2014 秋季学期总结

韩喆

iampkuhz@gmail.com

2015 年 1 月 23 日

Outline

- 研究生课程
- 实验室工作
  - 新华社
  - 基于中文维基百科的知识库
- 总结和展望

研究生课程

研究生课程 (15 学分)

- 必修：算法分析和复杂性理论（3）、自然辩证法概论（1）、国际英语视听说（2）
- 选修：语义计算与知识检索（3）、计算语言学（3）、模式识别（3）

Courses

- 计算语言学
  - 介绍基础的自然语言处理模型和方法，适合入门，内容不是很新，有相关基础的可能在课堂上学不到太多新的知识
  - 大作业：复述检测（77.27% 准确率，应该是班里最好的）
    - 方法：各种特征暴力叠加 + 分类器投票；state-of-art:77.6%
- 语义计算与知识检索
  - 信息检索有关的知识，可以在实验中提供不少思路和小经验
  - 大作业：中文问答（模型简单，效果不好...)；小作业：豆瓣电影评分
- 模式识别
  - 智能系任选，比较偏理论推导，没有基础，听得比较吃力。
  - 了解了一些神经网络在自然语言中的应用

Courses

- 个人总结
  - 各门课难度不一
    - 算法分析、模式识别难度较大，需要花一些时间
    - 全校必修、计算语言学很水，求过的话不用花什么时间...
  - 作业模式有了不少转变
    - 作业：查论文 -> 实现论文

实验室工作

- 新华社：新闻展示网站、新闻标注网站
- 基于中文维基百科的知识库
- 其余工作

Lab

- 新华社
  - 新闻展示网站
    - 新闻中的实体链接到中文维基百科知识库
    - 展示新闻中实体之间的关系
  - 新闻标注网站
    - 和曾颖一起负责。将组里抓取的各个网站的新闻汇集起来，交由志愿者进行 ccnc 类别标注
    - 我：生成新闻数据，半自动检测志愿者标注效果
    - zy：搭建新闻标注网站。
    - 进度：进行中。已标注 7000 篇，7 名志愿者

Lab

- 基于中文维基百科的知识库
  - 学期前进展（毕设工作）
    - 1100 多万三元组（830 多万高质量三元组）

低质量三元组:【中国：职位 1：国家主席】【中国：人物 1：习近平】

  - 问题/不足
    - 和百度百科的实体链接基于名称（字符串），且百度百科三元组较少
    - 未和其他中文知识库链接（但是已和英文维基百科知识库链接）
    - **谓词很多非中文（一半以上）：本学期工作**

```
{{Infobox NBA Player
| name       = 谢尔曼·道格拉斯<br/>{{lang|en|Sherman Douglas}}
| number      = 4&#x2c;11&#x2c;20{{#tag:ref||name=nbastat}}
| position    = 控球后卫{{#tag:ref|[http://stat-nba.com/player/820.html 谢尔曼·道格拉斯] - 数据NBA|name=nbastat}}
| birth_date  = {{Birth date and age|year=1966|month=9|day=15}}
```

Lab

- 基于中文维基百科的知识库
  - 本学期工作中心
    - 中文维基百科谓词标准化/归一化。

建筑商、建筑单位、承建商、主承建商、建商、设计建设团队、建造所

    - 启发式规则筛掉错误三元组/谓词；抽取谓词的特征，设计分类器判断任意两个谓词是否描述相同实体
    - 目标：生成一个内部统一的谓词体系（schema），合并其他知识库
  - 本学期工作（已完成）
    - 基于中文维基百科网页进行抽取
    - 650 万高质量三元组，谓词符合直观
    - 谓词规模：2.6w -> 1.5w（筛除错误三元组）
    - 基于网页抽取的三元组和基于 wikitext 抽取的三元组对应（启发规则）
    - 已抽特征：对应的 wikitext 的内容、上级标题、单词特征...
    - 制作了一个版本给彭宇新老师组使用

Lab

- 基于中文维基百科的知识库（进行中）
  - 模仿新闻标注网站写了一个谓词的标注网页，手工标注，排除错误谓词 (8k/15k)
  - 基于一些启发式规则标注训练数据（谓词二元组）
  - 抽取出谓词的相关特征，进行测试和改进

Id	Content	Frequencey	Cooccur	markedTag	examp
6701	发展为	19	1	0	普惠_F119:发展为:普惠 F135
6702	2	27	0	2	陈姓:2:冯
6703	学年制	3	1	0	中山火炬职业技术学院:学年制:全日制三年
6704	院系	13	1	0	中山火炬职业技术学院:院系:包装印刷系
6705	校刊	16	1	0	中山火炬职业技术学院:校刊:中山火炬学刊
6706	议会席次	29	1	4	溫和黨:议会席次:107/349
6707	省议会席次	4	1	4	溫和黨:省议会席次:376/1,656
6708	自治市议会席次	4	1	4	溫和黨:自治市议会席次:2,966/12,978
6709	dd-mm-yyyy	1	0	2	各地日期和时间表示法:dd-mm-yyyy:dd-mm-yyyy 和 yyyy-mm-dd

Lab

- 其余工作
  - 中英文所有的维基百科页面、dump 都已经抓下来了，大家可以使用
  - 帮助许坤和张晟做了一点和英文维基百科有关的抓取任务。
  - 参与冯老师小组的论文讨论班

总结和展望

“

latex，paper，计划/备忘录，托福...

Summary

- 优点
  - 养成使用 latex 的习惯了：所有的报告和大部分作业
  - 看了一些论文：仔细看过 20 多篇论文
    - 起因：写大作业、讨论班、找实验方法...
    - 速度提高：半天 ->2 个小时
- 失败
  - 每周两篇 paper
  - 每天的备忘录、背托福单词等诸多问题
- 缺点/不足
  - 做事缺少计划性
    - 开学的时候谓词任务做了不少，快期中的时候就突然不想做了，直到前一周才做

Summary

- 下学期展望
  - 背单词，考托福
  - 结束研究生课程内容
  - 完成中文知识库谓词归一，链接其余知识库
- （同之前的失败项）
  - 每周两篇 paper
  - 每天的备忘录、背托福单词等诸多问题

谢谢大家！