

基于HMM的中文名实体识别

主要内容

- ❖ 隐马尔可夫模型
- ❖ 规则与统计(HMM)相结合的音乐实体识别

规则与统计(HMM)相结合的音乐实体识别

- ❖ 1 问题描述
- ❖ 2 总体结构
- ❖ 3 训练语料标注
- ❖ 4 基于规则的音乐实体识别
- ❖ 5 基于统计(HMM)的音乐实体识别
- ❖ 6 音乐实体修正过程
- ❖ 7 实验结果

(张学清, 规则与统计相结合的音乐领域命名实体识别, 电子科技大学硕士论文, 2010.5)

1 问题描述

- ❖ 1) 定义
- ❖ 2) 实例
- ❖ 3) 歌手名和音乐组合名识别难点
- ❖ 4) 歌曲名和专辑名识别难点

1) 定义

- ❖ 歌手名（**Singer**）：是人名中特殊的一类，指特定歌手的固有名称、别名、英文名和译名，如：“刘德华”、“**Andy Lau**”、“周渝民”、“仔仔”、“鲍勃·迪伦”等。
- ❖ 音乐组合名（**Musical Band**）：指特定乐队、乐团和歌唱组合的固有名称，如：“小虎队”、“苏打绿”、“**Beyond**”等。
- ❖ 歌曲名（**Song**）：指特定歌曲的固有名称，如：“梦醒时分”，“月亮代表我的心”。
- ❖ 专辑名（**Album**）：指**特定专辑**的固有名称，如：“永远的邓丽君”、“旷世情歌全纪录”等。

2) 两个实例

近日，在EQ唱片歌手<Singer>肖飞</Singer>的个人博客中看到，由网络红人<Singer>小沈阳</Singer>演唱的歌曲《<Song>为什么呢</Song>》，实质上演唱者是<Singer>肖飞</Singer>。

<Band>至上励合< / Band>2009年全新首张专辑《<Album>齐天大盛</Album>》正式发行，于12月13日在北京星光现场举办了媒体发布会。去年，<Band>至上励合< / Band>凭借一首《<Song>棉花糖</Song>》成为乐坛当红励志偶像团体。

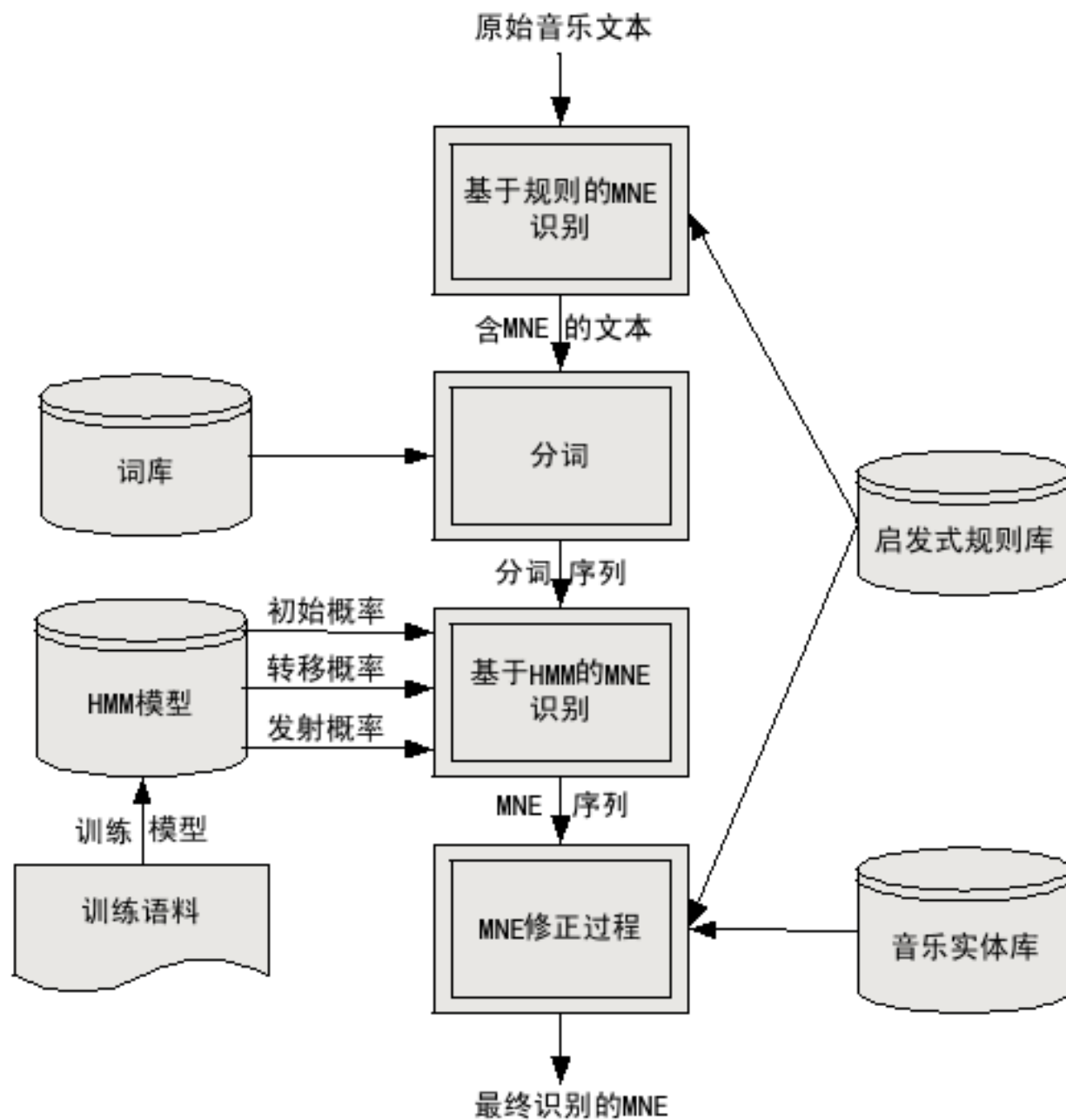
3) 歌手名和音乐组合名识别难点

- ❖ a) 歌手名和音乐组合名与上下文成词或自身成词的现象严重，容易产生歧义
 - 成词
 - ❧ 其首部与上文成词，如为何润东，做梦之旅的专访；
 - ❧ 其尾部与下文成词，如郑中基于、至上励合作为嘉宾等；
 - ❧ 其内部成词，如高峰、动力火车等。
- ❖ b) 歌手名和音乐组合名的长短不一，而且构成形式多样。歌手名的常见形式如下：
 - ❧ 歌手名的常见形式如下：
 - ❖ 姓和名组成，如张学友、任贤齐等；
 - ❖ 前缀和姓组成，如阿杜（杜成义）、小柯（柯肇雷）等；
 - ❖ 名和后缀组成，如华仔（刘德华）、周董（周杰伦）等；
 - ❖ 别名或艺名，如老狼、刀郎等。
 - ❧ 音乐组合名，用词比较随意，长度从一到十几个不等，如羽泉、南拳妈妈等，其中“南拳”和“妈妈”都是极其常见的词。
- ❖ c) 区分歌手名和普通人名比较困难

