计算语言学

第3讲 语料库

刘群
中国科学院计算技术研究所
liuqun@ict.ac.cn

中国科学院研究生院2002~2003学年第二学期课程讲义

什么是语料库

- 语料库是语言材料的集合
- 语料库的特点
  - 必须是真实语言环境中出现过的语言材料
  - 必须是以电子计算机为载体
  - 必须经过一定的分析、加工和处理

语料库的类型 1

- 按来源分类
  - 口语语料库
  - 书面语语料库
- 按语言分类
  - 单语语料库
  - 双语语料库

语料库的类型 2

- 按加工方式分
  - 单语
    - 原始语料库
    - 切分标注语料库
    - 句法树库
    - 语义标注语料库
    - ……
  - 双语
    - 篇章对齐语料库
    - 句子对齐语料库
    - 词语对齐语料库
    - 结构对齐语料库
    - ……

语料库研究的历史

- 第一代（1970－80年代）
  - 百万词级
  - 以语言研究为导向
- 第二代（1980－90年代）
  - 千万词级
  - 词典编纂 － 应用导向
- 第三代（1990年代－ ）
  - 超大规模（上亿词级）
  - 标准编码体系
  - 深度标注/多语种
  - NLP应用
- 第四代（？）
  - 互联网作为语料库

第一代语料库 1

- Brown语料库
  - 始建于1960年代初
  - W.N.Francis和H.Kucera发起
  - 美国Brown大学建立
  - 世界上第一个根据系统性原则采集样本的标准语料库
  - 主要代表当代美国英语
  - 规模100万词次

第一代语料库 2

- LOB语料库
  - 始建于1970年代初
  - 由英国Lancaster大学著名语言学家Geoffrey Leech倡议
  - 挪威Oslo大学Stig Johansson主持完成
  - 安装在挪威Bergen大学挪威人文科学计算中心
  - 规模于Brown语料库相当
  - 主要代表当代英国英语

第一代语料库 3

- 1960年代初，由Randolph Quirk主持
- 收集2000小时的谈话和广播等口语素材并整理成书面材料
- 由瑞典Lund大学J. Svartvik主持全部录入计算机
- 1975年建成

第二代语料库 1

- COBUILD语料库
  - 建于1980年代
  - 以词典编撰为应用背景
  - 有英国Birminghan大学与Collins出版社合作完成
  - 规模达2000万词次
  - 基于该语料库出版的Collins Cobuild词典（1987）受到了广泛的好评

第二代语料库 2

- Longman语料库
  - 建于1980年代
  - 包括三个语料库
    - LLELC语料库（Longman/Lancaster英语语料库）
    - LSC语料库（Longman口语语料库）
    - LCLE（Longman英语学习语料库）
  - 目标是编撰英语学习词典，为外国人学习英语服务
  - 词典规模达5000万词次

第三代语料库 1

- ACL/DCI语料库
  - 美国ACL倡议发起
  - 收集语料范围广泛
    - 华尔街日报
    - Collins英语词典
    - Brown语料库
    - PennTreeBank
    - 一些双语或多语文本等
  - 既有已标注的语料，也有未标注语料
  - 制定了语料库文件的格式标注
    - 采用统一的SGML标注语言
    - 语料标注依照TEI（Text Encoding Initiative）标准

第三代语料库 2

- PennTreeBank （宾州大学树库）
  - 美国Pennsylvania大学1980年代末开始发起
  - 由该校计算机系M.Marcus主持
  - 1993年，完成了对近300万英语词的句子语法结构标注
  - 2000年完成了中文树库（第一版）：10万词次，4185个句子

