自然语言处理的语义建模介绍-云栖社区-阿里云

在本文中，我将简单介绍自然语言处理(NLP)的语义建模思想。

语义建模(或语义语法)通常与语言建模(或语言语法)相比较，我们现在从二者的定义和对比来理解语义建模。

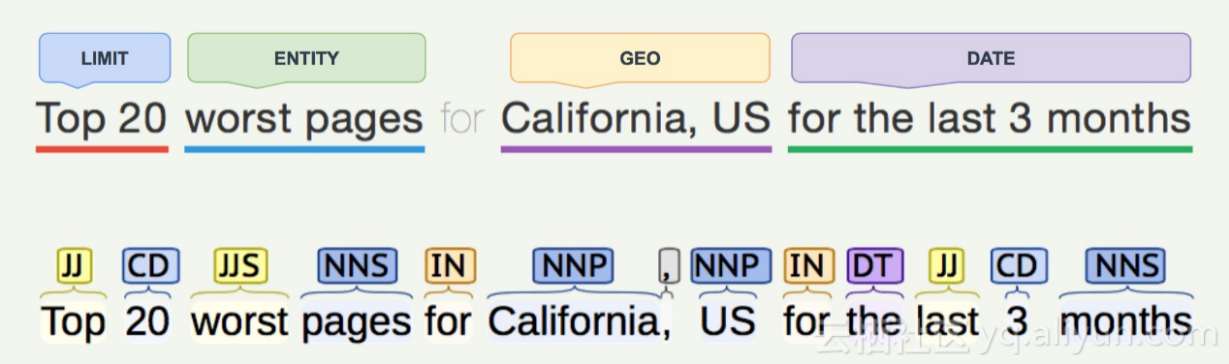
语言与语义

语义语法和语言语法都定义了理解自然语言句子的形式。语言语法涉及名词、动词等语言范畴。另一方面，语义语法是这样一种语法，它的非终端不是名词或动词等一般结构或语言类别，而是人或公司等语义类别。

语言和语义两种方法在20世纪70年代几乎同时出现。作为计算语言运动的一部分，语言建模一直备受关注，并且是整个自然语言处理开发的基础。

语义建模一开始很受欢迎，但由于技术复杂，很快就失败了。然而，近年来，语义建模经历了复兴，现在它几乎是等所有商业自然语言处理系统的基础。

理解语义语法和语言语法之间区别的最简单方法是看下面的插图:



在上图中，上下句子相同，但处理方式不同。下面的部分使用传统的语言语法进行分析，其中每个单词都用PoS (语音点)标记，如名词的NN，形容词的JJ，等等。但是，上面的部分使用语义语法进行解析，一个或多个单词形成高级语义类别，如日期或地理，而不是对单个单词进行PoS标记。

这当然是语言方法的高度简化的定义，因为我们忽略了共同参照分析、命名实体解析等。

这种将单个单词组合成高级语义实体的能力被引入来帮助解决困扰早期NLP系统的一个关键问题——语言歧义。

语言歧义

请看下图：

尽管两句的语言特征基本相同，但语义却完全不同。仅仅用语言语法来解决这种歧义将需要非常复杂的语境分析——如果甚至在这种语境可用的时候——而且在许多情况下，根本不可能确定地做到这一点。

另一方面，语义语法允许以简单和完全确定的方式清楚地解决这种歧义。使用构造得当的语义语法，Friday和Alexy这两个词将属于不同的类别，因此不会导致含义混乱。

请注意，这些词除了具有相同的PoS标签之外，还具有不同的“命名实体”分辨率。然而，在更复杂的现实生活中，名为实体解析的例子被证明远没有那么有效。

语义语法实例

让我们来看一下语义语法的简单定义。

不管配置的具体语法如何，语法通常被定义为语义实体的集合，其中每个实体至少具有一个名称和同义词列表，通过这些名称和列表可以识别该实体。

例如，以下是网站和用户实体及其同义词的简单定义:

根据这种语法，以下句子将全部分解成相同的两个语义实体:

