

语言分析小组 (LAGroup) 介绍

让计算机理解语言！

李正华（副教授）

苏州大学计算机学院

<http://hlt.suda.edu.cn/~zhli>

2018-6-6

个人简介

- 2002-2006-2008-2013：哈工大本硕博
 - 2006年保研第二名
 - 2008年保博
 - 2012年国家奖学金
- 2013-2016-至今：苏大讲师、副教授
- 热爱编程
 - 自2002年：C
 - 自2005年暑假：C++
 - 自2006年：Perl
 - 自2008年：Python

研究方向

- 分词（2014）
- 词性标注（2010）
- 句法分析（2007）
- 语义分析（2016）
- 词语关系挖掘（2016）

演示系统：句子分析平台

<http://hlt-la.suda.edu.cn>

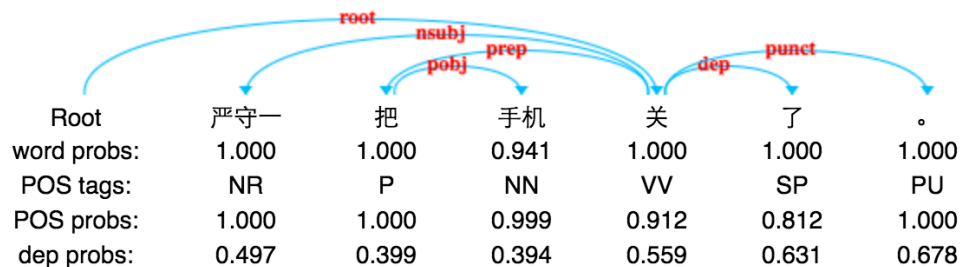
hlt-la.suda.edu.cn

语言分析

严守一/NR 把/P 手机关了。

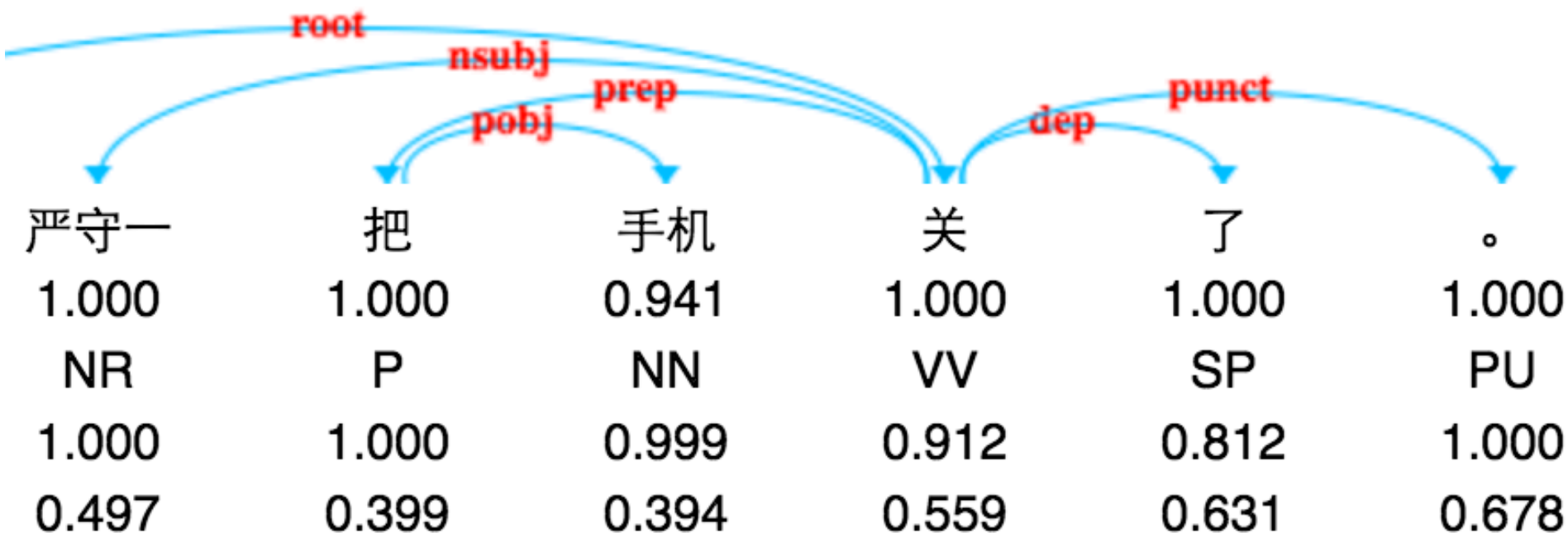
