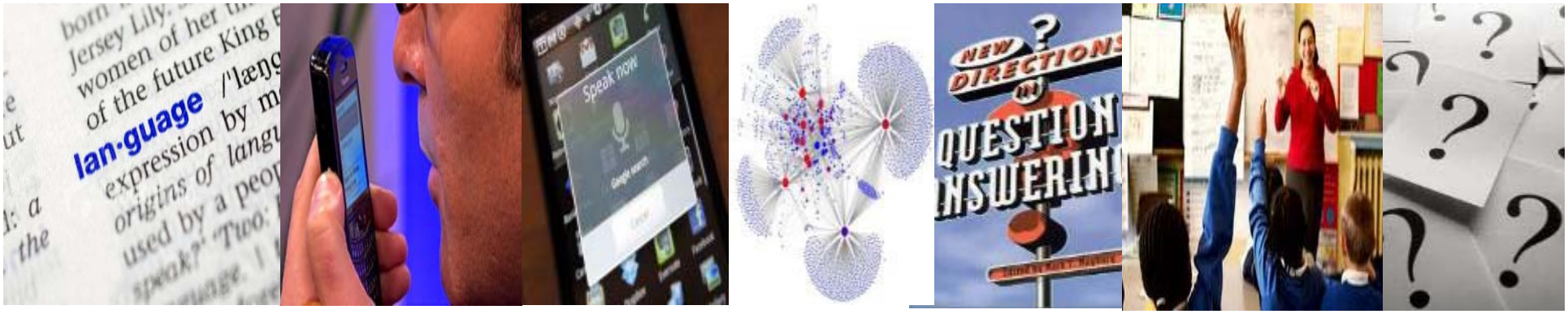



# 中文自然语言处理平台

# FudanNLP：从词法到句法，再到语义

报告人：邱锡鹏

xpqiu@fudan.edu.cn

http://jkx.fudan.edu.cn/~xpqiu/

FudanNLP

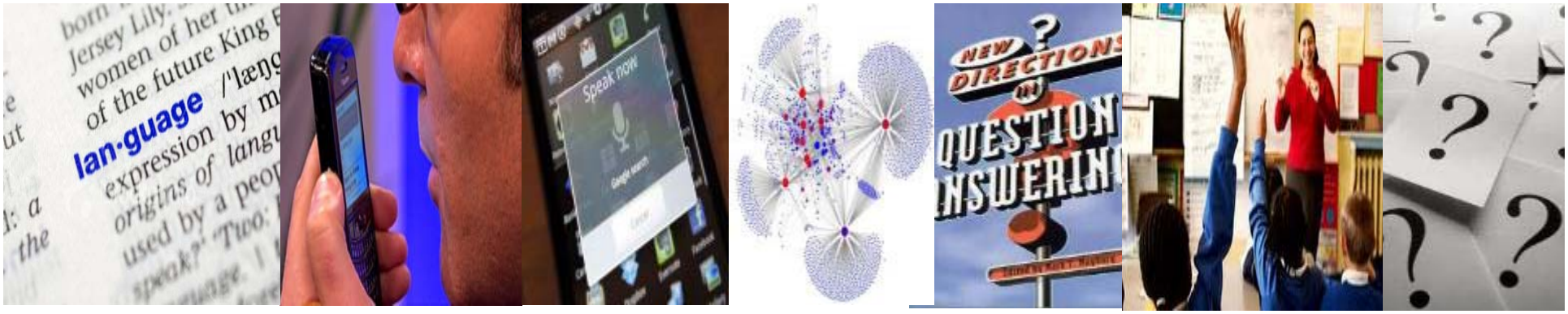
复旦自然语言处理

Java

LGPL v3 Free Software

# 提纲

1. FudanNLP系统介绍
2. 算法原理
3. 词法、句法、语义




# FudanNLP系统介绍

# 设计目标

- 为中文自然语言处理研发一个开源平台，使用统一框架，集成先进研究成果，降低中文分析门槛，促进中文自然语言处理的发展。
  - 算法
  - 数据集
  - LGPL3.0

