必要的开销 (例如内存副本) 的机会。特别是，从数据访问的角度来看，远程计算机只是作为 rdma 通道上的 "设备" 进行抽象，具有用于分配、读取和写入内存区域的简单内存接口。我们的图形分析器查看数据流图和张量，以便使用此接口优化内存分配和远程数据访问。与 tensorflow 中的标准 grpc 相比，具有代表性的深度学习基准的速度提高了 25 倍，即使是针对 rdma 优化的 rpc 实现也提高了 169%，从而加快了培训过程中的收敛速度.少

2018 年 5 月 22 日提交;最初宣布 2018 年 5 月。

293. 第: 1805 5.07978[pdf,其他] Cs。Lg

流媒体曼: 一种基于流的高效记忆增强神经网络推理

作者:seongsik park, jaehejang, seejoon kim, sungroh yoon

摘要: 随着人工智能在深度学习中的成功发展，人们对人工智能的部署越来越感兴趣。移动环境是最接近现实生活的硬件平台，已成为人工智能成败的重要平台。内存增强神经网络 (mann) 是一种神经网络，旨在有效地处理问答 (问答) 任务，非常适合移动设备。由于 mann 需要各种类型的操作和重复数据路径，因此很难在为其他传统神经网络模型设计的结构中加速推理，这是在移动设备中部署 mann 的最大障碍之一。环境。为了解决上述问题，我们建议流式曼。这是首次尝试实现和演示具有流**处理**概念的 mann 节能推理体系结构。为了充分发挥流**媒体处理**的潜力，我们提出了一种新的方法，称为推理阈值，采用贝叶斯方法，考虑到**自然语言处理**(nlp) 的特点任务。为了评估我们提出的方法，我们在一个适用于流**处理**的现场可编程门阵列 (fpga) 中实现了体系结构和方法。我们测量了 babi 数据集推理的执行时间和功耗。实验结果表明，与 nvidia titan v 的结果相比，流式曼 n 的能量性能效率提高了约 12 倍，如果应用推理阈值，则提高了 140 倍。少

2018 年 5 月 21 日提交;最初宣布 2018 年 5 月。

294. 第 1805 5.07340[pdf,其他] Cs。Lg

基于后缀双向 lstm 的句子建模改进

作者:悉达多·婆罗门

摘要:递归神经网络在序列数据的计算中已变得普遍存在，尤其是**在自然语言处理**中的文本数据。特别是，双向 lstm 是多个神经模型的核心，在广泛的.. 。更多

2018 年 9 月 10 日提交;v1 于 2018 年 5 月 18 日提交;最初宣布 2018 年 5 月。

295. 第 1805 5.0780[pdf,其他] Cs。艾

部分知识编译的近似模型计数

作者:赖勇

文摘 模型计数是计算给定命题公式的满意分配数的问题。尽管大多数知识编译 (kc) 方法**都可以自然**地提供精确的模型计数器，但实际上，由于大小的爆炸，它们无法为某些公式的模型的精确计数生成编译结果。决策 dnf 是一种重要的 kc **语言,**它捕获了大多数实际编译器。我们通过引入一类新的叶顶点 (称为未知顶点)，提出了一个广义的决策 dnnf (称为部分决策 dnnf)，然后提出了一种名为 PartialKC 的算法，从给定的

公式。模型数的无偏估计可以通过随机部分 dnnf 公式计算。partialkc 的每个调用都由 microkc 的多个调用组成，而后一个调用都是一个配备 kc 技术的重要采样**过程**。实验结果表明，par 武 alkc 比样本搜索和搜索树采样器都更准确，paratalkc 的尺度优于 searchtreespler，kc 技术可以显著加快采样速度。少

2018 年 5 月 18 日提交;最初宣布 2018 年 5 月。

296. 第 1805 5.07123[pdf，ps，其他] Cs。Lg

通过自适应符号嵌入进行树编辑距离学习：补充材料和结果

作者:benjamin pašen

摘要：公制学习的目的是通过学习距离测量来提高分类的准确性，这种距离测量使同一类的数据点更加紧密地结合在一起，并将不同类的数据点进一步分开。最近的研究表明，通过学习的成本函数，度量学习方法也可以应用于树，如分子结构、计算机程序的抽象语法树或**自然语言**的句法树。编辑距离，即替换、删除或插入树中节点的成本。然而，直接学习这样的成本可能会产生一个违反度量公理的编辑距离，难以解释，并且可能无法很好地概括。在这个贡献中，我们提出了一种新的树度量学习方法，通过嵌入树节点作为向量间接地学习编辑距离，从而使这些向量之间的欧几里得距离支持类判别。我们通过减少来自同一类的原型树的距离和增加来自不同类的原型树的距离来学习这种嵌入。在我们的实验中，我们表明，我们提出的度量学习方法改进了从计算机科学到生物医学数据到自然语言的六个基准数据集上的树木公制学习的最先进技术**处理**包含超过 300，000 个节点的数据集。少

2018 年 5 月 18 日提交;最初宣布 2018 年 5 月。

评论:通过自适应符号嵌入进行 icml 2018 纸树编辑距离学习的补充材料和其他结果

297. 第 1805 5.07030[pdf，其他] Cs。简历

半样式：学习使用未对齐的文本生成固定化图像字幕

作者:亚历山大·马修斯，谢乐兴，何旭明

摘要：语言风格是书面交际的重要组成部分，具有影响清晰度和吸引力的能力。随着视觉和**语言**的最新进步，我们可以开始解决生成图像字幕的问题，这些标题既是视觉基础的，也是适当的样式。现有方法要么需要与图像对齐的样式训练字幕，要么生成相关性较低的字幕。我们开发了一个模型，学习从大量的样式文本中生成视觉上相关的样式标题，而不需要对齐的图像。这个模型的核心理念叫 semstyle，它的核心思想是分离语义和风格。一个关键组件是使用**自然语言处理**技术和框架语义生成的新颖而简洁的语义术语表示。此外，我们还开发了一个统一的**语言**模型，用不同的单词选择和不同风格的句法来解码句子。自动和手动的评估显示 semstyle 保留图像语义中的标题，是描述性的，并且是样式的转移。更广泛地说，这项工作提供了从网络上大量的语言数据中学习更丰富的图像描述的可能性。少

2018 年 5 月 17 日提交;最初宣布 2018 年 5 月。

评论:2018 年 cvpr 会议接受

298. 第 1805.06150[pdf，其他] 反渗透委员会

追随网：机器人导航遵循自然语言方向与深层强化学习

作者:pararth shah，marek fiser，Aleksandra faust，j.chase kew，dilek hakkani-tur

摘要：了解和遵循人类提供的方向可以使机器人在未知的情况下有效地导航。我们提出了关注网，这是一种用于学习多模联运导航策略的端到端可微神经体系结构。关注网**映**

射自然语言说明以及运动原语的视觉和深度输入。跟随网使用以视觉和深度输入为条件的关注机制**处理**指令，以便在执行导航任务时专注于命令的相关部分。深度强化学习(rl) 稀疏奖励同时学习状态表现、注意力功能和控制策略。我们在复杂的**自然语言**方向数据集上评估我们的代理，这些数据集引导代理通过模拟住宅的丰富而现实的数据集。我们展示了示乐网代理学习执行以前看不到的指令，用类似的词汇描述，并成功地沿着训练中没有遇到的路径导航。在没有注意机制的情况下，该制剂比基线模型有 30% 的改善，在新的指令下成功率为 52%。少

2018 年 5 月 16 日提交;最初宣布 2018 年 5 月。

评论:7 页, 8 个数字

日记本参考:2018 年在 icra 举办的机器人运动规划与控制机器学习第三次研讨会

299. 第 1805.06087 [pdf,其他] Cs。CI

学习用合作的判别师写作

作者:ari holtzman, jan buys , maxwell forbes, antoine bosselut, david golub, yejin choi

摘要: 递归神经网络 (rnn) 是强大的自回归序列模型，但当用于生成**自然语言**时，它们的输出往往过于通用、重复和自相矛盾。我们假设，由 mn **语言**模型优化的目标函数，相当于一个文本的整体困惑，没有足够的表现力来捕捉语言原则描述的交际目标的概念，如格里斯的格言。我们建议学习多种判别模型的混合物，可用于补充 mn 生成器和指导解码**过程**。人类评价表明，我们的系统生成的文本比基线生成的文本有很大的优势，大大加强了生成的文本的总体一致性、风格和信息内容。少

2018 年 5 月 15 日提交;最初宣布 2018 年 5 月。

评论:在 acl 2018 论文集中

300. 第 1805.05670 [pdf,其他] Cs。Db

神经优化满足自然语言处理，以增强数据库教育

作者:刘思元, sourav s Bhowmick, w 骨折,音王舒, wanyi huang, shafiq joty

摘要: 关系数据库管理系统 (rdbms) 是世界各地许多大学教授的主要本科课程，作为其计算机科学课程的一部分。此类课程的核心组件是 rdbms 中查询优化器的设计和实现。查询优化器的目标是自动确定执行用户提交的声明性 sql 查询的最有效执行策略。查询优化**过程**生成一个查询执行计划 (qep)，该计划表示查询的执行策略。由于底层查询优化器的复杂性，对 qep 的理解要求学生了解与 rdbms 相关的特定于实现的问题。实际上，这是一个不现实的假设，因为大多数学生是第一次学习数据库技术。因此，他们往往很难理解 dbms 通过仔细阅读 qep 执行的查询执行策略，从而阻碍了他们的学习**过程**。在本演示中，我们提出了一个名为 neuron 的新系统，该系统促进了与 qep 的**自然语言**交互，以增强其理解。neuron 接受 sql 查询 (可能包括联接、聚合、嵌套等) 作为输入，执行它，并生成执行策略的基于自然 **语言**的描述 (文本和语音形式)由基础 rdbms 部署。此外，它还通过**基于自然语言**的问答框架，促进了对与 qep 相关的各种功能的了解。我们主张，这样的工具是世界上首创的，可以极大地提高学生对查询优化话题的学习。少

2018 年 8 月 20 日提交;v1 于 2018 年 5 月 15 日提交;最初宣布 2018 年 5 月。

301. 第 1805.05588 [pdf,其他] Cs。CI

用神经网络结合正则表达式：口语理解的一个案例研究

作者:罗炳峰,冯燕松,王正,黄松芳,瑞燕,赵东燕

摘要: 许多**自然语言处理**(nlp) 任务的成功取决于注释数据的数量和质量，但此类培训数据往往短缺。在本文中，我们提出了一个问题: "我们能否将神经网络 (nn) 与正则表达式 (re) 结合起来，以改进 nlp 的监督学习？"在回答中，我们开发了新的方法来利用在 nn 内不同层次的 res 的丰富表现力，表明当少量的培训实例可用时，这种组合会显著提高学习的有效性。我们通过将其应用于语音语言理解来评估我们的方法，以进行意向检测和插槽填充。实验结果表明，我们的方法在利用现有训练数据方面非常有效，对不了解 RE-unaware 的 nn 有了明显的推动作用。少

2018 年 5 月 15 日提交;最初宣布 2018 年 5 月。

评论:11 页, 2 数字，被 acl 2018 年接受

302. 第 1805. 05236[pdf,其他] Cs。铬

austra: 全球移动银行应用的大规模自动安全风险评估

作者:陈国柱,孟国柱,苏婷，范玲玲，薛银星,刘洋，徐丽华,薛敏辉,李波，双双豪

摘要: 现代金融技术 (fintech) 以其便利性和效率，被银行等金融机构广泛采用，实现无现金移动支付。然而，金融科技也使大规模和动态的交易容易受到安全风险的影响。鉴于此类漏洞造成的巨大财务损失，已开发了监管技术 (regtech)，但特别希望进行更全面的安全风险评估，以开展稳健、可扩展和高效的财务活动。在本文中，我们进行了首次自动化安全风险评估，并将重点放在全球银行应用上，以考察 fintech。首先，我们分析了大量的银行应用，并提出了一套全面的安全弱点，广泛存在于这些应用程序中。其次，我们设计了一个三相自动化安全风险评估系统 (ausera)，该系统将**自然语言处理**和数据和控制流的静态分析结合起来，以有效地识别银行应用程序。我们在 80 多个国家/地区的 693 个真实世界银行应用上进行了实验，并揭示了 2, 157 弱点。迄今为止，已有 21 家银行承认了我们报告的弱点。我们发现过时的银行应用程序版本、来自第三方库的污染以及薄弱的哈希函数都可能被攻击者利用。我们还显示，不同来源的银行应用表现出各种类型的安全弱点，主要原因是经济和法规正在形成。鉴于中介**性质**的急剧变化，雷加科技公司和所有利益相关者都有责任了解当代金融技术带来的安全风险的特点和后果。少

2018 年 6 月 7 日提交;v1 于 2018 年 5 月 14 日提交;最初宣布 2018 年 5 月。

303. 第 1805 5.05144[pdf,其他] si

三次飓风的推特故事: 哈维、伊尔玛和玛丽亚

作者:firoj alam, ferda ofli, muhammad imran, michael Aupetit

摘要: 人们在**自然灾害**和紧急情况下越来越多地使用微博平台，如推特。研究表明，微博上的数据对若干救灾任务很有用。然而，由于多种原因，例如用于分析大容量和高速数据流的可用工具有限，因此理解社交媒体数据是一项具有挑战性的任务。这项工作对三个灾害事件期间在微博上分享的数百万条推特的文字和多媒体内容进行了广泛的多维分析。具体而言，我们采用**了自然语言处理**和计算机视觉领域的各种人工智能技术，利用不同的机器学习算法来**处理**数据在灾难事件期间生成。我们的研究揭示了各种有用信息的分布情况，这些信息可以为危机管理人员和应对者提供信息，并促进未来灾害管理自动化系统的开发。少

2018 年 5 月 15 日提交;v1 于 2018 年 5 月 14 日提交;最初宣布 2018 年 5 月。

评论:参加 iscram 2018 会议

304. 第: 1805.0. 4793[pdf,其他] Cs。CI

神经语义解析中的编码到精细解码

作者:李东, mirella lapata

摘要: 语义解析的目的是将**自然语言**话语映射到结构化的意义表示中。在本文中，我们提出了一种结构感知神经体系结构，将语义分析**过程分解为**两个阶段。给定输入话语，我们首先生成其含义的粗略草图，其中低级别信息 (如变量名和参数) 被掩盖。然后，我们通过考虑**自然语言**输入和素描本身来填写缺失的细节。对具有不同域和意义表示特征的四个数据集的实验结果表明，尽管使用了相对简单的解码器，但我们的方法始终提高了性能，取得了有竞争力的结果。少

2018 年 5 月 12 日提交;最初宣布 2018 年 5 月。

评论:acl-18 接受

305. 第 1805.04617[pdf,其他] Cs。CI

图则银行: 一个人工收集的前链、调查提取和资源推荐语料库

作者:亚历山大·法布里, 李爱琳, 布拉特·特拉拉托沃拉库尔,何一娇, 魏大婷, 东德星, 凯特琳韦斯特菲尔德,德拉戈米尔·拉德夫

摘要: **自然语言处理**(nlp) 领域正在迅速发展，每天发表新的研究报告，同时还有大量的教程、代码库和其他在线资源。为了了解这一动态领域或保持最新的研究，学生以及教育工作者和研究人员必须不断筛选多个来源，以找到有价值的相关信息。为了解决这种情况，我们引入了 tutorialbank，这是一个新的、可公开提供的数据集，旨在促进 nlp 教育和研究。我们在 nlp 以及人工智能 (ai)、机器学习 (ml) 和信息检索 (ir) 等相关领域手动收集和分类了超过 6300 个资源。我们的数据集显然是用于 nlp 教育的最大的人工资源，不包括学术论文。此外，我们还为资源创建了搜索引擎和命令行工具，并对语料库进行了注释，以包括研究主题列表、每个主题的相关资源、主题之间的先决条件关系、除其他注释外，还有其他资源。我们正在发布数据集，并提出了进一步研究的几种途径。少

2018 年 5 月 11 日提交;最初宣布 2018 年 5 月。

评论:acl 2018，计算语言学协会第 56 届年会，澳大利亚墨尔本, 2018

306. 第 1805.0 04453[pdf,其他] Cs。CI

通过对话自动化的机器翻译引导多语言意图模型

作者:nicolas ruiz , srinivas bangalore, john chen

摘要: 随着基于聊天的对话系统在消费者和企业应用程序中的复兴，在开发数据驱动和基于规则的**自然语言**模型以了解人类意图方面取得了很大成功。由于这些模型需要大量的数据和领域内的知识，因此将等效服务扩展到新市场会被抑制对话自动化的**语言**障碍所破坏。本文介绍了一个用户研究，以评估开箱即用的机器翻译技术的效用, (1) 快速引导多语言口语对话系统, (2) 使现有的人类分析人员能够理解外语话语。此外，我们还评估机器翻译在人工辅助环境中的效用，在这种环境中，部分流量由分析师**处理**。在英语-& gt; 西班牙语实验中，我们观察到对话自动化的巨大潜力，以及人类分析人员处理**外语**话语的潜力，并具有较高的准确性。少

2018 年 5 月 11 日提交;最初宣布 2018 年 5 月。

评论:2018 年欧洲机器翻译协会会议 (eamt 2018) 可出版 6 页, 3 个数字

307. 第 1805.03838[pdf, ps,其他] Cs。CI

用于神经序列标记的混合半马尔可夫 crf

作者:叶志秀,凌振华

文摘: 本文提出了**用于自然语言处理**中神经序列标注的混合半马尔可夫条件随机场(scrf)。在传统条件随机字段 (crf) 的基础上，通过从要素中提取要素并描述线段之间的转换而不是单词，为将标签分配给线段的任务而设计了 scrf。本文通过同时使用字级和分段级信息，对现有的 scrf 方法进行了改进。首先，使用文字级别的标签来派生 scdf 中的段分数。其次，将 crf 输出层和 scrf 输出层集成到一个统一的神经网络中，并进行联合训练。在 conll 2003 命名实体识别 (ner) 共享任务上的实验结果表明，在不使用外部知识的情况下，我们的模型实现了最先进的性能。少

2018 年 5 月 10 日提交;最初宣布 2018 年 5 月。

评论:本文件已被 acl 2018 篇接受

308. 第 1805.3735[pdf,其他] Cs。铬

计算机网络流量异常检测的序列聚合规则

作者:benjamin j. radford , bartley d.richardson , shawn e. davis

文摘: 我们评估了将无监督异常检测应用于计算机网络流量数据或流量网络安全应用的方法。我们借鉴**了自然语言处理**文献，将流动概念化为机器之间使用的一种 "**语言**"。评估了五个序列聚合规则在标记流数据集中标记多个攻击类型 (cicidsse2017) 的有效性。对于序列建模，我们依赖于长的短期存储器 (lstm) 递归神经网络 (rnn)。此外，还描述了一个简单的基于频率的模型，并将其在攻击检测方面的性能与 lstm 模型进行了比较。我们的结论是，基于频率的模型在手头的任务中的性能往往与 lstm 模型相同或更好，但有几个明显的例外。少

2018 年 5 月 14 日提交;v1 于 2018 年 5 月 9 日提交;最初宣布 2018 年 5 月。

评论:为 2018 年美国统计协会数据科学和统计研讨会编写

309. 第 1805.3435[pdf,其他] Cs。艾

解码器: 为无监督相似性任务找到最佳表示空间

作者:vitalii zhelezniak, dan busbridge, 4 月 shen, samuel l.smith , nils y. hammerla

摘要: 实验证据表明，在许多无监督相似任务上，简单模型的性能优于复杂的深部网络。我们通过引入最优表示空间的概念，对这种行为提供了简单而严格的解释，在这个概念中，语义上的闭合符号映射到在模型的相似度下接近的表示。目标功能。此外，我们提出了一个简单的过程，无需任何再培训或架构修改，允许深度经常性模型与浅层模型相比表现良好 (有时更好)。为了验证我们的分析，我们进行了一组一致的经验评价，并在这个**过程**中引入了几个新的句子嵌入模型。尽管这项工作是在**自然语言处理**的背景下提出的，但这些见解很容易适用于依赖分布式表示来执行传输任务的其他域。少

2018 年 5 月 9 日提交;最初宣布 2018 年 5 月。

评论:iclr 2018 年研讨会轨道, 15 页, 3 个数字, 6 个表格

310. 第: 1805.03366[pdf,其他] Cs。Cl

利用 pu 学习学习低资源语言的词汇

作者:赵江,余祥福,谢朝杰,张启伟

摘要: word 嵌入是**处理自然语言**的许多下游应用程序中的关键组件。现有的方法往往假定存在大量的文本集合，用于学习有效的单词嵌入。但是，对于某些资源较低的**语言,**

可能无法使用这样的语料库。本文研究了如何有效地学习只有几百万令牌的语料库上的单词嵌入模型。在这种情况下，由于许多词对的共现没有观测到，共现矩阵是稀疏的。与现有的方法相比，通常只将几个未观察到的单词对作为负样本进行采样，我们认为共现矩阵中的零条目也提供了有价值的信息。然后，我们设计了一种积极的无标记学习(pu-learing) 方法来分解共发生矩阵，并验证四种不同**语言**的建议方法。少

2018 年 5 月 9 日提交;最初宣布 2018 年 5 月。

评论:发表于 naacl 2018

311. 第 1805.02917[pdf,其他] Cs。Lg

文本输入嵌入空间中的可解释的对抗性摄动

作者:佐藤元一,铃木君，广行新多，松本宇治

文摘: 在图像**处理**领域取得了巨大的成功之后，对抗性训练的思想已经应用到**了自然语言处理**(nlp) 领域的任务中。一种很有希望的方法是将图像**处理**领域开发的对抗训练直接应用于输入字嵌入空间，而不是文本的离散输入空间。但是，这种方法放弃了生成对抗文本以显著提高 nlp 任务的性能等可解释性。本文通过限制对输入嵌入空间中现有词的摄动方向，恢复了这些方法的可解释性。因此，我们可以通过考虑在保持甚至改进任务性能的同时将摄动作为句子中单词的替换来直接地将每个输入与摄动重构到实际文本中。少

2018 年 5 月 8 日提交;最初宣布 2018 年 5 月。

评论:8 页, 4 个数字

日记本参考:ijcai-ecai-2018

312. 第 1805.01554[pdf,其他] Cs。铬

分层 lst 与反网络钓鱼监督的深度学习模型

作者:明恩、 toan nguyen、thien huu nguyen

摘要: 反网络钓鱼旨在检测文本数据池中的网络钓鱼内容文档。这是网络安全中的一个重要问题，可以帮助保护用户免受虚假信息的影响。**天然的的语言加工过程**(nlp) 为这一问题提供了一个**自然**的解决方案，因为它能够分析文本内容以执行智能识别。在本工作中，我们将研究 nlp 中最先进的文本分类技术，以解决电子邮件的反网络钓鱼问题(即预测电子邮件是否为网络钓鱼)。这些技术是基于最近引起社会各界极大关注的深度学习模式。特别是，我们提出了一个框架，具有分层长短期记忆网络 (h-lstm) 和注意机制，在单词和句子级别同时建模电子邮件。我们的期望是为反网络钓鱼制定一个有效的模型，并展示深入学习网络安全问题的有效性。少

2018 年 5 月 3 日提交;最初宣布 2018 年 5 月。

评论:在: r. verma, a. das。(编辑): 第 4 届国际安全和隐私分析研讨会 (iwspa 2018) 第一次反网络钓鱼共享试点论文集，美国亚利桑那州坦佩, 21-03-2018

313. 第 1805.01542[pdf,其他] Cs。Cl

智能代理的自然语言理解功能的快速和可扩展扩展

作者:anuj goyal, angeliki metallinou, spyros matsoukas

文摘: 快速扩展智能虚拟代理的**自然语言**功能对于实现引人入胜且信息丰富的交互至关重要。然而，为新**的自然语言**领域开发准确的模型是一个时间和数据密集型**过程**。我们提出了高效的深度神经网络架构，通过转移学习最大限度地重用可用资源。我们的方法被应用于扩展流行的商业代理的理解能力，并在数百个新的领域进行评估，由内部

或外部开发人员设计。我们证明，我们提出的方法显著提高了低资源设置下的准确性，并能够以更少的数据快速开发准确的模型。少

2018 年 5 月 3 日提交;最初宣布 2018 年 5 月。

评论:出现在《 nacl-hlt 2018 》论文集 (行业跟踪)

314. 第 09iv:18005.01112[pdf,其他] Cs。Cl

semval-2018 任务 3 中的 binarizer: 分析依赖关系和深度学习以进行反讽检测

作者:Nikhil nikhil, muktabh mayank srivastava

摘要: 在本文中，我们描述了由 binarizer 团队提交给学期 2018 年任务 3 (英语推特中的反义检测) 子任务 a 的系统。反讽检测是许多**自然语言处理**工作的关键任务。我们的方法处理讽刺的推特，由包含不同情绪的较小部分组成。我们使用依赖关系解析器将推特分解为单独的短语。然后，我们使用基于 lstm 的神经网络模型嵌入这些短语，该模型经过预先训练，可以预测推特的表情符号。最后，我们训练一个完全连接的网络来实现分类。少

2018 年 5 月 3 日提交;最初宣布 2018 年 5 月。

评论:2018 年学期任务的解决方案 3

315. 第 1805.01083[pdf,其他] Cs。Db

文本的可扩展语义查询

作者:王晓兰,冯亚伦, behzad golshan, alon halevy, george mihaila , hidekazu oiwa, w· wan chiew tan

摘要: 我们提出了 koko 系统，通过将自然 **语言处理**技术的进步纳入其提取**语言**，将声明性信息提取提升到一个新的水平。koko 是一种新颖的方法，因为它的提取**语言**同时支持文本表面和依赖解析树的结构上的条件，从而允许更精细的提取。koko 还支持对表达概念的语言变体具有宽容性的条件，并允许聚合整个文档中的证据，以便过滤提取。为了扩展规模, koko 利用了多索引方案和启发式方法，实现了高效提取。我们对 koko 进行了广泛的评估，而不是公开的文本语料库。我们表明, koko 指数占用的空间最小，明显比以前的一些索引方案更快、更有效。最后，我们在 500 万篇维基百科文章中展示了 koko 的规模。少

2018 年 5 月 2 日提交;最初宣布 2018 年 5 月。

316. 第 09iv:1805. 00327[pdf,其他] Cs。Lg

一种神经记忆网络的分类方法

作者:马英,何塞

摘要: 本文根据内存网络的内存组织结构，提出了一种内存网络的分类方法。分类包括所有流行的记忆网络: 香草递归神经网络 (rnn)，长期短期记忆 (lstm)，神经堆栈和神经图灵机及其变种。分类学将所有这些网络置于一个保护伞下，并显示其相对表达能力，即香草 rnn & lt; = lst & lt; = 神经堆栈 & lt; = 神经 ram。分析了这些网络之间的差异和共性。这些差异还与不同任务的要求有关，这些任务可以为用户提供如何为特定任务选择或设计适当的内存网络的说明。作为一个概念上简化的问题，开发并测试了合成符号序列的四个任务: 计数、有干扰的计数、反转和重复计数，以验证我们的论点。我们使用两个**自然语言处理**问题来讨论这个分类如何帮助选择合适的神经记忆网络来解决现实世界的问题。少

2018 年 5 月 1 日提交;最初宣布 2018 年 5 月。

317. 第 09iv:18005.00145[pdf,其他] Cs。简历

基于对话的交互式图像检索

作者:郭晓晓,吴惠晓,程宇, 史蒂文·伦尼, 杰拉尔德·特索罗,罗杰里奥·施密特·费里斯

摘要: 现有的交互式图像检索方法证明了整合用户反馈、提高检索结果的优点。但是,大多数当前系统依赖于受限制的用户反馈形式, 如二进制相关性响应, 或基于一组固定的相对属性的反馈, 从而限制了它们的影响。在本文中, 我们介绍了一种新的交互式图像搜索方法, 使用户能够通过**自然语言**提供反馈, 从而实现更**自然**和有效的交互。我们将基于对话的交互式图像检索任务表述为强化学习问题, 并奖励对话系统在每次对话转向过程中提高目标图像的排名。为了在对话系统学习时减少收集人机对话的繁琐而昂贵的**过程**, 我们使用用户模拟器对我们的系统进行培训, 该模拟器本身就被训练来描述目标和候选图像之间的差异。我们的方法的有效性在鞋类检索应用中得到了证明。对模拟数据和真实世界数据的实验表明, 1) 我们提出的学习框架比其他监督和强化学习基线和 2) 基于**自然语言**的用户反馈实现了更好的准确性比预先指定的属性导致更有效的检索结果, 以及更**自然**和更有表现力的通信接口。少

2018 年 11 月 1 日提交;v1 于 2018 年 4 月 30 日提交;最初宣布 2018 年 5 月。

318. 第 xiv:1804. 10765[pdf, ps,其他] Cs。艾

在受控自然语言中指定和验证应答集程序

作者:罗尔夫·施维特

摘要: 我们展示了如何使用双向语法来指定和使用可控**自然语言**的答案集程序。我们从受控**自然语言**的程序规范开始, 并将该规范自动转换为可执行的答案集程序。生成的应答集程序可以按照特定的命名约定进行修改, 然后可以使用用作规范**语言的相同自然语言**子集对程序的修订版本进行重复化。双向语法被参数化以**进行处理**和生成, 处理引用表达式, 并在需要复制这些语法规则时利用语法规则数据结构中的对称性。我们证明,垂直化需要句子规划, 以便聚合类似的结构, 从而提高生成规范的可读性。在不进行修改的情况下, 生成的规范在语义上始终与原始规范等效;我们的双向语法是第一个允许在受控**自然语言处理**的上下文中进行语义往返的语法。本文件正在考虑接受 tplp。少

2018 年 4 月 28 日提交;最初宣布 2018 年 4 月。

评论:在第 33 届国际逻辑编程会议 (iclp 2018) 上发表的论文, 英国牛津, 2018 年 7 月 14 日, 15 页, latex, (arxiv: yymm)。nnnnn)

