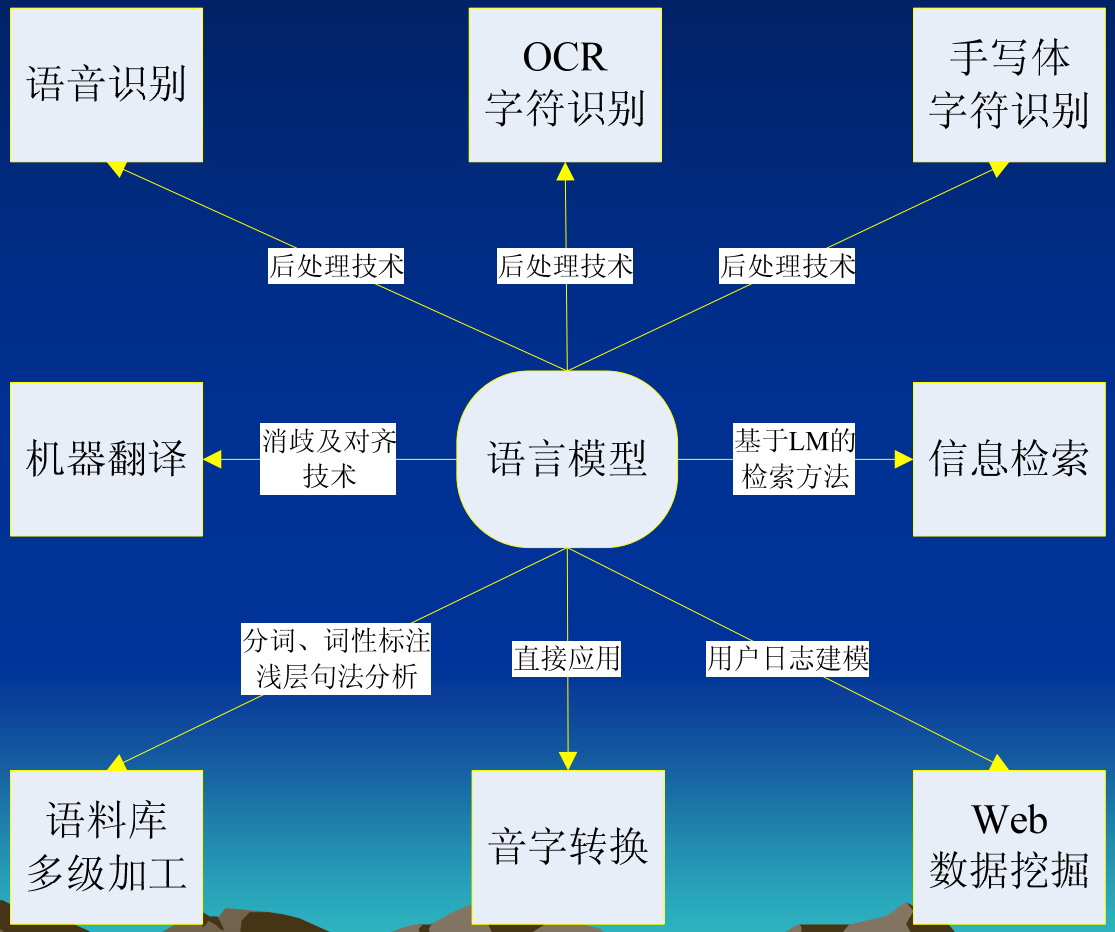
统计语言模型简介

报告人：肖镜辉
研究方向：语言模型
汉字键盘输入

提纲

- 研究意义
- 语言模型定义
- Ngram语言模型
- 指数语言模型
- 神经网络语言模型

研究意义


语音识别
OCR
字符识别
手写体
字符识别
后处理技术
后处理技术
后处理技术
机器翻译
消歧及对齐
技术
语言模型
基于LM的
检索方法
信息检索
分词、词性标注
浅层句法分析
直接应用
用户日志建模
语料库
多级加工
音字转换
Web
数据挖掘

语言模型定义

- 语言模型是针对某种语言建立的概率模型，使得正确词序列的概率值大于错误词序列的概率值——Goodman
- 对于词序列$w_1 \ldots w_m$，其概率为$P(w_1 \ldots w_m)$

$$p(w_1 w_2 \ldots w_m) = p(w_1) \prod_{i=2}^{m} p(w_i \mid w_1 \ldots w_{i-1})$$

