深度解读谷歌SyntaxNet：全新TensorFlow自然语言处理模型

本文作者：AI科技评论 2016-05-17 18:36

导语：SyntaxNet用来做什么？带来多大进步？下一步是什么？

今年夏天，雷锋网将在深圳举办一场盛况空前的“全球人工智能与机器人创新大会”（简称GAIR）。大会现场，雷锋网(公众号：雷锋网)将发布“人工智能&机器人Top25创新企业榜”榜单。目前，我们正在四处拜访人工智能、机器人领域的相关公司，从而筛选最终入选榜单的公司名单。如果你的公司也想加入我们的榜单之中，请联系：2020@leiphone.com。

图片来源：spaCy

编者注：spaCy是一个免费开源代码库，Matthew Honnibal是spaCy公司创始人及CTO。他在本科时学习语言学，从未想到未来自己会成为程序员。Honnibal获得悉尼大学计算机科学PHD学位并进行研究员工作后，在2014年离开学术界开始编写spaCy。本文中，Hobbibal深度解读了谷歌的自然语言处理模型。

上周，谷歌开源了其基于人工智能系统Tensorflow的自然语言解析模型分析库SyntaxNet。在过去的两年时间里，谷歌研究人员利用这个分析库发布了一系列神经网络分析模型。自从SyntaxNet发布以来，笔者就一直关注它，当然也一直也期待这个软件能够开源。不过，本文尝试围绕本次开源的相关背景做一些探讨，比如本次开源有什么新料，开源又有何重要意义？

在自然语言文本处理库中（比如spaCy），SyntaxNet提供了非常重要的模型。如果你把自然语言处理的概念”缩小”一点，就会意识到，这种你正在关注的技术可以拓展计算机的应用范围。即便是现在，你依然无法编写软件去控制一辆汽车，也无法用你的语气来回复电子邮件，更无法用软件来分析客户反馈，或为规避重大商业风险去监测全球新闻。诚然，自然语言处理无法操控无人驾驶汽车，但等下先，语言是人类最与众不同的能力，人类已经不可避免地掌握了这种技能，但是自然语言处理技术也很优秀，我们甚至难以预测它的潜力。谷歌搜索就是一种自然语言处理应用，所以你会发现这项技术其实已经在改变世界。不过

AI科技评论
编辑
发私信

当月热门文章
汤晓鸥：人工智能在中国有点过热了，我想泼泼冷水

最新文章
推荐美图的 Pinterest，如何靠机器学习吸睛？
社交媒体 AI 只会捅篓子？它可是新一代反恐精英
不瞒你说，企业才[illegible]步的超强助攻！
四周劲敌林立，霓虹国如何在第四次工业革命上发力？
谁都知道AI 将改变零售业，不过还有哪些注意事项？
AI 热潮注定会失败？机器学习智能才是王道？

热门搜索
HTC 黑科技 周鸿祎 Google play

，在笔者看来，自然语言处理还有很大发展空间。

在更大的价值链里，SyntaxNet其实算是一种较低级别的技术，它就像是一个改良的钻头，钻头本身无法给你石油，石油本身无法给你提供能量和塑料，能量和塑料本身也无法自动形成某种产品。但如果整个价值链的瓶颈是石油开采效率，那么大幅提高钻头技术（虽然是一种底层技术）也是非常重要的。

在笔者看来，在自然语言处理中语法解析就是一个瓶颈技术，如果它有四、五年时间做优化改进，将会对自然语言处理产生巨大影响。现在你可能会说，我之所以觉得这是个问题，是因为这项技术正从学术研究转变为商业化应用。但我所能说的就是，这其实是一种逆转因果关系：正是因为我理解问题的重要性，所以我投入其中，而不是相反。

好了，我知道即便某个技术遇到瓶颈，但也无法否定其重要性。SyntaxNet如何向前迈一大步呢？如果你已经在Stanford CoreNLP中使用了神经网络模型，那么可以肯定的是，你正在使用的其实是一种算法，在设计层面上这种模型和算法其实是完全一致的，但在细节上却不一样。使用spaCy语法解析模型也是如此。从概念上讲，SyntaxNet的贡献可能会让人觉得没那么大，毕竟它主要用于试验，优化和改进。然而，如果谷歌不做这项工作，可能就没有人会去做。可以说，SyntaxNet为神经网络模型打开了一扇窗，人们从中看到了一个充满各种想法创意的美丽风景，研究人员也正忙于探索这一切。当然啦，行业内也会有一种偏见，认为SyntaxNet会让研究人员看上去（感觉上）更聪明。可能，我们最终会有一个非常准确的语法分析模型，但是这个模型无法实现正确的假设（当然在系统设计的角度准确性是十分重要的），继而导致未来神经网络模型的发展越来越慢。在CoreNLP模型说明推出后的六个月，首个SyntaxNet论文才发布出来，他们使用了更大的网络，更好的激活函数，以及不同的优化方法，不仅如此，SyntaxNet还应用了更具原则性的定向搜索方法，进而取代了目前更多工作。使用 LSTM模型可以实现同样准确的并行工作，而不是按照SyntaxNet论文里描述的那样，同时发布前馈网络。