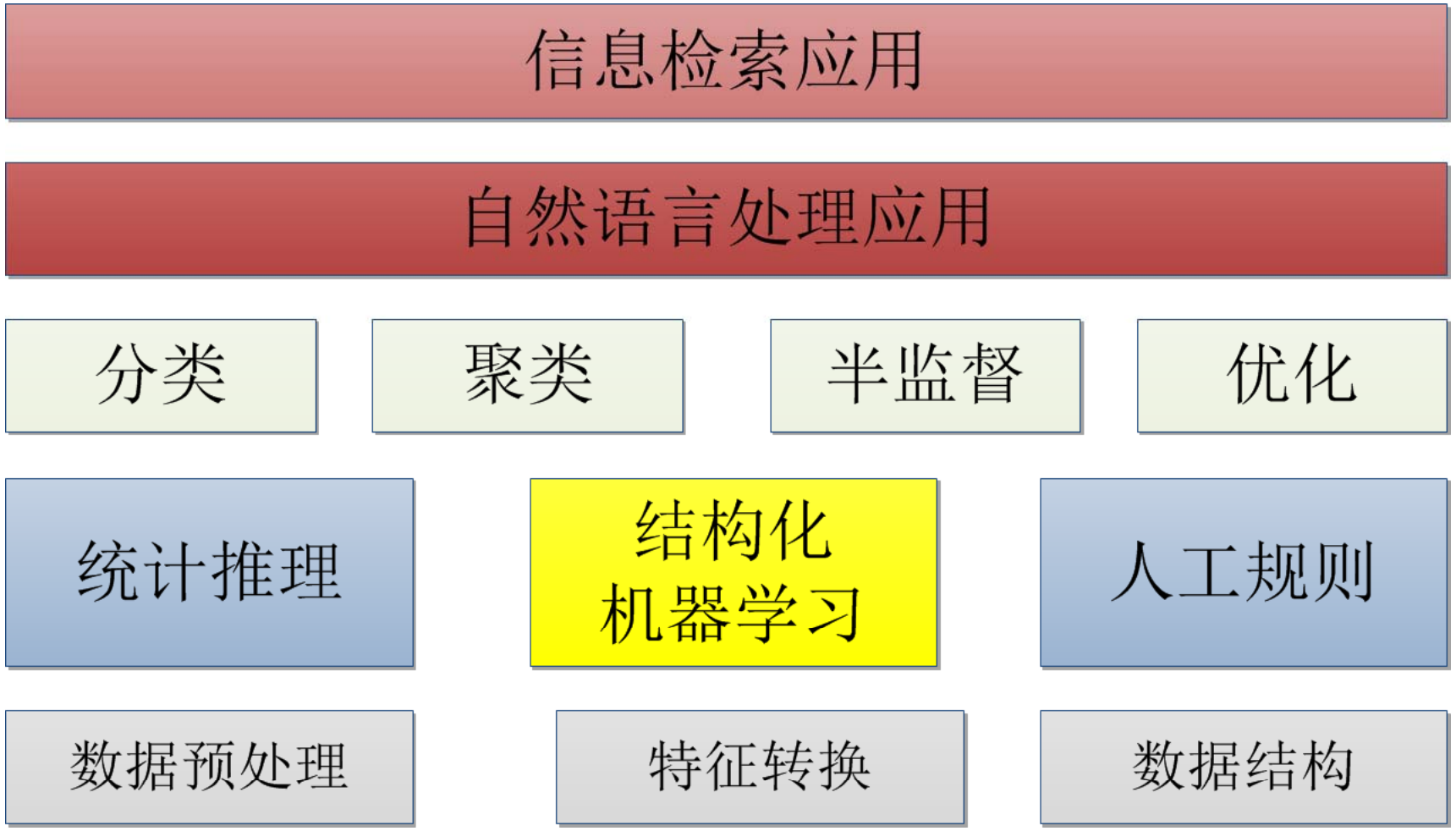
# 目前主要功能

- 中文自然语言处理
  - 中文分词
  - 词性标注
  - 实体名识别
  - 句法分析
  - 指代消解


FudanNLP

复旦自然语言处理

# FudanNLP框架


信息检索应用
自然语言处理应用
分类
聚类
半监督
优化
统计推理
结构化
机器学习
人工规则
数据预处理
特征转换
数据结构

# 研发路线图

- 2012.10(?) 发布1.5版，改进句法分析性能，增加指代消解模块，大幅改进新词识别率，增加详细的文档

- 2011.10.14 发布FudanNLP1.05版，增加程序注释，修正一些bug，支持并行化， 支持自定义词典，高速关键词抽取等
- 2011.8.1 发布FudanNLP 1.0版 （速度更快，内存占有更少）
- 2011.1.20 发布FudanNLP WebServices版
- 2010.12.22 发布FudanNLP 0.95版
