

第8章 词语切分与词性标注

宗成庆

中国科学院自动化研究所

cqzong@nlpr.ia.ac.cn

4. 课程内容





本章内容



1. 概述
2. 汉语分词要点
3. 汉语分词方法
4. 命名实体识别
5. 子词压缩
6. 词性标注
7. 习题



1. 概述

- ◆ 词是自然语言中能够独立运用的最小单位，是自然语言处理的基本单位。
- ◆ 不同的语言在词法层面需要完成不同的分析任务
 - 曲折语(如英语、德语、俄语等): 用词的形态变化表示语法关系，一个形态成分可以表示若干种不同的语法意义，词根和词干与语词的附加成分结合紧密。曲折语词法分析的任务就是词的形态分析(形态还原)(morphology analysis)。
 - 分析语(孤立语)(如汉语、越南语、苗语): 词语切分。
 - 黏着语(如日语、韩语、土耳其语等): 词语切分+形态还原。

本章主要关注汉语词语的切分、子词压缩和词性标注方法。

1. 概述

- ◆ 词性或称词类(Part-of-Speech, POS)是词汇最重要的特性，是语言中词的**语法分类**，具有相同句法功能、能够出现在同样的组合位置中的词，聚合在一起所形成的范畴。词类连接词汇到句法的桥梁。

如在汉语中，词类分为两大类：实词(content words)和虚词(functional words)，实词包括体词、谓词，体词又包括名词、代词等，谓词包括动词、形容词等。

词性标注的任务是让系统自动对词汇标注词性标记。



1. 概述

◆ 汉语自动分词和词性标注的重要性

- 词语切分是句子结构分析的基础
- 词语的分析具有广泛的应用，如词频统计，词典编纂，文章风格研究，文献处理，文本校对，简繁体转换等
- 即使在数据驱动的自然语言处理中，包括统计学习方法和神经网络方法，通常情况下基于词（具有较好的切分准确率）建立的模型性能优于以字或子词建立的模型
- 词性是反映句法结构信息的重要特征
- 词性在众多NLP任务中（如文本分类、情感分类、自动文摘等）具有重要作用



1. 概述

◆尝试：对下面的文字进行词语切分，并标注词性

克拉伦斯 威姆斯（Clarence Weems）和张宁相继上篮得手，
两队比分交替上升。暂停回来，张宁和费尔德连中2记三分，
88-88平！

克拉伦斯/nrg /x 威姆斯/nrf (/wkz Clarence/nrg Weems /nrf)
/wky 和/c 张/nrf 宁/nrg 相继/d 上篮/v 得手/v ， /wd 两/m 队/q
比分/n 交替/d 上升/v 。 /wj 暂停/v 回来/v ， /wd 张/nrf 宁/nrg
和/c 费尔德/nrg 连/d 中/v 2/m 记/v 三/m 分/q ， /wd 88/m -/wp
88/m 平/a ！ /wt

*参考北京大学计算语言学研究所制定的标注规范。



本章内容

1. 概述

 2. 汉语分词要点

3. 汉语分词方法

4. 命名实体识别

5. 子词压缩

6. 词性标注

7. 习题



2. 汉语分词要点

◆ 汉语自动分词中的主要问题

- 汉语分词规范问题（《信息处理用限定汉语分词规范（GB13715）》）

— 汉语中什么是词？两个不清的界限：

(1) 单字词与词素，如：新华社25日讯

(2) 词与短语，如：花草，湖边，房顶，鸭蛋，小鸟，担水，一层，翻过？



2. 汉语分词要点

● 歧义切分字段处理

(1) 交集型歧义

中国人为了实现自己的梦想

中国/ 人为/ 了/ 实现/ 自己/ 的/ 梦想

中国人/ 为了/ 实现/ 自己/ 的/ 梦想

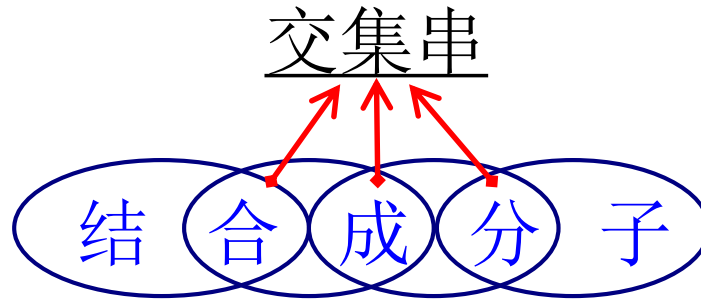
中/ 国人/ 为了/ 实现/ 自己/ 的/ 梦想

例如：“大学生”、“研究生物”、“从小学起”、“为人民工作”、“中国产品质量”、“部分居民生活水平”等等

➤ **定义：链长** 一个交集型切分歧义所拥有的交集串的集合称为交集串链，它的个数称为链长。

2. 汉语分词要点

例如：



“结合”、“合成”、“成分”和“分子”均构成词，交集串的集合为 {合，成，分}，因此，链长为3。

类似地，“为人民工作”中的公共交集字为：{人，民，工}，歧义字段的链长为 3；“中国产品质量”中的交集字为：{国，产，品，质}，歧义字段的链长为 4；“部分居民生活水平”中的交集字为：{分，居，民，生，活，水}，链长为 6。



2. 汉语分词要点

(2) 组合型歧义

门把手弄坏了。

门/ 把/ 手/ 弄/ 坏/ 了/ 。

门/ 把手/ 弄/ 坏/ 了/ 。

例如，“将来”、“现在”、“才能”、“学生会”等，都是组合型歧义字段。



2. 汉语分词要点

梁南元 曾于1987年对一个含有48,092字的自然科学、社会科学领域的样本进行了统计，结果交集型切分歧义有**518**个，多义组合型切分歧义有**42**个。据此推断，中文文本中切分歧义的出现频度约为**1.2次/100字**，交集型切分歧义与多义组合型切分歧义的出现比例约为**12:1**。