语言模型定义cont

- 理论评价标准：迷惑度
  - 信息论定义
  - 与测试语料相关

$$PP_c = 2^{-\frac{1}{N_c}\sum_{i=2}^{N_w}\log_2 p(w_i|w_1 \ldots w_{i-1})}$$

- 实践评价标准：错误率
  - 与测试系统相关

Ngram语言模型

- 基本Ngram模型
- 平滑技术
- 几个变种

基本Ngram模型

- 基本假设
  - **有限历史假设：** 当前词的条件概率仅仅与前*n-1*个词相关，而与该词序列的整个历史无关。
  - **齐次性假设：** 当前词的条件概率与当前词在词序列中的位置无关。
- 概率函数形式

$$p(w_1 w_2 \ldots w_m) = p(w_1)\prod_{i=2}^{m} p(w_i \mid w_{i-n} \ldots w_{i-1})$$

基本Ngram模型cont

- 模型训练
  - 原理：最大似然估计(Maximum Likelihood Estimation)
  - 公式

$$p(w_i \mid w_{i-n}...w_{i-1}) = \frac{C(w_{i-n}...w_{i-1}, w_i)}{C(w_{i-n}...w_{i-1})}$$

基本Ngram模型cont

- 优点
  - 简单(计算、训练)
  - 高效(MLE、Viterbi decoding)
- 存在的问题
  - 数据稀疏问题
  - 长距离约束问题
  - 自适应问题
  - 语言学知识利用问题
  - ……

工业界的宠儿；
学术界的弹靶
(baseline)！

Ngram模型的平滑技术

- 问题提出
- Additive smoothing
- Good-Turing smoothing
- Katz smoothing
- Interpolation smoothing

数据稀疏问题

- 问题描述： The problem of data sparseness, also known as the **zero-frequency problem** arises when analyses contain configurations(结构) that **never** occurred in the training corpus. Then it is not possible to estimate probabilities from observed frequencies, and some other estimation scheme that can generalize (that configurations) from the training data has to be used. —— Dagan

数据稀疏问题cont

- 实质：大规模统计方法与有限规模语料(训练集)之间的矛盾
  - 不可避免：Zipf's law
  - 非常普遍
    - IBM, Brown：366M英语语料训练trigram，在测试语料中，有14.7%的trigram和2.2%的bigram在训练语料中未出现
    - 我们的实验：500万字人民日报训练bigram模型，用150万字人民日报作为测试语料， 23.12%的bigram未出现
    - 对于音字转换：用上述bigram模型构建音字转换系统，45.33%的错误是由数据稀疏引起
  - 后果严重：尤其是零概率问题

Additive smoothing

- 思想：每个Ngram在训练语料中至少出现1次
- 公式：

$$p(w_i \mid w_{i-n}...w_{i-1}) = \frac{C(w_{i-n}...w_{i-1}, w_i)+1}{C(w_{i-n}...w_{i-1})+N}$$

- 扩展：Laplace law, Lidstone law, Jeffreys-Perkes law……

Good-Turing smoothing

- 思想：利用频率的类别信息来平滑频率
- 公式：

$$p_{GT}(w_i \mid w_{i-n}...w_{i-1}) = \frac{C_{GT}(w_{i-n}...w_{i-1}, w_i)}{C(w_{i-n}...w_{i-1})}$$

$$C_{GT}(w_{i-n}...w_{i-1}, w_i) = (C(w_{i-n}...w_{i-1}, w_i)+1) \times \frac{E(C(w_{i-n}...w_{i-1}, w_i)+1)}{E(C(w_{i-n}...w_{i-1}, w_i))}$$

$$E(C) \approx N(C)$$

- 扩展：对E(C)的更精确估计，e.g 逻辑回归方法

Katz smoothing

- 思想：当高阶模型不可靠时，采用低阶模型的概率
- 公式：

$$P(w_n \mid w_1 \ldots w_{n-1}) = \begin{cases} P_d(w_n \mid w_1 \ldots w_{n-1}) & if\ C(w_1 \ldots w_n) > 0 \\ \alpha(w_1 \ldots w_{n-1}) P_{GT}(w_n \mid w_1 \ldots w_{n-1}) & otherwise \end{cases}$$

$$\alpha(w_1 \ldots w_{n-1}) = \frac{\beta(w_1 \ldots w_{n-1})}{1 - \sum_{w_2 : C(w_1 \ldots w_n) > 0} P_{GT}(w_n \mid w_1 \ldots w_{n-1})} \qquad \beta(w_1 \ldots w_{n-1}) = 1 - \sum_{w_2 : C(w_1 \ldots w_n) > 0} P_d(w_n \mid w_1 \ldots w_{n-1})$$