319. 第 xiv:1804. 10669[pdf,其他] Cs。Sd

基于向量空间投影的深度语音去噪

作者:jeff hetherly, paul gamble, maria barrios, cory stephenson, karl ni

摘要: 我们提出了一种在非平稳和动态噪声存在的情况下从单个麦克风中去除扬声器的算法。我们的方法是由最近成功的神经网络模型的成功, 将扬声器与其他扬声器和歌手从器乐伴奏。与现有技术不同的是, 我们利用源对比估计生成的嵌入空间, 这是一种从**自然语言处理**中的负采样技术中获得的技术, 同时获得连续推理掩码。我们的嵌入空间通过共同建模扬声器和噪声的特性, 直接优化了它们的识别。这个空间是可以概括的,因为它不是扬声器或噪声特定的, 即使模型在训练集中没有看到扬声器, 也能去噪。参数具有双重目标: 一种是增强选择性带通滤波器, 可消除超过信号功率的时频位置的噪声, 另一种是在信号和噪声之间按比例拆分时频内容的滤波器。我们比较了最先进的算法以及传统的稀疏非负矩阵分解解。由此产生的算法通过提供更直观、更易于优化的方法, 同时实现具有竞争力的准确性, 从而避免了严重的计算负担。少

2018 年 4 月 27 日提交;最初宣布 2018 年 4 月。

评论:arxiv 管理说明: 文本与 arxiv:1705.0064662 重叠

320. 第 xiv:1804. 104486[pdf, ps,其他] lo c

功能需求的一致性检查

作者:西蒙娜·沃托

摘要: 需求是对系统预期行为的非正式和半正式描述。它们通常以**自然语言**句子的形式表示，并通过同行评审等方式手动检查错误。手动检查容易出错，耗时且不可扩展。随着网络物理系统日益复杂，需要在安全和安全关键环境中运行，自动化需求的一致性检查并构建工件以帮助系统工程师进行设计变得至关重要**程**。少

2018 年 4 月 27 日提交;最初宣布 2018 年 4 月。

321. 第 1804. 09779[pdf, ps,其他] Cs。Cl

基于自然语言推理的神经机器翻译语义现象评价研究

作者:adam poliak, y 奥纳坦·贝林科夫, james glass , benjamin van durme

文摘: 我们提出了一个程序来调查神经机器翻译 (nmt) 系统产生的句子表示编码不同语义现象的**程度**。我们使用这些表示作为特征来训练基于数据集从现有语义注释重铸的**自然语言**推理 (nli) 分类器。在将此**过程**应用于具有代表性的 nmt 系统时，我们发现它的编码器似乎最适合于在语法语义接口上支持推论，而不是需要世界知识的回指分辨率。最后，我们讨论了现有**过程**的优点和潜在缺陷，以及如何将其改进和扩展为评估语义覆盖的更广泛框架。少

2018 年 5 月 6 日提交;v1 于 2018 年 4 月 25 日提交;最初宣布 2018 年 4 月。

评论:将在 naacl 2018-11 页上提交

322. 第 1804. 09558[pdf,其他] Cs。Cl

wordnet 的视觉距离

作者:raquel pérez-arnal, armand vilalta, dario garcia-gasulla, ulises cortés, edward ayguadé, jesus labarta

摘要: 测量概念之间的距离是**自然语言处理**的一个重要研究领域，因为它可用于改进与解释这些相同概念有关的任务。wordnet 包含与单词 (即同步集) 相关的各种概念，通常用作计算这些距离的来源。本文探讨了基于视觉特征的 wordnet 同步的距离，而不是词汇同步的距离。为此，我们提取使用 imagenet 训练的深层卷积神经网络中生成的图形要素，并使用这些特征生成每个同步的代表。在这些表示的基础上，我们定义了同步的距离度量，它补充了传统的词汇距离。最后，我们提出了一些实验来评价它的性能，并将其与目前的最先进的方法进行了比较。少

2018 年 4 月 27 日提交;v1 于 2018 年 4 月 24 日提交;最初宣布 2018 年 4 月。

323. 第 1804. 09540[pdf,其他] Cs。Cl

困扰您的神经方法的命名实体? 构建 ne 表: 一种用于处理命名实体的神经方法

作者:janarthanan rajendran, jatin ganhotra, xiaoxiaoguo, mo yu, satinder singh

摘要: 许多**自然语言处理**任务需要处理命名实体 (ne) 在文本本身，有时也在外部知识来源。虽然这对人类来说通常很容易，但最近依靠学习过的单词嵌入进行 nlp 任务的神经方法很难使用它，尤其是词汇不足或罕见的 ne。本文针对这一问题提出了一种新的神经方法，并对结构化的问答任务、三个相关的面向目标的对话任务和基于阅读-理

解的任务进行了实证评价。它们表明，我们提出的方法可以有效地处理词汇和词汇外 (oov) ne。我们创建扩展版本的对话框 babi 任务 1, 2 和 4 和词汇外 (oov) 版本的 cbt 测试集，将在网上公开提供。少

2018 年 4 月 22 日提交;最初宣布 2018 年 4 月。

324. **第: 180008881[pdf,其他] Cs。Cl**

评估具有缩放属性的语言模型

作者:takahashi, kuriko tanaka-shii

摘要: **语言**模型主要是用困惑来评价的。虽然困惑量化了最容易理解的预测性能，但它没有提供模型成败的定性信息。因此，提出了另一种使用**自然语言**的缩放属性来评估**语言**模型的方法。考虑了五个这样的测试，前两个测试占词汇量的数量，其他三个解释**自然语言**的长期记忆。通过这些测试对以下模型进行了评估: n 克、概率无上下文语法 (pcfg)、simon 和 pitman-yor (py)**过程**、分层 py 和神经**语言**模型。只有神经**语言**模型表现出**自然语言**的长记忆特性，但程度有限。还讨论了这些模型的每次测试的有效性。少

2018 年 4 月 24 日提交;最初宣布 2018 年 4 月。

评论:14 页, 16 位数字

325. **第 1804.08186[pdf,其他] Cs。Cl**

文本中的自动语言识别研究综述

作者:tommi jauhiainen, marcolui , marcos zampieri, timothy baldwin, krister linden

文摘: **语言**识别 (li) 是确定文档或其一部分所用的**自然语言**的问题。自动 li 已经进行了五十多年的广泛研究。今天, li 是许多文本**处理**管道的关键部分，因为**文本处理**技术通常假定输入文本的**语言**是已知的。这方面的研究最近特别活跃。本文简要介绍了 li 研究的历史，并对 li 文献中迄今使用的特点和方法进行了广泛的调查。为了描述特征和方法，我们引入了一个统一的表示法。我们将讨论评估方法、li 的应用，以及不需要最终用户培训的现成 li 系统。最后，我们确定了悬而未决的问题，调查了迄今为止就每个问题开展的工作，并提出了今后在 li 中进行研究的方向。

2018 年 4 月 22 日提交;最初宣布 2018 年 4 月。

评论:在 jair-人工智能研究杂志上进行审查

326. **第 1804.08139[pdf,其他] Cs。Cl**

相同的表示，不同的注意: 可共享句子表示从多个任务学习

作者:郑仁杰,陈俊坤,邱锡鹏

摘要: 分布式表示在深度学习的**自然语言处理**中发挥着重要作用。然而，在不同的任务中，句子的表示方式往往各不相同，这通常是从零开始学习的，而且培训数据的数量有限。在本文中，我们声称一个好的句子表示应该是不变的，可以有利于后续的各种任务。为实现这一目标，我们提出了一种新的多任务学习信息共享方案。更具体地说，所有任务共享相同的句子表示形式，并且每个任务都可以从具有注意机制的共享句子表示中选择特定于任务的信息。每个任务关注的查询向量可以是静态参数，也可以是动态生成的。我们对 16 种不同的文本分类任务进行了广泛的实验，展示了我们的体系结构的优势。少

2018 年 4 月 22 日提交;最初宣布 2018 年 4 月。

评论:7 页

日记本参考:ijcai 2018

327. 第: 1804. 07942[pdf,其他] Cs。CI

生成股票问题回答

作者:赵鹏图,江勇,刘晓江, 雷舒 , 石树明

摘要: 我们研究股票相关问题的回答 (stockqa): 自动生成股票相关问题的答案, 就像专业股票分析师应用户要求向股票提供行动建议一样。stockqa 与以前的 qa 任务有很大的不同, 因为 (1) stockqa 中的答案是**自然语言**句子 (而不是实体或价值观), 由于 stockqa 的动态**性质**, 几乎不可能得到以从训练数据中提取的合理答案;(2) 股票 qa 要求正确分析 qa 对中的关键字与股票的数值特征之间的关系。我们建议使用内存增强编码器解码器体系结构来解决这个问题, 并集成不同的数字理解和生成机制, 这是 stockqa 的一个关键组成部分。我们构建了一个包含超过 180k stockqa 实例的大型数据集, 在此基础上对各种技术组合进行了广泛的研究和比较。实验结果表明, 具有独立字符分量的混合字符字符模型用于数字**处理**, 取得了最佳性能。通过对结果的分析, 我们发现, 我们最好的模型产生的答案中, 有 24.8% 仍然存在一般的答案问题, 而简单的混合检索生成模型可以缓解这一问题。少

2018 年 9 月 20 日提交;v1 于 2018 年 4 月 21 日提交;最初宣布 2018 年 4 月。

评论:数据: http://ai.tencent.com/ailab/nlp/data/stockQA.tar.gz

328. 第: 1804. 07911[pdf,其他] Cs。CI

通用句子嵌入的多任务学习: 一种使用传输和辅助任务的全面评估

作者:wasi uddin ahmad, xueying bai, Zhechao huang, chao jiang, nanyun peng, kai-wei chang

摘要: 学习分布式句子表示是**自然语言处理**中的关键挑战之一。先前的研究表明, 基于重复神经网络 (rnn) 的句子编码器在大量带注释的**自然语言**推理数据上接受训练, 可以有效地进行迁移学习, 以方便其他相关任务.本文通过对多任务和单任务学习的句子编码器进行了广泛的实验和分析, 证明了多任务的联合学习可以实现更好的句子表达。利用辅助任务进行的定量分析表明, 与单任务学习相比, 多任务学习有助于在句子表示中嵌入更好的语义信息。此外, 我们还将多任务句子编码器与上下文化的单词表示进行了比较, 并表明将它们结合起来可以进一步提高迁移学习的性能。少

2018 年 8 月 16 日提交;v1 于 2018 年 4 月 21 日提交;最初宣布 2018 年 4 月。

329. 第: 1804. 07847[pdf,其他] Cs。CI

多伊 10.1016/j.eswa.2018.07.032

作为多头选择问题的联合实体识别与关系提取

作者:giannis bekoulis, jones deleu, thomas demeester, chris develder

文摘: 最先进的联合实体识别和关系提取模型在很大程度上**依赖**于外部**自然语言处理** (nlp) 工具, 如 pos (词性部分) 分析器和依赖关系解析器。因此, 这种联合模型的性能取决于从这些 nlp 工具中获得的功能的质量。但是, 对于各种**语言**和上下文, 这些功能并不总是准确的。本文提出了一种联合神经网络模型, 该模型可同时进行实体识别和关系提取, 无需手动提取特征或使用任何外部工具。具体来说, 我们使用 crf (条件随机字段) 图层将实体识别任务建模, 并将关系提取任务建模为多头选择问题 (即, 可能识别每个实体的多个关系)。我们提供了一个广泛的实验设置, 以演示我们的方法的有效性, 使用来自不同上下文 (即新闻、生物医学、房地产) 和**语言**(即英语、荷兰语) 的数

据集。我们的模型的性能优于以前使用自动提取的特征的神经模型，而它的性能在基于特征的神经模型的合理范围内，甚至超过了它们。少

2018 年 8 月 16 日提交;v1 于 2018 年 4 月 20 日提交;最初宣布 2018 年 4 月。

评论:预打印-接受在具有应用程序的专家系统中发布

日记本参考:专家系统，第 114 卷, 2018 年 12 月 30 日, 34-45 页, issn 0957-4174 页

330. **第 1804. 07827[pdf,其他] Cs。CI**

高效的上下文表示：序列标记的语言模型修剪

作者:刘丽元,项仁,尚景波,彭健,韩嘉伟

摘要: 已经做出了许多努力，利用预先培训的**语言**模型 (lm) 促进**自然语言处理**任务，并对各种应用程序进行了重大改进。为了充分利用几乎无限的语料库和获取各种级别的语言信息，需要大尺寸的 lm;但对于特定的任务，只有这些信息的一部分是有用的。这样的大型 lm，即使在推理阶段，也可能会导致大量的计算工作负载，从而使其对于大规模应用程序过于耗时。在这里，我们建议压缩笨重的 lm，同时保留有关特定任务的有用信息。由于模型的不同层次保持不同的信息，我们开发了一种利用稀疏诱导正则化的模型修剪层选择方法。通过引入密集的连接，我们可以在不影响他人的情况下分离任何层，并将浅层和宽的 lm 拉伸得又深又窄。在模型培训中, lm 是通过分层辍学学习的，以获得更好的鲁棒性。在两个基准数据集上的实验证明了该方法的有效性。少

2018 年 9 月 10 日提交;v1 于 2018 年 4 月 20 日提交;最初宣布 2018 年 4 月。

评论:emnlp 2018

331. **第: 1804. 07686[pdf,其他] Cs。Db**

验证关系数据集的文本摘要

作者:saehanjo, immanuel trummer, wewelyu , daniel liu, xuezhiwang, congyu, niyati mehta

摘要: 我们提出了一个新**的自然语言**查询接口，攻击检查器，旨在对关系数据集的文本摘要。该工具侧重于转换为 sql 查询和声明查询结果的**自然语言**声明。在精神上类似于拼写检查器，攻击检查器标记的文本段落似乎与实际数据不一致。该系统的核心是一个概率模型，它以整体的方式解释输入文档的原因。根据声明关键字和文档结构，它将每个文本声明映射到关联查询转换的概率分布。通过为典型的输入文档有效地执行数万到几十万个候选翻译，系统将文本声明映射为正确性概率。此**过程**通过专门**的处理**后端变得实用，通过查询合并和结果缓存避免了冗余工作。验证是一个交互式**过程**，在这个过程中，向用户显示暂定结果，使他们能够在必要时采取纠正措施。我们的系统在一组 53 个公共文章中进行了测试，其中包含 392 项索赔。我们的测试案例包括主要报纸的文章、调查结果摘要和维基百科的文章。我们的工具揭示了大约三分之一的测试用例中的错误说法。详细的用户研究表明，与通用 sql 接口相比，使用我们工具的用户检查文本摘要的速度平均要快六倍。在全自动验证中，我们的工具比自然 **语言**查询接口和事实检查领域的基线实现了更高的召回和精度。少

2018 年 8 月 30 日提交;v1 于 2018 年 4 月 20 日提交;最初宣布 2018 年 4 月。

评论:18 页, 13 份, 11 张表格

332. **第: 1804. 07461[pdf,其他] Cs。CI**

glue：理解自然语言的多任务基准和分析平台

作者:alexwang, amampreet singh, julian michael, felixhill, omerlevy, samuel r. bowman

摘要: 要使**自然语言**理解 (nlu) 技术在实际和作为科学研究对象上发挥最大的作用，它必须是通用的: 它必须能够以不通用的方式**处理语言**专门针对任何一个特定的任务或数据集进行定制。为了实现这一目标，我们引入了通用**语言**理解评估基准 (glue)，这是一个评估和分析模型在各种现有 nlu 任务中的性能的工具。glue 是模型无关的，但它鼓励跨任务共享知识，因为某些任务的培训数据非常有限。我们还提供手工制作的诊断测试套件，可对 nlu 型号进行详细的语言分析。我们根据当前的多任务和转移学习方法评估基线，并发现它们并不能立即使每个任务的单独模型的培训的总体性能有实质性的改进，这表明在以下方面仍有改进的余地:开发通用和强大的 nlu 系统。少

2018 年 9 月 18 日提交;v1 于 2018 年 4 月 20 日提交;最初宣布 2018 年 4 月。

评论:https://gluebenchmark.com/

333. 第 1804. 07045[pdf,其他] Cs。Lg

语义对抗性深层学习

作者:tommaso dreossi, somesh jha, sanjit a. seshia

摘要: 在海量数据的推动下，机器学习 (ml) 算法 (特别是深度神经网络) 产生的模型正被用于值得关注的不同领域，包括汽车系统、金融、医疗保健、**自然语言处理**和恶意软件检测。特别令人关切的是，在自驾游和航空等网络物理系统 (cps) 中使用 ml 算法，对手可能会造成严重后果。但是，生成对抗示例和设计健壮 ml 算法的现有方法大多忽略了包含 ml 组件的整个系统的语义和上下文。例如，在使用深度学习进行感知的自主车辆中，并不是神经网络的每一个对抗性示例都可能导致有害的后果。此外，人们可能希望优先寻找对抗性的例子，而不是那些大大改变整个系统所需语义的例子。同样，构造健壮 ml 算法的现有算法忽略了整个系统的规范。本文认为，整个系统的语义和规范在这一研究领域发挥着至关重要的作用。我们提出了支持这一说法的初步研究结果。少

2018 年 5 月 18 日提交;v1 于 2018 年 4 月 19 日提交;最初宣布 2018 年 4 月。

334. 第: 1804.0770[pdf,其他] Cs。CI

用于可视化推理的双向匹配对象排序

作者:郝 tan, mohit bansal

摘要: 具有成分自然 **语言**指令的可视推理，例如，基于新发布的康奈尔**大学自然语言**可视推理 (nlvr) 数据集，是一项具有挑战性的任务，在这种情况下，模型需要能够在图像中放置在复杂排列中的多个对象之间创建准确的映射。此外，考虑到三个类似图像中对象的排序和关系，需要**处理**此映射以回答语句中的问题。本文针对 nlvr 任务提出了一种新的端到端神经模型，首先利用联合双向关注在视觉信息和**语言**短语之间建立双向条件。接下来，我们使用基于 rl 的指针网络对三个图像中每个图像中的无序对象的不同数量 (以便匹配语句短语的顺序) 进行排序和**处理**，然后在三个决策上进行池。我们的模型在数据集的结构化表示和原始图像版本上都实现了比最先进的改进 (4-6 的绝对值)。少

2018 年 9 月 6 日提交;v1 于 2018 年 4 月 18 日提交;最初宣布 2018 年 4 月。

评论:naacl 2018 (8 页; 增加了点对点订购的例子)

335. 第十四条: 1804. 05514[pdf,其他] Cs。DI

cl 学者: acl 选集知识图

作者:mayank singh, pradeep dogga, sohan patroo, diraj barnwal, ritam dutt, rajarshi haldar, pawan goyal, animesh mukherjee

文摘: 我们提出了 cl 学者, acl 选集知识图矿工, 以方便高质量的搜索和探索当前的研究进展, 在计算语言学界。与以前的工作不同, 在当前系统中, 定期对新传入文章进行爬网、索引和**处理**是完全自动化的。cl 学者利用文本信息和网络信息进行知识图构建。作为另一项新的举措, cl 学者支持超过 1200 学术**自然语言**查询, 以及基于关键字的构建知识图的标准搜索。它回答二进制、统计和基于列表的**自然语言**查询。当前系统部署在 http://cnerg.iitkgp.ac.in/aclakg。我们还提供 rest api 支持以及批量下载功能。我们的代码和数据可在 https://github.com/CLScholar。少

2018 年 4 月 16 日提交;最初宣布 2018 年 4 月。

评论:5 页

336. 第 1804. 05017 [pdf,其他] Cs。Cl

将词典纳入深层神经网络促进我国临床命名实体识别

作者:王琪,夏玉航,周阳明,唐然,高大奇, 何平

摘要: 临床命名实体识别 (cner) 旨在识别和分类临床术语, 如疾病, 症状, 治疗, 检查, 和身体部位在电子健康记录, 这是临床和翻译研究的一项基本和关键的任务。近年来, 深部神经网络在命名实体识别和许多其他**自然语言处理**(nlp) 任务方面取得了显著的成功。这些算法大多是端到端训练的, 可以自动从大规模标记的数据集中学习要素。但是, 这些数据驱动的方法通常缺乏**处理**稀有或看不见的实体的能力。以往的统计方法和特征工程实践表明, 人类知识可以为处理罕见和看不见的情况提供有价值的信息。本文通过将词典纳入汉语 c 纳任务的深度神经网络来解决这一问题。提出了两种扩展双向长期短期存储器 (双 lstm) 神经网络的体系结构和五种不同的特征表示方案来处理该任务。ccks-2017 task 2 基准数据集的计算结果表明, 与最先进的深度学习方法相比, 该方法具有极高的竞争性能。少

2018 年 4 月 13 日提交;最初宣布 2018 年 4 月。

评论:21 页, 6 个数字

337. 第 1804 04800 [pdf, ps,其他] si

从安全论坛中挖掘可操作的信息: 恶意 ip 地址的情况

作者:joobin gharibshah, taiching li, andre castro, konstantinos pelechrinis , evangelos e. papalexakis, michalis faloutsos

摘要: 这项工作的目的是系统地从黑客论坛中提取信息, 这些论坛的信息一般被描述为非结构化的: 帖子的文本不一定遵循任何写作规则。相比之下, 许多安全举措和商业实体正在利用现成的公共信息, 但它们似乎侧重于结构化的信息来源。在这里, 我们将重点关注在论坛中报告的 ip 地址中识别恶意 ip 地址的问题。我们开发了一种自动识别恶意 ip 地址的方法, 其设计目标是独立于外部源。一个关键的新特点是, 我们使用矩阵分解方法来提取用户行为信息的潜在特征, 并与相关帖子中的文本信息相结合。我们的技术的一个关键设计特点是, 它可以很容易地应用于不同的**语言**论坛, 因为它不需要一个复杂的**自然语言处理**方法。特别是, 我们的解决方案只需要新**语言**中的少量关键字以及特定功能捕获的用户行为。我们还开发了一个工具来自动收集安全论坛的数据。使用我们的工具, 我们从 3 个不同的论坛收集大约 600k 的帖子。该方法具有较高的分类精度, 而在所有三个论坛中, 在帖子中识别恶意 ip 的精度均大于 88%。我们认

为，我们的方法可以提供显著更多的信息：我们发现多达 3 倍的潜在恶意 ip 地址相比，参考黑名单 VirusTotal。随着网络战争越来越激烈，提前访问有用信息变得更加必要，以消除黑客的第一行动优势，我们的工作是朝着这个方向迈出的坚实一步。少

2018 年 4 月 13 日提交;最初宣布 2018 年 4 月。

评论:10 页

338. 第 xiv:1804.04589[pdf, ps,其他] Cs。CI

基于神经网络的摘要方法综述

作者:岳东

摘要: 自动文本摘要是在保留文档主要思想的同时缩短文本的自动化**过程**，是**自然语言处理**中的一个重要研究领域。本文献综述的目的是考察近年来在自动文本总结中基于神经元的模型的研究。我们详细研究了十个最先进的基于神经的总结器: 5 个抽象模型和 5 个萃取模型。此外，我们还讨论了可应用于总结任务的相关技术，并为今后基于神经的总结研究提供了有希望的途径。少

2018 年 3 月 19 日提交;最初宣布 2018 年 4 月。

评论:16 页, 4 张表格

339. 第 xiv:1804.04212[pdf,其他] Cs。红外

应用于推荐的 word2vec：超参数问题

作者:hugo caselles-dupré, florian lesaint, jimena royo-letelier

摘要: 带有负采样的 skip-ram 是 word2vec 的一个流行变体，最初设计和调整用于**创建自然语言处理**的单词嵌入，已被用于创建项目嵌入与成功的应用程序在推荐中。虽然这些字段不共享相同类型的数据，也不对相同的任务进行评估，但建议应用程序倾向于使用相同的已调整的超参数值，即使最佳超参数值通常已知为依赖于数据和任务。因此，我们通过对各种数据集进行大型超参数网格搜索，研究了每个超参数在推荐设置中的边际重要性。结果表明，优化忽略的超参数，即负采样分布、时代数、子采样参数和窗口大小，可显著提高推荐任务的性能，并可按大小。重要的是，我们发现**自然语言处理**任务和推荐任务的最佳超参数配置明显不同。少

2018 年 8 月 29 日提交;v1 于 2018 年 4 月 11 日提交;最初宣布 2018 年 4 月。

评论:本论文发表于 2018 年 10 月 2 日至 7 日在加拿大温哥华举行的第 12 届推荐人推荐人会议上

340. 第 xiv:1804.04177[pdf,其他] Cs。铬

使用深层神经网络检测恶意 powershell 命令

作者:danny hendler, shay kels, amir rubin

摘要: 微软的 powershell 是一种命令行外壳和脚本**语言**，默认情况下安装在 windows 计算机上。虽然管理员可以配置 powershell 以限制访问和减少漏洞，但可以绕过这些限制。此外, powershell 命令可以很容易地动态生成，从内存执行，编码和模糊，从而使 powershell 执行的代码的日志记录和取证分析具有挑战性。由于所有这些原因, powershell 越来越多地被网络犯罪分子用作其攻击工具链的一部分，主要用于下载恶意内容和横向移动。事实上，赛门铁克最近提交的一份专门针对 powershell 被网络卷曲滥用的全面技术报告报告说，他们收到的恶意 powershell 样本数量急剧增加，使用的渗透工具和框架数量急剧增加。动力壳牌。这突出表明迫切需要开发有效的方法来检测恶意 powershell 命令。在这项工作中，我们通过实现几种新型恶意

powershell 命令的检测器并评估其性能来应对这一挑战。我们实现了基于字符级卷积神经网络 (cnn) 的基于 "传统" 自然**语言处理**(nlp) 的探测器和探测器。探测器的性能是使用大型真实世界数据集进行评估的。我们的评估结果表明，尽管我们的探测器单独产生高性能，但将基于 nlp 的分类器与基于 cnn 的分类器相结合的集成探测器提供了最佳性能，因为后一种分类器能够检测到成功规避前一项命令的恶意命令。我们对这些回避命令的分析表明，cnn 分类器自动检测到的一些模糊模式在本质上很难使用我们应用的 nlp 技术来检测。少

2018 年 4 月 14 日提交;v1 于 2018 年 4 月 11 日提交;最初宣布 2018 年 4 月。

评论:19 页，5 个数字

341. **第 xiv:1804. 04087[pdf,其他] Cs。CI**

lstm 生成文本的自然语言统计特征

作者:marco lippi，marcelo a montemurro，mirko degli esosti，giampaolo cristadoro

摘要: 长期短期记忆 (lstm) 网络最近在处理**自然语言**生成的几项任务中表现出了显著的表现，例如图像字幕或诗歌创作。然而，只有少数作品分析了 lstm 生成的文本，以便定量地评估这种人工文本在多大程度上与人类生成的文本相似。我们将 lstm 生成的**语言**的统计结构与书面**自然语言**的统计结构以及各种顺序的马尔可夫模型所产生的统计结构进行了比较。特别是，我们通过评估词频统计、远程相关性和熵度量来描述**语言**的统计结构。我们的主要发现是，虽然 lstm 和 markof 生成的文本在其文字频率统计和熵度量中都能表现出类似于真实的特征，但 lstm 文本的维度与**自然**的**语言**。此外，对于 lstm 网络，控制生成**过程**的类似温度的参数显示了一个最佳值--生成的文本最接近于真实语言--在所有不同的统计**中**保持一致进行调查的特征。少

2018 年 4 月 10 日提交;最初宣布 2018 年 4 月。

342. **第 xiv:1804. 03673[pdf，ps,其他] Cs。CI**

数字文本分析的深度学习：情感分析

作者:reshma u，barathi ganeshh b，mandar kale，prachi mankame，gouri kulkarni

摘要: 在今天的场景中，想象一个没有消极的世界是非常不现实的，因为坏的新闻比好的新闻传播得更全面。虽然在现实生活中似乎不切实际，但这可以通过构建一个使用机器学习和**自然语言处理**技术的系统来实现，该系统使用负阴影识别新闻数据并通过以下方式对其进行筛选。只拿正面阴影 (好消息) 的新闻给最终用户。在这项工作中，大约有两个 lakhs 基准已被训练和测试使用的组合基于规则和数据驱动的方法。继使用文档术语矩阵 (表示) 和支持向量机 (分类) 的统计机器学习方法之后，vader 和过滤方法被用作注释工具。然后，深度学习算法进入图片，使这一系统可靠 (doc2vec)，最终与卷积神经网络 (cnn)，产生更好的结果比其他实验模块。训练准确率为 96%，测试准确率 (内外新闻数据) 超过 85%。少

2018 年 4 月 10 日提交;最初宣布 2018 年 4 月。

评论:8 页

msc 类: 68t50

343. **建议：1804. 03562[pdf] cs. cy**

多伊 10.1016/j.compenvurbsys.2018.01.010

大企业注册数据估算：支持我国产业的时空分析

作者:李法鹏,桂志鹏,吴华谊,龚建亚,王元,田思宇，张家文

摘要: 包含时间和位置信息的大、细粒度企业注册数据使我们能够对跨时间和空间的多个尺度上的行业模式进行定量分析、可视化和了解。但是，数据质量问题 (如不完整和歧义) 阻碍了此类分析和应用。当数据量巨大且不断增长时，这些问题就变得更具挑战性。高性能计算 (hpc) 框架可以解决大数据计算问题，但很少有研究机构系统地研究这种计算环境下企业注册数据的估算方法。本文提出了一种基于 apache spark 的大数据估算工作流程以及裸机计算集群，对企业注册数据进行了估算。我们集成了外部数据源，使用**了自然语言处理**(nlp)，并比较了几种机器学习方法，以解决企业注册数据中发现的不完整和歧义问题。实验结果说明了所提出的基于 hpc 的估算框架的可行性、有效性和可扩展性，也为其他大型地理参考文本**数据处理**提供了参考。利用这些估算结果，我们对中国工业的时空分布进行了可视化和简要讨论，展示了这些数据在质量问题得到解决时的潜在应用。少

2018 年 5 月 21 日提交;v1 于 2018 年 4 月 5 日提交;最初宣布 2018 年 4 月。

评论:15 页, 15 位数字

期刊参考:https://www.sciencedirect.com/science/article/pii/S0198971517302971, 2018 年

344. 第 1804.03540[pdf, ps,其他] Cs。CI

挖掘社交媒体进行新闻收集

作者:阿尔凯茨·祖比亚加

摘要: 社交媒体正在成为了解和跟踪突发新闻的日益重要的数据来源。这得益于连接到互联网的移动设备，它允许任何人从任何地方发布更新，进而导致公民新闻的日益存在。因此，社交媒体已成为记者在新闻收集过程中的首选资源。然而，利用社交媒体进行新闻收集具有挑战性，需要适当的工具，以便利获得有用的报道信息。本文综述了用于挖掘社交媒体的数据挖掘和**自然语言处理**的研究。我们讨论了研究人员为缓解社交媒体新闻收集所固有的挑战而开展的七项不同任务: 事件检测、摘要、新闻推荐人、内容验证、查找信息来源、开发新闻收集仪表板和其他任务。我们概述了迄今在这一领域取得的进展，总结了当前的挑战，并讨论了使用计算新闻协助社交媒体新闻收集的未来方向。本调查论文与研究社交媒体新闻的计算机科学家以及对计算机科学与新闻交叉感兴趣的跨学科研究人员有关。少

2018 年 4 月 10 日提交;最初宣布 2018 年 4 月。

345. 建议: 1804.03124[pdf,其他] Cs。CI

利用用户内部和用户间表示学习实现仇恨语音自动检测

作者:钱静,麦·埃尔谢利夫,伊丽莎白 m.

