现代汉语词语切分研究

常宝宝
北京大学计算语言学研究所
chbb@pku.edu.cn

什么是汉语自动切分？

- 通过计算机把组成汉语文本的字串自动转换为词串的过程被称为自动切分（segmentation）。
  - 例子:
    - 鱼在长江中游
    - →鱼/在/长江/中/游
- 汉语和英语等印欧语不同，词和词之间没有空格。
  - 例子:
    - I’m going to show up at the ACL

英语中的切分问题

- 英语中不是完全没有切分问题，不能仅仅凭借空格和标点符号解决切分问题。
  1. 缩写词 如：
     N.A.T.O. i.e. m.p.h Mr. AT&T
  2. 连写形式以及所有格词尾
     I’m He’d don’t Tom’s
  3. 数字、日期、编号
     128,236 +32.56 –40.23 02/02/94 02-02-94
     D-4 T-1-A B.1.2
  4. 带连字符的词
     text-to-speech text-based e-mail co-operate

英语中的切分问题

- 英语中的切分通常被叫做Tokenization。
- 同汉语相比，英语切分问题较为容易。

为什么要进行汉语的切分研究

- 对汉语进行切分是许多应用的要求
  1. TTS或语音合成
     - 只有正确切词，才能知道正确的发音，如: 的(de0) 目的(di4)
     - 只有正确切词，才能正确变音，如: (Third Tone Sandhi) 3+3→2+3 很好 好酒 小老鼠 3+3+3 → 2+3+3 or 3+2+3
     - 只有正确切词，才能正确解决轻声的问题，如：冬瓜 桌子

为什么要进行汉语的切分研究

2. 信息检索
   - 切分有助于提高信息检索的准确率，如：
     a.和服务于三日后裁制完毕，并呈送将军府中。
     b.王府饭店的设施和服务是一流的。
3. 词语的计量分析
   - 词频统计 (汉语中最常用的词是哪个词？)
4. …

- 汉语切词也是深层汉语分析的基础
  - 句法分析、语义分析等

基本方法

- 最大匹配法(MM)
  1. 正向最大匹配法(MM)
  2. 逆向最大匹配法(RMM)

正向最大匹配法

```
S←待切分的字串;
Segmentation←"";
len ← maxlen;
WHILE S≠"" DO
        W ← substr(S,0,len);
        IF (W∈D) THEN /*D 为电子词典*/
            S ← S – W;
            Segmentation ← Segmentation + W + "/";
            len ← maxlen;
        ELSE
            IF len = 1 THEN
                S ← S – W;
                Segmentation ← Segmentation + W + "/";
                len ← maxlen;
            ELSE
                len ← len – 1;
            ENDIF
        ENDIF
END WHILE
```

逆向最大匹配法

- 正向最大匹配法 从左向右匹配词典
- 逆向最大匹配法 从右向左匹配词典
- 例子
  - 输入:企业要真正具有用工的自主权
  - MM:企业/要/真正/具有/用工/的/自主/权
  - RMM:企业/要/真正/具有/用工/的/自/主权

最大匹配法

- 长词优先
  - 输入:他将来中国
  - MM:他/将来/中国
  - RMM:他/将来/中国
  - 正确:他/将/来/中国
- 算法非常简单

自动切分的评价

- 准确率(precision)

  准确率（P）＝切分结果中正确分词数/切分结果中所有分词数*100%

- 召回率(recall)

  召回率（R）＝切分结果中正确分词数/标准答案中所有分词数*100%

- F-评价(F-measure 综合准确率和召回率的评价指标)

  F-指标＝2PR/(P+R)

关键问题

- 切分歧义（消解）
  - 一个字串有不止一种切分结果
- 未登录词识别
  - 专有名词
  - 新词

切分歧义

1. 交集型歧义
   - 字串AJB中，若AJ∈D、JB∈D、A∈D、B∈D，则AJB为交集型歧义字段。此时，AJB有AJ/B、A/JB两种切分形式。其中J为交集字段。
   - 从小学
     从小/学/电脑 从/小学/毕业
2. 组合型歧义
   - 字串AB中，若AB ∈D、 A ∈D、 B ∈D，则AB为组合型歧义字段。此时，AB有AB、A/B两种切分形式。
   - 中将
     美军/中将/竟公然说 新建地铁/中/将/禁止商业摊点

切分歧义

3. 混合型歧义
   - 同时包含交集型歧义和组合型歧义的歧义字段
   - 人才能
     这样的/人才/能/经受住考验
     这样的/人/才/能/经受住考验
     这样的/人/才能/经受住考验
- 交集型歧义、组合型歧义分布
  - 中文文本中交集型切分歧义与组合型切分歧义的出现比例约为1：22[1]

