汉语自动分词

——从实践出发

杨沐昀

教育部-微软语言语音重点实验室

MOE-MS Joint Key Lab of NLP and Speech (HIT)

内容提要

- **词法分析：概念**
- **分词和分词标准**
- **中文分词：理性主义与经验主义**
- **中文分词：问题与发现**

1、词法分析

- 英文：Lexical or Morphological analysis
- Word Tokenization
  - 将字符序列转换为单词（Token）序列的过程
  - Tokens一般指单独的单词，但也可以是段落，句子
  - Look out, it's a dog. → look/ out/ ,/ it/ 's/ a/ dog/ ./
- Word Stemming（词干提取）
  - 词干提取（stemming）是抽取词的词干或词根形式（不一定能够表达完整语义）
- lemmatization（词形还原）
  - 词形还原（lemmatization），是把一个任何形式的语言词汇还原为一般形式（能表达完整语义）

1、词法分析

- 词干提取vs词形还原：分别用于IR 和 NLP
  - 在原理上：
    - 词干提取主要是采用“缩减”的方法，将词转换为词干，如将“cats”处理为“cat”，将“effective”处理为“effect”
    - 词形还原主要采用“转变”的方法，将词转变为其原形，如将“drove”和“driving”处理为“drive”
  - 在复杂性上：词干提取方法相对简单，词形还原更复杂；
    - 磁性还原需要返回词的原形，需要对词形进行分析，不仅要进行词缀的转化，还要进行词性识别，区分相同词形但原形不同的词的差别。
    - 词性标注的准确率也直接影响词形还原的准确率
  - 在实现方法上：主流方法类似，但具体实现上各有侧重
    - 词干提取的实现方法主要利用规则变化进行词缀的去除和缩减，从而达到词的简化效果。
    - 词形还原则相对较复杂，有复杂的形态变化，单纯依据规则无法很好地完成。其更依赖于词典，进行词形变化和原形的映射，生成词典中的词

1、词法分析

- 词性标注
  - 词性标注（part-of-speech tagging）,又称为词类标注或者简称标注，是指为分词结果中的每个单词标注一个正确的词性的程序，也即确定每个词是名词、动词、形容词或者其他词性的过程
  - 词性标注是很多NLP任务的预处理步骤，如句法分析，经过词性标注后的文本会带来很大的便利性，但也不是不可或缺的步骤
  - part-of-speech tagging→ part-of-speech/*noun* tagging/*verb*

序：词法分析

- 命名实体识别
  - 命名实体识别 (Named Entity Recognition, NER) 是在句子的词序列中定位并识别人名、地名、机构名等实体的任务
  - 通常包括两部分
    - 实体边界识别；
    - 确定实体类别（人名、地名、机构名或其他）
  - 实体类别主要识别三大类(实体类、时间类和数字类)和七小类(人名、地名、机构名、时间、日期、货币和百分比)命名实体
  - 对于中文来说，命名实体识别是未登录词识别的重要手段

内容提要

- **词法分析：概念**
- **分词和分词标准**
- **中文分词：理性主义与经验主义**
- **中文分词：问题与发现**

分词的提出

- 词：是自然语言中能够独立运用的最小单位，是语言信息处理的基本单位。
  - わたしはとうきょうだいがくりゅうがくせいです．
  - 남편이 한국어를 읽거나 쓸 줄 아세요?
- 分词：将句子转换成词序列
  - 中文、日文、韩语都存在这些问题

分词的意义

- 自动分词是正确的中文信息处理的基础
  - 文本检索
    - 和服 | 务 | 于三日后裁制完毕，并呈送将军府中。
    - 王府饭店的设施 | 和 | 服务 | 是一流的。

      如果不分词或者“和服务”分词有误，都会导致荒谬的检索结果。
  - 文语转换
    - 他们是来 | 查 | 金泰 | 撞人那件事的。(“查”读音为cha)
    - 行侠仗义的 | 查金泰 | 远近闻名。 (“查”读音为zha)

分词标准：什么是词

- 汉语词定义不明确
  - 牛肉是词，马肉是不是？
  - 打倒是词，打死、打伤、饿死、涂黑是不是？
- 采用“分词单位”的说法，建立词表
  - 取舍理由不够充分，人为色彩过重
  - 过于复杂，难于把握
- 为操作的方便，必须确定统一的标准或规范

