

# 基于混合模型的中文命名实体抽取系统的研究与实现

上海交通大学计算机系  
王睿, 张洁, 张由仪, 于禛  
导师: 姚天昉  
[mars198356@hotmail.com](mailto:mars198356@hotmail.com)

# 内容提要

- 引言
- 相关系统与技术
- 三个问题
  - 分词的问题
  - 领域的问题
  - 方法的选择
- 系统的架构与实现
- 测试结果与分析
- 结论与未来的工作

# 引言——研究动机

- 信息时代
  - 文本信息
    - 信息抽取 (Information Extraction)
  
- 搜索引擎
  - 引入自然语言技术
    - 命名实体抽取 (Name Entity Extraction)

# 引言——研究现状

- 中文分词
  - 不分词
  - 自动分词系统
  - 人工分词
  
- 限制领域 or 非限制领域（开放领域 Open Domain）
  - 准确率，难度
  - 适用范围
  
- 统计学 vs. 语言学
  - 基于规则（Rule-based）
  - 词频统计

# 相关系统与amp;技术

- 中科院分词系统ICTCLAS介绍

- 举例:

- 电影/n 《/w 王子/n 复仇/v 记/ng 》 /w 改编/v 自/p 莎士比亚  
/nr 的/u 什么/r 作品/n

- 基于支持向量机的文本分类系统

- 中文自然语言处理开放平台

- [www.nlp.org.cn](http://www.nlp.org.cn)

# 三个问题——分词的问题

- 为什么要分词
  - 词库匹配无法解决歧义情况
  - 抽取规则往往只能确定词的一个边界
  
- 一些分词错误的例子
  - 李文/和/在/操场/上.....（正确：李文和/在/操场/上.....）
  - 他们/同上/海/、/沈阳/方面/达成/协议/。（正确：他们/同/上海/、/沈阳/方面/达成/协议/。）
  - 德/比赛/的/战况/如下.....（正确：德比赛/的/战况/如下.....）

# 三个问题——分词的问题（续）

- 修正规则举例
  - $(n)/和/(prep) \rightarrow (n和)/(prep)$ （其中n为名词；prep为介词）
  - $(n和)/(n) \rightarrow (n)/和/(n)$ （同上）
  
- 未能处理的错误
  - 李文和于峥……（“于”可以作为介词也可以作为姓）
  - 李文和向荣华……（“向”可以作为介词也可以作为姓）
  
- 不同上下文
  - 尤文图斯/对抗AC米兰的比赛正在进行……（“尤文图斯”被识别成组织名称）
  - 尤文图斯站/在领奖台上……（“尤文图斯站”被识别成地名）