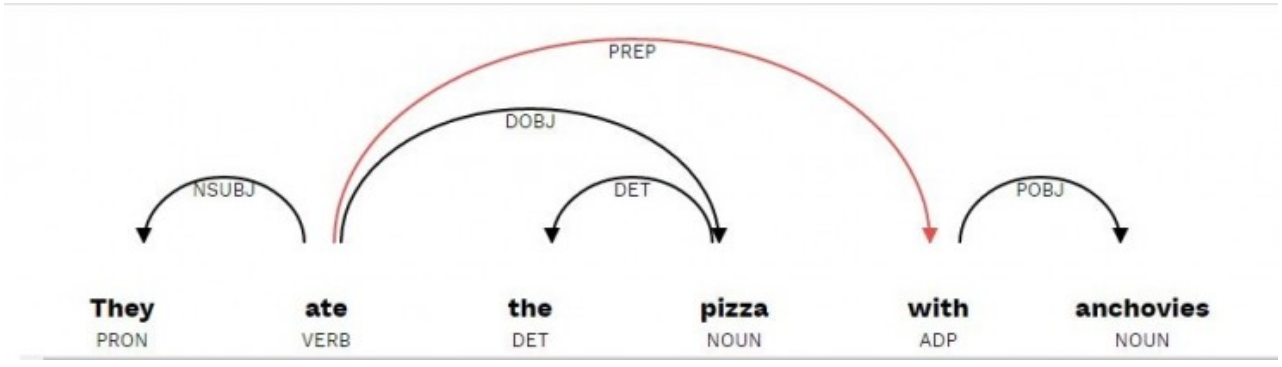
SyntaxNet用来做什么？

SyntaxNet语法解析器可以描述一个句子的语法结构，帮助其他应用程序理解这个句子。自然语言会产生很多意想不到的歧义，人们通常可以利用自己的知识过滤掉那些产生歧义的。举个大家比较喜欢的例子：

他们吃了加凤尾鱼的披萨（They ate the pizza with anchovies）

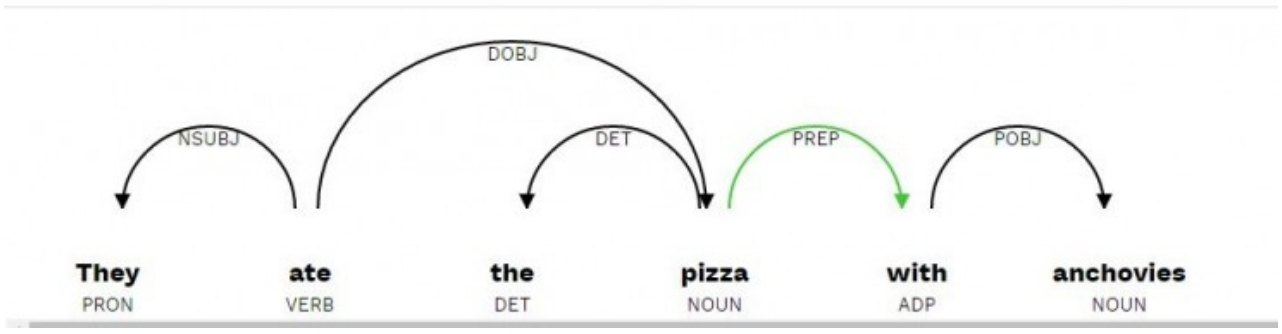
图片来源：spaCy

正确的语法分析是将“with”和“pizza”联系在一起，也就是他们吃了加凤尾鱼的披萨；



图片来源：spaCy

而不正确的语法分析是将“with”和“eat”联系在一起，他们和凤尾鱼一起吃了披萨。



图片来源：spaCy

如果你想要更形象地感受这个技术，不妨可以看下我们的displaCy demo，或是看一个简明的，基于规则方法的例子，去了解语法树是如何计算出来的。“单词与单词”关系熟也可以用来识别一些简单的语法语义，这样可以便于扩展形成“单词包”技术（比如word2vec，它是一个将单词转换成向量形式的工具。可以把对文本内容的处理简化为向量空间中的向量运算，计算出向量空间上的相似度，来表示文本语义上的相似度。）举个例子，我们解析去年Reddit论坛上的每一个评论，相比于严格限制空格分割单词的方法，使用word2vec显然更有帮助，因为后者可以分析短语，实体和单词生成一个很不错的概念图。