分词标准

- 汉语分词规范问题的提出
  - 分词是许多技术的基础：语音识别、信息检索、机器翻译等
  - 中文词之间没有明显分界符，不同人对同一句话词的界限有不同的看法，需要一个同一的标准。
  - 863/973和SIGHAN对计算机分词结果的评价都以人工分词结果作为标准，人工结果是否科学规范？
- 现有的规范
  - 《信息处理用现代汉语分词规范，中华人民共和国国家标准（GB/T13715）》
  - 《北京大学现代汉语语料库基本加工规范，北京大学，2002》~实践影响大
  - 《现代汉语语料库文本分词规范（Ver3.0），北京语言文化大学语言信息处理研究所、清华大学计算机科学与技术系，1998.12.09》
  - 《973当代汉语文本语料库分词、词性标注加工规范，山西大学，2003》
  - 《咨询处理用中文分词规范，台湾省，1998》

分词规范

- 信息处理用现代汉语分词规范
  - 规范总体分析
    - 分词规范是用来指导分词的，国家规范定义了词，词组，分词单位
    - 词：最小的能独立运用的语言单位
    - 词组：由两个或两个以上的词，按一定的语法规则组成，表达一定意义的语言单位
    - 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
    - 国家规范规定了现代汉语中“分词单位”的确定原则，给出了一套比较系统的规则，满足了信息处理的需要，是我国分词的主要规范

汉语分词规范

- 信息处理用现代汉语分词规范
  - 不同系统应用中分词规范的定位
    - 在校对系统中将含有易错字的词和词组作为“分词单位”单位
    - 检索系统相对注重专业术语和专业名词，并且一些检索系统倾向于将分词单位较小化。所以分词单位的粒度大小必须要考虑到查全率和查准率的矛盾
    - 语音合成系统需要把多音字所组成的词和词组作为分词单位，例如“出差”、“差遣”，因为在这些词或词组中，多音字“差”的音是确定的
    - 在简繁转换系统中一些简体字的繁体形式可能有多种，因此它的简繁转换是不确定的。但是从词和词组的层面上来看，它的转换又是确定的。所以为了提高简繁转换的正确率，简繁转换系统把这些词或词组收进词表
    - 在汉字输入系统中常常把一些互现频率高的相互邻接的几个字也作为输入的单位来提高输入速度

汉语分词规范

- 信息处理用现代汉语分词规范
  - 分词标准实例
    - 二字或三字词，以及结合紧密、使用稳定的：发展　可爱　红旗　对不起　自行车　青霉素
    - 四字成语一律为分词单位：胸有成竹　欣欣向荣

      四字词或结合紧密、使用稳定的四字词组:社会主义　春夏秋冬　由此可见
    - 五字和五字以上的谚语、格言等，分开后如不违背原有组合的意义，应予切分:

      时间/就/是/生命/

      失败/是/成功/之/母

汉语分词规范

- 信息处理用现代汉语分词规范
  - 分词标准实例
    - 结合紧密、使用稳定的词组则不予切分:不管三七二十一
    - 惯用语和有转义的词或词组，在转义的语言环境下，一律为分词单位:

      妇女能顶/半边天/

      他真小气，象个/铁公鸡/
    - 略语一律为分词单位：科技 奥运会 工农业
    - 分词单位加形成儿化音的“儿”：花儿 悄悄儿 玩儿

汉语分词规范

- 信息处理用现代汉语分词规范
  - 分词标准实例
    - 阿拉伯数字等，仍保留原有形式:1234 7890
    - 现代汉语中其它语言的汉字音译外来词，不予切分：巧克力 吉普
    - 不同的语言环境中的同形异构现象，按照具体语言环境的语义进行切分：

      把/手/抬起来

      这个/把手/是木制的

基础工作之烦冗辛苦

内容提要

- **词法分析：概念**
- **分词和分词标准**
- **中文分词：理性与经验**
- **中文分词：问题与发现**

理性主义的分词方法

- 使用预先建立的词典
- 依赖人的语言观察和经验直觉设计算法
  - 长度和频率
- 从研究角度来看，启发式函数设计过于主观
  - 假设条件过强
  - 并未建立与问题本质的联系
  - 均属于贪心策略，未及考虑全局最优