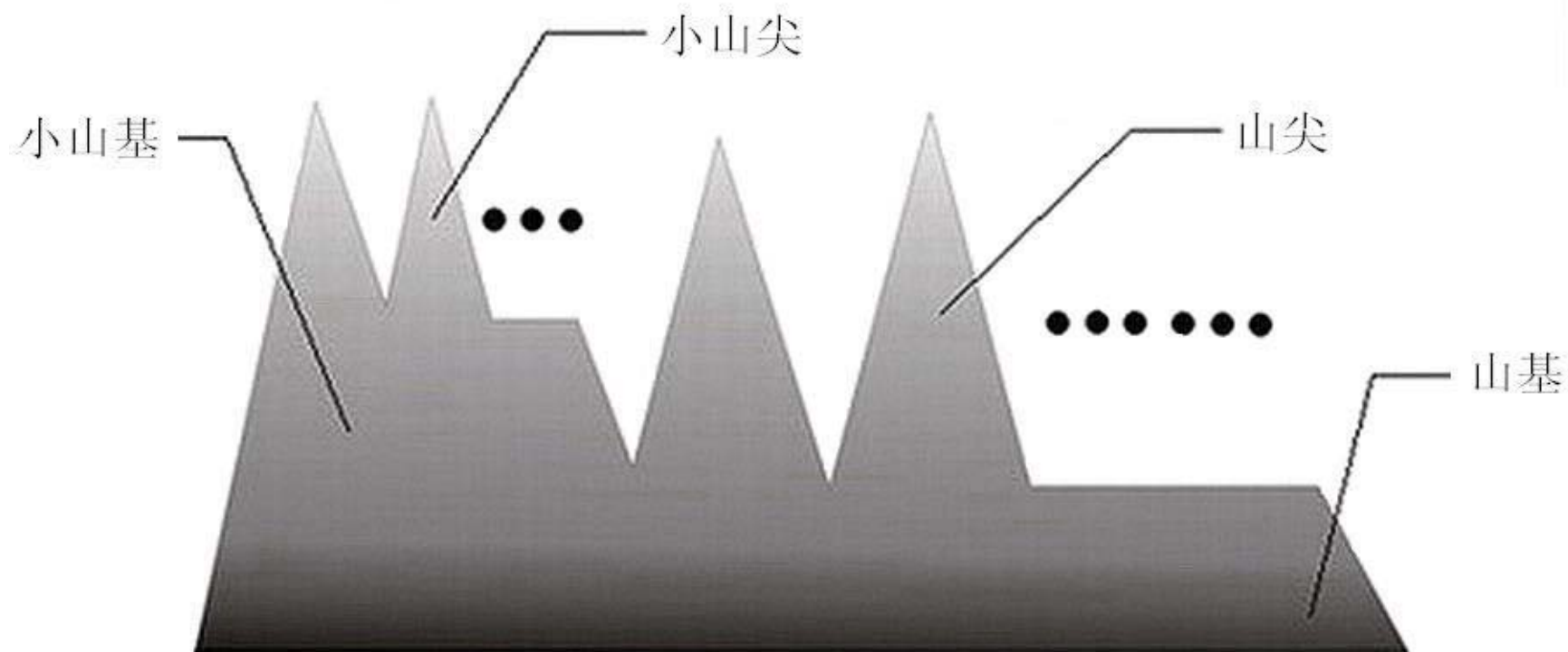
# 三个问题——领域的问题

## ■ 一词多义

- 在对大连中路无法打开局面的情况下，申花队只有采取边路进攻。（“大连中路”指大连方球场的中间区域）
- “……黄花忠魂，以励来兹”。（“黄花”指“黄花岗七十二烈士”）
- 老鹰大战雄鹿。（“老鹰”和“雄鹿”均为NBA队名；也可以均指动物）