文摘: 仇恨语音检测是**自然语言处理**(nlp) 中一个关键但具有挑战性的问题。尽管有许多专门研究开发 nlp 仇恨言论检测方法，但准确性仍然很低。核心问题是，社交媒体帖子简短嘈杂，大多数现有的仇恨言论检测解决方案将每个帖子视为一个孤立的输入实例，这很可能产生高假阳性和负率。在本文中，我们通过提出一种新的模型，利用用户内部和用户间的表示学习，在 twitter 上进行强大的仇恨语音检测，从根本上改进了自动仇恨语音检测。除了目标推特之外，我们还收集和分析用户的历史帖子，以模拟用户内部的推文表示。为了抑制单个推文中的噪音，我们还使用强化的用户间表示学习技术对所有其他用户发布的类似推文进行建模。实验表明，利用这两种表示可以显著提高一个强的双向 lstm 基线模型的 f-分数 10.1%。少

2018 年 9 月 13 日提交;v1 于 2018 年 4 月 9 日提交;最初宣布 2018 年 4 月。

346. 第 xiv:1804.02956[pdf] cse

面向可重复的研究：经验需求工程论文的自动分类

作者:克林顿·伍德森,简·赫夫曼·海斯,萨拉·格里菲斯

文摘: 研究必须是可复制的，以便对科学产生影响，并为我们领域的知识体系做出贡献。然而，研究表明，来自学术实验室的研究有 70% 无法复制。在软件工程中，更具体地说，对工程 (re) 的要求，可重复研究很少见，数据集并不总是可用的，或者方法没有得到充分描述。这种缺乏可复制的研究阻碍了进展，研究人员不得不从零开始复制实验。从 re 开始的研究人员必须筛选会议论文，找到经验性的会议论文，然后必须查阅经验论文 (如果有的话) 中的数据，以初步确定论文是否可以复制。本文讨论了这一问题的两个部分，即识别可再生能源文件和识别可再生能源纸张中的经验文件。最近的可再生能源和经验会议论文被用来学习特征，并建立一个自动分类器来识别可再生能源和经验论文。我们引入了经验需求研究分类器 (erc) 方法，该方法利用**自然语言处理**和机器学习对会议文件进行监督分类。我们将我们的方法与基于基准关键字的方法进行比较。为了评估我们的方法，我们检查 ieee 需求工程会议和 ieee 软件测试和分析国际研讨会上的一组论文。我们发现，除少数情况外，erc 方法在所有情况下的表现都优于基线方法。少

2018 年 4 月 9 日提交;最初宣布 2018 年 4 月。

评论:这项工作得到了国家安全基金赠款 ccf-1511117 的部分支持; 7 页

347. 第 xiv:1804.02816[pdf, ps,其他] cs. ne

多伊 10.1109/SMC.2015.465

利用受限玻尔兹曼机在克隆选择算法中生成免疫记忆的一种方法

作者:shinkamada, takumi ichimura

文摘: 近年来，实时提取图像特征需要一种较高的图像**处理**技术。在我们的研究中，旅游主题数据是从基于手机的参与式传感 (mpps) 系统中收集的。每条记录都包含带有 gps 的图像文件、地理位置名称、用户的数字评估以及在用户真正访问的观光景点以**自然语言**编写的注释。在我们之前的研究中，利用具有免疫记忆细胞 (csaim) 的克隆选择算法可以探测到观光景点的著名地标。但是，以前的方法没有正确检测到某些地标，因为它们没有足够的信息来提取特征。为了改善这一弱点，我们提出了由限制玻尔兹曼机器产生免疫记忆的方法。为了验证该方法的有效性，利用机器学习工具对主观数据进行了分类实验。少

2018 年 4 月 9 日提交;最初宣布 2018 年 4 月。

评论:6 页, 10 个数字, 2015 年 ieee 系统、人和控制论国际会议 (ieee smc 2015) 的产品目录

348. 第 xiv:1804.02042[pdf,其他] Cs。CI

semval-2018 任务 7 中的 eth-ds3lab：将循环神经网络和卷积神经网络有效地结合起来进行关系分类和提取

作者:jonathan rotsztejn, nora hollenstein, e zhang

摘要: 在非结构化文本中可靠地检测实体之间的相关关系是知识提取的宝贵资源，这也是它在**自然语言处理**领域引起人们极大兴趣的原因。本文提出了一个基于卷积和递归神经网络的关系分类和提取系统，该系统在 2018 年 semval 任务 7 的 4 个子任务中

排名第 3 位。我们为最相关功能背后的设计选择提供详细的解释和理由，并分析其重要性。少

2018 年 4 月 5 日提交;最初宣布 2018 年 4 月。

评论:参加 2018 年学期 (第 12 届语义评估国际研讨会)

日记本参考:第 12 届语义评价国际研讨会论文集, 2018 年，计算语言学协会, 689-696 页, http://aclweb.org/anthology/S18-1112

349. **第 xiv:1804.01772[pdf, ps,其他] Cs。CI**

不仅仅是规模--分布式词语表示在科学出版物分析中的作用研究

作者:安德烈斯·加西亚,何塞·曼努埃尔·戈麦斯-佩雷斯

摘要: 学术传播领域的知识图的出现以及人工智能和**自然语言处理**方面的最新进展使我们更接近于智能系统可以提供帮助的场景科学家在一系列知识密集型任务。本文介绍了学术出版物中用于从科学图中提取的科学文本的智能**处理**的单词嵌入生成的实验结果。我们将特定于域的嵌入的性能与从非常大的通用语料库生成的现有预训练载体进行比较。我们的研究结果表明，语料库的特异性和体积之间存在着一种权衡。来自特定于域的科学语料库的嵌入有效地捕获了域的语义。另一方面，通过一般语料库获得可比结果也是可以实现的，但前提是存在着非常大的形成良好的文本语料库。此外，我们还表明，知识领域之间的重叠程度与域评估任务中嵌入的性能直接相关。少

2018 年 4 月 5 日提交;最初宣布 2018 年 4 月。

350. **第 1804.01653[pdf] Cs。Lg**

深度学习回顾

作者:张荣,李伟平,唐莫

摘要: 近年来，中国、美国等国家、谷歌等高科技公司加大了对人工智能的投资。深度学习是当前人工智能研究的重点领域之一。本文分析和总结了深度学习的最新进展和未来的研究方向。首先，概述了深度学习的三个基本模型，包括多层感知器、卷积神经网络和递归神经网络。在此基础上，进一步分析了卷积神经网络和递归神经网络的新模型。然后总结了深度学习在人工智能许多领域的应用，包括语音**处理**、计算机视觉、**自然语言处理**等。最后，本文讨论了深度学习中存在的问题，并给出了相应的可能解决方案。少

2018 年 8 月 28 日提交;v1 于 2018 年 4 月 4 日提交;最初宣布 2018 年 4 月。

评论:在中文。已发表在《信息与控制》杂志上

351. **第 1804.01622[pdf,其他] Cs。简历**

场景图中的图像生成

作者:贾斯汀·约翰逊,阿格里姆·古普塔,李飞飞

摘要: 为了真正了解视觉世界，我们的模型不仅应该能够识别图像，还应该能够生成图像。为此，最近在从**自然语言**描述中生成图像方面取得了令人激动的进展。这些方法在有限的领域 (如鸟类或花朵的描述) 上给出了惊人的结果，但却难以忠实地再现具有许多对象和关系的复杂句子。为了克服这种限制，我们提出了一种从场景图生成图像的方法，从而能够显式地推理对象及其关系。我们的模型使用图形卷**积来处理**输入图，通过预测对象的边界框和分割掩码来计算场景布局，并将布局转换为具有级联细化网络的图像。该网络通过对一对歧视者的对抗训练，以确保逼真的输出。我们验证了我们关

于视觉基因组和 COCO-Stuff 的方法，在这种方法中，定性结果、消融和用户研究证明了我们的方法能够生成具有多个对象的复杂图像。少

2018 年 4 月 4 日提交;最初宣布 2018 年 4 月。

评论:将出席 2018 年 cvpr 会议

352. 第十四条: 1804. 01486[pdf,其他] Cs。CI

从多模医学数据的大量来源中学习的临床概念嵌入

作者:andrew l. beam, benjamin kmpa, ingbar fred , nathan p.palmer, xu shi, tixi cai, isaac s. kohane

摘要: 单词嵌入是一种流行的方法，可以在不受监督的情况下学习在**自然语言处理**中广泛使用的单词关系。在本文中，我们提出了一套新的嵌入医学概念学习使用一个非常大的多式联运医疗数据集合。根据最近的理论见解，我们展示了如何将由 6000 万成员组成的保险索赔数据库、2000 万临床笔记和 170 万篇全文生物医学期刊文章结合起来，将概念嵌入到一个共同的空间中,为 108, 477 个医学概念提供了有史以来最大的一套嵌入。为了评估我们的方法，我们提出了一种新的基于统计能力的基准方法，专门用于测试医学概念的嵌入。我们的方法被称为 cui2vec，在大多数情况下，与以前的方法相比，获得了最先进的性能。最后，我们提供了一套可下载的预培训嵌入，供其他研究人员使用，以及一个在线工具，用于交互式探索 cui2vec 嵌入。少

2018 年 5 月 18 日提交;v1 于 2018 年 4 月 4 日提交;最初宣布 2018 年 4 月。

353. 第 1804.01189[pdf,其他] Cs。Sy

分配系统中断持续时间的实时预测

作者:aaron jaech, baosen zhang , mari ostendorf , daniel s. kirschen

摘要: 本文利用历史中断记录对一系列神经网络预测值进行了训练，解决了预测计划外停电持续时间的问题。根据环境因素进行初始持续时间预测，并根据传入的现场报告进行更新，使用**自然语言处理**自动分析文本。使用 15 年中断记录的实验显示了良好的初始结果和利用文本的性能改进。案例研究表明,**语言处理**标识指向停机原因和修复步骤的短语。少

2018 年 7 月 29 日提交;v1 于 2018 年 4 月 3 日提交;最初宣布 2018 年 4 月。

评论:出现在电力系统上的 ieee 事务

354. 第 xiv: 18004.00968[pdf] Cs。CI

基于卷积神经网络的深度问题分类

作者:prudhvi raj Dachapally, srikanth ramanam

摘要: 用于计算机视觉的卷积神经网络是相当直观的。在图像分类中使用的典型 cnn 中，第一层学习边缘，下面的图层学习一些可以识别对象的过滤器。但 cnn 自然**语言处理**并不经常使用，也不是完全直观。我们对卷积滤波器为文本分类任务学习的内容有了很好的了解，为此，我们提出了一个能够在更短的时间内产生良好结果的神经网络结构。我们将使用卷积神经网络来预测问题的主要或更广泛的主题，然后对每个预测的主题使用单独的网络来准确地对它们的子主题进行分类。少

2018 年 3 月 31 日提交;最初宣布 2018 年 4 月。

评论:4 页，短纸

355. 第 xiv:804.00832[pdf,其他] Cs。CI

应序列到序列学习对语文本的音符性恢复

作者:iroro orife

摘要: Yorùbá是一种广泛使用的西非**语言**，其书写系统具有丰富的声调和正字法。除了极少数例外，由于设备和应用支持有限，电子文本中省略了变音。变音提供形态信息，对词汇消歧、发音至关重要，对任何约纳语文本到语音 (tts)、自动语音识别 (asr) 和**自然语言处理**都至关重要(nlp) 任务。将自动分形恢复 (adr) 重构为机器翻译任务，我们用两种不同的注意序列到序列神经模型对未变音文本进行**了处理**。在我们的评估数据集中，这种方法产生的变光错误率小于 5%。我们发布了预先培训的模型、数据集和源代码，作为一个开源项目，以推进 yoryba **语言**技术方面的努力。少

2018 年 10 月 29 日提交;v1 于 2018 年 4 月 3 日提交;最初宣布 2018 年 4 月。

评论:6 页, 3 个数字。讨论了 2018 年互动预印，并发表了额外的数字和评审者的评论

356. 第 xiv:804.00806[pdf] Cs。Cl

利用资源丰富语言的代码混合语言的情感分析

作者:nurendra choudhary, rajat singh, ishita Bindlish, manish shrivastava

文摘 代码混合数据是**自然语言处理**的一个重要挑战，因为它的特点与标准**语言**的传统结构完全不同。本文提出了一种新的方法--代码混合文本的情感分析 (sacmt)，利用对比学习将句子分为相应的情绪--正、负或中性。我们利用暹罗网络的共享参数将代码混合和标准**语言**的句子映射到一个共同的情感空间。此外，我们还介绍了一种基于聚类的预处理方法来捕获代码混合转写词的变化。我们的实验表明, sacmt 在代码混合文本的情绪分析中的表现优于最先进的方法，准确率为 7.6%, f-分数为 10.1%。少

2018 年 4 月 2 日提交;最初宣布 2018 年 4 月。

评论:在第 19 届计算语言学和智能文本处理国际会议上接受长纸, 2018 年 3 月，越南河内. arxiv 管理说明: 文本与 arxiv:804.008005 重叠

357. 第 xiv:18004.00401[pdf,其他] Cs。Db

数据库的端到端神经自然语言接口

作者:prasetya utama, nathaniel weir, fuat basik, carsten binnig, ugur cetintemel , benjamin hettasch, amir ilkhechi, shekarPrasetya, arif usta

摘要: 从新数据集中提取见解的能力对于决策至关重要。可视化交互工具在数据探索中发挥着重要作用，因为它们为非技术用户提供了直观地撰写查询和理解结果的有效方法。**天然**的**语言**作为数据库的替代查询接口，最近获得了吸引力，有可能使非专家用户能够高效和有效地制定复杂的问题和信息需求。然而，理解**自然语言**问题并将其准确地转换为 sql 是一项具有挑战性的任务，因此数据库的**自然语言**接口 (nlidb) 尚未进入实用工具和商业产品。本文提出了一种具有**自然语言**界面的新数据探索工具 dbpal。dbpal 利用深度模型的最新进展，通过以下方式使查询理解更加可靠: 首先, dbpal 使用深层模型将**自然语言**语句转换为 sql，从而实现**翻译过程**更有力的解释和其他语言的变化。其次，为了在不了解数据库架构和查询功能的情况下支持用户措辞问题，dbpal 提供了一个经验习得的自动完成模型，该模型在查询制定过程中向用户建议部分查询扩展，从而帮助编写复杂的问题查询。少

2018 年 4 月 2 日提交;最初宣布 2018 年 4 月。

358. 第 1804. 00079[pdf,其他] Cs。Cl

通过大规模多任务学习学习通用分布式句子表示

作者:sandeep subramanian, adam trischler, yeshua bengio, christopher j pal

摘要: 最近在**自然语言处理**(nlp) 方面取得的许多成功都是由以无监督方式在大量文本上训练的单词的分布式矢量表示所驱动的。这些表示形式通常用作跨一系列 nlp 问题的单词的通用功能。然而，将这一成功扩大到学习句子等单词序列的表示，仍然是一个悬而未决的问题。最近的工作探索了无监督和监督学习技术与不同的培训目标，以学习通用固定长度句子表示。在这项工作中，我们提出了一个简单，有效的多任务学习框架的句子表示，结合了各种训练目标的归纳偏差在一个单一的模型。我们在几个数据源上对这一模型进行培训，在 1 亿多句子上实现了多个培训目标。广泛的实验表明，在薄弱相关的任务中共享单个重复的句子编码器，与以前的方法相比，可以实现一致的改进。我们在使用我们学习到的通用表示形式的转移学习和低资源设置方面提出了实质性的改进。少

2018 年 3 月 30 日提交;最初宣布 2018 年 4 月。

评论:2018 年国际 lr 会议接受

359. 第 18 第三条. 11544[pdf,其他] Cs。简历

引导我: 与深度网络互动

作者:christian ruprecht, iro laina, nassir navab, gregory d.hager, fedico 雷莫 Tombari

摘要: 随着机器学习方法进入涉及最终用户的现实世界应用，人类和智能机器之间的交互和协作变得越来越重要。虽然以前的许多工作都在**自然语言**和视觉的交汇点上，比如图像字幕或从文本描述中生成图像，但对语言的使用却较少的关注 。学习的视觉**处理**算法的性能。本文探讨了通过用户输入灵活地引导训练的卷积神经网络以提高推理过程中性能的方法。我们通过在网络中插入一个充当空间语义指南的图层来实现这一目标。本指南经过培训，可以直接通过能量最小化方案，也可以通过将人类**语言**查询转换为交互权重的循环模型间接修改网络的激活。学习口头交互是全自动的，不需要手动文本注释。我们对两个数据集的方法进行了评估，表明指导预先培训的网络可以提高性能，并提供对指南与 cnn 之间的交互的广泛见解。少

2018 年 3 月 30 日提交;最初宣布 2018 年 3 月。

评论:cvpr 2018

360. 第 xiv:1803. 10609[pdf, ps,其他] Cs。Sd

第五个 "chime" 语音分离和识别挑战: 数据集、任务和基线

作者:jon barker, sh 真司 watanabe, emmanuel vincent, jan trmal

文摘: chime 挑战系列旨在通过促进语音和**语言处理**、信号**处理**和机器学习接口的研究，推进强大的自动语音识别 (asr) 技术。本文介绍了第 5 届 chime 挑战赛，该挑战赛考虑了远程多麦克风会话 asr 在实际家庭环境中的任务。语音材料是使用晚餐聚会场景获得的，它致力于捕获**代表自然**对话语音的数据，并由 6 个 kinect 麦克风阵列和 4 个双耳麦克风对记录。挑战的特点是单阵列轨道和多阵列轨道，对于每个轨道，将为侧重于远程麦克风捕获方面的鲁棒性的系统和试图解决任务所有方面的系统生成不同的排名包括会话**语言**建模。我们讨论了这一挑战的基本原理，并详细描述了数据收集过程、任务以及阵列同步、语音增强以及常规和端到端 asr 的基准系统。少

2018 年 3 月 28 日提交;最初宣布 2018 年 3 月。

361. 第 xiv:1803.09875[pdf,其他] Cs。红外

一种绕过推特 api 限制的网络刮擦方法

作者:a. hernandez-suarez, g. sanchez-perez, k. toscano-medina, v. martinez-hernandez, v.sanchez, h. perez-meana

摘要: 从社交网络中检索信息是许多数据分析领域的第一步，也是最基本的一步，如**自然语言处理**、情感分析和机器学习。重要的数据科学任务在历史数据收集上进行中继，以获得进一步的预测结果。最近的作品大多使用推特 api，这是一个收集公共信息流的公共平台，允许按时间顺序查询不超过三周的推特。在本文中，我们提出了一种新的方法来收集历史推特在任何日期范围内使用网络刮擦技术绕过 twitter api 限制。少

2018 年 3 月 26 日提交;最初宣布 2018 年 3 月。

362. 第 xiv:18009641 [pdf, ps,其他] Cs。CI

马拉雅拉姆语中可翻译词和原语的无监督分离

作者:deepak p

摘要: 区分内**语言**词和可翻译词是辅助涉及不同**自然语言**的文字**处理**任务的关键步骤。我们考虑了在马拉雅拉姆**语**中，在不受监督的情况下将可翻译单词与文本的母语单词分离的问题。概述了词干以外字符多样性的关键观察，提出了一种基于词根的词评优化方法。我们的方法依赖于使用概率分布的字符 n 克，这是与在迭代优化公式中的本土性测试一致的细化。通过实证评价，说明了我们的方法 dtim 为马拉雅拉姆的本土性评分提供了显著的改进，将 dtim 确立为任务的首选方法。少

2018 年 3 月 26 日提交;最初宣布 2018 年 3 月。

评论:10 页，第 14 届自然语言处理国际会议论文集，印度加尔各答，2017 年 12 月 18 日至 21 日

类:I.2。7

363. 第 xiv:1803. 09668 [pdf,其他] Cs。Lg

针对人工神经网络的无剪切攻击

作者:boussad addad, jerome kdjabachian , christophe meyer

摘要: 在过去的几年里，由于人工深层神经网络在计算机视觉、**自然语言处理**的许多机器学习任务中取得了巨大的成功，在人工智能领域取得了显著的突破、语音识别、恶意软件检测等。然而，它们极易受到容易制作的对抗性例子的影响。许多调查都指出了这一事实，并提出了不同的方法来制造攻击，同时对原始数据增加有限的扰动。到目前为止，已知的最可靠的方法是所谓的 c & amp; w 攻击 [1]。尽管如此，一个被称为功能挤压加上整体防御的对策表明，这些攻击的大部分可以被摧毁 [6]。在本文中，我们提出了一种新的方法，我们称之为中心初始攻击 (cia) 的优点是双重的: 首先，它通过构造保证最大扰动小于事先固定的阈值，而不进行裁剪**过程**,降低攻击的质量。其次，它对最近引入的防御措施 (如功能压缩、jpeg 编码，甚至针对投票组合的防御) 都是稳健的。虽然它的应用并不局限于图像，但我们使用 imagenet 数据集上当前最好的五个分类器来说明这一点，其中两个是针对攻击的强健进行的对手重新训练。在任何像素上的固定最大扰动仅为 1.5%，大约 80% 的攻击 (有针对性) 愚弄投票合奏防御，当扰动只有 6% 时，几乎 100% 的攻击。虽然这表明了抵御中情局攻击是多么困难，但本文的最后一节给出了一些限制其影响的指导方针。少

2018 年 3 月 28 日提交;v1 于 2018 年 3 月 26 日提交;最初宣布 2018 年 3 月。

评论:12 页

364. 第 1803.08966[pdf,其他] 反渗透委员会

结构化语言中机器人规划解释的反例

作者:陆峰, mahsa ghasemi, kai-wei chang, ufuk topcu

摘要: 已使用模型检查等自动化技术来验证基于马尔可夫决策**过程**(mdp) 的机器人任务计划模型，并生成可能有助于诊断需求违规情况的反例。但是，此类工件可能过于复杂，人类无法理解，因为反例的现有表示形式通常包括大量路径或复杂的自动机。为了帮助提高反例的可解释性，我们定义了一个可解释的反例的概念，其中包括一组结构化**的自然语言**句子，用来描述导致需求的机器人行为在机器人任务计划的 mdp 模型中违规。我们提出了一种基于混合整数线性规划的方法，用于生成最小、健全和完整的可解释的反例。通过对仓库机器人规划的案例研究，证明了该方法的有效性。少

2018 年 3 月 23 日提交;最初宣布 2018 年 3 月。

评论:接受在 2018 年机器人与自动化国际会议 (icra) 会议上发表

365. 第 xiv:18008896[pdf,其他] Cs。简历

视觉问题回答的端到端神经体系结构的显式推理

作者:somak aditya, ye 周 yyang , Aditya baral

摘要: 许多视觉和**语言**任务需要的是常识推理，而不仅仅是数据驱动的图像和**自然语言处理**。在这里，我们采用可视问题回答 (vqa) 作为一个示例任务，其中一个系统需要用**自然语言**回答有关图像的问题。目前最先进的系统试图使用深度神经架构来解决任务，并实现了很有希望的性能。然而，由此产生的系统一般是不透明的，它们难以理解需要额外知识的问题。本文在一组基于倒数神经网络的系统的基础上，提出了一个显式推理层。推理层能够在需要额外知识的情况下进行推理和回答问题，同时为最终用户提供可解释的接口。具体而言，推理层采用基于概率软逻辑 (psl) 的引擎对一篮子输入进行推理: 视觉关系、问题的语义解析以及来自 word2vec 和脸网的背景本体论知识。对在 vqa 数据集上生成的答案和关键证据谓词进行的实验分析验证了我们的方法。少

2018 年 3 月 23 日提交;最初宣布 2018 年 3 月。

评论:9 页, 3 个数字, aaai 2018

366. xiv:18008850[pdf] Cs。红外

临床注意事项中的自动特征生成检测手术部位感染

作者:沈飞辰, david w larson, james m. naessens, elizabeth b. habermann, h 宏方 liu, sunghwan sohn

摘要: 术后并发症 (pscs) 被称为偏离正常的术后疗程，并按严重程度和治疗要求分类。手术部位感染 (ssi) 是主要的 psc 之一，也是最常见的与卫生保健相关的感染，导致住院时间和费用增加。在这项工作中，我们评估了一种自动的方法，利用带有启发式的自动语言分析来检测 ssi，并与医学专家一起评估这些关键字，从而从临床叙事中生成词典 (即关键字特征)。为了进一步验证我们的方法，我们还使用自动生成的关键字对队列进行了决策树算法。结果表明，我们的框架能够从临床叙事中识别 ssi 关键字，并通过增加搜索查询来支持基于搜索的**自然语言处理**(nlp) 方法。少

2018 年 3 月 26 日提交;v1 于 2018 年 3 月 23 日提交;最初宣布 2018 年 3 月。

367. 第 1803.08793[pdf,其他] cse

利用递归神经网络探讨错误密码的自然性

作者:杰克·兰尚廷,高吉

摘要: 统计**语言**模型是强大的工具，已用于**自然语言处理**中的许多任务。最近，它们被用于其他顺序数据，如源代码。(ray 等人, 2015 年) 表明，可以训练 n-gram 源代码**语言模式**，并使用它通过与**语言**模型相关的熵确定 "非自然" 行，从而预测代码中的错误行。在这项工作中，我们建议使用一种更先进的**语言**建模技术，长期短期记忆递归神经网络，建模源代码，并根据熵对错误线进行分类。在使用 auc 的错误行分类任务中，我们的方法略胜出 n-gram 模型。少

2018 年 3 月 21 日提交;最初宣布 2018 年 3 月。

368. 建议: 1803.08314[pdf,其他] Cs。简历

显示、告诉和区分: 通过使用部分标记数据的自检索进行图像字幕

作者:刘锡辉,李洪生,邵静,陈大鹏,王晓刚

摘要: 图像字幕的目的是通过机器生成描述图像内容的字幕。尽管做出了许多努力，但为图像生成歧视性字幕仍然不是微不足道的。大多数传统的方法模仿**语言**结构模式，因此往往陷入复制频繁短语或句子的刻板印象，忽视了每个图像的独特方面。在这项工作中，我们提出了一个图像字幕框架，以自我检索模块作为培训指导，鼓励生成判别字幕。它具有独特的优势: (1) 自检索指导可以作为标题判别的度量和评价，以保证生成字幕的质量。(2) 生成的字幕和图像之间的对应关系**自然**地包含在生成**过程中**，而没有人工注释，因此我们的方法可以利用大量未标记的图像来提升字幕性能，无需额外的费力注释。我们展示了所提出的检索引导方法在 coco 和 flickr30k 字幕数据集上的有效性，并以更有鉴别性的字幕显示了其优越的字幕性能。少

2018 年 7 月 22 日提交;v1 于 2018 年 3 月 22 日提交;最初宣布 2018 年 3 月。

评论:被 eccv 2018 接受

369. 第 1803.08193[pdf,其他] lo c

非决定论的认识论

作者:adam bjorndahl

摘要: 本文提出了非确定性程序执行的新语义，取代了命题动态逻辑 (pdl) 的标准关系语义。在这些新的语义下，程序执行被表示为从根本上确定性 (即功能性)，而非确定性则作为代理和系统之间的认知关系出现: 直观地说，给定的非确定性结果**过程**正是那些不能事先排除的。我们使用拓扑和动态拓扑逻辑 (dtl) 框架将这些概念形式化。我们证明了 dtl 可以用来解释 pdl 的**语言**，以捕获上面的直觉的方式，而且这个设置中的连续函数与确定性过程完全对应.我们还证明了对动态拓扑模型的相应类别, pdl 的某些公理化仍然是正确和完整的。最后，我们扩展了使用子集空间逻辑机制的知识整合的框架，并表明公共公告的拓扑解释与测试程序的**自然**解释完全吻合。少

2018 年 3 月 21 日提交;最初宣布 2018 年 3 月。

评论:21 页, 2 个数字

370. 第 xiv:1803.07724[pdf,其他] Cs。Cl

注意: 视觉问题回答 (vqa) 的体系结构

作者:jasdeep singh, vincent ying, alex nutkiewicz

摘要: 视觉问题回答 (vqa) 是深度学习研究中越来越流行的话题，需要**将自然语言处理**和计算机视觉模块协调成一个单一的体系结构。我们在 vqa 挑战中排名第一的模型的基础上，开发了 13 个新的关注机制，并引入了一个简化的分类器。我们进行了 300

个 gpu 小时的大量超参数和体系结构搜索，实现了 64.78 的评估分数，超过了现有最先进的单一模型的验证分数 64.78。少

2018 年 3 月 20 日提交;最初宣布 2018 年 3 月。

评论:视觉问题回答项目

msc 类: 68txx

371. 第 1803. 07292[pdf, ps,其他] cse

多伊 10.114/3393983196444

自然语言与否 (nln)-软件工程文本分析管道的软件包

作者:mika v. mäntylä, fabio calefato, maelick claes

文摘: 自然语言处理(nlp) 的使用在软件工程中越来越受欢迎。为了正确执行 nlp，我们必须**预处理**文本信息，以便将**自然语言**与其他信息 (如日志消息) 分开，而日志消息通常是软件工程中通信的一部分。我们提出了一个简单的方法来分类某些文本输入是否自然**语言**。虽然我们的 nlon 软件包只依赖于 11 个**语言**功能和字符三图，但我们能够在三个不同的数据源上实现 0.976-987 之间的 roc 曲线性能区域，而 lasso 从 glmnet 返回作为我们的学习者和两个人类的评价者提供地面真相。跨源预测性能较低，波动较大，最高中华民国性能从 0.913 到 0.913 不等。与以前的工作相比，我们的方法提供了类似的性能，但更轻，使其更容易应用于软件工程文本挖掘管道。我们的源代码和数据是作为 r 包提供的，以便进一步改进。少