- 2010.06.28 发布FudanNLP 0.8版

# 使用信息

信息来源: https://www.google.com/analytics

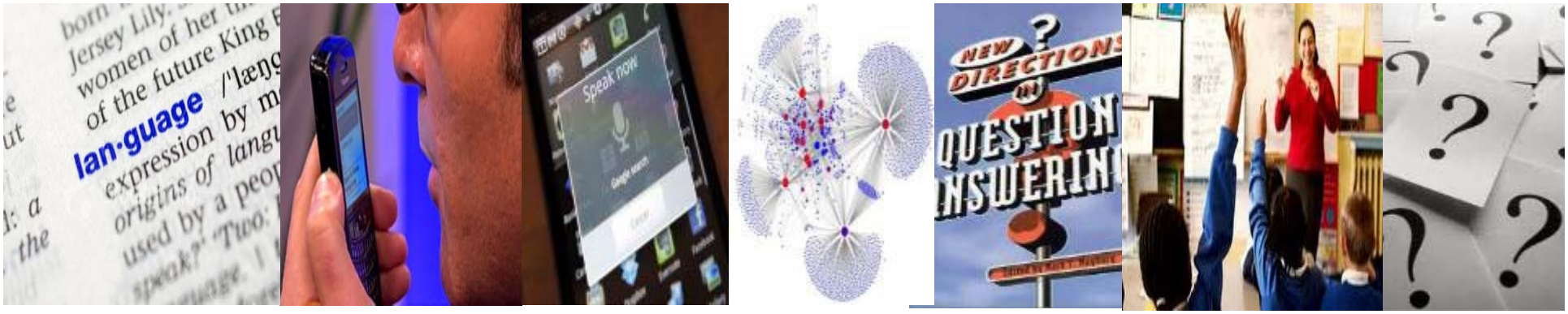
## Google Code

| 国家/地区 | 访问次数 | 平均访问持续时间 |
|---|---|---|
| China | 14,331 | 00:04:45 |
| United States | 508 | 00:03:28 |
| Taiwan | 279 | 00:02:50 |
| Hong Kong | 208 | 00:05:32 |
| Japan | 144 | 00:03:39 |
| Singapore | 61 | 00:05:19 |
| South Korea | 49 | 00:03:07 |
| New Zealand | 38 | 00:08:13 |
| France | 37 | 00:09:09 |
| Germany | 32 | 00:03:54 |