## ■ 训练语料的选取

# 三个问题——领域的问题（续）



# 三个问题——方法的选择

## ■ 历史上

- 语言学派与统计学派
- 计算/语言/学

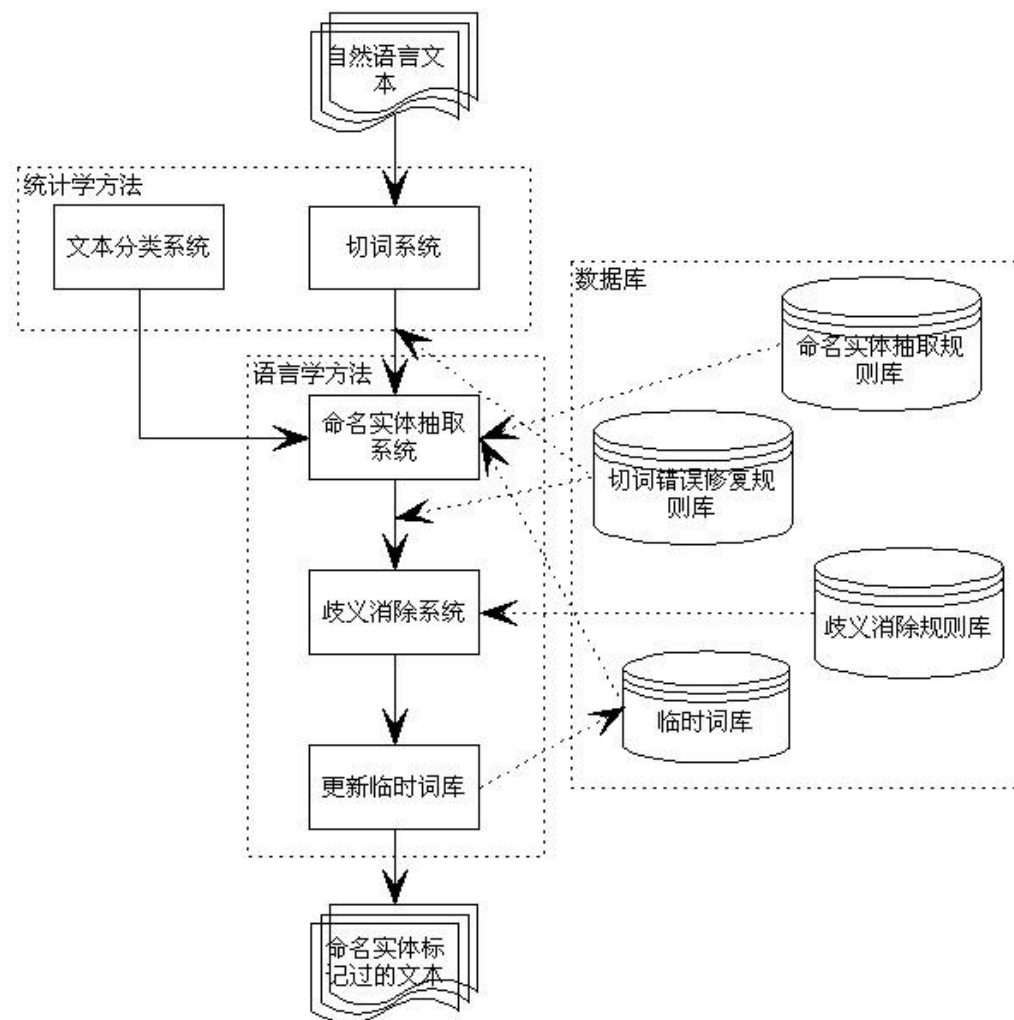
## ■ 人类认知世界

- 图象、声音、气味.....
- 词汇、上下文信息、生活经验、别人的建议.....

# 三个问题——方法的选择（续）

- 人名：汉字在姓和名中出现的概率+上下文规则
- 地名：地名库+上下文规则
- 组织名：组织机构名、地名库+已经识别出的地名、人名+上下文规则
- 消除歧义
  - 为规则设置权值
  - 举例：
    - 里昂是法国的一个城市。
    - 里昂是一支实力不凡的球队。
    - 里昂是法国里昂的一支实力不凡的球队。

# 系统的架构与实现



# 系统的架构与实现（续）

## ■ 规则举例

- 0100 013 50#T:ns @1/T:n @1/W:俱乐部 @1

## ■ 规则说明

- 0100是分领域编号，位数越往后分类越细致，这里是体育领域；
- 013是领域内的编号；
- 50是规则的权值；
- #是分隔符，后面表示正式的规则；
- :也是分隔符，后面表示初步识别出的不同类型，这里ns表示地名；T表示词类；
- @n表示n个分词单位；
- W表示具体的词。