观察到什么？

分词算法

- 基于字符串匹配的分词算法—理性主义
  - 正向最大匹配
  - 逆向最大匹配
  - 双向最大匹配
  - 最短路径分词法

正向最大匹配分词(Forward Maximum Matching method, FMM)

- 基本思想：将当前能够匹配的最长词输出
  - 1. 设自动分词词典中最长词条所含汉字个数为I
  - 2. 取被处理材料当前字符串序数中的I个字作为匹配字段，查找分词词典。若词典中有这样的一个I字词，则匹配成功，匹配字段作为一个词被切分出来，转6
  - 3. 如果词典中找不到这样的一个I字词，则匹配失败
  - 4. 匹配字段去掉最后一个汉字，I--
  - 5. 重复2-4，直至切分成功为止
  - 6. I重新赋初值，转2，直到切分出所有词为止

正向最大匹配分词(FMM)

输入：S1="计算语言学课程是三个课时"

W=S1

::W= 计算语言学课程是三个小时

Search in Dic for : W

If fail W= W[:-1]

::W=计算语言学课程是三个小

……..

词语
…
计算语言学
课程
课时
…

大规模真实语料中99%的词例（token）的长度都在5字以内[1]

[1] 黄昌宁、赵海，2007，中文分词十年回顾，《中文信息学报》2007年第3期，8-19页。

最大匹配法

- FMM伪代码

```
for ( int i = 0; i < N; i ++) {
    for (int j = N; j > i; j --) {
        if (IsWord(input [i..j])) {
        output(input[i..j]);
        i = j;
        }
    }
    output(input[i]);
}
```

- N：句子长度
- Isword()是查字典过程;

正向最大匹配分词(Forward Maximum Matching method, FMM)

- “市场/中国/有/企业/才能/发展/”
- 错误切分率为1 / 169
- 往往不单独使用，而是与其它方法配合使用

逆向最大匹配分词(Backward Maximum Matching method, BMM法)

- 分词过程与FMM方法相同，不过是从句子(或文章)末尾开始处理，每次匹配不成功时去掉的是最前面的一个汉字
- “市场/中/国有/企业/才能/发展/
- 实验表明：逆向最大匹配法比最大匹配法更有效，错误切分率为1／245

