211院校研究生规划教材

计算语言学

JISUANYUYANXUE

张霄军 编著

陕西师范大学出版总
司

计算语言学

张霄军 编著

JI SUAN YU YAN XUE

陕西师范大学出版总社有限公司

图书代号 JC11N1066

图书在版编目(CIP)数据

计算语言学/张霄军编著. —西安:陕西师范大学出版总社有限公司, 2011.10
ISBN 978-7-5613-5815-3

Ⅰ.①计… Ⅱ.①张… Ⅲ.①计算语言学 Ⅳ.①H087

中国版本图书馆 CIP 数据核字(2011)第 203572 号

计算机语言学

编　　著/ 张霄军
责任编辑/ 颜　红
责任校对/ 张　立
封面设计/ 鼎新设计
出版发行/ 陕西师范大学出版总社有限公司
(西安市长安南路 199 号　邮编 710062)
网　　址/ http://www.snupg.com
经　　销/ 新华书店
印　　刷/ 西安交通大学印刷厂
开　　本/ 787mm×960mm　1/16
印　　张/ 13
字　　数/ 204 千
版　　次/ 2011 年 10 月第 1 版
印　　次/ 2011 年 10 月第 1 次印刷
书　　号/ ISBN 978-7-5613-5815-3
定　　价/ 25.00 元

陕西师范大学研究生教材建设项目资助

目 录

第1章 计算语言学简介

1.1 什么是计算语言学?

Grishman(1986:4)将“计算语言学(computational linguistics)”定义为“一门研究如何利用计算机来理解和生成自然语言的科学(Computational linguistics is the study of computer systems for understanding and generating natural language)”。这个定义指明了计算语言学的研究目标和研究手段。“理解和生成自然语言”,是计算语言学的研究目标;“利用计算机”,是计算语言学的研究手段。更确切地说,应该是“利用计算机建立传输说话者所表述和听话者所理解的信息的计算模型(to reproduce the natural transmission of information by modeling the speaker's production and the hearer's interpretation on a suitable of computer)”(Hausser, 2001:1)。Allen(1995:1)则认为计算语言学的目标应该表述为“利用计算机科学的算法和数据结构来建立语言的计算理论(to develop a computational theory of language, using the notions of algorithms and data structures from computer science)”。

要实现语言的“生成”,首先得要实现计算机对自然语言的“理解”。有人认为,现阶段提出“理解”目标不切实际,因为目前占主流地位的统计方法所达到的目标只是“处理”,还谈不上“理解”。更进一步说,并非经过“理解”才能“处理”。但是,统计方法只是解决问题的方法之一,它不能“处理”所有语言的问题,只有计算机真正“理解”了语言,才能实现语言的生成。目前计算语言学

研究主要致力于自然语言的理解，自然语言生成方面的研究相对薄弱。

要使计算机理解自然语言，必须使之具备以下自然语言知识（Allen，1995：10）：

语音和音系学知识（phonetic and phonological knowledge）：主要关注语音（sounds）怎样转化为词（words）。

形态学知识（morphological knowledge）：主要关注词素（morphemes）怎样构成词。

句法知识（syntactic knowledge）：主要关注词怎样构成句子（sentences）。

语义知识（semantic knowledge）：主要关注词义（word meanings）怎样构成句义（sentence meanings）。

语用知识（pragmatic knowledge）：主要关注句子在不同语境（situations）中的使用。

语篇知识（discourse knowledge）：主要关注上下句之间的关系。

世界知识（world knowledge）：主要指说话者和听话者所具备的对外部世界的认知。

乍看上去，计算机要具备的自然语言知识似乎与传统语言学（traditional linguistics）和现代语言学（modern linguistics）的内容大致相当。传统语言学着重语言事实的描写，经验性质比较突出。现代语言学，例如乔姆斯基语言学（Chomskyan linguistics），理论性非常强，已经脱离了经验科学的范畴，我们称之为“理论语言学（theoretical linguistics）”。但计算语言学和这二者是有本质区别的。