2018 年 3 月 20 日提交;最初宣布 2018 年 3 月。

日记本参考:msr ' 18:15 国际采矿软件存储库会议, 2018 年 5 月 28 日至 29 日，瑞典哥德堡

372. 第: 1803.07179[pdf,其他] Cs。简历

基于关注的动作识别时间加权卷积神经网络

作者:曾金良,王乐,刘子义，张启林，牛振兴, 华刚, 郑南宁

摘要: 自卷积神经网络 (cnn) 等强大的机器学习工具引入以来，人类行动识别的研究显著加快。然而，最近的文献仍在积极探讨将时间信息纳入 cnn 的有效和高效方法。在**自然语言处理**研究领域流行的反复关注模型的推动下，提出了基于关注的时间加权 cnn (atw)，将视觉注意力模型嵌入到一个时间中加权多流美国有线电视新闻网。这种关注模型只是作为时间加权来实现的，但它有效地提高了视频表示的识别性能。此外，建议的 atw 框架中的每个流都能够进行端到端训练，网络参数和时间权都通过随机梯度下降 (sgd) 和反向传播进行优化。我们的实验表明，所提出的注意机制通过关注更相关的视频片段，极大地促进了更具鉴别性的片段的性能提升。少

2018 年 3 月 19 日提交;最初宣布 2018 年 3 月。

评论:第十四届人工智能应用与创新国际会议 (aiai 2018), 2018 年 5 月 25 日至 27 日，希腊罗兹

373. 第: 1803. 07136[pdf,其他] Cs。CI

具有递归量化分析的动态自然语言处理

作者:rick dale, nicolas d. duran, moreno coco

摘要: 写作和阅读是动态**的过程**。当作者组成文本时，会产生一系列的单词。作者希望，这个序列会导致对其他人的某些思想和想法的重新审视。读者的这些作曲**和复习过程**都是及时安排的。这意味着文本本身可以在动力系统的镜头下进行研究。一种常用的

分析动力系统行为的技术，称为递归量化分析 (rqa)，可作为分析文本顺序结构的一种方法。rqa 将文本视为顺序度量，很像时间序列，因此可以被视为一种动态**自然语言处理**(nlp)。扩展有几个好处。因为它是一组时间序列分析工具的一部分，所以可以在一个共同的框架中提取许多度量值。其次，这些措施与**自然语言处理**中一些常用的措施有着密切的关系。最后，利用递推分析通过开发从复杂动态系统中得出的理论描述，为文本的扩展分析提供了一个机会。我们展示了古腾堡项目的 8,000 个文本的示例分析，将其与众所周知的 nlp 方法进行了比较，并描述了一个 r 包 (crqanlp)，该包可与 r 库 crqa 一起使用。少

2018 年 3 月 19 日提交;最初宣布 2018 年 3 月。

374. **第: 1803.06397[pdf,其他] Cs。CI**

情感计算的深度学习：决策支持中基于文本的情感识别

作者:bernhard kratzwald, suzana ilic, mathias krous, stefan feuerriegel, helmut prendinger

摘要: 情绪广泛影响人类的决策。情感计算考虑到了这一事实，目的是根据个人的情绪状态量身定制决策支持。然而，由于**语言**的复杂性和模糊性，在叙述性文档中准确识别情感是一项具有挑战性的任务。通过深入学习可以提高性能;然而，正如本文所证明的，这项任务的具体**性质**要求在双向**处理**、作为正则化的形式的辍学率和加权损失方面定制经常性神经网络。功能。此外，我们还提出了一种量身定制的情感计算转移学习形式：在这里，网络是针对不同任务 (即情绪分析) 进行预训练的，而输出层随后被调整到情感识别的任务。由此产生的性能在 6 个基准数据集的整体设置中进行评估，我们发现，重复神经网络和转移学习的性能始终优于传统的机器学习。总之，这些发现对情感计算的使用有相当大的影响。少

2018 年 9 月 10 日提交;v1 于 2018 年 3 月 16 日提交;最初宣布 2018 年 3 月。

评论:被决策支持系统 (dss) 接受

375. **第 xiv:18005662[pdf, ps,其他] Cs。CI**

中国文学文本实体关系分类的结构正则化神经网络

作者:季文,孙旭,任宣城,苏琪

摘要: 关系分类是**自然语言处理**领域的一项重要语义**处理**任务。本文提出了中国文学文本的关系分类任务。构建了一个新的中国文学文本数据集，为这一任务的研究提供了便利。提出了一种新的模型，即结构正则化双向循环卷积神经网络 (sr-brcnc)，用于识别实体之间的关系。该模型沿着从结构正则化依赖树中提取的最短依赖路径 (sdp) 学习关系表示，具有降低整个模型复杂度的优点。实验结果表明，该方法可显著提高 f1 分数 10.3 分，优于中国文学文本的最新方法。少

2018 年 3 月 15 日提交;最初宣布 2018 年 3 月。

评论:接受在 naacl hlt 2018. arxiv 管理说明: 实质性文本重叠与 arxiv:1711. 02509

376. **第 18005526[pdf,其他] Cs。简历**

按语言旋转的未配对图像字幕

作者:顾九祥,沙菲克·乔蒂,蔡建飞,王刚

摘要: 图像字幕是一项涉及计算机视觉和**自然语言处理**的多模式任务，其目标是学习从图像到自然 **语言**描述的映射。通常，映射函数是从一组图像标题对的训练集中学习的。但是，对于某些**语言**，大规模图像标题配对语料库可能不可用。我们提出了一种通

过**语言**透视来解决这个未配对图像字幕问题的方法。我们的方法可以有效地从枢轴**语言**(中文) 中捕获图像隐藏器的特征，并使用另一个枢轴目标**(**汉英) 句子并行语料库将其与目标语言 (英语) 对齐。我们评估了两个图像到英语的基准数据集的方法: mscoco 和 flickr30k。与几种基线方法进行的定量比较证明了我们方法的有效性。少

2018 年 7 月 18 日提交;v1 于 2018 年 3 月 14 日提交;最初宣布 2018 年 3 月。

评论:17 页, 4 位数字, 2018 年 eccv 会议接受

377. 第 xiv:18004596[pdf] Cs。Cl

网络圣战仇恨语音的自动检测

作者:tom de smedt, guy de pauw, pieter van staeyen

摘要: 我们开发了一个系统，自动检测在线圣战仇恨言论超过 80% 的准确性，通过**使用从自然语言处理**和机器学习的技术。该系统接受了 2014 年 10 月至 2016 年 12 月收集的 45 000 条颠覆性推特信息的培训。我们对语料库中的圣战修辞进行了定性和定量分析，考察了推特用户网络，概述了用于培训系统的技术程序，并讨论了使用实例。少

2018 年 3 月 12 日提交;最初宣布 2018 年 3 月。

评论:31 页

报告编号:ctrs-007

日记本参考:clips 技术报告系列 7 (2018) 1-31

378. 第 xiv:18004469[pdf,其他] Cs。简历

生成对抗性网的图像合成简介

作者:何黄,余菲普,王长虎

文摘: 在过去的几年里，生成对抗性网 (gans) 的研究急剧增加。gan 于 2014 年提出，已应用于计算机视觉和**自然语言处理**等各种应用，并取得了令人印象深刻的性能。在 gan 的众多应用中，图像合成是研究最多的领域，这一领域的研究已经证明了 gan 在图像合成中的巨大潜力。本文提供了图像合成方法的分类，综述了文本到图像合成和图像转换的不同模型，讨论了一些评价指标以及图像合成中未来可能的研究方向。和甘在一起少

2018 年 3 月 12 日提交;最初宣布 2018 年 3 月。

379. 第 xiv:18004329[pdf,其他] Cs。Cl

将自然语言解析为 sparql 的语义分析: 用神经注意的方法提高目标语言的表示

作者:fabiano ferreira luz, marcelo finger

摘要: 语义解析是将**自然语言**句子映射为其意义的正式表示的**过程**。在本文中，我们使用神经网络方法将**自然语言**句子转换为 sparql **语言**中的本体数据库查询。此方法不依赖于手工规则、高质量词典、手动构建的模板或其他手工制作的复杂结构。我们的方法是基于向量空间模型和神经网络。该模型以两个学习步骤为基础。第一步生成自然 **语言**和 sparql 查询中的句子的矢量表示形式。第二步使用此矢量表示作为神经网络 (带有注意机制的 lstm) 的输入，生成能够对**自然语言**进行编码和解码 sparql 的模型。少

2018 年 3 月 12 日提交;最初宣布 2018 年 3 月。

380. 第 xiv:18000770[pdf,其他] Cs。Cl

反评级: 一种用于诈骗 ico 识别的深度学习系统

作者:bi 某,郑鹏 deng, fei li, will mon 露, pengshi, sunzijun, weiwu, sikuang wang, williamwang wang, arianna yuan,张天伟,李继伟

摘要: 加密货币 (或数字令牌、数字货币, 如 btc、eth、xrp、近地天体) 在使用、价值和公众理解方面迅速取得进展, 为投资者带来了惊人的利润。与其他货币和银行系统不同, 大多数数字令牌不需要中央当局。分散对信用评级构成重大挑战。大多数 ico 目前不受政府规定的约束, 这使得国际工商管理组织项目的可靠信用评级制度成为必要和紧迫。本文介绍了第一个基于学习的加密货币评级系统--反评级。我们利用**自然**语言 **处理**技术来分析到目前为止的 2, 251 种数字货币的各个方面, 如白皮书内容、创建团队、github 存储库、网站等。监督学习模型用于将加密货币的寿命和价格变化与这些特征联系起来。为了获得最佳设置, 所提出的系统能够精确地识别 0.83 精度的诈骗 ico 项目。我们希望这项工作将帮助投资者识别诈骗 ico, 并在自动评估和分析 ico 项目中吸引更多的努力。少

2018 年 3 月 8 日提交;最初宣布 2018 年 3 月。

381. **第 xiv:1800.03585[pdf,其他] Cs。Cl**

循环化在层次结构建模中的重要性

作者:ke tran, arianna bisazza, christof monz

摘要: 最近的研究表明, 当被训练来解决常见的**自然语言处理**任务 (如**语言**建模) 时, 递归神经网络 (rnn) 可以隐式捕获和利用分层信息。linzen 等人, 2016 年) 和神经机器翻译 (shi 等人, 2016 年)。相比之下, 使用非递归神经网络对结构化数据进行建模的能力尽管在许多 nlp 任务中取得了成功, 但却很少受到关注 (gehring 等人, 2017 年;vaswani 等人, 2017 年)。在这项工作中, 我们比较了这两种体系结构--经常性的和非递用的--它们对层次结构进行建模的能力, 并发现递归对于此目的确实很重要。少

2018 年 8 月 28 日提交;v1 于 2018 年 3 月 9 日提交;最初宣布 2018 年 3 月。

评论:emnlp 2018

382. **建议: 1803.03503[pdf, ps,其他] Cs。Lg**

实现局部深度学习的神经网络构建

作者:崔永元, 林少波, 周丁轩

摘要: 深度学习的主题最近吸引了来自不同学科的机器学习用户, 包括: 医学诊断和生物信息学、金融市场分析和在线广告、语音和笔迹识别、计算机视觉和**自然语言处理**、时间序列预测和搜索引擎。然而, 深度学习的理论发展仍处于起步阶段。本文的目的是引入一种深度神经网络 (也称为深网) 的局部流形学习方法, 每个隐藏层都有一个特定的学习任务。为了图解的目的, 我们只关注具有三个隐藏层的深网, 第一层用于降维, 第二层用于偏差缩小, 第三层用于减少方差。反馈组件还设计用于消除异常值。本文的主要理论成果是 O(米-2s / (2 秒)+ D))有规律性的回归函数的近似值 s, 根据数字米采样点, 其中 (未知) 流形维度 D 替换维度 D 浅层网的采样 (欧几里得) 空间。少

2018 年 3 月 9 日提交;最初宣布 2018 年 3 月。

评论:22 页

383. **第 xiv:1800.02994[pdf,其他] Cs。Cl**

意象如何激发诗歌: 用记忆网络从意象中生成中国古典诗歌

作者:徐林丽,梁江,川琴,王哲,东芳杜

摘要: 随着神经模型和**自然语言处理**的最新发展，中国古典诗歌的自动生成因其艺术和文化价值而备受关注。以往的作品主要集中在生成给定关键字或其他文本信息的诗歌，而视觉灵感的诗歌很少被探索。从意象中生成诗歌比从文本中生成诗歌更具挑战性，因为图像包含非常丰富的视觉信息，不能完全使用几个关键字来描述，一首好的诗应该准确地传达图像。本文提出了一种利用图像生成诗歌的基于记忆的神经模型。具体而言，提出了一种具有主题记忆网络的编码解码器解码器模型，从图像中生成中国古典诗歌。据我们所知，这是首次尝试利用神经网络的图像生成中国古典诗歌的作品。通过人的评价和定量分析的综合实验研究表明，该模型能够生成准确地传达图像的诗歌。少

2018 年 3 月 8 日提交;最初宣布 2018 年 3 月。

评论:2018 年获 aaai 接受

384. 第 xiv:1800272[8pdf,其他] Cs。CI

建立一个经鉴定的临床笔记的大语料库

作者:wye boag, tristan naumann, peter szolov 多彩的

摘要: 临床笔记通常描述患者生理的最重要的方面，因此对医学研究至关重要。然而，研究人员通常无法获得这些注释，而无需事先删除敏感的受保护健康信息 (phi)，这是一种被称为去识别的**自然语言处理**(nlp) 任务。需要自动取消识别临床笔记的工具，但如果无法访问那些包含 phi 的相同笔记，则很难创建这些工具。这项工作提出了第一步，以创建一个大型综合识别的临床笔记和相应的 phi 注释，以促进开发去识别工具。此外，还根据该语料库对其中一个工具进行了评估，以了解这种方法的优点和缺点。少

2018 年 3 月 7 日提交;最初宣布 2018 年 3 月。

385. 第 xiv:1800271[0pdf,其他] Cs。CI

生成矛盾、中性和内篇句子

作者:沈一康,谭少云，黄金伟,阿伦·考维尔

摘要: 学习分布式句子表示在**自然语言处理**(nlp) 领域仍然是一个有趣的问题。我们想要学习一个模型，它近似于给定语句的逻辑前置点的表示条件潜在空间。在我们的论文中，我们提出了一种生成句子的方法，该方法以输入句子和逻辑推理标签为条件。我们通过将输出句子的不同可能性建模为潜在表示的分布来做到这一点，我们使用对抗目标对其进行训练。我们使用两个最先进的模型来评估该模型，用于识别文本分配 (rte) 任务，并根据实际句子测量 BLEU 分数，作为我们模型产生的句子多样性的探针。实验结果表明，在我们的框架下，我们有明确的方法来提高生成句子的质量和多样性。少

2018 年 3 月 7 日提交;最初宣布 2018 年 3 月。

386. 第 xiv:1800232[9pdf,其他] Cs。Lg

学习内存访问模式

作者:milad hashemi, kevin swersky, jamie a. smith, grant ayers, heinerlitz, Jichuan chang, christos kozyrakis, partharathy ranganathan

摘要: 工作负载复杂性的激增和摩尔定律扩展的最近放缓要求采用高效计算的新方法。研究人员现在开始在软件优化、增强或取代传统的启发式和数据结构方面使用机器学习的最新进展。然而，计算机硬件体系结构的机器学习空间只是略有探索。在本文中，我们展示了深度学习解决冯·诺依曼内存性能瓶颈的潜力。我们关注学习内存访问模式的关键问题，目的是构建准确有效的内存预取器。我们将当代预取策略与自然 **语言处**

理中的 n-gram 模型联系起来，并展示了递归神经网络如何作为一种即时替代。在一套具有挑战性的基准数据集上，我们发现神经网络在精度和召回方面始终表现出卓越的性能。这项工作代表了向实用的基于神经网络的预取迈出的第一步，为计算机体系结构研究中的机器学习开辟了广泛的激动人心的方向。少

2018 年 3 月 6 日提交;最初宣布 2018 年 3 月。

387. 第 xiv:18002129[pdf, ps,其他] Cs。简历

深卷神经网络体系结构的非技术研究

作者:felix altenberger, claus lenz

摘要: 人工神经网络最近在许多学科和各种应用中取得了巨大的成果，包括**自然语言**理解、语音**处理**、游戏和图像数据生成。在其中，人工神经网络的强大性能被证明的一个特殊应用是图像中的对象识别，其中通常应用于深层卷积神经网络。在本次调查中，我们对这一主题 (具有深层卷积神经网络的对象识别) 进行了全面的介绍，重点介绍了网络体系结构的演变。因此，我们的目标是以简单和非技术性的方式压缩这一领域最重要的概念，以便未来的研究人员能够快速地对一般情况进行了解。本文的结构如下: 1. 阐述 (卷积) 神经网络和深度学习的基本思想，并考察它们在图像分类、对象定位和目标检测三个目标识别任务中的应用。2. 我们回顾了深层卷积神经网络的演变，对最重要的网络体系结构按时间顺序呈现的形式进行了广泛的概述。少

2018 年 3 月 6 日提交;最初宣布 2018 年 3 月。

评论:17 页 (包括参考资料), 23 个后记人物，使用 ieeetran

388. 第 18001384[pdf,其他] Cs。Db

深度学习的数据恢复 [愿景]: 实现自我驱动数据的实现

作者:saravanan thirumuruganathan, nan tang, mourad ouzzani

摘要: 过去。数据管理--发现、集成和清理数据的**过程**--是最古老的数据管理问题之一。不幸的是，这仍然是数据科学家最耗时和最不愉快的工作。到目前为止，成功的数据管理故事主要是特定于域 (例如, etl 规则) 或特定于任务的临时解决方案 (例如，实体解析)。目前。当前数据管理解决方案的力量在数量、速度、多样性和准确性方面跟不上不断变化的数据生态系统，主要原因是提供上述临时解决方案所需的人力成本较高，而不是机器成本以上。同时，深度学习在图像识别、**自然语言处理**、语音识别等领域取得显著成功方面也取得了长足进步。这主要是由于它能够理解既不特定于领域也不特定于任务的功能。未来。数据管理解决方案需要跟上快速变化的数据生态系统的步伐，在快速变化的数据生态系统中，主要的希望是设计领域无关和任务无关的解决方案。为此，我们启动了一个名为 autodc 的新研究项目，以释放深度学习自驾游数据策划的潜力。我们将讨论如何调整和扩展不同的深度学习概念，以解决各种数据管理问题。我们展示了一些关于在 autodc 中发生的深度学习和数据管理之间的早期接触的低垂成果。我们相信，这项工作所指出的方向不仅将推动 autodc 实现数据管理民主化，而且还将成为研究人员和从业人员转向数据管理解决方案新领域的基石。少

2018 年 3 月 4 日提交;最初宣布 2018 年 3 月。

389. 第 xiv:18001335[pdf,其他] Cs。Cl

caesar: 启用上下文感知摘要关注读取器

作者:陈龙辉, kshitiz tripathi

摘要: 理解**自然语言**的意义是 **自然语言处理**(nlp) 的主要目标，而文本理解是实现这一目标的基石，所有其他语言都是在这一目标上实现的。可以解决聊天机器人、**语言**翻译等问题。我们报告了一个摘要关注阅读器，我们设计的目的是更好地模拟人类阅读**过程**，以及基于字典的解决方案，涉及数据中的词汇外 (oov) 单词，以产生基于机器理解阅读的答案从 squad 基准的段落和问题。我们通过两个流行的模型 (匹配 lstm 和动态 coco 执着) 实现这些功能，能够接近于与从人类获得的结果相匹配。少

2018 年 3 月 4 日提交;最初宣布 2018 年 3 月。

390. 第 xiv:1800111 64[pdf] Cs。简历

从亚历克网开始的历史：深层学习方法的综合调查

作者:md zahangir alom, tarek m. taha, christopher yakopcic, stefopher westberg, paheding sidike, mst shamima nasrin, brian c van esesn, abdul a. a. aw 瓦尔, vijayan k。阿萨里

摘要: 在过去几年中，深度学习在各种应用领域取得了巨大的成功。机器学习的这一新领域发展迅速，并通过一些新的应用模式应用于大多数应用领域，这有助于开辟新的机会。在不同的学习方法类别上，提出了不同的方法，包括监督学习、半监督学习和非监督学习。实验结果表明，在图像**处理**、计算机视觉、语音识别、机器翻译、艺术、医学成像、医学等领域，深度学习的性能优于传统的机器学习方法。信息处理、机器人与控制、生物信息学、**自然语言处理**(nlp)、网络安全等。本报告简要介绍了 dl 方法的发展，包括深神经网络 (dnn)、卷积神经网络 (cnn)、递归神经网络 (rnn) (lstm) 和门式递归单元 (gru)、自动编码器 (ae)、深度信仰网络 (dbn)、生成对抗性网络 (gan) 和深度强化学习 (drl)。此外，我们还介绍了基于上述 dl 方法的高级变型 dl 技术的最新发展。此外，dl 方法在不同的应用领域进行了探索和评估，并纳入了本调查。我们还包括最近开发的框架、sdk 和基准数据集，用于实施和评估深度学习方法。有一些调查发表在神经网络中的深度学习 [1, 38] 和 rl 调查 [234]。然而，这些论文还没有讨论培训大规模深度学习模型的个别先进技术和最近开发的生成模型的方法 [1]。少

2018 年 9 月 12 日提交;v1 于 2018 年 3 月 3 日提交;最初宣布 2018 年 3 月。

评论:39 页，46 个数字，3 个表。arxiv 管理说明：文本与第十四条重叠，1408.3264，arxiv:1411. 4046

391. 第 xiv:1803. 00985[pdf,其他] Cs。Cl

基于朴素贝叶斯和潜在信息的词预测混合模型

作者:henrique x. goulart, mauro d. l. tosi, daniel soares gonçalves, rodgo f.maia, guilherme a. wachs-lopes

摘要: 从历史上看，**自然语言处理**区域已经受到了许多研究人员的太多关注。超越这种兴趣的主要动机之一与单词预测问题有关，该问题指出，给一个句子中的集合词一个固定的词，可以推荐下一个词。在文献中，这个问题是通过基于句法或语义分析的方法解决的。仅此类分析都无法为最终用户应用程序实现实际结果。例如，潜在语义分析可以处理文本的语义特征，但不能建议考虑句法规则的单词。另一方面，有一些模型将这两种方法结合起来，并取得最先进的结果，例如深度学习。这些模型需要很高的计算工作量，这使得该模型在某些类型的应用中不可行。随着技术和数学模型的进步，有可能以更高的精度开发更快的系统。本文提出了一种基于朴素贝叶斯和潜在语义分析的混合词建议模型，并考虑了围绕未填补的空白的相邻词。结果表明，该模型在 msr 句子完成挑战中的精度达到了 44.2。少

2018 年 3 月 2 日提交;最初宣布 2018 年 3 月。

392. 第 xiv:1803.0012[pdf, ps,其他] Cs。CI

多伊 10.114/345458.319191535

越南人的事实问答系统

作者:phong Duc-Thien, duc-thien bui

文摘: 本文介绍了越南**语**端到端事实问答系统的开发。该系统将统计模型和基于本体**的方法**结合在一系列处理模块中，以提供从**自然语言**文本到实体的高质量映射。我们提出了在开发这样一个智能用户界面的挑战，为孤立**的语言**，如越南语，并表明为屈折语言开发的技术不能应用 "按原样"。我们的问答系统可以在测试集中以很有希望的准确性回答广泛的一般知识问题。少

2018 年 3 月 28 日提交;v1 于 2018 年 3 月 1 日提交;最初宣布 2018 年 3 月。

评论:在 hqa18 讲习班的会议记录中，网络会议同伴，法国里昂

393. 第 xiv:1803.00202[pdf] Cs。红外

协同指标学习推荐系统在戏剧电影发行中的应用

作者:miguel campo, jj Espinoza, julie rieger , abhinav taliyan

摘要: 产品推荐系统对于各大电影制片厂在电影绿**光过程**中以及作为机器学习个性化管道的一部分都很重要。协作过滤 (cf) 模型已被证明能够有效地为具有明确客户反馈数据的在线流媒体服务的推荐系统提供支持。cf 模型在无法提供反馈数据的情况下、在新产品发布等冷启动情况下以及在客户层明显不同的情况下 (例如，高频客户与临时客户) 的性能不佳。生成的**自然语言**模型可用于表示新产品描述，如新的电影情节。当与 cf 结合使用时，它们已证明可以提高冷启动情况下的性能。在这些情况下，虽然有明确的客户反馈，推荐引擎必须依靠二进制购买数据，这大大降低了性能。幸运的是，购买数据可以与产品描述相结合，在一个方便的产品空间中生成有意义的产品和客户轨迹表示形式，在这种空间中，接近表示相似。学习测量这个空间中的点之间的距离可以通过一个深度神经网络来完成，该神经网络可以训练客户历史和产品描述的密集矢量化。我们开发了一个基于协作 (深度) 计量学习 (cml) 的系统来预测新戏剧版本的购买概率。我们使用大量的客户历史数据集对模型进行了培训和评估，并对在培训窗口之外发布的一组电影测试了该模型。初步实验显示，相对于不在合作偏好方面进行训练的模型而言，会有收获。少

2018 年 2 月 28 日提交;最初宣布 2018 年 3 月。

评论:6 页, 3 个数字, 3 个表

msc 类: 68t05;68t50

394. 第 xiv:1803.00124[pdf] Cs。CI

多伊 10.1109/ASAR.2018.8480191

用文字表示法改进阿拉伯语中的情感分析

作者:abdulaziz m. alayba, vasile palade, matthew england, raat iqbal

摘要: 阿拉伯语在形态、拼写法和方言方面的复杂性使得阿拉伯语的情感分析更具挑战性。另外，文本功能从推特等短信中提取，为了衡量情绪，让这项任务更加困难。近年来，深度神经网络的应用较多，在情感分类和**自然语言处理**应用中表现出非常好的效果。单词嵌入 (即单词分发方法) 是一种当前的功能强大的工具，用于从上下文文本中捕获最接近的单词。在本文中，我们描述了如何从不同阿拉伯国家的十份报纸上获得

的一个大型阿拉伯语料库中构建 word2vec 模型。通过应用不同的机器学习算法和具有不同文本特征选择的卷积神经网络，我们报告了在我们公开的阿拉伯语健康状况下情绪分类的准确性 (91%- 95%) 的提高情绪数据集 [1]少

2018 年 3 月 30 日提交;v1 于 2018 年 2 月 28 日提交;最初宣布 2018 年 3 月。

评论:提交人接受的 2018 年 asar 提交版本

类:I.2.7;I.2。6

日记本参考:proc. 第二届阿拉伯语和衍生脚本分析与识别国际讲习班 (asar ' 18)，第 13-18 页。ieee, 2018

395. 第 xiv:1803. 00105[pdf] cs. cy

计算国际关系：编程、编码和互联网研究能为学科做些什么?