- 扩展：对低阶模型的选择，KN平滑算法，Rosenfeld1992工作，肖镜辉的工作，等

Interpolation smoothing

- 思想：将高阶模型和低阶模型作线性组合
- 公式：

$$P_{inter}(w_i \mid w_{i-n+1},...w_{i-1}) = \lambda_n \times P_{inter}(w_i \mid w_{i-n+2},...w_{i-1}) + (1-\lambda_n) \times P(w_i \mid w_{i-n+1},...w_{i-1})$$

- λ_n优化值：EM算法
- 扩展：对组成模型的选择及λ_n的优化方法，Witten-Bell算法，肖镜辉的工作，等

小结

- 经典平滑方法采用统计学手段来解决一个工程问题
- 忽略了语言学的作用
- Stanley F. Chen and Joshua Goodman. An empirical study of smoothing techniques for language modeling. Computer Speech and Language, 13:359-394, October 1999.

Class-based Ngram Model

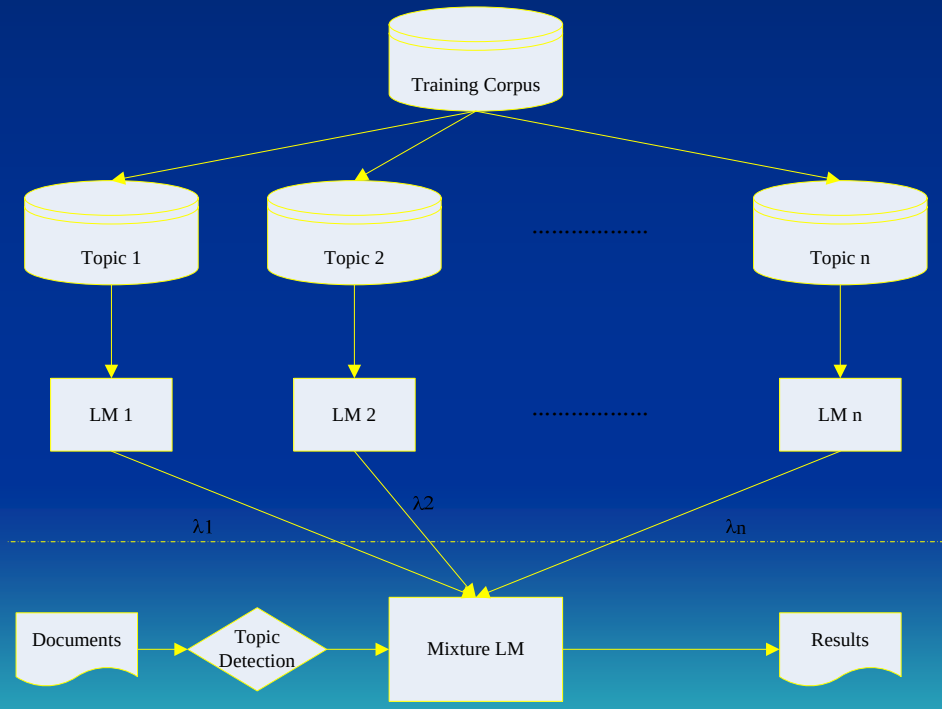
- 思想：基于词类建立语言模型
  - 缓解数据稀疏问题
  - 融合部分语法信息
- 公式：

$$p(c_1 c_2 \ldots c_m) = p(c_1)\prod_{i=2}^{m} p(c_i \mid c_{i-n} \ldots c_{i-1})$$