理论语言学和计算语言学都是研究自然语言的，但服务对象有所不同：前者是面向人的（human-oriented），后者是面向计算机的（computer-oriented）。计算语言学是一门实验科学，所以它提出的问题既要符合自然语言处理的实际需要，又要用现有的计算机技术能够解决。超出计算机的能力，就不具有可行性。此外，计算语言学中研究对象的定义必须明确，不能含糊。例如汉语“词”的定义，理论语言学上的定义是：“词是最小的、能独立运用的语言单位”，这一定义并不清楚。语言学家也分析了词的一些特征，例如“结合紧密、使用稳定”等等，但没有定量标准。这样的定义对计算机来说是没有用的，那么计算语言学中“词”的定义是什么呢？“能在分词词表中找到的就是词，否则就不是词，或者是未登录词（unknown words）。”这样，计算机就去词表中查找，能找到的就是词，找不到的就划归到“未登录词”里去做下一步处理。

理论语言学研究主要不是考虑计算机的应用，因此无法提出自然语言处理

的问题和理论。例如,汉语自动分词问题就是从中文信息处理角度提出来的,汉语理论语言学研究从来没有、也不可能提出这样的问题。另一方面,理论语言学不一定要形式化(formalized),也没有为形式化提供任何手段。形式化(formalization)是数学表示的问题,包括两个方面:一是问题本身的形式化描述,二是解决问题的方法的形式化描述,后者通常用数学模型来体现。对于语言形式化的研究,本书在第2章有详细介绍。

而且我们认为,对于计算机而言,从"词"到"句"的直接理解跨度有点大。因此,我们增加了一个"过渡知识",即"短语知识(phrasal knowledge)",它主要关注从词到短语的转化。限于篇幅,本书涉及的语言知识主要是指形态学知识(第3章)、短语知识(第4章)和句法知识(第5章)。

当然,要让计算机掌握和具备以上的语言知识,计算语言学研究者首先得将这些知识形式化,并将其用算法的形式在计算机上加以实现。因此,冯志伟(2005:1)将计算语言学研究划分为以下四个阶段:

(1)把需要研究的问题在语言学上加以形式化,建立语言的形式化模型,使之能以一定的数学形式,严密而规整地表示出来。

(2)把这种严密而规整的数学形式表示为算法,使之在计算上形式化。

(3)根据算法编写计算机程序,使之在计算机上加以实现,建立各种实用的自然语言处理系统。

(4)对于建立的自然语言处理系统进行评测,使之不断地改进质量和性能,以满足用户的要求。

这四个阶段可以简单概括为"数学模型—算法表示—程序实现—质量评测"。对于计算语言学研究,最困难也最有挑战性的是第一步,即提出问题和理论并且将问题的解决办法用数学模型来表示。本课程的主要任务正在于此。至于算法表示,那是"算法分析"和"数据结构"等课程要解决的问题;至于程序实现,则是"程序设计语言"课程要解决的问题;而质量评测则贯穿于具体的研究任务的各个阶段。

从宏观上看,计算语言学的基本方法有两种:基于规则的(rule-based)方法和基于经验的(empirical-based)方法。前者的理论基础是语言学上的理性主义(rationalism),以乔姆斯基理论为代表。乔姆斯基认为人的语言能力是由遗传决定的,体现为若干条原则,在不同的自然语言中带上了不同的参数而已。语言学研究的目标就是人类的这种语言能力。至于言语,那只是语言能力的具体表现,不是语言学应该关注的重点。理性主义方法的特点是演绎法(deduction),从原则和参数演绎出规则,从规则推导出具体的句子。乔姆斯基语言学

虽然不属于计算语言学,但对于计算语言学的形成和发展有重大影响。基于规则的计算语言学研究方法中的理性主义体现在两个方面:第一,目标定位于"自然语言理解",希望在理解的基础上来处理自然语言。第二,方法的核心是"基于规则",希望根据通过内省和演绎得到的一整套规则来处理自然语言。

