



音字转换技术

刘秉权

计算机科学与技术学院

新技术楼612室

Email: liubq@hit.edu.cn



主要内容

- 语音输入
- 汉字智能键盘输入
- 音字转换的主要方法
- 系统开发中的问题



音字转换

- 音字转换技术就是利用计算机这个工具采用人工智能、统计学、语言学等方法充分利用各种语言单位的上下文关系处理汉语语音音节串或拼音串到对应汉字串的自动转换。
- 应用
 - 汉语语音输入
 - 拼音键盘输入



汉语语音输入

- 语音识别阶段：把自然的语音信号转换为机器可以处理的数字表达的音节形式（或拼音形式）
- 语音理解阶段或音字转换阶段：把音节转换为汉字形式

先语音识别、
再语音理解



语音识别

- 语音识别单位

- 单音节（字识别）
- 多音节（词识别）

识别阶段

- 单音节语音识别多被采用

- 汉语是单音节语言，口语中，词与词之间很少有清晰的停顿，主要是单音节之间有间隔，单音节输入较易为人们接受
- 汉语中音节只有400个，音调节也只有1200个，词是由音节组合成，只要音节可以正确输入，则这些音节能构成无限的词组和语句。况且语音输入训练时只训练单音节是比较容易的，而在大词汇量（如几万个）情况下训练词的输入是难以让用户接受的
- 单音节输入比多音节输入变化范围小，对系统资源的要求较低



语音理解阶段

- 字处理：把语音识别器给出的和输入音节相近的几个音所包含的近音字在计算机屏幕上显示出来，让用户选择所需的汉字
- 词处理：通过用户读入音节停顿时间的不同确定词的长短，再和系统词库进行近音匹配，把近音词在屏幕上显示出来让用户选择。字词停顿难以控制
- 语句级处理：明显优于字词输入形式
 - 从操作心理学上看，操作人员倾向于按有一定意义的短语或句子为单位进行短时记忆
 - 从信息论角度讲，汉字的多维熵要少于一维熵，语句输入法比字词输入法需要较少的输入信息