# 测试结果与分析——评价指标

## ■ 召回率 (Recall)

$$Recall = \frac{Correct}{Require}$$

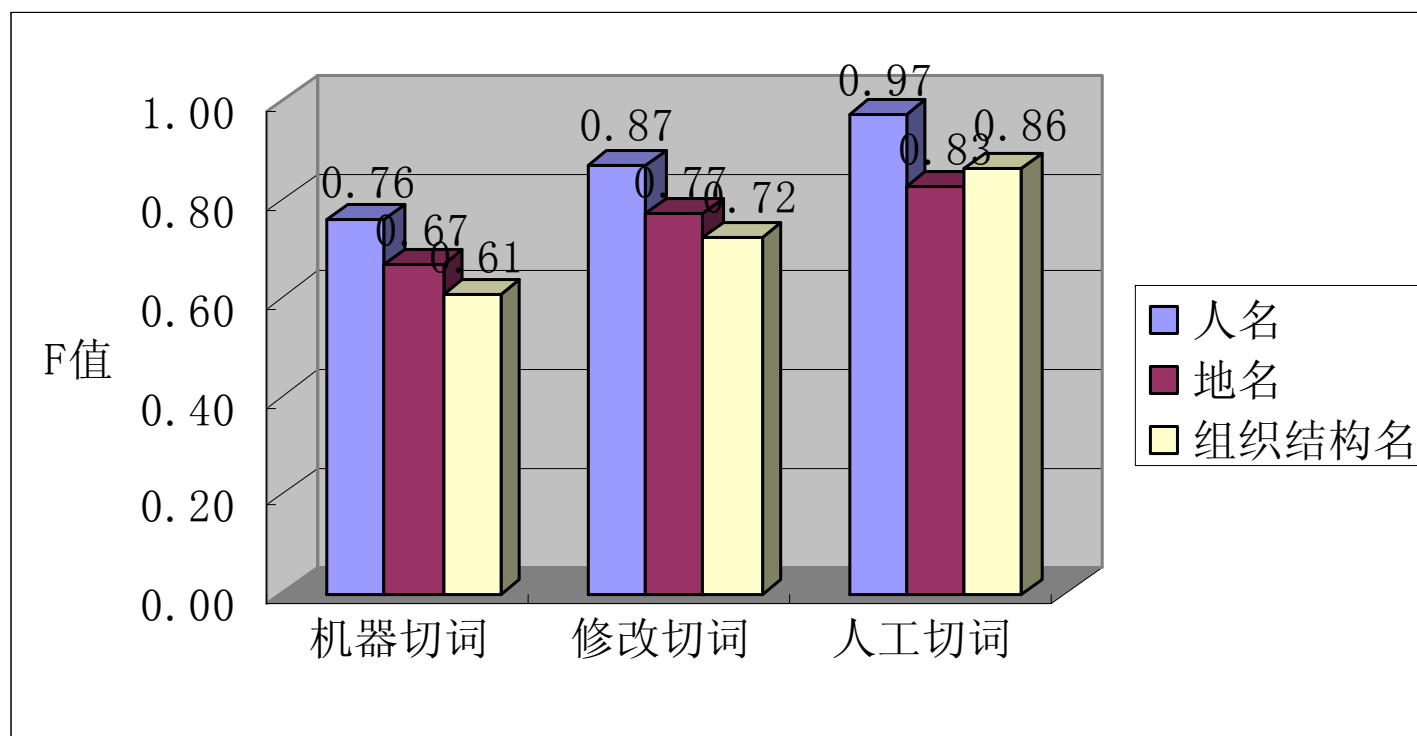
## ■ 准确率 (Precision)