2. 汉语分词要点

● 未登录词的识别

(1)人名、地名、组织机构名等命名实体，例如：

盛中国，张建国，李爱国，蔡国庆；

高升，高山，夏天，温馨，温泉，武夷山，时光，程序；

彭太发生，朱李月华；赛福鼎·艾则孜，爱新觉罗·溥仪；

平川三太郎，约翰·斯特朗

(2)新出现的词汇、术语、个别俗语等，例如：

博客，非典，禽流感，恶搞，微信，给力，内卷，新冠

(3)与新冠肺炎有关的新词

阿尔法、贝塔、德尔塔、奥米克戎、单采、混采、

核酸异样、阳性感染者、初筛阳性感染者、方舱、

绿码、黄码、加码、密接、次密接、二类密接...

封城、静默、静默7+3...，弹窗、吹哨人 ...

2. 汉语分词要点

我们的统计结果：

错误类型			错误数	比例(%)			例子
集外词	命名实体	人名	31	25.83	55.0	98.33	约翰·斯坦贝克
		地名	11	9.17			米苏拉塔
		组织机构名	10	8.33			泰党
		时间和数字	14	11.67			37万兆
	专业术语		4	3.33		脱氧核糖核酸	
	普通生词		48	40.00		致病原	
	切分歧义		2	1.67			歌名为
合计			120	100			

从互联网上随机摘取了418个句子，共含11,739个词，19,777个汉字（平均每个句长约为28个词，每个词约含1.68个汉字）。



2. 汉语分词要点

◆ 汉语自动分词的基本原则

- **合并原则1**：语义上无法由组合成分直接相加而得到的字串应该合并为一个分词单位。

例如：不管三七二十一(成语)，或多或少(副词片语)，十三点(定量结构)，六月(定名结构)，谈谈(重叠结构，表示尝试)，辛辛苦苦(重叠结构，加强程度)，进出口(合并结构)

- **合并原则2**：语类无法由组合成分直接得到的字串应该合并为一个分词单位。

(a)字串的语法功能不符合组合规律，如：好吃，好喝，好听，好看等

(b)字串的内部结构不符合语法规律，如：游水等



2. 汉语分词要点

◆ 汉语自动分词的辅助原则

操作性原则，富于弹性，不是绝对的。

- 切分原则1：有明显分隔符标记的应该切分之。

分隔标记指标点符号或一个词。如：

上、下课 → 上/ 下课

洗了个澡 → 洗/ 了/ 个/ 澡



2. 汉语分词要点

- 切分原则2: 结构复杂、合并起来过于冗长的词尽量切分。
 - (1) 词组带接尾词，如：太空/ 计划/ 室、塑料/ 制品/ 业
 - (2) 动词带双音节结果补语，如：看/ 清楚、讨论/ 完毕
 - (3) 复杂结构：自来水/公司；中文/分词/规范/研究/计划
 - (4) 正反问句：喜欢/ 不/ 喜欢、参加/ 不/ 参加
 - (5) 动宾结构、述补结构的动词带词缀时。如：写信/ 给；
取出/ 给；穿衣/ 去
 - (6) 词组或句子的专名，多见于书面语，戏剧名、 歌曲名。
如：鲸鱼/的/生/与/死；那/一/年/我们/都/很/酷
 - (7) 专名带普通名词。如：胡/ 先生、京沪/ 铁路



2. 汉语分词要点

- **合并原则1**：附着性语(词)素与前后词合并为一个单位。

例如：“吝”是一个附着语素，“不吝”、“吝于”等合并成一个词；“员”：检查员、邮递员、技术员等；“化”：现代化、合理化、多变化、年轻化、民营化等。

- **合并原则2**：使用频率高或共现率高的字串尽量合并。

如：“进出”、“收放”（**动词并列**）；“大笑”、“改称”（**动词偏正**）；“关门”、“洗衣”、“卸货”（**动宾结构**）；“春夏秋冬”、“轻重缓急”、“男女”（**并列结构**）；“象牙”（**名词偏正**）；“暂不”、“毫不”、“不再”、“早已”（**副词并列**）等。



2. 汉语分词要点

- **合并原则4：**双音节加单音节的偏正式名词尽量合并。

如：“线、权、车、点”等所构成的偏正式名词：“国际线、分数线、贫困线”、“领导权、发言权、知情权”、“垃圾车、交通车、午餐车”、“立足点、共同点、着眼点”等。

- **合并原则5：**双音节结构的偏正式动词应尽量合并。

这条原则只适合于少数偏正式的动词，如：“紧追其后”、“组建完成”等，不适合动宾及主谓式复合动词。

2. 汉语分词要点

◆ 分词结果测试

- 封闭测试 vs. 开放测试
- 专项测试 vs. 总体测试



- ✓ 歧义字段切分能力
- ✓ 集外词(生词)处理能力
- ✓ 人名、地名、组织机构名等命名实体识别能力等



2. 汉语分词要点

◆评价指标

- 正确率(Correct ratio/Precision, P): 测试结果中正确切分或标注的个数占系统所有输出结果的比例。假设系统输出 N 个, 其中, 正确的结果为 n 个, 那么,

$$P = \frac{n}{N} \times 100\%$$

- 召回率(找回率) (Recall ratio, R): 测试结果中正确结果的个数占标准答案总数的比例。假设系统输出 N 个结果, 其中, 正确的结果为 n 个, 而标准答案的个数为 M 个, 那么,