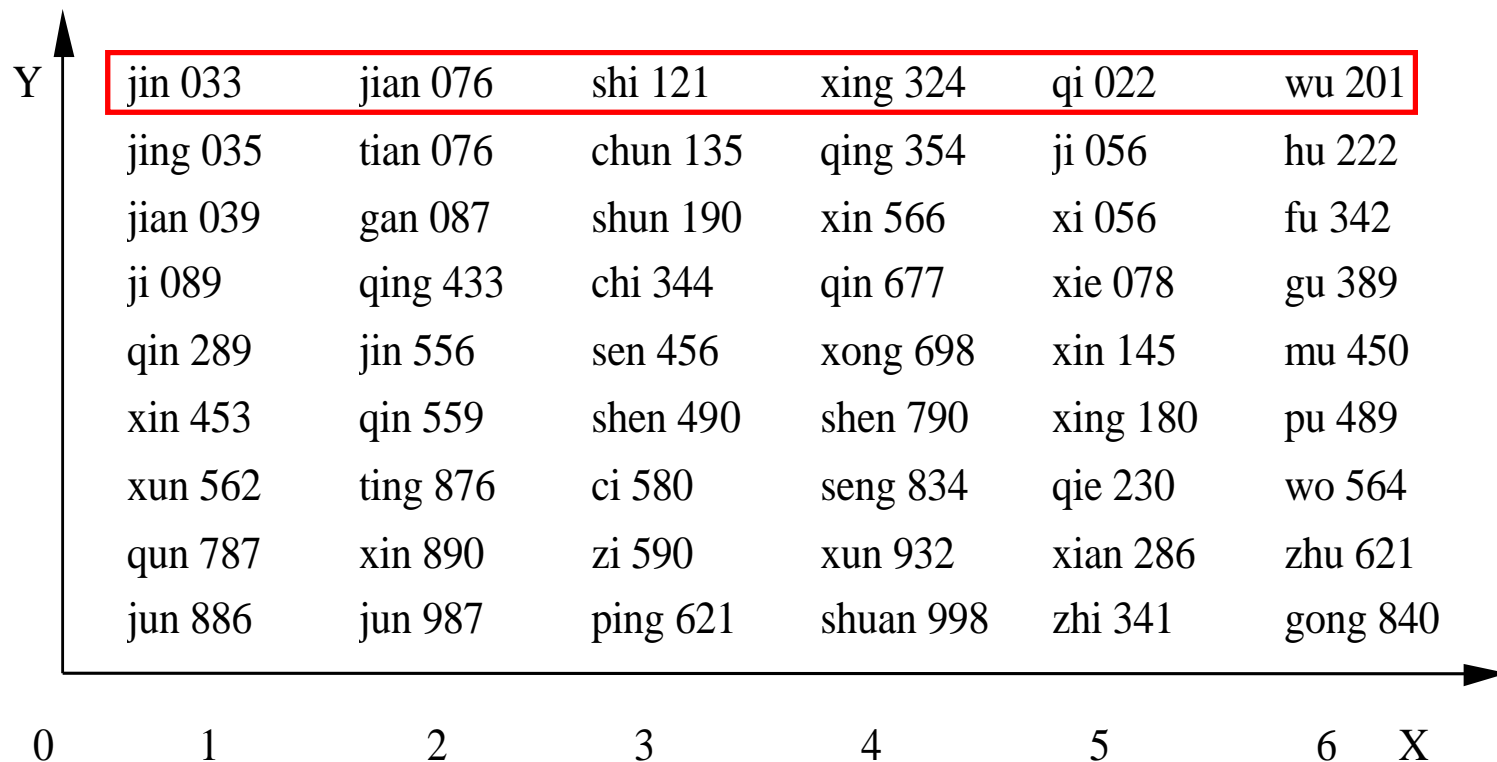


声音语句输入分阶段处理



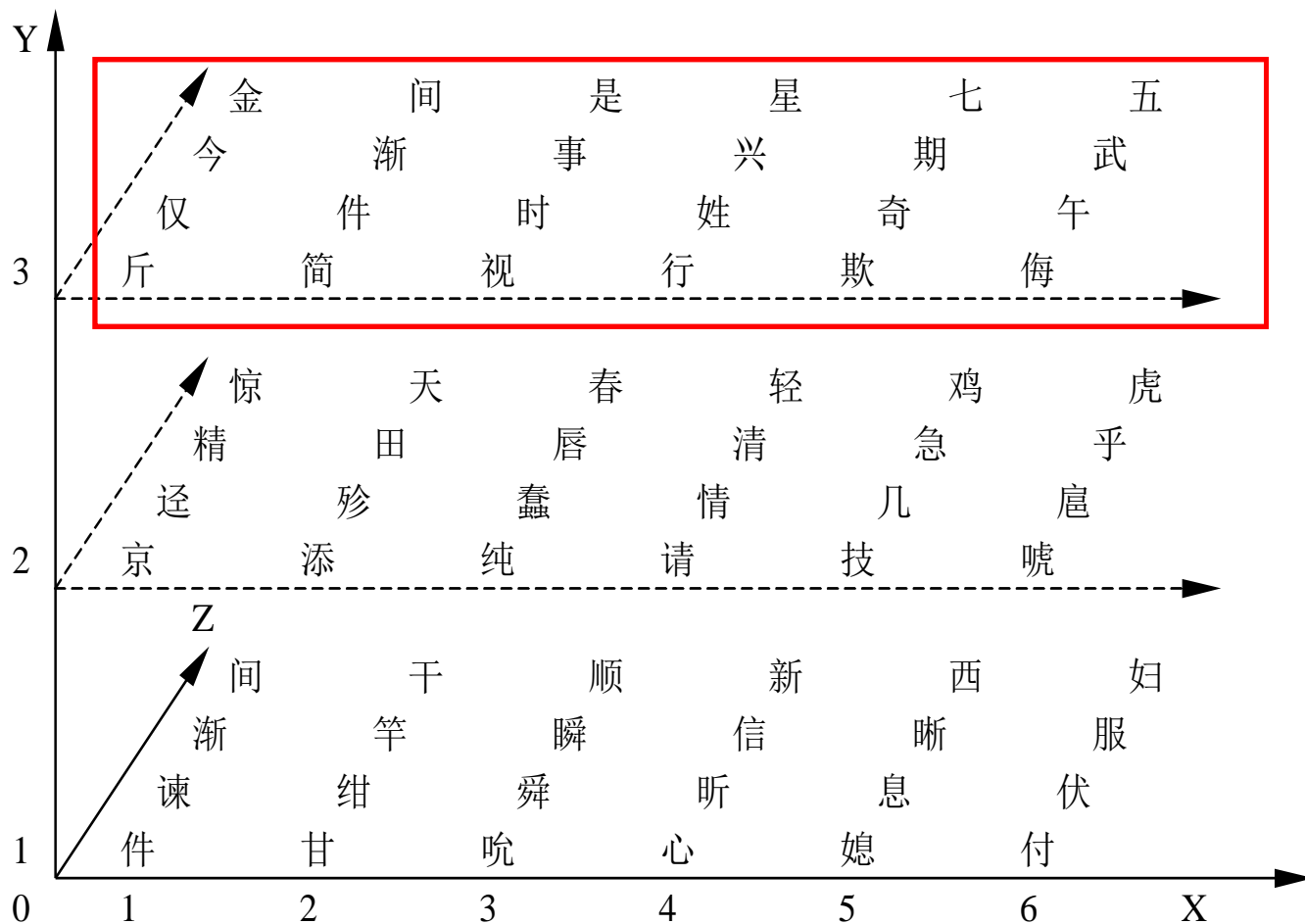
对“今天是星期五”

语音识别给出的音节候选向量



jin 033	jian 076	shi 121	xing 324	qi 022	wu 201
jing 035	tian 076	chun 135	qing 354	ji 056	hu 222
jian 039	gan 087	shun 190	xin 566	xi 056	fu 342
ji 089	qing 433	chi 344	qin 677	xie 078	gu 389
qin 289	jin 556	sen 456	xong 698	xin 145	mu 450
xin 453	qin 559	shen 490	shen 790	xing 180	pu 489
xun 562	ting 876	ci 580	seng 834	qie 230	wo 564
qun 787	xin 890	zi 590	xun 932	xian 286	zhu 621
jun 886	jun 987	ping 621	shuan 998	zhi 341	gong 840