1.Website user

2.HTTP address online user

3.Website online user

4.<WEBSITE> <USER>

语义实体序列可以进一步绑定到用户定义的意图，以便最终采取行动。这种用户定义意图的集合通常构成完整的NLP流水线。

当然，现实生活中的系统支持更复杂的语法定义。同义词的定义有很多不同的方式，因为同义词本身有很多不同的类型；语义实体可以有数据类型，并且可以被组织成分层组来帮助短期记忆处理——不幸的是，所有这些都超出了这个博客所涉及的范围。你可以在这里找到这种语法支持的一个例子。

决定论与概率论

我们强调了上述语义语法方法的确定性。尽管语言和语义语法应用的具体实现可以是确定性的和概率性的，但是语义语法几乎总是导致确定性处理。

原因在于语义语法本身的性质，它基于简单的同义词匹配。正确定义的语义语法允许对语义实体进行完全确定性的搜索。根本没有“猜测”——语义实体要么被毫不含糊地找到，要么没有。

由此产生的语义语法决定论是一个惊人的品质。虽然概率方法可以在许多众所周知的场景中工作，如情感分析、支持聊天机器人或文档理解，但它根本不适合NLP / NLU驱动的业务数据报告和分析。例如，的反馈是85 %还是86 %是正面的并不重要，只要它朝着正确的方向发展。然而，另一方面，报告销售数字必须准确无误，必须与会计系统的数据精确匹配。即使像“你上一季度的总销售额是1亿美元，概率是97%”这样的高概率结果在任何情况下都是毫无价值的。

尽管语义语法有很多好处，但有一个明显的限制阻碍了它的发展(至少最初是这样)，即它只能应用于狭窄的数据域。

通用与特定领域

尽管语言语法对所有数据域都是通用的(因为它处理动词和名词等通用语言结构)，但语义语法及其基于同义词的匹配仅限于特定的、通常非常狭窄的数据域。原因在于，为了创建语义模型，需要拿出一个所有实体的详尽集合，最令人畏惧的是，所有同义词的集合。

对于一个特定的数据域来说，这是一项可管理的任务，并且在很大程度上得益于复杂的现实系统。对于一般的NLU来说，它和一般的人工智能(AGI)一样，语义建模根本不起作用。

在过去的十年中，有很多研究致力于推进具有闭环人类管理和监督式自学习能力的语义建模，但事实上，语义建模在处理特定的、定义良好的和可理解的数据域时应用得最好。

有趣的是，流行的NLP / NLU深度学习(DL)方法对于特定的数据域几乎没有足够好的效果。这是因为缺乏DL模型训练所需的足够大的预先存在的训练集。这就是传统的闭环人类管理和自学ML算法在语义建模系统中盛行的原因。

监督式自主学习

人类管理(或人类切换)和监督式自学习算法是两种相互关联的技术，有助于在开发新语义模型时，减少为语义实体提供详尽同义词集的问题。

这两项工作如下。首先创建语义模型，其中包含语义实体的基本同义词集，这可以相当快地完成。一旦使用此模型的NLP / NLU应用程序开始操作此模型无法自动“理解”的用户语句，将进入固化。在人类管理中，用户句子将被修改以适应模型，并且自学习算法将“学习”该修改，并且下次将自动执行该修改，而不需要人工切换。

在此过程中有两个关键属性:

Â·Â 人类活动改变了用户输入，以适应现有的语义模型，也就是说，改变了用户句子，使其可以自动回答。通常，它包括纠正拼写错误、口语化、俚语、删除停止词或添加缺少的上下文。

Â·Â 用户句子中的这种变化(即固化)被输入到自学习算法中，以便将来“记住”。因为这种改变最初是由一个人执行的，这个人使这种自学习成为一个有监督的过程，并消除了累积学习错误的引入。

在所有这些中，重要的是监督允许在语义建模“进一步学习”时保持语义建模的确定性。语义模型通过监控和有监督的自我学习，在每次监控中学习更多，最终可以比开始时学到更多的知识。因此，这个模型可以从小处开始，通过人的交互来学习——这个过程和许多现代人工智能应用程序没有什么不同。

以上为译文。

本文由北邮@爱可可-爱生活 老师推荐，阿里云云栖社区组织翻译。

文章原标题《Introduction Into Semantic Modeling for Natural Language Processing**》，译者：**Mags，审校：袁虎。

文章为简译，更为详细的内容，请查看原文。