$$R = \frac{n}{M} \times 100\%$$

两种标记: R_{OOV} 指集外词的召回率;
 R_{IV} 指集内词的召回率。



2. 汉语分词要点

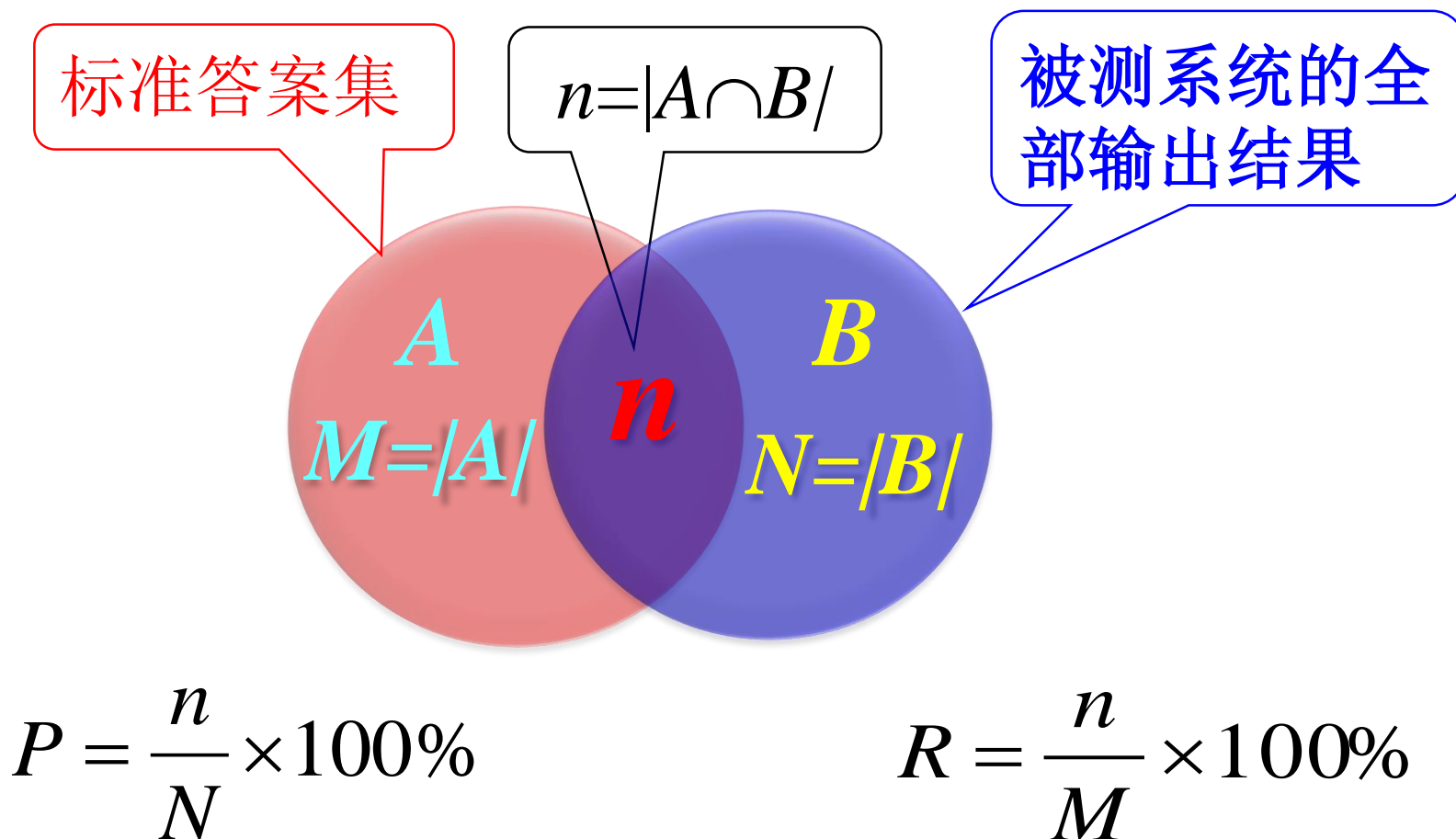
- F-测度值(F-Measure): 正确率与找回率的综合值。计算公式为:

$$F - measure = \frac{(\beta^2 + 1) \times P \times R}{\beta^2 \times P + R} \quad (8-1)$$

一般地, 取 $\beta=1$, 即:

$$F1 = \frac{2 \times P \times R}{P + R} \quad (8-2)$$

2. 汉语分词要点





2. 汉语分词要点

假设某个汉语分词系统在一测试集上输出 5260 个分词结果，而标准答案是 4510 个词语，根据这个答案，系统切分出来的结果中有 4120 个是正确的。那么：

$$P = \frac{4120}{5260} \times 100\% = 78.33\%$$

$$R = \frac{4120}{4510} \times 100\% = 91.35\%$$

$$\begin{aligned} F1 &= \frac{2 \times P \times R}{P + R} \\ &= \frac{2 \times 78.33 \times 91.35}{78.33 + 91.35} \\ &= 84.34\% \end{aligned}$$



本章内容

1. 概述
2. 汉语分词要点
- ➡ 3. 汉语分词方法
4. 命名实体识别
5. 子词压缩
6. 词性标注
7. 习题



3. 汉语分词方法

- ◆ 有词典切分 vs. 无词典切分
- ◆ 基于规则的方法 vs. 基于统计的方法



3. 汉语分词方法

① 最大匹配法(Maximum Matching, MM)

是一种有词典的切分方法，也称机械切分方法。

按照切分方向分为：

- 正向最大匹配算法 (Forward MM, FMM)
- 逆向最大匹配算法 (Backward MM, BMM)
- 双向最大匹配算法 (Bi-directional MM)

● 基本思路：

给定字符串 $S = c_1 c_2 \dots c_n$ ，某一词 $w_i = c_1 c_2 \dots c_m$ ， m 为词典中最长词的字数。假设 $m = 7$ 。

3. 汉语分词方法

输入字符串：他是研究生物化学的一位科学家。

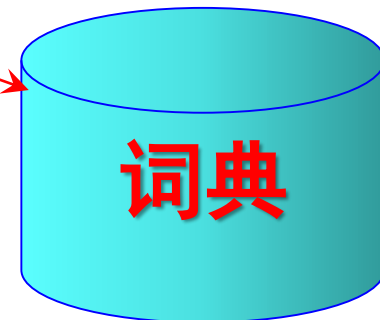
切分过程：

↑ — 7 — → |

| — 6 — → |

| — 5 — → |

...



