自然语言理解

(4)

宗成庆

中科院自动化研究所

模式识别国家重点实验室

cqzong@nlpr.ia.ac.cn

http://www.nlpr.ia.ac.cn/English/cip/cqzong.htm

第四章 语料库语言学

4.1 概述

语料库（corpus）就是存放语言材料的仓库（数据库）。

“语料库语言学（Corpus Linguistics）已经成为语言研究的主流。基于语料库的研究不再是计算机专家的独有领域，它正在对语言研究的许多领域产生愈来愈大的影响。”

—J. Thomas 等人为祝贺语料库语言学的主要奠基人和倡导者 G. Leech 六十岁生日而出版的语料库语言学研究论文集的开场白。

—[丁信善，1998]

4.1 概述

- 语料库语言学的定义
  - 根据篇章材料对语言的研究称为语料库语言学。

    - [Aijmer, 1991]

  - 基于现实生活中语言运用的实例进行的语言研究称为语料库语言学。

    - [McEnery, 1996]

  - 以语料为语言描写的起点或以语料为验证有关语言的假说的方法称为语料库语言学。

    -[Crystal, 1991]

4.2 语料库语言学研究的内容

- 四个方面：
  - 语料库的建设与编纂
  - 语料库的加工和管理技术
  - 语言研究中语料库的使用
  - 语料库语言学在计算语言学中的应用

4.3 语料库语言学的发展

- ～20世纪50年代中期：早期的语料库语言学
  - 语料库在语言研究中被广泛使用：语言习得、方言学、语言教学、句法和语义、音系研究
- 1957～20世纪80年代初期：沉寂时期
  - 1957年Chomsky 的《句法理论》及其以后一系列著作的发表，根本改变了语料库语言学的发展状况。
  - Chomsky 及其转换生成语法学派批判早期的语料库研究方法：
    - 基于语料库的研究方法有误
    - 语料的不充分性

4.3 语料库语言学的发展

- **20世纪80年代～：复苏与发展时期**
  - 第二代语料库相继建成：
  - 1983年英国Lancaster大学建成 Lancaster-Oslo / Bergen Corpus (LOB语料库): 研究英国英语，500语篇，每个语篇约2000词。
  - 法国国家科学研究中心与美国芝加哥大学联合建成法语语料库（Tremor de la Language Francaise, TLF语料库）：2000书面法语文本，1.5亿词。
  - 芬兰赫尔辛基大学建成历史英语语料库（The Helsinki Corpus of Historical English）:850-1720年, 1600万词。

4.3 语料库语言学的发展

• 1988年伦敦大学建成国际英语语料库（The International Corpus of English, ICE）：语料来自所有英语国家，各100万词，1990－1993年，口语和书面语各一半，18岁以上接受英语教育的成人。

➢ 基于语料库的研究项目增多

4.3 语料库语言学的发展

- 复苏原因

1）计算机的迅速发展；

2）转换生成语言学派对语料库语言学的批判不都正确，有的是片面的甚至是错误的。

4.4 语料库类型

- 按内容构成和目的划分
  - 异质的（heterogeneous）　　一[黄昌宁，2002]

    最简单的语料收集方法，没有事先规定和选材原则
  - 同质的（homogeneous）

    与“异质”正好相反，比如美国的 TIPSTER 项目只收集军事方面的文本。
  - 系统的（Systematic）

    充分考虑语料的动态和静态问题、代表性和平衡问题以及语料库的规模等问题。
  - 专用的（specialized）

    如：北美的人文科学语料库

4.4 语料库类型

- 按语言种类划分：
  - 单语的
    - （已切分的）具有词性标注
    - 句法结构信息标注（树库）
    - 语义信息标注
  - 双语的
    - 篇章对齐 / 句子对齐 / 结构对齐
  - 多语的

两个术语：生语料，熟语料

4.5 语料库建设

❑ 语料库的设计

◆ 静态与动态

语料库建设的另一种主张是动态的，或监督语料库（monitor corpus）：动态文本集，数据的收集通常是随遇的，而不是平衡的

◆ 代表性和平衡

一个语料库具有代表性，是指在该语料库上获得的分析结果可以概括成为这种语言整体或其指定部分的特性。

—[Leech, 1991]

如何达到不同部分之间的平衡？

4.5 语料库建设

- 规模

第一代语料库100万词次

1990s 1000－2000 万词次小型的一般语料库

一般而言，在保证质量的前提下应足够大。

4.5 语料库建设

- 语料库的管理与维护

  错误修正或改善

  版本升级

  语料库的检索系统、分析和处理工具的维护等

4.6 几个典型语料库

❑ 布朗语料库 (Brown Corpus)

- ➢ 20世纪60s，Francis 和 Kucera 在布朗（Brown）大学建立，是世界上第一个根据系统性原则采集样本的标准语料库
- ➢ 100万词规模
- ➢ 选自1961年美国人撰写出版的普通语体的文本
- ➢ 15种题材，共500个样本，每个样本不少于2000词
- ➢ 1961年布朗大学出版了当代英语词频词典
- ➢ 1970s Greene 和 Rubin 设计了TAGGIT词性标注系统（词类标记81种，上下文约束规则3300条），自动标注正确率77%。

