# 第3章 词性标注

在上一章中，我们对自己所要做的所有预处理步骤进行了讨论，以便在工作中可以应对任何文本语料库。我们现在应该可以放心地对任何种类的文本进行解析和清理了。我们应该执行所有的文本预处理，譬如针对任意文本的标识化处理、词干提取以及停用词移除等。我们可以根据自己的需要执行和定制所有相关的预处理工具。到目前为止，我们已经重点讨论了针对文本型文档的一般性预处理工作。现在，让我们将焦点转向那些动作更为激烈的NLP预处理步骤吧。

在本章，我们将具体讨论何谓词性标注，以及词性（POS）在NLP应用环境中的意义。我们也会学习如何用NLTK标注有意义的信息，并介绍可用于NLP密集型应用程序的各种标注器。最后，我们还将学习如何用NLTK来标注命名实体。我们会详细讨论各种NLP标注器，并且还会提供一些代码片段来帮助您理解它们。我们也将会看到这些标注器的最佳实践，以说明在什么地方应该使用哪种标注器。在读完本章之后，读者应了解：

* 何谓词性标注，以及其在NLP中的重要性。
* 如何使用NLTK中形形色色的词性标注。
* 如何用NLTK创建自定义的词性标注。

## 何谓词性标注

其实，我们可能在小时候就已经听说过**词性（POS）**这个术语了。尤其在形容词和副词的实际使用上，要想掌握其中的窍门还是很花时间的。这两者的区别究 竟是什么？或许，我们可以考虑将所有这方面的知识进行编码以创建一个系统。这件事看起来好像挺容易的的，但这几十年来，将这些知识转化为可编码的机器学习模型一直都是一个非常难解的NLP问题。在我个人看来，虽然目前最先进的词性标注算法在预测给定单词的词性上已经有了较高的精确度（约97％）。但词性标注领域中仍有大量的研究在等着我们。

对于像英语这样的语言来说，它们在新闻和其他领域往往都有许多已被标注的语料库。这为我们带来了许多先进的算法。尽管在一般情况下，这其中的一些标注器应该足以应付各种跨不同领域的、文本化的使用环境了。但在某些特定的用例中，POS的预判可能还是有些不尽如人意。对于这些用例，我们可能就得要从头开始建立一个标注器了。然而，如果我们想要深入了解POS，就得先要对机器学习领域中的一些技术有一个基本的了解。虽然这部分有些是我们要在*第6章：文本分类*中要讨论的内容，但在这里，我们必须讨论一下相关基础知识，以便我们可以创建一个自定义的POS标注器。

首先，我们要学习一些现成可用的POS标注器，及其相配的token集。在此，我们会看到一些以**元组**形式存在的、独立单词的POS。然后，我们再将焦点转移到其中一些标注其的内部工作原理上来，最后，我们还将讨论如何从头开始创建一个自定义的标注器。

我们在讨论POS时，总少不了会用到Penn Treebank[[1]](#footnote-1)这个最常用到的POS标记库：

|  |  |
| --- | --- |
| **标签** | **相关说明** |
| NNP | 专用名词的单数形式 |
| NNPS | 专用名词的复数形式 |
| PDT | 前置限定词 |
| POS | 所有格结束符 |
| PRP | 人称代词 |
| PRP$ | 所有格代词 |
| RB | 副词 |
| RBR | 相对副词 |
| RBS | 最高级副词 |
| RP | 小品词 |
| SYM | 符号（数学符号或特殊符号） |
| TO | To |
| UH | 叹词 |
| VB | 动词的基本形式 |
| VBD | 动词的过去式 |
| VBG | 动词的动名词用法 |
| VBN | 动词的过去分词 |
| WP | Wh-代词 |
| WP$ | 所有格wh-代词 |
| WRB | Wh-副词 |
| # | 井号符 |
| $ | 美元符 |
| . | 句号 |
| , | 逗号 |
| : | 分号，分隔符 |
| ( | 左括号 |
| ) | 右括号 |
| " | 直双引号 |
| ' | 左单引号 |
| " | 左双引号 |
| ' | 右单引号 |
| " | 右双引号 |

这些看起来就是我们在小学英语课堂上所学的东西，对吧？现在，既然我们已经了解了这些标签所代表的含义，下面就可以来运行一个实验了：

>>> import nltk

>>> from nltk import word\_tokenize

>>> s = "I was watching TV"

>>> print nltk.pos\_tag(word\_tokenize(s))