他/ 是研究生物化学的一位科学家。

↑ — 7 — → |

...

FMM 切分结果：他/ 是/ 研究生/ 物化/ 学/ 的/ 一/ 位 / 科学家/ 。

BMM 切分：他是研究生物化学的一位科学家。

... ← 7 — ↑

BMM 切分结果：他/ 是/ 研究/ 生物/ 化学/ 的/ 一/ 位/ 科学家/ 。



3. 汉语分词方法

● FMM 算法描述

- (1) 令 $i=0$, 当前指针 p_i 指向输入字符串初始位置, 执行以下操作:
- (2) 计算当前指针 p_i 到字符串末端的字数 n , 如果 $n=1$, 转(4), 结束算法。
否则, 令 m =词典中最长单词的字数, 如果 $n<m$, 令 $m=n$;
- (3) 从当前 p_i 起取 m 个汉字作为词 w_i , 判断:
 - (a) 如果 w_i 是词典中的词, 则在 w_i 后添加一个切分标志, 转(c);
 - (b) 如果 w_i 不是词典中的词且 w_i 的长度大于1, 将 w_i 从右端去掉一个字, 转(a)步; 否则(w_i 的长度等于1), 则在 w_i 后添加一个切分标志, 将 w_i 作为单字词添加到词典中, 执行 (c)步;
 - (c) 根据 w_i 的长度修改指针 p_i 的位置, 如果 p_i 指向字符串末端, 转(4), 否则, $i=i+1$, 返回 (2);
- (4) 输出切分结果, 结束分词程序。



3. 汉语分词方法

● 方法评价

➤ 优点：

- 程序简单易行，开发周期短；
- 仅需要很少的语言资源（词表），不需要任何词法、句法、语义资源。

➤ 弱点：

- 歧义消解的能力差；
- 切分正确率不高，一般在95%左右。



3. 汉语分词方法

② 基于语言模型的分词方法

无词典切分

● 基本思路

设对于待切分的句子 S , $W = w_1w_2\ldots w_k$ ($1 \leq k \leq n$) 是一种可能的切分。

$$W^* = \arg \max_W p(W | S)$$

$$= \arg \max_W p(W) \times p(S | W)$$

语言模型

生成模型

详见第4章举例



3. 汉语分词方法

● 方法评价

➤ 优点:

- 在训练语料规模足够大和覆盖领域足够多时，可以获得较高的切分正确率。

➤ 弱点:

- 模型性能较多地依赖于训练语料的规模和质量，训练语料的规模和覆盖领域不好把握；
- 计算量较大。



3. 汉语分词方法

③ 由字构词的分词方法(Character-based tagging)

(或称“基于字标注/词位的分词方法”)

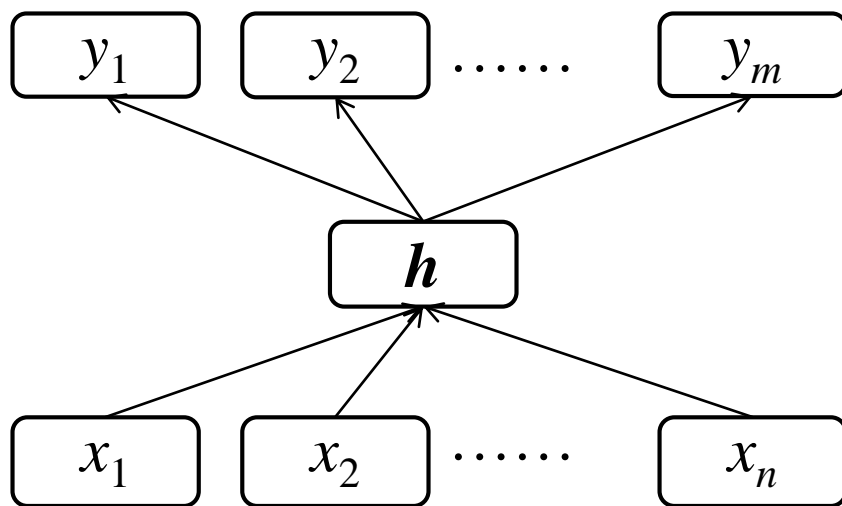
- 基本思想：将分词过程看作是字的分类问题。该方法认为，每个字在构造一个特定的词语时都占据着一个确定的构词位置(即词位)。假定每个字只有4个词位：词首(B)、词中(M)、词尾(E)和单独成词(S)，那么，每个字归属一特定的词位。