## FudanNLP WS

平均每日访问次数3000+

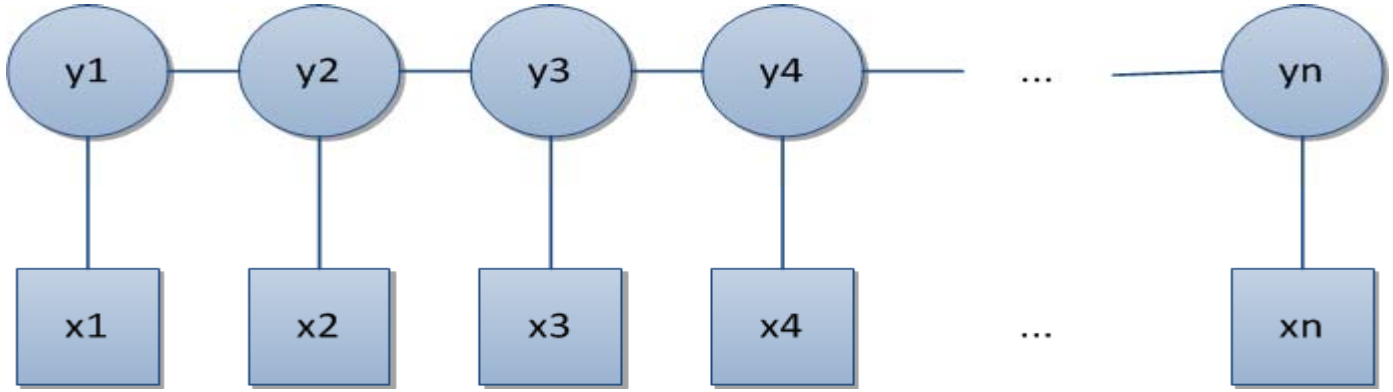
工具包累计下载7000+次




# 主要算法

# 结构化学习

- 在结构化学习中，预测不再局限在一个数，而可以是复杂的结构化对象，比如一副图像，一个标签序列，或是分析树等。




$$\overline{Y} = \underset{Y}{\operatorname{argmax}} f(\varphi(X, Y), W)$$

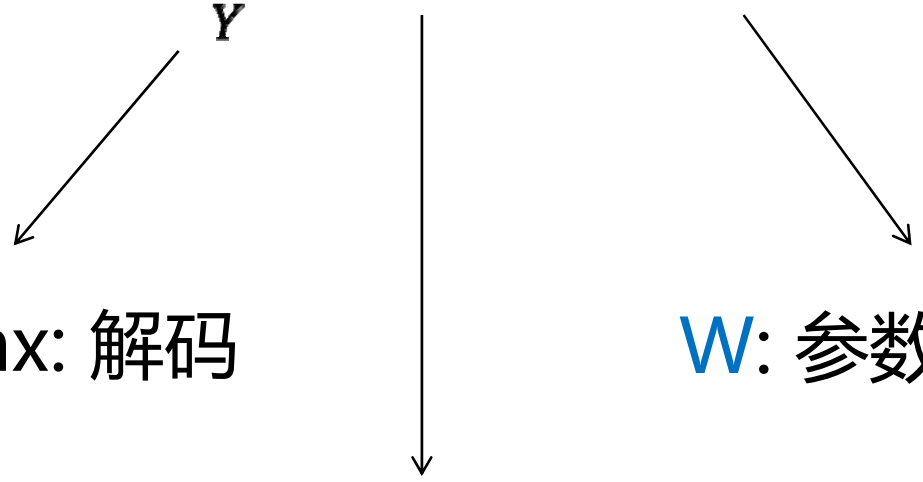
C.M. Bishop, Pattern recognition and machine learning, Springer New York, 2006.

# 结构化学习的三个基本问题

- 模型表示
  - 定义一个模型，将自然语言处理任务转换为结构化学习问题。
- 解码问题
  - 给定一个模型，计算最可能的解。
- 参数估计
  - 给定训练语料，学习模型参数。

# 结构化学习形式化表示

$$\overline{Y} = \underset{Y}{\operatorname{argmax}}\, f(\varphi(X, Y), W)$$



argmax: 解码

W: 参数估计

$f$, $\varphi$：模型表示

# 结构化学习的表示问题

- 将自然语言处理任务转换为结构化学习问题。
- 定义合适的目标函数。

Hierarchical Multi-Class Text Categorization with Global Margin Maximization (ACL 2009)
Hierarchical Text Classification with Latent Concepts (ACL 2011)

# 结构化学习的参数估计问题

- 最大似然估计
  - 条件随机场（CRF）、隐马尔可夫模型（HMM）
- 最大边际距离
  - PA算法、$M^3N$
- 在线学习
  - Perceptron

$$\overline{Y} = \underset{Y}{\operatorname{argmax}} f(\varphi(X, Y), W)$$

Labelwise Margin Maximization for Sequence Labeling (CICLING 2011)

# 结构化学习的解码问题

- 精确推理
  - Max-Sum算法: Belief Propagation, Viterbi
- 近似推理
  - 优化
    - 线性规划
  - MCMC, Loopy Belief Propagation, Expectation Propagation
  - 基于动态特征的Viterbi算法