最大匹配法的问题

- 存在分词错误：增加知识、局部修改
- 局部修改：增加歧义词表，排歧规则
  - 三/ 个人→ 三/ 个/ 人

规则示例

IF W = "个人", W_{Left} = 数词 THEN W = "个/ 人/" ENDIF

最大匹配法的问题

- 存在分词错误 → 增加知识，局部修改
- 无法发现分词歧义 → 从单向最大匹配改为双向最大匹配

A. 正向最大匹配和逆向最大匹配结果不同

FMM 有意/ 见/ 分歧/

BMM 有/ 意见/ 分歧/

B. 正向最大匹配和逆向最大匹配结果相同

FMM & BMM 原子/ 结合/ 成分/ 子时/

双向匹配法（Bi-direction Matching method, BM法）

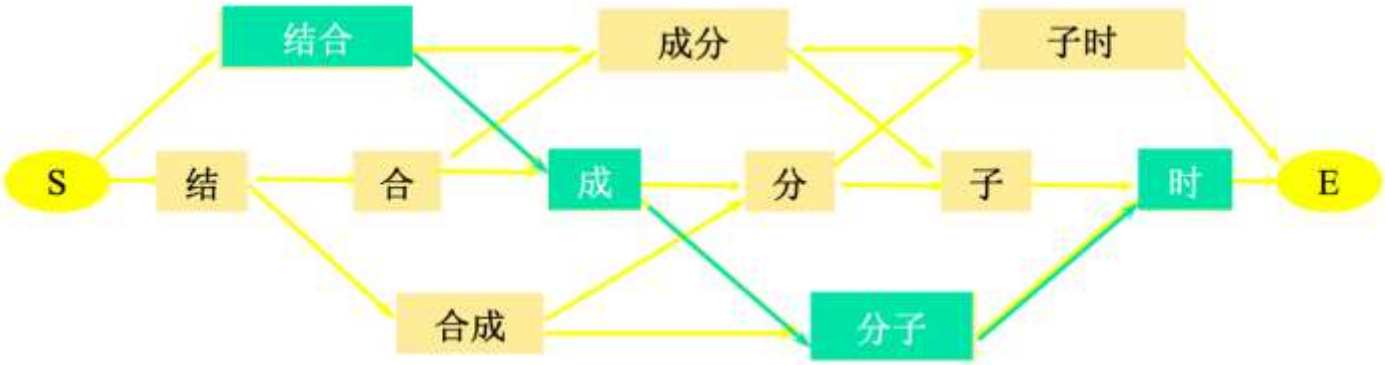
- 双向最大匹配法是将正向最大匹配法（FMM）得到的分词结果和逆向最大匹配法（BMM）得到的结果进行比较，从而决定正确的分词方法
- 据Sun M.S. 和 Benjamin K.T.（1995）的研究表明
  - 中文中90.0%左右的句子，正向最大匹配法和逆向最大匹配法完全重合且正确
  - 只有大概9.0%的句子两种切分方法得到的结果不一样，但其中必有一个是正确的（歧义检测成功）
  - 只有不到1.0%的句子，或者正向最大匹配法和逆向最大匹配法的切分虽重合却是错的，或者正向最大匹配法和逆向最大匹配法切分不同但两个都不对（歧义检测失败）。
- 这正是双向最大匹配法在实用中文信息处理系统中得以广泛使用的原因所在

最少分词法

- 分词结果中含词数最少
  - 优化代替了贪心
  - 等价于最短路径
- 算法：
  - 动态规划算法
  - 优点：好于单向的最大匹配方法
    - 最大匹配：独立自主/和平/等/互利/的/原则
    - 最短路径：独立自主/和/平等互利/的/原则
  - 缺点：忽略了所有覆盖歧义，也无法解决大部分交叉歧义
    - 结合成分子时
      - 结合|成分|子 {} 结|合成|分子 {} 结合|成|分子

最大词频分词法—经验主义的萌芽

- 基本思想：出现频率越高的词越可靠
  - 正确率可达到92%，效果一般好于基于长度信息的方法
  - 实现中再次需要：搜索技术（动态规划、有向图求最优）



词图给出了一个字符串的全部切分可能性

分词任务：寻找一条起点S到终点E的最优路径

内容提要

- **词法分析：概念**
- **分词和分词标准**
- **中文分词：理性与经验**
- **中文分词：问题与发现**

分词问题：歧义

- 交集型切分歧义
  - 汉字串AJB被称作交集型切分歧义，如果满足AJ、JB同时为词(A、J、B分别为汉字串)。此时汉字串J被称作交集串。
    - [例] “结合成分子”
      - 结合 | 成 分|子 |
      - 结合|成|分子|
      - 结 | 合成 |分子|
    - [例] “美国会通过对台售武法案”
    - [例] “乒乓球拍卖完了”

分词问题：歧义

- 组合型切分歧义
  - 汉字串AB被称作组合型切分歧义，如果满足条件：A、B、AB同时为词
    - [例]组合型切分歧义：“起身”
    - 他站 | 起 | 身 | 来。
    - 他明天 | 起身 | 去北京。

分词问题：歧义

- 交集型歧义字段中含有交集字段的个数，称为链长
  - 链长为1： 和尚未
  - 链长为2： 结合成分
  - 链长为3： 为人民工作
  - 链长为4： 中国产品质量
  - 链长为5： 鞭炮声响彻夜空
  - 链长为6： 努力学习语法规则
  - 链长为7： 中国企业主要求解决
  - 链长为8： 治理解放大道路面积水
  - ……

真实文本中分词歧义的分布情况

交集型歧义：组合型歧义 = 1: 22 语料规模：17,547字 [1]

语料规模：500万字新闻语料 [2]

链长 / 歧义字段	1	2	3	4	5	6	7	8	总计
Token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
Type种数	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] 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。

[2] 刘开瑛，2000，《中文文本自动分词和标注》，商务印书馆，65页。

真实文本中分词歧义的分布情况

- “真歧义” 和 “伪歧义”
  - 真歧义指存在两种或两种以上的可实现的切分形式，如句子 “必须/加强/企业/中/国有/资产/的/管理/”和 “中国/有/能力/解决/香港/问题/”中的字段 “中国有”是一种真歧义
  - 伪歧义一般只有一种正确的切分形式，如 “*建设/有*” 、 “*中国/人民*” 、 “*各/地方*” 、 “*本/地区*” 等

真实文本中分词歧义的分布情况

78248个[1]交集型歧义字段

- 伪歧义：94%
- 真歧义：6%
  - 多种切分均匀分布 12% (甲)
  - 一种切分切分占优 88% (乙)

(甲) 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

(乙) 上级/解除/了/他的职务
方程的/解/除了/零以外还有...