[1]刘挺、王开铸，1998，关于歧义字段切分的思考与实验。《中文信息学报》第2期，63-64页。

切分歧义

- 交集型歧义的链长
  - 交集型歧义字段中含有交集字段的个数，称为链长。
  - 从小学　链长是1
  - 结合成分 链长是2
  - 为人民工作 链长是3
  - 中国产品质量 链长是4
  - 部分居民生活水平 链长是6
  - 治理解放大道路面积水 链长是7

切分歧义

- 真实文本中交集型歧义字段分布[1]。(510万新闻语料)

链长	1	2	3	4	5	6	7	8	总计
词次数	47402	28790	1217	608	29	19	2	1	78248
比例	50.58	47.02	1.56	0.78	0.04	0.02	0.00	0.00	100
字段数	12686	10131	743	324	22	5	2	1	23914
比例	53.05	42.36	3.11	1.35	0.09	0.02	0.01	0.01	100

[1] 中文文本自动分词和标注，刘开瑛著，商务印书馆，2000，66~67

歧义的分类

1. 真歧义
   - 歧义字段在不同的语境中确实有多种切分形式
   - **地面积**
     这块/地/面积/还真不小
     地面/积/了厚厚的雪
   - **和平等**
     让我们以爱心/和/平等/来对待动物
     阿美首脑会议将讨论巴以/和平/等/问题
   - **把手**
     锌合金/把手/的相关求购信息
     别/把/手/伸进别人的口袋里

歧义的分类

2. 伪歧义

- 歧义字段单独拿出来看有歧义，但在(所有)真实语境中仅有一种切分形式可接受。
- **挨批评**
  挨/批评(√)　挨批/评(×)
  学生/挨/批评/挥拳打老师
- **平淡**
  平淡(√)　平/淡(×)
  平淡/生活感动人

歧义的分类

- 根据文献[1]，对于交集型歧义字段，真实文本中伪歧义现象远远多于真歧义现象。
  - 伪歧义 94%
  - 真歧义 6%
    - 多种切分形式均匀分布 12%
      - 应用于
        将信息技术/应用/于/教学实践
        信息技术/应/用于/教学中的哪个方面
    - 一种切分形式占优 88%
      - 解除了
        上级/解除/了/他的职务 (大多数)
        方程的/解/除了/零以外还有…

[1] 中文文本自动分词和标注，刘开瑛著，商务印书馆，2000，66~67

歧义的发现

- 歧义消解的前提是发现歧义。切分算法应该有能力检测到输入文本中何时出现了歧义切分现象。
- MM和RMM法均没有检测歧义的能力。
  - 只能给出一种切分结果。
- 最短路径法
  - 选择词数最少的切分结果
  - 没有歧义检测能力，尤其组合歧义

歧义的发现

- 双向最大匹配(MM+RMM)
  - 同时采用MM法和RMM法
  - 若果MM法和RMM法给出同样的结果，则认为没有歧义，若不同，则认为发生了歧义。
  - 输入:企业要真正具有用工的自主权
    MM:企业/要/真正/具有/用工/的/自主/权
    RMM:企业/要/真正/具有/用工/的/自/主权

歧义的发现

- 双向最大匹配法不能发现所有的歧义，存在盲点
  - 最大匹配法不能发现组合型歧义（长词优先）
    - 输入:他从马上下来
      MM、RMM:他/从/马上/下来
  - 在一定条件下（链长为偶数），双向最大匹配法也不能发现交集型歧义
    - 输入: 原子结合成分子时
      MM:原子/结合/成分/子时
      RMM:原子/结合/成分/子时

歧义的发现

- 统计数据[1]
  - 文本中90%左右的句子，MM和RMM结果相同且正确。
  - 文本中1%左右的句子，MM和RMM结果相同且不正确。
  - 文本中9%左右的句子， MM和RMM结果不相同(其中一个正确或两者均不正确)（检测到歧义）
- 双向最大匹配法使用较为广泛的原因。