[('I', 'PRP'), ('was', 'VBD'), ('watching', 'VBG'), ('TV', 'NN')]

如果我们只想使用新闻类或与其类似语料库的POS，那么您就只需要知道前三行代码即可。在这段代码中，我们先对一段文本进行了标识化处理，然后对其调用了NLTK库中的pos\_tag方法，得到了一组（词形，词性标签）形式的元组。这就是一个NLTK库内置的POS标注器。

|  |
| --- |
| 该方法在内部使用的是由maxent分类器（我们会在后面章节中讨论分类器这个话题）所调校出来的模型，以预测特定单词属性所属于的类型标签。  如果您想了解更多其中的细节，可以参考下面这个链接： https://github.com/nltk/nltk/blob/develop/nltk/ tag/\_\_init\_\_.py |

由于NLTK库使用了Python中强大而有效的数据结构，所以我们在使用由NLTK库处理之后所输出的成果时会具有更多的灵活性。现在，您一定很想见识一下POS在真实应用中的典型用法究竟会是怎样的。在一个典型的预处理操作中，我们通常要尽可能地找出所有的名词。下面我们来看一段代码，它会帮我们找出给定句子中的所有名词：

>>> tagged = nltk.pos\_tag(word\_tokenize(s))

>>> allnoun = [word for word,pos in tagged if pos in ['NN','NNP'] ]

现在，我们来试着回答一下下面的问题：

* 我们可以在进行词性标注之前先删除掉停用词吗？
* 我们要如何获取到该句子中所有的动词？

### Stanford标注器

NLTK库还有一个非常棒的的特性，就是它还有许多针对其它的预置标注器的封装器，例如**Stanford工具包**。下面的POSTagger就一个很常见的例子：

>>> from nltk.tag.stanford import POSTagger

>>> import nltk

>>> stan\_tagger = POSTagger('models/english-bidirectional-distdim. tagger','standford-postagger.jar')

>>> tokens = nltk.word\_tokenize(s)

>>> stan\_tagger.tag(tokens)

|  |
| --- |
| 如果您想要使用上述代码，就必须先从下面的链接中获取Stanford标注器： http://nlp.stanford.edu/software/stanford-postagger-full-2014-08-27.zip.  然后将其中的jar文件和模型文件解压到一个文件夹中，然后在POSTagger()的参数中指定其绝对路径。 |

现在我们来总结一下，用NLTK库实现标注任务的方式主要有两种：

1. 使用NLTK库或其他库中的预置标注器，并将其运用到测试数据上。这两种标注器应该足以应付纯英语文本环境，以及非特殊领域语料库中所有的词性标注任务了。
2. 基于测试数据来创建或调校出适用的标注器。这就等于我们要处理一个非常特殊的用例或者开发一个自定义的标注器了。

下面，我们来更深入地了解一下一个典型的POS标注器在其内部究竟做了些什么。

### 深入了解标注器

一个典型的标注器通常要用到大量的调校数据，它主要被用于标注出句子中的各个单词，并为其标上POS标签。标注是个纯手动的操作，看起来应该像这样：

Well/UH what/WP do/VBP you/PRP think/VB about/IN the/DT idea/NN of/IN ,/, uh/UH ,/, kids/NNS having/VBG to/TO do/VB public/JJ service/NN work/NN for/IN a/DT year/NN ?/.Do/VBP you/PRP think/VBP it/PRP 's/BES a/DT ,/,

上面这些样例取自Penn Treebank总机的语料库。人们已经针对一些大型语料库的做了大量的手动标注工作。甚至我们还有一个叫**做语言数据联盟**（**简称LDC**）[[2]](#footnote-2)组织，那里的人们就花了很多时间来研究不同语言的标注、不同的文本种类以及不同的标注操作，譬如词性标注，句法分析标注、以及对话标注等（我们会在后面讨论这些话题）。

|  |
| --- |
| 读者可以从<https://www.ldc.upenn.edu/>处获得上述所有的资源以及更多的相关信息。（尽管LDC可以免费提供一小部分数据的一小部分，但如果需要我们也可以购买去整个标注语料库。NLTK库中大概内置了PTB的大约10％的数据。） |

如果我们还希望调校出属于自己的POS标注器，就需要针对特定的领域来执行相关的标注操作。这种标注操作就需要有这些领域的专家来协助。

通常情况下，像词性标注这样的标注问题往往会被视同为顺序性的标签化问题或者某种分类问题，后者特指人们为特定token所生成的正确标签，并用相关判别模型对其进行预判的那一类问题。