[1] 刘开瑛，2000，《中文文本自动分词和标注》，商务印书馆，66-67页。

真实文本中分词歧义的分布情况

在一个1亿字真实汉语语料库中抽取出的前4,619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，占92.63%，如"和软件"、" 充分发挥"、"情不自禁地"，这部分伪歧义类型的实例对语料的覆盖率高达53.35%。[1]

[1] 孙茂松 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》，载《中文信息学报》1999年第1期。

分词问题：未登录词

- 虽然一般的词典都能覆盖大多数的词语，但有相当一部分的词语不可能穷尽地收入系统词典中，这些词语称为未登录词
- 已知但未尽收（必要性、可行性）
  - 重叠词：“高高兴兴”、“研究研究” 一般认为已解决
  - 专有名词：时间词 、中文人名、地名、机构名称、外国译名

分词问题：新词

- 中国自改革开放的20年来平均每年产生800多个新词语
- 新词的出现，使得自动分词结果中出现过 多的“散串”，从而影响了分词的准确率
- 最近的研究还显示，60%的分词错误是由新词导 致的
- 大部分未知，无法尽收词典

未登录词（新词）种类

- 数字类复合词(numeric type compounds)，即组成成分中含有数字，包括时间、日 期、电话号码、地址、数字等，如“2005年”、“三千”
- 专有名词(Proper names)，主要包括人名、地名、机构名。如“张三”、“北京”、“微软”
- 缩略词(abbreviation)，如“中油”、“日韩”
- 派生词(derived words)，主要指含有后缀词素的词，如“电脑化”
- 复合词(compounds)，由动词或名词等组合而成，如“获允”、“搜寻法”、“电脑桌”

新词发现难点

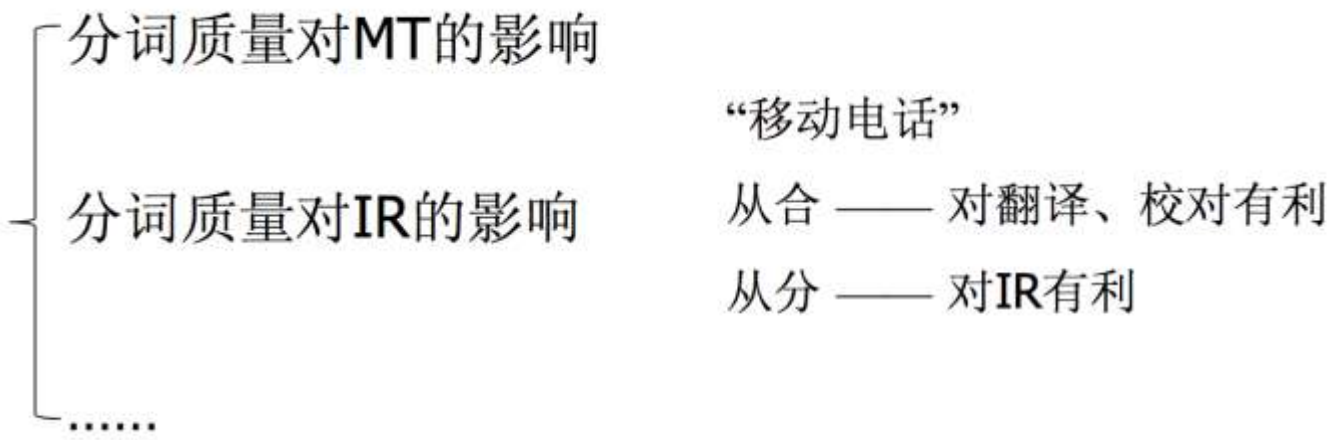
- 由于中文词语定义的模糊性，新词没有统一的定义标准，且涵盖面广，很难找到一种通用的有效的方法
- 新词尤其是非命名实体，在构成方面没有普遍的规律
- 对于低频新词由于数据稀疏，识别难度很大
- 很难根据词语的词形、词义和词语用法的变化以及利用时间信息发现新词

