Utilizing Knowledge Graphs for Text-Centric Information Retrieval

大型知识图（KG）和可扩展的实体链接技术是深入理解文本语义的强大工具。 虽然这些工具可用于广泛的任务，但在本教程中，我们将重点关注知识图和实体链接如何最有效地用于以文本为中心的信息检索（IR）。  
我们使用术语实体来表示KG中的任何条目，同时将其与文本中的实体（在文献中称之为关于命名实体识别的实体）进行区分。 因此，我们利用实体的通用扩展定义来涵盖KG中的任何条目，例如，可能包括人员和地点，还包括化合物，疾病以及无形概念，例如“信息”。 恢复”。

这种关系表示为KG中的边缘（或超边界）。 它们可以手动输入到图表中，来自各种数据库，或者从文本中提取，用于自动知识图填充。 然后，实体检索指的是响应用户查询检索相关KG条目的任务。 实体链接是指文本的注释，使得所有实体提及都使用KG条目的标识符进行注释

从INEX [13]，TREC和TAC KBP计划[4,29]开始，实体链接和检索的任务已经取得了进展。 虽然先前已经证明KG可以作为扩展术语和平滑的来源[1,2,6,33,57]，实体链接和检索方法的改进已经导致实体利用的一系列成功 临时以文本为中心的检索场景中的关系，描述和类型[12,38,49,55]。 准确的实体链接方法在这种情况下起着关键作用，因为它们提供了非结构化文本和KG中实体的结构化信息之间的桥梁。

这些成功导致了如何有效利用以实体为中心的知识库来理解文本数据并估计基于实体的与给定信息需求的相关性的新研究线。 该领域的大量最新进展使其成为总结和报告社区最先进方法的理想时间。 本教程中概述的方法和方法为几个以文本为中心的检索任务的未来发展奠定了基础，包括发现新兴实体[28]，解决查询方面[50]，将内容组织成主题[3， 16]以及实体感知的特设文档检索。

我们还通过提供实体检索的新颖和最新进展的概述来讨论语义搜索的问题，这些进展在此主题的先前教程中未涉及。 本教程重点介绍如何使用KG进行以文本为中心的信息检索，更具体地说，是如何利用KG提供的不同类型的数据进行临时文档检索和其他搜索系统。 我们参考KG4IR研讨会，了解该领域正在进行的工作[17] .1本教程分为四个部分：a）实体链接，b）实体检索，c）利用以文本为中心的信息检索实体，d）开放式研究 区域，将在以下部分中讨论。

实体链接  
存在各种各样的通用百科知识知识图，例如Freebase，DBpedia，WikiData，Yago，Microsoft's  
Satori和Google的知识图以及特定领域的知识库，例如统一医学语言系统。 语言和其他知识也可以在受控词汇和语义网络中编码，例如MeSH，WordNet，Babelnet和ConceptNet。 大多数知识图具有指定类型，关系和元数据的基础本体，而其他知识图则简单地构建为概念网络或实体集合。

虽然每个KG具有独特的特征，但KG实体通常与不同的名称相关联，并且可能与分类法或类别系统中的类型相关联，并且与其他实体有关系。一些幼稚园还包含对实体的明确文本描述和/或指向“关于”每个实体的文本文档的链接。在整个教程中，我们将讨论如何使用这些不同类型的信息中的每一种：a）检索一组实体，以便将信息需求表述为关键字查询或问题，或者更广泛地说：如何评估KB元素与给定主题的相关性，b）如何识别文本片段中KG的实体提及c）如何利用这些提及来评估文本片段的相关性。  
实体链接[48]是在文本中识别实体提及并将它们与知识图中的对应实体对齐的任务。实体链接系统通常被构造为管道。第一步是识别可链接短语，即可以提及实体的文本片段。在第二步中，检索每个这样的短语的候选集，当然必须考虑所提到的实体未包含在知识图中的可能性（所谓的NIL实体）。最后一步是根据提及的背景来区分提及中提到的哪些候选实体。我们讨论了各种最佳实践和方法，例如主题模型和单词嵌入。

