103 | The Web 2018论文精读：如何从文本中提取高元关系?

2018-05-30 洪亮劼

AI技术内参 进入课程 >

讲述：初明明
时长 06:04 大小 2.79M

今天我们来看万维网大会 2018 的最佳论文，标题是“HighLife: Higher-arity Fact Harvesting”。作者都来自德国著名的“马克斯·普朗克计算机科学研究所”（Max Plank Institute for Informatics）。这个研究所是德国最大的基础科学研究组织“马克斯·普朗克学会”（Max-Planck-Gesellschaft）的分支研究机构，致力于在科学刊物上发表新的研究成果，开发软件系统和培养新的科学研究工作者。马克斯·普朗克学会因其杰出的科研成果在德国甚至全世界都获得了很高的声誉。

什么是高元关系?

这篇论文主要是涉及到**高元（Higher-Artiy）关系**的提取。那什么是高元关系呢?

传统的信息提取和知识库主要是关注二元关系的提取和存储。例如，我们可以知道居里夫人分别于 1903 年和 1911 年获得了诺贝尔奖。但是关系数据库中并不知道这两年的奖项分别是物理和化学。同理，我们可以在知识库中存放居里夫人获得过诺贝尔物理奖以及诺贝尔化学奖的信息，但是就无法和 1903 年和 1911 年这两个信息进行配对。通过这个例子我们可以看出，基于二元关系的信息提取和知识库虽然简单易行，但是有其先天的局限性。

这篇论文要讨论的高元关系，就是希望能够直接对“居里夫人在 1903 年获得了诺贝尔物理学奖”这样的三元甚至更高元的关系进行提取和表征。作者们认为这篇论文是较少的关注高元关系提取的先驱工作。

论文的主要贡献

我们刚才说了，这篇论文的一个重要贡献就是针对高元关系的提取所作出了很多努力。

具体来说，作者们使用“**种子事实**”（Seed Facts）作为一种监督信息来学习**模式**（Patterns），并且利用这些学习到的模式来寻找更多的“**候选事实**”（Facts Candidates），如此循环。这是把过去的一种针对二元关系提取的方法给扩展到高元关系。这个方法的潜在问题是：在能够保证“高召回”（High Recall）的情况下，得到的很多关系可能存在“**噪声**”和“**目标浮动**”（Target Drifts）。这里所说的目标浮动指的是我们提取的事实有可能存在主题上的偏差。

为了解决这个问题，作者们在这篇论文里利用了“**限制推理**”（Constraint Reasoning），来对已经得到的事实进一步筛选以得到最后的结果。这里的限制可以是“类型”（Type）上的，比如，我们限制提取到的普利策奖为“书籍”而非“电影”或“音乐”。通过这些在取值或者类型上的限制，我们可以对获取到的事实进行清理。

论文解决的另外一个难点就是很多高元信息在原始的文本中就是缺失的，或者是不完全的。比如，“Google 于 2014 年收购了 Nest”这个事实就没有提及金额，而“Google 以 32 亿美元收购了 Nest”这个事实又没有提及时间。作者们针对这个情况，把整个框架给扩展到了缺失信息中，从而能够从原始文本中拼凑多元关系。

论文的核心方法

文章提出了一个由好几个组件组成的系统用于信息的提取。

首先，有一个叫作 **NERD**的组件，即“人名识别和去歧义”组件，用于从句子中提取不同的“实体”。这里面运用到了很多外部的信息库，比如医疗生物实体库“联合医疗语言系统”（Unified Medical Language System）、支持新闻实体的 AIDA 系统以及 WordNet 语料库。同时，在这个部分，NERD 还依赖于“斯坦福自然语言处理核心库”（Stanford CoreNLP）提供“人名识别”以及“词类分析”（Part of Speech）等基础功能。

在提取了人名和实体名之后，作者们就开始构建一个从词类分析得到的**树型数据结构**。这个数据结构的目的是反映 N 元关系和内部信息的架构。这个部分基本上也是依赖传统的自然语言处理所得到的树结构，只不过进行了简单的修正。

得到树结构之后，接下来的一系列工作都是**在这个树结构上获取不同的模式，从而能够得到想要的高元关系**。这里面有很多细节，我们在这里就不赘述了。比如，作者们利用“**树挖掘**”（Tree Mining）技术来发现频繁出现的子树结构，从而认定某个子树模式是不是一个好的候选事实。这里的思路其实和经典的“**频繁模式挖掘**”（Frequent Pattern Mining）一样，都是去不断地计算一个结构的“**支持度**”（Support）和“**置信度**”（Confidence），从而通过两个值来决定是不是要把这个模式给留下来。

除此以外，这一部分的部件还需要支持“**部分 N 元候选事实**”（Partial N-ary Fact Candidate）的匹配。之前我们也讲过了，这个功能也算是这篇论文的一个贡献。这里面的重要职能就是能够对树的一部分结构进行匹配，而不需要对所有的部分都能够完全一致。

当作者们通过树挖掘从而发现了基本的候选事实之后，下面需要做的工作就是针对这些候选事实进行推理盘查，看是不是所有的事实都能经得住推敲。也就是说，我们需要查看有没有存在多个事实不一致的地方。

需要指出的是，从整体上来看，所有组件的流程基本上都是**无监督的数据挖掘操作**。也就是说，整个系统并不需要依赖于什么训练数据。

方法的实验效果

作者们在纽约时报数据集以及 PubMed 数据集上都进行了实验，主要观测的指标是“**精度**”（Precision）。我们之前提到过，这篇文章所研究的高元关系提取，这个问题很新颖。因此，作者们还利用 CrowdFlower 众包平台来获取了数据的标签，用于检测所提取关系的准确度。当然这部分数据量相对来说是比较小的。

从实验的效果上来说，文章提出的方法能够达到平均接近 80%~90% 的精度，这可以说是非常令人振奋的结果了，而达到这样的结果仅仅需要几百个种子事实。

小结

今天我为你讲了今年万维网大会的优秀论文。文章介绍了如何从文本中提取高元关系，这是一个比较新的研究领域。

一起来回顾下要点：第一，我们简单讨论了高元关系的含义；第二，我们重点介绍了论文的主要贡献和核心思路；第三，我们简单分享了提出方法的实验成果。

最后，给你留一个思考题，在什么样的应用中，我们可以利用到这篇文章提出的高元关系？

欢迎你给我留言，和我一起讨论。

上一篇　102 | The Web 2018论文精读：如何改进经典的推荐算法BPR?

下一篇　104 | 如何快速学习国际顶级学术会议的内容?

精选留言 (2)

写留言

刘军

2018-07-29

是否可以用在上市公司公告分析中?

展开

sky

2018-06-02

感觉可以用到推荐系统

展开