4) 歌曲名和专辑名识别难点

- ❖ a) 歌曲名和专辑名的组成方式非常随意。
 - ❧ 许多歌曲名和专辑名都是很常见的词、短语或句子，如周华健的《朋友》、满文军的《懂你》、宋祖英《长大后我就成了你》等。
 - ❧ 近来流行歌曲名中还出现了标点符号，如张韶涵的《亲爱的，那不是爱情》，这给区分歌曲和普通句子带来了很大难度。
- ❖ b) 歌曲名、专辑名中含有歌手名，如安又琪的《你好，周杰伦》。这很可能会导致歌曲名、专辑名的漏选，以及歌手名的错选。
- ❖ c) 歌曲名和专辑名的长度极其不固定。这将导致歌曲名和专辑名的边界很难确定。
- ❖ d) 很多情况下，一首歌曲和一个专辑使用相同的名称，如齐秦的专辑《又见溜溜的她》，其主打曲为《又见溜溜的她》。这将导致很难辨别一个实体到底是歌曲名还是专辑名。

2 总体结构



3 训练语料标注

- ❖ (1) 手工标注：在UltraEdit-32软件中定义和录制4个宏，分别用来标注不同的音乐实体在文本中的位置。
- ❖ (2) 利用改进后的分词程序对已标识出音乐实体的文本进行分词。
- ❖ (3) 使用机器标注的方法生成最终的训练语料。生成的训练语料除了用于HMM模型的训练外，也用于基于规则的MNE识别过程。

各宏的功能描述表

宏的名称	宏功能描述	快捷键
歌手名	用标签对 (Singer_nrb, Singer_nre) 标注歌手名	F3
组合名	用标签对 (Band_nzb, Band_nze) 标注组合名	F4
歌曲名	用标签对 (Song_nqb, Song_nqe) 标注歌曲名	F5
专辑名	用标签对 (Album_njb, Album_nje) 标注专辑名	F6

步骤1标注示例

在推出最新大碟《以你为荣》之后，古巨基于4月23日在台北举行演唱会。他将演唱20余首经典曲目，有《爱得太迟》、《白玫瑰》等。

在推出最新大碟《Album_njb 以你为荣 Album_nje》之后，Singer_nrb 古巨基 Singer_nre 于4月23日在台北举行演唱会。它将演唱20余首经典曲目，有《Song_nqb 爱得太迟 Song_nqe》、《Song_nqb 白玫瑰 Song_nqe》等。

步骤2分词标注示例

在/p 推出/v 最新/a 大/a 碟/ng 《/w Album_njb/nx
以/p 你/r 为/v 荣/ag Album_nje/nx 》 /w 之后/f , /w
Singer_nrb/nx 古/a 巨/ag 基 Singer_nre 于/p 4月/t
23日/t 在/p 台北/ns 举行/v 演唱会/n 。 /w 他/r 将/d
演唱/v 20/m 余/m 首/q 经典/n 曲目/n , /w 有/v 《/w
Song_nqb/nx 爱/v 得/u 太/d 迟/a Song_nqe/nx 》 /w
、 /w 《/w Song_nqb/nx 白/a 玫瑰/n Song_nqe/nx
》 /w 等/u 。 /w

训练语料中的实体标注符号表

实体类型	实体举例	实体标注符号
简单歌手名	[刘德华]	(nhb,nhe)
与上文成词的歌手名	为 <u>[何润东]</u>	(nib,nie)
与下文成词的歌手名	[张学友] <u>好帅</u>	(njb,nje)
简单组合名	[动力火车]	(nob,noe)
与上文成词的组合名	做 <u>[梦之旅]</u> 的专访	(npb,npe)
与下文成词的组合名	[飞轮海] <u>边唱</u>	(nqb,nqe)
歌曲名	[月亮代表我的心]	(nxb,nxe)
专辑名	[罗生门]	(nyb,nye)

步骤3标注后的结果示例

在/p 推出/v 最新/a 大/a 碟/ng 《/w [/nyb 以/p 你/r
为/v 荣/ag]/nye 》 /w 之后/f , /w [/njb 古/a 巨/ag
基于/p]/nje 4月/t 23日/t 在/p 台北/ns 举行/v
演唱会/n 。 /w 他/r 将/d 演唱/v 20/m 余/m 首/q
经典/n 曲目/n , /w 有/v 《/w [/nxb 爱/v 得/u 太/d
迟/a]/nxe 》 /w 、 /w 《/w [/nyb 白/a 玫瑰/n]/nye
》 /w 等/u 。 /w