语料库的收集、整理和应用

语料库三方面	属性	值
A. 语料本身	规模	百万词级 | 千万词级 | 亿万词级 | …
	领域	政治 | 经济 | 体育 | 心理学 | …
	体裁	文学 | 应用文 | 新闻 | …
	时代	共时 | 历时
	语体	书面语 | 口语
	语种	单语 | 双语 | 多语 双语平行语料库 | 双语比较语料库
	语言层次	语音（音节，韵律） | 语法（词，句，…）
B. 语料加工	数据形式	Text文本 | HTML文本 | 数据库 | …
	编码体系	TEI标准 | 自定义编码体系 | …
	加工层次	词性 | 句法 | 语义 | 语篇 | … 双语句子对齐 | 词对齐 | …
	加工方式	自动 | 人机互助 | 人工
C. 语料应用	应用领域	通用 | 词典编纂 | 机器翻译 | …
	辅助软件	检索工具 | 人机界面 | 数据接口 | …

语料的选取

- Summers, Longman/Lancaster English Corpus: Criteria and Design, Harlow: Longman
  - 精品原则
  - 有影响力原则
  - 随机挑选原则
  - 高流通度原则
  - 典型性原则
  - 易于获得原则
  - 具有统计样本意义原则
  - 符合语言规范原则
- 平衡性：主观性强

《人民日报》语料库 1

- 北京大学、富士通公司、人民日报社共同开发
- 含《人民日报》1998年上半年全部文本（约1千7百万字）
- 完整的词语切分和词性标注信息
- 高准确率

《人民日报》语料库 2

- 样例

  历史/n 将/d 铭记/v 这个/r 坐标/n ：/w 北纬/b ４１．１/m 度/q 、/w 东经/b １１４．３/m 度/q ；/w 人们/n 将/d 铭记/v 这/r 一/m 时刻/n ：/w １９９８年/t １月/t １０日/t １１时/t ５０分/t 。/w

  ……

  [中国/ns 政府/n]nt 顺利/ad 恢复/v 对/p 香港/ns 行使/v 主权/n ，/w 并/c 按照/p “/w 一国两制/j ”/w 、/w “/w 港人治港/l ”/w 、/w 高度/d 自治/v 的/u 方针/n 保持/v 香港/ns 的/u 繁荣/an 稳定/an 。/w

《人民日报》语料库 3

部分标记集

代码	名称	帮助记忆的诠释
Ag	形语素	形容词性语素。形容词代码为a，语素代码 g 前面置以A。
a	形容词	取英语形容词adjective的第1个字母。
ad	副形词	直接作状语的形容词。形容词代码a和副词代码d并在一起。
an	名形词	具有名词功能的形容词。形容词代码a和名词代码n并在一起。
b	区别词	取汉字“别”的声母。
c	连词	取英语连词conjunction的第1个字母。
Dg	副语素	副词性语素。副词代码为d，语素代码 g 前面置以D。
d	副词	取adverb的第2个字母，因其第1个字母已用于形容词。
e	叹词	取英语叹词exclamation的第1个字母。
f	方位词	取汉字“方”
g	语素	绝大多数语素都能作为合成词的“词根”，取汉字“根”的声母。
h	前接成分	取英语head的第1个字母。
i	成语	取英语成语idiom的第1个字母。
j	简称略语	取汉字“简”的声母。
k	后接成分	
l	习用语	习用语尚未成为成语，有点“临时性”，取“临”的声母。
……	……	……

London-Lund英语口语语料库 1

^what a_bout a cigar\ette# . /
((4 sylls)) /
I ^w\on't have one th/anks# - - - /
^aren't you .going to sit d/own# - /
^[/\m]# - /
^have my _coffee in p=eace# - - - /
^quite a nice .room to !s\it in ((actually))# /
^\isn't it# /
^y/\es# - - - /