音字候选三维向量

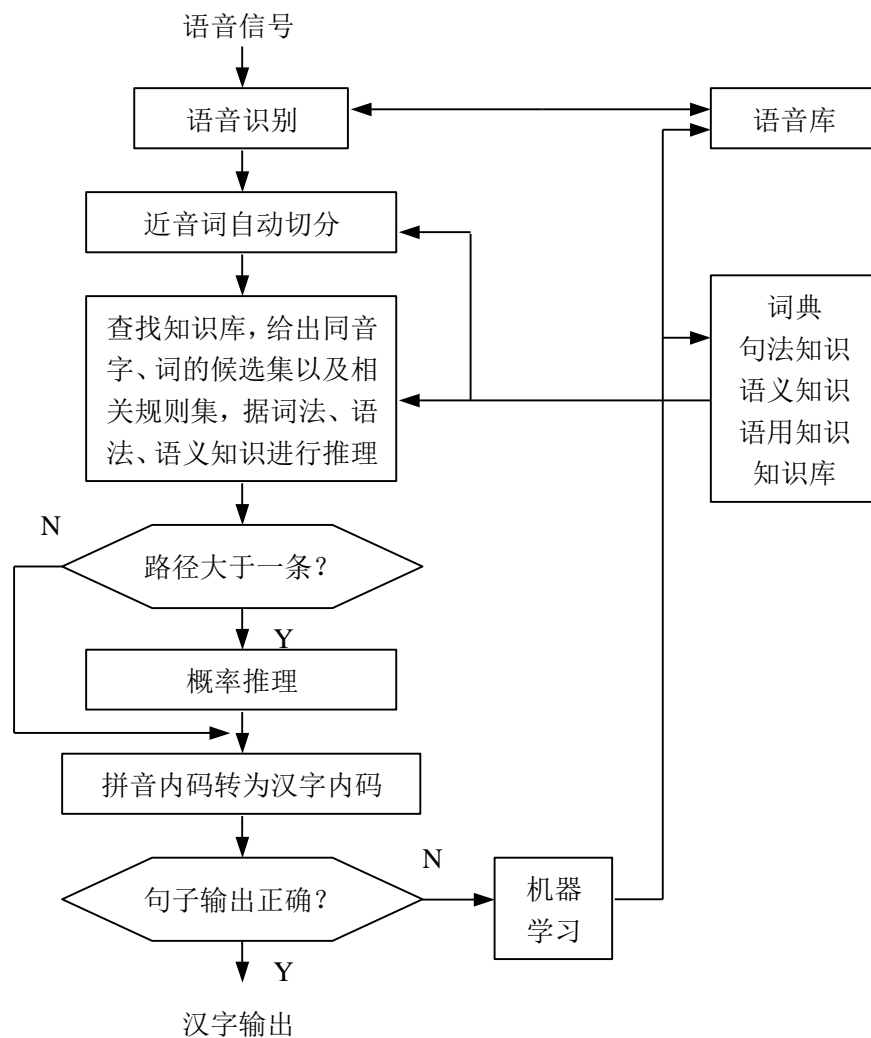




语音理解

- 模拟人们听写行为的思维过程，运用多种知识排除歧义性，在音字候选向量中找到一条沿X轴方向从1到N的最佳路径作为输出

声音语句输入的系统实现





汉字智能拼音键盘输入

- 基于字形的输入法:重码率低，难以掌握，适合专职打字员
- 基于拼音的输入法:易学，字词输入法重码率高，输入速度慢
- 智能输入：通常基于拼音，机器自动处理重码选择，易学，快速，输入过程更符合人的思维规律，适合广大非专职打字员



拼音输入的多种表达形式

- 拼音助学和提示输入：声母提示、拼音提示
- 拼音的压缩表达（简拼、双拼、三拼、声母输入）
- 用户自定义简拼
- 模糊拼音输入
- 逐键提示输入
- 面向数字键盘的数字拼音输入



拼音输入预处理-拼音流的切分

- 手工切分
- 自动切分
 - 切分规则(实用、简单)
 - 拼音统计模型
 - 与音字转换结合
- 自动切分，人工干预



拼音输入预处理-拼音纠错

■ 引起错误的原因

- 与人们日常使用语言文字的习惯有关
- 不同地域的人的发音存在着一定的差异
- 用户真正关心的是转换后汉字的正确，一般情况下并不对输入拼音的正确与否进行检查
- 对键盘的熟练程度

■ 拼写自动修改方法

- 通过对用户输入过程中所犯各种错误的分析，建立一种有效可行的打字模型，通过收集用户真实输入的数据，统计得到用户打字模型的参数
- 同时基于大量的中文文本，训练得到一个强大的中文语言模型，并与中文的打字模型相结合，采用类似语音识别的技术，修改用户输入中的各种错误，并得到最合适的汉字
- 拼写纠正不仅可以进行用户自适应，而且还适用于各种语言