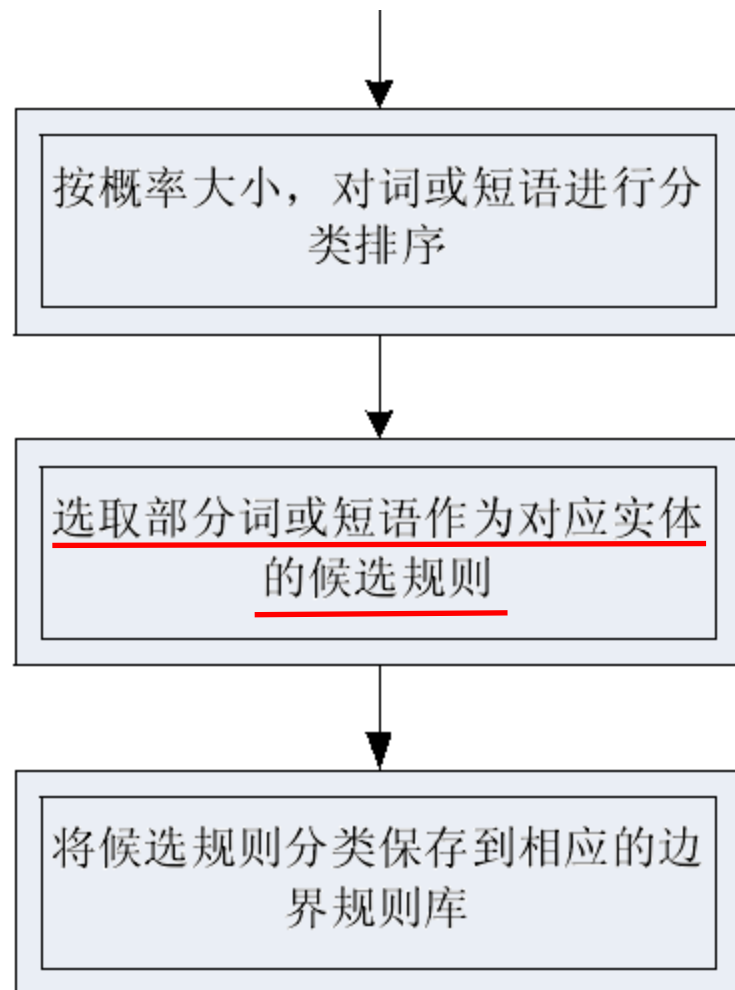
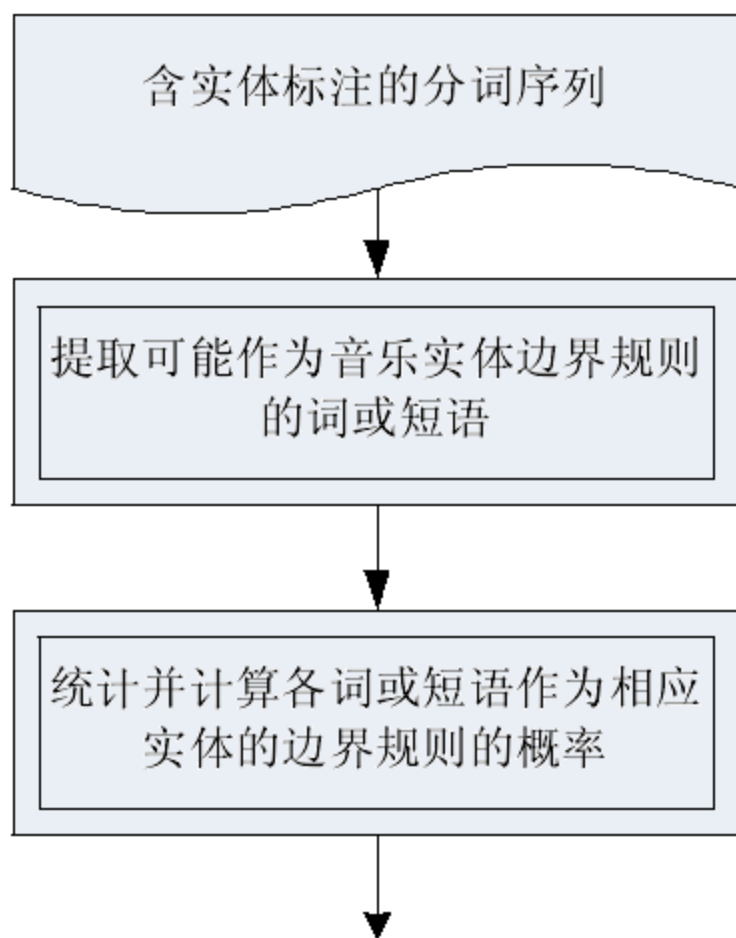
4 基于规则的音乐实体识别

- ❖ 1) 规则库构建
- ❖ 2) 规则自动提取过程
- ❖ 3) 基于规则的歌手名识别算法
- ❖ 4) 识别结果

1) 规则库构建

规则库类型	缩写	组成部分
<u>歌手名左边界规则库</u>	SIN_LBI	<u>经常出现在歌手名左边的词或短语</u>
歌手名右边界规则库	SIN_RBI	经常出现在歌手名右边的词或短语
组合名左边界规则库	BAN_LBI	经常出现在组合名左边的词或短语
组合名右边界规则库	BAN_RBI	经常出现在组合名右边的词或短语
歌曲名左边界规则库	SON_LBI	经常出现在歌曲名左边的词或短语
歌曲名右边界规则库	SON_RBI	经常出现在歌曲名右边的词或短语
专辑名左边界规则库	ALB_LBI	经常出现在专辑名左边的词或短语
专辑名右边界规则库	ALB_RBI	经常出现在专辑名右边的词或短语

2) 规则自动提取过程



3) 基于规则的歌手名识别算法

- ❖ (1) 读入未分词的音乐文本;
- ❖ (2) 通过匹配的方式, 从文本中提取出所有位于歌手名左边界和歌手名右边界之间的文本作为候选歌手名。
- ❖ (3) 判断候选歌手名是否符合歌手名的要求, 如长度等。若不符合, 则丢弃。剩下的就是识别出来的歌手名。

4) 识别结果

- ❖ 假设歌手名左边界规则库中包含“主唱”、“鼓手”、“吉他手”，歌手名右边界规则库中包含“三人”，则利用上述算法，可以从下图的文本中识别出“唐平”、“王澜”和“李剑”三个歌手名。

[/nob 艳/a 乐队/n]/noe 由/p 主/ag 唱/v [/nhb 唐/nr 平/nr]/nhe
、/w 鼓手/n [/nhb 王/nr 澜/g]/nhe 和/c 吉他/n 手/n [/nhb 李/nr
剑/nr]/nhe 三/m 人/n 组成/v ， /w 自/p 2002年/t 成立/v 以来/f
共/d 发行/v 了/y 《/w [/nyb 艳/a]/nye 》 /w 和/c 《/w [/nyb
惊/v /m 艳/a]/nye 》 /w 两/m 张/q 专辑/n 。 /w