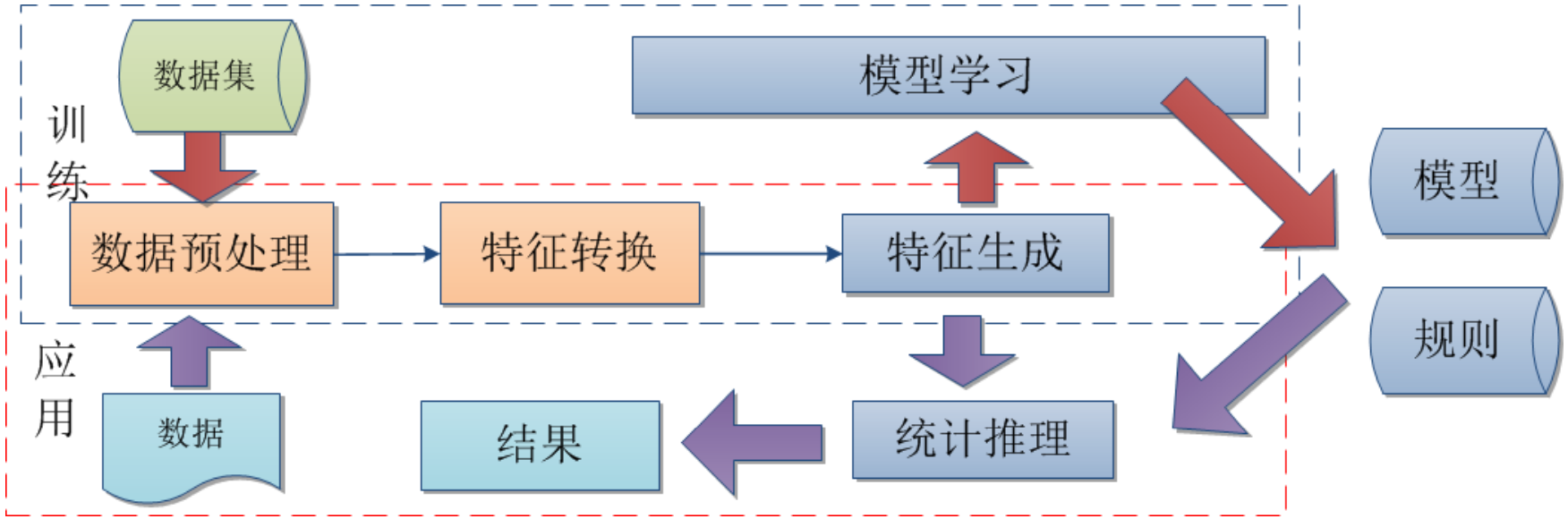
$$\overline{Y} = \underset{Y}{\operatorname{argmax}}\, f(\varphi(X, Y), W)$$

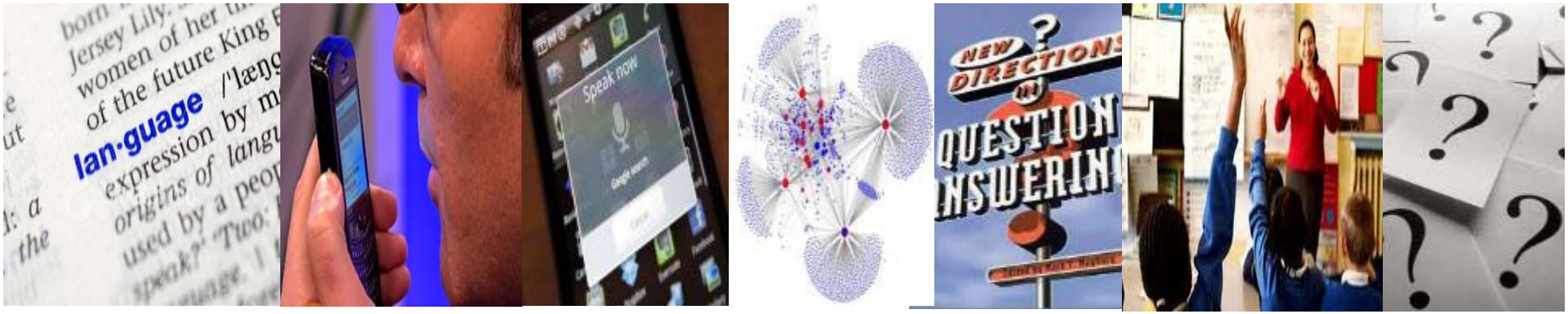
Part-of-Speech Tagging for Chinese-English Mixed Texts with Dynamic Features (EMNLP 2012)