不同类别未登录词（新词）识别难度的差异

- 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

分词质量评价

- 计算分词正确率的不同标准
  - 以词数算
  - 以句数算
- 分词质量对NLP应用系统的影响



分词质量评价

- 准确率(precision)

$$准确率（P）= \frac{切分结果中正确分词数}{切分结果中所有分词数} *100\%$$

- 召回率(recall)

$$召回率（R）= \frac{切分结果中正确分词数}{标准答案中所有分词数} *100\%$$

- F-评价(F-measure 综合准确率和召回率的评价指标)

$$F\text{-指标} = \frac{2PR}{P+R}$$

分词质量评价

- 2003年国家863评测部分结果
- 最好成绩
  - P = 93.44 %, R = 93.69 %, F1 = 93.46 %
- 最差成绩
  - P = 91.42 %, R = 89.27 %, F1 = 90.33 %

分词质量评价

- 2005年SIGHAN 汉语分词评测结果(使用MSR语料)

评测方式	系统排名	性能指标				
		召回率	精确率	*F*1	R_{oov}	R_{iv}
封闭测试	最好	0.962	0.966	0.964	0.717	0.968
	最差	0.898	0.896	0.897	0.327	0.914
开放测试	最好	0.980	0.965	0.972	0.59	0.99
	最差	0.788	0.818	0.803	0.37	0.8

R_{oov} 表示集外词（未登录词）的召回率
R_{iv} 表示集内词（词典词）的召回率
封闭测试是指模型训练和测试只允许使用SIGHAN提供的数据

SIGHAN

- SIGHAN是国际计算语言学会（ACL）中文语言处理小组的简称，其英文全称为“Special Interest Group for Chinese Language Processing of the Association for Computational Linguistics”，又可以理解为“SIG汉“或“SIG漢“
- Bakeoff则是SIGHAN所主办的国际中文语言处理竞赛，第一届于2003年在日本札幌举行（Bakeoff 2003),第二届于2005年在韩国济州岛举行(Bakeoff 2005), 而2006年在悉尼举行的第三届（Bakeoff 2006）则在前两届的基础上加入了中文命名实体识别评测

SIGHAN

年份	论文数
2002	20
2003	31
2004	21
2005	35
2006	41
2008	33
2010	73
2012	41
2103	20
2014	35
2015	28
2017	5

思考题（每题计1分，本次作业4分）

- 1. 汉语分词存在歧义，那么对应的英语任务Tokenization是否也存在类似的问题？（3学分）
- 请找到一个英文的tokenization工具，分析期代码中如何处理这些问题？（4.5学分做此条）
- 2. 从最长匹配到最大频率分词，体现了什么工程实践中的普遍规律？
- *1. 站在工程技术高度，分词/tokenization于NLP的意义是什么（提示：方法论角度，此题可在第1讲课后再提交）
- *2. 可否证明最长匹配分词的合理性？（要求超越直觉说明和个例说明的层次，要更客观可信；此题可在第2讲课后再提交）

THANKS

参考阅读

- 刘开瑛，2000，《中文文本自动分词和标注》，商务印书馆，第1－6章
- 赵铁军，2000，《机器翻译原理》，哈尔滨工业大学出版社，第3章
- 冯志伟，2001，《计算语言学基础》，商务印书馆，第2章
- 何克抗 等，1991，《书面汉语自动分词专家系统设计原理》，载《中文信 息学报》，1991年第2期。
- 白栓虎，1995，《汉语词切分及标注一体化方法》，载陈力为、袁琦主编《计算语言学进展与应用》，清华大学出版社。
- 孙茂松 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》，载《中文信息学报》1999年第1期。
- 陈小荷，2000，《现代汉语自动分析》，北京语言文化大学出版社，第7章

部分分词工具

分词工具名称	下载地址
CoreNLP	https://nlp.stanford.edu/software/segmenter.shtml
LTP	https://www.ltp-cloud.com/
jieba	https://github.com/fxsjy/jieba
NLPIR	http://ictclas.nlpir.org/
THULAC	http://thulac.thunlp.org/
SnowNLP	https://github.com/isnowfy/snownlp