为了不让读者直接掉入过于复杂的应用实例中，我们先从一些简答的标注方法开始。

下面我们来看一段代码，它将会告诉我们Brown语料库中各POS标签的分布频率：

>>> from nltk.corpus import brown

>>> import nltk

>>> tags = [tag for (word, tag) in brown.tagged\_words(categories='news')]

>>> print nltk.FreqDist(tags)

< FreqDist: 'NN': 13162, 'IN': 10616, 'AT': 8893, 'NP': 6866, ',': 5133,

'NNS': 5066, '.': 4452, 'JJ': 4392 >

如您所见，NN是这里出现频繁最高的标签，我们可以基于此来创建一个非常幼稚的POS标注器，用于给所有的测试文本分配NN标签。 当然，我们也可以使用NLTK中的一个叫做DefaultTagger的函数来做这件事。DefaultTagger函数就是个顺序性标注器，下面我们就来讨论一下这个标注器。该标注器会去调用一个evaluate()函数，主要用于评估相关单词POS的准确度。这是我们针对Brown语料库的标注器所用的基准。在default\_tagger这个案例中，我们的预测准确率大约在13％左右。下面，我们将会把同样的基准进一步推广到所有的标注器上。

>>> brown\_tagged\_sents = brown.tagged\_sents(categories='news')

>>> default\_tagger = nltk.DefaultTagger('NN')

>>> print default\_tagger.evaluate(brown\_tagged\_sents)

0.130894842572

### 顺序性标注器

毫无疑问，上面那个标识器简直是弱爆了。其实，DefaultTagger本质上只是基类SequentialBackoffTagger的一部分而已，后者是一个顺序性的标注服务。标注器会试着基于其所处的上下文环境来模型化相关的标签。而且如果它不能进行正确的标签预测，就会去咨询BackoffTagger。通常情况下，DefaultTagger参数都可以被当作一个BackoffTagger实体来使用。

下面，让我们继续来看一些更复杂的顺序性标注器。

#### N-gram标注器

**N-gram**标注器是SequentialTagger的一个子类，它会在其所在的上下文环境中标注出前n个单词，并预测给定token的POS标签。这些标注器中还包含了一些人们常用的变体，即UnigramsTagger、BigramsTagger和TrigramTagger：

>>> from nltk.tag import UnigramTagger

>>> from nltk.tag import DefaultTagger

>>> from nltk.tag import BigramTagger

>>> from nltk.tag import TrigramTagger # we are dividing the data into a test and train to evaluate our taggers.

>>> train\_data = brown\_tagged\_sents[:int(len(brown\_tagged\_sents) \* 0.9)]

>>> test\_data = brown\_tagged\_sents[int(len(brown\_tagged\_sents) \* 0.9):]

>>> unigram\_tagger = UnigramTagger(train\_data,backoff=default\_tagger)

>>> print unigram\_tagger.evaluate(test\_data)

0.826195866853

>>> bigram\_tagger = BigramTagger(train\_data, backoff=unigram\_tagger)

>>> print bigram\_tagger.evaluate(test\_data)

0.835300351655

>>> trigram\_tagger = TrigramTagger(train\_data,backoff=bigram\_tagger)

>>> print trigram\_tagger.evaluate(test\_data)

0.83327713281

其中，基于一元模型的标注将只考虑相关标签的条件频率，以及针对每个给定token所能预测到的、频率最高的标签。而BigramTagger参数将会考虑给定的单词和该单词的前一个单词，其标签将以元组的形式来相关被测试单词所得到的标签。类似地，TrigramTagger参数会将查找过程扩展到给定单词的前两个单词。

很明显，TrigramTagger参数的覆盖范围会是比较小的，而实例精度则会是高的。从另一方面来说，UnigramTagger的覆盖范围则会比较大一些。为了让准确率与反馈率之间保有一定的平衡，我们在上述代码片段中结合了这三种标注器。首先，它会进行对给定单词的三元模型查找，以预测标签的顺序；前提是如果没有发现其Backoff有指向BigramTagger参数和UnigramTagger参数，以及最后的NN标签。