5 基于HMM的音乐实体识别

- ❖ 1) 隐马尔科夫模型的定义
- ❖ 2) 训练隐马尔科夫模型
- ❖ 3) 解码算法
- ❖ 4) 根据状态组合模板查找音乐实体

1) 隐马尔科夫模型的定义

- ❖ 采用二元(Bigram)隐马尔科夫模型来识别各种音乐命名实体
- ❖ HMM模型的输入文本是含有音乐实体的分词序列
- ❖ HMM模型的观察值应该包括词和SIN、BAN、SON、ALB 标识串，则定义模型的观察值集合 $O = \{W_i, SIN, BAN, SON, ALB\}, 1 \leq i \leq m$ ，其中， W_i 表示词库中的一个词， m 为词库中词的数目，SIN表示已识别的歌手名，BAN表示已识别的组合名，SON表示已识别的歌曲名，ALB表示已识别的专辑名。若用 M 表示所有可能观察值的数目，则 $M = m + 4$

隐马尔科夫模型的定义

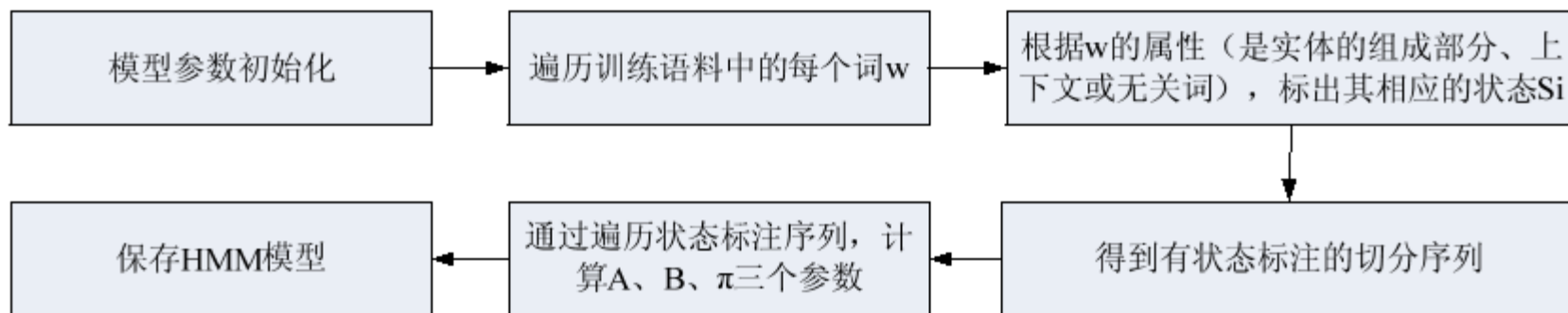
- ❖ HMM模型的状态集合是那些能够反映命名实体组成的属性集合，即每一个状态都能代表某一类命名实体的内部组成成分、上下文或无关成分
- ❖ 本文的 HMM 模型的状态集合 S 包含 $N=29$ 个状态

那么，HMM 模型的状态转移矩阵 $A = (a_{ij})_{29 \times 29}$, $a_{ij} = P(s_j | s_i)$ 表示从状态 s_i 转移到状态 s_j 的概率；发射矩阵 $B = (b_{ij})_{29 \times M}$, $b_{ij} = P(o_j | s_i)$ 表示状态 s_i 对应观察值 o_j 的概率；初始状态概率为 $\pi = \{\pi_1, \pi_2, \dots, \pi_{29}\}$, $\pi_i = P(s_i)$ 表示初始状态是 s_i 的概率。

HMM模型的状态集合

状态 s_i	含义	例子	状态 s_i	含义	例子
RB	歌手名的首部词	[郭]富城	QE	歌曲名的尾部词	月亮代表我的[心]
RI	歌手名的中部词	郭[富]城	QS	歌曲名自身成词	[烟火]
RE	歌手名的尾部词	郭富[城]	JB	专辑名的首部词	[星光]依旧灿烂
RS	歌手名自身成词	[高峰]	JI	专辑名的中部词	星光[依旧]灿烂
RU	上文与歌手名的首部成词	[为何]润东	JE	专辑名的尾部词	星光依旧[灿烂]
RV	歌手名的尾部与下文成词	张学[友好]帅	JS	专辑名自身成词	[寓言]
RN	被嵌套的歌手名、组合名	你好 [周][杰][伦]	KA	歌手名和组合名的上文	艺人
ZB	组合名的首部词	[至上]励合	LA	歌手名和组合名的下文	演唱
ZI	组合名的中部词	至上[励]合	KB	歌曲名的上文	插曲
ZE	组合名的尾部词	至上励[合]	LB	歌曲名的下文	这首
ZS	组合名自身成词	[零点]	KC	专辑名的上文	发行
ZU	上文与组合名的首部成词	[做梦]之旅的专访	LC	专辑名的下文	收录
ZV	组合名的尾部与下文成词	飞轮[海边]唱	ML	音乐实体间联结词	和、与
QB	歌曲名的首部词	[月亮]代表我的心	WL	与实体组成及上下文无关的词	
QI	歌曲名的中部词	月亮[代表]我[的]心			

2) 训练隐马尔科夫模型



❖ 自动训练过程主要包括两个部分：

❧ 状态标注过程：就是将训练语料中包含实体标注的切分序列的词性标注转换为状态标注

❧ 模型参数计算过程：就是通过遍历带有状态标注的切分序列，计算出初始状态概率 π 、状态转移概率矩阵 **A** 和发射概率矩阵 **B** 的取值。

输入的切分序列是根据“规则”来进行实体识别的

状态标注过程

n v a这些

- ❖ 输入：已经有音乐实体标注且带有词性标注的切分序列
- ❖ 输出：具有状态标注的切分序列
- ❖ 结果示例：

以“词”为切分单位

在/WL 推出/WL最新/WL 大/WL 碟/WL 《/KC 以/JB 你/JI 为/JI 荣/JE
》/LC 之后/WL ， /KA 古/RB 巨/RI 基于/RV 4月/WL 23日/WL 在/WL
台北/WL 举行/WL 演唱会/WL 。 /WL 他/WL 将/WL 演唱/WL 20/WL
余/WL 首/WL 经典/WL 曲目/WL ， /WL 有/WL 《/KB 爱/QB 得/QI
太/QI 迟/QE 》 /LB 、 /WL 《/KB 白/QB 玫瑰/QE 》 /LB 等/WL 。 /WL