分析

清除



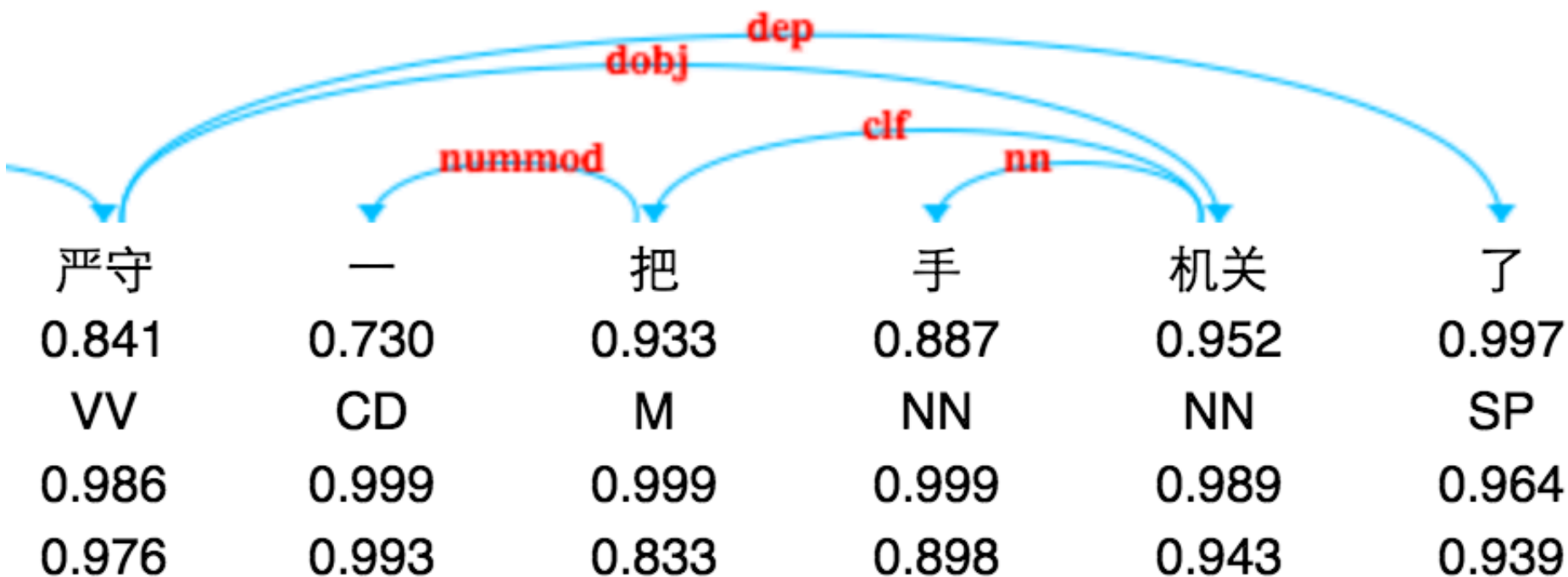
演示系统：句子分析平台

<http://hlt-la.suda.edu.cn>



词法、句法、语义分析

- 看起来比较**简单**，甚至**枯燥**？
- 其实**不简单**



词法、句法、语义分析

- 看起来比较简单，甚至枯燥？
- 其实不简单

王冕死了父亲



词法、句法、语义分析

- 看起来比较简单，甚至枯燥？
- 其实不简单

用毒毒毒蛇，
毒蛇会不会被毒毒死



词法、句法、语义分析

- 看起来比较简单，甚至枯燥？
- 其实不简单

冬天能穿多少穿多少，
夏天能穿多少穿多少



词法、句法、语义分析

- 涵盖了机器学习的方方面面
 - 问题：分类、序列标注、树/图结构
 - 方法：最大熵，SVM，CRF、前馈神经网络，CNN，RNN，注意力机制，对抗学习
 - 算法：动态规划、柱搜索、贪心搜索、A*搜索
 - 学习方式：有/半/无监督、强化学习
- 打下坚实的机器学习、自然语言处理基础
- 其他自然语言处理任务：非常容易上手