而基于经验的方法的理论基础是经验主义(empiricism),来源于 Shannon 的信息论(information theory)。信息论认为语言事件(语言表现)是有概率(probability)的,可以通过统计而得到这些概率,从而对自然语言处理(natural language processing, NLP)的各种具体问题进行决策。经验主义方法的特点是归纳法(induction),集中体现为"语料库语言学(corpus linguistics)"。与理性主义相对立,经验主义认为,完成自然语言处理任务不一定要经过"理解"的阶段,通过内省和演绎得到的"规则"往往是颗粒度较大的语言知识,只有通过运用统计方法,才能自动获得大量的、带概率的小颗粒度语言知识,从而处理大规模真实文本。有关信息论的内容,在第 2 章有更加详细的介绍。

从学科属性上来说,计算语言学到目前为止,其理论体系尚未建立起来,还不能算是一门理论科学。其主流方法(统计方法)是经验主义的,这充分表明计算语言学还是一门经验科学。另一方面,计算语言学又的确是一门实验科学,其理论和方法的正确性都需要通过在计算机上做实验来得到证明。而理论语言学则不是一门实验科学,有些问题本质上无法通过实验来研究,例如语言的发展规律。

1.2 计算语言学与自然语言处理

一般情况下,对于"计算语言学"和"自然语言处理"这两个术语是不加区分的。因为二者的本质是基本相同的,区别可能仅仅在于自然语言处理更注重实践,而计算语言学较重视理论。也可以说,计算语言学是建构自然语言处理系统的理论基础(刘海涛,2001:23)。Manaris(1998:5)认为自然语言处理可以定义为"研究在人与人交际中以及在人与计算机交际中的语言问题的一门学科,即研究表示语言能力和语言应用的模型,建立计算框架来实现这样的语言模型,提出相应的方法不断地加以完善,根据模型设计各种实用系统,并探讨这些实用系统的评测技术(the discipline that studies the linguistic aspects of human-human and human-machine communication, develops models of linguistic compe-

tence and performance, employs computational frameworks to implement processes incorporating such models, identifies methodologies for iterative refinement of such processes/models, and investigates techniques for evaluating the resultant systems)"。之前我们分析过计算语言学的四个研究阶段,该定义也印证了自然语言处理的主要内容就是这四个阶段。我们说本课程主要关注第一个阶段,也说明计算语言学是建构自然语言处理系统的基础。

计算机速度的提高和存储量的增加,使得计算语言学在语音合成(speech synthesis)、语音识别(speech recognition)、文字识别(character recognition)、拼写检查(spelling check)、语法检查(grammar check)这些应用领域,得到了商品化的开发。除了早期就开始的机器翻译(machine translation)和信息检索(information retrieval)等应用研究进一步得到发展之外,计算语言学在信息抽取(information extraction)、问答系统(question answering system)、自动文摘(text summarization)、术语的自动抽取和标引(term extraction and automatic indexing)、文本数据挖掘(text data mining)、自然语言接口(natural language interaction)、计算机辅助语言教学(computer-assisted language learning)等新兴的应用研究中,都有了长足的进展。此外,计算语言学的技术在多媒体系统(multimedia system)和多模态系统(multimodal system)中也得到了应用。

汉字识别的核心技术是字形特征的抽取和模式识别,识别结果是否能组织为有意义的文本,取决于对自然语言的理解。语音识别和语音合成则需要用到文语转换技术,即从文本到标音符号的相互转换,其中多音字的处理是关键。但汉字识别属于模式识别,语音识别和语音合成则属于数字信号处理,都不是计算语言学的研究内容。只有在识别之后进行语言处理时才是计算语言学发挥作用的时候,也就是说,其后处理属于计算语言学,后处理关系到整个识别系统的精度。