具体步骤

- ❖ 1. 从已分词且有音乐实体标注的训练语料中依次读入切分序列;
- ❖ 2. 根据音乐实体标注 **nhb**、**nob**、**nxb**、**nyb**, 定位简单歌手名、简单组合名、歌曲名和专辑名。
 - ∞ a) 将简单歌手名和简单组合名前后的词标注为 **KA**、**LA**, 歌曲名前后的词标注为 **KB**、**LB**、专辑名前后的词标注为 **KC**、**LC**;
 - ∞ b) 判断 **nhb** 与 **nhe** 之间的词串长度 (即词的个数)。若大于等于 3, 将首词标注为 **RB**、中间的词都标注为 **RI**、尾词标注为 **RE**; 若为 2, 各词分别被标注为 **RB**、**RE**; 若为 1, 该词被标注为 **RS**;
 - ∞ c) 判断 **nob** 与 **noe** 之间的词串长度。若大于等于 3, 将首词标注为 **ZB**、中间的词都标注为 **ZI**、尾词标注为 **ZE**; 若为 2, 各词分别被标注为 **ZB**、**ZE**; 若为 1, 该词被标注为 **ZS**;
 - ∞ d) 判断 **nxb** 与 **nxe** 之间的词串是否包含歌手名、组合名、**SIN** 和 **BAN**。若包含, 则将这些词都标为 **RN**。然后, 判断 **nxb** 与 **nxe** 之间的词串的长度。若为 1 时, 且未被标注为 **RN**, 则标为 **QS**; 若大于 1, 且首词和尾词未被标注为 **RN**, 则将其分别标注为 **QB**, **QE**, 并将剩下的未被标为 **RN** 的词都标注为 **QI**; RN 优先级高于 QB、QI、QE、QS 这些
 - ∞ e) 采用与 d) 步相同的方法, 对 **nyb** 与 **nye** 之间的词串进行状态标注, 只是使用的状态不同, 将 **QS**、**QB**、**QI**、**QE** 分别替换为 **JS**、**JB**、**JI**、**JE**;

具体步骤

- ❖ 3. 根据音乐实体标注nib、njb、npb、nqb分别定位出与上文成词歌手名、与下文成词歌手名、与上文成词组合名、与下文成词组合名。
 - ❧ a) 将nib与nie之间的词串的首词标注为RU；词串的尾词标注为RE，所有未标注的词都标注为RI；词串的后一个词标注为LA；
 - ❧ b) 将njb与nje之间的词串的尾词标注为RV；词串的首词标注为RB，所有未标注的词都标注为RI；词串前面的一个词标注为KA；
 - ❧ c) 采用与a)步相同的方法，对npb与npe之间的词串进行状态标注；采用与b)步相同的方法，对nqb与nqe之间的词串进行状态标注。只是使用的状态不同，将RU、RV、RB、RE、RI分别替换为ZU、ZV、ZB、ZE、ZI；
- ❖ 4. 将音乐实体之间的联结词标注为ML，并将其余未标注的词标注为WL。

模型参数计算过程

- ❖ 利用最大似然估计计算 **A**, **B**, **π** 三个概率

$$P(o_j | s_i) \approx c(o_j, s_i) / c(s_i), 1 \leq i \leq 29, 1 \leq j \leq M$$

$$P(s_i | s_{i-1}) \approx c(s_{i-1}, s_i) / c(s_{i-1}), 2 \leq i \leq 29$$

$$P(s_i) \approx c(s_i) / \sum_{j=1}^{29} c(s_j), 1 \leq i \leq 29$$

- ❖ 采用加法平滑对状态转移概率进行平滑

$$P(o_j | s_i) = \begin{cases} P_{ML}(o_j | s_i), c(s_i) \geq 5 \\ \alpha \cdot P_{GT}(o_j | s_i), 1 \leq c(s_i) < 5 \\ \beta \cdot P(o_j), c(s_i) = 0 \end{cases}$$

3) 解码算法

- ❖ 假定分词序列 $O = (o_1, o_2, \dots, o_m)$ ，音乐实体识别问题就相当于求解：

$$S^* = \arg \max_S P(S | O)$$

- ❖ 针对二元 HMM 模型：

$$S^* = \arg \max_S \prod_{i=1}^m P(o_i | s_i) P(s_i | s_{i-1})$$

- ❖ 取对数：

$$S^* = \arg \min_S \left[- \sum_{i=1}^m (\ln P(o_i | s_i) + \ln P(s_i | s_{i-1})) \right]$$

4) 根据状态组合模板查找音乐实体

歌手名: $RB+RI^*+RE$, $RI+RE$, $RE+RE$, $RU+RI^*+RE$, $RB+RI^*+RV$, RS ;

组合名: $ZB+ZI^*+ZE$, $ZI^{\wedge}+ZE$, $ZE^{\wedge}+ZE$, $ZU+ZI^*+ZE$, $ZB+ZI^*+ZV$, ZS ;

歌曲名: $QB+QI^*+RN^*+QE$, $QB+QI^*+RN^{\wedge}$, $RN^{\wedge}+QI^*+QE$, RN^{\wedge} , $QI^{\wedge}+QE$, $QE^{\wedge}+QE$, QS ;

专辑名: $JB+JI^*+RN^*+JE$, $JB+JI^*+RN^{\wedge}$, $RN^{\wedge}+JI^*+JE$, $JI^{\wedge}+JE$, $JE^{\wedge}+JE$, JS ;

其中, 星号表示 0 次或多次, 尖号表示至少出现一次。当出现状态为“RU”或“ZU”的词时, 取其中最后一个字作为音乐实体的首部。当出现状态为“RV”或“ZV”的词时, 取其中第一个字作为音乐实体的尾部。

我们来看一个例子, 假设已经获得最优状态序列: “老牌/WL 歌手/KA 林/RB 忆/RI 莲/RE 首/LA 唱/WL 新/WL 歌/ WL 《/KB 柿子/QS 》/LB”, 通过匹配模板“ $RB+RI+RE$ ”和“ QS ”, 可以识别出歌手名“林忆莲”, 歌曲“柿子”。

6 音乐实体修正

❖ 错误识别的原因

- ❧ 歌手名和普通人名的区分比较困难，可能会将普通人名误识别为歌手名
- ❧ 歌手名和音乐组合名具有相似的上下文，歌曲名和专辑名常相同，可能会导致实体类型错误

❖ 修正方案

- ❧ 针对歌手名和音乐组合名：由于其数量比较少且容易收集，所以采用通过音乐实体库（包含歌手名库、组合名库）来过滤各种错误实体的方法。
- ❧ 对于歌曲名和专辑名：由于其数量非常多，利用实体上下文的启发式规则来更正实体类型。

