2015 春季学期总结

韩喆

iampkuhz@gmail.com

PIE 组

2015 年 12 月 10 日

Outline

- 工作
  - 研究生课程
  - 实验室工作
    - 新华社新闻类别标注
    - 中文维基百科谓词归一化
    - 杂项
- 总结、展望

研究生课程

- 面向对象分析与设计
- 高等计算机体系结构
- 中国特色社会主义理论与实践研究
- 自然语言处理高级专题

- 高体不挂的话，就完成授课类课程要求的学分了
- 认真程度不及上学期
- OO 课件比较漂亮，和冯老师的课件有的拼
- 高体比较像考历史（必考 + 生僻知识点）
- 政治课都差不多，按时交作业就好
- NLP 讲的比较难，但作业比较简单

实验室工作

— 新华社新闻类别标注

新华社新闻类别标注

- 和曾颖 (网站搭建) 负责
- 上学期完成样例新闻标注
- 迭代增加功能

1. 完成一级类别、二级类别标注任务
2. 增加了多种新闻来源 新浪、中国新闻网、人民网、新华网、网易新闻、腾讯
3. 自动生成标注任务评价
   - 准确率: 标注语料随机混合部分已标注语料
   - 一致性: 不同标注者含有部分相同新闻，与其他人比较
4. 二级类别根据个人意愿半自动生成
   - 手工记录标注者擅长类别优先级，自动从已标注一级类别的新闻中选择给二级类别给标注者，控制不同人的重合新闻数量
5. (?) 总体觉得比新华社的样例数据更好 （一级类别新闻分布类似长尾）
   - 标注了 5.7w 个一级类别，1.9w 个 2 级类别

新华社新闻类别标注

- 登陆

新华社新闻类别标注

- 登陆
- 主页
  -

新华社新闻类别标注

- 登陆
- 主页
  - 公告

新华社新闻类别标注

- 登陆
- 主页
  - 公告
  - 反馈

新华社新闻类别标注

- 登陆
- 主页
  - 公告
  - 反馈
- 标注任务

新华社新闻类别标注

- 登陆
- 主页
  - 公告
  - 反馈
- 标注任务
- 标注新闻

新华社新闻类别标注

- 登陆
- 主页
  - 公告
  - 反馈
- 标注任务
- 标注新闻
- 管理员

新华社新闻类别标注

- 登陆
- 主页
  - 公告
  - 反馈
- 标注任务
- 标注新闻
- 管理员

实验室工作

— 中文维基百科谓词归一化

中文维基百科谓词归一化

张家輝

2010年8月24日參加電影《歧人》江蘇首映禮。

男演員

羅馬拼音 Cheung Ka Fai
英文名 Nick Cheung
國籍 中国（香港）
籍貫 廣東番禺
出生 1967年12月2日（47歲）
英屬香港
語言 粵語、英語、普通話
配偶 關詠荷（2003年至今）
兒女 張童（Brittany Cheung）
- 2006年01月24日（9歲）
活躍年代 1987年至今
經紀公司 鍾珍[1]
獎項 最佳男演員 – 上海國際電影節
2013年《激戰》– 程輝
最佳男演員 – 香港電影評論學會

Subject	Predicate	Object
張家輝	罗马拼音	Cheung Ka Fai
張家輝	英文名	Nick Cheung
張家輝	国籍	中国(香港)
張家輝	籍贯	广东番禺
張家輝	出生	1967年12月2日 (46岁)
張家輝	出生	英属香港
張家輝	语言	粤语
張家輝	语言	英语
張家輝	语言	普通话
張家輝	配偶	关咏荷(2003年至今)
張家輝	儿女	张童(Brittany Cheung)
張家輝	活跃年代	1989年至今
張家輝	经纪公司	钟珍

- 背景/回顾：基于中文维基百科的知识库：330w 三元组，1.6w 个谓词

中文维基百科谓词归一化

”邮政”

INSEE/邮政编码、ISO 3166-2 邮政简写、美国邮政编号、美国邮政编码、邮政、邮政代码、邮政信箱、邮政分区、邮政区号

- Motivation: 消除语义重复的谓词
- 问题转化：任给两个谓词，判断是否为相同语义
- 实验效果：在正负样本 1:1 的数据集上有监督训练，将近 70% 的准确率

中文维基百科谓词归一化

”邮政”

INSEE/邮政编码、ISO 3166-2 邮政简写、美国邮政编号、美国邮政编码、邮政、邮政代码、邮政信箱、邮政分区、邮政区号

- Motivation: 消除语义重复的谓词
- 问题转化：任给两个谓词，判断是否为相同语义
- 实验效果：在正负样本 1:1 的数据集上有监督训练，将近 70% 的准确率
- **进展**：投了 NLPCC2015，被拒。。

杂项

1. 提供了 NLPCC2015 测评的知识库
   - NLPCC2015 task: Entity Recognition and Linking in Chinese Search Queries
2. 大组 + 小组讨论班 (报告 4 次)

总结展望

Summary

上学期计划

- 每日备忘录
- 每周两篇论文
- 考托福

本学期完成

- 每日备忘录
- 维基百科谓词的归一化

暑期计划

- 完善中文维基百科知识库的体系
  - 借鉴 dbpedia，整合中英文维基百科、freebase
- 看一下有关的书：shell，Java
- 修改 NLPCC 的论文，继续投稿

下学期计划

- 尽量在开学时完成知识库体系构建，整理出一篇期刊
- 完成助教学分
- 百度百科的知识库提取和整合
- 动手做点小东西（展示知识库）

谢谢大家！