基于条件随机场(CRFs)的序列标注方法。详见第3章。

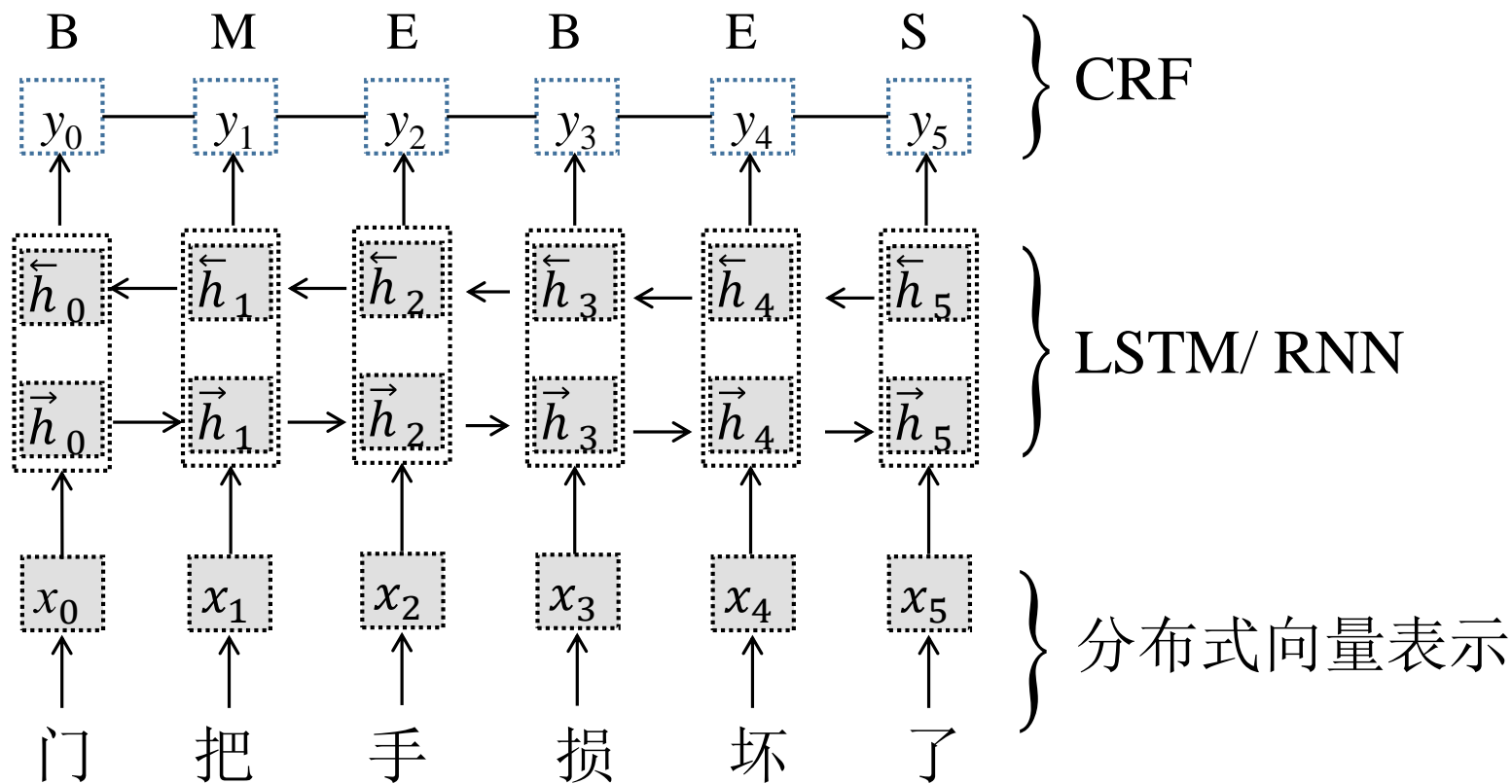
3. 汉语分词方法

④ 基于神经网络的分词方法

把分词看作序列标注任务，输入输出均为序列， $n:m$ 的对应关系。

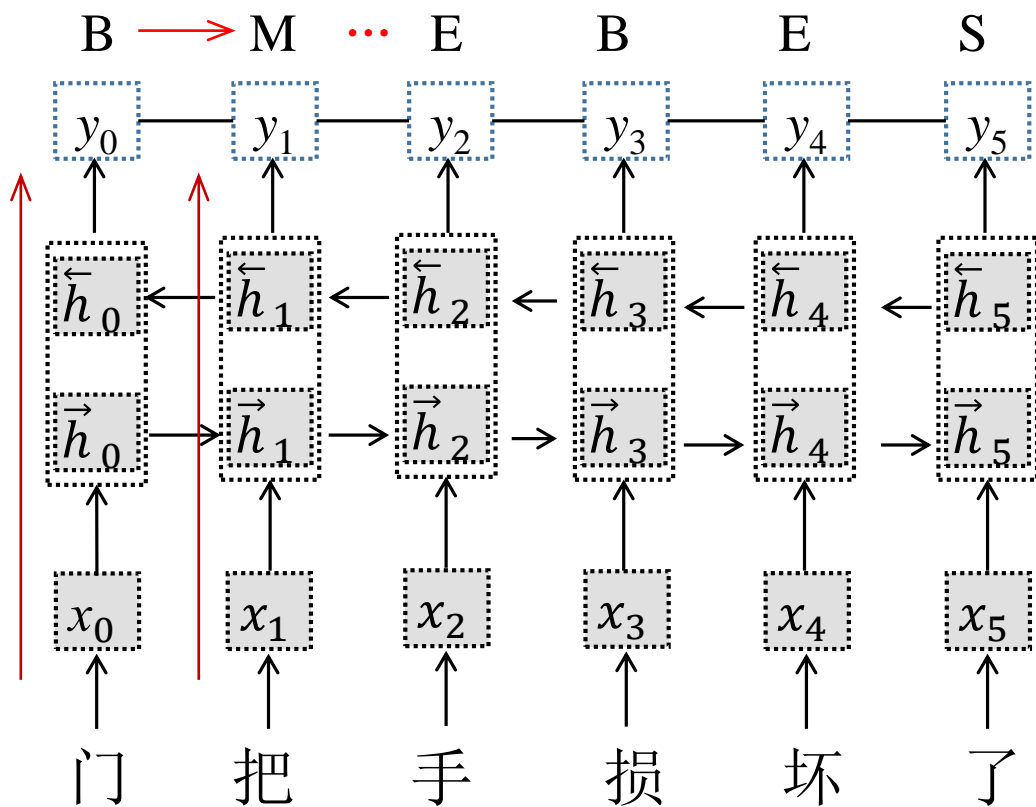


3. 汉语分词方法



切分结果：门把手/ 损坏/ 了

3. 汉语分词方法



切分结果：门把手/ 损坏/ 了

$$Y \in \{B, M, E, S\}, \mathbf{h}$$

$$p(Y | X, \lambda) = \frac{1}{Z(X)} \exp\left(\sum_j \lambda_j \cdot F_j(Y, X)\right)$$

$$Z(X) = \sum_Y \exp\left(\sum_j \lambda_j \cdot F_j(Y, X)\right)$$

- ① 4个标签分别用向量表示；
- ② 纵向由 \mathbf{h} 预测标签；
- ③ 横向获得标签转移得分；
- ④ 上述两个得分相加后用 Softmax 归一化，确定预测标记。