$$Precision = \frac{Correct}{Fact}$$

## ■ F值 (F-Measure)

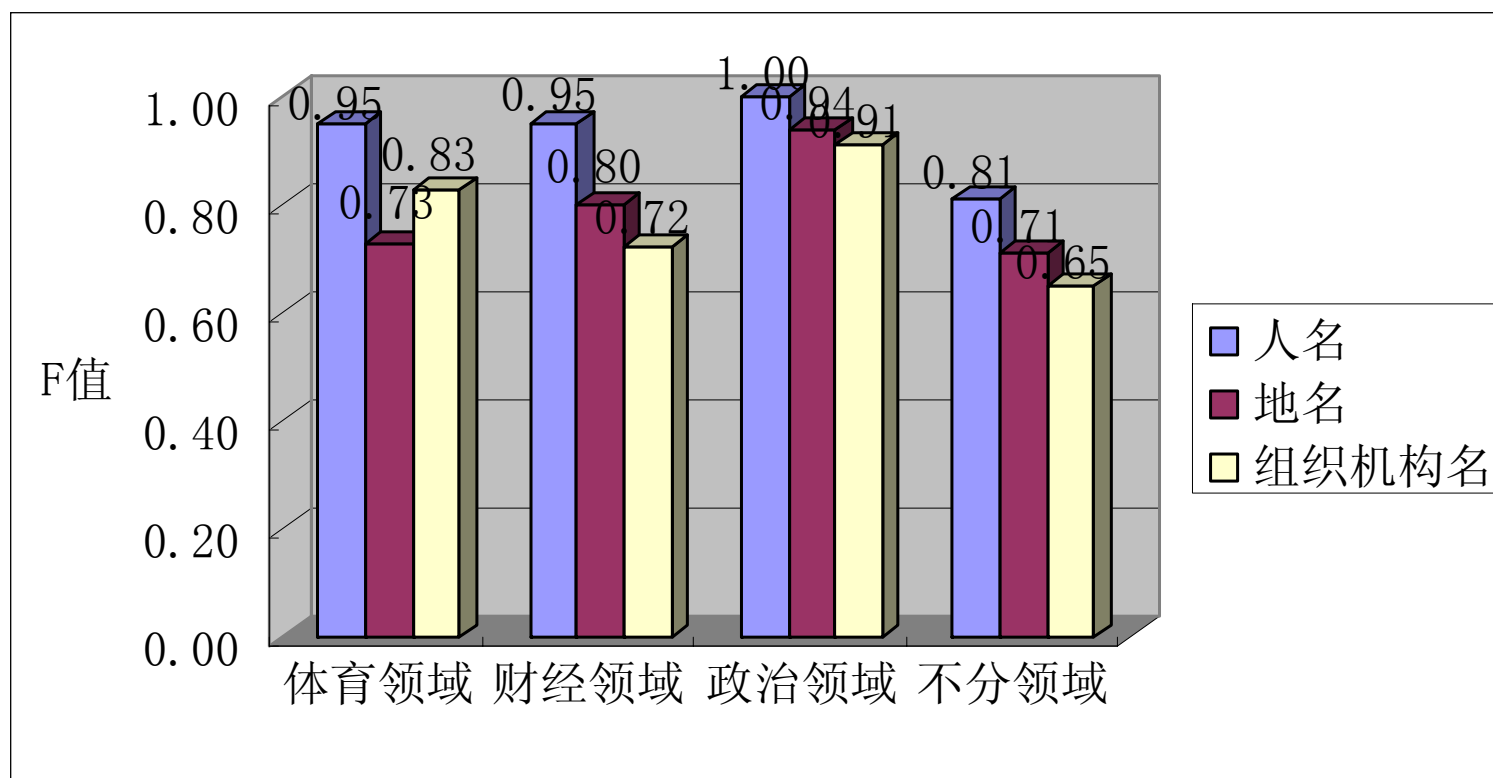
$$F - Measure = \frac{(\beta^2 + 1.0) \times Precision \times Recall}{\beta^2 \times Precision + Recall}$$