作者:h. akin unver

摘要: 在过去几年里，由于通信技术的大量进步和大量个人数据的日常生产，计算社会科学成为一门技术性很强、很流行的学科。由于过去十年人均数据产量大幅增加，无论是从规模、字节还是从细节、心率监测器、互联网连接的电器、智能手机、社会科学家提取有意义的社会政治能力来看，都是如此数字数据中的人口统计信息也有所增加。计算国际关系 (comintt) 在方法上存在巨大差距, comin 是指使用一种或多种工具，如数据挖掘、**自然语言处理**、自动文本分析、网络刮，地理空间分析和机器学习，以提供更大和更好的组织数据，以测试更先进的红外理论。本文在概述了计算 ir 的潜力以及红外学者如何建立计算机科学的技术熟练程度 (如从 python、r、qgis、arcgis 或 github 开始) 之后，重点介绍了作者的一些作品。红外学生如何思考计算红外的想法。本文认为，计算方法超越了定性方法和定量方法之间的方法论分歧，为构建真正的多方法研究设计奠定了坚实的基础。少

2018 年 2 月 28 日提交;最初宣布 2018 年 3 月。

396. 决议: 1802. 10229[pdf,其他] Cs。CI

结构梯度树推送的集体实体消歧

作者:杨毅,奥赞伊尔索伊,卡齐谢费特·拉赫曼

摘要: 我们提出了一个基于渐进树引导的结构化学习模型，用于在文档中共同消除命名实体的歧义。梯度树提升是一种广泛使用的机器学习算法，是许多表现最好的**自然语言处理**系统的基础。令人惊讶的是，尽管**语言**具有结构化**的性质,**但大多数作品都限制了使用梯度树提升作为常规分类或回归问题的工具。据我们所知，我们的工作是第一个采用结构化梯度树提升 (sgtb) 算法进行集体实体消歧的工作。通过定义全局特征而不是以前的消歧决策，并使用本地特征对其进行联合建模，我们的系统能够生成全局优化的实体分配，以便在文档中提及。对于我们的全局规范化模型来说，精确推理的成本高得令人望而却步。为了解决这一问题，我们提出了一种具有黄金路径的双向波束搜索算法，它是标准波束搜索算法的一种变体。bibsg 利用过去和未来的全球信息来进行更好的本地搜索。在标准基准数据集上的实验表明, sgtb 在已发布的结果基础上有了显著改善。具体而言, sgtb 在流行的 aida-conll 数据集上的绝对精度接近 1%，优于以前最先进的神经系统。少

2018 年 4 月 23 日提交;v1 于 2018 年 2 月 27 日提交;最初宣布 2018 年 2 月。

评论:被 naacl 2018 部接受

397. 第 1802.09968[pdf,其他] Cs。CI

一种混合文字符的摘要总结方法

作者:张志登,黄志嘉,杨志远,许永珍

摘要: 自动抽象文本摘要是**自然语言处理**的一个重要而具有挑战性的研究课题。在许多广泛使用**的语言中**,**汉语**有一个特殊的属性，一个汉字包含了与一个词相当的丰富信息。现有的中文文本摘要方法，无论是采用完全基于字符的表示，还是基于描述的表示，都未能充分利用这两种表述所携带的信息。为了准确地捕捉文章的本质，我们提出了一种混合字符格学方法 (hwc)，它保留了基于描述和基于字符表示的优点。我们通过将建议的 hwc 方法应用于两种现有方法来评估其优势，并发现它在广泛使用的数据集 lcsts 上以 24 个 rouge 点的优势生成最先进的性能。此外，我们还发现 lcsts 数据集中包含一个问题，并提供一个脚本来删除重叠对 (摘要和短文本)，以便为社区创建一个干净的数据集。建议的 hwc 方法还可在新的、干净的 lcsts 数据集上生成最佳性能。少

2018 年 9 月 8 日提交;v1 于 2018 年 2 月 27 日提交;最初宣布 2018 年 2 月。

398. **建议: 1802. 09059[pdf,其他] Cs。Lg**

用于文本数据的语感消歧的一种单深双向 lstm 网络

作者:ahmad pesaranghader, ali pesaranghader, stan matwin, marina sokolova

摘要: 由于最近的技术和科学进步，我们有大量的信息隐藏在非结构化文本数据中，如离线在线叙述、研究文章和临床报告。由于这些数据固有的模糊性，正确挖掘这些数据，一种语感消歧算法可以避免**自然语言处理**(nlp) 管道中出现的一些困难。不过，考虑到一种**语言**或技术领域中的大量不明确的词语，我们可能会遇到适当部署现有水务署模式的限制限制。本文试图通过提出一个单一的双向长期短期存储器 (blstm) 网络来解决单词单的 wsd 算法问题，该网络通过考虑感官和上下文序列对所有不明确的词共同作用。在 senseval-3 基准上进行了评估，我们证明了我们的模型的结果与性能最好的 wsd 算法是可比较的。我们还讨论了应用额外的修改如何减少模型故障和对更多培训数据的需求。少

2018 年 2 月 25 日提交;最初宣布 2018 年 2 月。

评论:12 页, 1 个数字，刊登在 2018 年 5 月 8 日至 11 日在加拿大多伦多举行的第三十一届加拿大人工智能会议论文集上

399. **建议: 1802.09055[pdf] Cs。Hc**

特定于域的设计模式: 针对对话用户界面的设计

作者:艾哈迈德·法迪勒

摘要: 设计会话用户界面体验是复杂的，因为对话带来了许多期望。当这些期望得到满足时，我们会觉得界面是**自然**的，但一旦被侵犯，我们就会觉得出了问题。在过去十年中，人类**的语言**技术和行为使人类能够利用口语对话与软件对话，以获取、创建和**处理**信息。对设计聊天机器人交互的实用性了解较少。在本文中，我们介绍了会话用户界面 (cui) 的**性质**，并描述了它们所基于的基础技术。此外，我们还定义了在各个领域设计会话接口的准则。本文特别关注 cui 设计模式中使用的元素和技术的分类。在总结了 ui 的某些挑战后，我们讨论了特定域的 cui 设计中需要考虑的重要功能和聊天机器人状态。我们设想这项研究支持 mui 研究人员设计适用于特定领域的定制聊天机器人，并改善人工智能和会话剂领域的研究挑战的当前状态。少

2018 年 2 月 25 日提交;最初宣布 2018 年 2 月。

评论:7 页

400. 第 1802.08395[pdf,其他] Cs。Cl

实现端到端口语理解

作者:d 南京市 serdyuk, yon 强 wang, christian fuegen, anuj kumar, 白阳 liu, y 雅库 ua bengio

文摘: 口语理解系统传统上被设计为一个由多个组件组成的管道。首先，音频信号由自动语音识别器**处理**，用于转录或最佳假设。根据识别结果 ，**自然语言**理解系统将文本划分为结构化数据 ，作为式用户的域、意图和插槽 ，如对话系统、免提应用程序。这些组件通常是独立开发和优化的。本文介绍了一种用于口语理解的端到端学习系统的研究 。通过这种统一的方法，我们可以直接从音频特征推断语义意义，而不需要中间文本表示。研究表明，训练模型能取得合理的良好效果，并证明该模型可以直接从音频特征中捕捉语义注意力。少

2018 年 2 月 23 日提交;最初宣布 2018 年 2 月。

评论:提交给 icassp 2018

401. 第 1802.08148[pdf,其他] Cs。Cl

习语: 多语言链接的习语数据集

作者:diego moussallem, mohamed ahmed sherif, diego esteves, marcos zampieri , axel-cyrille ngonga ngomo

摘要: 在本文中，我们描述了 lidioms 数据集，这是一种多语言 rdf 表示的习语，目前包含五**种语言**: 英语、德语、意大利语、葡萄牙语和俄语。数据集旨在通过提供跨**语言**习语之间的链接来支持**自然语言处理**应用程序。对基础数据进行了爬网，并从各种来源进行了集成。为了确保爬网数据的质量，所有习语都由至少两名母语人士进行评估。在此，我们介绍了为构建数据而设计的模型。我们还提供了将 li 追究系统链接到著名的多语言数据集 (如 babelnet) 的详细信息。根据语言链接开放数据社区，生成的数据集符合最佳实践。少

2018 年 2 月 22 日提交;最初宣布 2018 年 2 月。

评论:接受在 2018 年语言资源和评价会议上发表

402. 建议: 1802.07459[pdf,其他] Cs。Cl

通过图形卷积网络匹配长文本文档

作者:刘邦,张婷,迪牛,林景红, 赖昆峰,徐宇

摘要: 识别两个文本对象之间的关系是许多**自然语言处理**任务背后的核心研究问题。提出了一种广泛的文本匹配深度学习方案，主要侧重于句子匹配、问答或查询文档匹配。我们指出，现有的方法在匹配长文档方面效果不佳，这对基于 ai 的新闻文章理解和事件或故事形成至关重要。原因是这些方法要么省略，要么未能在长文档中充分利用复杂的语义结构。在本文中，我们提出了一种文本匹配的图形方法，特别是针对长文档匹配，例如确定两个新闻文章是否在现实世界中报道相同的事件，可能具有不同的叙述。我们建议概念交互图为文档生成图形表示形式，顶点表示不同的概念，每个顶点是文档中的一个或一组连贯关键字，边缘表示不同之间的交互概念，由文档中的句子连接。基于文档对的图形表示，我们进一步提出了一种 siamese 编码图形卷积网络，该网络通过 siamese 神经网络学习顶点表示，并通过图形卷积网络聚合顶点特征。生成匹配

的结果。基于腾讯为其智能新闻产品创建的两个标记新闻文章数据集，对所提出的方法进行了广泛的评价，表明所提出的长文档匹配图方法明显优于广泛的最先进的方法。少

2018 年 2 月 21 日提交;最初宣布 2018 年 2 月。

评论:9 页, 6 个数字

403. 建议: 1802. 07370[pdf,其他] Cs。Cl

使用后缀编码的宇宙句子表示

作者:悉达多·婆罗门

摘要: 计算句子的通用分布表示是自然**语言处理**中**的**一项基本任务。我们提出了一种学习这种表示的方法，方法是对句子中单词序列的后缀进行编码，并对斯坦福**自然语言**推理 (snli) 数据集进行训练。我们通过在 senteval 基准上对其进行评估，改进关于多个传输任务的现有方法，从而证明了我们的方法的有效性。少

2018 年 2 月 20 日提交;最初宣布 2018 年 2 月。

评论:4 页，提交给 iclr 2018 年讲习班

404. 修订: 180006893[pdf, ps,其他] Cs。Cl

学习 157 种语言的单词向量

作者:edouard grave, piotr bojanowski, prakhar gupta, armand joulin, tomas mikolov

文摘 分布式字表示 (或字向量) 最近被应用于**自然语言处理**中的许多任务，从而获得了最先进的性能。成功应用这些表示的一个关键要素是在非常大的语料库上对它们进行培训，并在下游任务中使用这些预先培训的模型。在本文中，我们描述了我们如何为 157 **种语言**训练这种高质量的单词表示。我们使用两个数据来源来训练这些模型：免费的在线百科全书维基百科和来自共同抓取项目的数据。我们还引入了三个新的单词类比数据集来评估这些单词向量，用于法语、印地语和波兰语。最后，我们在有评估数据集的 10 **种语言**上对预先训练的单词向量进行评估，与以前的模型相比，它表现出非常强的性能。少

2018 年 3 月 28 日提交;v1 于 2018 年 2 月 19 日提交;最初宣布 2018 年 2 月。

评论:接受 lrec

405. 修订: 1802.0 6829[pdf] Cs。艾

本体驱动计算机系统的设计原理及软件开发模型

作者:a. v. palagin, n. g. petrenko , v. yu。velychko, k. s. malakhov, o. v. karun

摘要: . 基于本体论方法的面向知识的信息系统方法论。这种系统实现了以技术为导向的**从自然语言**文本集中提取知识及其正式和逻辑的呈现和应用**处理**

2018 年 2 月 13 日提交;最初宣布 2018 年 2 月。

评论:在俄语

日记本参考:信息化与管理问题第 2 卷第 34 号(2011) 96-101

406. 建议: 1802. 06368[pdf,其他] Cs。Lg

节点中心化与节点嵌入算法特征的分类性能

作者:kto nozawa, masanari kim ura, atsunori kanemura

摘要: 将图形节点嵌入到向量空间中可以允许使用机器学习来预测节点类，但与**自然语言处理**领域相比，节点嵌入算法的研究还不成熟，因为图形的**多样性**。通过四个节点

嵌入算法、四、五个图形中心和六个数据集的系统实验，研究了节点嵌入算法在描述不同图形的图形中心度量方面的性能。实验结果对节点嵌入算法的性质进行了深入的研究，为进一步研究这一课题提供了依据。少

2018 年 2 月 18 日提交;最初宣布 2018 年 2 月。

评论:iclr 2018 年研讨会轨道正在审查中

407. 建议: 1802. 06196[pdf,其他] Cs。CI

分布式术语词库的网络嵌入是否可以与 word 向量结合起来，以获得更好的表示能力?

作者:abhik j 长官, pawan goyal

摘要: 从文本中学习到的单词**的**分布式表示在最近的**时代被**证明是成功的。虽然有些方法将单词表示为使用预测模型 (word2vec) 或密集计数模型 (glove) 从文本计算的向量，但另一些方法则试图在分布词库网络结构中表示这些词，其中一个词的邻域是一组具有足够上下文重叠的词。在网络嵌入技术 (deepwalk、line、node2vec 等) 研究的推动下，我们将分布词库网络转化为密集的词向量，并研究了分布词库嵌入在改进中的作用整体文字表示。这是我们第一次尝试，我们表明，结合通过分布词库嵌入获得的单词表示与最先进的单词表示，有助于在评估时显著提高性能针对 nlp 任务，如单词相似性和相关性、同义词检测、类比检测等。此外，我们还表明，即使不使用任何手工制作的词法资源，我们也可以得出在单词相似性和相关性任务中具有与使用词汇资源的表示的类似性能的表示。少

2018 年 2 月 17 日提交;最初宣布 2018 年 2 月。

408. 建议: 1802. 06 194[pdf,其他] Cs。简历

hwnet v2: 手写文档的有效 word 图像表示

作者:Praveen krishnan, c. v. jawahar

摘要: 我们提出了一个框架，用于学习手写单词图像的高效整体表示。该方法采用了具有传统分类损耗的深层卷积神经网络。我们工作的主要优势在于: (一) 有效利用合成数据预训练深度网络, (二) 带有兴趣池区域 (称为 hwnet v2) 的 resnet-34 体系结构的改编版本，该架构学习具有变量的判别特征大小的单词图像，和 (iii) 现实的增强训练数据与多个尺度和弹性失真，模仿自然**过程**的**笔迹**。我们进一步研究了在各个层次进行微调的**过程**，以缩小合成域和真实域之间的域差距，并使用文献中提出的最新可视化技术分析了在不同层学到的差异.我们的表现使我们能够以最小的表示尺寸在标准手写数据集和历史手稿上的最先进的单词发现性能。在具有挑战性的 iam 数据集上，我们的方法首先是报告 0.90 以上的 map，用于仅有 32 个维度的表示大小的单词发现。此外，我们还介绍了英文和 indic 脚本打印文档数据集的结果，这些结果验证了所建议的单词图像表示框架的一般**性质**。少

2018 年 2 月 17 日提交;最初宣布 2018 年 2 月。

评论:17 页, 13 位数字

409. 建议: 1802. 05934[pdf,其他] Cs。CI

基于实例的基于局部敏感哈希的交叉数据集查询的深度转移学习

作者:somnath basu roy chowdhury, k m annervaz , ambedkar dukkipati

摘要: 监督学习模型通常是在单个数据集上进行培训的，这些模型的性能在很大程度上取决于数据集的大小，即具有地面真相的可用数据量。学习算法尝试仅根据培训期间

提供的数据进行泛化。在这项工作中，我们提出了一种归纳转移学习方法，它可以通过在**自然语言处理**(nlp) 领域从不同的学习任务中注入相似的实例来增强学习模型。我们建议使用来自源数据集的实例表示形式，即 \ textimt，而无需从源学习模型继承任何东西}。学习了 \ textin{源} \ & amp; \ textythnetn 数据集的实例的表示形式，使用软关注机制和 \ textin{位点敏感哈希} 检索相关源实例，然后扩展到模型中在目标数据集的训练中。我们的方法同时利用本地回文 (实例级别信息) 以及数据集的宏观统计观点。使用这种方法，我们在基线上显示了三个主要新闻分类数据集的显著改进。实验评价还表明，该方法大大减少了对标记数据的依赖，从而使可比业绩有了很大的优势。通过我们提出的交叉数据集学习过程，我们表明，与从单个数据集学习相比，可以实现更好的竞争力性能。少

2018 年 2 月 16 日提交;最初宣布 2018 年 2 月。

410. **建议: 1802.05930[pdf,其他] Cs。Cl**

超越数据集的学习: 用于自然语言处理的知识图增强神经网络

作者:k m annervaz, somnath basu roy chowdhury, ambedkar dukkipati

摘要: 机器学习一直是许多 ai 问题的典型解决方案，但学习仍然在很大程度上依赖于特定的培训数据。一些学习模型可以在贝叶斯建立的先验知识中结合起来，但这些学习模型没有能力根据需要获得任何有组织的世界知识。在这项工作中，我们建议以知识图 (kg) 事实的形式，以自然 **语言处理**(nlp) 任务的形式，加强具有世界知识的学习模型。我们的目标是开发一个深入的学习模型，可以根据任务使用注意力机制，从知识图中提取相关的事先支持事实。为了减少注意力空间，我们引入了一种基于卷曲的知识图实体和关系群的学习表示模型。我们证明了该方法是高度可扩展的，以达到必须**处理**的先验信息量，并可应用于任何通用 nlp 任务。利用该方法，我们可以显著提高《新闻纵横》、《 dbpedia 》数据集的文本分类性能，以及斯坦福**自然语言**推理 (snli) 数据集的自然**语言**推断性能。我们还证明，当一个深度学习模型能够以知识图的形式获得有组织的世界知识时，它可以用大量的标记培训数据来很好地训练。少

2018 年 5 月 20 日提交;v1 于 2018 年 2 月 16 日提交;最初宣布 2018 年 2 月。

评论:2018 年接受

411. **建议: 1802.05667[pdf,其他] Cs。Cl**

使用词汇数据库和语料库统计信息计算单词和句子之间的相似性

作者:ate pawar , vijay mago

摘要: 句子间语义相似度的计算是**自然语言处理**领域中一个长期存在的问题。语义分析领域在文本分析相关研究中发挥着至关重要的作用。语义相似性因操作域的不同而不同。本文结合语义相似性和语料库统计，提出了一种解决这一问题的方法。为了计算单词和句子之间的语义相似性，该方法采用基于边缘的方法，使用词汇数据库。该方法可应用于多个领域。该方法已在基准标准和平均人类相似度数据集上进行了测试。在这两个数据集上进行测试时，它给出了比其他类似模型更高的单词和句子相似性相关值。对于单词相似性，我们得到了皮尔逊相关系数为 0.8753，对于句子相似性，相关性为 0.8753。少

2018 年 2 月 20 日提交;v1 于 2018 年 2 月 15 日提交;最初宣布 2018 年 2 月。

412. **建议: 1802. 05583[pdf,其他] Cs。Cl**

用于罗马尼亚文本到语音和语音到文本应用程序的工具和资源

作者:Tiberiu boros, stefan daniel Dumitrescu, vasile pais

文摘: 本文介绍了一套旨在为罗马尼亚**语的自然语言处理**、文本到语音合成和语音识别提供支持的资源和工具。虽然这些工具是通用的，可以用于任何**语言**(我们成功地培训了我们的系统的 50 多种**语言**，并参与了通用依赖共享任务)，但这些资源仅与罗马尼亚语相关**的语言处理**。少

2018 年 2 月 15 日提交;最初宣布 2018 年 2 月。

413. 建议: 1802.0512[pdf,其他] Cs。铬

基于本机 api 系统调用的恶意软件检测的机器学习方法

作者:陈宇金

摘要: 随着计算系统的日益先进和用户越来越多地参与技术，安全性从未像现在这样受到更大的关注。在恶意软件检测、静态分析、分析潜在恶意文件的方法中，一直是突出的方法。但是，随着恶意程序变得更加高级，并采用模糊其二进制文件来执行相同的恶意函数的功能，这种方法很快就会失败，这使得静态分析对于较新的变体极为困难。本文评估的方法是一种新的动态恶意软件分析方法，它可以比静态分析更好地推广到较新的变体。在**自然语言处理**(nlp) 最近取得成功的启发下，通过对包含有用信息的系统调用进行此类分析，评估了广泛使用的文档分类技术在检测恶意软件时的应用关于程序的操作，作为程序对内核的请求。所考虑的功能是从良性和恶意程序的系统调用跟踪中提取的，对这些跟踪进行分类的任务被视为系统调用跟踪的二进制文档分类任务。**处理**系统调用跟踪是为了删除参数，而只保留系统调用函数名称。这些功能被分组为各种 n 克，并加权与期限频率反向文档频率。本文表明，由随机梯度下降和传统的坐标下降优化线性支持向量机 (svm) 在 svm 的 wolfe 双形式上是有效的，达到了 96% 的最高精度, 95% 的召回分数。其他贡献包括确定重要的系统调用序列，这些序列可以成为进一步研究的途径。少

2018 年 5 月 19 日提交;v1 于 2018 年 2 月 15 日提交;最初宣布 2018 年 2 月。

评论:8 页，英特尔国际科学与工程博览会项目-SOFT006T

414. 建议: 1802.05340[pdf,其他] Cs。艾

从游戏到符号推理: 学习 sat 求解启发式在阿尔法风格 (去) 零

作者:王飞, tiark rompf

摘要: 尽管深度神经网络最近在图像和语音识别、**自然语言处理**和强化学习等各个领域取得了成功，但在带来数值优化到符号推理。研究人员提出了不同的途径，如证明合成的神经机器翻译、符号的矢量化和表示符号模式的表达式，以及用于降维的神经后端与符号的耦合决策的前端。然而，这些最初的探索仍然只是点解决方案，并存在着缺乏正确性保证等其他缺点。本文提出了将符号推理转换为游戏的方法，并在符号问题的上直接利用了阿尔法 (go) 零的深层强化学习的力量。以布尔满意度 (sat) 问题为展示，论证了该方法的可行性，以及模块化、效率和正确性保证的优点。少

2018 年 2 月 14 日提交;最初宣布 2018 年 2 月。

415. xiv:1802. 05300[pdf,其他] Cs。Lg

利用可信数据在严重噪声破坏的标签上训练深部网络

作者:dan hendrycks, mantas mateika, duncan wilson, kevin gimpel

摘要: 随着深度学习的出现，海量数据集的重要性日益增加，使得对标签噪声的鲁棒性成为分类器所具有的一个关键属性。标签噪声的来源包括大型数据集的自动标记、非

专家标记以及数据中毒对手的标签损坏。在后一种情况下，腐败可能是任意的坏的，甚至是如此糟糕，分类器可以高度自信地预测错误的标签。为了防止这种噪音源，我们利用了这样一个事实，即一小套干净的标签往往很容易采购。我们证明，通过使用一组具有干净标签的受信任数据，可以实现对标签噪声具有严重强度的鲁棒性，并建议进行损失校正，以数据高效的方式利用受信任的示例，以减轻标签噪声对深度的影响神经网络分类器。在视觉和**自然语言处理**任务中，我们在几个优势上试验各种标签噪声，并表明我们的方法明显优于现有的方法。少

2018 年 10 月 30 日提交;v1 于 2018 年 2 月 14 日提交;最初宣布 2018 年 2 月。

评论:将于 2018 年 NIPS 出现。https://github.com/mmazeika/glc 提供 pytorch 代码

416. 建议: 1802. 04609[pdf,其他] Cs。CI

基于网络特征的共称低聚检测

作者:abhik j 长官, pawan goyal

摘要: 词汇关系的区分一直是**自然语言处理**(nlp) 领域的长期追求。近年来，为了检测词汇关系，如超义、同义、共称等，分布语义模型正被广泛用于其他形式。尽管在检测超我的关系方面做了大量的努力，但共低位基因检测的问题却很少被调查。本文提出了一种新的监督模型，利用各种网络措施来识别具有较高精度、性能更好或与最先进模型相当的共同低联关系。少

2018 年 2 月 13 日提交;最初宣布 2018 年 2 月。

417. 建议: 1802. 04162[pdf,其他] Cs。Lg

一般上下文强盗的策略梯度

作者:潘飞阳,蔡庆鹏，唐平忠,富镇庄,河青

文摘: 上下文土匪算法已成功地部署到各种工业应用中，以便在勘探和开发与最先进的性能之间进行权衡，从而最大限度地降低在线成本。但是，适用性受到对问题的过于简化的假设的限制，例如假设奖励线性取决于上下文，或者假设状态不受以前操作影响的静态环境。在这项工作中，我们提出了一种替代方法，一般的上下文土匪使用演员-评论家神经网络直接优化的政策空间，创造的策略梯度的上下文土匪 (pgcb)。它优化了一类政策，在这些政策中，选择武器的边际概率 (在其他武器的预期中) 具有简单的封闭形式，因此目标是可微的。特别是，这类政策的梯度是简洁的。此外，我们还提出了两种有用的启发式技术，即依赖时间的贪婪和演员退出。前者确保 pgcb 在极限中具有经验贪婪，而后者则通过使用带辍学率的电抗网络作为贝叶斯近似值，在勘探和开发之间取得平衡。pgcb 可以解决标准案例中的上下文土匪，也可以解决马尔可夫决策**过程**中的泛化问题，因为在这种情况下，有一种状态决定着武器上下文的分布，并影响在选择手臂时的即时奖励，因此可以适用于广泛的现实设置，如个性化的推荐系统和**自然语言**世代。我们评估了玩具数据集以及音乐推荐数据集上的 pgcb。实验表明，pgcb 收敛速度快，遗憾度低，优于经典的上下文土匪方法和香草策略梯度法。少

2018 年 5 月 22 日提交;v1 于 2018 年 2 月 12 日提交;最初宣布 2018 年 2 月。

418. 建议: 180004051[pdf,其他] cs. ne

一个深刻的音乐表现来统治他们? * 对不同的代表性学习策略进行比较分析

作者:jaehun kim, julián urbano , cynthia c. s.liem, alan hanjalic

摘要: 在计算机视觉和**自然语言处理**领域成功部署深度学习的启发下，这种学习范式也进入了音乐信息检索领域。为了从深度学习中受益，同时也是高效的方式，深度转移学习已经成为一种常见的方法。在这种方法中，可以重用预先训练的神经网络的输出，作为新的学习任务的基础。潜在的假设是，如果初始和新的学习任务显示出共性并应用于相同类型的输入数据（例如音乐音频），则生成的数据深度表示也会为新任务提供信息。但是，由于大多数用于生成深度表示的网络都是使用单一的初始学习源进行训练的，因此，对于任意的未来任务而言，上述假设的有效性是值得怀疑的。本文介绍了对音乐领域的数据和学习任务产生深度表示的最重要因素的研究结果。我们通过广泛的实证研究进行了这项调查，涉及多个学习来源，以及多个深度学习架构，其来源之间的信息共享水平不同，目的是学习音乐表现。然后，我们在考虑多个目标数据集进行评估时验证这些表示形式。我们的实验结果提供了一些关于如何在音乐领域中进行广泛部署的深度数据表示的方法设计的见解。少

2018 年 10 月 26 日提交;v1 于 2018 年 2 月 12 日提交;最初宣布 2018 年 2 月。

评论:这项工作已提交给 "神经计算与应用：关于音乐和音频深度学习的特刊"，目前正在审查中

419. **建议: 1802. 03656[pdf, ps,其他] Cs。CI**

文本动物园是重新思考文本分类的新基准

作者:王本友,王力,魏启康,刘立春

摘要: 文本表示**是自然语言处理**中的一个基本问题，尤其是在文本分类中。最近，许多神经网络方法与微妙的表示模型（如 fasttext, cnn, mn 和许多混合模型与关注机制）声称，他们在特定的文本分类数据集中达到了最先进的。但是，它缺乏一个统一的基准来比较这些模型，并揭示了每个子组件在各种设置中的优势。我们重新实现了 20 多个流行的文本表示模型，以便在 10 多个数据集中进行分类。本文从神经网络的角度对文本分类任务进行了重新思考，并通过对上述结果的分析，得到了一定的效果。少

2018 年 3 月 18 日提交;v1 于 2018 年 2 月 10 日提交;最初宣布 2018 年 2 月。

评论:需要完成一个基准

420. **建议: 1802. 03436[pdf,其他] Cs。佛罗里达州**

锤子类型过程的语言（和系列）

作者:cosmin bonchis, gabriel istt, vlad Bonchis

文摘: 我们研究与哈默斯利**过程**相关的**语言**和形式权力系列。我们表明，普通的哈默斯利**进程**产生了一种规则的**语言**，而哈默斯利树**进程**产生的是确定性的无上下文（但非规则）语言。对于哈默斯利**过程**间隔的扩展，我们表明有两种相关**的正式语言**。其中之一导致与普通哈默斯利树**过程**相同的**语言**类别。另一个生成非上下文**语言**。研究结果的动机是研究著名的乌兰-哈默斯利问题的类似物的可跨越序列的问题。为了实现这个目标，我们还给出了一个计算与哈默斯利**过程**变体相关的形式功率序列的算法。我们使用这些算法来解决缩放常数的**性质**，在以前的工作中推测是黄金比率。我们的结果为这一猜想提供了实验支持。少

2018 年 4 月 9 日提交;v1 于 2018 年 2 月 9 日提交;最初宣布 2018 年 2 月。

评论:出现在机器、计算和普遍性中（mcu' 2018）

421. **建议: 180003238[pdf,其他] Cs。CI**

基于递归神经网络的语义变分自动编码器用于序列到序列学习

作者:张明君, seo seung wan, pilsung kang

文摘: 序列到序列 (seq2seq) 模型在最近各种**自然语言处理**方法 (如机器翻译、文本摘要和语音识别) 的成功中发挥了重要作用。然而，目前的 seq2seq 模型很难从一长串单词中保存全球潜在信息。变分自动编码器 (vae) 通过学习输入句子的连续语义空间来缓解这一问题。然而，这并不能完全解决问题。本文提出了一种新的基于递归神经网络 (rnn) 的 seq2seq 模型--rnn 语义变分自动编码器 (rn-svae)，以更好地捕获一系列单词的全局潜在信息。为了正确地反映句子中单词的含义，而不考虑它在句子中的位置，我们使用编码器的最终状态和之前的每一个隐藏状态之间的注意信息构造一个文档信息向量。然后，利用该向量来利用变分方法，学习连续语义空间的均值和标准差。利用文档信息向量寻找句子的语义空间，可以更好地捕捉句子的全局潜在特征。三**种自然语言**任务 (即语言建模、缺失词法归因、释义识别) 的实验结果证实，所提出的 rn-svae 比两个基准模型具有更高的性能.少

2018 年 6 月 2 日提交;v1 于 2018 年 2 月 9 日提交;最初宣布 2018 年 2 月。

评论:14 页

422. 建议: 18002210[pdf,其他] Cs。简历

视觉刺激诱发的脑活动语义表达

作者:erri Asoh, ichiro kobayashi, sh 真司 nishimoto, satoshi nishida, hideki asoh

文摘: 基于**语言**表征的人脑活动定量建模在神经科学系统中得到了积极的研究。然而，以前的研究考察了单词级的表达，对我们能否从大脑活动中恢复结构化句子了解甚少。本研究试图从视觉刺激引起的人脑活动中生成语义内容的**自然语言**描述。为了有效地利用少量可用的大脑活动数据，我们提出的方法采用了使用深度学习框架的预先训练的图像字幕网络模型。为了将大脑活动应用于图像字幕网络，我们训练了学习大脑活动与深层图像特征之间关系的回归模型。结果表明，该模型可以解码大脑活动，并使用**自然语言**句子生成描述。我们还对来自大脑不同亚群的数据进行了几次实验，这些子集已知是**处理**视觉刺激的。结果表明，句子世代的语义信息在整个皮层中普遍存在。少

2018 年 1 月 19 日提交;最初宣布 2018 年 2 月。

评论:11 页, 8 个数字

423. 建议: 1802. 01174[pdf] Cs。DI

从科学生物医学出版物中发现和提取作者贡献信息的一种方法

作者:dominika tkaczyk, andrew collins, joeran beel

摘要: 创建科学出版物是一个复杂的**过程**，通常由许多不同的活动组成，如设计实验、数据编制、编程软件以及撰写和编辑手稿。关于论文个别作者贡献的信息在评估作者的科学成就方面很重要。生物医学学科的一些出版物以**自然语言**编写的简短章节的形式描述了作者的作用，通常题为 "作者的贡献"。本文对这些章节的内容中常见的角色进行了分析，提出了科学出版物中作者从**自然语言**文本中自动提取角色的算法。在研究的第一部分，我们使用聚类技术，以及开放信息提取 (openie)，在从 pubmed 中心资源获得的 2, 000 个贡献部分中，半自动地发现最受欢迎的角色。我们的方法发现的角色包括: 实验 (1, 743 实例，占主体内整个角色集的 17%)、分析 (1, 343, 16%)、研究设计 (1, 132, 13%)、解释 (879, 10%)、概念化 (879, 10%)、纸质阅读 (879, 10%),论文写作 (724, 8%)，论文审查 (501, 6%)，论文起草 (351, 4%)，协调 (319, 4%)，数据收集 (76, 1%)，论文审查 (41, 0.5%) 和文献审查 (41, 0.5%)。然后利用发现的角色，基

于朴素贝叶斯算法，自动为受监督角色提取者构建训练集。根据我们所进行的评价，所提出的角色提取算法能够精确地从文本中提取角色 0.71、召回 0.71 和 f1 0.71。少