3. 汉语分词方法

● 实验结论:

- RNN与CRF相比，CRF取词的窗口作为输入，特征只在窗口范围内选取，而神经网络可以学习长距离关系，但是RNN难以训练，存在梯度消失/爆炸现象；
- 在序列标注任务上，RNN(LSTM)优于CNN；
- LSTM无需使用外部词表资源，效果依然很好；可同时应用到多种语言，多种序列标注任务上；但是，LSTM变种结构多、参数多、调参过程困难。



3. 汉语分词方法

◆ 最新的改进工作:

- Tzu Hsuan Chou et al. Advancing Multi-Criteria Chinese Word Segmentation Through Criterion Classification and Denoising. *Proc. ACL-2023*, pp. 6460–6476
- Dedong Li et al. CWSeg: An Efficient and General Approach to Chinese Word Segmentation. *Proc. ACL-2023 Industry Track*, pp. 1-10
- Rian He et al. Weighted self Distillation for Chinese word segmentation. *Findings of ACL-2022*, pp. 1757–1770
- Yuanhe Tian et al. Improving Chinese Word Segmentation with Wordhood Memory Networks. *Proc. of ACL-2020*, pp. 8274-8285



3. 汉语分词方法

◆ 可用的分词工具:

- (1)FastHan: <https://github.com/fastnlp/fastHan> (BERT+CRF)
- (2)WMSeg: <https://github.com/SVAIGBA/WMSeg> (ZEN + CRF)
- (3)Urheen (自动化所): <http://www.nlpr.ia.ac.cn/cip/software.htm> (n -gram + ME)
- (4)Jieba (Andy Sun): <https://github.com/fxsjy/jieba>
<https://pypi.org/project/jieba/>
- (5)HanLP(何晗): <https://github.com/hankcs/HanLP>
<https://www.hanlp.com/>
- (6)THULAC(清华): <https://github.com/thunlp/THULAC-Python>



本章内容

1. 概述
2. 汉语分词要点
3. 汉语分词方法
- ➡ 4. 命名实体识别
5. 子词压缩
6. 词性标注
7. 习题



4. 命名实体识别

◆ 命名实体(Named Entity, NE)

- 通常指：人名、地名、组织机构名、数字、日期、货币和数量。
- 在特定领域，如医学领域，有时也包含专业术语，如疾病名称、药物名称、化学成分等。

命名实体识别(named entity recognition, NER)被简称为NER任务。



4. 命名实体识别

◆关于汉语人名

- 台湾出版的《中国姓氏集》收集姓氏 5544 个，其中，单姓 3410 个，复姓 1990 个，3字姓 144 个。
- 中国目前仍使用的姓氏共 737 个，其中单姓 729个，复姓 8 个。
- 根据我们收集的 300 万个人名统计，姓氏有974个，其中，单姓 952个，复姓 23 个，300万人名中出现汉字4064个。

[曹文洁，2002]

- ✧名字用字范围广，分布松散，规律不很明显，没有标记。
- ✧姓氏和名字都可以单独使用用于特指某一人。
- ✧许多姓氏用字和名字用字(词)可以作为普通用字或词被使用。



4. 命名实体识别

◆关于汉语地方名

- 《中华人民共和国地名录》(1994)收集88026个，不包括相当一部分街道、胡同、村庄等小地方的名称。
- 真实语料中地名出现情况复杂。如地名简称、地名用词与其他普通词冲突、地名是其他专用名词的一部分，地名长度不一等。



4. 命名实体识别

◆NER与汉语分词的关系

- 在汉语分词的基础上以词为单位使用规则、统计、神经网络等各种方法
- 以汉字为单位直接使用序列标注方法

◆NER方法

- 基于规则的识别方法
- 统计学习方法(n -gram/ CRFs等)
- CRFs + 神经网络
- 利用规则方法进行识别后校正（可选步骤）



4. 命名实体识别

推荐参阅:

- [1] Y. Chen et al. A Joint Model to Simultaneously Identify and Align Bilingual Named Entities. *Computational Linguistics*, 39(2): 229-266
- [2] Y. Chen et al. On Jointly Recognizing and Aligning Bilingual Named Entities. *Proc. ACL'2010*, pp. 631-639
- [3] C. Dong et al. December. Character-Based LSTM-CRF with Radical-Level Features for Chinese Named Entity Recognition. *Proc. NLPCC'2016*, pp. 239-250. [Radical-LSTM:在 LSTM-CRF结构基础上, 对中文汉字做偏旁部首级别的 LSTM 变换。]



4. 命名实体识别

◆分词与NER存在的主要问题

- 过于依赖训练样本，而标注大规模训练样本费时费力，且仅局限于个别领域，由此导致分词和NER系统对新词的识别能力差，往往在与训练样本差异较大的测试集上性能大幅度下降。
- 现有的训练样本主要在新闻领域，而实际应用千差万别：网络新闻、微博/ 微信/ QQ 等非规范文本、不同的专业领域(中医药、生物、化学、能源)。

