《多语自然语言处理：从原理到实践》

书籍信息

版 次：1
页 数：
字 数：
印刷时间：2015年02月01日
开 本：16开
纸 张：胶版纸
包 装：平装
是否套装：否
国际标准书号ISBN：9787111484912
丛书名：计算机科学丛书?

内容简介

《多语自然语言处理：从原理到实践》是第一本全面阐述如何构建健壮和准确的多语自然语言处理系统的图书，由两位资深专家编辑，集合了该领域众多尖端进展以及从广泛的研究和产业实践中总结出的实用解决方案。
第一部分介绍现代自然语言处理的核心概念和理论基础，展示了如何理解单词和文档结构、分析语法、建模语言、识别蕴涵和检测冗余。第二部分彻底阐述与构建真实应用有关的实际考量，包括信息抽取、机器翻译、信息检索、文摘、问答、提炼、处理流水线等。

作者简介

nielM.Bikel现为Google公司高级研究科学家，正在开发用于自然语言处理和语音识别的新方法。在IBM工作期间，他为IBM的GALE多语种信息抽取和自动应答系统构架了拦截系统。在宾夕法尼亚大学攻读博士后期间，他建造了第一个可扩展的多语种语法分析引擎。
ImedZitouni现为微软公司高级研究员。2004~2012年，他是IBM公司高级研究科学家，领导IBM公司的阿拉伯语信息抽取和数据资源工作组。在此之前，他还曾领导DIALOCA的语音/自然语言处理组和Bell实验室/阿尔卡特朗讯的语言建模和呼叫路由工作。他的研究涉及机器翻译、自然语言处理和口语对话系统。"

目 录

译者序
前言
关于作者
第一部分理论
第1章找出词的结构
11词及其部件
111词元
112词形
113词素
114类型学
12问题和挑战

121不规则性
122歧义性
123能产性
显示全部信息

在线试读部分章节

第一部分
Multilingual Natural Language Processing Applications:From Theory to Practice
理论
第1章“找出词的结构”，描述如何识别人类语言中不同类型的词，如何建立词的内部结构、语法性质、词法概念的模型。
第2章“找出文档的结构”，讨论如何找出文档结构，并将其分解为更容易处理的单位，例如句子或表示同一话题的文本段。
第3章“句法”，描述如何找出句子的结构。
1～2
第4章“语义分析”，探索找出句子意义表示的自动方法。
第5章“语言模型”，讨论如何建立一个模型，该模型可对每个可能的有限长度的词串赋以一个概率估算或分数。
第6章“文本蕴涵识别”，讨论确定一段文本中的指定事实是否为另一段文本中的事实所蕴涵的方法。
第7章“多语情感与主观性分析”，探索确定句子是否是主观的并确定所表达的意见的倾向性和其他性质的方法。
第1章
Multilingual Natural Language Processing Applications:From Theory to Practice
找出词的结构
Otakar Smr, HyunJo You
人类语言很复杂。我们用语言来表示思想，获取信息，推断出意义。语言表达并非没有组织。其结构多样，复杂程度千差万别，复杂结构由基本部件组成，在一定的上下文中通过共现来表示比其孤立使用时更精细的意义及其意义间的关系。 第一部分 Multilingual Natural Language Processing Applications:From Theory to Practice 理论 第1章“找出词的结构”，描述如何识别人类语言中不同类型的词，如何建立词的内部结构、语法性质、词法概念的模型。 第2章“找出文档的结构”，讨论如何找出文档结构，并将其分解为更容易处理的单位，例如句子或表示同一话题的文本段。
第3章“句法”，描述如何找出句子的结构。 1～2
第4章“语义分析”，探索找出句子意义表示的自动方法。 第5章“语言模型”，讨论如何建立一个模型，该模型可对每个可能的有限长度的词串赋以一个概率估算或分数。 第6章“文本蕴涵识别”，讨论确定一段文本中的指定事实是否为另一段文本中的事实所蕴涵的方法。 第7章“多语情感与主观性分析”，探索确定句子是否是主观的并确定所表