2018 年 2 月 4 日提交;最初宣布 2018 年 2 月。

424. **建议: 1802: 0000[pdf,其他] Cs。CI**

基于目标的聊天机器人对话框管理引导与转移学习

作者:vladimir llievski, clclu musat, and 列 ea hossmann, michael baeriswyl

摘要: 面向目标 (go) 对话系统，俗称目标导向聊天机器人，可帮助用户在封闭的域内实现预定义的目标 (例如预订电影票)。第一步是通过使用**自然语言**理解技术来理解用户的目标。一旦知道了这个目标，机器人就必须管理对话来实现这个目标，这个目标是根据学习到的政策进行的。对话系统的成功取决于政策的质量，而这又取决于政策学习方法是否有高质量的培训数据，例如深度强化学习。由于领域的特殊性，可用数据的数量通常太低，无法培训良好的对话政策。本文介绍了一种转移学习方法，以减轻域内数据可用性低的影响。我们基于转移学习的方法将机器人在远程域中的成功率提高了 20%，与没有转移学习的模型相比，我们对接近域的成功率提高了一倍多。此外，转学学习聊天机器人学习政策的速度高达 5 到 1 0 倍。最后，由于转移学习方法是对其他**处理**(如预热启动) 的补充，我们表明它们的联合应用能产生最佳的结果。少

2018 年 7 月 24 日提交;v1 于 2018 年 2 月 1 日提交;最初宣布 2018 年 2 月。

评论:7 页 (6 页加 1 页参考), 5 个数字, 1 个伪代码图

425. **建议: 1802. 0000[pdf,其他] Cs。红外**

生物医学自然语言处理中的构词比较

作者:王燕山,刘思佳,纳维德·阿夫扎尔,马吉德·拉斯特格-莫贾拉德,王利伟, 沈费晨,保罗·金斯伯里,刘洪芳

摘要: 词嵌入在生物医学**自然语言处理**(nlp) 应用中得到了广泛的应用，因为它们提供了词的向量表示，捕捉单词的语义属性和语言关系之间的单词。许多生物医学应用使用不同的文本资源 (如维基百科和生物医学文章) 来训练单词嵌入，并将这些单词嵌入应用于下游的生物医学应用。然而，在评估从这些资源中培训的嵌入一词方面几乎没有开展什么工作。在本研究中，我们提供了从临床笔记、生物医学出版物、维基百科和新闻这四种不同资源中训练的单词嵌入的实证评估。我们进行了定性和定量的评估。为了进行定性评估，我们手动检查了五个与一组给定的目标医学单词最相似的医学单词，然后通过这些单词嵌入的可视化分析单词嵌入。在定量评价方面，我们进行了内在评价和外在评价。根据评价结果，我们可以得出以下结论。首先，在临床笔记和生物医学出版物上训练的 "嵌入" 一词可以更好地捕捉医学术语的语义，找到更相关的类似医学术语，并且与维基百科上训练的医学术语相比，更接近人类专家的判断。新闻。其次，下游生物医学 nlp 应用的词嵌入质量并不存在一致的全局排名。但是，添加单词嵌入作为额外的功能将提高大多数下游任务的结果。最后，在生物医学领域语料库培训的 "嵌入" 一词不一定比在其他一般领域语料库上接受过任何下游生物医学 nlp 任务培训的词具有更好的性能。少

2018 年 7 月 18 日提交;v1 于 2018 年 2 月 1 日提交;最初宣布 2018 年 2 月。

426. **建议: 1802. 00396[pdf,其他] Cs。CI**

分裂的国家？一种利用文本和投票检测偏好亲和力块的多网络方法

作者:caleb pomeroy, niheer dasandi, slava j. mikhaylov

摘要: 本文为一个新兴的文献做出了贡献，该文献将选票和文本与投票和文本同时建模，以更好地理解所表达的偏好的两极分化。根据**自然语言处理**和网络科学文献的发展，提出了一种在国际关系等多维环境下估计偏好两极分化的新方法--即单词嵌入，它保留了人类**语言**的宝贵的语法品质，以及多层网络中的社区检测，它将密集连接的行为者定位在多个复杂的网络中。我们发现，同时使用这些工具有助于更好地估计各国在联合国投票和发言中表达的外交政策偏好，而超出了仅投票允许的范围。这些定位的亲和力集团的效用通过在国际关系中的冲突爆发的应用得到了证明，尽管这些工具将引起所有在多层面上衡量偏好和两极分化的学者的兴趣。设置。少

2018 年 2 月 1 日提交;最初宣布 2018 年 2 月。

427. **决议: 1202. 00272[pdf, ps,其他] Cs。Hc**

基于三维人类活动识别和人样决策机制的服务机器人实时人机交互

作者:康丽,孙世英, 吴金婷，赵晓光,谭敏

文摘: 本文介绍了一种基于三维人体活动识别和人形决策机制的服务机器人实时人机交互 (hri) 系统的开发。人机交互 (hri) 系统允许一个人使用**自然肢体语言**与服务机器人互动，收集 3d 骨架关节序列，其中包含关于用户的丰富的人体运动信息，通过微软 kinect。此信息用于训练一个三层长时间短期内存 (lstm) 网络，用于人类行动识别。机器人根据在线 lstm 网络测试了解用户意图，并通过机器人手臂或机箱的运动对用户做出响应。此外，类似人类的决策机制也被融合到这个**过程**中，这使得机器人可以本能地根据任务优先级决定是否中断当前的任务。在机器人操作系统 (ros) 平台上建立了整个系统的框架。我们的服务机器人与用户进行了实际活动互动，以展示开发的 hri 系统的有效性。少

2018 年 2 月 1 日提交;最初宣布 2018 年 2 月。

428. **建议: 1801 1.10437[pdf,其他] Cs。艾**

深度学习在实践中的作用。但它在理论上行得通吗?

作者:lénguyn hoang, rachid guerraoui

摘要: 深度学习依赖于一种非常特殊的神经网络：那些叠加了几个神经层的神经网络。近年来，深度学习在图像分析、语音识别、**自然语言处理**等诸多方面取得了重大突破。然而，对这一成功没有任何理论解释。特别是，不清楚为什么网络越深，其实际表现就越好。我们认为，这个解释与从我们周围的宇宙中收集的数据的一个关键特征密切相关，这些数据为机器学习算法提供了信息：大的非并行逻辑深度。粗略地讲，我们推测，宇宙中最短的计算描述是具有固有的计算时间大的算法，即使有大量计算机可用于并行化。有趣的是，这个猜想，结合民间传说在理论计算机科学中,P"nC，解释深度学习的成功。少

2018 年 1 月 31 日提交;最初宣布 2018 年 1 月。

评论:6 页, 4 个数字

429. **建议: 1801. 10314[pdf,其他] Cs。Cl**

复杂的顺序问题回答：学习用知识图对链接的问题回答对进行转换

作者:amrita sahar, vardaan pahuja, mitesh m. khapra, karthik sankaranarayanan, sarath chandar

摘要: 在与聊天机器人交谈时，人类通常倾向于提出许多问题，其中很大一部分可以通过参考大规模的知识图 (kg) 来回答。虽然问题回答 (qa) 和对话系统是独立研究的，

但有必要仔细研究它们，以评估机器人所面临的涉及这两个任务的实际场景。为此，我们介绍了复杂顺序 qa 的任务，它结合了两个任务，即: (i) 通过复杂的推理，在数以百万计的实体的现实大小的 kg 上回答事实问题，以及 (ii) 学习通过一系列的一致地链接 qa 对。通过涉及内部和众包工人的劳动密集型**半自动流程**，我们创建了一个包含约 200k 对话框的数据集，总转弯时间为 160 万转。此外，与现有的大型 qa 数据集不同，这些数据集包含可从单个元组回答的简单问题，因此，我们对话框中的问题需要更大的 kg 子图。具体而言，我们的数据集存在需要逻辑、定量和比较推理及其组合的问题。这就需要一些模型: (i) 解析复杂的**自然语言**问题, (二) 使用会话上下文来解决话语中的相关和省略号, (ii) 要求澄清不明确的查询，最后 (四)检索 kg 的相关子图，以回答此类问题。然而，我们的实验结合了最先进的对话框和 qa 模型，表明他们显然没有达到上述目标，不足以处理如此复杂的现实世界设置。我们认为，这个新的数据集加上本文所报告的现有模型的局限性，应该会鼓励对复杂序列 qa 的进一步研究。少

2018 年 10 月 4 日提交;v1 于 2018 年 1 月 31 日提交;最初宣布 2018 年 1 月。

评论:接受于 aaaie18

430. **建议: 1801. 10296[pdf,其他] Cs。CI**

增强自知网络: 序列建模的硬注意力和软注意力的混合

作者:沈涛,周天一一,龙国东, 姜静,王森,张成奇

摘要: 许多**自然语言处理**任务完全依赖于句子中几个标记之间的稀疏依赖关系。软注意机制在通过每两个令牌之间的软概率建模本地/全局依赖关系方面表现出很有希望的性能，但在应用于长句时，它们并不有效和高效。相比之下，硬关注机制直接选择了令牌的子集，但由于其组合**性质**，难以训练，效率低下。为了相互的互惠互利，本文将软关注和硬关注整合到一个语境融合模型中, "增强自我关注 (resa)"。在 resa 中，硬注意力修剪一个序列，让软性的自我关注进入**过程**，而软注意力则反馈奖励信号，以方便硬注意力的训练。为此，我们开发了一种新的硬关注，称为 "强化序列采样 (rss)", 通过策略梯度并行和训练令牌。使用两个 rss 模块, resa 可以有效地提取每一对选定令牌之间的稀疏依赖关系。我们最后提出了一个无 rnn cnn 的句子编码模型, "增强的自我注意力网络 (resan)", 完全基于 resa。它在斯坦福**自然语言**推理 (snli) 和涉及合成知识 (sick) 的句子数据集上都取得了最先进的性能。少

2018 年 7 月 5 日提交;v1 于 2018 年 1 月 30 日提交;最初宣布 2018 年 1 月。

评论:9 页, 2 个数字; 在 ijcai-ecai-18 接受

431. **建议: 1801. 09896[pdf,其他] Cs。CI**

成本行动 "重新组装共和国" 的试点研究: 对 hartlib 论文信函的语言驱动的网络分析

作者:barbara mcillvray, fedico ssati

摘要: 本报告概述了我们在成本行动 is1310 "重新集结共和国, 1500-1800" 范围内进行的一项探索性研究，该研究与第 3 工作组 "案文和议题" 和第 2 工作组的活动有关。"人与网络"。在这项研究中，我们调查了**自然语言处理**(nlp) 和网络文本分析的使用在一个小样本的十七世纪的信件从 hartlib 文件中挑选，其记录是在一个目录早期现代在线信件 (emlo)，其在线版本可在谢菲尔德大学人文研究所的网站上查阅 (http://www.hrionline.ac.uk/hartlib/)。我们概述了用于将文本**自动处理**为网络表示形式的 nlp 管道，以确定文本的 "叙事中心"，即文本中最核心的实体，以及它们之间的关系。少

2018 年 1 月 30 日提交;最初宣布 2018 年 1 月。

评论:18 页, 7 个数字

432. 建议: 1801. 09718[pdf,其他] Cs。简历

vqa 中基于对象的推理

作者:mikyas t. desta, larry chen , tomasz kornuta

摘要: 视觉问题回答 (vqa) 是一个新的问题领域, 必须**处理**多模态输入, 以解决以**自然语言**形式给出的任务。由于解决方案本质上需要将视觉和**自然语言处理**与抽象推理结合起来, 因此该问题被认为是 ai 完全的。最近的进展表明, 使用从输入中提取的高层次、抽象的事实可能有助于推理。按照这个方向, 我们决定开发一个结合最先进的目标检测和推理模块的解决方案。在平衡的 klevr 数据集上取得的结果证实了承诺, 并在复杂的 "计数" 任务上显示出显著的、有几个百分点的准确性提高。少

2018 年 1 月 29 日提交;最初宣布 2018 年 1 月。

评论:10 页, 15 位数字, 作为会议文件发表在 2018 年 ieee 冬季会议版关于计算机视觉应用 (wacv' 2018)

433. 建议: 1801. 09041[pdf,其他] Cs。简历

告诉和回答: 使用属性和字幕实现可解释的视觉问题

作者:李青,傅建龙,余东飞,陶梅,罗洁波

文摘: 视觉问题回答 (vqa) 引起了计算机视觉和**自然语言处理**界的关注。大多数现有的方法采用通过预先培训的 cnn 表示图像的管道, 然后将不可解释的 cnn 功能与问题结合起来预测答案。尽管这样的端到端模型可能会报告有希望的性能, 但除了答案之外, 它们很少为 vqa **过程**提供任何见解。在这项工作中, 我们建议将端到端 vqa 分为两个步骤: 解释和推理, 试图通过揭示这两个步骤之间的中间结果来实现更可解释的 vqa。为此, 我们首先提取属性并生成描述, 分别使用预先训练的属性探测器和图像字幕模型作为图像的解释。接下来, 推理模块利用这些解释代替图像来推断问题的答案。这种故障的优点包括: (1) 属性和说明可以反映系统从图像中提取的内容, 从而为预测的答案提供一些解释;(2) 这些中间结果可以帮助我们在预测答案错误的情况下, 识别图像理解部分和答案推理部分的不能力。我们在一个流行的 vqa 数据集上进行了广泛的实验, 并根据对解释质量的几次测量对所有结果进行了剖析。我们的系统实现了与最先进的性能相当, 但具有更显著的可解释性和内在能力, 以更高质量的解释进一步改进的好处。少

2018 年 1 月 27 日提交;最初宣布 2018 年 1 月。

434. 建议: 1801. 09036[pdf, ps,其他] Cs。Cl

矛盾与分歧的一种模式。初步报告和讨论

作者:wlodek zadrozny, lu 西亚 a garbayo

摘要: 我们引入了一种新的形式化模型--基于鞘的数学构造--用于在文本源中表示矛盾信息。这种模式的优点是让我们 (a) 找出不一致的原因;(b) 衡量它的强度;(c) 并为此做些什么, 例如提出调和不一致的建议的方法。这种模式**自然代表了**矛盾和分歧之间的区别。它基于将**自然语言**句子表示为带有位于格子上的参数的公式的思想, 基于理论共享的谓词创建部分顺序, 并在这些部分顺序上与产品建立护套格子作为秸秆。不同程度是以全球和地方部分的存在来衡量的。还讨论了鞘方法的局限性和与自然 **语言处理**方面近期工作的联系, 以及物理、数据融合、拓扑数据分析和认识论方面的语境问题。少

2018 年 1 月 27 日提交;最初宣布 2018 年 1 月。

评论:本文在 isaim 2018 国际人工智能与数学研讨会上发表。2018 年 1 月 3 日，佛罗里达州劳德代尔堡。轻微的排版错误已得到纠正

435. 建议: 1801. 07757[pdf,其他] si

savitr: 一种在紧急情况下从微博中实时提取位置的系统

作者:ritam dutt, khakhah hiware, avijit ghosh,rameshwar bhaskaran

摘要: 我们介绍了 savitr，这是一个利用推特微博网站上发布的信息监测和分析紧急情况的系统。鉴于只有很小比例的微博被地理标记，因此这样的系统必须从微博的文本中提取位置。我们采用**自然语言处理**技术，以无监督的方式推断微博文本中提到的位置，并将其显示在基于地图的界面上。该系统设计用于高效性能，f 分达到 0.79，比其他可用的位置提取工具快两个数量级。少

2018 年 1 月 23 日提交;最初宣布 2018 年 1 月。

评论:smer-www 2018 提交

436. 建议: 1801. 07737[pdf] Cs。Cl

情感: 波斯人的情感分析语料库

作者:pe 徽 ram hosseini, ali ahahian ramaki, hassan maleki , Mansourehanvari , seyed Abolghasem mirroshandel

文摘: 情感分析 (sa) 是**自然语言处理**、计算语言学和信息检索领域的一个主要研究领域。近年来，学术界和产业界对南航的兴趣不断增长。此外，越来越需要生成适当的资源和数据集，特别是为包括波斯语在内的低资源**语言**生成适当的资源和数据集。这些数据集在使用监督、半监督或非监督方法设计和开发适当的意见挖掘平台方面发挥着重要作用。在本文中，我们概述了开发一个手动注释情感语料库 "情感语料库" 的整个**过程**，该语料库涵盖了正式和非正式的当代波斯语书面。据我们所知，sentipers 是一个独特的情感语料库，在三个不同的层面上都有如此丰富的注释，包括文档级、句派级和波斯语的 "纵横"。语料库包含超过 26000 句用户的意见，从数字产品领域，并受益于特殊的特点，如量化的正面或消极的意见，通过分配一个数字在特定范围内的任何给定句子。此外，我们还提供了关于语料库各组成部分的统计数据，并研究了注释者之间的相互注释协议。最后，我们在注释**过程**中面临的一些挑战也将得到讨论。少

2018 年 1 月 23 日提交;最初宣布 2018 年 1 月。

437. 建议: 1801. 06613[pdf,其他] Cs。Cl

构建一种具有椭圆意识的 web 文本依赖树库

作者:任宣成,孙旭,纪文 , 魏炳真,张伟东, 张志远

摘要: web2.0 带来了大量用户生成的数据，揭示了一个人的想法、经验和知识，这些都是许多任务的重要来源，如信息提取和知识库构建。然而，文本的口语性质给当前**的自然语言处理**技术带来了新的挑战，这些技术更适应语言的形式. 省略是一种常见的语言现象，有些词被忽略，因为它们被理解为上下文，特别是在口头话语中，阻碍了依赖分析的改进，这对依赖关系的意义是非常重要的。句子。为了促进这一领域的研究，我们将发布一个由 319 个小叶组成的中国依赖树库，其中包含 572 项判决，并保留了遗漏和上下文。少

2018 年 1 月 22 日提交;v1 于 2018 年 1 月 19 日提交;最初宣布 2018 年 1 月。

评论:树库可在
https://github.com/lancopku/Chinese-Dependency-Treebank-with-Ellipsis

438. **xiv:1801. 06480[pdf,其他] Cs。CI**

利用卷积神经网络进行文本分类迁移学习的实践者指南

作者:tushar semwal, gaurav mathur , profodyenigalla , Shivashankar b. nair

摘要: 当给定的数据集没有足够的标记示例来训练准确的模型时，迁移学习 (tl) 起着至关重要的作用。在这种情况下，在源数据集上预先训练的模型中积累的知识可以转移到目标数据集，从而改进目标模型。尽管 tl 在基于图像的应用领域取得了成功，但它在**自然语言处理**(nlp) 应用中的影响和实际应用仍是一个研究课题。由于其层次结构，深神经网络 (dnn) 在调整其参数和层深度方面提供了灵活性和定制性，从而形成了利用 tl 的适当区域。本文报道了利用卷积神经网络 (cnn) 进行的大量实证实验的结果和结论，并试图揭示拇指规则，以确保成功的正迁移。此外，我们还强调了可能导致负转让的有缺陷的手段。我们探讨了各层的可转移性，并描述了变化的超参数对转移性能的影响。此外，我们还提出了一个比较精度值和模型大小与最先进的方法。最后，我们从实证结果中得出推断，并提供最佳实践，以实现成功的正转移。少

2018 年 1 月 19 日提交;最初宣布 2018 年 1 月。

评论:2018 年 sdm 接受 9 页, 2 个数字

439. **建议: 1801. 06 271[pdf,其他] cse**

挖掘 android 应用程序用于生成可操作的基于 gui 的执行方案

作者:mario linares-vasquez, martin white, carlos bernal-cardenas, kevin moran, denys poshyvanyk

摘要: 从 android 应用执行跟踪、事件或源代码中提取的基于 gui 的模型对于具有挑战性的任务 (如方案或测试用例的生成) 非常有用。然而，提取有效的模型可能是一个昂贵的**过程**。此外，用于自动派生基于 gui 的模型的现有方法无法生成包含在执行 (也不是事件) 跟踪中未观察到的事件的方案。在本文中，我们讨论了这些和其他主要的挑战，在我们的新的混合方法，创造为猴子实验室。我们的方法基于记录 -mine-gene-verat-审评框架，该框架依赖于记录产生执行 (事件) 跟踪、挖掘这些事件跟踪并使用统计**语言**建模生成执行方案的应用用法,静态和动态分析，并使用在实际设备上交互式执行应用来验证生成的方案。该框架旨在挖掘能够为给定应用生成可生存和完全可重播 (即可操作) 方案的模型，这些方案反映**了自然**用户行为或不常见的用法 (例如，角落案例)。我们评估了 monkeylab 在一个案例研究涉及几个中型到大型开源 android 应用程序。我们的研究结果表明, monkeylab 能够挖掘基于 gui 的模型，这些模型可用于为 google nexus 7 平板电脑上的**自然**和非自然事件序列生成可操作的执行场景。少

2018 年 1 月 18 日提交;最初宣布 2018 年 1 月。

评论:12 页，接受第 12 届 ieee 采矿软件资料库工作会议 (msr' 15)

440. **建议: 1801. 06261[pdf,其他] Cs。CI**

文本分类器工作的研究

作者:devendra singh sachan, manzil zaheer, ruslan salakhutdinov

摘要: 文本分类是**自然语言处理**中研究最广泛的任务之一。在组合原理的推动下，采用了大型多层神经网络模型，试图有效地利用组成表达式。几乎所有报道的工作都使用判

别方法训练大型网络，这些方法伴随着没有适当容量控制的警告，因为它们往往会锁定任何可能无法概括的信号。利用最近各种最先进的文本分类方法，我们探讨这些模型是实际学习组成句子的含义，还是仍然只关注一些关键字或词典来对文档进行分类。为了检验我们的假设，我们仔细构建了训练和测试拆分没有直接重叠的数据集，但整体**语言**结构将是相似的。我们研究了各种文本分类器，并观察到这些数据集的性能下降很大。最后，我们证明，即使是简单的模型与我们提出的正则化技术，这破坏了集中在关键词典，可以大大提高分类的准确性。少

2018 年 8 月 5 日提交;v1 于 2018 年 1 月 18 日提交;最初宣布 2018 年 1 月。

评论:2018 年学术会议论文集，第 27 届国际计算语言学会议：技术论文 (coling 2018)，2017 年深度学习研讨会 NIPS：衔接理论与实践

441. **建议: 180006052[pdf] cs. cy**

高等教育中的大数据和学习分析：揭开多样性、获取、存储、nlp 和分析的神秘面纱

作者:阿迈勒·阿尔布拉维

摘要: 不同部门试图利用投资于大数据分析和**自然语言处理**的机会，以提高生产力和竞争力。高等教育部门目前面临的挑战包括迅速变化和不断变化的环境，这就需要制定新的思维方式。因此，作为解决高等教育许多问题的一部分，分析的兴趣增加了，包括学生自然减员率和学习者支持率。本研究对大数据、学习分析和 nlp 在高等教育中的应用进行了全面的探讨。此外，它还引入了一个集成的学习分析解决方案，利用分布式技术系统，支持教育机构的学术当局和顾问就个别学生作出决定。少

2018 年 1 月 3 日提交;最初宣布 2018 年 1 月。

评论:6 页, 2017 年 ieee 大数据与分析会议 (icbda)

442. **建议: 1801. 05574[pdf,其他] Cs。简历**

准离散度量和离散度量之间的最优传输方法

作者:陆英,陈立明, 赛迪，顾贤峰

摘要: 正确估计两种数据分布之间的差异一直是机器学习中的一项重要任务。最近，cuturi 提出了辛克霍恩距离，利用两个分布之间的近似最优运输成本作为距离来描述分布差异。虽然自那时以来，辛克霍恩的距离也受到两个不可忽视的限制，但它已成功地应用于各种机器学习应用 (例如自然**语言处理**和计算机视觉)。.第一个是辛克霍恩距离只给出真实沃瑟斯坦距离的近似值，第二个是在矩阵缩放过程中经常发生的 "除以零" 问题，当熵正则化系数设置为较小的值时。在本文中，我们引入了一种新的 brenier 方法来计算两个离散分布之间更精确的 wasserstein 距离，这种方法成功地避免了上述对辛克霍恩距离的两个限制，并给出了另一种方法。估计分配差异。少

2018 年 1 月 17 日提交;最初宣布 2018 年 1 月。

443. **建议: 1801. 05568[pdf,其他] Cs。简历**

多伊 10.1109/ICIIECS.2017.8276124

使用深层神经体系结构的图像字幕

作者:parth shah, Vishvajit bakarola, supriya 帕蒂

摘要: 使用任何**自然语言**句子 (如英语) 自动创建图像的描述是一项非常具有挑战性的任务。它需要图像**处理**和**自然语言处理**方面的专业知识。本文讨论了图像字幕任务的不同可用模型。我们还讨论了近年来对象识别和机器翻译任务的进展如何大大提高了

图像字幕模型的性能。除此之外，我们还讨论了如何实现此模型。最后，我们还使用标准评价矩阵对模型的性能进行了评价。少

2018 年 1 月 17 日提交;最初宣布 2018 年 1 月。

评论:2017 年信息嵌入式和通信系统创新国际会议 (iciecs) 接受的预印版论文

444. 新建: 1801 1.05206[pdf, ps,其他] Cs。Db

序列，但功能: 数据流处理的双重性质

作者:sebastian herbst, johes tenschert, anderas m. wahl, klaus Meyer-Wegener

摘要: 数据流**处理**的重要性不断增加，因为在过去十年中，可用数据量一直在稳步增加。除了传统的领域 (如数据中心监控和单击分析) 外，还有越来越多的支持网络的生产机器生成连续的数据流。由于其连续性,对数据流的查询可能会更加复杂，而且与数据库查询相比显然很难理解。由于用户必须考虑操作细节，维护和调试变得具有挑战性。目前的方法将数据流建模为序列，因为这就是数据流的物理接收方式。这些模型产生了一个以实现为中心的视角。我们通过关注时间切片语义来探索一种建模数据流的替代方法。这个焦点产生了一个基于函数的模型，它更适合于查询语义的推理。通过将流**处理**中相关概念的定义与模型相适应，说明了我们的方法的实际用途。因此，我们将数据流和查询原语链接到函数式编程和数学中的概念。最值得注意的是，我们证明了数据流是单元的，并展示了如何为当前数据流模型派生单声道定义。我们基于合理、一致的查询模型，为数据流相关主题提供了一个抽象而实用的视角。我们的工作可以为未来的数据流查询语言奠定坚实的基础。少

2018 年 1 月 16 日提交;最初宣布 2018 年 1 月。

445. xiv:1801. 04821[pdf,其他] Cs。直流

改善多面体过程网络中的通信模式

作者:christophe alias

摘要: 嵌入式系统性能受功耗的限制。趋势是卸载 gpu、xeon phi 或 fpga 等硬件加速器上的贪婪计算。fpga 芯片将可编程芯片的灵活性和专业硬件的能效结合起来，成为**一种天然**的解决方案。需要高级**语言**的硬件编译器 (高级合成、hls) 来利用 fpga 的所有功能，同时满足严格的上市时间限制。编译器对并行性和数据局部性的优化深入地重构了**进程**的执行顺序，从而在通信通道中形成了读写模式。这将中断大多数 fifo 通道，这些通道必须使用可寻址缓冲区来实现。实施同步需要昂贵的硬件，这通常会导致严重的性能损失。本文提出了一种对通信进行分区的算法，使大多数 fifo 信道在循环平铺后都能恢复，这是并行性和数据局部性的关键优化。实验结果表明，常规内核的 fifo 检测有了很大的改进，而需要一些额外的存储。另外，在某些情况下，存储甚至可以减少。少

2018 年 1 月 15 日提交;最初宣布 2018 年 1 月。

评论:2018 年在 hip3 展会上提交

报告编号:hip3es2015/

446. 建议: 1801 1.04223[pdf,其他] cs. it

下一代上下文感知无线网络的机器智能技术

作者:tadilo Endeshaw bogale, xibin wang, long bao le

摘要: 下一代无线网络 (即 5g 及更高) 由于异构网络 (hetnet) 的超密集部署，将是极其动态和复杂的，这对网络规划、运营、管理和网络构成了许多严峻挑战。故障 排除。

与此同时，随着从以人为本向面向机器的通信不断转变，无线数据的生成和使用日益分布，使得未来无线网络的运营更加复杂。在降低未来网络操作的复杂性方面，智能利用分布式计算资源并提高上下文感知的新方法变得极为重要。在这方面，新出现的旨在将计算、存储、控制、通信和网络功能分布在更接近最终用户的位置的迷雾（边缘）计算架构，对于实现未来无线网络的高效运行具有巨大潜力。这些有前途的架构使采用人工智能（ai）原则，其中包括学习、推理和决策机制，作为设计紧密集成的网络的**自然**选择。为此，本文全面介绍了集成机器学习、数据分析和**自然语言处理**(nlp）技术的人工智能在提高无线效率方面的应用情况。网络操作。特别是，我们全面讨论了这些技术在下一代无线网络的高效数据采集、知识发现、网络规划、运营和管理方面的应用。还为这一网络提供了利用 ai 技术的简要案例研究。少

2018 年 1 月 12 日提交;最初宣布 2018 年 1 月。

评论:国际电联特刊 n.1 人工智能（ai）对通信网络和服务的影响（现已出版）

447. **建议: 1801. 03911[pdf,其他] Cs。CI**

自然语言建模中的非平稳内核随机学习

作者:sahalil garg, greg ver steeg, aram galstyan

摘要: **天然的的语言处理**通常涉及使用语义图或句法图进行计算，以促进基于结构关系的复杂推理。卷积内核为基于节点（单词）级别关系比较图形结构提供了强大的工具，但它们很难自定义，并且可能会产生很高的计算开销。我们提出了卷积内核的推广，一个非平稳的模型，以更好地表达**自然语言**在监督设置。为了对我们模型引入的参数进行可扩展的学习，我们提出了一种新的算法，该算法利用 k-最近邻域图上的随机采样，以及基于位置敏感哈希的近似。我们展示了我们的方法在一个具有挑战性的现实世界（结构化推理）问题上的优势，即从科学论文的文本中自动提取生物模型。少

2018 年 2 月 1 日提交;v1 于 2018 年 1 月 11 日提交;最初宣布 2018 年 1 月。

448. **建议: 1801. 03564[pdf, ps,其他] Cs。CI**

无监督的语音部分归纳

作者:奥米德·卡希菲

摘要: 语音部分（pos）标记是**自然语言处理**中一项古老而基本的任务。虽然受监督的 pos 标签已显示出很有希望的准确性，但由于缺乏标记数据，使用受监督的方法并不总是可行的。在这个项目中，我们试图通过迭代地通过分层聚集聚类分析过程寻找反复出现的单词模式来不出意外地诱导 pos 标记。与最先进的无监督 pos 标签的标记结果相比，我们的方法显示了很有希望的结果。少