领域差异和生词识别是分词和NER面临的最大挑战



4. 命名实体识别

● 举例1

李时珍（约1518～1593），字东璧，晚号濒湖山人，蕲州（今湖北蕲春）人。世业医，父言闻，有医名。幼习儒，三次应乡试不中。自嘉靖三十一年（1552年）至万历六年（1578年），历时二十七载，三易其稿，著成《本草纲目》五十二卷，初刊于金陵。

公开的分词系统切分准确率为：57.3%～94.8%

4. 命名实体识别

● 举例2

类别	类别描述
事件报道	特定事件/具体事件
新闻内容	新闻消息/格式较规范
观点传播	观点词汇多/日常闲谈/观点评论
信息共享	分享的信息或者链接/为他人提供的建议
私人会话	帖子开头有“@某人”/日常闲谈
交易信息	帖子中出现金钱、比例词汇

根据对2011年微博内容的统计，大约75%的内容为个人心情和感受方面的。



4. 命名实体识别

补充词汇：

包括各种符号、表情符、数字串等

词典来源	词语数量
维基百科+常用在线词典(普通词汇)	1301320
以下5项经合并筛选后形成的网络用语词典	541941
(1)微博用语词库	10330
(2)网络用语大全	294
(3)网络关键词以及词频数据	500000
(4)《人民日报》微博词频统计	42315
(5)百度百科对于网络用语的解释	1051
网络情感词典+传统情感词典（情感词汇）	26207
经过合并筛选后的词汇总数：175万	



4. 命名实体识别

分词性能:

分词方法	准确率(%)	召回率(%)	F1值(%)
Stanford	80.40	76.52	78.41
Urheen	80.46	77.43	78.92
ICTCLAS(+微博处理)	82.62	83.52	83.07
CWS	80.12	73.24	76.52
CWS(+词典+符号处理)	90.52	90.73	90.62

CWS: Chinese word segmentation based on ME model



本章内容

1. 概述
2. 汉语分词要点
3. 汉语分词方法
4. 命名实体识别
5. 子词压缩
6. 词性标注
7. 习题



5. 子词压缩

句子表示和生成的基本粒度：单词和字符/字。

◆ 以单词为基本粒度的缺点：

- “长尾”分布，低频词的代表较差
- 词表的单词较多，计算复杂性高

◆ 以字符为基本粒度的缺点：

- 字符的歧义性较高
- 字符的序列长度较大

◆ 寻找一种介于单词和字符之间的粒度： 子词(sub-word)

综合单词和字符粒度的优势，使其在句子表示和生成中最好地发挥作用。

单词1	↑ 单词
单词2	
...	
单词29999	↓
<unk>	

字符1	↑ 字符/字
字符2	
...	
字符2999	↓
<unk>	



5. 子词压缩

● 基本思路

- 对于英语等屈折语文本，可直接用双字节编码算法(Pair Encoding, BPE) 算法进行字符压缩。
- 对于汉语文本，如果有很好的分词工具，先对文本进行词语切分，在切分结果的基础上利用BPE算法进行单字压缩，合并那些最大次数的相邻汉字或字符。

5. 子词压缩

● BPE算法

- ① 对邻近的两个字符(汉字)合并，统计被合并的两个邻近字符(汉字)在整个文本中出现的次数 α ；
- ② 将 α 最大的两个邻近字符(汉字)用原文本中不存在的符号替换(压缩)，重复进行上面的操作。直到没有被合并的字符(汉字)为止，或者达到限定合并的次数。

例1: aaabdaaababc China

① a a a b d a a a b a b c \rightarrow ② XabdXababc \rightarrow ③ XYdXYYYc

aa 出现4次
ab 出现3次
其他出现1次

X=aa

ab出现次数
最多: 3次
Y=ab

XY出现次数最多:
2次。Z=XY

④ ZdZYc

aaab@@ d@@ aaab@@ ab@@ c \rightarrow 还原, 标记

5. 子词压缩

例2:

这 / 本 / 童 话 书 / 已 / 翻 译 / 成 / 中 文 / 了
请 / 用 / 中 文 / 复 述 / 这 / 篇 / 故 事
中 文 / 的 / “ / 危 机 / ” / 分 / 为 / 两 / 个 / 字
春 / 因 / 繁 花 / 而 / 美 丽 /
繁 花 似 锦 / 的 / 深 圳 / 洋 溢 / 着 / 欢 乐 / 的 / 气 氛



第一次迭代，合并“中文”（3次）

这 / 本 / 童 话 书 / 已 / 翻 译 / 成 / 中文 / 了
请 / 用 / 中文 / 复 述 / 这 / 篇 / 故 事
中文 / 的 / “ / 危 机 / ” / 分 / 为 / 两 / 个 / 字
春 / 因 / 繁 花 / 而 / 美 丽 /
繁 花 似 锦 / 的 / 深 圳 / 洋 溢 / 着 / 欢 乐 / 的 / 气 氛



第二次迭代，合并“繁花”（2次）

这 / 本 / 童 话 书 / 已 / 翻 译 / 成 / 中文 / 了
请 / 用 / 中文 / 复 述 / 这 / 篇 / 故 事
中文 / 的 / “ / 危 机 / ” / 分 / 为 / 两 / 个 / 字
春 / 因 / 繁花 / 而 / 美 丽 /
繁花 似 锦 / 的 / 深 圳 / 洋 溢 / 着 / 欢 乐 / 的 / 气 氛

“/” 为分词标记

循环修改过程，直到：1)达到最大迭代次数；或者2)双字符的最大出现次数为1（约定数）。



5. 子词压缩

在机器翻译中，WMT14 的训练语料为450万英德双语对照的平行句对，采用子词压缩合并之后，抽取出的词表为3.2万个子词（源语言端和目标语言大约都是这个数目）。

参考文献：


Rico Sennrich, Barry Haddow, and Alexandra Birch, 2016. Neural Machine Translation of Rare Words with Subword Units. In Proceedings of ACL 2016, pages 1715–1725.