4.6 几个典型语料库

- LLC口语语料库 (London-Lund Corpus of Spoken English)
  - 1960s 伦敦大学著名语言学家Quirk组织
  - 2000小时的对话和广播等口语素材
  - 瑞典隆德（Lund）大学教授 Svartvik 主持录入计算机
  - 英语口语调查（The Survey of Spoken English, SSE）
  - SSE 于1981年完成，建成 London-Lund Corpus of Spoken English (LLC)
  - 87个文本，每个文本约5000词，最终规模50万词
  - 5大类：面对面交谈，电话交谈，讨论、采访、辩论，未经准备的当众评论、论证、演讲，经准备的当众演讲
  - 标注：语调、节律、关键词（语段），词类、出现次数、搭配关系等

4.6 几个典型语料库

- 朗文语料库（Longman Corpus）
  - 朗文语料库委员会（Longman Corpus Committee）
  - January 1981- November 1990
  - 设计原则：1）尊重本族语言者的直觉和语料库权威
    2）向研究人员提供语料（英国50%，美国40%，其它国家10%）
    3）书面语
  - 选自1900～的20世纪英语：知识性（informative）文本60%，想象性（imaginative）文本40%
  - 10个分布广泛的领域：自然和纯科学、应用科学、社会科学、世界事务等
  - 2800 万词

4.6 几个典型语料库

❑宾州(Pennsylvania)大学语料库 (UPenn Tree Bank) (http://www.cis.upenn.edu/~treebank/home.html)

- 美国宾州大学计算机系 M.Marcus 教授主持
- 1993年完成约300万词次英语句子的语法结构标注
- 2000年完成第一版中文树库，约10万词次，4185个句子

例子：原始句子：他还提出一系列具体措施的政策要点。

词性标注：他/PN 还/AD 提出/VV 一/CD 系列/M 具体/JJ 措施/NN 和/CC 政策/NN 要点/NN 。/PU

4.6 几个典型语料库

例子：(IP (NP-SBJ (PN 他))
(VP (ADVP (AD 还))
(VP (VV 提出))
(NP-OBJ (QP (CD 一)
(CLP (M 系列)))
(NP (NP (ADJP (JJ 具体)
(NP (NN 措施)))
(CC 和)
(NP (NN 政策)
(NN 要点))))))
(PU 。))

4.6 几个典型语料库

- 北京大学语料库（ http://icl.pku.edu.cn/ ）
  - 北大计算语言学研究所俞士汶教授主持，北大、富士通、人民日报社共同开发
  - 《人民日报》1998年上半年全部文本（约1700万字）
  - 100万字切分及词性/注音标注
  - 完整的词语切分和词性标注信息

例子：

咱们/r 中国/ns 这么/r 大/a 的/u 一个/m 多/a 民族/n 的/u 国家/n 如果/c 不/d 团结/a ，/w 就/d 不/d 可能/v 发展/v 经济/n ，/w 人民/n 生活/n 水平/n 也/d 就/d 不/d 可能/v 得到/v 改善/vn 和/c 提高/vn 。/w

4.6 几个典型语料库

❑ 台湾中研院平衡语料库

(http://rocling.iis.sinica.edu.tw/ROCLING/corpus98/)

- 台湾中央研究院平衡语料库（Sinica Corpus）：世界上第一个带有完整词类标记的汉语平衡语料库
- 目标：500万词次汉语平衡语料库
- 设计思想：1) 遵循台湾计算语言学会的分词标准
  2) 采样时以自然段落为准，不看文章长度
  3) 语料采用多重分类法

4.6 几个典型语料库

Chinese LDC

- 国家 973 项目资助（图象、语音、自然语言理解与知识挖掘，编号：G1998030504）
- 语音，文字（口语，书面语）
- 单语：分词及词性标注语料，树库语料
- 双语：汉英句子对齐
- 规模： 汉语通用词表 ：8－10万词

  汉语信息词典：2.5-3.0 万词

  分词词性标注语料：500万字

  汉语句法树库：100万字 … …

4.6 几个典型语料库

- LC-STAR 项目 (NLPR-Nokia)
  - 14 国语言：英文、俄语、中文、西班牙语 …
  - 文本语料不少于100M words （中文约3000万字）
  - 领域：新闻 612万字，19%、财经418万字，14%、
    文化娱乐 374万字，12%、体育829万字，27%、
    消费 499万字，16%、个人通讯 355万字，12%
    共计约：3087 万字
  - 抽取常用词汇：4.5 万词
  - 另外收集专用词汇：5000词
  - 人名：5 万个
  - 词典标注：拼音、词性等

小结

- 语料库语言学的基本定义、研究内容和发展历程
- 语料库类型
- 语料库建设中的基本问题
- 典型语料库

习题

1. 思考一下，如果让你评价一个语料库，并给出定量的分值，你将如何建立评分方法？
2. 查阅或通过网页下载有关北京大学语料库和宾州大学语料库（UPenn Tree Bank）的文献资料，了解语料库的设计、加工过程。

Thanks

谢谢!