2018 年 1 月 10 日提交;最初宣布 2018 年 1 月。

449. **arxiv:1801. 0306[pdf,其他] Cs。红外**

深度搜索：基于内容的图像搜索和检索

作者:tanya piplani, david bamman

摘要: 今天的互联网大多由包括视频和图像在内的数字媒体组成。随着像素成为大多数交易在互联网上发生的货币，有一种相对轻松地浏览这片信息海洋的方式变得越来越重要。youtube 每分钟上传 400 个小时的视频，在 instagam、facebook 等网站上浏览了数百万张图片。在深刻学习和成功领域最近取得的进步的启发下，我们提出了一种"寻找一种**自然 语言"，这些进步在图像字幕、机器翻译、单词 2vec、跳过思想等各种问题上都取得了进步。基于处理**的深度学习模型，允许用户输入他们要搜索的图像类型

的描述，作为响应，系统检索所有在语义上和上下文中与查询相关的图像。以下各节介绍了两种方法。少

2018 年 1 月 11 日提交;v1 于 2018 年 1 月 9 日提交;最初宣布 2018 年 1 月。

评论:arxiv 管理说明: 文本重叠与 arxiv:1706.0 6064 由其他作者

450. 建议: 1801. 02607[pdf,其他] Cs。红外

web2 文本: 深层结构化木板去除

作者:thijs vogels, octavian-eugen ganea, carsten eickhoff

摘要: 网页是许多**自然语言处理**和信息检索任务的宝贵信息来源。从这些文档中提取主要内容对于派生应用程序的性能至关重要。为了解决这个问题，我们引入了一个新的模型，该模型执行序列标记，将 html 页面中的所有文本块统称为样板或主要内容。我们的方法使用一个隐藏的马尔可夫模型，在从 dom 树特征中获得的潜力之上，使用卷积神经网络。该方法为在 cleaneval 基准上的样板去除提供了新的最先进的性能。作为信息检索管道的一个组成部分，它提高了对 clueweb12 集合的检索性能。少

2018 年 3 月 27 日提交;v1 于 2018 年 1 月 8 日提交;最初宣布 2018 年 1 月。

评论:出现在 ecir 2018 中

451. 建议: 1801. 02581[pdf, ps,其他] Cs。Cl

分类器和代码混合因子在情绪识别中的作用分析

作者:soumil mandal, Dipankar das

摘要: 讲多种语言的人经常在**语言**之间切换，以便在社交平台上表达自己的意见。有时，**保留语言**的原始脚本，而对所有**语言**使用通用脚本也相当流行，因为方便。在这种情况下，多种**语言**混合着不同的语法规则，使用相同的脚本，即使在准确情绪的情况下，这也是**自然语言处理**的一项具有挑战性的任务识别。本文报道了在具有英语和孟加拉语两**种语言**的代码混合特性的电影评论数据集上进行的各种实验的结果，这两种语言都是用罗马文字输入的。我们测试了仅在代码混合数据上训练过的各种机器学习算法，并使用朴素贝叶斯 (nb) 模型达到了 99.00% 的最高精度。我们还测试了在代码混合数据上训练的各种模型，以及英语功能，通过支持向量机 (svm) 模型获得了 72.50 的最高精度。最后，我们分析了错误分类的代码段，并讨论了需要解决的挑战，以提高准确性。少

2018 年 3 月 15 日提交;v1 于 2018 年 1 月 8 日提交;最初宣布 2018 年 1 月。

评论:第十八届计算语言学与智能文本处理国际会议, cicling 2017 (rcs)

452. xiv:1801.02107[pdf,其他] Cs。Cl

米山: 一个大的波斯-英平行语料库

作者:奥米德·卡希菲

摘要: **自然语言处理**中最主要和最重要的任务之一是机器翻译，它现在高度依赖多语言并行语料库。通过本文介绍了从文学名著中收集到的超过一百万对的最大的波斯-英平行语料库。我们还提出了语料库的采集**过程**和统计，并利用语料库对基线统计机器翻译系统进行了实验。少

2018 年 1 月 10 日提交;v1 于 2018 年 1 月 6 日提交;最初宣布 2018 年 1 月。

453. xiv:1801. 02054[pdf] Cs。Cl

英语诗歌语料库中的探索: 一种神经认知诗学的视角

作者:arthur m. jacobs

摘要: 本文描述了一个由大约3000个英国文学文本组成的语料库，其中约有2.5亿字从古腾堡项目中提取，涵盖了 130 多位作者 (如达尔文、狄更斯、莎士比亚) 撰写的小说和非小说的一系列流派。定量叙事分析 (qna) 被用来探索一个干净的子语料库，古腾堡英国诗歌语料库 (gepc)，其中包括100多个诗意文本与大约200万字约来自50位作者 (例如，济慈，乔伊斯，华兹华兹)。一些典型的 qna 研究显示作者的相似之处基于潜在的语义分析，重要的主题为每个作者或各种文本分析指标为乔治·艾略特的诗 "丽莎如何爱国王" 和詹姆斯·乔伊斯的 "室内乐"，例如。词汇多样性或情绪分析。gepc 特别适用于数字人文、**自然语言处理**或神经认知诗学的研究，例如作为培训和测试语料库，或用于刺激开发和控制。少

2018 年 1 月 6 日提交;最初宣布 2018 年 1 月。

评论:27 页, 4 个数字

454. **新建: 1801.0 1900[pdf,其他] Cs。Cl**

基于知识的词感消歧--基于主题模型的

作者:devendra singh chaplot, ruslan salakhutdinov

摘要: 在**自然语言处理**中，语感消歧是一个悬而未决的问题，在无监督环境中，任何给定文本中的所有单词都需要在不使用的情况下消除歧义，这是一个特别具有挑战性和有用的问题任何标记的数据。通常，wsd 系统使用句子或目标单词周围的一个小窗口作为消除歧义的上下文，因为它们的计算复杂性会随着上下文的大小呈指数级扩展。本文利用主题模型的形式主义，设计了一个水务署系统，该系统可根据上下文中的单词数量线性扩展。因此，我们的系统能够利用整个文件作为上下文，消除一个词的歧义。建议的方法是潜在 dirichlet 分配的变体，在该变量中，文档的主题比例被同步比例所取代。我们进一步利用 wordnet 中的信息，在对单词进行同步分发之前分配一个非均匀的信息，在同步上分发文档之前分配一个逻辑法线。我们对 SemEval-、SemEval-、semeval-2007、semeval-2013 和 semeval-2015 英语全字水务系统的拟议方法进行了评估，并显示其性能大大优于最先进的无监督知识的水务署系统。少

2018 年 1 月 5 日提交;最初宣布 2018 年 1 月。

评论:出现在 aaai-18 中

455. **建议: 1801. 01828[pdf, ps,其他] Cs。Cl**

屏蔽谷歌对抗攻击的语言毒性模型

作者:nestor rodriguez, sergio rojas-galeano

摘要: 网络社区缺乏节制使参与者能够遭受个人侵犯、骚扰或网络欺凌，这些问题在当代后真相政治情景中因极端主义激进化而加剧。这种敌意通常是通过有毒**的语言**、亵渎或辱骂性的言论来表达的。最近，谷歌开发了一种基于机器学习的毒性模型，试图评估评论的敌意;不幸的是 ，有人认为 ，上述模式可能会纵评论文本序列的对抗性攻击所欺骗。本文首先将这种对抗性攻击描述为使用混淆和极性变换。前者通过排版编辑来腐蚀有毒触发内容，而后者则通过对有毒内容的语法否定来欺骗。然后，我们提出了一个两阶段的方法来对抗这些异常，在最近提出的文本去混淆方法和毒性评分模型的基础上。最后，我们进行了一个实验，约 24000 条扭曲的评论，展示了如何以这种方式恢复对抗变种的毒性是可行的，同时导致处理时间的**大约**增加了两倍。尽管新的对手挑战将不断地来自于书面**语言**的多才多艺，但我们预计，将机器学习和文本模式识别方

法结合起来的技术，每一种方法都是针对的需要不同层次的语言特征来实现对有毒**语言**的鲁棒检测，从而促进无侵略的数字交互。少

2018 年 1 月 5 日提交;最初宣布 2018 年 1 月。

456. 建议: 1801.01331[pdf,其他] Cs。CI

多伊10.186537/18-12 个月

越南自然语言处理工具包

作者:thanh vu, dat quoc nguyen, dai quoc nguyen, mark dras, mark johnson

摘要: 我们提出了一个易于使用和快速的工具包，即 vncornlp---越南语的 java nlp 注释管道。我们的 VnCoreNLP 支持关键**的自然语言处理**(nlp) 任务，包括分词、词性 (pos) 标记、命名实体识别 (ner) 和依赖关系分析，并获得最先进的 (sota) 这些任务的结果。我们发布 vncornlp 以提供丰富的语言注释，以便利越南 nlp 的研究工作。我们的 vncornlp 是开源的，可在: https://github.com/vncorenlp/VnCoreNLP

2018 年 4 月 1 日提交;v1 于 2018 年 1 月 4 日提交;最初宣布 2018 年 1 月。

评论:计算语言学协会北美分会 2018 年会议论文集: 示威, naacl 2018，将出现

457. 第 1712.09495[pdf,其他] lo c

多伊10.4204/EPTCS.263。2

在自由超图类别中重写

作者:法比奥·扎纳西

摘要: 我们研究了在对称单线范畴的背景下，在每个物体上都有一个可分离的弗罗比尼斯一元数的情况下，对方程理论进行重写。这些类别，也称为超图类别，越来越重要: frobenius 结构最近出现在跨学科应用中，包括量子**过程**、动力系统和**自然**的研究**的语言处理**。在这项工作中，我们给出了一个组合的箭头的自由超图类别的共同示称为标记超图的共同应，并建立了一个精确的对应之间重写模量 frobenius 结构和双推推重写超图上的其他。这种解释允许在超图上使用结果，以确保可在自由超图类别中重写融合的可识别性。我们的结果概括了以前的方法，其中只考虑由单个对象 (道具) 生成的类别。少

2018 年 1 月 3 日提交;v1 于 2017 年 12 月 27 日提交;最初宣布 2017 年 12 月。

评论:《2017 年程序中的 gam 》, arxiv:1712.08 345

日记本参考:eptcs 263, 2017，第 16-30 页

458. 第: 1712.08841[pdf,其他] Cs。CI

用于子字符表示学习的双长短期内存网络

作者:韩河、吴磊、杨晓坤、华燕、高志敏、易峰、 乔治·汤森

文摘: 在**自然语言处理**(nlp) 中，字符通常被认为是最小的**处理**单元。但许多非拉丁**语都**有象形文字书写系统，涉及一个包含数千个或数百万个字符的大字母表。每个字符都由更小的部分组成，而这些部分往往被以前的工作所忽略。在本文中，我们提出了一个新的架构，使用两个堆叠的长期短期记忆网络 (lstm) 来学习子字符级别表示和捕捉更深层次的语义意义。为了建立一个具体的研究和证实我们的神经结构的效率，我们以汉语词分割为例。在这些**语言**中，汉语是一个典型的例子，每个字符都包含几个叫做自由基的成分。我们的网络采用共享的激进水平嵌入来解决简体中文和繁体中文的单词分割，没有额外的繁体中文到简体中文转换，在这种高度的端到端方式下，分词可以显著与以前的工作相比简化了。激进层次的嵌入也可以捕捉到字符级别以下更深层次的语义

意义，提高系统的学习性能。通过将激进和字符嵌入捆绑在一起，减少了参数计数，而语义知识则在两个层次之间共享和转移，大大提高了性能。在 4 个 bakeoff 2005 数据集中中，有 3 个数据集的最先进结果超过了最先进的结果，最高可达 0.4%。我们的结果是可重复的，源代码和语料库可在 github 上获得。少

2018 年 1 月 4 日提交;v1 于 2017 年 12 月 23 日提交;最初宣布 2017 年 12 月。

评论:2018 年接受及推出

459. 第 (xiv:1712.08 207)[pdf,其他] Cs。Cl

序列到序列模型的变化注意

作者:hareesh bahuleyan, lili ou , olga vechtomova, pascal poupart

摘要: 变分编码解码器 (ved) 使用神经网络将源信息编码为一组随机变量，而神经网络又使用另一个神经网络解码为目标数据。在**自然语言处理**中，序列到序列 (seq2seq) 模型通常用作编码器解码器网络。当与传统的 (确定性) 注意机制相结合时，注意力模型可能会绕过变分的潜在空间，从而变得无效。本文提出了一种 ved 的变分注意机制，将注意向量也建模为高斯分布随机变量。两个实验结果表明，在不损失质量的情况下，我们提出的方法可以缓解旁路现象，因为它增加了生成句子的多样性。少

2018 年 6 月 21 日提交;v1 于 2017 年 12 月 21 日提交;最初宣布 2017 年 12 月。

评论:在《2018 年涂装论文集》中。同样被 tardgm 研讨会 @ icml 2018 年接受，以供介绍

460. 第 1712.07008[pdf,其他] Cs。Cl

预期 BLEU 分数的可区分下限

作者:vlad zhukov, eugene golikov, maksim kretov

摘要: 在**自然语言处理**任务中，模型的性能通常是用一些不可区分的度量来衡量的，比如 BLEU 的分数。为了使用高效的基于梯度的方法进行优化，优化一些代理丢失函数是一种常见的解决方法。如果这种损失的优化也导致目标度量的改进，这种方法是有效的。相应的问题称为损失评估不匹配。在本工作中，我们提出了一种计算预期 BLEU 分数的可微下限的方法，该方法不涉及计算成本高昂的采样过程，例如使用增强学习 (rl) 中的 reforce 规则时所需的采样过程框架。少

2018 年 8 月 23 日提交;v1 于 2017 年 12 月 13 日提交;最初宣布 2017 年 12 月。

评论:在 2017 年 NIPS 对话 ai 研讨会上发表: 今天的实践和明天的潜力

461. 第 1712.03607 Cs。Lg

基于渐变归一化和深度的深度学习衰变

作者:robert kwiatkowski, oscar chang

摘要: 本文介绍了一种新的梯度归一化和衰变的深度方法。我们的方法利用了在深度神经网络中对所有梯度进行归一化的简单概念，然后根据它们在网络中的深度对所述梯度进行衰减。我们建议的规范化和衰变技术可以与最新的艺术优化器结合使用，并且是对任何网络的一个非常简单的补充。这种方法虽然简单，但在图像分类任务的最先进网络 (如 densenet 和 resnet) 以及用于**自然语言处理**任务的 lstm 上的融合时间有所改善。少

2018 年 2 月 28 日提交;v1 于 2017 年 12 月 10 日提交;最初宣布 2017 年 12 月。

评论:当时的结果似乎更有希望

462. 第 1712.01336 si

听混乱的窃窃私语：面向新的股票趋势预测的深层学习框架

作者:胡子牛,刘伟清,姜边,刘玄哲,刘铁燕

摘要: 股票趋势预测在从股票投资中寻求利润最大化方面发挥着至关重要的作用。然而，由于股市的高度波动和非平稳**性质**，精确的趋势预测非常困难。互联网上信息的爆炸化，以及**自然语言处理**和文本挖掘技术的不断发展，使投资者能够从在线内容中揭示市场趋势和波动性。不幸的是，与股票市场相关的在线内容的质量、可信度和全面性差别很大，其中很大一部分是低质量的新闻、评论，甚至谣言。为了应对这一挑战，我们模仿人类面对如此混乱的网络新闻的学习过程，其驱动原则有三个：顺序内容依赖、多样化的影响以及有效和高效的学习.本文以前两个原理为中心，设计了一种混合注意力网络，根据近期相关新闻的顺序预测股票走势。此外，我们还运用自定进度的学习机制来模仿第三个原则。对真实世界股市数据的大量实验证明了我们方法的有效性。少

2018 年 3 月 16 日提交;v1 于 2017 年 12 月 6 日提交;最初宣布 2017 年 12 月。

评论:(1) ms 形式 (作者的组织) 计划对该技术申请专利，本文不包括相应的确认，因此我们需要暂时撤销该专利。(二) 实验细节不完整，可能会让读者感到困惑，我们需要细化细节，避免给读者带来不必要的麻烦 "

463. 第: 1712.0 2007[pdf,其他] Cs。Hc

多伊 10.114/317292944.3173007

将故事与可视化耦合：利用文本分析作为数据与解释之间的桥梁

作者:ronald meteyer, qiyu zhi, bart janczuk, walter cheirer

摘要: 网络作家和新闻媒体越来越多地将可视化 (和其他多媒体内容) 与叙事文本结合起来，以创建叙事可视化。然而，这两个元素往往是相互独立地呈现的。我们提出了一种自动集成文本和可视化元素的方法。我们从一个作家的叙述开始，大概可以用视觉数据证据来支持。我们利用**自然语言处理**、定量叙事分析和信息可视化来 (1) 自动从数据丰富的故事中提取叙述成分 (谁、什么、什么、什么时候、在哪里)，以及 (2)将支持数据证据与文本集成，以形成叙事可视化。我们还采用双向交互，从文本到可视化和可视化，再到文本，以支持读者在两个方向上的探索。我们通过在体育新闻数据丰富领域的案例研究来演示该方法。少

2018 年 1 月 6 日提交;v1 于 2017 年 12 月 5 日提交;最初宣布 2017 年 12 月。

评论:18, 3 个数字, 5 页

464. 第: 1712.0027[pdf,其他] Cs。Lg

基于内核的采样自适应采样软值

作者:guy blanc, steffen rendle

摘要: softmax 是用于多级问题的最常用的输出函数，广泛应用于视觉、**自然语言处理**和推荐等领域。softmax 模型在类的数量上具有线性成本，这使得它对于许多实际问题来说过于昂贵。加快培训的常见方法是在每个培训步骤中只对部分课程进行抽样。已知这种方法是偏置的，偏置增加的采样分布偏离输出分布越多。然而，几乎所有最近的工作都使用简单的采样分布，这些分布需要较大的采样大小来缓解偏差。在本文中，我们提出了一类新的基于内核的采样方法，并开发了一种有效的采样算法。基于内核的采样在模型训练时对其进行调整，从而导致低偏差。基于内核的采样可以很容易地应用于许多模型，因为它只依赖于模型的最后一个隐藏层。我们对偏置、采样分布和采样大

小的权衡进行了实验研究，结果表明，基于核的采样结果在样本较少的情况下具有较低的偏置。少

2018 年 8 月 1 日提交;v1 于 2017 年 12 月 1 日提交;最初宣布 2017 年 12 月。

465. 第: 1711. 10233[pdf, ps,其他] lo c

定时系统的行为等效性

作者:tomasz brbrgos, marco peressotti

摘要: 定时转换系统是行为模型，包括对时间流的显式处理，并用于将几个基础**过程**演算和自动机的语义形式化。尽管它们具有相关性，但对定时过渡系统及其行为理论的一般数学特征仍然缺失。我们引入了第一个定时行为模型的统一框架，其中包括已知的行为等效，如定时二重计算、定时**语言等价以及**弱和时间抽象的对应。所有这些等价的概念都是由它们在光谱中的判别力量**自然组织起来**的。我们证明，这一结果并不取决于所审查的系统的类型：它适用于任何对定时过渡制度的概括。我们实例化了我们的框架，以实现定时转换系统及其定量扩展，如定时概率系统。少

2018 年 10 月 1 日提交;v1 于 2017 年 11 月 28 日提交;最初宣布 2017 年 11 月。

类:F.1。1

466. 第: 1711. 10203[pdf,其他] Cs。Cl

可视化和 "诊断分类器" 揭示了递归和递归神经网络如何处理层次结构

作者:Dieuwke hupkes, sara veldhoen , willem zuidema

文摘: 我们研究神经网络如何使用分层的组合语义来学习和**处理语言**。为此，我们定义了**处理**嵌套算术表达式的人工任务，并研究了不同类型的神经网络是否可以学习计算其含义。我们发现递归神经网络可以找到这个问题的通用解决方案，我们通过三个步骤将它分解来可视化这个解决方案：项目、求和和壁球。作为下一步，我们将研究递归神经网络，并显示以增量方式**处理**其输入的门控递归单元在执行此任务时也表现得非常好。要了解反复出现的网络编码的内容，光靠可视化技术是不够的。因此，我们开发了一种方法，在这种方法中，我们对网络编码和**处理**的信息进行了多项假设的制定和测试。对于每个假设，我们都会对每个时间步长的隐藏状态表示特征进行预测，并训练 "诊断分类器" 来测试这些预测。我们的研究结果表明，这些网络遵循的策略类似于我们假设的 "累积策略"，这解释了网络在新表达式上的高精度，对比训练中看到的表达时间更长的概括，以及温和的随着长度的增加而恶化。这表明，诊断分类器可以是一个有用的技术，打开神经网络的黑匣子。我们认为，与大多数可视化技术不同的是，诊断分类确实从玩具领域中的小型网络扩展到处理真实数据的更大、更深的重复网络，因此可能有助于更好地了解**自然语言处理**领域当前最先进模型的内部动态。少

2018 年 4 月 20 日提交;v1 于 2017 年 11 月 28 日提交;最初宣布 2017 年 11 月。

评论:20 页

日记本参考:人工智能研究杂志 61 (2018) 907-926

467. 第: 1711.07950[pdf,其他] Cs。Cl

掌握地牢：机械车的基础语言学习

作者:杨志林,张赛正,杰克·乌尔班内克，威尔·冯,亚历山大·h·米勒,亚瑟·斯庄，杜维·基拉,杰森·韦斯顿

摘要: 与大多数**利用**静态数据集的自然**语言处理**研究相反，人类以基于环境的方式以交互方式学习**语言**。在这项工作中，我们提出了一个互动学习程序称为机械涡轮下降

(mtd)，并使用它来训练代理执行**自然语言**命令接地在一个幻想文本冒险游戏。在 mtd 中，turkers 在短期内竞相培训更好的代理商，并通过长期分享其代理的技能进行合作。与静态数据集相比，这将为 turkers 带来一种游戏化、引人入胜的体验，并为代理提供更高质量的教学信号，因为 turkers **自然会**根据代理的能力调整训练数据。少

2018 年 4 月 16 日提交;v1 于 2017 年 11 月 21 日提交;最初宣布 2017 年 11 月。

468. 第 711.07 280[pdf,其他] Cs。简历

视觉和语言导航：解释实际环境中的可视化导航指令

作者:peter anderson, qi wu, damien teney , jake bruce, mark johnson , niko sünderhauf, ian reid, stephen gould, anton van den hngel

摘要：一个能够进行**自然语言**教学的机器人一直是一个梦想，因为在杰森卡通系列想象的休闲生活由一群细心的机器人助手调解之前。这是一个顽固的梦想。然而，最近在视觉和**语言**方法方面取得的进展在密切相关的领域取得了令人难以置信的进展。这一点很重要，因为机器人根据所看到的情况来解释**自然语言**导航指令，正在进行类似于视觉问题回答的视觉和**语言过程**。这两个任务都可以解释为视觉上接地的序列序列转换问题，并且许多相同的方法都适用。为了支持和鼓励视觉和**语言**方法的应用，以解释基于视觉的导航指令的问题，我们提出了 matterport3d 模拟器-一个大规模的增强学习环境，基于真实的图像。使用这个模拟器，它可以在未来支持一系列体现的视觉和**语言**任务，我们提供第一个基准数据集的可视接地自然 **语言**导航在真正的建筑物-房间到房间 (r2r) 数据集。少

2018 年 4 月 5 日提交;v1 于 2017 年 11 月 20 日提交;最初宣布 2017 年 11 月。

评论:cvpr 2018 聚焦演示

469. 建议: 1711.06744[pdf,其他] Cs。CI

学习用 n-gram 机器组织知识和回答问题

作者:范阳,聂家中，威廉·科恩,聂老妮

文摘：尽管深度神经网络在**自然语言处理**方面取得了巨大的成功，但它们在知识密集型 ai 任务 (如开放域问答 (qa) 方面是有限的。现有的端到端深度 qa 模型需要在观察问题后对整个文本进行**处理**，因此它们在回答问题时的复杂性在文本大小中是线性的。这对于实际任务 (如维基百科的 qa、小说或 web) 来说是令人望而却步的。我们建议通过使用符号意义表示来解决这个可伸缩性问题，它可以在与文本大小无关的复杂性下有效地编制索引和检索。我们将我们的方法 (称为 n-gram 机器 (ngm)) 应用于三项具有代表性的任务。首先，作为概念的验证，我们证明了 ngm 成功地解决了合成文本的 babi 任务。其次，我们通过对 "终身 babi" 的实验表明，ngm 可以扩展到大型语料库，这是一个包含数百万句话的特殊版本 babi。最后，在维基电影数据集上，我们利用 ngm 诱导潜在结构 (即模式)，并回答**来自自然语言**维基百科文本的问题，只有 qa 对作为弱监管。少

2018 年 7 月 1 日提交;v1 于 2017 年 11 月 17 日提交;最初宣布 2017 年 11 月。

评论:提交给 NIPS

470. 第: 1711. 06420[pdf,其他] Cs。简历

外观、想象和匹配：使用生成模型改进文本-视觉跨模检索

作者:顾九祥,蔡建飞,乔蒂，李牛,王刚

文摘: 文本视觉交叉模态检索一直是计算机视觉和**自然语言处理**界的研究热点。学习多模态数据的适当表示对于跨模态检索性能至关重要。与现有的图像文本检索方法将图像文本对嵌入到一个共同的表示空间中作为单个特征向量不同，我们建议将生成**过程合并**到交叉模态特征嵌入中，通过这种嵌入，我们不仅能够学习全局抽象特征，还可以了解局部接地特征。大量实验表明，该框架能够很好地匹配具有复杂内容的图像和句子，并在 mscoco 数据集上实现了最先进的跨模态检索结果。少

2018 年 6 月 13 日提交;v1 于 2017 年 11 月 17 日提交;最初宣布 2017 年 11 月。

评论:10 页, 6 位数字，在 2018 年 cvpr 会议上被列为焦点

471. **建议: 1711. 06 128[pdf, ps,其他] Cs。艾**

使用法律规则启用推理

作者:林浩盘,穆斯塔法·哈什米

摘要: 为了实现验证过程的自动化,需要**将用自然语言**编写的法规规则转换为机器能够理解的格式。然而，现有的任何形式主义都不能充分代表法律规范中出现的要素。例如，这些形式主义大多不提供捕获神性效果行为的功能，这是自动合规性检查的一个重要方面。本文提出了一种将使用 legal 示 ml 表示的法律规范转换为模态 defeaseic 变体 (反之亦然) 的方法，以便将使用 legal 示 ml 表示的法律声明转换为机器可读的格式，从而可以理解和推理取决于客户的喜好。少

2018 年 4 月 7 日提交;v1 于 2017 年 11 月 11 日提交;最初宣布 2017 年 11 月。

评论:25 页。逻辑规划理论与实践 (tplp) 中的出版考虑

类:F.4.2;F.1。1

472. **第 1711. 05408[pdf,其他] Cs。佛罗里达州**

作为加权语言识别器的递归神经网络

作者:陈一宁,索查·吉尔罗伊,安德烈亚斯·马莱蒂,乔纳森·梅,凯文·耐特

文摘: 我们研究了简单递归神经网络 (rnn) 作为识别加权**语言**的形式化模型的各种问题的计算复杂度。我们专注于单层、刷新 u 激活、具有 softmax 的自重 rnn，这些技术通常用于**自然语言处理**应用。我们证明了此类 rnn 的大多数问题是不可判定的，包括一致性、等价性、最小化性和最高权重字符串的确定。但是，对于一致的 rnn，最后一个问题可以判定，尽管解决方案长度可以超过所有可计算的边界。如果另外串被限制为多项式长度，问题变得 np 完全和 apx 坚硬。总之，这表明，在这些 rnn 的实际应用中，近似和启发式算法是必要的.