自动校对可大大减轻人工校对的工作量,使得这一环节跟出版业的其他环节的自动化相适应。计算机辅助语言教学属于现代教育技术,如果没有自然语言处理技术的支持,电子教案就基本上是纸质教案的翻版。好的教学软件应该包括更多的人机交互活动,例如习题的自动生成、作业的自动批改。

机器翻译的意义毋庸赘言,这是一种综合性最强的应用。仅就文本形式的翻译而言,就需要用到知识表示方法、机译词典构造、源语言的分析、目标语言的生成等各种技术。如果是口语现场翻译,还需要有语音识别、语音合成以及人机接口技术的配合。

智能检索,包括信息检索、信息抽取、文本挖掘、话题跟踪、文本分类、文本

过滤、问答系统等，是当前最热门的应用。文本分类是智能检索的一个重要方面，对于网站新闻频道的自动更新具有特殊意义。例如，中国搜索在线报告，它们的新闻频道就是实用文本分类技术而自动更新的，其他网站的最新消息可在两分钟内在它们的频道中得到反映。

自动文摘可帮助人们快速、准确、全面地获取信息，特别是因特网上的信息。简单的原文浓缩，就能起到一定的作用。哪些句子最能代表原文内容，需要根据其出现位置、所含词语来进行计算。如果要用不同于原文的句子来表示，则还需要用到语句分析和语句生成技术。

但计算语言学的研究内容和其主要应用不是一一对应的，后者需符合市场需要。有些基础研究本来就不是瞄准直接应用的，例如句法分析技术可在多种应用系统中起作用，但不可能独立成为一种社会大众所需要的应用。

Grishman(1986:4-5)将计算语言学的应用归结为三个大类：机器翻译、信息检索和人机界面(man-machine interface)。这三大类内容将分两章介绍，即机器翻译(第6章)和智能检索(第7章)，智能检索就涵盖了文本挖掘、信息检索和人机对话等内容。

第2章 计算语言学基础

2.1 概率论与信息论

在各种各样的信息中，人类的语言可以说是最复杂、最动态的一部分。如果人类对自己的语言知识规律认识不清，语言学家就无法为计算机制定完备的语言学规则，因此让机器模拟人类学习人类的语法、分析语句等活动在计算语言学领域就得不到有力的语言学支持。在这种情况下，美国数学家、信息论的创始人 Claude Shannon 大胆地提出来用数学的方法来处理自然语言的想法（Shannon，1948）。遗憾的是当时的计算机条件根本无法满足大量信息处理的需要，所以他的这个想法当时并没有被人们重视。20 世纪 70 年代初，有了大规模集成电路的快速计算机后，Shannon 的梦想才得以实现。而首先成功利用数学方法解决自然语言处理问题的是美国语音和语言处理专家 Fred Jelinek，他领导了一批杰出的科学家利用大型计算机来处理人类语言问题，统计语言模型（statistical language model）就是在那个时候提出的（Jelinek et al.，1975；Jelinek，1976）。

举一个数学在计算语言学领域的应用实例：在机器翻译、语音识别、印刷体或手写体识别、拼写纠错、文字输入和文献查询等自然语言处理领域中，我们都需要知道一个单词序列（words string）是否能构成一个大家能理解的句子，显示给使用者。对这个问题，我们可以用一个简单的数学统计模型来解决：如果 S 表示一连串特定顺序排列的词 $w_1, w_2, \cdots, w_n$，换句话说，S 可以表示某一个由一连串特定顺序排列的词而组成的一个有意义的句子。现在，机器对语言的识别

从某种角度来说，就是想知道 S 在文本中出现的可能性，也就是数学上所说的 S 的概率(probability)，用 $P(S)$ 来表示。怎样求解这个概率，这就需要用到概率论(probability theory)的知识。