SyntaxNet是一个训练和运行句法依赖解析的模型库。这个模型可以较好地权衡语义分析速度和准确度。可能是为了显得更时髦些，谷歌给这个模型起了个很酷的名字——Parsey McParseface。希望他们能够继续延续这种时髦的命名方式，我觉得未来应该有个更好的方式，让模型发展时间轴显得更清楚一些，自然语言处理技术也应如此。

SyntaxNet带来多大进步？

虽然打上了“当今世界上最准确的语义分析”标签，但Parsey McParseface其实只比最近的相关语义分析研究领先了一点点而已，如今的语义分析模型使用了更加复杂的神经网络架构，但有更多限制性的参数调整。因此，很多相似的技术也将不再会局限在学术圈。另一方面，如果你关心这种模型是否能够实实在在做一些事情，那么现实可能会让你有些失望了，目前这些技术还无法真正去“做事”。自从去年SyntaxNet论文发布之后，笔者本人一直在断断续续的研究神经网络模型spaCy，但是效果并不太好，我们想要让spaCy便于安装，我们想要它在单CPU上快速运行，我们还想要它保持多线程，不过所有这些要求目前都很难实现。

对于语义分析基准，Parsey McParseface在每秒600个单词的速度下，准确度可以超过94%。同样地，spaCy每秒识别1.5万字的精准度为92.4%。这个准确度可能听上去不是很高，但对于应用程序来说，其实已经算非常好的了。

任何预测系统，通常最重要的考虑因素就是基准预测的差异，而不是绝对进度。一个预测天气的模型，今天和昨天的准确度可能是一样的，但是它不会增加任何价值。关于依存关系语法分析，大约80%的依赖关系都很简单明确，这意味着，一个只正确预测那种依附关系的系统正在注入少量额外信息，这种能力不是只查看每个单词就能做到的，而是要考虑词和词之间的关系。

总而言之，我认为在目前人工智能的大趋势下，Parsey McParseface是一个非常好的里程碑。重要的是，它可以实现多快的速度，以及能实现多么先进的自然语言处理技术。我觉得以前有很多想法不能实现，但是肯定会有那一刻的到来，一瞬间所有都变得可行了。

下一步是什么？

最让我兴奋的是，通过Parsey McParseface模型设计，自然语言处理技术有了一个非常清晰的方向，这时你可能会说：“好的，如果它有作用就太好了。”2004年，语义分析领域的领军人物之一 Joakim Nivre表示，这种类型的语法解析器可以一次性读句子，继而减少错误理解。它适用于任何状态表达，任何行为集合，任何概率模型架构。举个例子，如果你解析一个语音识别系统的输入，你可以让语法解析器优化语音识别器，在基于句法环境下猜测对方要说的话。如果你使用知识库，那么可以扩展状态表达，使其中包含你的目标语义，让它学习语法。

联合模型和半监督学习一直是自然语言理解研究最完美的体现。从来没有人怀疑它们的优点——但是如果没有一个具体的方法，这些技术也只是陈词滥调罢了。很明显，理解一个句子需要正确地拆分单词，但这样做会带来很多问题，更难以找到一个满意的解决方案。此外，一个自然语言理解系统应该可以利用现

有的大量未标注文本，这同样需要不同类型的模型支持。我认为，针对上述两个问题，一个过渡的神经网络模型能够给出答案。你可以学习任何架构，你看到的文本越多，你学习的就越多，而且神经网络模型也不需要添加任何新参数。

显然，我们想要在Parsey McParseface和spaCy模型之间构建一座桥梁，这样在spaCy应用程序接口的支持下，你才能使用更加准确的模型。不过，对于任何单独用例，让这种技术真正发挥作用总是会出现一些变数。特别是每一个应用程序中总会存在不同类型的文本，如果数据模型能调整到域，准确度才能够有实质提升，比如一些完整编辑的文本，像财务报告，你必须要让语义分析模型把“市值”这个词考虑成决定性指标，才能更好地理解全文；但是如果在理解Twitter上的推文时，你让语义分析模型将“市值”理解成决定性指标，通常是没有什么意义的。

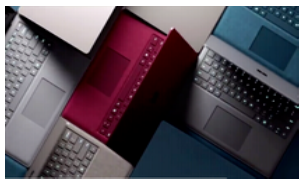
我们的目标就是要提供一系列预先训练模式，去解决这一问题，让语义分析模型适应不同的语言和风格。我们也有一些令人非常兴奋的想法，尽可能轻松地帮助每个用户训练属于自己的自定义模型。我们认为，在自然语言处理中，算法总是冲在最前面，而数据往往滞后。我们希望解决这个问题。

via spaCy

相关文章

谷歌 人工智能 自然语言处理 NLP

Windows 10 S宣战谷歌？听听微软高管怎么说的



微软发布Windows 10 S、Surface Laptop，要和谷

为了从谷歌、微软等挖角，BAT都使出了哪些奇

谷歌AI商业化又进一步：快速辨别糖尿病视网膜病变，