苏州大学数据标注平台（SUDAP）

<http://101.132.166.249/anno-sys>

- 2014：初版（大三陆芳丽、王效静）
- 2015-2018：张月、严秋怡、朱运、黄德朋
- 2017：暑假大规模使用，30人在线标注
- 2018：暑假兼职50多人报名

看准了一件事，就要坚持做好！

标注界面

苏州大学自然语言处理标注工具

[deptest](#) / [注销](#)

deptest,您好! 您当前正在进行: dep_parsing 任务,content-search-500-10-15数据批次,第7个任务您已标注: 4个任务

ROOT 引领 多效 柔肤 时代 , 满足 你 对 bb 霜 的 各种 需求 !

重 做

提 交

标注界面

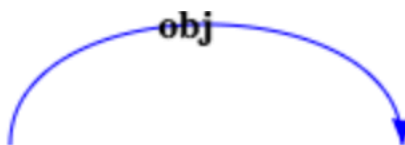
ROOT 引领 多效 柔肤 时代 ， 满足 你 对 bb 霜 的 各种 需求 ！

sbj
obj
pobj
att
adv

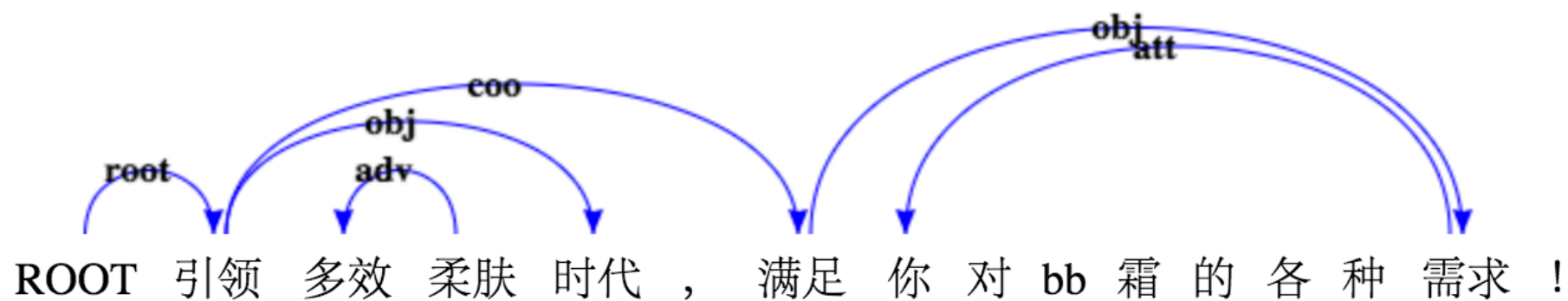
重 做

提 交

ROOT 引领 多效 柔肤 时代 ， 满足 你 对 bb 霜 的 各种 需求 ！



标注界面



重 做

提 交

科研论文（2014-至今）

- 顶级国际会议11篇
 - ACL-2014/**2015/2016/2018**
 - EMNLP-**2016/2017**
 - COLING-2014/**2016/2018**
 - AACL-**2018**
 - IJCNLP-**2017**
- 顶级国际期刊2篇
 - IEEE Transactions on Audio, Speech, and Language Processing 2014/**2017**
- 我的科研观
 - 珍惜自己的精力，做有价值、重要的研究
 - 不过分强调论文数量（论文只是科研产出的一种）

科研项目（2014-至今）

- 国家纵向项目
 - 2016-2018: NSFC青年基金 句法分析 主持 20万
- 企业横向项目（前瞻性科研）
 - 2016: 腾讯犀牛鸟 语义分析 主持 12万
 - 2016: 百度NLP组 句法分析 主持 50万
 - 2017: 阿里NLP组 数据标注 主持子课题 50万
 - ...