音字转换(键盘输入)的实现方法

- 基于理解的方法
- 基于语用统计的方法
- 基于模板匹配的方法



基于理解的方法

- 利用汉语语法知识来消化同音字、词，以及化解歧义分词
- 学科分类中属于人工智能的自然语言理解分支
- 根据自动分词得到同音字、词的候选集，查找知识库得到相关的规则，再经过归约推理，得出转换结果
- 特点：优点-系统正确率比较稳定。缺点-语言覆盖面较小，当输入语句的语法不规范时，不能有效处理；建立知识库时知识表达和知识获取均非常困难



词法规则

- $\langle \text{姓} \rangle \langle \text{名} \rangle | \langle \text{小} \rangle \langle \text{姓} \rangle | \langle \text{老} \rangle \langle \text{姓} \rangle | \dots \rightarrow \langle \text{人名} \rangle$
- $\langle \text{张} \rangle | \langle \text{王} \rangle | \langle \text{李} \rangle | \langle \text{赵} \rangle | \dots \rightarrow \langle \text{姓} \rangle$
- $\langle \text{基数} \rangle | \langle \text{序数} \rangle \rightarrow \langle \text{数词} \rangle$
- $\langle \text{系数} \rangle | \langle \text{系数} \rangle \langle \text{位数} \rangle | \langle \text{系数} \rangle \langle \text{位数} \rangle \langle \text{基数} \rangle \rightarrow \langle \text{基数} \rangle$
- $\langle \text{个} \rangle | \langle \text{十} \rangle | \langle \text{百} \rangle | \langle \text{千} \rangle | \langle \text{万} \rangle | \dots \rightarrow \langle \text{系数} \rangle$



短语规则

- <形容词><名词>|<名词><名词> → <名词短语>
- <副词><动词>|<动词><动态助词> → <动词短语>
- <着>|<了>|<过> → <动态助词>



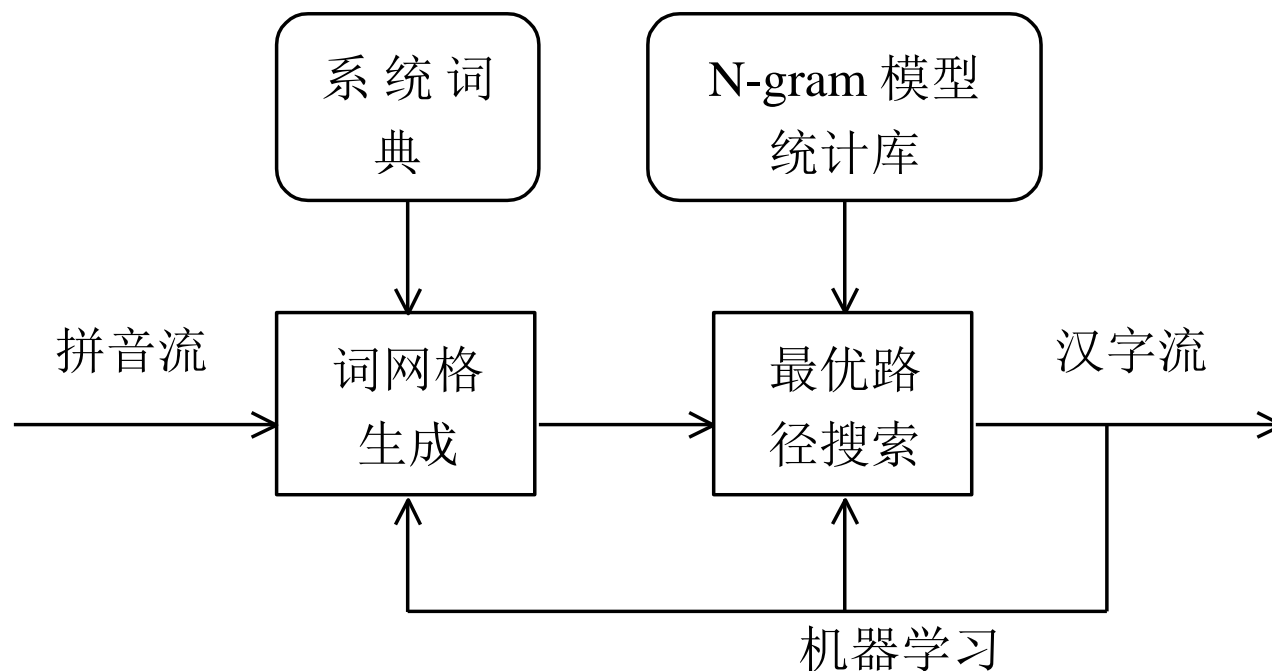
句法规则

- $\langle \text{主语} \rangle \langle \text{动词} \rangle | \langle \text{主语} \rangle \langle \text{状语} \rangle \langle \text{动词} \rangle | \dots \rightarrow S$
- $\langle \text{动物} \rangle \langle \text{吃} \rangle \langle \text{食物} \rangle \rightarrow S$

基于语用统计的方法

- 主要利用语用统计的数据来消化同音字、词，以及化解歧义分词
- 通过汉语字与字或词与词之间的同现概率来完成汉语语用统计库的构造
- 根据拼音输入构造词网格
- 通过概率计算求得词网格中的最佳路径
- 为减少状态空间的搜索时间，必须采用有效的搜索方法，比如采用动态规划的Viterbi算法或各种A*算法
- 特点：转换率较高，是主流方法，对不同领域文本具有偏向性
- 基于理解和基于语用统计相结合的方法：实现两种方法的取长补短