2018 年 3 月 4 日提交;v1 于 2017 年 11 月 14 日提交;最初宣布 2017 年 11 月。

473. **第: 1711.05066[pdf,其他] Cs。Cl**

学习可执行的神经语义解析器

作者:jenpengcheng, siva reddy, vijay saraswat, mirella lapata

文摘: 本文介绍了一种神经语义解析器，该解析器将**自然语言**话语映射到逻辑形式上，这些逻辑形式可以针对特定于任务的环境 (如知识库或数据库) 执行，以生成响应。解析器使用基于转换的方法生成树状结构的逻辑形式，该方法将通用树生成算法与逻辑**语言**定义的域通用操作结合起来。生成**过程**由结构化的递归神经网络建模，它为预测提供了丰富的句子上下文和生成历史编码。为了解决**自然语言**和逻辑形式标记之间的不匹配问题，探讨了各种注意机制。最后，我们考虑神经语义解析器的不同训练设置，包括提供注释逻辑形式的完全监督训练、提供外延的弱监督训练和仅提供外延的远程监

控没有标记的句子和知识库是可用的。在广泛的数据集中进行的实验证明了我们的解析器的有效性。少

2018 年 8 月 12 日提交;v1 于 2017 年 11 月 14 日提交;最初宣布 2017 年 11 月。

评论:在计算语言学杂志

474. 第: 1711. 04903[pdf,其他] Cs。Cl

通过对抗培训进行强大的多语言语音部分标记

作者:yasunaga michihiro, jungo kasai, dragomir radev

摘要: 对抗性训练 (at) 是一种强大的神经网络正则化方法，旨在实现对输入扰动的鲁棒性。然而，在**自然语言处理**的背景下，从 at 获得的鲁棒性的具体影响仍不清楚。本文提出并分析了一种利用 at 的神经 pos 标记模型。在我们对 penn treebank wsj 语料库和通用依赖 (ud) 数据集 (27 **种语言**) 的实验中，我们发现 at 不仅提高了整体标记的准确性，而且 1) 可以防止低资源语言中的过度拟合和 2) 提高了罕见/看不见的单词的标记精度。我们还证明了 3) at 改进的标记性能有助于依赖关系分析的下游任务, 4) at 有助于模型学习更清晰的单词表示。5) 所提出的 at 模型在不同的序列标记任务中通常是有效的。这些积极的结果促使进一步使用 at 进行**自然语言**任务。少

2018 年 4 月 20 日提交;v1 于 2017 年 11 月 13 日提交;最初宣布 2017 年 11 月。

评论:naacl 2018

475. 第: 1711. 09090[pdf,其他] Cs。Cl

莫吉特: 大规模产生情感反应

作者:周贤达,王洋先生

摘要: 产生情感**语言是**构建富有同情心的**自然语言处理**代理的关键一步。然而，这一系列研究的一个主要挑战是缺乏大规模的标记培训数据，以前的研究仅限于少量的人类附加注释的情感标签。此外，明确控制生成文本的情绪和情绪也很困难。在本文中，我们采取了更激进的方法: 我们利用推特数据的想法，这些**数据自然被标记为**表情符号。更具体地说，我们收集了大量的推特对话，其中包括在反应中的表情符号，并假设表情符号传达了句子的潜在情绪。然后，我们引入了一个强化的条件变分编码器方法来训练这些对话的深层生成模型，这使得我们能够使用表情符号来控制生成的文本的情绪。我们在定量和定性分析中通过实验表明，所提出的模型能够根据指定的情绪成功地生成高质量的抽象会话响应。少

2018 年 5 月 12 日提交;v1 于 2017 年 11 月 11 日提交;最初宣布 2017 年 11 月。

476. 第: 1711.02799[pdf,其他] Cs。Lg

有限的加权学习

作者:mostafa dehghani, arash mehrjou, stephan gouws, jaap kamps, bernhard schölkopf

摘要: 训练深度神经网络需要许多训练样本，但实际上培训标签的获取成本很高，质量可能不同，因为有些标签可能来自值得信赖的专家标签，而另一些标签可能来自启发式或其他薄弱监督来源例如众包。这就在学习**过程**中创造了一个基本的质量与数量权衡。我们是从少量高质量数据中学习，还是从潜在的大量弱标记数据中学习？我们认为，如果学习者能够以某种方式知道并在学习数据表示时考虑到标签质量，我们就能得到两个世界中最好的结果。为此，我们提出了 "功能加权学习" (fwl)，这是一种半监督的学生-教师方法，用于使用弱标记数据训练深层神经网络。fwl 根据教师 (可访问高质量

标签) 估计的标签质量后置，在每个样本的基础上对学生网络 (在我们关心的任务上接受培训) 的参数更新进行调制。学生和老师都是从数据中学习的。我们对 fwl 进行信息检索和**自然语言处理**这两个任务的评估，在这两个任务中，我们的性能优于最先进的替代半监督方法，这表明我们的方法能够更好地利用强和弱标签，并导致更好地依赖任务的数据表示。少

2018 年 5 月 23 日提交;v1 于 2017 年 11 月 7 日提交;最初宣布 2017 年 11 月。

评论:作为会议文件在 iclr 2018 年发表

477. 第: 1711.01731[pdf, ps,其他] Cs。Cl

对话制度研究: 新的进展与新领域

作者:陈红申,刘晓瑞，尹大伟,唐继良

摘要: 对话系统越来越受到人们的关注。对话系统的最新进展绝大多数是由深度学习技术推动的，这些技术被用来加强广泛的大数据应用，如计算机视觉、**自然语言处理**、和推荐系统。对于对话系统，深度学习可以利用大量数据来学习有意义的特征表示和响应生成策略，同时需要最少的手工制作。本文从不同的角度对对话系统的最新进展进行了综述，并讨论了一些可能的研究方向。特别是，我们通常将现有的对话系统划分为面向任务和非任务导向的模型，然后详细介绍深度学习技术如何帮助他们使用代表性算法，最后讨论一些吸引人的研究方向，这些策略可以把对话系统的研究带入一个新的前沿。少

2018 年 1 月 11 日提交;v1 于 2017 年 11 月 6 日提交;最初宣布 2017 年 11 月。

评论:13 页. arxiv 管理说明: 文本重叠与 arxiv:170001008 由其他作者

478. 建议: 1711.00331[pdf,其他] Cs。Cl

多伊 10.1109/TASLP.2018.2837384

构词的语义结构与解释

作者:lutfi kerem senel, ihsan utlu, veysel yucesoy, aykut koc , tolga Aykut

摘要: 密集词嵌入由于其在许多 nlp 中的最先进的性能，在自然**语言处理**(nlp) 研究中非常流行，它将词的语义意义编码到**低**维向量空间中任务。单词嵌入在捕获单词之间的语义关系方面非常成功，因此在相应的向量空间中必须存在一个有意义的语义结构。然而，在许多情况下，这种语义结构在嵌入维度中分布广泛且异质，这使得解释成为一个很大的挑战。在本研究中，我们提出了一种统计方法来揭示密集词嵌入中潜在的语义结构。为了执行我们的分析，我们引入了一个新的数据集 (semcat)，其中包含超过 6500 字的语义分组在 110 个类别。我们进一步提出了一种量化 "嵌入" 一词可解释性的方法;该方法是传统的入侵词汇测试的一种实用的替代方法，需要人工干预。少

2018 年 5 月 16 日提交;v1 于 2017 年 11 月 1 日提交;最初宣布 2017 年 11 月。

评论:11 页, 8 个数字，被 ieeeeicm 音频、语音和语言处理事务所接受

日记本参考:l. k., utlu, v. yücesoy, a. koc and t. cukur, "文字嵌入的语义结构和解释", 载于 ieeeicm 《音频、语音和语言处理交易》，第 26 卷，第 10 期, 2018 年 10 月。

479. 决议: 1711.00279[pdf,其他] Cs。Cl

深度强化学习的释义生成

作者:李子超,新疆,李峰,李航

摘要: 从给定句子自动生成释义是**自然语言处理**(nlp) 中一项重要但具有挑战性的任务，在许多应用中发挥着关键作用，如问题回答、搜索和对话。本文提出了一种深层强化学

习的方法来解释生成。具体来说，我们为任务提出了一个新的框架，它由一个 \ textho {出名} 和一个 \ text% {评估者} 组成，这两个框架都是从数据中学习的。生成器，作为序列到序列的学习模型，可以产生转译给出一个句子。作为深度匹配模型构建的评价者可以判断两个句子是否相互解释。发电机首先通过深度学习进行培训，然后通过强化学习进行进一步微调，其中奖励由评价者给予。为了评价评价者的学习，我们根据现有培训数据的类型，分别提出了两种基于监督学习和逆强化学习的方法。实证研究表明，所学评价者可以指导生成器产生更准确的释义。实验结果表明，所提出的模型 (生成器) 在自动评价和人工评价中的表现优于最先进的转译生成方法。少

2018 年 8 月 23 日提交;v1 于 2017 年 11 月 1 日提交;最初宣布 2017 年 11 月。

评论:emnlp 2018

480. 第 1710.10280[pdf,其他] Cs。CI

一次性和最少的单词嵌入学习

作者:andrew k. lampinen , james l. mcclelland

摘要: 标准的深度学习系统需要数千个或数百万个实例来学习一个概念，并且无法轻松集成新概念。相比之下，人类有令人难以置信的能力进行一次性或少枪的学习。例如，从仅仅听到一个句子中使用的单词，人类可以通过利用周围单词的语法和语义告诉我们的东西来推断很多。在这里，我们从中汲取灵感，以突出一个简单的技术，通过这种技术，深度经常性网络可以类似地利用他们以前的知识，从少量数据中学习一个新词的有用表示。这可以让**自然语言处理**系统不断地从遇到的新词中学习，从而变得更加灵活。少

2018 年 1 月 2 日提交;v1 于 2017 年 10 月 27 日提交;最初宣布 2017 年 10 月。

评论:正在审查 15 页, 7 个数字，作为 2018 年国际航天中心会议文件

类:I.2。7

481. 第 1710.06280[pdf,其他] 反渗透委员会

使用不受约束的口语说明交互式地选取真实世界的对象

作者:jun hatori, yuta kikuchi, sosuke kobayashi, kuniyuki takahashi, yutatsuboi, yuya unno, wilson ko , jethro tan

文摘: 对**口语**自然**语言**的理解是机器人有效地与人沟通的重要组成部分。然而，处理不受约束的口语指令具有挑战性，因为 (1) 复杂的结构，包括在口语**中**使用的各种表达，(2) 在解释人的指令时固有的模糊。本文提出了第一个能够处理不受约束的口语的综合系统，能够有效地解决口语教学中的歧义问题。具体而言，我们将基于深度学习的对象检测与**自然语言处理**技术集成在一起，以处理不受约束的语音指令，并提出了一种机器人解决问题的方法指令模糊通过对话。通过我们在模拟环境和物理工业机器人手臂上的实验，我们展示了我们的系统有效理解人类操作人员的自然指令的能力，以及物体的成功率如何提高采摘任务可以通过互动澄清**过程**来完成。少

2018 年 3 月 27 日提交;v1 于 2017 年 10 月 17 日提交;最初宣布 2017 年 10 月。

评论:9 页。2018 年机器人与自动化国际会议 (icra)。附带视频可通过以下链接获得: https://youtu.be/_Uyv1XIUqhk (该系统提交给 icra-2018) 和 http://youtu.be/DGJazkyw0Ws (提交 icra-2018 后有所改善)

482. 第: 1710.05916[pdf,其他] Cs。Ce

利用神经网络从 pmu 数据中检测线路故障

作者:李清培,斯蒂芬·赖特

摘要: 我们提出了一种基于神经网络和交流潮流方程的方法，利用只放置在总线子集上的相量测量单元传感器 (pmu) 中的信息来识别电网中的单线和双线断电。我们的方法不是通过反转物理模型来从传感器数据中推断停机情况，而是使用 ac 模型来模拟传感器在多个需求和季节性条件下对所有感兴趣的停机的响应，并使用生成的数据来训练神经网络分类器，直接从传感器数据中识别和区分不同的中断事件。培训结束后，分类器的实时部署只需要少量矩阵向量产品和简单的矢量操作。这些操作的执行速度比基于交流潮流的模型的反演要快得多，该模型由非线性方程和可能的整数/二进制变量组成，以及代表电压和功率的变量流。我们有动力使用神经网络，成功地应用于计算机视觉和**自然语言处理**等领域。神经网络会自动查找原始数据的非线性变换，这些变换突出了使分类任务更容易的有用功能。我们描述了一种选择传感器位置的原则性方法，并表明即使在广泛的需求配置文件中，也可以通过一组受限的测量结果来准确地分类线路中断。少

2018 年 3 月 27 日提交;v1 于 2017 年 10 月 16 日提交;最初宣布 2017 年 10 月。

483. 第 1709.09590[pdf, ps,其他] Cs。CI

多伊 10.1016/j.eswa.2018.02.031

一种注重联合分割和解析的神经结构及其在房地产广告中的应用

作者:giannis bekoulis, jones deleu, thomas demeester, chris develder

摘要: 在使用**自然语言处理**(nlp) 技术**处理**人工生成的文本时，出现的两个基本子任务是 (i) 将纯文本分割为有意义的子单元 (例如，实体) 和 (ii) 依赖关系分析，以建立子单位之间的关系。在本文中，我们开发了一个相对简单有效的神经联合模型，该模型既执行分割，又执行依赖关系解析，而不是像大多数最先进的作品那样一个接一个地执行分析。我们将特别关注房地产广告设置，旨在将广告转换为结构化描述，我们将其命名为属性树，包括 (1) 从分类广告中识别财产的重要实体 (如房间) 和 (2) 结构的任务他们成一个树的格式。在这项工作中，我们提出了一个新的联合模型，它能够同时处理这两个任务，并通过 (i) 避免一个接一个的子任务以流水线的方式产生错误传播来构造属性树，以及 (ii) 利用子任务之间的交互。为此，我们对管道方法和新的联合模型进行了广泛的比较研究，报告了属性树的整体边缘 f1 分数的提高超过 3 个百分点。另外，我们还提出了注意的方法，以鼓励我们的模型在属性树的构建过程中关注突出的令牌。因此，我们实验证明了细心的神经结构对所提出的关节模型的有效性，并为我们的应用展示了边缘 f1 分数的两个百分点的进一步改进。少

2018 年 3 月 19 日提交;v1 于 2017 年 9 月 27 日提交;最初宣布 2017 年 9 月。

评论:预打印-接受在具有应用程序的专家系统中发布

日记本参考:专家系统，第 102 卷，2018 年 7 月 15 日，100-112 页，issn 0957-4174 页

484. 第: 1709.08600[pdf,其他] Cs。CI

ezlearn: 在大规模数据注释中开发有机监督

作者:maxim grechkin, hoifung poon, bill howe

摘要: 许多实际应用需要自动数据注释，例如基于基因表达识别组织起源和将图像分类为语义类别。注释类通常很多，并且会随着时间的推移而发生变化，注释示例已成为监督学习方法的主要瓶颈。在科学和其他高价值领域中，通常可以使用大量的数据示例存储库，以及两个有机监督源: 注释类的词典和一些数据示例所附带的文本描述。远程监控已成为利用这种间接监督的一种有希望的范例，它通过自动注释文本描述在词典

中包含类的示例。但是，由于语言的差异和歧义，此类训练数据本质上是嘈杂的，这限制了这种方法的准确性。本文介绍了一种用于文本模式的辅助**自然语言 处理**系统，并结合联合训练，在远程监控中降低噪声，增强信号。在不使用任何手动标记数据的情况下，我们的 ezla 系统学会了在功能基因组学和科学图形理解中准确地对数据样本进行注释，大大优于经过数万人训练的最先进的监督方法注释的示例。少

2018 年 7 月 1 日提交;v1 于 2017 年 9 月 25 日提交;最初宣布 2017 年 9 月。

485. **第 (xiv:170 009. 08294)[pdf,其他] Cs。CI**

用于文本处理的上下文敏感卷积滤波器

作者:沈定根,林仁强敏,李一通，李登华

摘要: 卷积神经网络 (cnn) 最近已成为**自然语言处理**(nlp) 的一个流行组成部分。尽管取得了成功，但大多数在 nlp 中使用的现有 cnn 模型都为所有输入句子共享相同的学习 (和静态) 过滤器集。在本文中，我们考虑了一种使用小元网络来学习上下文相关的卷积滤波器进行文字处理的方法. 元网络的作用是将句子或文档的上下文信息抽象到一组输入感知筛选器中。我们进一步将这个框架推广到模型句子对，在这个模型中，引入了双向滤波器生成机制来封装相互依赖的句子表示。在我们关于四个不同任务的基准，包括本体分类、情绪分析、答案句子选择和释义识别，我们提出的模型，一个修改后的 cnn 与上下文相关的过滤器，始终优于标准 cnn 和基于关注的 cnn 基线。通过可视化学习的上下文相关筛选器，我们进一步验证并合理化了建议框架的有效性。少

2018 年 8 月 30 日提交;v1 于 2017 年 9 月 24 日提交;最初宣布 2017 年 9 月。

评论:被 emnlp 2018 作为完整文件接受

486. **第 xiv:170 9.09 5583[pdf,其他] Cs。铬**

基于区域分类的减少对深部神经网络的规避攻击

作者:曹晓宇,龚振强

文摘: 深神经网络 (dnn) 已经改变了几个人工智能研究领域，包括计算机视觉、语音识别和**自然语言处理**。然而，最近的研究表明, dnn 在测试时容易受到对抗操纵。具体来说，假设我们有一个测试示例，其标签可以通过 dnn 分类器正确预测。攻击者可以向测试示例中添加一个精心打造的小噪声，以便 dnn 分类器预测不正确的标签，其中精心编制的测试示例称为对抗性示例。这种攻击被称为逃避攻击。在自驾游等安全和安保关键应用中部署 dnn 时，规避攻击是最大的挑战之一。在这项工作中，我们制定了防范逃避攻击的新方法。我们的主要观察是，对抗性示例接近分类边界。因此，我们建议基于区域的分类对对抗示例具有鲁棒性。对于一个良性/对抗测试示例，我们将信息集成在以示例为中心的超立方体中，以预测其标签。相反，传统的分类器是基于点的分类，即在给定测试示例的情况下，分类器仅根据测试示例预测其标签。我们对 mnist 和 cibar-10 数据集的评估结果表明，我们基于区域的分类可以显著减轻规避攻击，而不会牺牲良性示例的分类准确性。具体来说，我们基于区域的分类在测试良性示例时实现了与基于点的分类相同的分类精度，但我们基于区域的分类比基于点的分类对各种分类的鲁棒性要高得多逃避攻击。少

2018 年 1 月 11 日提交;v1 于 2017 年 9 月 16 日提交;最初宣布 2017 年 9 月。

评论:第 33 届计算机安全应用年会 (acsac), 2017

487. **第: 1709.03082[pdf,其他] cs. ne**

多伊 10.114/31955106.3 195117

一种将目标递归单元 (gru) 与支持向量机 (svm) 相结合的神经网络体系结构，用于网络流量数据中的入侵检测

作者:abien fred agarap

文摘: 门控递归单元 (gru) 是一种最近发展起来的长期短期记忆 (lstm) 单元的变体，这两种类型都是递归神经网络 (rnn)。通过经验证据，这两种模型已被证明在各种机器学习任务中都是有效的，如**自然语言处理**(wen 等人, 2015 年)、语音识别 (chorowski 等人, 2015 年) 和文本分类 (yang 等人, 2016 年)。通常，与大多数神经网络一样，上述两个 rnn 变体都使用 softmax 函数作为其预测的最终输出层，并使用交叉熵函数来计算其损耗。本文通过引入线性支持向量机 (svm) 作为 gru 模型最终输出层 softmax 的替代品，对该规范进行了修正。此外，交叉熵函数应替换为基于边缘的函数。虽然也有类似的研究 (alalshekmubarak & amp; smith, 2013;tang, 2013)，本建议主要用于使用京都大学蜜罐系统的 2013 年网络流量数据对入侵检测进行二进制分类。结果表明, gru-支持向量机模型的性能相对高于传统的 gr 鲁-softmax 模型。该模型的训练准确率约为 81.54，测试准确率约为 81.54，而后者的训练准确率为约 63.07%，测试准确率约为 70.75。此外，这两个最终输出层的并列表明，支持向量机在预测时间上的性能将优于 softmax--这是一项理论含义，在研究中的实际训练和测试时间得到了支持。少

2018 年 3 月 10 日提交;v1 于 2017 年 9 月 10 日提交;最初宣布 2017 年 9 月。

评论:在 2018 年机器学习和计算国际会议上, 5 页, 4 个数字, 5 张表格，接受论文

488. 第 xiv:170 08.06665[pdf,其他] cse

软件工程与 sp 智能理论

作者:j 杰拉德·沃尔夫

摘要: 本文从 "sp 智能理论" 及其在 "sp 计算机模型" 中的实现，提出了一种新的软件工程方法。尽管表面上出现了，但事实表明，软件工程中的许多关键思想在 sp 系统的结构和工作原理上都有对应的观点。这种新的软件工程方法的潜在好处包括: 软件开发的自动化或半自动化，并在必要时支持 sp 系统的编程;允许程序员专注于 "面向世界" 的并行性，而无需担心并行性，以加快**处理**速度;支持通过书面或口头 **自然语言**编程 sp 系统的长期目标;减少或消除 "设计" 和 "实施" 之间的区别;减少或取消编译或解释等操作;减少或消除对软件验证的需要;减少对软件验证的需求;程序和数据库之间没有正式的区别;有可能大幅度减少数据文件的类型和计算机**语言**的数量;版本控制的好处;并减少技术债务。少

2018 年 8 月 5 日提交;v1 于 2017 年 8 月 18 日提交;最初宣布 2017 年 8 月。

489. 第 07:170 8.07009[pdf,其他] Cs。CI

基于深度学习的自然语言处理的最新趋势

作者:tomyoung, Devamanyu 语区 hazarika, soujanya poria , erik cambria

摘要: 深度学习方法采用多个**处理**层来学习数据的分层表示，并在许多领域产生了最先进的结果。近年来，各种模型设计和方法在**自然语言处理**(nlp) 的背景下蓬勃发展。在本文中，我们回顾了大量 nlp 任务所采用的重要的深度学习相关模型和方法，并提供了它们演变的过程。我们还对各种模型进行了总结、比较和对比，并对 nlp 中深度学习的过去、现在和未来提出了详细的了解。少

2018 年 10 月 10 日提交;v1 于 2017 年 8 月 9 日提交;最初宣布 2017 年 8 月。

评论:增加了小队

490. 第 xiv:170 8.01525[pdf,其他] Cs。CI

语言设计与重正化

作者:angel j. galleo, roman orus

摘要: 在这里，我们从物理学的角度考虑了语法中的一些众所周知的事实，这使得我们能够建立一些显著的等价。具体而言，我们观察到, n. chomsky 在 1995 年提出的 merge 行动可以解释为物理信息粗粒度。因此, merge 在语言学中需要根据不同的时间尺度在物理学中重新规范化信息。我们用**语言**模型，即在词序上的概率分布，在自然**语言处理**中广泛使用，以及其他词条，使这一点在**数学**上是形式化的。在此设置中, merge 对应于实现粗粒度的 3-索引概率张量，类似于概率上下文无关语法。有意义句子的概率**向量**自然是由随机张量网络 (tn) 给出的，这些网络大多是无环的，如树张量网络和矩阵产品状态。这些结构在句法距离上有短距离的相关性，由于人类**语言**的特殊性，它们在计算上的操作效率极高。我们还提出了如何从某些 tn 量子态的概率分布中获得这样的语言模型，并证明了量子计算机可以有效地制备这些**模型**。此外，利用纠缠理论中的工具，我们利用这些量子态来证明句子中一组单词的概率分布的经典下限。这些结果的意义在理论和计算语言学、人工智能、编程**语言**、ma 和蛋白质测序、量子多体系统等方面进行了讨论。我们的工作展示了上世纪的许多关键语言思想，包括计算语言学的发展，与与重新规范化相关的已知物理概念完美契合。少

2018 年 3 月 15 日提交;v1 于 2017 年 8 月 4 日提交;最初宣布 2017 年 8 月。

评论:22 页, 21 个数字, 1 个表。修订版本，新标题、附录中包含一些正式的语言细节，以及其他改动

491. 第 xiv:170 08.01425[pdf,其他] Cs。CI

争论推理理解任务：隐性认股权证的识别与重构

作者:ivan habernal, henning wachsmuth , iryna gurevych, benno stein

摘要: 推理是**自然语言**论证的重要组成部分。要理解一个论点，我们必须分析它的保证，这解释了为什么它的主张来自它的前提。由于参数具有高度的背景化，因此通常是预先假定的，而不是隐式的。因此，理解不仅需要**语言**理解和逻辑技巧，还取决于常识。本文系统地提出了一种建立权证的方法。我们在一个可扩展的众包**过程**中实施它，从而形成了一个自由许可的数据集，并从新闻评论中获得了 2k 真实论据的授权令。在此基础上，提出了一项新的具有挑战性的任务--论证推理理解任务。考虑到有声明和前提的参数，目标是从两个选项中选择正确的隐式保证令。这两个认股权证都是可信的，在词汇上也是接近的，但却导致了相互矛盾的说法。这项任务的解决办法将确定自动重建权证的一个实质性步骤。然而，使用几个神经关注和**语言**模型的实验表明，目前的方法是不够的。少

2018 年 2 月 27 日提交;v1 于 2017 年 8 月 4 日提交;最初宣布 2017 年 8 月。

评论:被接受为 naacl 2018 长文件；详情见首页

492. 第 xiv:170 08. 00107[pdf,其他] Cs。CI

翻译中的学习：语境化的词汇向量

作者:bryan mccann, james bradbury, c 仁 seong, richard socher

摘要: 计算机视觉受益于初始化多个深层，并在像 imagenet 这样的大型监督培训集中预先训练权重。**天然**的**的语言加工过程**(nlp) 通常只看到使用预先训练的单词向量的

底层深层模型的初始化。在本文中，我们使用一个深 lstm 编码器从一个注意序列到序列模型训练机器翻译 (mt) 上下文词向量。我们表明，添加这些上下文向量 (cowe) 比在各种常见的 nlp 任务中只使用无监督的单词和字符向量提高了性能：情绪分析 (sst, imdb)、问题分类 (trec)、包络 (snli) 和问题回答 (squad)。对于细粒度情绪分析和包包，coe 提高了我们的基线模型的性能，以达到最先进的水平。少

2018 年 6 月 20 日提交;v1 于 2017 年 7 月 31 日提交;最初宣布 2017 年 8 月。

493. **第 1707. 09952[pdf] 中心**

多伊 10.1109/JETCAS.2018.2796379

基于 reram 模拟神经训练加速器的能量、延迟、面积和精度的多尺度协同设计分析

作者:matthew j. marinella , sapan Agarwal,亚历山大? 夏, isaac richter, robin 雅各布斯-盖德林 (robin jacobs-gedrim), john niroula, steven j .plimpton, engin ibek, conrad d. james

摘要: 神经网络是**一种对自然语言处理**和模式识别越来越有吸引力的算法。具有和 gt;50M 参数的深度网络是通过在 lt;50 pj/业务运行的现代 gpu 集群实现的，最近，生产加速器能够在董事会一级 lt;5pJ 每个操作。然而，随着 cmos 扩展速度的放缓，将需要新的范式来实现每瓦增益性能的下一个几个数量级。使用模拟电阻存储器 (reram) 横杆在加速器中执行关键矩阵操作是一个很有吸引力的选择。本工作提供了一个详细的设计，使用的状态的 14, 16 纳米 pdk 模拟横杆电路块，旨在**处理**神经网络的训练和推理所需的三个关键内核。给出了能量、延迟、面积和精度的详细电路和设备级分析，并与使用标准数字 reram 和 sram 操作的相关设计进行了比较。结果表明，与仅使用数字 reram 的类似块相比，模拟加速器具有 270x 的能量和 540x 的延迟优势，并且每次乘法和累积 (mac) 只需要 11 fj。与基于 sram 的加速器相比，能量更好，延迟更好，延迟为 34x。尽管模拟加速器中的训练精度降低，但还是提供了几个改进选项。在这个加速器块的类似数字纯版本上可能获得的收益表明，继续优化模拟电阻存储器是有价值的。训练加速器的这种详细电路和设备分析可作为进一步的体系结构级研究的基础。少

2018 年 2 月 16 日提交;v1 于 2017 年 7 月 31 日提交;最初宣布 2017 年 7 月。

494. **第 1707. 09751[pdf,其他] Cs。CI**

技能 2vec：从职位描述中确定相关技能的机器学习方法

作者:le van-duyet, vo minhquan, dang quang an

摘要: 近年来，在许多**自然语言处理**(nlp) 任务中，在无监督的学习单词嵌入中取得了巨大的成功。本文的主要贡献是开发一种名为 skill2vec 的技术，该技术在招聘中应用机器学习技术，以加强搜索策略，找到具有适当技能的候选人。skill2vec 是由 mikolov 等人在 2013 年开发的由 word2vec 启发的神经网络体系结构。它将技能转化为新的向量空间，具有计算的特点，呈现技能关系。我们由招聘公司的领域专家手动进行了实验评估，以展示我们的方法的有效性。少

2018 年 3 月 29 日提交;v1 于 2017 年 7 月 31 日提交;最初宣布 2017 年 7 月。

495. **第 0707. 08608[pdf, ps,其他] Cs。CI**

具有输出约束的网络基于梯度的推理

作者:jay yoonlee, michael wick, sanket vaibhav mehta, jean-baptiste tristan , jaime Carbonell

摘要: 实践者将神经网络应用于**自然语言处理**(nlp) 中日益复杂的问题，例如具有丰富输出结构的句法分析。许多这样的结构预测问题需要对输出值的确定性约束;例如，在序列到序列的语法分析中，我们要求顺序输出对有效的树进行编码。虽然隐藏的单元可能会捕获这些属性，但网络并不总是能够仅仅从培训数据中了解到这些限制，从业人员随后必须诉诸后**处理**。本文提出了一种神经网络的推理方法，该方法在不执行基于规则的后**处理**或昂贵的离散搜索的情况下，对输出强制进行确定性约束。相反，本着基于梯度的训练精神，我们通过基于梯度的推理来实施约束：对于测试时的每个输入，我们都会推动连续权重，直到网络的无约束推理过程生成满足约束。我们将我们的方法应用于三个具有硬性约束的任务：序列转导、选区解析和语义角色标记 (srl)。在每种情况下，该算法不仅满足约束，而且提高了精度，即使底层网络是最先进的。少

2018 年 8 月 26 日提交;v1 于 2017 年 7 月 26 日提交;最初宣布 2017 年 7 月。

496. **特别报告: 1707. 07435[pdf,其他] Cs。红外**

基于深度学习的推荐人系统：一个调查与新的展望

作者:张帅,姚丽娜,孙爱新，易泰

文摘: 随着在线信息量的不断增加，推荐系统已成为克服此类信息超载的有效策略。由于推荐系统在许多 web 应用程序中得到广泛采用，而且它可能对改善与过度选择有关的许多问题产生潜在影响，因此，推荐系统的效用怎么强调也不为过。近年来，深度学习在计算机视觉和**自然语言处理**等许多研究领域引起了相当大的兴趣，这不仅是由于其出色的性能，也是由于其具有吸引力的特性。从零开始学习功能表示。深度学习的影响也很普遍，最近证明了其应用于信息检索和推荐系统研究的有效性。显然，推荐制度中的深度学习领域蓬勃发展。本文旨在对近年来在深度学习推荐系统上的研究工作进行全面回顾。更具体地说，我们提供和设计了基于深度学习的推荐模型的分类，同时提供了最先进的综合摘要。最后，我们扩展了当前的趋势，并提供了与这一领域的这一新的激动人心的发展相关的新观点。少

2018 年 9 月 4 日提交;v1 于 2017 年 7 月 24 日提交;最初宣布 2017 年 7 月。

评论:35 页，提交到期刊

497. **第 xiv:170 7.005928[pdf,其他] Cs。CI**

用于命名实体识别的深层主动学习

作者:沈燕耀,云贤,扎卡里·利普顿,雅科夫克罗德,阿尼马什里·阿南库马尔

摘要: 深度学习在许多**自然语言处理**任务 (包括命名实体识别 (ner)) 上都产生了最先进的性能。但是，这通常需要大量标记的数据。在这项工作中，我们证明，当深度学习与主动学习相结合时，标记的训练数据的数量可以大幅减少。虽然主动学习是一种样品效率，但它在计算上可能非常昂贵，因为它需要迭代再培训。为了加快这一速度，我们引入了一个轻量级的架构，即 cnn-无关-lstm 模型，由卷积字符和单词编码器以及长期的短期内存 (lstm) 标记解码器组成。该模型在任务的标准数据集上实现了近乎最先进的性能，同时在计算上比性能最佳的模型效率要高得多。我们在培训**过程**中进行增量主动学习，并且能够将最先进的性能与仅 25% 的原始培训数据相匹配。少

2018 年 2 月 3 日提交;v1 于 2017 年 7 月 18 日提交;最初宣布 2017 年 7 月。