博士生

- 2017级
 - 李英（昆明理工考博）
 - 研究方向：句法分析、领域移植
- 2018级
 - 龚晨（苏大保研、直博）
 - EMNLP-2017（第一作者），国家奖学金
 - 研究方向：结构化分析
 - 夏庆荣（苏大保研软工第一、直博）
 - 研究方向：语义分析

硕士生（2014级）

- 巢佳媛
 - 苏大考研
 - 微软（北京）实习半年
 - 阿里巴巴（杭州）就职
 - 合作发表3篇顶级文章
 - ACL-2015, EMNLP-2016, IEEE Trans ASLP-2017
 - 校优秀硕士论文

硕士生（2015级）

- 陈伟
 - 南阳理工考研，专硕
 - 爱奇艺（北京）实习半年；转正
- 凡子威
 - 滁州学院考研，专硕
 - 科大讯飞（北京）实习5个月；搜狗（北京）工作
- 张月
 - 苏大保研（信管第一），学硕
 - 阿里巴巴（杭州）实习4个月；转正
 - 合作发表两篇顶级论文：ACL-2016/IJCNLP-2017

硕士生（2016级）

- 郭丽娟（江西财经保研）
 - 将去科沃斯机器人公司（苏州）实习
- 孙佳伟（北航）
 - 目前在搜狗（北京）实习
- 朱运（山西大学）
- 2位直博：龚晨、夏庆荣

硕士生（2017级）

- 黄德朋（苏科技）
 - 江心舟（苏大保研，计科第一）
 - 目前在百度自然语言处理组实习
 - 彭雪（山东农大）
 - 章波（苏大，计科第一）
-
- 顶级论文ACL-2018：江心舟、章波（共同第一作者）、李正华、张民、李生、司罗

硕士生（2018级）

- 蒋炜（苏大）
- 刘亚慧（山东农大）
- 陆凯华（苏大）
- 吴锴（浙江理工）
- 张宇（苏大）

硕士生 (2019级)

- 招收中 (~4位)
- 欢迎保研或考研同学邮件 (附简历和成绩单) 联系我。

学生指导和培养风格

- 做好本职工作：认真指导学生，不因为学生能力差就放弃
- 重视提高学生的基本功
 - Coding、NLP、ML
 - 第一学期有一个coding任务列表
- 做有价值的工作
 - 深入了解前沿，适度创新
 - 发表高水平论文
 - 实用系统

学生管理规则（2015.1起执行）

<http://hlt.suda.edu.cn/~zhli/LAGroupRules>

- 每周工作6天（每月一周双休，随时请假1天）
- 每天写日志，每周周志
- QQ签到，鼓励运动
- 和985高校的优秀学生相比，我们的学生在代码、英语、数学、沟通、眼界、独立思考等各方面，都有差距。如果不能在有限的几年里（专注科研的时间最多一年半）努力提高，打好NLP基础，这辈子估计也没机会缩小差距，毕业后会一直在IT企业底层卖苦力。

最后的话

- 硕士阶段，研究方向和指导老师最重要

985 一般课题组 前景一般的方向

远不如

苏大 好课题组 好方向 靠谱的老师

- 只要你愿意学习和沟通，一定可以打下坚实的NLP/ML基础，做出高水平成果
 - 公司会抢着要你
 - 为读博或直博做好充分准备

让计算机理解语言！

谢谢大家的聆听
欢迎提问（线下发邮件也可）

<http://hlt.suda.edu.cn/~zhli>

C最基础的内容

<http://open.163.com/special/opencourse/paradigms.html>

- 各种数据类型在内存中的存储
 - char char* int short float double
 - 数组，结构体等复杂类型
- 指针
 - 内存分配和释放
 - 链表（单向、双向、环形）
- 排序算法
 - 快速排序、堆排序
- ...