信息是个很抽象的概念。我们常常说信息很多，或者信息较少，但却很难说清楚信息到底有多少，比如一本 50 万字的中文书到底有多少信息量。一条信息的信息量大小和它的不确定性有直接的关系。比如说，我们要搞清楚一件非常不确定的事，或是我们一无所知的事情，就需要了解大量的信息。相反，如果我们对某件事已经有了较多的了解，我们不需要太多的信息就能把它搞清楚。所以，从这个角度，我们可以认为，信息量的度量就等于不确定性的多少。那么我们如何量化地度量信息量呢？这就需要用到信息论(information theory)的知识。

本节简单介绍概率论和信息论的基础知识及其在自然语言处理中的应用。

2.1.1 概率论基础知识

2.1.1.1 经典概率论

概率的研究始于 17 世纪中期，但到 18 世纪就有了很大的发展。将概率作为一个体系进行整理研究是在 19 世纪初由法国天文学家、数学家拉普拉斯(Pierre Simon Laplace，1749—1827)完成的。拉普拉斯在 1812 年出版的《概率的分析理论》中，首先明确地对经典概率(classic probability)作了定义："全体共有 N 个事件，假定它们都是以相同程度确定的，发生 E 情形的有 r 个事件，那么 E 情形发生的概率是 r/N。"

在单词序列中识别出一个正确句子就是一个典型的概率问题。假设 S 是由 8 个单词组成的序列，那么根据全排列数公式①，应该会有 $8! = 40320$，即 40320 种 S，而正确的序列只有一种，即正确句子出现的概率 $P(S)$ 是 1/40320。

上例中每一种 S 都是 1 个事件，所以共有 40320 个事件。定义中的"E 情形"在例子中是指"正确的序列"，即实验者期望得到的正确句子，因此"r 个事件"只能是其中的 1 个事件，也就是说，正确序列发生的概率是 1/40320。在每次组合单词序列的试验中，这 1 个正确事件都可能出现，也可能不出现，这就是定义中所说的"它们都是以相同程度确定的"，即随机的。

我们将在试验(experiment)中可能出现也可能不出现的事情称为随机事件

① 全排列数公式：从 n 个不同元素中取出 $m(m \leqslant n)$ 个元素的所有排列，当 $m = n$ 时，为全排列。计算公式为：$P_n^n = n(n-1)(n-2)\cdots 1 = n!$

(random event),而每次试验必定发生的事情称为必然事件(certain event),如由这8个单词组成的任一序列。每次试验都不可能发生的事件称为不可能事件(impossible event),如由其他单词组成的序列、多于或少于8个单词组成的序列等。随机事件、必然事件和不可能事件统称为事件(event)。试验中直接观察到的最简单的结果称为基本事件(basic event),记为 A。构成基本事件的元素为样本点(sample point),试验中所有样本点构成的集合称为样本空间(sample space),记为 Ω。

概率具有三个基本性质:

(1) $1 \geqslant P(A) \geqslant 0, P(\Omega) = 1$。

(2)同一样本空间中互不相容的事件(incompatible event) A_1、A_2 之和的概率等于两事件概率之和,即 $P(A_1 + A_2) = P(A_1) + P(A_2)$,该性质又被称为概率的"加法原则(addition principle)"。

(3)两个独立事件(independent event)同时出现的概率等于该两事件概率的乘积,即 $P(A_1 A_2) = P(A_1) \cdot P(A_2)$,该性质也被相应地称为概率的"乘法原则(multiplication principle)",两个独立事件同时出现的概率又叫"联合概率(joint probability)"或者"同现概率(co-occurrence probability)"。

但是句子序列的组成是有规律可循的。有这样一种现象,有些词在出现时其后或者其前总会出现另外一个词,也就是说,这两个词是互为条件——你出来我必定出来。当然很多词并没有这样100%的同现概率,但也有一定的概率。基本事件的发生是有条件的,这就用到另外一种概率——"条件概率(conditional probability)"。在已知事件 B 发生条件下,事件 A 发生的概率称为事件 A 的条件概率,记为 $P(A|B)$。条件概率 $P(A|B)$ 与无条件概率 $P(A)$ 通常是不相等的。那么,如何计算条件概率呢?