# 测试结果与分析——分词实验



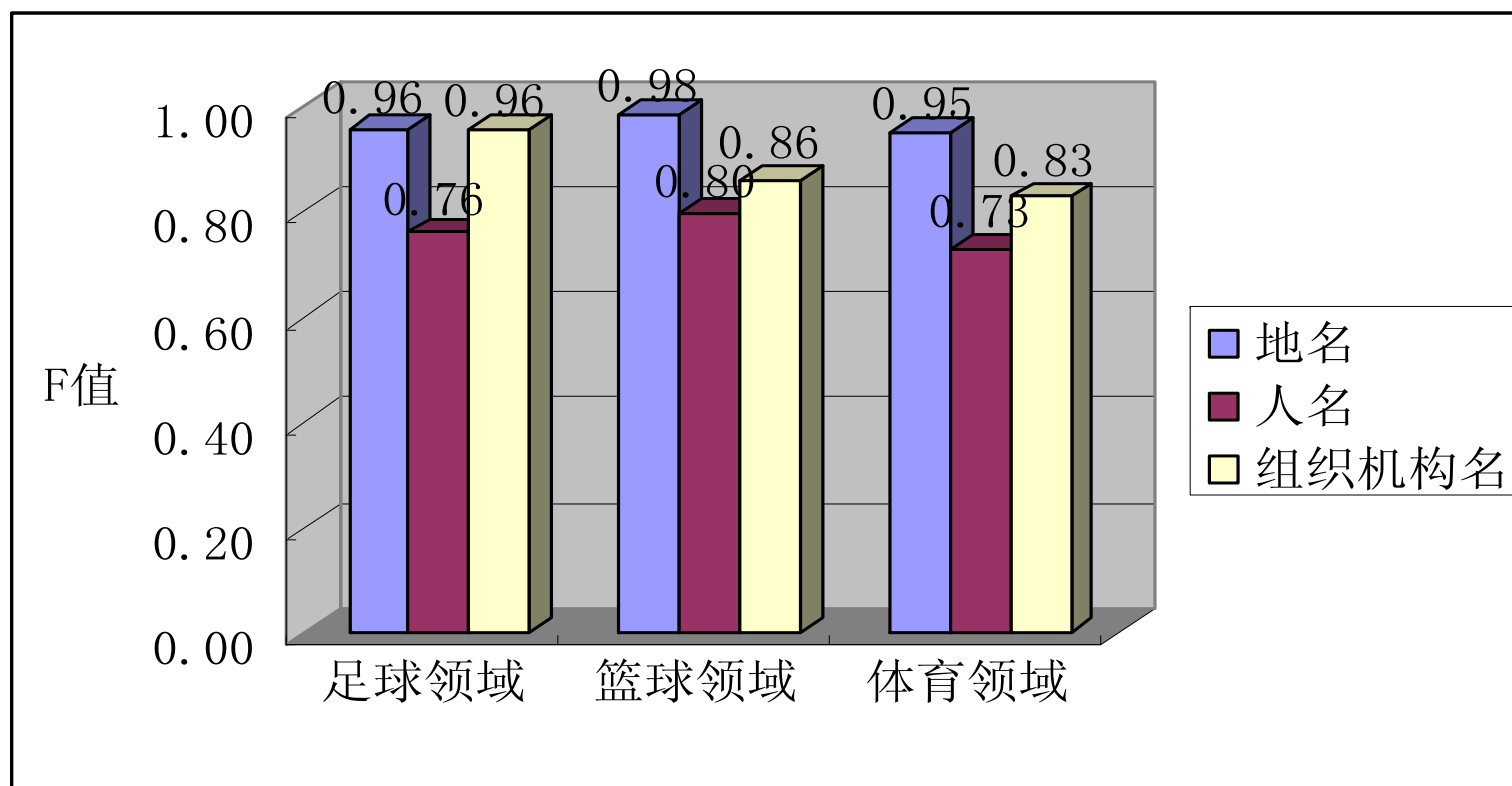
- 凯日/nr 曼妙/a 传/n 古德约翰逊/nr (正确: 凯日曼/nr 妙/a 传/v 古德约翰逊/nr)
- 年薪/n 加/j 奖金/n 达/v 340万/m 美元/q (正确: “加”不应该是缩写词, 而是动词v)

