nlp 最新科研论文进展

方建勇 2018.10.29

1，图像可以具有包含文本的元素和与它们相关联的边界框，例如，通过计算机屏幕图像上的光学字符识别识别的文本，或具有标记对象的自然图像。我们提出了一种端到端的可训练架构，以整合来自这些元素和图像的信息，以分割/识别自然语言表达所指的图像部分。我们计算每个元素的嵌入，然后将其投影到图像特征图的相应位置（即，相关联的边界框）。我们表明，这种体系结构改进了解析引用表达式，仅使用图像和其他包含元素信息的方法。我们演示了基于 COCO 的引用表达数据集的实验结果，以及我们开发的表达数据集的网页图像。

2，我们提出了一个少数关系分类数据集（FewRel），由来自维基百科的 100 个关系的 70,000 个句子组成，并由众筹工作者注释。每个句子的关系首先由远程监督方法识别，然后由群众工作者过滤。我们采用最新的最先进的少数几种学习方法进行关系分类，并对这些方法进行全面评估。实证结果表明，即使是最具竞争力的少数学习模型也在努力完成这项任务，特别是与人类相比。我们还表明，需要一系列不同的推理技巧来解决我们的任务。这些结果表明，少数关系分类仍然是一个悬而未决的问题，仍需要进一步的研究。我们的详细分析指出了未来研究的多个方向。有关数据集和基线的所有详细信息和资源都在此 http URL 上发布。

3，我们提出了一种自然语言理解的新方法，其中我们将输入文本视为图像，并应用 2D 卷积神经网络从单词视觉模式的变化中学习句子的局部和全局语义。我们的方法表明，可以从带有文本的图像中获取具有语义意义的特征，而无需使用传统的自然语言理解算法所需的光学字符识别和顺序处理管道。为了验证我们的方法，我们提出了两个应用的结果：文本分类和对话建模。使用 2D 卷积神经网络，我们能够超越基于非拉丁字母的文本分类的最新准确度结果，并为八个文本分类数据集取得了有希望的结果。此外，当从 bAbI 对话框数据集的任务 4 中使用词汇表实体时，我们的方法优于内存网络。

4，基于词频的提取摘要方法易于实现，并且可以跨语言产生合理的结果。然而，它们具有显着的局限性 - 它们忽略了上下文的作用，它们在文档中提供不均匀的主题覆盖，有时是脱节和难以阅读的。我们使用语言类型学的简单前提 - 英语句子是实体之间潜在交互的完整描述，通常以主语 - 动词 - 宾语的顺序 - 来解决这些困难的一个子集。我们开发了一种提取摘要的混合模型，它将基于词频的关键词识别与来自自动生成的实体关系图的信息结合起来，为摘要选择句子。使用词频和基于主题词的方法进行的比较评估表明，所提出的方法具有竞争性，通过传统的 ROUGE 标准，并且通过大型小组（N = 94）的人类评估者评估，平均产生适度更多的信息摘要。

5，在自然语言处理中，使用递归神经网络成功地解决了许多任务，但是这样的模型具有大量参数。这些参数中的大多数通常集中在嵌入层中，嵌入层的大小与词汇长度成比例增长。我们提出了一种用于 RNN 的贝叶斯稀疏化技术，它允许压缩 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化模型，以过滤掉不必要的单词并进一步压缩 RNN。我们表明，保留单词的选择是可以解释的。

6，在大数据时代，物流计划可以通过数据驱动来利用过去积累的知识。而在电影行业，电影策划也可以利用现有的在线电影知识库来取得更好的效果。然而，由于大量现有电影和各种真实因素有助于每部电影的成功，例如电影类型，可用预算，制作团队（仅限于电影规划，完全依赖传统启发式方法进行电影规划是无效的。在本文中，我们研究了一个“大片规划”（BP）问题，以便从以前的电影中学习，并以完全数据驱动的方式规划低预算但高回报的新电影。在对在线电影知识库进行彻底调查后，本文介绍了一种新颖的电影规划框架“大片电影配置熟人大片规划”（BigMovie）。从投资的角度来看，BigMovie 使用给定的预算最大化计划电影的估计总量。它能够以 0.26 平均绝对百分比误差（预算为 0.16）精确估计电影总数。同时，从制作团队的角度来看，BigMovie 能够形成一个优化的团队，其中包含团队成员熟悉的人/电影类型。历史协作记录用于通过熟人张量来估计电影配置因子的熟人得分。我们将 BP 问题表述为非线性二元规划问题并证明其

NP-硬度。为了在多项式时间内解决它，BigMovie 放松了硬二进制约束并将 BP 问题作为一个三次规划问题来解决。在 IMDB 电影数据库上进行的大量实验证明了 BigMovie 能够进行有效的数据驱动大片规划。

7，知识库（KB）在 NLP 中是最重要的。我们采用多视图学习来提高 KB 中实体类型信息的准确性和覆盖范围。我们依赖于两种元视图：语言和表达。对于语言，我们考虑来自维基百科的高资源和低资源语言。为了表示，我们考虑基于实体的上下文分布（即，在其嵌入上），实体的名称（即，在其表面形式上）和在维基百科中的描述的表示。两种元视图语言和表示可以自由组合：每对语言和表示（例如，德语嵌入，英语描述，西班牙语名称）是一个独特的视图。我们使用细粒度类进行实体类型化的实验证明了多视图学习的有效性。我们发布了 MVET，一个大型多视图 - 特别是多语言 - 我们创建的实体类型数据集。可以在此数据集上评估单语言和多语言细粒度实体类型系统。