转引自Tony McEnery & Andrew Wilson, 1996, Corpus Linguistics, p55,

London-Lund英语口语语料库 2

部分标记集

标记	含义
#	语调群的结束 (end of tone group)
^	语音开始 (onset)
/	上升型核心语调 (rising nuclear tone)
\	下降型核心语调 (falling nuclear tone)
^	先升后降型核心语调 (rise-fall nuclear tone)
_	平型核心语调 (level nuclear tone)
[ ]	不完整的词语和音节符号 (enclose partial words and phonetic symbols)
.	标准重音 (normal stress)
!	高音高于前一个音节的重音 (booster: higher pitch than preceding prominent syllable)
=	高音跟前一个音节相当的重音 (booster: continuance)
(( ))	不清晰的音节 (unclear)
* *	同步发音 (simultaneous speech)
-	一个重音单位的停顿 (pause of one stress unit)

Chinese PennTreeBank 1

原始数据：

他还提出一系列具体措施和政策要点。

词性标注结果：

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

Chinese PennTreeBank 2

句法树标注结果：

```
(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 。))
```


中科院计算所双语语料库

语料库的编码体系

- SGML（标准置标语言）

  http://www.w3.org/MarkUp/SGML/
- XML（可扩展的置标语言）

  http://www.w3.org/TR/REC-xml
- TEI（文档编码计划）

  http://www.tei-c.org/
- CES（语料库编码标准）

  http://www.tei-c.org/Applications/index-co02.html

冯志伟，标准通用置标语言SGML及其在自然语言处理中的应用》，载《当代语言学》1998年第4期。

语料库检索

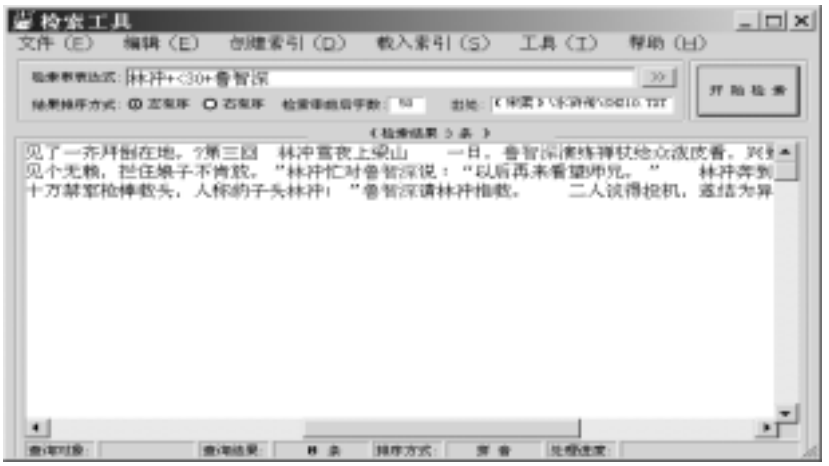
- 通常倒排表实现词语到文本的快速检索

term_table

term_ID	posting_list
Term_1	doc_1, ... , doc_i
Term_2	doc_1, ... , doc_ j
⋮	⋮
Term_n	doc_1, ... , doc_n

- 演示：语料库检索和集列（concordance）

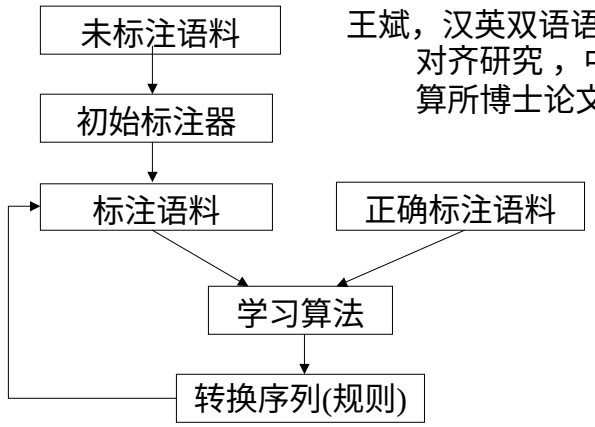
语料库检索(Demo)



英语句子边界的识别

- 困难所在：句点的多义性
- 例子：
  - The group includes Dr. J.M. Freeman and T. Boone Pickens Jr.
  - “ This issue crosses partly lines and crosses philosophical lines!” said Rep. John Rowland (R., Conn.).
  - The machine is 1.1 kilograms weight.
  - Our web site address is http://www.ict.ac.cn .
  - One of them looked sideways with his drunken eyes and said, "You...you are Minister Feng?"