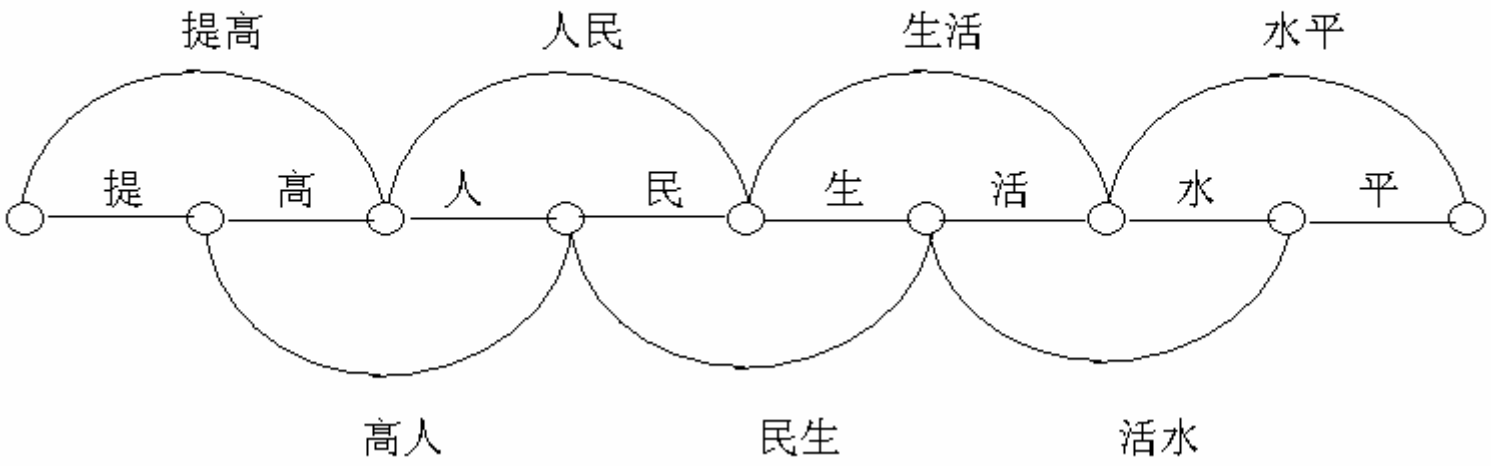
[1] Sun,M.S.and Benjamin K. T. 1995. Ambiguity resolution in Chinese word segmentation. Proceedings of the 10th Asia Conference on Language,Information and Computation, 121 -126.Hong Kong.

歧义的发现

- MM+逆向最小匹配法
- 全切分算法
  - 依据词表，给出输入文本的所有可能的切分结果
  - 效率低于MM法
  - 可以检测到所有的歧义现象
  - 输入：　提高人民生活水平
    输出：　提/高/人/民/生/活/水/平
    提高/人/民/生/活/水/平
    提高/人民/生/活/水/平
    提高/人民/生活/水/平
    提高/人民/生活/水平
    ……

数据结构

- 歧义切分的表示—词图



歧义消解

- 基于记忆的歧义消解
  - 伪歧义所占比例非常大
  - 文献[1]从一个1亿字真实汉语语料库中抽取出的前4619 个高频交集型歧义切分覆盖了该语料库中全部交集型歧义切分的59.20 %，其中4279个属伪歧义，覆盖率高达53.35%。鉴于伪歧义的消解与上下文无关，对伪歧义型高频交集型歧义切分，可以把它们的正确（唯一）切分形式预先记录在一张表中，其歧义消解通过直接查表即可实现。

[1]孙茂松、左正平等，1999， 高频最大交集型歧义切分字段在汉语自动分词中的作用。《中文信息学报》第1期，27-34页。

歧义消解

- 基于规则的歧义消解
  - P[+R+M+Q+A|Z]+”马上”→马+上
    他从大红/马/上/下来
    这件事需要/马上/办
  - “一起”+~V→一+起
    我们/一起/去故宫
    一/起/恶性交通事故

歧义消解

- 基于统计的歧义消解
  - 在词图上寻找统计意义上的最佳路径
  - 统计词表中每个词的词频，并将其转换为路径代价
    - C = - log(f/N)
  - 切分路径的代价为路径上所有词的代价之和
  - 寻求代价最小的路径

未登录词识别

- 中国人名：李素丽 老张 李四 王二麻子
- 中国地名：定福庄 白沟 三义庙 韩村河 马甸
- 翻译人名：乔治·布什 叶利钦 包法利夫人 酒井法子
- 翻译地名：阿尔卑斯山 新奥尔良 约克郡
- 机构名：方正公司 联想集团 国际卫生组织 外贸部
- 商标字号：非常可乐 乐凯 波导 杉杉 同仁堂
- 专业术语：万维网 主机板 模态逻辑 贝叶斯算法
- 缩略语：三个代表 五讲四美 打假 扫黄打非 计生办
- 新词语：卡拉OK 波波族 美刀 港刀

未登录词识别

- 未登录词识别困难
  - 未登录词没有明确边界
  - 许多未登录词的构成单元本身都可以独立成词
- 每一类未登录词都要构造专门的识别算法
- 识别依据
  - –内部构成规律（用字规律）
  - –外部环境（上下文）