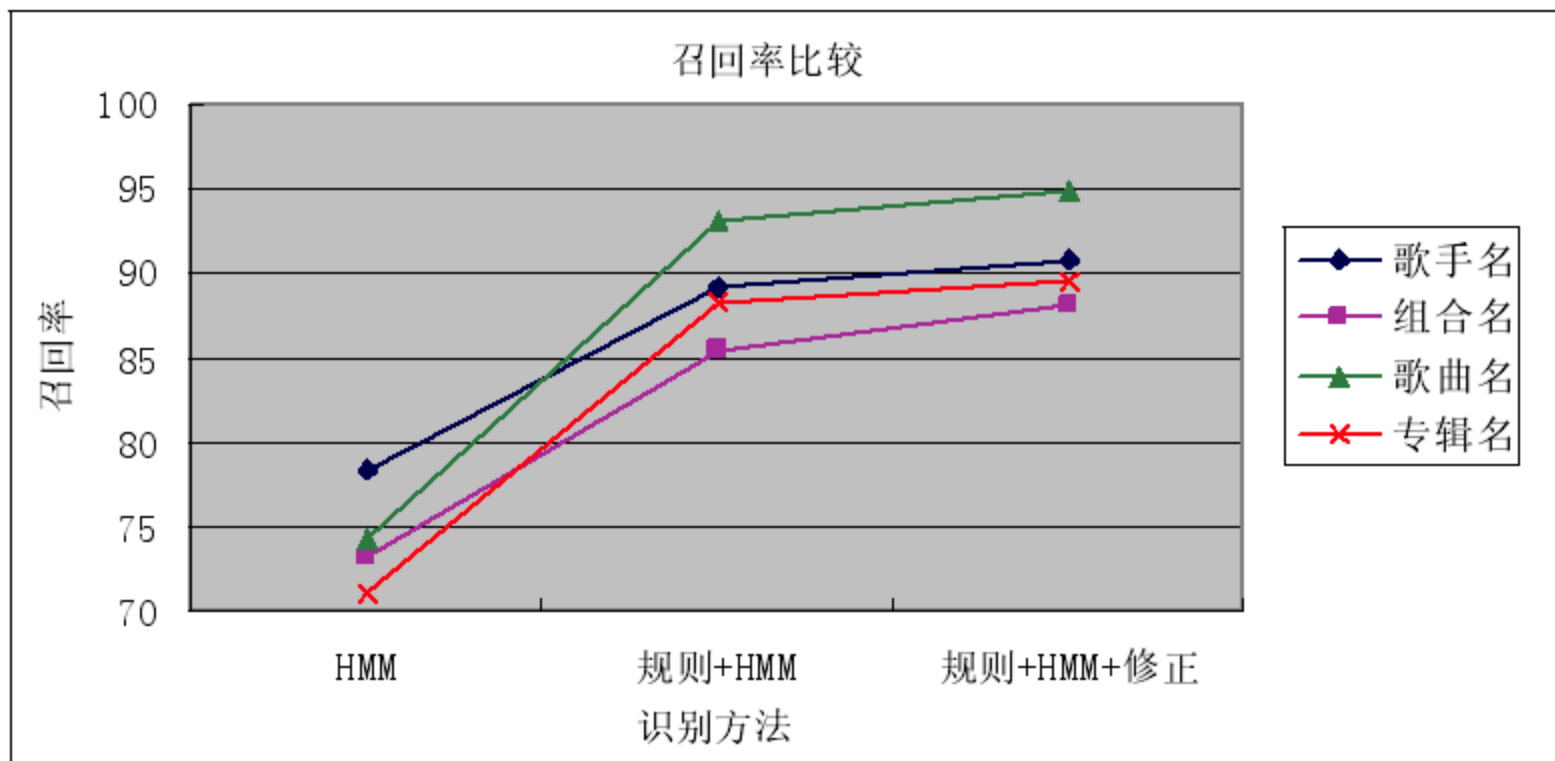
7 实验结果

- ❖ 1) 实验设计
- ❖ 2) 实验结果
- ❖ 3) 音乐名实体识别示例

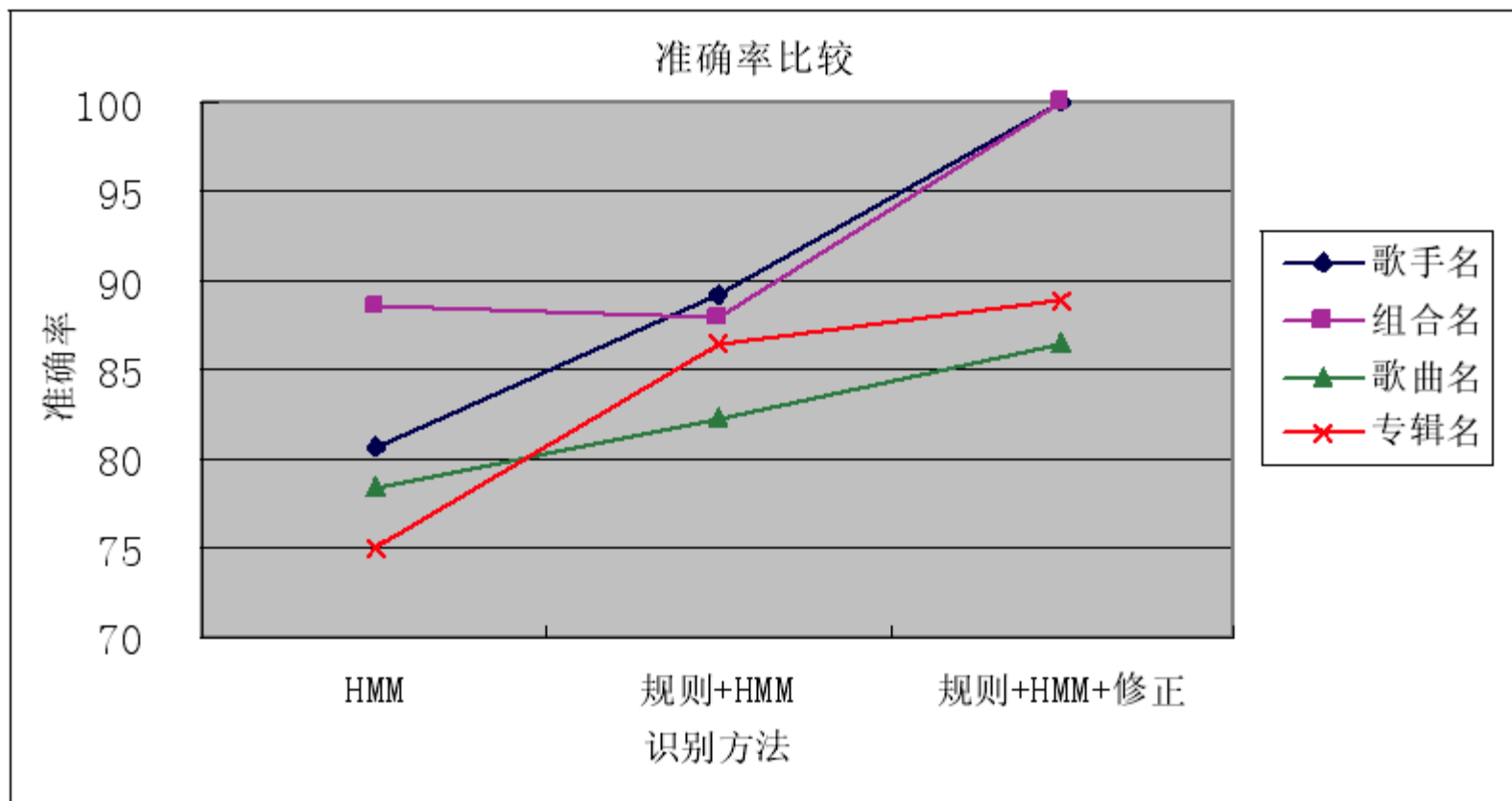
1) 实验设计

- ❖ 实验一：测试基于HMM的音乐实体识别方法的识别性能。
- ❖ 实验二：测试基于规则的前处理和基于HMM相结合的音乐实体识别方法的识别性能。
- ❖ 实验三：测试本文提出的规则和统计相结合的音乐实体识别方法的识别性能。该方法先使用基于规则的方法识别部分音乐实体，再使用基于HMM的方法来识别，最后对前两步识别出的结果进行修正处理。