# 性能优化

- 性能问题
  - 高维特征数量
    - 词性标注中特征数：115M，非零参数数量：1.5M
  - 解码问题
    - 状态数多
      - 词性标注中有148个状态
  - 人工规则
    - 结合机制
  - 容错性

# 系统流程






# 词法、句法、语义

# 词法难点：标准不明确+新词

- 《信息处理用现代汉语分词规范 GB/T 13715-92》
  - 词：最小的能独立运用的语言单位。
  - 汉语分词： 从信息处理需要出发，按照特定的规范，对汉语按分词单位进行划分的过程。
  - 基本原则：结合紧密、使用稳定
- 例子：
  - /吃饭/　　/吃/鱼/
  - /不管三七二十一/
  - /由此可见/
  - 自干五

# 句法分析难点：更灵活

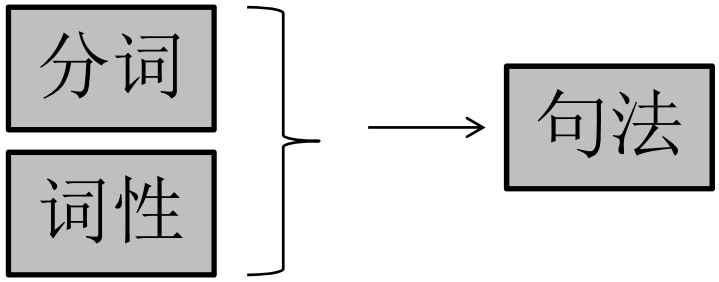
- 语法和词性不同于欧洲语言
- 大量省略成分
- 与语义紧密结合

# 主流方法

- 流水线方式：

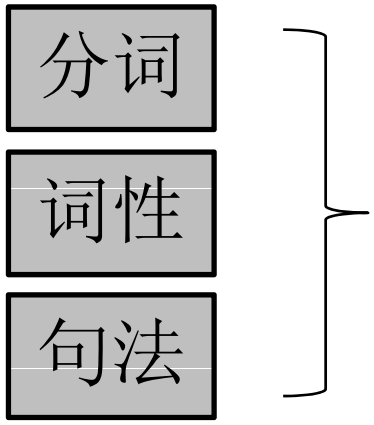

√

- 联合方法：




√√

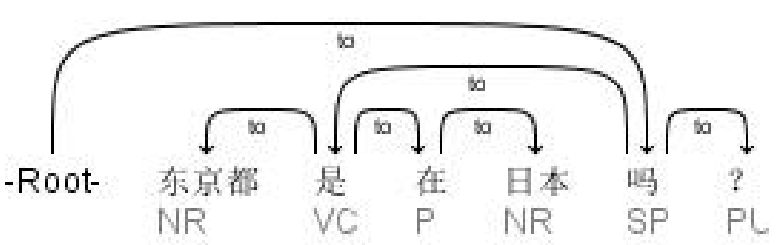
如何进一步提高？

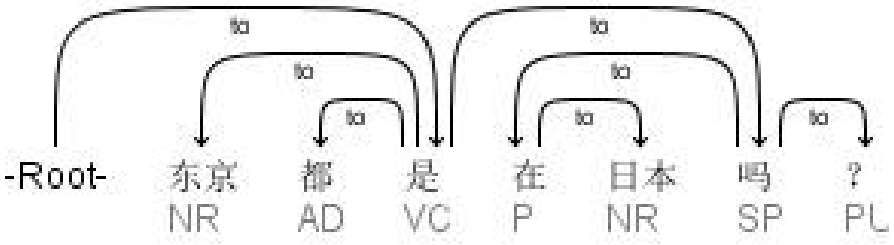



√√√

# 引入语义信息

东京都是在日本吗？




都：词义
都是：单复数
东京都：实体名

# 怎样结合语义信息到统计模型？

- 语义信息
  - 人工语义，需要语言学专家（代价高）
  - 众包
  - 自动抽取的含噪声语义
- 马尔科夫逻辑网
  - 将概率图模型和一阶逻辑结合
  - 概率图模型能有效地应对不确定性
  - 一阶逻辑容易表达知识。