未登录词识别

- 未登录词识别进展
  - 较成熟
    - –中国人名、译名
    - –中国地名
  - 较困难
    - –商标字号
    - –机构名
  - 很困难
    - –专业术语
    - –缩略语
    - –新词语

中文人名识别

- 在汉语的未登录词中，中国人名是规律性最强，也是最容易识别的一类；
  - 中国人名一般由以下部分组合而成：
    - 姓：张、王、李、刘、诸葛、西门
    - 名：李素丽，王杰、诸葛亮
    - 前缀：老王，小李
    - 后缀：王老，赵总
  - 中国人名各组成部分用字比较有规律

中文人名识别

- 根据统计, 汉语姓氏大约有1000多个(数量有限), 姓氏中使用频度最高的是“王”姓, “王, 陈, 李, 张, 刘”等5个大姓覆盖率达32%, 姓氏频度表中的前14个高频度的姓氏覆盖率为50%, 前400个姓氏覆盖率达99%。人名的用字也比较集中。频度最高的前6个字覆盖率达10.35%, 前10个字的覆盖率达14.936%, 前15个字的覆盖率达19.695%, 前400个字的覆盖率达90%

中文人名识别

- 一个识别模型
  - r1: word→name
  - r2: name→1-hanzifamily 2-hanzigiven
  - r3: name→1-hanzifamily 1-hanzigiven
  - r4: name→2-hanzifamily 2-hanzigiven
  - r5: name→2-hanzifamily 1-hanzigiven
  - r6: 1-hanzifamily→ $hanzi_i$
  - r7: 2-hanzifamily→ $hanzi_i$ $hanzi_j$
  - r8: 1-hanzigiven→ $hanzi_i$
  - r9: 2-hanzigiven→ $hanzi_i$ $hanzi_j$

中文人名识别

- 计算一个可能的人名字串的概率，若其概率大于某个阈值，则判别为人名。

$$
\begin{aligned}
&P(C_1C_2C_3)\\
&= P(r_1)\cdot P(r_2)\cdot P(r_6)\cdot P(r_9)\\
&= P(name)\cdot P(\text{1-}hanzifamily\ \text{2-}hanzigiven \mid name)\\
&\cdot P(C_1 \mid \text{1-}hanzifamily)\cdot P(C_2C_3 \mid \text{2-}hanzigiven)
\end{aligned}
$$

评测

- 国内863、973
- 国际SIGHAN

Site	word count	R	c_r	P	c_p	F	OOV	R_{oov}	R_{iv}
S01	17,194	0.962	±0.0029	0.940	±0.0036	0.951	0.069	0.724	0.979
S10	17,194	0.955	±0.0032	0.938	±0.0037	0.947	0.069	0.680	0.976
S09	17,194	0.955	±0.0032	0.938	±0.0037	0.946	0.069	0.647	0.977
S07	17,194	0.936	±0.0037	0.945	±0.0035	0.940	0.069	0.763	0.949
S04	17,194	0.936	±0.0037	0.942	±0.0036	0.939	0.069	0.675	0.955
S08	17,194	0.939	±0.0037	0.934	±0.0038	0.936	0.069	0.642	0.961
S06	17,194	0.933	±0.0038	0.916	±0.0042	0.924	0.069	0.357	0.975
S05	17,194	0.923	±0.0041	0.867	±0.0052	0.894	0.069	0.159	0.980

什么是词？

- 词是由语素构成的、能够独立运用的最小的语言单位。
- 词就是说话的时侯表示思想中一个观念的词。
- 缺乏操作标准。
- 汉语中语素、词和词组的界线模糊。
  - 象牙 是词？ 兔牙？
  - 吃饭 吃鱼
  - 毁坏 打坏

什么是词？

- 关于什么是词，不同的人有不同的把握。

	M1	M2	M3	T1	T2	T3
M1		0.77	0.69	0.71	0.69	0.70
M2			0.72	0.73	0.71	0.70
M3				0.89	0.87	0.80
T1					0.88	0.82
T2						0.78

100个句子（4372字），6个人 人工切分，两两比较

汉语分词规范

- 《信息处理用汉语分词规范》GB/T13715-92，中国标准出版社，1993
  - **分词单位**：汉语信息处理使用的、具有确定的语义或语法功能的基本单位。包括本规范的规则限定的词和词组。
  - 结合紧密、使用稳定
  - 不但有规范 还要有词表
  - 什么是切分单位和应用有关
  - 工程观点
- 《资讯处理用中文分词规范》台湾中研院，1995

阅读文献

[1] 汉语自动分词研究评述

[3] A Stochastic Finite State Word Segmentation Algorithm for Chinese