#### 正则表达式标注器

接下来，我们还要再来看一个顺序性的标注器类，这是一个以标注器为基础的正则表达式。在该类中，我们的任务不再是去寻找确切的单词了。现在，我们可以在定义一个正则表达式的同时定义出给定表达式所对应的标签。例如在下面的代码中，我们会看到一些最常见的正则表达式模式是如何获取不同词性的。我们知道其中有一些模式都有其各自所关联的POS类别。例如我们知道在英语文章中，任何以*ness*结尾的词都会是一个形容词。如果不做这样的标注，我们就会需要去写一堆正则表达式和纯Python代码来做这件事，而NLT K中的RegexpTagger参数为我们提供了基于POS的模式的这样一种优雅的方式。而且这种方式也可以被用于所有与POS模式相关的领域。

>>> from nltk.tag.sequential import RegexpTagger

>>> regexp\_tagger = RegexpTagger(

[( r'^-?[0-9]+(.[0-9]+)?$', 'CD'), # cardinal numbers

( r'(The|the|A|a|An|an)$', 'AT'), # articles

( r'.\*able$', 'JJ'), # adjectives

( r'.\*ness$', 'NN'), # nouns formed from adj

( r'.\*ly$', 'RB'), # adverbs

( r'.\*s$', 'NNS'), # plural nouns

( r'.\*ing$', 'VBG'), # gerunds

(r'.\*ed$', 'VBD'), # past tense verbs

(r'.\*', 'NN') # nouns (default)

])

>>> print regexp\_tagger.evaluate(test\_data)

0.303627342358

正如您所看到的，我们在这里只是使用了一些基于POS的显模式，就能达到大约30％的准确度。如果我们以混合的方式来使用正则表达式标注器，譬如说如通过BackoffTagger，就有可能会使性能得到改善。在预处理步骤中，我们还有另一个正则表达式标注器的使用用例，就是用它来替换原生的Python函数string.sub()，我们可以用该类标注器来标注日期模式、资金模式、位置模式等信息。

* 我们能否修改N-gram标注器那一节中hybrid标注器的代码，使之成为一个可用的正则表达式标注器？这样做能否改善性能？
* 我们能否基于标注日期和货币的正则表达式来写一个标注器？

### Brill标注器

Bril标注器是一种基于转换操作的标注器，其思路是我们可以先对给定标签做一个猜测，然后在下一轮迭代中，基于标注器接下来所学到的规则设置返回到原先的错误上并修复它。这也是一种监督型的标注方式，但与N-gram标注不同的是，后者会在调校数据中对N-gram模式来进行计数，而在这里我们要查找的是转换规则。

如果该标注器是从Unigram / Bigram这两个准确度可接受的标注器开始做起的，而后再到Brill标注器，我们要查找的就不再是一个三元组，而是一些基于标签、位置以及单词本身的规则。

例如我们可以有这样一条规则：

当目标单词的前置词被标注为TO时，其标注NN 就应该被替换为VB。。

在基于UnigramTagger打了一些标签之后，如果我们面对的只是一个简单的规则，那就可以去持续地改进这些标注。这种改进是一个交互的过程。通过几次迭代操作和一些更优的规则，Brill标注器是可以超越某些N-gram标注器的。我们在这里唯一的建议是要意标注器在调校数据集上被过度配适（over-fitting）的问题。

|  |
| --- |
| 读者可以在下面链接中找到更多在工作中被使用的规则实例：  <http://stp.lingfil.uu.se/~bea/publ/megyesiBrillsPoSTagger.pdf> |

* 您是否能基于自己的观察写出更多的规则？
* 请试着将Brill标注器与UnigramTagger搭配使用。

### 基于机器学习的标注器

到目前为止，我们所用到的都是一些来自于NLTK或Stanford的、用于预先调校的标注器。虽然我们在上一节的一些实例中使用了它们，但这些标注器的内部对我们来说仍然是一个黑盒子，例如pos\_tag内部使用的是**最大熵分类器（MEC）**。而StanfordTagger所采用的也是最大熵算法的一个修改版本。但它们属于不同的模型。其中有许多基于**隐马尔可夫模型（HMM）**和**条件随机场（CRF）**的标注器，它们所使用的都是生成性模型。

当然，本书所要介绍的内容覆盖不了上述所有的话题。但我会尽力推荐一些相关的NLP类，以便读者对这些概念能有一个大致上的了解。我们将会在*第6章：文本分类*中重点介绍一些分类技术，其中也不乏一些非常高级的NLP话题，它们也会需要更多的关注。