# 测试结果与分析——领域实验



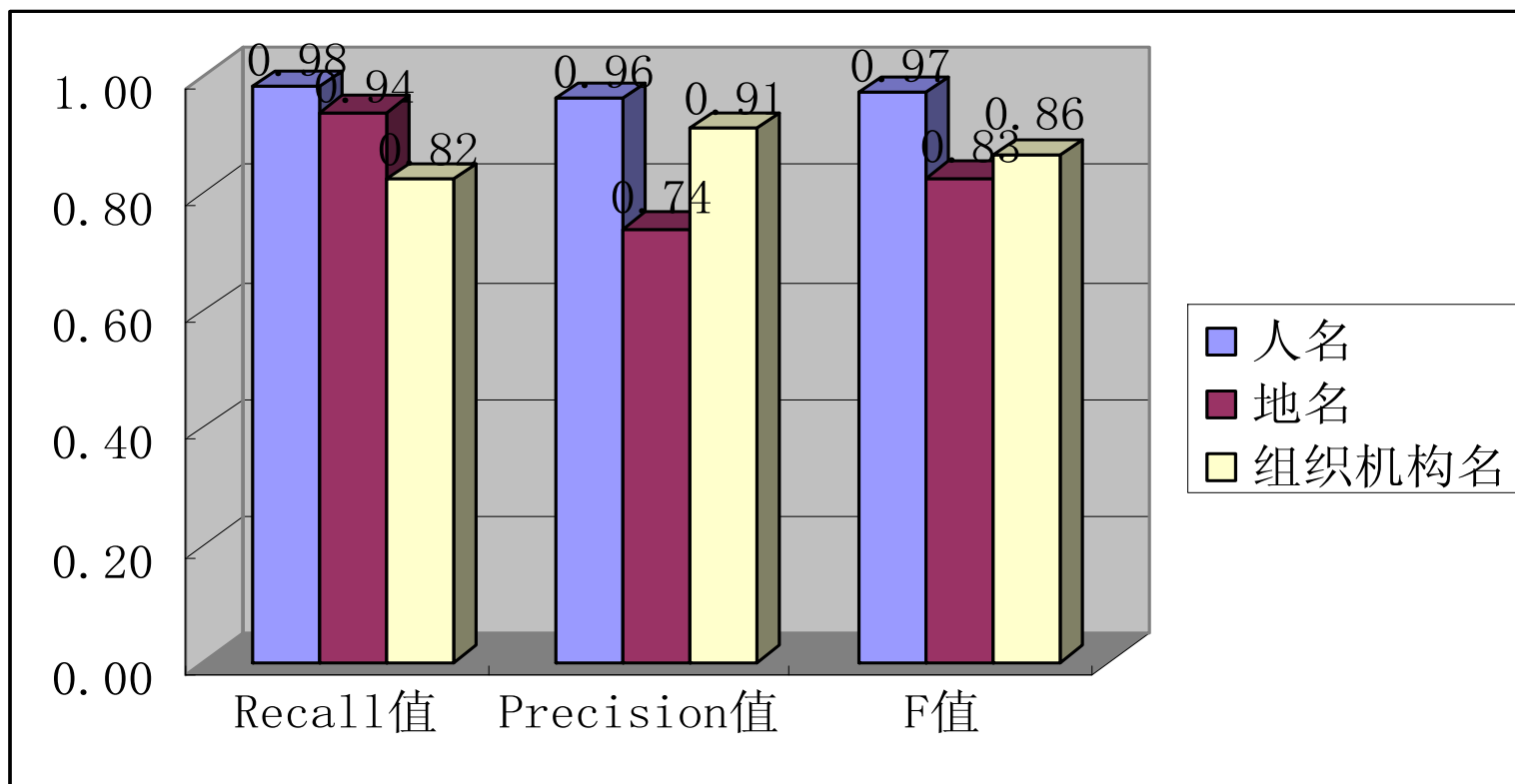
■ 说明：基于人工分词

# 测试结果与分析——领域实验



■ 说明：基于人工分词

# 测试结果与分析——总体实验



■ 评价：人名比较好，地名和组织名相对差

# 结论与未来的工作

- 分词的错误将直接影响到命名实体抽取的效果；
- 分词与命名实体抽取结合是有助于提升两者的效果的；
- 确定领域，应用“群山”模型结构的规则库是有助于提升命名实体抽取效果的；
- 方法不应当单一；
- 混合应用统计学和语言学的方法可以提升效果。

## 结论与未来的工作（续）

- 一是分词与命名实体抽取交替迭代式地进行如何实现；
- 二是在抽取效果与领域细致程度中寻求一个平衡点；
- 三是对不同方法的不同组合进行探索和讨论。

# 致谢

- 感谢姚老师的指导与同组同学的帮助和支持!
- 谢谢大家!