英语句子边界的识别
基于转换的错误驱动的学习 1



王斌，汉英双语语料库自动对齐研究，中科院计算所博士论文，2000

英语句子边界的识别
基于转换的错误驱动的学习 2

- 激发环境

  〖左单词〗〖前缀〗〖句点〗〖后缀〗〖右单词〗

  leftword prefix . suffix rightword

- 例子

  I have never seen Mr. Wang since 1994.

  “Mr”后的句点的以上各值为：

  leftword=“seen”, prefix=“Mr”,

  suffix=NULL, rightword=“Wang”

英语句子边界的识别
基于转换的错误驱动的学习 3

- 激发环境函数（布尔值）
  - prefix：是否为空(isNull)、右端是否数字(isRdigit)、右端是否句点(isRdot)、右端是否其他标点(isRpunct)、是否英语单词(isEnglishword)、首字母是否大写(isCapitalized)
  - suffix: 是否为空、左端是否数字(isLdigit)、左端是否句点(isLdot)、左端是否其他标点(isLpunct)、是否英语单词、首字母是否大写
  - leftword：是否为空、是否英语单词、左端是否句点、左端是否其他标点符号、首字母是否大写
  - rightword: 是否为空、是否英语单词、右端是否句点、右端是否其他标点符号、首字母是否大写

英语句子边界的识别
基于转换的错误驱动的学习 4

- 转换规则模板

  如果 某句点的prefix各属性取值为布尔集合A，同时suffix各属性取值为布尔集合B，leftword各属性取值为布尔集合A，rightword各属性取值为布尔集合D

  那么 该句点将由表示结尾改变为不表示结尾（或者由不表示结尾改为表示结尾）

  其中A、B、C、D分别表示各属性的取值集合。

英语句子边界的识别
基于转换的错误驱动的学习 5

- 初始标注器
  - 任意标注，或者
  - 全部句点标注为句子边界，或者
  - 全部句点标注为非句子边界
- 规则组织形式：一个规则序列
- 标注算法：对初始标注的语料中的每一个句点，依次执行规则序列中的每一条规则，对该句点的标注进行修改

英语句子边界的识别
基于转换的错误驱动的学习 6

- 规则评价函数

$$F(T)=\begin{cases}\text{Num\_good\_transform}_T, \\ \qquad \text{当Num\_bad\_transform}_T = 0 \\ \text{Num\_good\_transform}_T/\text{Num\_bad\_transform}_T, \\ \qquad \text{当Num\_bad\_transform}_T \neq 0\end{cases}$$

英语句子边界的识别 基于转换的错误驱动的学习 7

- 学习算法的基本思想：
  - 每次循环时，在所有可能应用的规则中，寻找一条最好的规则，使其评价函数最高
- 开放测试结果（2.2M训练语料）：

	测试集1	测试集2	测试集3
文章类型	计算机文献	英文小说	杂类
句点个数	1757	814	4060
正确率	98.1%	97.2%	97.9%

英语句子边界的识别 基于转换的错误驱动的学习 8

- 学习到的规则样例

  if
  - prefix满足：isRdigit=Yes
  - suffix满足：isLdigit=Yes
  - leftword满足：isEnglishword=Yes，isRdot=No，isRpunc=No，isCapitalized=No
  - rightword满足：isEnglishword=Yes，isLdot=No，isLpunc=No，isCapitalized=No

  then
  - 该句点表示结尾→该句点不表示结尾

复习思考题

- 到http://www.icl.pku.edu.cn下载北京大学《人民日报》切分标注语料库，研究汉语“动词＋名词”可能构成哪些歧义结构？
- 阅读Chinese PennTreeBank的语料库加工规范，并翻译成中文，了解语料库的加工过程。