8，跨语言文本摘要（CLTS）以与源文档的语言不同的语言生成摘要。最近的方法使用来自两种语言的信息来生成具有最丰富的句子的摘要。但是，这些方法的性能可能因语言而异，这会降低摘要的质量。在本文中，我们提出了一个压缩框架来生成跨语言摘要。为了分析性能，特别是稳定性，我们在四种语言（英语，法语，葡萄牙语和西班牙语）的数据集上测试了我们的系统和提取基线，以生成英语和法语

摘要。自动评估表明，我们的方法优于采用最先进的 CLTS 方法，所有语言的 ROUGE 分数都更好，更稳定。

9，语义文本相似度（STS）是自然语言处理（NLP）中许多应用的基础。我们的系统结合了卷积和递归神经网络来衡量句子的语义相似性。它使用卷积网络来考虑单词的本地上下文和 LSTM 来考虑句子的全局上下文。这种网络组合有助于保留句子的相关信息，并改善句子之间相似性的计算。我们的模型取得了良好的效果，并且与最先进的系统相比具有竞争力。

10，文本相似度计算是自然语言处理和相关领域中的基本问题。近年来，已经开发出深度神经网络来执行该任务并且已经实现了高性能。神经网络通常在监督学习中用标记数据进行训练，并且标记数据的创建通常非常昂贵。在这篇简短的论文中，我们讨论了用于文本相似度计算的无监督学习。我们提出了一种新的方法，称为基于编辑距离的字嵌入（WED），它将字嵌入到编辑距离中。在三个基准数据集上的实验表明，WED 优于现有技术的无监督方法，包括编辑距离，基于 TF-IDF 的余弦，基于字嵌入的余弦，Jaccard 索引等。

11，诸如 COCO 和 Flickr30k 之类的标准图像字幕任务在事实上是中性的，并且（对于人类）状态是明显的（例如，“一个人弹吉他”）。虽然这些任务对于验证机器是否理解图像的内容是有用的，但它们并不

像人物那样与人类互动。考虑到这一点，我们定义了一个新任务，即 Personality-Captions，其目标是通过结合可控的风格和个性特征，尽可能吸引人类。我们收集并发布了 201,858 个这样的字幕的大型数据集，其中包含 215 种可能的特征。我们建立的模型将（i）句子表示（Mazare et al。，2018）的现有工作与受过 17 亿次对话示例训练的变形金刚相结合; （ii）图像表示（Mahajan 等，2018），ResNets 对 35 亿社交媒体图像进行了培训。我们在 Flickr30k 和 COCO 上获得了最先进的性能，并在我们的新任务中获得了强大的性能。最后，在线评估验证我们的任务和模型对人类有吸引力，我们的最佳模型接近人类表现。

12，我们提出了一种从单个输入深度图像重建，完成和语义标记 3D 场景的方法。我们通过基于对抗性学习的新颖架构来提高回归语义 3D 地图的准确性。特别是，我们建议使用多个对抗性损失术语，这些术语不仅可以强制实现与实际相关的实际输出，还可以有效嵌入内部特征。这是通过将处理部分 2.5D 数据的编码器的潜在特征与从训练重建完整语义场景的变分 3D 自动编码器提取的潜在特征相关联来完成的。此外，与完全通过 3D 卷积操作的其他方法不同，在测试时我们在下采样期间保留输入的原始 2.5D 结构，以提高模型内部表示的有效性。我们在语义场景完成的主要基准数据集上测试我们的方法，以定性和定量评估我们提案的有效性。

13，本文考察了双边论证对在线消费者评论的感知有用性的影响。与之前的作品相比，我们的分析从基于语言的角度阐述了对评论的接受。为此，我们提出了一种基于分布式文本表示和多实例学习的有趣文本分析方法，以实现评论文本中论证的两面性。随后使用大量亚马逊评论的实证分析表明，评论中的双边论证显着增加了他们的帮助。我们发现这种效果对于积极评价比对负评价更强，而更高程度的情绪语言会削弱效果。我们的研究结果对零售商平台具有直接影响，可以利用我们的结果优化其客户反馈系统并提供更有用的产品评论。

14，深度学习几乎彻底改变了 NLP 的所有领域。然而，在不利方面，通常认为它依赖于大量的培训数据。因此，这些技术似乎不适合注释数据有限的区域，如情感分析，其中有许多细微差别且难以获取的注释格式，或资源贫乏的语言中遇到的其他低数据场景。与这种流行的观念相反，我们提供了三种不同类型语言的经验证据，而今天最受欢迎的神经架构只能通过几百种观察来训练。我们的结果表明，尽管存在如此强大的数据限制，但高质量，预训练的字嵌入对于实现高性能至关重要。

15，自动提取临床概念是将临床记录中的非结构化数据转换为结构化和可操作的信息的重要步骤。在这项工作中，我们提出了一种临床概念提取模型，用于利用特定领域的语境词嵌入在临床记录中自动注释临床问题，治疗和测试。首先在语料库上训练上下文单词嵌入模型，

其中临床报告和临床领域中的相关维基百科页面的混合。接下来，使用上下文词嵌入模型训练双向 LSTM-CRF 模型用于临床概念提取。我们在 I2B2 2010 挑战数据集上测试了我们提出的模型。我们提出的模型在报告的基线模型中实现了最佳性能，并且在 F1 得分方面优于最先进模型 3.4%。