135 | ACL 2018论文精读：什么是对话中的前提触发？如何检测？

2018-08-13 洪亮劼

AI技术内参 进入课程 >

讲述：初明明
时长 07:33 大小 3.46M

今天，我来和你分享 ACL 2018 的第二篇最佳论文，题目是《让我们“再”次做到：检测副词前提触发词的第一种计算方法》（Let’s do it “again”: A First Computational Approach to Detecting Adverbial Presupposition Triggers）。

这篇论文的作者都来自加拿大麦吉尔大学（McGill University）的计算机系。前三位学生作者是这篇论文的共同第一作者，对论文的贡献相同。他们的导师张智杰（Jackie Chi Kit Cheung）助理教授是这篇论文的最后一个作者。张智杰于 2014 年从多伦多大学博士毕业，之前曾两次在微软研究院实习过，他长期从事自然语言处理的研究。

论文的主要贡献

这篇论文的背景要从“语用学”（Pragmatics）说起。语用学是语言学的一个分支学科，与符号学理论相互交叉、渗透，研究语境对语言含义产生的影响和贡献。语用学包括言语行为理论、对话内涵义、交流中的对话，以及从哲学、社会学、语言学以及人类学等角度解析人类语言行为的研究。

语用学分析研究语言行为（如招呼、回答、劝说）的文化准绳和发言规则。不同的文化之间皆有约定俗成、客套的对话，在跨文化交流中，为了避免因为语言规范的差异而在交谈之中产生误解，社会语言学的知识与务实能力是语言学习者所不能忽视的。

在语用学中，“前提”（Presuppositions）是交谈的参与者共同约定的假设和认知，而且在谈话中被广泛使用。同时，在这篇论文中，作者们把提示“前提”的“表达”（Expression）定义为“**前提触发**”（Presupposition Triggers），包括一些动词、副词和其他短语。为了更加清晰地说明这些概念，作者们举了这么一个例子。

假设我们现在有两句话：

1. 约翰再次要去那家餐厅（John is going to the restaurant again）。
2. 约翰已经去过了那家餐厅（John has been to the restaurant）。

第一句话要能够成立必须要建立在第二句话的基础上。特别是“前提触发”词“再”（Again）的使用，是建立在第二句话真实的情况下。换句话说，第一句话必须在第二句话的上下文中才能够被理解。值得一提的是，即便我们对第一句话进行否定，“约翰不打算再去那家餐厅了”（John is not going to the restaurant again），依然需要第二句话的支持。也就是说，“前提触发”词在这里并不受到否定的影响。

这篇论文的核心贡献就是对以副词为主的前提触发词进行检测。这里面包括“再”（Again）、“也”（Also）和“还”（Still）等。再此之前，还没有对这方面词汇进行检测的学术研究工作。能够对这类前提触发词进行检测，可以应用到**文本的归纳总结**（Summarization）和**对话系统**等场景中。

为了更好地研究这个任务，作者们还基于著名的自然语言处理数据 Penn Treebank 和 English Gigaword，建立了两个新的数据集从而能够进行触发词的分类检测工作。最后，作者们设计了一个基于“关注”（Attention）机制的时间递归神经网络（RNN）模型来针对前提触发词进行检测，达到了很好的效果。

论文的核心方法

现在，我们来讨论这篇论文的一些细节。

首先，我们来看看**数据集是如何生成的**。数据中的每一个数据点都是一个**三元组**，分别是标签信息（正例还是负例），文本的单词，文本单词所对应的“词类标签”或简称为 POS 标签（例如动词、名词）。

数据点正例就表明当前数据包含前提触发词，反之则是负例。另外，因为我们需要检测的是副词性的前提触发词，因此我们还需要知道这个词所依靠的动词。作者们把这个词叫作副词的“**管理词**”（Governor）。

作者们首先针对文档扫描，看是否含有前提触发词。当发现有前提触发词的时候，提取这个触发词的管理词，然后提取管理词前 50 个单词，以及管理词后面到句子结束的所有的单词。这就组成了正例中的单词。当找到了所有的正例之后，作者们利用管理词来构建负例。也就是说，在文本中寻找哪些句子含有一样的管理词，但并不包括后面的前提触发词，这样的句子就是负例。

下面，我们来看一下作者们提出模型的一些构成。从大的角度来说，为了识别前提触发词，作者们考虑了一个**双向 LSTM**的基本模型架构，在此之上有一个“关注机制”，在不同的情况下来选择 LSTM 的中间状态。

具体来说，整个模型的输入有两部分内容。

第一部分，是**文本的单词进行了词向量（Embedding）的转换**。我们已经反复看到了，这是在自然语言处理场景中利用深度学习模型必不可少的步骤。这样做的好处就是把离散数据转换成了连续的向量数据。

第二部分，是**输入这些单词相对应的 POS 标签**。和单词不一样的是，POS 标签依然采用了离散的特性表达。

然后，连续的词向量和离散 POS 标签表达合并在一起，成了双向 LSTM 的输入。这里，利用双向 LSTM 的目的是让模型针对输入信息的顺序进行建模。跟我们刚才提到的例子一样，前提触发词和其所依靠的动词，在一个句子的段落中很明显是和前后的其他单词有关联的。因此，双向 LSTM 就能够达到对这个结构进行记忆的目的，并且提取出有用的中间变量信息。

下面需要做的就是**从中间变量信息到最终的分类结果的变换**。这里，作者们提出了一个叫“**加权池化网络**”（Weighted Pooling Network）的概念，并且和“关注”机制一起来进行这一步的中间转换。

可以说，作者们这一步其实是借助了计算机视觉中的经常使用的卷积神经网络 CNN 中的池化操作来对文档进行处理。具体来说，作者们把所有 LSTM 产生的中间状态堆积成一个矩阵，然后利用同一个矩阵乘以其自身的转置就得到了一个类似于相关矩阵的新矩阵。可以说，这个新矩阵是完全抓住了当前句子通过 LSTM 中间变量转换后所有中间状态的两两关系。

然后，作者们认为最后的分类结构就是从这个矩阵中抽取信息而得到的。至于怎么抽取，那就需要不同的权重。这种根据不同的情况来设置权重的机制就叫作“关注”机制。经过矩阵中信息的抽取，然后再经过全联通层，最终就形成了标准的分类输出。

论文的实验结果

作者们在我们上面提到的两个新数据集上进行了实验，并且和一系列的方法进行了比较。其他的方法包括简单的对数几率回归方法（Logistic Regression），简化了的但是依然利用了双向 LSTM 结构的模型，还有一个利用 CNN 来进行提取信息的模型。

在两个数据集上，论文提出的方法比对数几率回归以及 CNN 的方法都要好 10%～20% 左右。和简化的 LSTM 模型相比，优势并没有那么大，但依然有统计意义上的好效果。

小结

今天我为你讲了 ACL 2018 的另外一篇最佳论文。

一起来回顾下要点：第一，这篇论文的背景是语用学，核心贡献是对以副词为主的前提触发词进行检测；第二，论文的核心方法是提出一个双向 LSTM 的基本模型架构，并利用“关注机制”，根据不同的情况来设置权重；第三，论文构建了两个数据集，取得了较好的实验结果。

最后，给你留一个思考题，这篇论文使用了双向 LSTM 的架构，能不能使用单向 LSTM 呢？

欢迎你给我留言，和我一起讨论。

上一篇　134 | ACL 2018论文精读：问答系统场景下，如何提出好问题？

下一篇　136 | ACL 2018论文精读：什么是“端到端”的语义哈希？

精选留言

写留言

> 由作者筛选后的优质留言将会公开显示，欢迎踊跃留言。