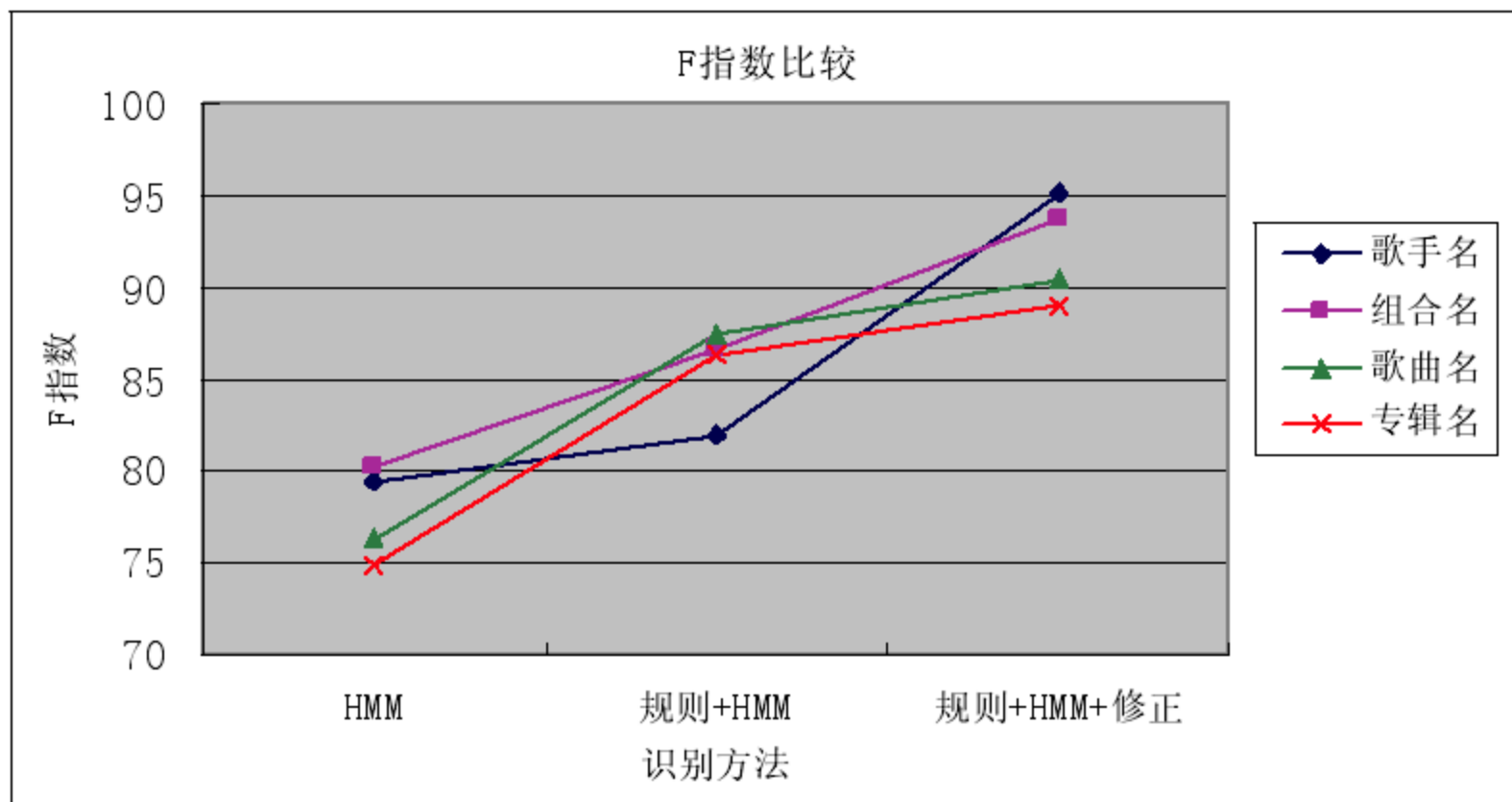
2) 实验结果:召回率比较




2) 实验结果:准确率比较



2) 实验结果: F值比较



3) 音乐名实体识别示例

 音乐命名实体识别系统

Web主题信息抽取

自动分词

音乐命名实体识别

原始文本：

一向极少在演唱会上翻唱他人作品的林忆莲，此次还将破天荒重新演绎华语乐坛几位年轻唱作人的作品。张震岳《爱我别走》、苏打绿《小情歌》、方大同《爱爱爱》均已被列入演唱会选曲中。此外，久未有新歌出炉的林忆莲，此次演唱会将首唱新歌《柿子》。

