Chinese Corpus Linguistics: What Does the Future Hold? Panel Presentation by Maosong Sun

Department of Computer Science & Technology
Tsinghua University

Round-table Conference on Linguistic Corpus and Corpus Linguistics, Hong Kong, May 6-8, 2011

一、目标：面向Web的中文计算

- 要在此目标下研究各个层面的中文自动分析方法

  比谷 歌 地球 更 牛 的 3D 地图

  越狱 最后 一 集 了， 好 舍不得舍 不得 啊！

  对 药家 鑫案 判决 的 立场 和 意见

  别克 大军 去 往辰山植物园 的 路上

  想 叫 个外 卖 都 不 容易， 打 了 几个 电话 都 不 送

- Stanford Parser 句法分析任重道远
- 当务之急：面向Web的汉语分词和词性标注系统 (Web条件下数据稀疏，可信计算)

二、手段：基于Web的中文计算

- 将Web看做一个几乎无限大的语料库、资源库的大一统(百衲衣=>袈裟; +学贯中西)
- 思考之一：基于自然标注Web资源的自然语言处理（资源建设，算法设计）

【例】中文Ontology的自动扩充　（可与英文资源相互印证）

乔布斯本人宣称iPad是一种全新种类的产品

iPad是一种娱乐加办公的时尚潮流产品

ipad是一种简易的手持设备

iPad是一种更浸入式的设备

iPad是一种混合设备

iPad是一种全新类型的电脑

iPad是一种触摸屏平板电脑

二、手段：基于Web的中文计算

- 思考之二：机器与人工相结合的策略

【例】汉语句法分析

人工标注百万级的常用结构（二八定律）；

基于实例的汉语句法分析（“不变”与“变”）

替补门将平托被红牌罚下

梅西被红牌罚下

守门员被红牌罚下

马佩佩恶意蹬踏对手被红牌罚下

英格兰门将格林被红牌罚下

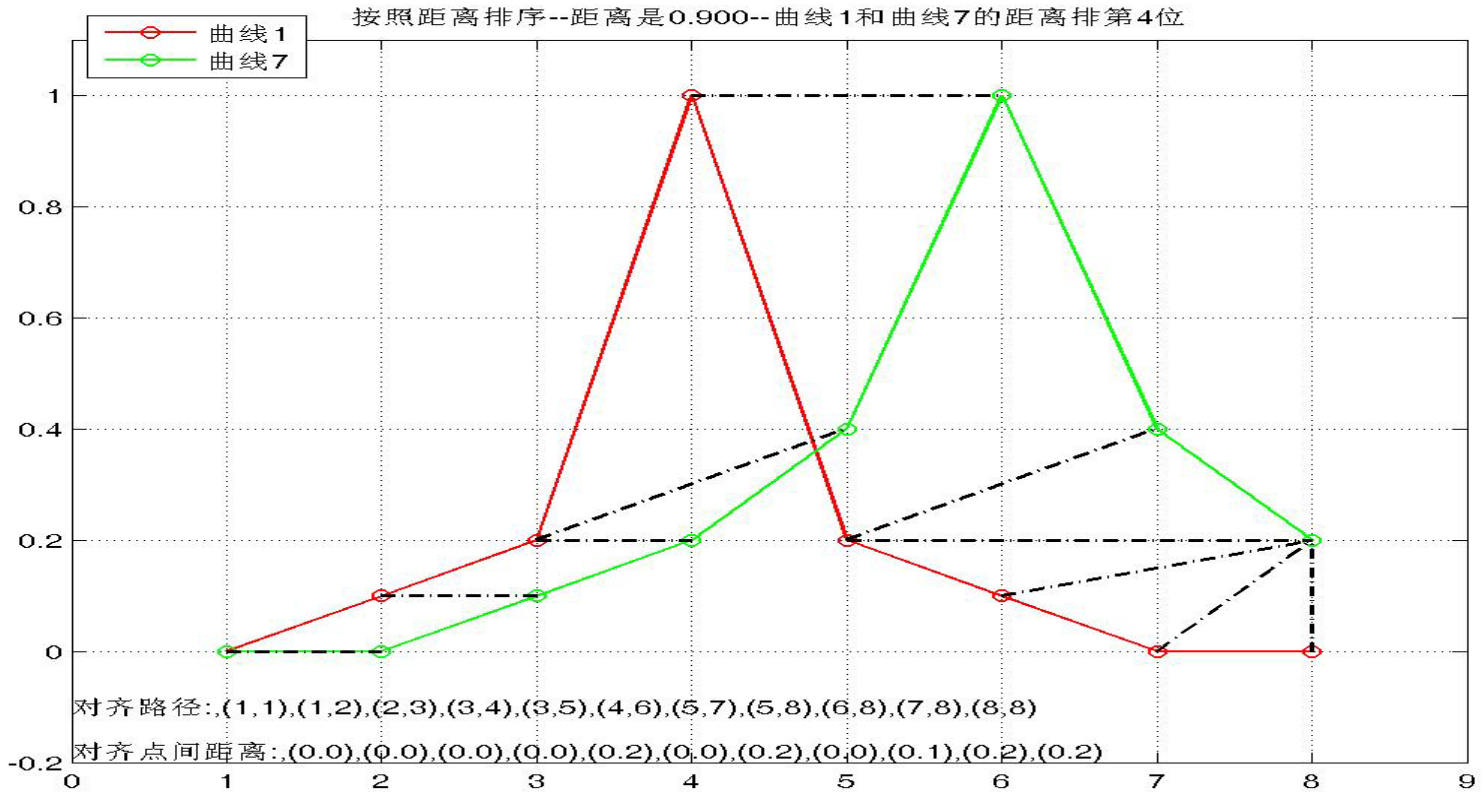
=>

常用模式+统计+规则“三位一体”的汉语句法分析?

三、必用其利器而善其事

- 计算尺vs.电脑　电脑vs.云（并行）计算

【例】热点词发现：对query日志的自动分类



几个基本点

（1）基本任务：**We should develop Chinese NLP technologies of the Web, for the Web and by the Web.**

（2） 基本原则之一

Prabhakar Raghavan：shallow knowledge, but massive scale

（3）基本原则之二

钱钟书: 能够帮助人的电脑，需要人的更多帮助

Tim Berners-Lee: Social machine(WWW2008 in Beijing)

Jeff Bezos: Artificial Artificial Intelligence

（4）基本原则之三

系统的观点：去粗取精，去伪存真，由此及彼，由表及里

（5）基本计算框架

Web-scale数据+适当深度的Web-scale内容计算+形式化的Web-scale群体智慧+Web-scale机器学习+云计算环境

and Q&A