如果非要我做一个简短说明，我们也可以将这种归类型的词性标注问题看作是一种分类问题，我们会为其指定某个具体的单词以及其它相关的特性，譬如它的前置词、上下文语境、的形态变化等。我们会将给定单词归类到相应的POS类别中，而它的其它部分也会被根据相似的特性被模型化，以形成某种生成性模型。读者们可以参考下面提示框中列出的链接，以查看其中某些话题的资料。

|  |
| --- |
| NLP CLASS：<https://www.coursera.org/course/nlp>  HMM：<http://mlg.eng.cam.ac.uk/zoubin/papers/ijprai.pdf>  MEC: <https://web.stanford.edu/class/cs124/lec/Maximum_Entropy_Classifiers.pdf>  <http://nlp.stanford.edu/software/tagger.shtml> |

## 命名实体识别（NER）

除了POS之外，找出文本中的实体项也是最常见的标签化问题之一。通常情况下，NER主要由实体名、位置和组织。当然，有些NER系统要标注的实体不只有这三项。这也可以被视作是一个顺序性的标签化问题，我们会利用上下文语境和其它相关特性来标签化这些命名实体。当然，NLP领域中还有更大量的研究在推进当中，人们正试图研究如何标签化生物实体、零售产品实体等。和之前，NLTK库中NER标注的方式也主要有两种。一种方式是使用预先调校好的NER模型，我们只要读取测试数据的得分即可。另一种方式是建立一个机器学习基本模型。为此，NLTK库为我们提供了ne\_chunk()方法和一个封装了Stanford NER标注器的命名实体识别系统。

### NER标注器

NLTK库提供的的命名实体提取方法是ne\_chunk()。我们已经用一小段代码给读者演示了如何用它来对任意语句进行标注。这种方法需要我们先进行文本的预处理，即先对语句进行标识化处理，然后再进行语块分解和词性标注的处理顺序，之后才能进行命名实体的标注。虽然NLTK库在这里会用到ne\_chunking，且并没有拆分任何东西，但它标注了多个分词，以便能通过它们调用到某个有意义的实体项。

对于命名实体来说，NE语块分解的方式是基本相同的：

>>> import nltk

>>> from nltk import ne\_chunk

>>> Sent = "Mark is studying at Stanford University in California"

>>> print(ne\_chunk(nltk.pos\_tag(word\_tokenize(sent)), binary=False))

(S (PERSON Mark/NNP) is/VB Z studying/VBG at/IN (ORGANIZATION Stanford/NNP University/NNP) in/IN NY(GPE California/NNP)))

如您所见，ne\_chunking方法主要用于识别相关的人员（即姓名），以及他所在的地点（即位置）和组织。如果其binary参数被设置为True，该方法就会给出整个句子的树结构以及上面的每一个标签。而当该参数被设置为False时，该方法就给我们提供具体的人员、位置和组织的信息，其情况与之前我们使用Stanford NER标注器的例子一样。

和POS标注器的情况类似，NLTK库中也封装了Stanford NER。该NER标注器具有更高的准确度。下面，我们就通过一段代码来看看这种标注器的具体使用。您将会在这个例子中看到，我们可以只用三行代码标注出所有的实体：

>>> from nltk.tag.stanford import NERTagger

>>> st = NERTagger('<PATH>/stanford-ner/classifiers/all.3class.distsim. crf.ser.gz',... '<PATH>/stanford-ner/stanford-ner.jar')

>>> st.tag('Rami Eid is studying at Stony Brook University in NY'.split())

[('Rami', 'PERSON'), ('Eid', 'PERSON'), ('is', 'O'), ('studying', 'O'),

('at', 'O'), ('Stony', 'ORGANIZATION'), ('Brook', 'ORGANIZATION'),

('University', 'ORGANIZATION'), ('in', 'O'), ('NY', 'LOCATION')]

如果我们仔细观察，就会发现即使对一段非常小的测试语句来说，Stanford标注器的执行性能也要好于NLTK ne\_chunk标注器。

如今，虽然这些各式各样的NER标注器已经成为了一种非常受欢迎的、用来执行各种通用实体标注的解决方案，但我们还是有必要调校出属于自己的标注器，以应对像生物学、医学这样特定领域的标注实体任务，因此我们需要创建出属于自己的NER系统。当然，我我在这里也会向您推荐一个NER：Calais。它的标注方式不只有典型的NER，还有更多实体项。而且这个标注器的性能也非常不错：

https://code.google.com/p/python-calais/

## 该让你练练手了

下面我们来回答一下之前章节中提出的问题：

* 我们能否在执行词性标注之前移除掉停用词？

不能；如果我们移除掉这些停用词，就等于丢失了上下文语境，而有一些POS标注器（预先调校模型）是要以单词的上下文语境为特征来标出给定单词的POS的。

* 我们要如何获取相关语句中的所有动词？

我们可以用pos\_tag来获取该语句中的所有动词

>>> tagged = nltk.pos\_tag(word\_tokenize(s))