- 缺点
  - 描述能力低
  - 合适的词类难以获取

Topic-Based Ngram Model

- 思想：将训练集划按主题分成多个子集，并分别建立Ngram模型，以解决语言模型的主题自适应问题。



Topic-Based Ngram Model cont

- 公式：

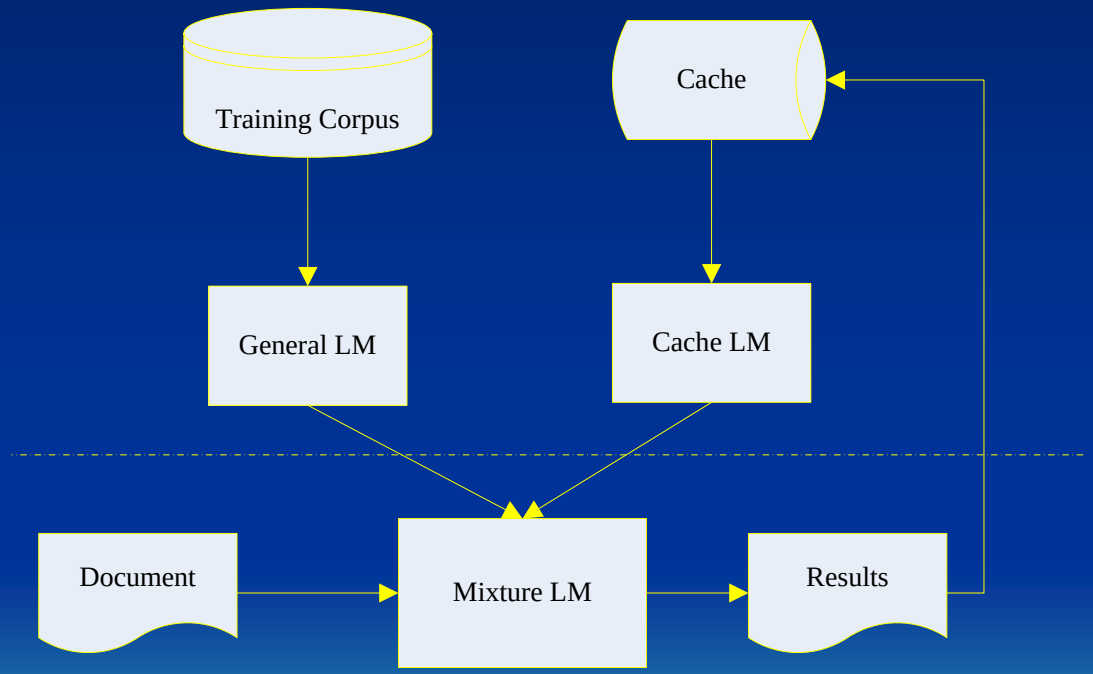
$$P(w_i \mid w_{i-n+1},\ldots w_{i-1}) = \sum_{Topic=i} \lambda_i P_i(w_i \mid w_{i-n+1},\ldots w_{i-1})$$

- 难点：
  - 主题划分方法
  - 优化权值确定

Cache-based Ngram Model

- 思想：利用cache缓存前一时刻的信息，以用于计算当前时刻概率，解决语言模型动态自适应问题
- 根据：
  - People tends to use words as few as possible in the article.
  - If a word has been used, it would possibly be used again in the future

Cache-based Ngram Model cont



Cache-based Ngram Model cont

- 公式：

$$P_{Mixed}(w_i|w_{i-n+1},...w_{i-1})=\lambda\times P(w_i|w_{i-n+2},...w_{i-1})+(1-\lambda)\times P_{Cache}(w_i|w_{i-n+1},...w_{i-1})$$

- 难点：
  - Cache大小
  - 优化权重
  - 模型评测

Skipping Ngram Model

- 思想：The current word is constrained by the skipped words in the word history, other than the adjacent words.
- 优点：Exploiting more information of history words and avoid the curse of dimensionality meanwhile.
- 公式：

$$P_{Mixed}(w_i \mid w_{i-n+1}, \ldots w_{i-1}) = \sum_{k=1}^{n-1} \lambda_k \times P(w_i \mid w_{i-k})$$

Trigger-based Ngram Model

- 思想：利用trigger刻画远距离约束关系
- 难点：
  - Trigger的抽取
  - 与Ngram模型的融合

指数语言模型

- 最大熵模型
- 最大熵马尔科夫模型
- 条件随机域模型

最大熵模型

- Rosenfeld 1994
- 根据最大熵原理估计Ngram概率
- 优点：
  - 融合多种知识源
- 缺点：
  - 训练算法时间复杂度过高，不适合处理大标记集问题