达的意见的倾向性和其他性质的方法。 第1章 Multilingual Natural Language Processing Applications:From Theory to Practice 找出词的结构 Otakar Smr, HyunJo You 人类语言很复杂。我们用语言来表示思想，获取信息，推断出意义。语言表达并非没有组织。其结构多样，复杂程度千差万别，复杂结构由基本部件组成，在一定的上下文中通过共现来表示比其孤立使用时更精细的意义及其意义间的关系。 整体上理解语言不可行。语言学家从不同的角度、不同的细节层次来考察语言，比如形态学研究词的可变形式和功能，而句法则研究词如何排列构成短语、子句和句子。由于发音而导致的词结构限制由语音学描述，而书写的规则则构成了语言的正字法。语言表达式的意义属于语义学的内容，词源学和词汇学则研究词的演变并解释词之间的语义、形态和其他联系。 词可能是语言最直观的单位，但实际上定义什么是词颇为棘手。词的研究是句法、语义抽象及其他与语言相关的高级话题的前提。形态学是语言处理的必要部分，尤其在多语的环境下变得越来越重要。 本章将探索如何识别人类语言中不同类型的词，如何建立词的内部结构、语法性质、词法概念的模型。词结构的发现称为形态分析（morphological parsing）。
这个任务有多困难？决定因素有很多。 3 在某些语言中，词由空格或标点分割；但是在另一些语言中，书写系统使读者区分词或者确定其精确的语音形式。有些语言的词不随上下文变化，而另一些语言的词会根据句法和语义有不同的词形变化。 11词及其部件 在大多数语言中，词被定义为能形成完整言语的最小语言单位。词的最小语义部分称为词素(morpheme)。根据交流方式的不同，词素可用形素(grapheme)（比如字母和字符等书写符号）拼写出或用音素（phoneme）（口语中可区分的语音单位）说出
在手语中用的符号也由称为音素的元素构成。
。确定词、词素和短语之间精确的分界并不总是很容易［1,2］。 111词元
假设英语中的词只由空格和标点隔开［3］，考虑例11： 例11Will you read the newspaper? Will you read it? I wont read it 如果我们懂词源和句法知识，那么我们注意到这里有两个词可能和假设有些冲突：newspaper和wont。前者是一个复合词，有明显的派生结构。如果有词典或其他语言证据可佐证该词的来源的假设，我们可能会更详细地描述它。书面上，newspaper及其相关概念和单独的news与paper是不同的。然而，在口语中其区别却不甚明显，词的识别成了一个问题。 为了一般性，语言学家喜欢把wont分解为两个语法词，或称词元，其中每个词元有其独立的作用并有规范形式。从结构上说，wont可被分析为will后面跟随not。在英语中，这种词的切分(tokenization)和规范化(normalization)也许很少，而在其他语言中，这种现象可能很多。 在阿拉伯语或希伯来语中［4］，某些词元在书写时需要与前后的词元连写，也可改变其形式。其内在的词法或句法单位可能体现在紧缩的一串字母中，并非能明晰地分解为词。很多语言中的词元有这种行为，这种词元经常被称为附着词。 在汉语、日语［5］、泰语的书写系统里，不采用空格来隔开词。在某种程度上形式地可区分的单位是句子或子句。在韩语中，字符串称为eojeol（词节），粗略地对应于语音或认知单位，比词大，比子句小［6］，如例12所示： 4 例12 haksayngtuleykeyman cwusyessnunte使用耶鲁拼音表示韩文，通过点号标出原始的字符。使用连字号标记形态学边界，加号分开词元。 haksayngtuleykeyman cwusiessnunte student+plural+dative+only give+honorific+past+while while(he/she)gave(it)only to the students 尽管如此，基本的形态单位被视为有其句法地位［7］。在这些语言中，词的切分，或称分词(word segmentation)，是形态分析的基础性步骤，也是大多数语言处理应用的前提。 112词形原文lexeme按照字面意义是指词典的基本单位，实际就是“ 词 ”。当强调其基本意义时，也翻译为“ 语素 ”。这里为了和“ word ”相区分，译为“ 词形 ”。不采用目前

