陈晓阳

自然语言处理 & 文本挖掘

联系方式

北京市海淀区
中关村南大街5号
北京理工大学
100081

+86 156 5299 5998

ling032x@gmail.com
http://ling0322.info
Twitter: @ling0322
Github: ling0322

编程语言

♥ C/C++(从高二开始)
Python
Java, C#
Js & CoffeeScript
HTML & CSS
Scala, Go, Rust

模型&算法

最大熵, 逻辑回归
感知器, SVM
HMM, LDA

库&框架

Hadoop
MongoDB, LevelDB
node.js, tornado
cedar, darts, opal
MaltParser, ZPar
nltk, word2vec

工具

♥ Sublime Text
Visual Studio
Eclipse, Vim
automake
gdb, valgrind
LaTeX

教育情况

2012至今 **硕士** – 北京理工大学
专业: 计算机科学与技术, 研究方向: 自然语言处理
主要课程: 计算语言学, 机器学习, 数据挖掘, Web智能与社会计算.

2008 – 2012 **学士** – 苏州科技学院
专业: 计算机科学与技术
主要课程: 数据结构, 操作系统, 编译原理, 网络原理, Java, C#

个人简介

WEBDICT中文词库计划发起者, 熟悉自然语言处理和文本挖掘, 熟悉常用机器学习的模型和算法. 兴趣是使用C++实现中文分词, 新词发现, 词性标注以及依存句法分析等自然语言处理基础工具.

实习经历

2014.5至今 **百度** 研发工程师
自然语言处理部门, 智能交互应用基础研究(HCI-BASE)小组, 参与的项目:
- 爱奇艺智能交互query改写
- 知识图谱中实体标签提取
- 使用机器翻译结合依存句法优化Query-Title匹配

2012.3 **豌豆实验室(豌豆荚)** 前端工程师
前端项目组实习生, 公司内部协作工具的编写与维护.

项目经历

2012-2014 **MilkCat中文依存句法分析以及相关工具包** 开源项目 – goo.gl/uN3MXX
开源中文依存句法分析工具包, C++编写, 提供C/Python/Go/Java语言接口.
- 分词模型: bigram & CRF, 2.5MB/s, F1 = 0.968
- 词性标注: TnT & CRF (OOV only), 2MB/s
- 依存分析: arc-eager, 200KB/s, UAS=0.801 (beam-search in dev)

2013 **Goldeye互联网数据挖掘软件** 实验室项目
网页新闻信息的数据挖掘平台(财经杂志已部署使用), 主要负责:
- 数据挖掘部分可视化展示(D3.js)
- 实体以及词语的关联分析以及趋势分析算法设计

2013-2014 **WEBDICT中文词库计划** 开源项目 – webdict.info
使用机器学习以及众包方式构建中文词库, 已收集到20GB纯文本数据以及22万词语.
- 网易&腾讯新闻、贴吧、Twitter中文圈定向爬虫
- 网站的设计(前端Boopstrap+jQuery, 后台RoR)

2013-2014 **新浪微博爬虫CoCoCat** 实验室开源项目 – goo.gl/5gbiHj
支持微博用户信息抽取, 好友与关注着抽取, 用户时间线抽取以及搜索结果的抽取.