>>> allverbs = [word for word,pos in tagged if pos in

['VB','VBD','VBG'] ]

* 我们能否修改N-gram标注器那一节中hybrid标注器的代码，使之成为一个可用的正则表达式标注器？这样做能否改善性能？

是的，我们可以修改N-gram标注器那一节中hybrid标注器的代码，使之成为一个可用的正则表达式标注器：

>>> print unigram\_tagger.evaluate(test\_data,backoff= regexp\_tagger)

>>> bigram\_tagger = BigramTagger(train\_data, backoff=unigram\_ tagger)

>>> print bigram\_tagger.evaluate(test\_data)

>>> trigram\_tagger=TrigramTagger(train\_data,backoff=bigram\_tagger)

>>> print trigram\_tagger.evaluate(test\_data)

0.857122212053

0.866708415627

0.863914446746

而且其性能也得到了改善，因为我在这里加入了一些基于模式的基本规则，替代了预测出现频率最高的标签。

* 我们能否基于标注日期和货币的正则表达式来写一个标注器？

是的，我们可以基于标注日期和货币的正则表达式来写一个标注器。其代码如下：

>>>date\_regex = RegexpTagger([(r'(\d{2})[/.-](\d{2})[/.-](\d{4})$'

,'DATE'),(r'\$','MONEY')])

>>>test\_tokens = "I will be flying on sat 10-02-2014 with around

10M $ ".split()

>>>print date\_regex.tag(test\_tokens)

|  |
| --- |
| 最后两个问题并没有确切的答案。  因为这里有许多规则要取决于读者自己的观察，所以并没有正确/错误的答案。 |

现在，您能否再来挑战一下*第1章：自然语言处理简介*中词云这样的题目？当然，其范围只局限于名词和动词。

参考资料：

<https://github.com/japerk/nltk-trainer>

<http://en.wikipedia.org/wiki/Part-of-speech_tagging>

<http://en.wikipedia.org/wiki/Named-entity_recognition>

<http://www.inf.ed.ac.uk/teaching/courses/icl/nltk/tagging.pdf>

<http://www.nltk.org/api/nltk.tag.html>

## 本章小结

本章的主旨是要向读者介绍NLP预处理步骤中最有用的标注操作。一般情况下，我们所讨论的词性问题中包含了POS在NLP语境中所具有的意义。当然，我们还具体讨论了在NLTK库中各种不同的、用于预先调校POS标注器的方式，它们是如此的简单易用，并且能创建出如此精彩的应用。然后，我们有对所有可用的词性标注选项进行了探讨，譬如N-gram标注、正则表达式标注等。另外，我们还开发了一些混合型的标注器，以应对一些特定领域的语料库。

其中，我们简单的介绍了一个典型的预先调校标注去的构建过程。我们也探讨了各种可能用于解决标注问题的方法。最后，我们还讨论了NER标注器，以及该标注器在NLTK库中的工作方式。我希望读者在阅读完本章之后，能理解POS和NER在通用NLP语境中的重要性，并且知道如何运行并使用本章中所列出的NLTK代码，这样本章的目的就达到了。但旅程并没有就此结束。我们现在只是了解了一些粗浅的NLP预处理步骤，以及在大多数实践应用中POS、NER的主要运用。而在问答系统、文本综述和语音应用这一类更为复杂的NLP应用中，我们会需要用到语块分解、语法解析、语义学这些更深层的NLP技术，而这些正是我们下一章将要讨论的话题。

1. 译者注：Penn Treebank原本是一个NLP项目的名称，该项目主要用于对相关语料进行标注，标注内容包括词性标注以及句法分析。其语料来源是1989年的华尔街日报，包含了2499篇文章。这里指代该项目所标注的结果。 [↑](#footnote-ref-1)
2. 译者注：LDC，全名Linguistic Data Consortium，这是一个由大学、图书馆、企业、政府、研究机构共同合办的联合企业，成立于1992年，目前由宾夕法尼亚大学负责主要运营。LDC最初的角色只是保存与分发科研要用到的语言数据，后来有了资金，就开始自行收集并构建一些数据，如今该组织已经拥有了非常多的语言数据资源，成为了最主要的科研语言资源管理分发机构之一。 [↑](#footnote-ref-2)