N-gram模型实现音字转换的系统结构图



词网格示意图

观测序列

o_0

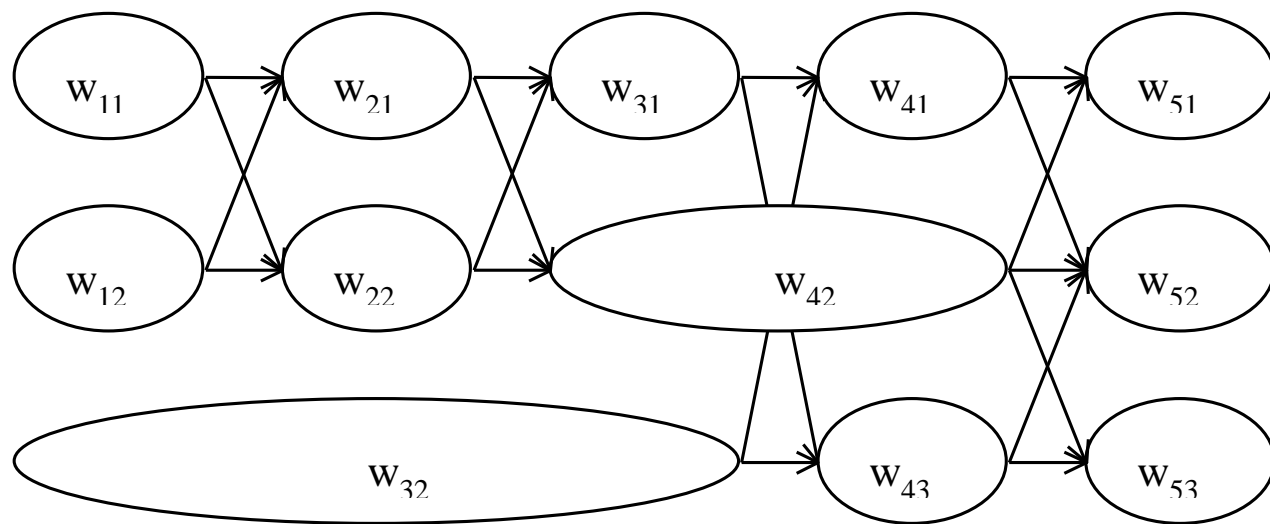
o_1

o_2

o_3

o_4

候选词





动态规划，Viterbi算法

- 动态规划(dynamic programming)是运筹学的一个分支，是求解决策过程(decision process)最优化的数学方法。
- 20世纪50年代初美国数学家R.E.Bellman等人在研究多阶段决策过程(multistep decision process)的优化问题时，提出了著名的最优化原理(principle of optimality)，把多阶段过程转化为一系列单阶段问题，利用各阶段之间的关系，逐个求解，创立了解决这类过程优化问题的新方法——动态规划。
- 1957年出版了他的名著Dynamic Programming，这是该领域的第一本著作。
- Viterbi算法是一种动态规划算法。



Markov模型

$$M = \langle N, \pi, A \rangle,$$

N ：为状态数目

状态转移概率矩阵： $A = a_{ij}$

其中， $a_{ij} = P(q_t = S_j | q_{t-1} = S_i)$ ， $1 \leq i, j \leq N$

$$a_{ij} \geq 0, \sum_{j=1}^N a_{ij} = 1$$

初始状态概率分布： $\pi = \pi_i$

其中， $\pi_i = P(q_1 = S_i)$ ， $1 \leq i \leq N$ ， $\pi_i \geq 0$ ， $\sum_{i=1}^N \pi_i = 1$



Viterbi最优路径搜索

Viterbi 算法可以在给定的 Markov 模型 $M = \langle N, \pi, A \rangle$ 的情况下, 从词网格中求出一在该模型下最可能的状态序列 $q^* = q_1 q_2, \dots, q_T$ 。定义 $\delta_t(i)$ 为时刻 t 时, 从起始结点到当前时刻第 i 个候选 w_i 的最佳路径的权值, $\varphi_t(i)$ 纪录概率最大路径上当前状态的前一个状态。



Viterbi算法

1 • 初始化 $\delta_1(i) = \pi_i$,

$$\phi_1(i) = 0, \quad 1 \leq i \leq N_1$$

2 • 递归 $\delta_t(j) = \max_{1 \leq i \leq N_{t-1}} \delta_{t-1}(i) \cdot a_{ij}$

$$\phi_t(j) = \operatorname{argmax}_{1 \leq i \leq N_{t-1}} \delta_{t-1}(i) \cdot a_{ij}, \quad 2 \leq t \leq T, \quad 1 \leq j \leq N_t$$