的流行翻译“ 词位 ”。——译者注 词这个术语，通常我们不但指其在给定上下文中的语言形式，而且表示其形式背后的概念，以及可表示该概念的其他形式的集合。该集合被称为词形，或词项，它们构成了一个语言的词典。词可根据其行为分为动词、名词、形容词、连词、小品词等词类（词性）。词形的引用形式也称为原形(lemma)。 当我们把词转化为其他形式时，比如把单数的mouse转为复数mice或mouses，我们说对该词形进行了屈折变化。当把一个词形变化为形态上相关的另一个词，而不管其词类是否相同时，我们称对该词形进行了派生。例如，名词receiver和reception是由动词 to receive 派生而来。 例13Did you see him? I didnt see him I didnt see anyone
例13提出了didnt的切分和anyone的内部结构问题。在释义I saw no one中，词to see被屈折变化成saw以表示其过去时态的语法功能。同样，him是he或甚至表示所有人称代词的更抽象的语素的从格形式。在上述释义中，no one可以被认为是和词nobody同义的最小词。如果我们把两个紧密相关的词元no one当作一个固定的词理解，那么，对于用语法描述什么是一个词的困难就不复存在了。在例子13的捷克语翻译中，词vid ě t “ to see ”屈折变化为过去时，而形式是由第一人称和第二人称的两个词元组成（即vid ě la jsj ‘ youFEMSG saw ’ and nevid ě la jsem ‘ IFEMSG did not see ’）。捷克语的否定是一个屈折变化参数，而不仅是句法的，需同时在动词及其相关代词中标记，正如例14所示：
例14Vid ě las ho? Nevid ě la jsem ho Nevid ě la jsem nikoho saw+youare him? notsaw Iam him notsaw Iam noone 这里，vid ě las 是 vid ě la jsi “ youFEMSG saw ”的紧缩形式。jsi “ you are ”中的s是附着词，由于捷克语的自由语序，可以附着在几乎任何词的后面。因此我们可提问：Nikphos nevid ě la? “ Did you see no one? ”，此处代词nikoho “ no one ”后面跟了这个附着词。 113词素 形态理论的主要差别在于是否并且如何将词形的性质与其结构部件联系起来［8,9,10,11］。 5 这些部件通常称为“ 节 ”(segment)或“ 形元 ”(morph)。词的表意形元称为某种功能的词素（morpheme）。 人类语言采用很多手段，可将形元或词素合并成词形。最简单的形态过程将形元一个接一个连接起来，如disagree ments，其中agree是一个自由词素，其他三个是表达语法意义的黏着词素，合起来表示词的整体意义。 在更复杂的情形中，形元间可互相作用，其形式可有语音或书写的变化，称为“ 形音 ”（morphophonemic）变化。词素的其他形式称为变体词素（allomorph）。在韩语中，形态变化和词素的形式依赖于语音的例子比比皆是。很多词素随着其语音上下文不同而系统地改变其形式。下面的例15列出了表示过去时态的时态标记的变体词素e ss、ass、yess。前两个根据其前面动词词干的语音而变化，最后1个经常和动词ha “ do ”一起使用。适当的变体可直接跟在词干后面，也可以进一步紧缩，如例12中siess紧缩为sy ess。在形态分析中，变体词素规范化为词素的正规形式是有益的，尤其是当形元的紧缩与简单的切分相干扰的时候。 例15 紧缩形式(a）,(b)是普通的，但是需要引起注意，因为两个字符缩成了一个。其他类型（c）,(d),(e）语音上不可预测，或与具体词相关。例如，cohass “ have been good ”永远不能紧缩，而noh和ass被合并成了 nwass，如例15（e）所示。 还有形成词的其他语言手段需要加以解释，因形态分析过程本身并不是小事。连接操作可能伴有形元的嵌入或交缠，这在阿拉伯语中很普遍。即使在英语中，也存在将词内部的元音进行改变的非连接的屈折变化：请比较mouse和 mice、see和saw、read和read的音变。 在阿拉伯语中，内部的屈折变化经常发生，并且具有不同的性质。词内部的一部分，称为词干，可由词根和词素模式来描述。词的结构因此可由抽象了词根的、只显示模式和附着在其左右的其他形元来描述。使用Buckwalter标

记直译原来的阿拉伯文字。为了方便阅读，也给出了标准的语音转写，以减少歧义。"
显示全部信息