3 ENTITY RETRIEVAL  
提交给Web搜索系统的大部分查询旨在找到一个实体或一组实体，可以直接从KG中检索[47]。 此类查询可以通过其名称，属性或相关实体来引用目标实体，并以关键字或问题的形式表示[32]。 得到的排序实体可以直接呈现给用户，或者用作用于以文本为中心的检索的查询扩展术语的源。

此外，通过考虑知识图中的距离，可以使检索到的实体多样化[34]。  
还可以使用基于语料库的伪相关性反馈方法来检索实体，其中针对实体链接分析反馈文档[51]。使用实体检索相关文本可以被视为通过相关文本检索实体的反向问题[11]。  
由于知识图实体与文本文档或网页不同，因此需要新的检索模型，这些模型通常称为对象检索模型。规范方法是将关于实体的异构和半结构化信息（例如，其名称别名，属性，类别，传出/进入链接和内容）组合成静态[42,45,58]或动态[23]。多场实体表示。可以使用专门的结构化文档检索模型来检索这样的实体表示，例如Fielding Sequential Dependence Model [58]及其基于特征的变体[45]。实体检索模型还可以有效地利用查询[24]中的实体链接或类型提示[22,25]。

4利用文本检索中的实体

在本教程中，我们重点关注以文本为中心的IR系统的三个核心角度：a）关键字匹配和平滑模型，b）使用伪相关反馈和查询日志查询扩展模型，以及  
c）用于多样化和冗余移除的组件。 这些方面的大多数工作都是在术语和短语的层面上运作。 然而，实体链接算法和对象检索的最新发展使得有效利用KG提供的丰富信息变得可行。

对垂直，复合和聚合搜索的研究提供了对该问题的另一种观点，其中主要任务是组合来自各种资源的信息。一个中心思想是通过使用实体作为信息的枢纽和多样化来形成信息包[9]。在某些情况下，知识库被解释为检索的进一步垂直[44]。  
先前提出的系统成功地利用知识库来改进临时文档检索。这些系统一方面将实体检索和语义搜索的概念结合起来，另一方面将文本检索模型和实体链接结合起来。有时用户可能会发现将KG实体明确地包含在他们的自由文本查询[5]中，或者随着时间的推移跟踪实体是有帮助的[15]。  
KG感知文档检索模型包含实体名称，上下文术语和实体链接的匹配。与寻找相关实体的方法一起，这些方法产生了有效的可推广检索方法

不同的机器学习方法有助于解决此任务。 Concept Feedback [33]使用基于特征的系统和图形遍历。潜在实体空间[38]使用生成语言模型。 EsdRank [55]和实体查询特征扩展[12]将基于KG的查询扩展的实体检索，文本检索和不同指标与监督的学习到排名方法相结合。对语义网络，潜在相关实体，实体类型的推断可以集成到这样的系统中[31,57]。语言模型建立在不确定的实体链接之上[49]。此外，具有知识库的统计术语关联图为查询扩展提供了额外的来源[2]。将关系提取用于文本检索具有潜力，但仍然是一个挑战[30,52]。  
神经网络的几种方法最近已经引入了信息，包括用于结合实体注释的扩展。一个例子是二重唱模型中实体和文档预测的组合[56]。

**Weakly-supervised Contextualization of Knowledge Graph Facts**

我们引入神经事实上下文化方法（NFCM）来解决KG事实上下文化任务。 NFCM首先在给定事实的邻域中生成一组候选事实，然后使用监督学习对候选事实进行排序以对模型进行排名。排名模型结合了我们自动从数据中学习的功能，并使用我们为此任务设计或调整的一组手工制作的功能来表示查询候选事实。为了获得训练学习大规模排名模型所需的注释，我们使用对大型实体标记文本语料库的远程监督自动生成训练数据。我们表明，对这些数据学习的排名函数在KG事实的背景下是有效的。使用人类评估员进行的评估表明，它显着优于多个竞争基线。