3 • 终结 $p^* = \max_{1 \leq i \leq N_T} \delta_T(i)$

$$q_T^* = \operatorname{argmax}_{1 \leq i \leq N_T} \delta_T(i)$$

4 • 状态序列求解: $q_t^* = \phi_{t+1}(q_{t+1}^*)$, $t = T-1, T-2, \dots, 1$

其中 N_t 为结束于时刻 t 的候选词（状态）的个数。



基于模板匹配的方法

- 寓汉语语法知识于巨量的短语串中，进而利用这些短语串来消化同音字、词，以及化解歧义分词。这种短语串通常称之为“模板词”。
- 根据分词后的输入语句查找模板词库和句法规则库，然后进行匹配处理。如果匹配结果唯一，则不必再用概率推理；若存在两个以上的候选结果时，则根据句法规则或概率推理进一步判定，选出一个最有希望的可能结果作为输出。
- 特点：对于已经搜索过模板词的或者具有相同类型的领域，系统的转换正确率比较高。但由于模板词数量巨大，对计算机存储空间要求较高。



系统开发中的问题

- 功能设计
- 界面设计
- 词表构建
- 统计模型数据结构
- 词网格构建及最优路径计算
- 操作系统挂接



功能设计

- 拼音输入方式
- 句内编辑
- 中英文切换
- 机器学习



界面

- 逐键跟随
- 编辑条嵌入编辑器
- 每页显示候选数
- 更换皮肤
- 各种新体验



词表构建

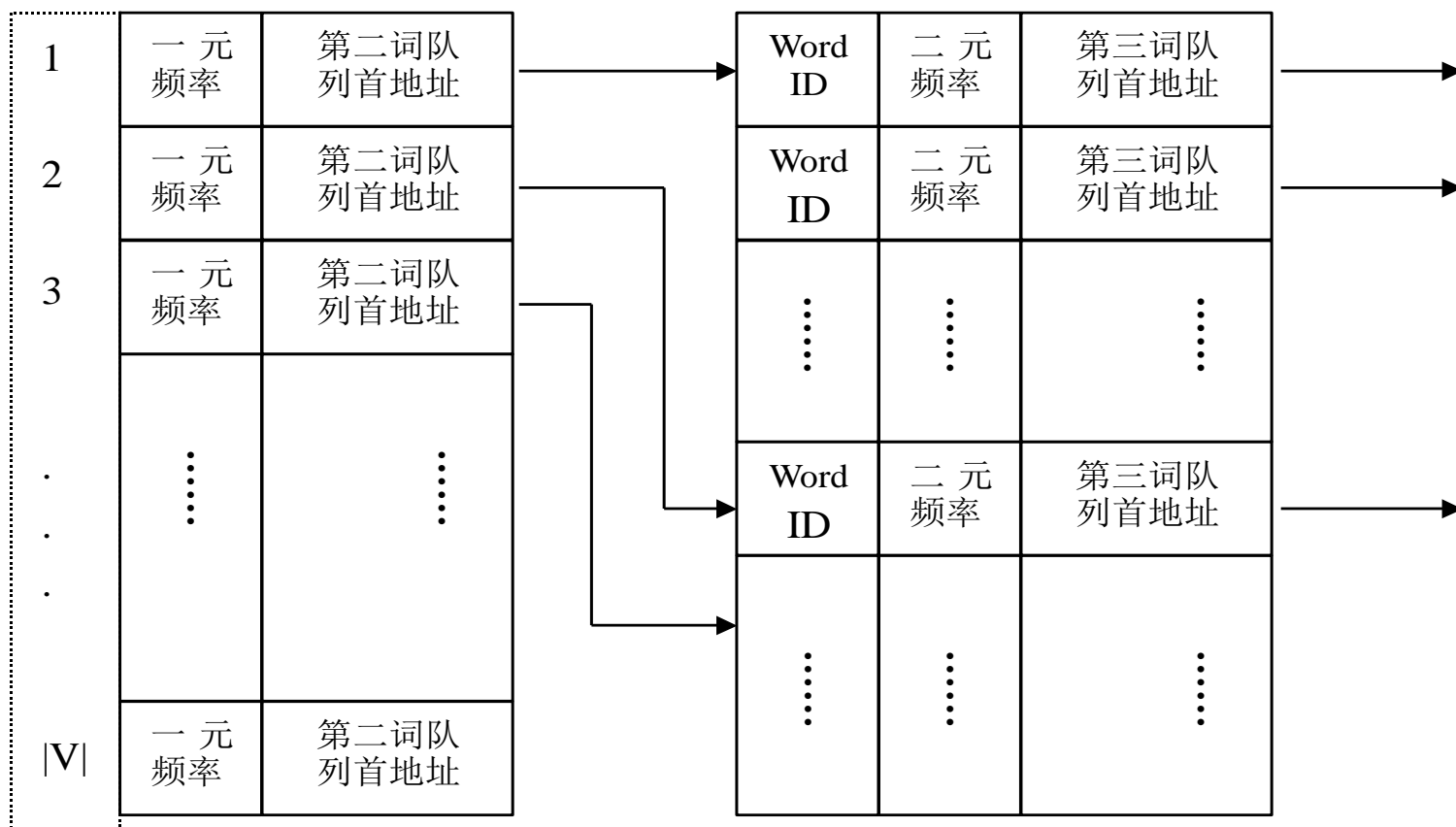
- 汉字编码确定
- 词条的表示形式
- 词表规模
- 词表存储形式
- 词表加工