开源代码：

<https://github.com/rsennrich/subword-nmt>



本章内容

1. 概述
2. 汉语分词要点
3. 汉语分词方法
4. 命名实体识别
5. 子词压缩
-  6. 词性标注
7. 习题



6. 词性标注

◆ 面临的问题

词性(part-of-speech, POS)标注(tagging)的主要任务是消除词性兼类歧义。在任何一种自然语言中，词性兼类问题都普遍存在。例如：

(1) Time flies like an arrow.

(2) I want you to web our annual report.

对 Brown 语料库的统计，55%词次兼类。根据《现代汉语八百词》，兼类词占 22.5%。

6. 词性标注

- 在汉语中

(1) 形同音不同，如：“好(hao3, 形容词)、好(hao4, 动词)”。

例句：这个人什么都**好**，就是**好**酗酒。

(2) 同形、同音，但意义毫不相干，如：“会(会议，名词)、会(能够、动词)”。例句：每次他都**会**在**会**上制造点新闻。

(3) 具有典型意义的兼类词，如：“典型(名词或形容词)”、“教育(名词或动词)”。例句：用那种方式**教育**孩子，简直是对**教育**事业的嘲笑。

(4) 上述情况的组合，如：“行(xing2, 动词/形容词; hang2, 名词/量词)”。例句：每当他走过那**行**白杨树时，他都感觉好像每一棵树都在向他**行**注目礼。



6. 词性标注

◆ 标注集的确定原则

不同语言中，词性划分基本上已经约定俗成。

自然语言处理中对词性标记要求相对细致。

● 一般原则：

- 标准性：普遍使用和认可的分类标准和符号集；
- 兼容性：与已有资源标记尽量一致，或可转换；
- 可扩展性：扩充或修改。

6. 词性标注

- UPenn Treebank 的词性标注集

➤ **33类**: NN 名词、NR 专业名词、NT 时间名词、VA 可做谓语的形容词、VC “是”、VE “有” 作为主要动词、VV 其他动词、AD 副词、M 量词，等等。

- 北大计算语言研究所的词性标注集

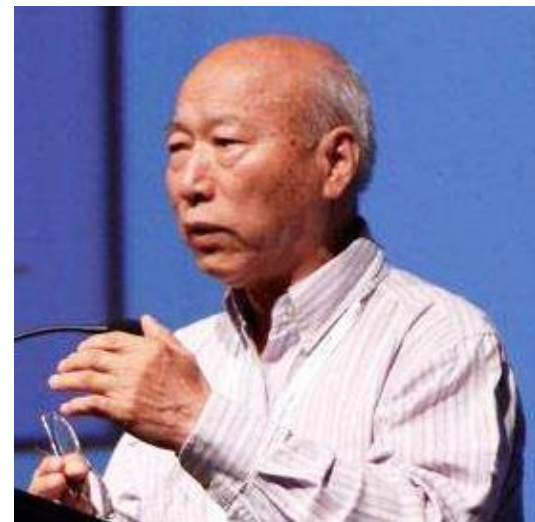
➤ **26个基本词类代码，74个扩充代码，标记集中共有106个代码**: 名词(n)、时间词(t)、处所词(s)、方位词(f)、数词(m)、量词(q)、区别词(b)、代词(r)、动词(v)、形容词(a)、状态词(z)、副词(d)、介词(p)、连词(c)、助词(u)、语气词(y)、叹词(e)、拟声词(o)、成语(i)、习用语(l)、简称(j)、前接成分(h)、后接成分(k)、语素(g)、非语素字(x)、标点符号(w)。

6. 词性标注

● 综合性语言知识库CLKB

北京大学计算语言学研究所成立于1986年。著名语言学家、前北京大学副校长朱德熙先生担任研究所第一任所长。计算机系俞士汶教授长期担任研究所常务副所长，中文系陆俭明教授任副所长。

该研究所长期致力于基础资源的研究与建设，在计算语言学基础理论、NLP模型和方法以及应用技术研发等方面取得了一批优秀成果，在国内外具有重要影响。综合型语言知识库于2011年荣获国家科技进步奖二等奖。



俞士汶

(1938.12.8 – 2021.11.4)



6. 词性标注

◆ 标注方法

- 基于规则/有限状态机的词性标注方法
- 基于统计模型的词性标注方法
 - HMM: 分词与词性标注一体化方法
 - CRFs: 序列标注方法
- 规则和统计方法相结合的词性标注方法

◆ 性能评价指标: 准确率



本章小结

◆汉语分词要点

- 汉语分词中的主要问题
- 两种歧义
- 切分原则：基本原则和辅助原则
- 性能评价方法

◆分词方法

- MM、最少分词法、统计法、由字构词法（CRFs）等

◆命名实体识别：人名、地名、组织机构名识别

◆子词压缩

◆词性标注



本章内容

1. 概述
2. 汉语分词要点
3. 汉语分词方法
4. 命名实体识别
5. 子词压缩
6. 词性标注

7. 习题



7. 习题

1. 阅读《信息处理用现代汉语分词规范》(中华人民共和国国家标准 GB13715), 了解规范的基本内容。
2. 利用已经学习过的理论方法和北京大学标注的《人民日报》分词和词性标注语料, 设计实现至少两种不同的汉语词语自动切分方法, 进行性能测试和分析。然后利用不同类型的网络文本测试你的分词系统, 对比分析分词方法和不同测试样本的性能变化。
3. 在上一题目得到的分词结果的基础上, 实现子词压缩。
4. 设计实现一个人名识别系统(针对中英文均可)。
5. 设计实现一个组织机构名识别系统(针对中英文均可)。
6. 设计实现一个基于CRFs模型的汉语词性标注系统。

第2题和第3题做作业。

谢谢!

Thanks!