在本文中，我们讨论KG事实情境化的任务，即，给定一个由两个实体和连接它们的关系组成的KG事实，从KG检索与该事实相关的其他事实。此任务类似于临时检索：（i）“查询”是KG事实，（ii）“文档”是KG中“查询”附近的其他事实。我们提出了一种神经事实语境化方法（NFCM），这种方法首先生成一组候选事实，这些事实属于来自主要事实的实体的{1,2} -hop路径的一部分。然后，NFCM根据它们对主要事实进行背景化的相关性对候选事实进行排序。我们估计我们学习使用监督数据对模型进行排名。排名模型结合了（i）我们自动从数据中学习的特征和（ii）代表查询候选事实的特征以及我们为此任务设计或调整的一组手工制作的特征。由于幼稚园的规模和异质性，即大量实体和关系类型，我们转向远程监督以收集训练数据。使用另一个经人工验证的测试集合，我们测量我们提出的方法的性能，并将其与几个基线进行比较。我们总结我们的贡献如下。

我们介绍了KG事实情境化的任务，其中目标是，由两个实体和一个连接它们的关系组成的事实，对KG中与该事实相关的其他事实进行排序。  
我们提出NFCM，一种使用远程监督和学习排名来解决KG事实情境化的方法。 我们的研究结果表明：（i）远程监督是收集此任务的训练数据的有效手段;（ii）神经学习对端到端训练的模型进行排名，优于人类策划评估的几个基线 组。  
我们提供详细的结果分析和对任务性质的见解。

在本文中，我们介绍了知识图事实情境化任务，并提出了NFCM，一种弱监督的方法来解决它。 NFCM首先通过查看1或2跳邻居来生成查询事实的候选集，然后使用监督机器学习对候选事实进行排序。 NFCM将手工制作的功能与使用深度学习自动识别的功能相结合。我们使用远程监督来通过使用与我们使用的KG中的实体高度重叠的大型实体标记文本语料库来增强训练数据的收集。我们的实验结果表明：（i）远程监督是收集此任务的训练数据的有效手段，（ii）NFCM明显优于此任务的几个启发式基线，以及（iii）手工和自动学习的特征有助于到NFCM的检索效果。对于未来的工作，我们的目标是探索更复杂的方法，将手工制作与自动学习的排名功能相结合。此外，我们还希望探索其他数据源以收集培训数据，例如新闻文章和点击日志。最后，我们希望探索以多种方式在搜索引擎结果页面中组合和呈现排名事实的方法。

**Improving Exploratory Search Experience through Hierarchical Knowledge Graphs**

最近的工作探索了层次结构和网络的相对优势，并指出这些好处在很大程度上是互补的：层次结构为用户提供了对中心主题的一些理解，使他们能够开发出更为全面的信息概述;网络允许人们从表示中收集具体信息，而不是需要广泛阅读单个文档[39]。鉴于知识图和层次结构的互补优势，本文的主要研究问题是，我们是否可以通过算法生成无缝数据结构，将两个层次结构和网络的优势结合到一个统一的结构中。  
在本文中，我们评估层次知识图（HKG）作为低级实体关系和高级中心概念的组合表示的效率。我们使用简单的解析算法自动生成这些知识图[37]，然后使用动态阈值方法提取层次结构。我们使用混合方法方法评估这些HKG。定量数据认为HKG保留了知识图的透明度优势和层次结构的结构优势。定性数据与定量观察三角化，并提供对层次和网络可视化的优缺点的额外见解。 。