统计模型的数据结构

- 数组结构
- 哈希表
- 其它压缩方式
- 领域模型、自适应模型的存储

N-gram语言模型的一种存储结构





词网格构建及最优路径计算

- 基本数据结构
- 变长节点如何处理
- 词网格更新
- 动态最优路径计算



操作系统挂接

■ 钩子函数

- WINDOWS的主要特性之一
- 捕捉自己进程或其它进程发生的事件
- 通过“钩挂”，给WINDOWS一个处理或过滤事件的回调函数，该函数也叫做“钩子函数”
- 每次发生感兴趣的事件时，WINDOWS都将调用该函数

■ IME

- 输入法编辑器(Input Method Editor)，一种专门的应用程序，用来输入代表东亚地区书面语言文字的不同字符
- 无法用普通的编程环境来单步调试



网络环境汉字智能输入

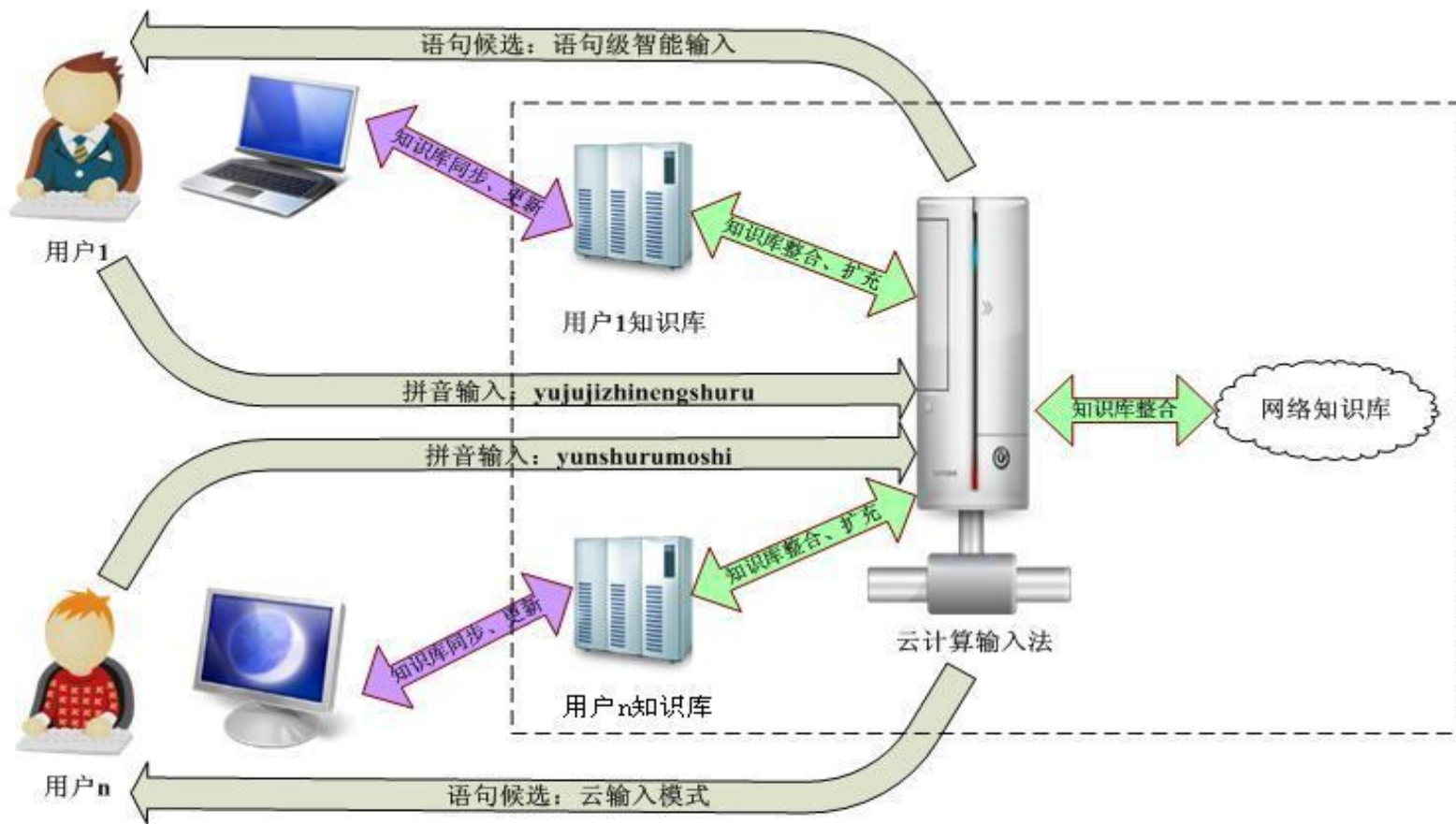
- 大规模、超大规模语料获取与加工
 - 定向网络爬虫
 - 大规模文本分类
- 领域词表的加工
 - 生词识别
 - 领域术语识别
- 领域语言模型的训练
- 个性化学习与热词推荐
- 信息安全（用户隐私）问题
- 云计算输入模式
 - 存储分工
 - 计算分工