例1　某工厂有职工500人,男女各一半,男女职工中非熟练工人分别为40人和10人,求任一女非熟练职工的概率。

解:令 A = {选出的职工为非熟练工人},B = {选出的职工为女职工},

则 $P(A) = \dfrac{50}{500}, P(B) = \dfrac{250}{500}, P(AB) = \dfrac{10}{500}$,

$$P(A|B) = \frac{10}{250} = \frac{\frac{10}{500}}{\frac{250}{500}} = \frac{P(AB)}{P(B)}, P(B|A) = \frac{10}{50} = \frac{\frac{10}{500}}{\frac{50}{500}} = \frac{P(AB)}{P(A)},$$

由此,我们可以给出定义:

设 A、B 为两事件，如果 $P(B)>0$，则称 $P(A|B)=\frac{P(AB)}{P(B)}$ 为在事件 B 发生的条件下，事件 A 的条件概率。同样，如果 $P(A)>0$，则称 $P(B|A)=\frac{P(AB)}{P(A)}$ 为在事件 A 发生条件下，事件 B 的条件概率。

这样我们就可以根据条件，对上述 8 个单词重新排序。这次就不需要 40320 次排序了，而是“有条件”地排序，即利用条件概率的公式，S 这个序列出现的概率等于每一个词出现的概率相乘，于是 $P(S)$ 可展开为：

$$P(S)=P(w_1)P(w_2|w_1)P(w_3|w_1w_2)\cdots P(w_n|w_1w_2\cdots w_{n-1})。$$

（公式 2.1）

其中 $P(w_1)$ 表示第一个词 w_1 出现的概率；$P(w_2|w_1)$ 是在已知第一个词的前提下，第二个词出现的概率；以此类推。不难看出，到了词 w_n，它的出现概率取决于它前面所有词。如果是 8 个单词，那么第 8 个单词的出现概率就要取决于它前面所有 7 个词。就是说，事件 A_1，A_2，…，A_8 的联合概率，等于 A_1 的概率连续乘以 $A_2 \sim A_8$ 的条件概率，而每个条件概率都是把各自的“历史”作为已知条件。从计算上来看，如果不对“历史”的长度加以限制，则计算量太大，各种可能性太多，无法实现。如果将“历史”长度限制为 $N-1$，就是 $N-1$ 阶马尔科夫链（$N-1$ order Markov chain），或者叫做 N 元语法（N-gram）。马尔科夫链的基础是马尔科夫假说（Markov assumption），即“[在一个句子中]某一个单词出现的概率只受到其之前几个单词的影响（the probability of the next word depends only on the previous words in the input）”（Manning & Schütze，1999：77）。用一个通俗的比喻来形容，一只被切除了大脑的白鼠在若干个洞穴间的蹿动就构成一个马尔科夫链。因为这只白鼠已没有了记忆，瞬间而生的念头决定了它从一个洞穴蹿到另一个洞穴；当其所在位置确定时，它下一步蹿往何处只与此刻的位置相关，与它以往经过的路径（path）无关。

我们可以用诗歌中元音（vowel）字母和辅音（consonant）字母交替变化的规律来验证这个假设。选取普希金（1799—1837）的长诗《叶甫盖尼·奥涅金》开头的两句，意为：“我不想取悦骄狂的人生，只希望博得朋友的欣赏（Не мысля гордый свет забавить，Вниманье дружбы возлюбя）”。这两句诗行可以看成是一个元音和辅音混合而成的“锁链”：在这条锁链上只有两种链环，C 代表辅音、V 代表元音（为了使问题简化起见，把两个元音字母算作辅音）。分别统计了在 C 后面出现 C 和 V 的概率 $P(C|C)$ 和 $1-P(C|C)$，以及在 V 后出现 C 和 V 的概率 $P(C|V)$ 和 $1-P(C|V)$，把结果与按照俄语拼音规则计算出的结果进行比