最大熵马尔科夫模型

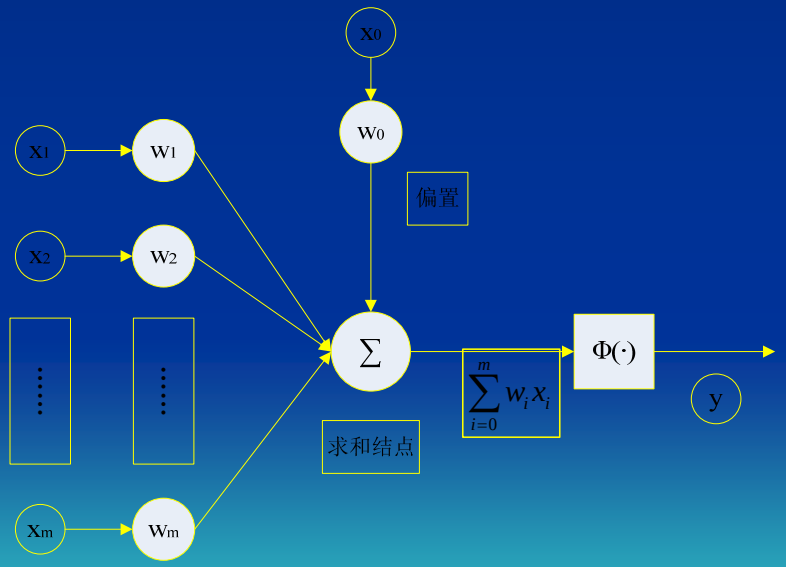
- McCallum 2000
- 根据最大熵原理估计HMM中的发射概率和转移概率
- 优点：
  - 使ME更适用于解决序列问题
- 缺点：
  - 局部概率归一化——标记偏置问题

条件随机域模型

- Lafferty 2001
- 根据Markov随机域相关理论描述语言序列
- 优点：
  - 全局归一化：避免MEMM标记偏置问题
- 缺点：
  - 时间复杂度过高

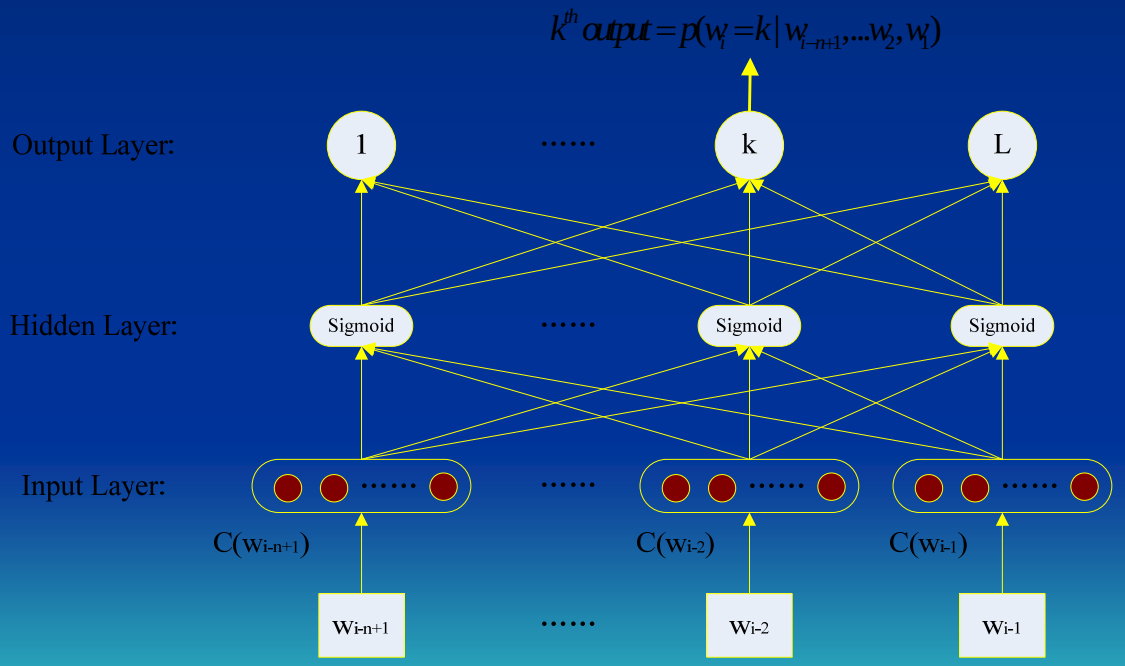
神经网络语言模型

- Bengio 2000
- 思想：利用神经网络估计Ngram概率
- 神经元模型：



神经网络语言模型cont

- 神经网络语言模型



神经网络语言模型cont

- 优点
  - 避免数据稀疏问题
- 缺点
  - 计算量过大，往往需要大规模并行机群支持