**Entity Set Expansion via Knowledge Graphs**

实体集扩展问题是将一小组种子实体扩展为更完整的类似实体集。 它可以应用于Web搜索，项目推荐和查询扩展等应用程序。 传统上，人们通过利用网页内实体的共现来解决这个问题，其中种子实体之间的潜在语义相关性不能被揭示。 我们提出了一种新方法，通过考虑知识图的不足（例如，不完整性），使用知识图来解决问题。 我们基于种子的语义特征设计了一个有效的排序模型来检索候选实体。 对公共数据集的大量实验表明，所提出的解决方案明显优于最先进的技术。

在本文中，我们提出了一种基于知识图中实体之间的语义相关性的分类模型（称为k-松弛的共同语义特征），以有效地评估实体与少数给定种子之间的相似性。 该模型旨在处理知识地图的不完整性。

591/5000

在本文中，我们通过使用KG来解决实体集扩展的问题。 我们提出了一种称为通用语义特征的概念，用于描述种子实体共有的共同特征，作为发现和排序候选实体的基础。 通过广泛的实验研究，我们发现所提出的解决方案非常适合于具有良好的实体和谓词（关系）覆盖范围的ESE主题。 即使对于那些在解决方案中没有良好信息覆盖的主题，也可以通过发现种子共享的一些常见语义特征来工作

**Open-World Knowledge Graph Completion**

作为解决此任务的第一次尝试，我们引入了一个名为ConMask的开放世界KGC模型。该模型学习实体名称及其文本描述部分的嵌入，以将看不见的实体连接到KG。为了减少噪声文本描述的存在，ConMask使用依赖于关系的内容屏蔽来提取相关片段，然后训练完全卷积神经网络以将提取的片段与KG中的实体融合。对新旧大型数据集的实验表明，ConMask在开放世界的KGC任务中表现良好，甚至在标准的封闭世界KGC任务上优于现有的KGC模型。

在目前的工作中，我们引入了一个新的开放世界知识图完成模型ConMask，它使用依赖于关系的内容掩蔽，完全卷积神经网络和语义平均来从实体和关系的文本特征中提取依赖于关系的嵌入。 在幼儿园。 开放世界和封闭世界KGC任务的实验表明，ConMask模型在两个任务中都具有良好的性能。 由于标准KGC数据集中存在问题，我们还发布了两个用于KGC研发的新DBPedia数据集。  
ConMask模型是一种提取模型，如果在实体的描述中表达了必要的信息，那么它目前只能预测关系。 未来工作的目标是扩展ConMask，使其能够发现新的或隐含的关系。

**Variational Reasoning for Question Answering with Knowledge Graph**

为了应对这些挑战，我们提出了一种新颖且统一的深度学习架构，以及端到端的变分学习算法，可以处理问题中的噪声，同时学习多跳推理。我们的方法在文献中最近的基准数据集上实现了最先进的性能。我们还推导出一系列新的基准数据集，包括多跳推理的问题，神经翻译模型解释的问题以及人类语音中的问题。

**Complex Sequential Question Answering: Towards LearningtoConverse Over Linked Question Answer Pairs with a Knowledge Graph**

为此，我们介绍了复杂顺序质量保证的任务，该任务结合了以下两个任务：（i）通过对数百万实体的实际大小的KG进行复杂的假设来回答事实问题，以及（ii）学习通过一系列实体进行交谈相关联的QA对。

在本文中，我们介绍了复杂顺序问答（CSQA）的任务，其中包含由链接QA对上的对话组成的大规模数据集。 该数据集包含200,000个对话，转速为1.6M，并通过手动密集的半自动化过程收集。 据我们所知，这是第一个包含复杂问题的数据集，这些问题需要对包含数百万元组的大型知识图进行逻辑，定量和/或比较推理。 我们提出了一个CSQA模型，它是对话和质量保证的最先进模型之间的交叉，并突出了该模型在处理CSQA任务时的不足之处。 很明显，CSQA有几个挑战，解决/解决所有这些挑战超出了单一论文的范围。 我们希望这项任务和数据集的引入能够激发研究的兴趣  
社区开发复杂顺序问答的模型。