云计算输入模式

- 基于云存储的输入模式实现用户的领域专业属性和输入习惯的网络存储、同步和更新
 - 免安装模式完全通过服务器完成知识库的管理和整个输入过程
 - 协同式输入模式通过本地客户端输入法和服务器端云输入法同时完成输入过程，客户端对输入结果进行整合
 - 服务器提供云存储输入模式

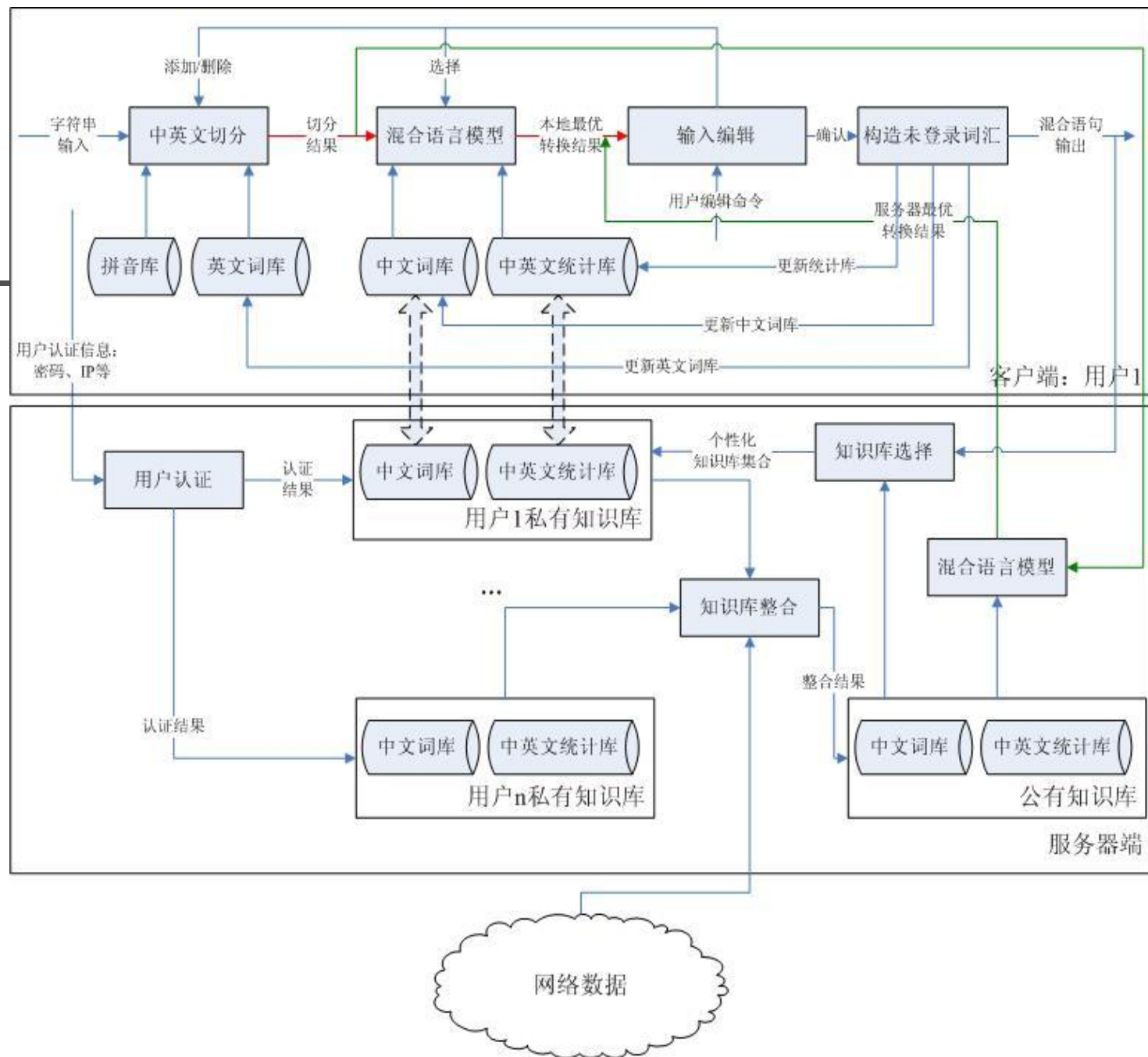
图示

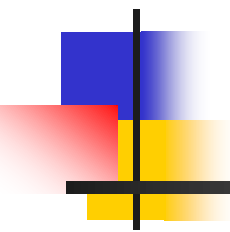


图中描绘出了基于云计算的三种输入模式：

- 1、紫色箭头表示基于云存储的输入模式：输入过程由本地客户端机器完成，服务器对用户私有知识库进行存储用来同步和更新。
- 2、灰色箭头表示免安装的云输入模式：输入过程由服务器完成并返回给用户。
- 3、蓝色箭头和紫色箭头一起表示基于云计算的协同式输入模式：本地客户端输入法和服务器端云输入法同时完成输入过程，对输入结果进行整合返回给用户。服务器对所有用户的私有知识库和网络知识库进行整合，并根据用户用户领域专业属性和输入习惯对用户私有知识库进行自动扩展和更新。

语句级汉字输入的系统结构





谢谢！