Discovering Logical Knowledge for Deep Question Answering, (CIKM, 2012)
Recognizing Inference in Texts with Markov Logic Networks, (ACM TALIP 2012)

# 相关论文

1. Sparse Higher Order Conditional Random Fields for improved sequence labeling (ICML 2009)
2. Hierarchical Multi-Class Text Categorization with Global Margin Maximization (ACL 2009)
3. Detecting Hedge Cues and their Scopes with Average Perceptron (CONLL 2010)
4. 2D Trie for fast parsing (COLING 2010)
5. Joint Training and Decoding Using Virtual Nodes for Cascaded Segmentation and Tagging Tasks ( EMNLP 2010)
6. Hierarchical Text Classification with Latent Concepts (ACL 2011)
7. Labelwise Margin Maximization for Sequence Labeling (CICLING 2011)
8. Fusion of Multiple Features and Supervised Learning for Chinese OOV Term Detection and POS Guessing (IJCAI 2011)
9. Part-of-Speech Tagging for Chinese-English Mixed Texts with Dynamic Features (EMNLP 2012)
10. Discovering Logical Knowledge for Deep Question Answering, (CIKM 2012)
11. Recognizing Inference in Texts with Markov Logic Networks, (ACM TALIP 2012)

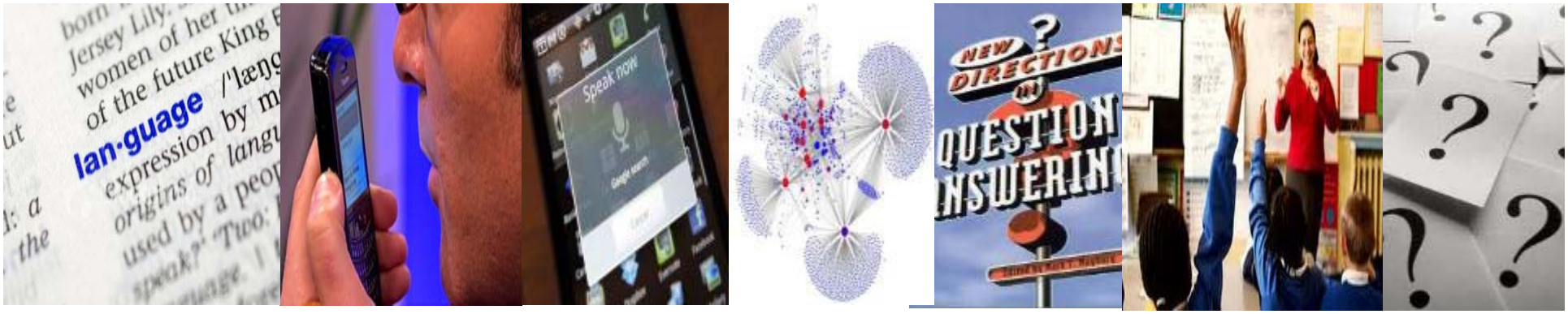
# 下一步工作

- 自然语言处理
  - 增加语义分析
  - 语义获取
  - Web Services
  - 文档
  - 众包

# 致谢

- **项目负责**
  - 邱锡鹏、黄萱菁
- **当前开发人员**
  - 曹零 (2010-)
  - 赵嘉亿 (2010-)
  - 田乐 (2010-)
  - 刘昭 ( 2010- )
- **过去开发人员**
  - 计峰 (2009-2012)
  - 高文君 (2009-2010)
  - 缪有栋 (2009-2010)
  - 沈超 (2009)

希望有兴趣的老师、同学一起参与开发！




# 谢谢

Q&A