识别结果：

歌手名：林忆莲、张震岳、方大同
组合名：苏打绿
歌曲名：爱我别走、小情歌、爱爱爱、柿子
专辑名：

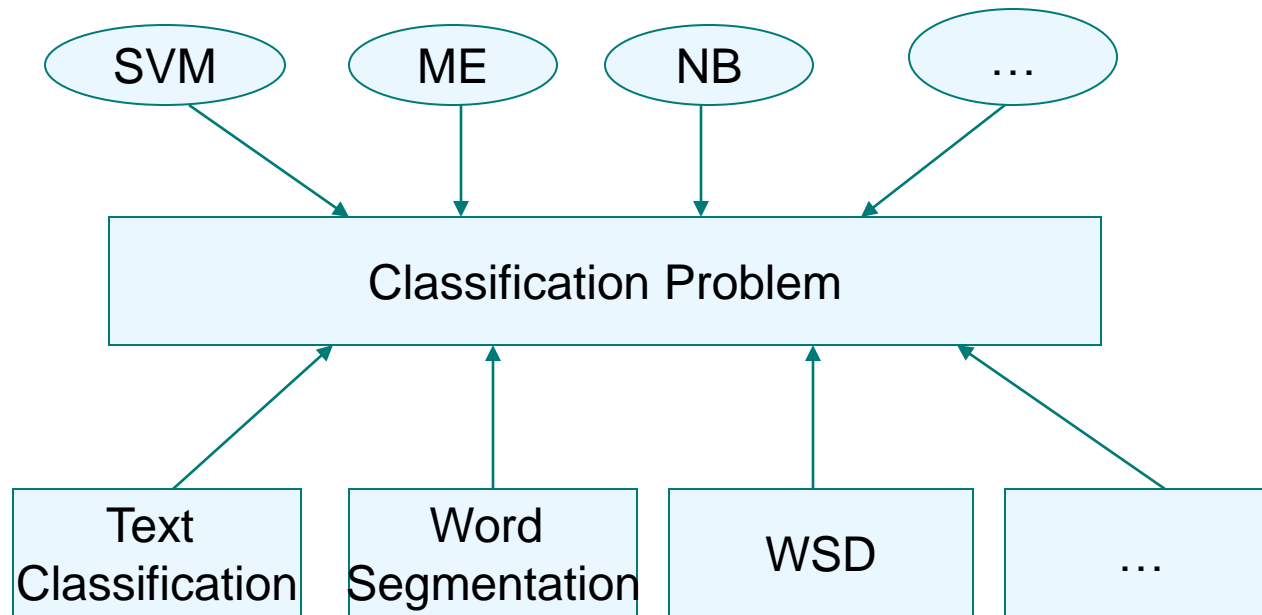
实体识别

保存结果

用机器学习方法解决NLP问题

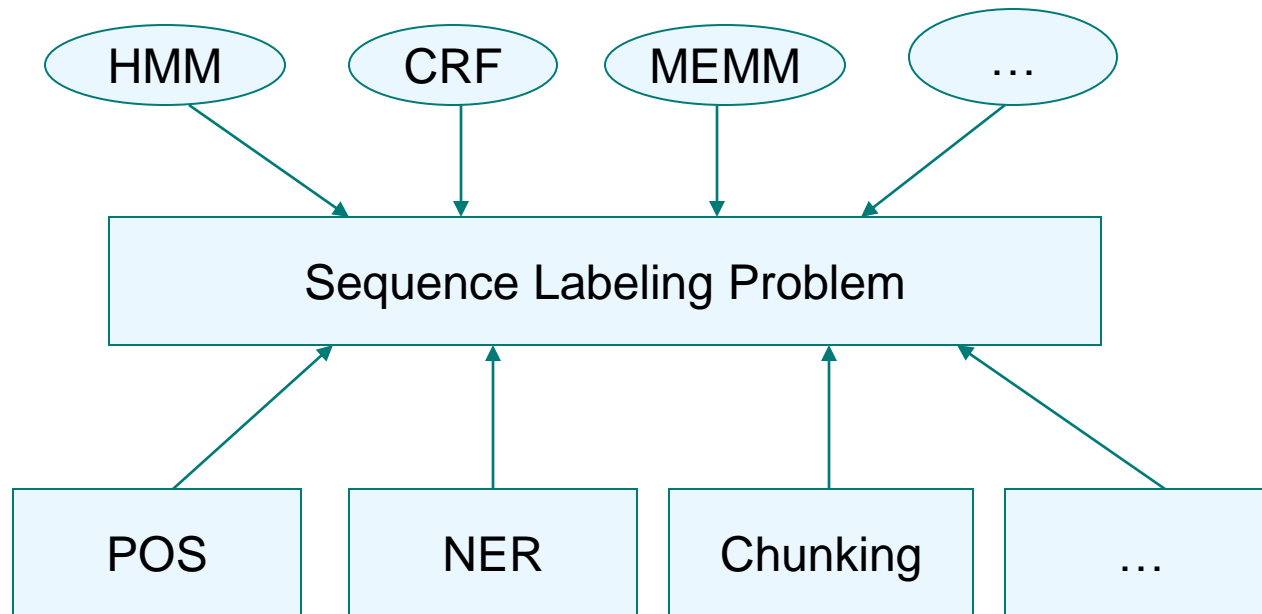
- ❖ 1) 确定方法和问题的关系
- ❖ 2) 特征工程
- ❖ 3) 评价方法

1) 确定关系 (E.g. 1)



Given a sample X , assign a class label Y to X according to $P(Y|X)$

确定关系 (E.g. 2)



Given a sequence $x_1x_2\dots x_n$, assign a class label sequence $y_1y_2\dots y_n$ according to $P(y_1y_2\dots y_n | x_1x_2\dots x_n)$

2) 特征工程

- ❖ Feature engineering
 - ❧ Feature selection
 - ❧ Find new features
- ❖ Need to understand current problem deeply
- ❖ Systematic method (Filter or Wrapper)
 - ❧ Filter method (such as MI) independent of concrete machine learning method
 - ❧ Wrapper dependent on concrete machine learning method
- ❖ Intuition → Experiment → Explanation

3) 评价方法

❖ Common metrics

- ∞ Precision

- ∞ Recall

- ∞ Accuracy

- ∞ F-measure (micro, macro)

- ∞ ROC

本章结束
谢 谢！